Jogo Computacional sobre Dobramento de Proteínas: Minerando regras com auxílio da Inteligência Humana

Renan M. Luz[1], **Diana F. Adamatti**[1], **Adriano V. Werhli**[1]

[1] C3 – Universidade Federal do Rio Grande (FURG)
Caixa Postal 15.064 – 96.203-900 – Rio Grande – RS – Brazil

renanml@hotmail.com, {dianaada,werhli}@gmail.com

Abstract. *Proteins play a fundamental role in the nature and the discovery of their functionalities and behaviors, raise the interest in several areas. Therefore, this study aims to search for knowledge of protein folding techniques through human intelligence using protein structures in the HP model. Data were acquired through a serious game where they were tested by two statistical data mining algorithms to extract possible rules: decision trees and Naıve Bayes. With this study, some rules were obtained and we concluded that it is possible to acquire protein folding rules through human intelligence.*

Resumo. *As proteínas desempenham um papel fundamental na natureza e a descoberta de suas funcionalidades e seus comportamentos, despertam muito o interesse de diversas áreas. Portanto, este trabalho tem como objetivo a busca por conhecimentos de técnicas de dobramento de proteínas através da inteligência humana utilizando estruturas de proteínas no modelo HP. Foram adquiridos dados através de um jogo sério onde foram submetidos a algoritmos estatísticos de mineração de dados para extrair possíveis regras, são eles, Árvore de decisão e Naïve Bayes. Com este trabalho, foram obtidas algumas regras e chegou-se a conclusão que é possível adquirir regras de dobramento de proteínas através da inteligência humana.*

1. Introdução

As proteínas desempenham um papel fundamental na natureza e essas estruturas, compostas por 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 interesse em áreas envolvidas como a biologia, empresas de fabricação de medicamentos e até mesmo a produção de outras proteínas [Ptitsyn 1996]. Segundo Cooper et al. (2010b) ainda existe uma vasta gama de problemas ainda não resolvidos a respeito da predição dessas estruturas, soluções que são essenciais para a vida do ser humano.

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], disponibilizando aos jogadores, estruturas completas de proteínas em 3D. A inteligência humana, em um âmbito científico, é a capacidade de seres humanos de raciocinar, planejar, solucionar problemas e abstrair ideias por conta própria [Miranda 2002] sem mesmo ter conhecimento do problema científico a sua frente e a real solução do mesmo, sendo

assim, até mesmo pessoas comuns conseguem gerar e elaborar respostas de problemas ainda não solucionados pela ciência.

Devido as dificuldades de estudar as complexas estruturas das proteínas com seus inúmeros aminoácidos, foi criado por Dill (1985) o modelo HP (Hidrofóbico-Polar), que simplifica as estruturas em um modelo 2D apresentando-as apenas em forma de sequência binária de aminoácidos, com isso, tornando as estruturas mais simples e facilitando sua predição.

Tendo em vista a importância do assunto abordado e seus resultados positivos do uso da inteligência humana para buscar soluções de dobramento de proteína com suas estruturas completas, este trabalho visa utilizar a inteligência humana para tentar adquirir conhecimentos e regras de dobramentos de proteínas utilizando estruturas simplificadas no modelo HP proposto por Dill (1985) através de um jogo sério. Para extrair estratégias da inteligência humana, foram utilizadas técnicas de mineração de dados alimentadas com as jogadas adquiridas através do jogo. Foram escolhidas duas técnicas que ao longo do projeto se mostraram funcionais, são elas: árvores de decisão e Naïve Bayes.

2. Referencial Teórico

2.1. 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 uma forma tridimensional específica, sendo capazes de realizar a sua função biológica [Baker 2000, Chandru et al. 2000, Dinner et al. 2000, Plotkin and 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 mal dobradas, que não desempenham corretamente sua função no organismo [Baker 2000, Chandru et al. 2000, Dinner et al. 2000, Plotkin and Onuchic 2000, Trovato et al. 2005].

2.2. 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].

Apesar da simplicidade do modelo HP, o processo de dobramento tem 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].

Tabela 1. Sequências HP compiladas [Hart and 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	$H_2P_2HP_2HP_2HP_2HP_2HP_2H_2$	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	$H_2PHPHPHPH_4PHP_3HP_3HP_4HP_3HP_3HPH_4PHPHPHPHP_2$	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

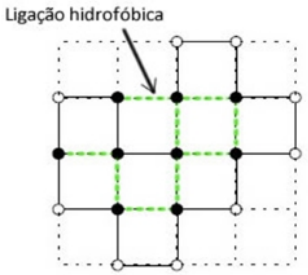

Figura 1. Ligação hidrofóbica em uma estrutura 2D modelo HP.

