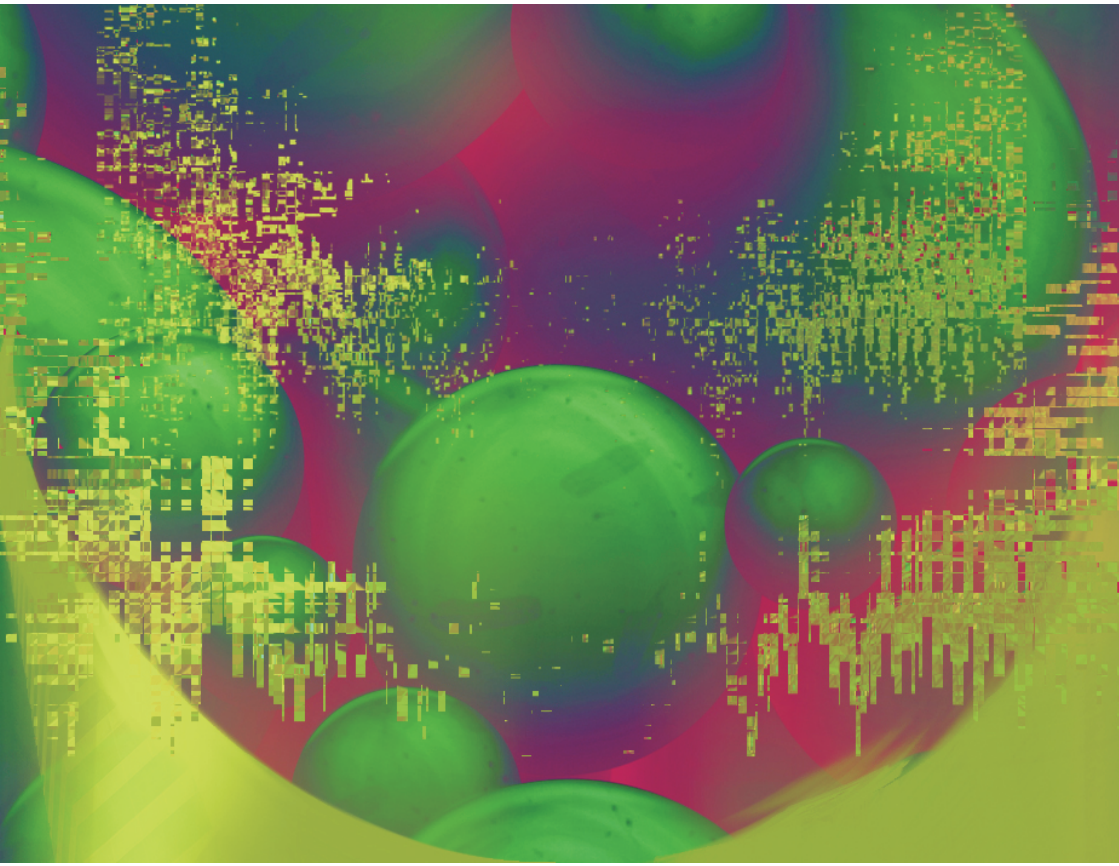
Boletim de Pesquisa 20
e Desenvolvimento dezembro, 2008

ISSN 1677-9266

Reconhecimento de padrões de pontes de hidrogênio – preliminares do desenvolvimento de uma metodologia baseada em TI para a predição da posição de átomos de hidrogênio em proteínas



Embrapa

ISSN 1677-9266
Dezembro, 2008

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura, Pecuária e Abastecimento

Cooperação Internacional da Embrapa e International Potash Institute

Boletim de Pesquisa e Desenvolvimento 20

Reconhecimento de padrões de pontes de hidrogênio – preliminares do desenvolvimento de uma metodologia baseada em TI para a predição da posição de átomos de hidrogênio em proteínas