A energia de uma estrutura no modelo HP depende do número e posição de aminoácidos hidrofóbicos que ocupam os pontos de uma grade onde a estrutura se encontra. A Figura 1 mostra um modelo HP 2D em que a estrutura da proteína ilustrada contém 18 aminoácidos simplificados, que ocupam os pontos da grade, sendo que 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 simbolizadas pelos traços verdes [Dill 1985, Dill et al. 1995].

Benchmark 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 1 pode-se ver a quantidade de resíduos e a menor energia encontrada para cada sequência [Hart and Istrail 2012].

2.3. Jogos sérios

Atualmente, existe uma grande tendência no âmbito científico de criar projetos de pesquisa que utilizam jogos computacionais como ferramenta para se obter resultados ci-

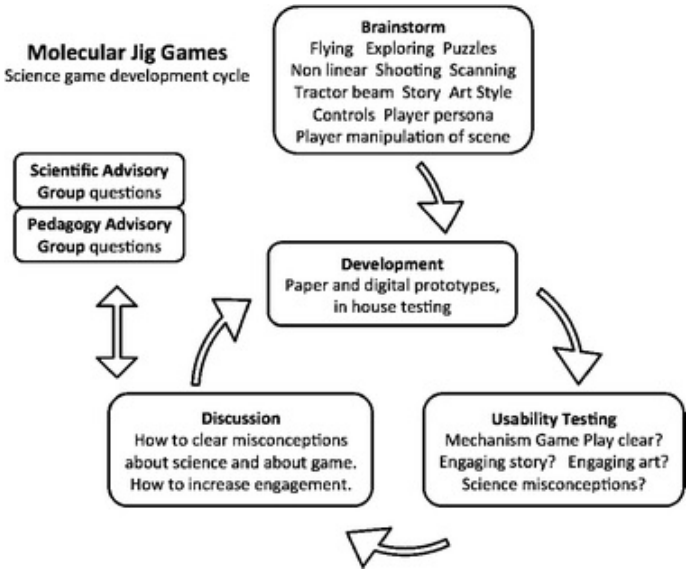

Figura 2. Template do cronograma de produção do jogo Molecular Jig. Otimizado para envolver cientistas e professores em todos os processos de desenvolvimento de jogos [Montes 2014].

entíficos. Um jogo sério transforma problemas científicos em quebra-cabeças e os fornece em um mecanismo de jogo, fazendo com que jogadores não-especialistas resolvam estes problemas [Allé 1999, Cooper et al. 2010b, Burnett et al. 2016].

Dr. Montes, criador do jogo sério Molecular Jig[1] compartilhou na discussão Montes (2014) um pouco de sua experiência em jogos sérios dizendo que a criação de um jogo de computador de sucesso, muitas vezes requer tempo e esforço considerável, tanto no desenvolvimento dos conceitos, bem como na implementação e explicou também o processo de desenvolvimento utilizado por ele para desenvolver o jogo Molecular Jig.

No início, elabora os objetivos de aprendizagem e escolhe um mecanismo de jogo. Para ensinar de forma mais eficaz, escolhe um mecanismo de jogo que vai exigir aos estudantes utilizar fatos e conceitos a serem aprendidos, a fim de ganhar o jogo. Uma vez que os objetivos de aprendizagem e mecanismo de jogo são escolhidos, são criados protótipos e testados pela equipe e em seguida com membros do público-alvo. Neste processo são realizadas observações e acompanhamento de dúvidas com os membros para avaliar o mecanismo de jogo. Caso os jogadores não conseguirem identificar o objetivo do jogo por conta própria e não obter condições de vitória no ambiente, então, o mecanismo é muito complexo ou mal elaborado. Durante cada ciclo de iteração é realizada consultas com cientistas e educadores a respeito da experiência adquirida.

2.4. 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 a grande quantidade desses dados. Portanto, novas e mais complexas estruturas de armazenamento foram e estão sendo desenvolvidas, tais como: banco de dados, data warehouses e bibliotecas virtuais [Cios et al. 2007, Han and Kamber 2006, Larose 2005].