Adauto Luiz Mancini
Roseli Aparecida Francelin Romero

Campinas, SP
2008

Embrapa Informática Agropecuária
Área de Comunicação e Negócios (ACN)
Av. André Tosello, 209
Cidade Universitária "Zeferino Vaz" – Barão Geraldo
Caixa Postal 6041
13083-970 – Campinas, SP
Telefone (19) 3211-5700 – Fax (19) 3211-5754
URL: http://www.cnptia.embrapa.br
e-mail: sac@cnptia.embrapa.br

Supervisão editorial: *Suzilei Almeida Carneiro*
Normalização bibliográfica: *Marcia Izabel Fugisawa Souza*
Revisão de texto: *Adriana Farah Gonzalez*
Editoração eletrônica: *Área de Comunicação e Negócios (ACN)*

1ª. edição on-line - 2008

Mancini, Adauto Luiz

Reconhecimento de padrões de pontes de hidrogênio – preliminares do desenvolvimento de uma metodologia baseada em TI para a predição da posição de átomos de hidrogênio em proteínas / Adauto Luiz Mancini. - Campinas : Embrapa Informática Agropecuária, 2008.
20 p. : il. – (Boletim de pesquisa e desenvolvimento / Embrapa Informática Agropecuária ; 20).

ISSN 1677-9266

1. Reconhecimento de padrões. 2. Pontes de hidrogênio. 3. Predição da posição de hidrogênio. 3. Bioinformática. I. Título. II. Série.

CDD – (21st .ed.)
006.4
570.285

Embrapa 2008

Sumário

Reconhecimento de padrões de pontes de hidrogênio – preliminares do desenvolvimento de uma metodologia baseada em TI para a predição da posição de átomos de hidrogênio em proteínas

Adauto Luiz Mancini [1]
Roseli Aparecida Francelin Romero [2]

Resumo

Os primeiros resultados do desenvolvimento de um novo método para a localização do átomo de hidrogênio contido em grupos hidroxila da cadeia lateral dos aminoácidos é apresentado neste artigo. Os métodos existentes utilizam campos de força para esse problema de localização. Os autores propõem uma abordagem computacional para esse problema, pelo reconhecimento de padrões de pontes de hidrogênio agrupados por similaridade em clusters. Os resultados das primeiras tentativas foram ruins e ajustes ao método foram necessários. Os novos experimentos foram bem sucedidos e mostram que o desenvolvimento da metodologia deve ser continuado. A pesquisa está em andamento e os próximos desafios para viabilizar o método são: dado um padrão de entrada composto de um grupo hidroxila (sem considerar a posição do hidrogênio) e seus possíveis átomos receptores contidos na vizinhança, determinar qual dos clusters previamente calculados contém dados mais similares ao padrão de entrada; estender o método para cadeias laterais contendo dois ou três átomos de hidrogênio com liberdade de rotação.

Termos para indexação: reconhecimento de padrões, cluster, rede neural, posição do átomo de hidrogênio.

[1] Mestre em Ciências da Computação, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP.
(e-mail: adauto@cnptia.embrapa.br).
[2] Phd em Ciências da Computação, Presidente da CPq do ICMC-USP
(e-mail: rafrance@icmc.sc.usp.br).

Initial results of a new methodology for predicting the position of hydrogen atoms in proteins

Abstract

The first results of the developing of a new method for the location of hydrogen atoms contained in the hydroxyl groups in the side chain of amino acids is presented in this article. The existing methods use force fields to the problem of location. The authors propose a computational approach to this problem, througth pattern recognition of hydrogen bonds grouped by similarity in clusters. The results of the first attempts were bad and adjustments were needed to the method. The new experiments were successful and show that the development of the methodology should be pursued. The search is ongoing and the future challenges are given a standard input composed of a hydroxyl group (without considering the position of hydrogen) and its possible receivers atoms in the neighborhood, determine which of the clusters containing previously calculated data is more similar to the pattern of entry; extend the method to side chains containing two or three hydrogen atoms with freedom of rotation.

Index terms: pattern recognition; cluster; neural networks; predicting hydrogen positions.

Introdução

O conhecimento correto da posição espacial dos átomos permite o cálculo, análise e quantificação de diversas propriedades físicas e espaciais das moléculas, como potencial eletrostático, pontes de hidrogênio, volume, superfícies e cavidades. O conhecimento dessas características permitem um melhor entendimento do funcionamento das proteínas e suas interações com outras moléculas, com aplicações na pesquisa de dobramento, atracamento e determinação de sítio ativo.

Muitas das estruturas macromoleculares mantêm sua conformação espacial por um grande número de diferentes interações não covalentes. Pontes de hidrogênio e outras interações envolvendo dipolos têm natureza direcional e assim ajudam a definir a forma macromolecular e a especificidade das interações moleculares (TURNER et al., 2000). Apesar de a energia de uma interação média de uma ponte de hidrogênio ser pequena (20kJ/mol) quando comparada a uma ligação covalente (200kJ/mol), o fato de haver uma grande ocorrência de pontes de hidrogênio tem uma forte influência no dobramento das proteínas. A ponte de hidrogênio é uma das interações interatômicas mais importantes no dobramento (folding) de proteínas, apesar de esse papel ser mais frequentemente atribuído à interação hidrofóbica. Os elementos da estrutura secundária de proteínas (hélices alfa e folhas beta) são formados essencialmente por padrões de pontes de hidrogênio. Consequentemente, pontes de hidrogênio têm sido extensivamente estudadas estatisticamente, experimentalmente e teoricamente (MCDONALD; THORNTON, 1994).

Para uma análise adequada das pontes de hidrogênio é desejável saber a posição dos átomos constituintes da ponte, porque são aplicados critérios geométricos sobre as posições dos átomos envolvidos na formação da ponte. Assim, conhecer as posições dos átomos de hidrogênio possibilita um estudo mais completo das forças que mantém a estabilidade protéica. Vários artigos sobre pontes de hidrogênio são encontrados na literatura científica nas áreas da química, física e biologia molecular (GANCIA et al., 2001; NIKOLIC et al., 2008). A ponte de hidrogênio, definida em sua forma tradicional, é uma interação entre dois átomos eletronegativos, doador e receptor, por um átomo de hidrogênio intermediário que é covalentemente ligado ao doador. A densidade eletrônica da ligação átomo doador - hidrogênio é deslocada para o átomo doador, assim uma carga positiva é induzida no átomo de hidrogênio. Essa carga parcial interage com a nuvem eletrônica do átomo receptor. Diferente da ligação covalente, a ponte de hidrogênio é uma interação multipolar envolvendo pelo menos três átomos (doador, receptor, hidrogênio) (TORSHIN et al., 2000).

Em proteínas, aproximadamente metade dos átomos são do tipo hidrogênio (ENGLER; PARAK, 2003). O grande avanço tecnológico das últimas décadas permitiu a determinação experimental da estrutura tridimensional das proteínas em laboratório. Apesar disso, temos uma grande deficiência de dados sobre a localização dos átomos de hidrogênio em termos experimentais. O método de cristalografia por difração de neutrons, que é o

mais adequado para a determinação dos átomos de hidrogênio, uma vez que registra a colisão do feixe de neutrons com os núcleos dos átomos, requer o uso de um reator nuclear sendo pouco acessível aos pesquisadores. O método RNM (ressonância nuclear magnética) determina a estrutura de proteínas em solução, mas é limitado às moléculas menores que 30 kD, sendo o método de escolha para proteínas pequenas que não são facilmente cristalizadas. Apenas 7.535 estruturas foram resolvidas por RNM, que consegue determinar a posição dos átomos de hidrogênio, de um total de 53.794 estruturas registradas no PDB em 21 de outubro de 2008. Usando cristalografia de raio X com alta resolução, temos 331 estruturas com resolução em Angstroms entre 0,5 -1,0 e 3709 estruturas com resolução de 1,0 – 1,5. Somente para estruturas com resolução superior a 1,2 Å é possível localizar a posição de átomos de hidrogênio usando difração de raio-X. Em função dessas dificuldades tecnológicas para a determinação da posição de átomos de hidrogênio, várias metodologias são disponíveis para posicionar hidrogênio em modelos de proteínas de átomos pesados, quando há ausência de informação experimental.