[1] http://www.molecularjig.com/

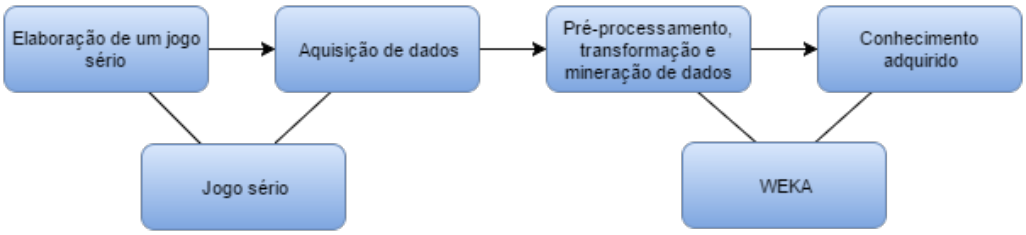

Figura 3. Fluxograma da metodologia aplicada neste projeto

Segundo CABENA et al. (1998), de uma perspectiva de banco de dados, a mineração de dados é um campo interdisciplinar que une técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatística, banco de dados e visualização, para conseguir extrair informações de grandes bases de dados. Os principais objetivos das práticas de mineração de dados são 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 passíveis de interpretação pelos humanos. Os objetivos da predição e descrição podem ser alcançados usando uma variedade de métodos de mineração de dados [Fayyad et al. 1996]. Desse trabalho, os métodos de mineração de dados utilizados são:

- Classificação por árvore de decisão, que funciona como um fluxograma em forma de árvore, onde cada nó indica um teste sobre o valor. Esse método tem com o objetivo reduzir a impureza ou incerteza dos dados o máximo possível [Hall et al. 2009].
- Classificação Bayesiana, que faz uso de fórmulas estatísticas e cálculo de probabilidades para realizar a classificação. Assim, possibilita aquisição de conhecimento das probabilidades em uma base de dados, levando em consideração que os atributos fornecidos serão igualmente importantes e estatisticamente independentes. Desta forma, o valor de um atributo não influencia no valor de outro [Mitchell 1997].

Estes métodos foram escolhidos por serem de fácil entendimento para análise de resultados, e por apresentarem dados quantitativos como a matriz de confusão e dados qualitativos como caminhos da árvore e regras no Bayes. Fora isso, são técnicas com bom desempenho para grandes quantidades de instâncias.

3. Materiais e Métodos

Na Figura 3 pode-se observar graficamente o fluxograma dos processos realizados ao longo do projeto assim como os materiais e ferramentas utilizadas para a obtenção dos resultados.

3.1. Elaboração de um jogo sério biológico

Para coletar informações da inteligência humana, uma alternativa existente é 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

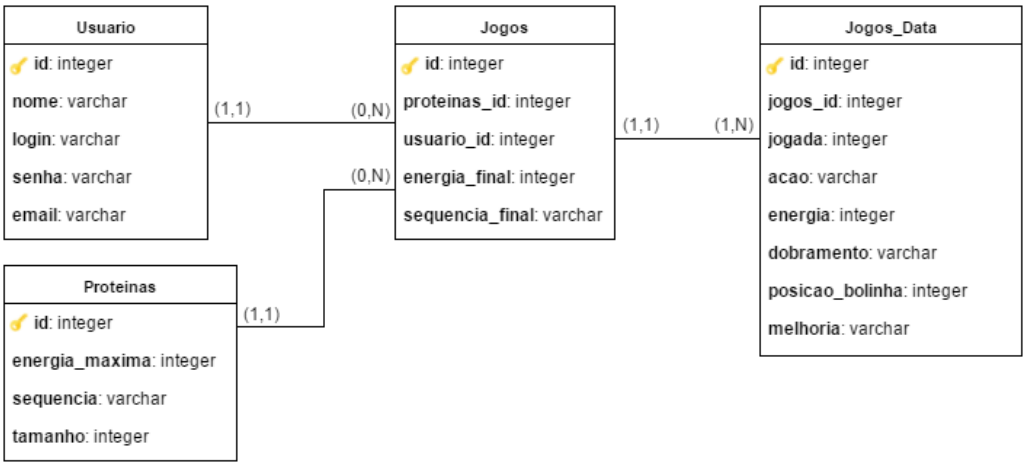

Figura 4. Modelo de relacionamento do banco de dados do jogo de dobramento de proteína

o interesse dos jogadores para que se possa obter um grande número de dobramentos fornecidos pelos mesmos. Para isso, foram utilizadas linguagens de programação web, fazendo com que se tenha um acesso facilitado.