Algumas dessas metodologias são softwares individuais, como MCCE (GEORGESCU et al., 2002), enquanto outras são componentes de pacotes maiores, como CHARMM (BROOKS et al., 1983), CNS (BRÜNGER et al., 1998), GROMACS (LINDAHL et al., 2001), MolProbity (WORD et al., 1999), WHAT IF (VRIEND, 1990) e XPLOR (BRÜNGER, 1992). Essas ferramentas utilizam diferentes algoritmos para gerar modelos iniciais, e incorporam uma variação de funções de energia empíricas e fisicoquímicas, e em alguns casos refinamento adicional por protocolos de minimização de energia (FORREST; HONIG, 2005). Exemplos da caracterização de pontes de hidrogênio em um campo de força são apresentados por Fabiola et al. (2002) e Kortemme et al. (2003).

O ponto comum entre todas essas metodologias é o uso de algum campo de força. As fórmulas desenvolvidas para o campo de força produzem como resultado um valor (energia) que dependendo do objetivo pode representar a intensidade de um tipo de interação molecular como pontes de hidrogênio ou indicar uma medida da estabilidade da estrutura quando a molécula está em uma determinada conformação. Para o problema de posicionamento dos átomos de hidrogênio, diversas conformações contendo os átomos de hidrogênio são geradas por algum método e a partir do cálculo do campo de força para cada uma das conformações é escolhida a configuração considerada mais estável (menor energia). O processo pode ser feito em um único passo, mas geralmente é iterativo ajustando-se o melhor resultado obtido em cada passo. Em algumas situações, como no caso em que os hidrogênios estão ligados a um átomo de carbono, com uma distribuição tetraédrica de suas ligações, pode-se optar por posicionar os átomos de hidrogênio em critérios puramente geométricos, sem o uso do campo de força. A definição do campo de força é construída como um somatório de termos envolvendo restrições geométricas às ligações covalentes entre átomos (comprimento e ângulo entre ligações, ângulos diedrais entre partes da molécula, planaridade de anéis aromáticos) e termos envolvendo interações não covalentes (pontes de hidrogênio, eletrostática, van der Walls, hidrofobicidade, etc).

Este trabalho descreve um novo método para a predição de átomos de hidrogênio de grupos hidroxilas (OH) presentes na cadeia lateral dos aminoácidos serina, treonina ou tirosina. A predição é feita a partir do reconhecimento de padrões de pontes de hidrogênio. Os padrões são agrupados por similaridade, alinhados espacialmente e então usados para o treinamento de redes neurais.

A ideia principal é que o átomo de hidrogênio, ligado covalentemente ao átomo doador da ponte de hidrogênio, tem sua posição definida principalmente em função da presença de um (ou mais) átomo(s) receptor(es) da ponte. Podemos agrupar as ocorrências similares das pontes de hidrogênio encontradas nas proteínas e posteriormente alinhar espacialmente as instâncias contidas em cada grupo. Assim, uma rede neural poderia ser treinada para reconhecer os padrões de pontes de hidrogênio contidos no grupo e calcular a posição do hidrogênio.

Metodologia

O problema foi particionado em 3 etapas principais: extração dos dados sobre padrões de pontes de hidrogênio da base de dados PDB; agrupamento de padrões similares e treinamento da rede para previsão do hidrogênio.

Consideremos um padrão de ponte de hidrogênio como sendo o conjunto de átomos (A) - antecessor do doador, (D) - doador do hidrogênio, (H) - hidrogênio e possíveis receptores do átomo de hidrogênio, do tipo oxigênio (Oi) e nitrogênio (Ni) contidos em uma dada vizinhança. Vamos denominar de vizinhança cada um desses conjuntos de átomos que representa a ocorrência de uma ponte de hidrogênio encontrada na proteína. Assim, de modo geral, uma vizinhança é representada por um conjunto V = {A,D,H,O1,...,Om,N1,...,Nn}, onde m =0 e n =0. Se m = 0, então não existem átomos receptores do tipo oxigênio na vizinhança e se n = 0 não existem átomos receptores tipo nitrogênio.