3.1.1. Implementação

Para implementar a interface gráfica foram utilizado HTML5 juntamente com CSS, que são duas das principais tecnologias para a construção de páginas web, possibilitando confeccionar o site e a interface do jogo, fazendo com que funcione em qualquer navegador e que possa ser hospedado em qualquer servidor web.

Também foram utilizadas 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; 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.

3.1.2. Banco de dados

Na Figura 4 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.

- **Tabela Usuario:** utilizada para o armazenado dos dados de cadastro dos jogadores no site. 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 estruturas das proteínas da tabela 1 que são fornecidas aos jogadores. Cada estrutura possui seu tamanho, sua sequência HP e sua energia máxima seguindo a metodologia das estruturas. Esta Tabela contém um

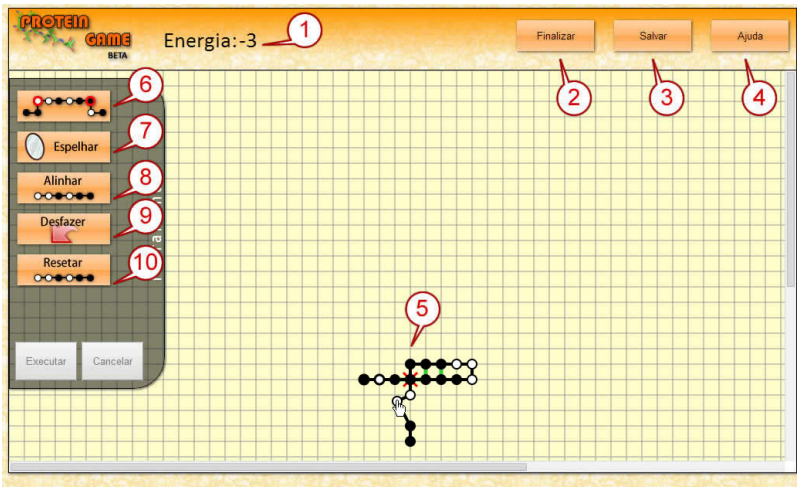

Figura 5. Interface da tela do jogo com a uma proteína de estrutura HP após sofrer alguns dobramentos feitos pelo jogador

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 linha 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 realizaram em seus jogos. Com isso é possível saber qual o número da jogada, a ação feita nesta jogada, a energia atingida calculada de acordo com a metodologia do modelo HP, 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”, que serão utilizadas nas técnicas de classificação. Portanto são os dados armazenados nesta tabela, que serão utilizados nas técnicas de mineração de dados.

3.1.3. Interface

Como foi dito no início desta seção, para se obter um grande volume de entradas no banco de dados, é necessário despertar o interesse dos jogadores facilitando o aprendizado e o entendimento de vitória, assim como, despertando nos jogadores a vontade de retornar ao jogo e realizar novas partidas. Portanto, foi desenvolvida uma interface simples e amigável como mostra a figura 5 que é obtida após o jogador selecionar a proteína que deseja.

1. Quantidade de energia em que a estrutura se encontra
2. Botão para finalizar o jogo
3. Botão para salvar a partida caso queira continua-la em outro momento
4. Botão de ajuda que mostra uma documentação que auxilia o aprendizado da jogabilidade

5. A estrutura da proteína em modelo HP esticada, que é o estado inicial da estrutura dentro do jogo

   Do item 6 ao 10, são ferramentas que foram criadas para auxiliar o jogador possibilitando efetuar uma série de dobramentos mais rapidamente:

6. Possibilita ao jogador selecionar um intervalo de aminoácidos e efetuar um dobramento igual ao da figura 6a
7. Efetua um espelhamento na estrutura da proteína, como mostra a figura 6b
8. Possibilita o jogador a selecionar um intervalo de aminoácidos fazendo com que todos neste intervalo fiquem alinhados, como mostra a Figura 6c
9. Caso o jogador não goste da última jogada feita, essa ferramenta possibilita voltar a estrutura ao seu estado anterior
10. Faz com que a estrutura da proteína volte ao seu estado inicial como mostra a Figura 6d.