Inicialmente considerou-se uma vizinhança com formato esférico com um raio de 3,5 Å centrado no átomo doador, por causa da simplicidade geométrica do cálculo. O valor do raio foi escolhido considerando: um comprimento de ligação de 1 Å entre o doador e o hidrogênio; até 2,5 Å de distância entre o hidrogênio e um potencial átomo receptor.

O hidrogênio tem liberdade de rotação em torno do eixo antecessor-doador respeitando: o comprimento da ligação covalente entre o átomo doador e o hidrogênio; o ângulo compreendido entre os átomos antecessor, doador e hidrogênio. Essa liberdade de rotação ocorre porque a ligação covalente entre o antecessor e o doador é simples. O formato esférico privilegia os receptores mais próximos do átomo doador, ao invés de uma proximidade uniforme em relação às possíveis posições do hidrogênio. Um novo formato de toro com um raio de 2,5 Å em torno da possível trajetória circular do

hidrogênio foi posteriormente adotado como formato mais adequado para a vizinhança.

Os arquivos textos da base de dados PDB referentes ao ano de 2004 foram lidos sequencialmente e quando um aminoácido lido for do tipo serina, treonina ou tirosina, é verificado se as coordenadas dos átomos antecessor, doador e hidrogênio do aminoácido estão presentes. Se esses átomos estiverem presentes, a busca dos átomos oxigênio e nitrogênio contidos na vizinhança, com o formato escolhido, é executada. Em função do número de átomos receptores, do tipo oxigênio e nitrogênio, é escolhido o arquivo texto de saída apropriado, em que as coordenadas dos dados da vizinhança serão adicionadas, quando houver pelo menos um átomo receptor potencial.

Dado um conjunto de vizinhanças com a mesma composição (mesmo número de receptores oxigênio, mesmo número de receptores nitrogênio), um método que permita agrupar essas vizinhanças é desejado, de forma que o arranjo espacial das composições contidas em um grupo seja similar. Se tentarmos usar como atributo diretamente as coordenadas espaciais dos átomos não teremos sucesso, uma vez que não sabemos a priori como as vizinhanças estão deslocadas/rotacionadas umas em relação às outras. Se considerarmos duas vizinhanças com a mesma quantidade m de receptores oxigênio e mesma quantidade n de receptores nitrogênio, e soubermos para cada átomo receptor de uma vizinhança qual o átomo correspondente da outra, podemos saber se o arranjo espacial das vizinhanças é similar pela comparação das distâncias entre pares de átomos contidos em uma estrutura e as distâncias dos pares correspondentes na outra estrutura, uma vez que a distância é uma medida independente de como a vizinhança está deslocada/rotacionada.

Cada vizinhança contém um único átomo antecessor, doador e hidrogênio. Porém, se existirem dois receptores do tipo oxigênio, por exemplo, é preciso um critério para associar cada oxigênio de uma vizinhança com apenas um oxigênio da outra, para que posteriormente medidas de similaridade possam ser calculadas entre as duas vizinhanças. Denominamos esse procedimento de nomeação dos receptores.

O primeiro método estabelecido para nomeação dos receptores contidos em uma vizinhança esférica, centrada no átomo doador, consiste em dado um receptor, fazer a projeção vetorial escalar do vetor com origem na extremidade do diâmetro que contém os átomos antecessor-doador e destino no átomo receptor, sobre o vetor com comprimento igual ao diâmetro da esfera, tendo como ponto médio o átomo doador e que contém também o átomo antecessor. A origem do vetor é a extremidade do diâmetro mais próxima do antecessor e o destino é o outro extremo do diâmetro. Os receptores são indexados na ordem crescente do valor de suas projeções vetoriais (Fig. 1).