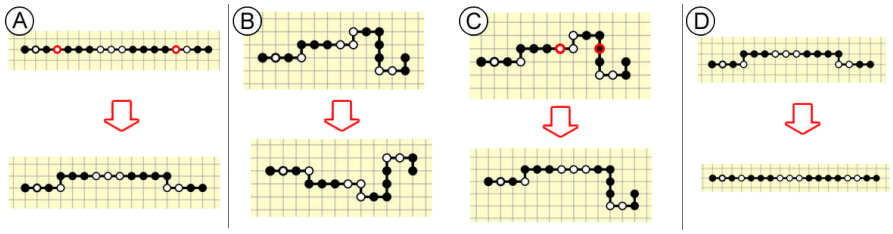

Figura 6. Exemplos do uso das ferramentas oferecidas no ambiente do jogo de dobramento de proteína

3.2. Aquisição dos Dados

Após os últimos ajustes na implementação do jogo, foram realizados testes com alunos de graduação em aulas de professores que cederam uma parte do seu horário. Primeiramente foi realizada uma breve introdução para os alunos a respeito do projeto e algumas explicações das funcionalidades do jogo. Foi priorizado nas aplicações do jogo, as proteínas S1 (HPHPHHHPPPHHHHPPHH) e S2 (PHPPHPHHHPHHPHHHHH) para que fosse possível obter um volume maior de dados nessas duas estruturas. Essas foram escolhidas, devido a suas menores quantidades de aminoácidos que as demais, com o objetivo de se obter um maior número de partidas de uma mesma proteína. Ao todo, foi possível aplicar o jogo em duas aulas, com uma duração de 1 hora cada, contendo um total de aproximadamente 40 alunos.

No geral, 44 jogadores foram registrados no banco de dados, obtendo um valor de 401 jogos criados, totalizando 12.059 entradas na tabela Jogos_Data.

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 formas de lapidar as entradas e submete-las novamente aos testes, até que se consiga adquirir resultados estatísticos satisfatórios.

Foi utilizado o software WEKA [2], que é gratuito, desenvolvido pela Universidade de Waikato na Nova Zelândia, onde pode-se facilmente utilizar técnicas de mineração de dados, apenas fornecendo as entradas com os dados através de arquivos de diferentes formatos. Após a leitura das entradas é possível selecionar as colunas que o usuário deseja trabalhar, e através de abas é possível selecionar o tipo de técnica de mineração de dados e também diversos tipos de algoritmos para cada técnica. Os algoritmos de mineração de dados utilizados no presente trabalho foi J48, que efetua a técnica de mineração de dados chamada Classificação por árvore de decisão [Hall et al. 2009]; e Naïve Bayes, que faz uso de fórmulas estatísticas e cálculo de probabilidades para realizar a classificação [Mitchell 1997]. Estes algoritmos foram escolhidos pois apresentam uma resposta de fácil predição.

Os dados adquiridos através do jogo foram exportados do banco de dados em um arquivo CSV. 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 estão os atributos separados por vírgula e que serão interpretados pelo WEKA. Cada linha restante é um dobramento que a estrutura sofreu por intermédio do jogo.

Cada coluna do arquivo é interpretada pelo WEKA como um atributo. Os atributos utilizados foram:

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 , Z(Ferramenta que efetua o espelhamento da estrutura), L(Ferramenta que efetua o alinhamento do intervalo de aminoácidos selecionados) e a R(Ferramenta que reseta a estrutura deixando em sua posição inicial);
5. **Posicao:** Contém a posição do aminoácido que foi efetuado o dobramento (ordem na sequência);
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.

3.4. Características dos testes

Ao longo do projeto, foi efetuado diversos testes, devido ao maior volume de dados, foram utilizados os dados referentes a proteína S1 (HPHPHHHPPPHHHHPPHH). Nos testes feitos anteriormente, foi descoberto que as jogadas ”zero”e ”um”podem ser excluídas, porque a jogada zero é a posição inicial e nunca se altera, e também a jogada um, se efetuada, sempre deixa a estrutura apenas em um ângulo de 90 graus sem causar alteração na energia da proteína. Também foram ignoradas todas as jogadas de melhoria = igual, de forma a balancear a quantidade de instâncias da base de conhecimento utilizada devido ao seu elevado número de instâncias quando comparados com as melhorias ”melhorou”e

[2]http://www.cs.waikato.ac.nz/ml/weka/index.html

Tabela 2. Matriz de confusão J48 versão 1

	A	B
A = melhorou	**851**	247
B = piorou	377	**465**

"piorou". Portanto, os testes tinham um total de 1.940 instâncias e tiveram como seu atributo classificador o atributo melhoria.

Utilizando o algoritmo J48, também se definiu o número mínimo por folha de 19 instâncias por folha (uma taxa de 1% do total de instâncias utilizadas), pois nos testes iniciais chegou-se a conclusão de que este número deixa o tamanho da árvore e as taxas de acerto balanceadas.

Para utilizar o algoritmo J48 as instâncias foram classificadas de duas formas, contendo diferentes atributos, e para o algoritmo Naïve Bayes foram mantido os atributos da versão dois, devido a irrelevância do atributo jogada que se obteve nos resultados dos primeiros testes. Assim, foram realizados 3 conjuntos de testes, com os seguintes atributos:

1. J48 Versão 1: jogada, movimento, posição e melhoria.
2. J48 Versão 2: movimento, posição e melhoria.
3. Naïve Bayes: movimento, posição e melhoria.

3.5. Resultados Obtidos

Dos 3 conjuntos de dados, foram obtidos os seguintes resultados:

3.5.1. J48 Versão 1

Neste teste foi obtida a árvore classificatória apresentada na Figura 7, que contém uma taxa de 67.83% de acerto, e uma matriz de confusão como mostra a Tabela 2.

Conforme a Tabela 2, pode-se observar que:

- As instâncias com melhoria = melhorou tiveram 851 instâncias classificadas corretamente;
- As instâncias com melhoria = piorou tiveram 465 instâncias classificadas corretamente.

Pode-se observar que na Figura 7 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 piora; **Caminho Azul:** Este caminho mostra que efetuando um movimento D(direita) nos aminoácidos ≤ 27 a energia da estrutura melhora; **Caminho Laranja:** Este caminho mostra que a energia melhora quando o jogado utiliza a ferramenta U(voltar).

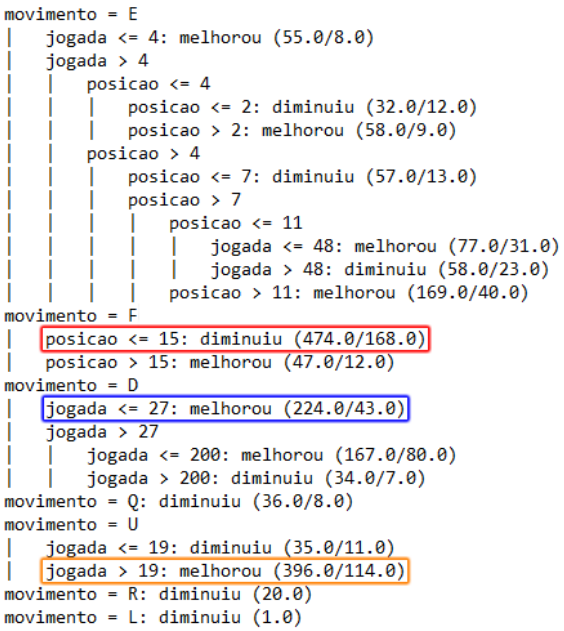

Figura 7. Árvore de classificação gerada na versão 1

Este teste obteve uma taxa de acerto razoavelmente boa e pode-se observar que a matriz de confusão (Tabela 2) obteve uma boa igualdade nas classificações melhorou e piorou.

3.5.2. J48 Versão 2

Neste teste foi obtida a árvore classificatória apresentada na Figura 8, que contém uma taxa de 68.29% de acerto, e uma matriz de confusão como mostra a Tabela 3.

Conforme a Tabela 3, pode-se observar que:

- As instâncias com melhoria = melhorou obteve 887 instâncias classificadas corretamente;
- As instâncias com melhoria = piorou obteve 438 instâncias classificadas corretamente.

Pode-se observar que na Figura 8 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 piora; **Caminho Azul:** Estes caminhos mostram que efetuando um movimento E(esquerda) ou D(direita) nos aminoácidos > 6 a energia da estrutura melhora; **Caminho Laranja:** Este caminho mostra que a energia melhora quando o jogado utiliza a ferramenta U(voltar).

Este teste obteve uma taxa de acerto razoavelmente boa e pode-se observar que a matriz de confusão (Tabela 3) obteve uma boa igualdade das classificações melhorou e piorou.