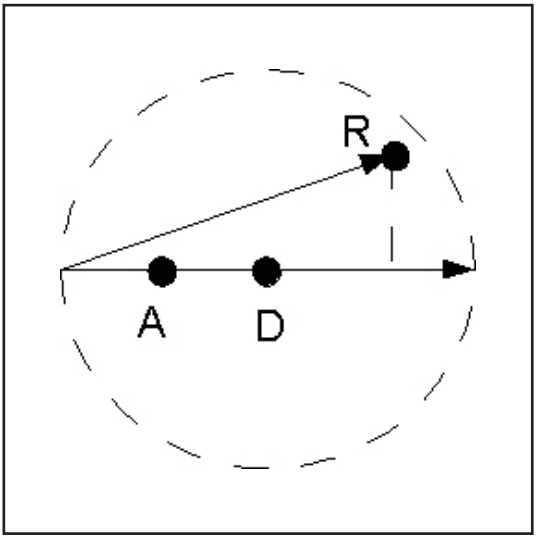


Fig. 1. Nomeação de receptores por projeção escalar de vetores.

A primeira tentativa para o estabelecimento de uma metodologia usando técnicas inteligentes para a previsão dos átomos de hidrogênio contidos nas hidroxilas (-OH) das cadeias laterais foi:

1. a seleção de receptores contidos em uma vizinhança esférica centrada no átomo doador com raio de 3,5 Å (1 Å para a ligação covalente entre o doador e o hidrogênio; 2,5 Å para a ponte entre o hidrogênio e o receptor);
2. a nomeação dos receptores dentro da vizinhança por projeção escalar;
3. o agrupamento dos padrões pela rede neural do tipo SOM (self organizing map, também conhecido como mapa de Kohonen) (KOHONEN, 1988) usando como entrada listas de distâncias entre as posições dos átomos que compõem a vizinhança;
4. o alinhamento espacial dos padrões agrupados em um cluster;
5. dado um cluster de padrões alinhados espacialmente, é feito o treinamento de uma rede neural tipo MLP (multilayer perceptron) para os dados do cluster, usando: como entrada as coordenadas dos átomos antecessor, doador e receptores; como saída 3 neurônios (unidades de processamento da rede neural) representando as coordenadas x, y, z do átomo de hidrogênio contido na hidroxila. Foi utilizada uma arquitetura contendo duas camadas intermediárias de neurônios. Variações na quantidade de neurônios dessas duas camadas foram testadas em experimentos distintos para verificar qual arquitetura apresentava maior índice de acertos nas previsões da rede neural.

Quando um padrão de ponte de hidrogênio é apresentado à rede MLP e a rede faz a previsão da posição do átomo de hidrogênio, calcula-se o erro linear da previsão como sendo a distância euclidiana entre a posição

prevista e a posição correta do átomo de hidrogênio. O resultado é classificado em 3 faixas de valores em função do erro:

• ótimo - erro < 0,2 Å;
• bom - 0,2 Å =erro 0,6 Å;
• ruim - erro > 0,6 Å.

Para os experimentos da rede MLP, os dados de um cluster foram divididos em 3 grupos:

- treino – dados utilizados para o treinamento supervisionado da rede MLP. O algoritmo utiliza os valores de entrada e de saída dos dados de treino para ajustar os pesos das conexões que interligam os neurônios da rede, em função do erro entre os valores de saída produzidos pela rede e os valores de saída corretos dos dados de treino. Como os dados de treino são usados para ajustar a rede, pode ocorrer o fato da rede memorizar o mapeamento entre as entradas e as saídas do conjunto de dados, ao invés de generalizar a função que mapeia todo o domínio dos dados de entrada a partir da amostra dos dados de treino, que é o desejado;
- validação – os dados de validação são usados para verificar o nível de generalização obtido pela rede neural. Podem ser usados, por exemplo, para se determinar o momento de encerramento do treino da rede neural.
- teste – os dados de teste são utilizados após o término da fase de treinamento da rede neural e servem para verificar o nível de acerto da rede neural.MLP. O algoritmo utiliza os valores de entrada e de saída dos dados de treino para ajustar os pesos das conexões que interligam os neurônios da rede, em função do erro entre os valores de saída produzidos pela rede e os valores de saída corretos dos dados de treino. Como os dados de treino são usados para ajustar a rede, pode ocorrer o fato da rede memorizar o mapeamento entre as entradas e as saídas do conjunto de dados, ao invés de generalizar a função que mapeia todo o domínio dos dados de entrada a partir da amostra dos dados de treino, que é o desejado;