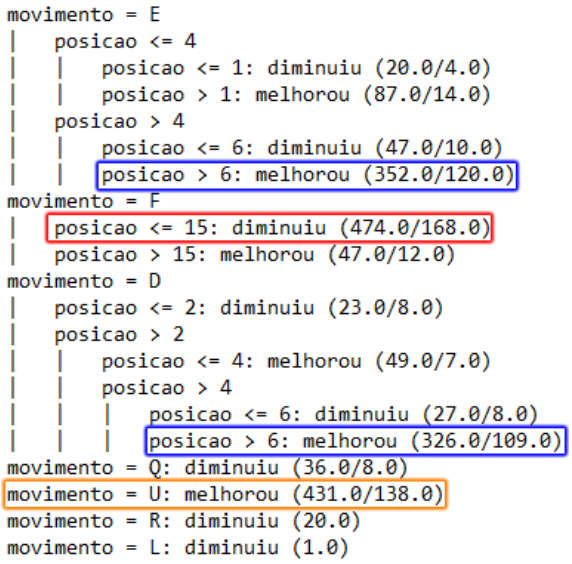

Figura 8. Árvore de classificação gerada na versão 2

Tabela 3. Matriz de confusão J48 versão 2

	A	B
A = melhorou	**887**	211
B = piorou	404	**438**

3.5.3. Naïve Bayes

Com o algoritmo Naïve Bayes foram obtidas classificações de probabilidades conforme a Figura 4, que contém uma taxa de 66.03% de acerto, sendo semelhante as taxas de acertos obtidas com o algoritmo J48. A matriz de confusão com esse algortimo é apresentada na Tabela 5.

Conforme a Tabela 5, pode-se observar que:

- As instâncias com melhoria = melhorou obtiveram 902 instâncias classificadas corretamente;
- As instâncias com melhoria = piorou obtiveram 379 instâncias classificadas corretamente.

Pode-se observar pela Tabela 4 que existem índices de probabilidade de cada tipo de dobramento efetuado, podendo ser: E(Esquerda), D(Direita), U(Ferramenta Desfazer), Q(Ferramenta que efetua o dobramento mostrado na Figura 6a), Z(Ferramenta que efetua o espelhamento da estrutura como mostrado na Figura 6b), L(Ferramenta que efetua o alinhamento do intervalo de aminoácidos selecionados como mostrado na Figura 6c) e a L(Ferramenta que reseta a estrutura deixando em sua posição inicial como mostrado na Figura 6d).

Nota-se que foram obtidas 1105 jogadas que melhoraram a energia da proteína e que em 849 jogadas, a energia piorou. Também foram extraídos os seguintes conhecimentos:

- Efetuando um dobramento para esquerda(E), de 508 jogadas, em 320 a energia melhorou (sendo 62,99 % das instâncias);
- Mantendo a estrutura mais esticada(F), de 523 jogadas, em 319 a energia piorou (sendo 60,99 % das instâncias);

Tabela 4. Tabela de probabilidades Naïve Bayes

Movimento	Melhorou	Piorou
E	**320**	188
F	204	**319**
D	276	151
Q	09	29
U	294	139
R	01	21
L	01	02
Total	**1105**	**849**

Este teste também obteve uma taxa de acerto razoavelmente bom e pode-se observar que a matriz de confusão (Tabela 5) obteve resultado semelhante das matrizes de confusões obtidas com o algoritmo J48.

3.6. Considerações finais

Com os resultados adquiridos nos testes, foram obtidas algumas estratégias de dobramento de proteína e devido à obtenção de uma matriz de confusão e taxas de certos semelhantes com dois algoritmos utilizadas, pode-se afirmar que a base de dados se mostra válida e coerente. As regras de dobramento de proteína mais relevantes obtidas através dos testes foram:

- Com o movimento F(frente) nos aminoácidos ≥ 15, a energia da estrutura piora.
- Efetuando um movimento E(esquerda) nos aminoácidos maior que 6, a energia da estrutura melhora.
- A energia melhora quando o jogado utiliza a ferramenta U(voltar).

É possível analisar que se obteve 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 ocorre piora na energia. Assim, os jogadores desfazem sua jogada, voltando a posição anterior, onde a estrutura encontra-se com uma energia melhor.

Foi observado também que, por algum motivo, as classificações melhoria = piorou contém muitas classificações erradas, como mostram as matrizes de confusão dos testes. Ainda não se tem uma resposta relacionada a essa questão, que deve ser mais bem investigada.

4. Conclusões e Trabalhos Futuros

Existem muitos comportamentos ainda não desvendados a respeito das proteínas, e sempre será importante buscar respostas às perguntas sobre este assunto. Devido às limitações

Tabela 5. Matriz de confusão Naïve Bayes

	A	B
A = melhorou	**902**	196
B = piorou	463	**379**

tecnológicas existentes, ainda não se consegue desvendar alguns destes problemas. Assim, muitos cientistas estão em busca de conhecimentos sobre o assunto, como a extração de conhecimentos sobre dobramento de proteína com a ajuda da inteligência humana, onde já existem resultados positivos.

Os trabalhos relacionados buscaram conhecimentos com estruturas de proteínas mais complexas, deixando suas predições também complexas. Portanto este trabalho buscou realizar a predição de estruturas de uma maneira mais simplificada através do modelo HP, mas bem próximo da realidade.

Analisando os resultados obtidos através dos testes realizados no 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 apresentarem 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 de conhecimentos para esse tipo de aplicação, mas ainda existe uma vasta gama de algoritmos a serem explorados e submetidos a testes.

Desta forma, como trabalhos futuros, pretende-se obter um maior volume de dados para a base de dados (realizar mais jogos); melhorar as ferramentas fornecidas para os jogadores, facilitando a jogabilidade; tornar o jogo mais dinâmico e multiplayer (capaz de deixar 2 ou mais jogadores interagirem juntos na mesma partida em tempo real); e aplicar os dados obtidos em outras técnicas de mineração de dados, buscando assim , descobrir novas regras.

Também se espera, a partir de boas estratégias de dobramento, criar um jogador artificial, levando a obtenção de possíveis dobramentos, ainda não foram descobertos.

Referências

Allé, J. M. (1999). *O Grande Livro dos Jogos*. Leitura.

Baker, D. (2000). A surprising simplicity to protein folding. *Nature*, 405(6782):39–42.

Burnett, S., Furlong, M., Melvin, P. G., and Singiser, R. (2016). Games that enlist collective intelligence to solve complex scientific problems. *Journal of Microbiology e Biology Education*, 17(1):133–136.

Chandru, V., DattaSharma, A., and Anil Kumar, V. (2000). A surprising simplicity to protein folding. *Discrete Applied Mathematics*, 405(127):145–161.

Cios, K. J., Pedrycz, W., Swiniarski, R. W., and Kurgan, L. A. (2007). *Data Mining - A Knowledge Discovery Approach*. Springer.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popovic, Z., and players, F. (2010a). Predicting protein structures with a multiplayer online game. *Nature*, 466:756–760.

Cooper, S., Treuille, A., Barbero, J., Leaver-Fay, A., Tuite, K., Khatib, F., Snyder, A. C., Beenen, M., Salesin, D., Baker, D., Popovic, Z., and players, F. (2010b). The challenge of designing scientific discovery games. *Foundations of Digital Games*, pages 40–47.

Dill, K., Bromberg, S., Yue, K., Fiebig, K., Yee, D., Thomas, P., and e Chan, H. (1995). Principles of protein folding - a perspective from simple exact models. *Protein science*, 4:561–602.

Dill, K. A. (1985). Theory for the folding and stability of globular proteinst. *Biochemistry*, 24:1501–1509.

Dinner, A., Sali, A., Smith, L., Dobson, C., and Karplus, M. (2000). Understanding protein folding via free-energy surfaces from theory and experiment. *Trends in biochemical sciences*, 25:331–339.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). *From Data Mining to Knowledge Discovery in Databases*. American Association for Artificial Intelligence.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). *The weka data mining software: An update*. SIGKDD Explorations.

Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques*. Elsevier.

Hart, W. and Istrail, S. (2012). Hp benchmarks. Sandia Web Site.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley and Sons*. John Wiley and Sons, Inc.

Miranda, M. J. (2002). A inteligência humana: contornos da pesquisa. *Paidéia (Ribeirão Preto)*, 12(23):19–29.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math.

Montes (2014). Applications and serious games: from docking to protein folding: general discussion. *Faraday Discuss.*, 169:501–519.

Plotkin, S. and Onuchic, J. (2000). Investigation of routes and funnels in protein folding by free energy functional methods. *Proceedings of the National Academy of Sciences*, 97:6509–6514.

Ptitsyn, O. B. (1996). A determinable but unresolved problem. *The FASEB Journal*, 10:3–4.

Trovato, A., Hoang, T., Banavar, J., Maritan, A., and Seno, F. (2005). What determines the structures of native folds of proteins? *Journal of Physics: Condensed Matter*, 17:1515–1522.