Após os dados de uma vizinhança serem lidos, são deslocados de forma que: todas as coordenadas fiquem positivas; o menor valor das coordenadas em qualquer direção seja posicionado com uma folga de 10% do valor máximo – mínimo, a partir da origem. Os dados são comprimidos proporcionalmente de forma que fiquem no intervalo (0,1) em cada coordenada (porque esse é o intervalo de saída da função de ativação sigmóide da rede neural, atualmente usada nos experimentos), e um fator de expansão é calculado para permitir o reescalamento dos valores comprimidos para os valores originais. O treinamento da rede é feito após essa transformação dos dados. A distância média entre os átomos doador-hidrogênio e centro do orbital-hidrogênio dos padrões de entrada são calculadas para permitir que, além da posição do hidrogênio prevista diretamente pela rede neural, uma posição normalizada possa ser calculada a partir da posição prevista usando o valor do comprimento médio da ligação doador-hidrogênio.

Utilizando a metodologia descrita, os resultados obtidos não foram satisfatórios, mesmo após o treino da rede com 100.000 iterações para ajuste dos pesos. Posteriormente, o número de iterações foi reduzido para 10.000, porque a quantidade anterior estava consumindo horas de uso de processamento por experimento. Considerou-se como possível causa do mau desempenho o formato da vizinhança. Uma vez que a esfera é centrada no átomo doador, podem ser considerados alguns receptores contidos no hemisfério que contém o átomo antecessor que podem estar distantes do átomo de hidrogênio e introduzir ruído nos cálculos.

Um novo formato de vizinhança foi implementado, composto de um toro com raio de 2,5 Å em torno da possível trajetória circular em que o átomo de hidrogênio pode rotacionar. Contrariamente ao esperado, não houve uma melhoria significativa nos resultados dos experimentos.

Após uma análise criteriosa da metodologia empregada, foram percebidos 2 possíveis pontos fracos: o critério de nomeação dos átomos receptores e o algoritmo de agrupamento.

O processo de nomeação de receptores por projeção escalar, reduzindo o posicionamento de dados espaciais em uma ordenação unidimensional, poderia estar induzindo a erros na nomeação. Várias configurações espaciais diferentes podem ser reduzidas a uma única configuração unidirecional. O algoritmo de agrupamento SOM, que é um método tradicional, validado e amplamente utilizado para clusterização de dados, pode não ser adequado aos dados da aplicação proposta neste trabalho. O método SOM supõe uma distribuição uniforme dos dados, o que não é garantido ocorrer com os dados da aplicação.

Um novo critério para nomeação dos receptores foi desenvolvido, denominado nomeação por paridade de vizinhanças. Nesse novo método, dada uma vizinhança à nomeação dos receptores, não é fixa e definitiva como ocorre na nomeção por projeção escalar de vetores. Para cada par de vizinhanças é estabelecida uma nomeação dos receptores específica para determinar a correspondência dos receptores de uma vizinhança do par com os receptores da outra vizinhança do par. Computacionalmente esse método é mais caro, porém obtém resultados mais precisos em relação à nomeação por projeção escalar, uma vez que a identificação dos receptores deixa de ser um atributo geral calculado em função dos dados da vizinhança isolada, tornando-se um atributo específico entre um par de vizinhanças para o qual se deseja estabelecer a correspondência entre os receptores.

Dado um par de vizinhanças A e B, a nova nomeação é obtida fixando-se uma vizinhança A e movendo a outra B de forma a alinhar espacialmente os átomos antecessor, doador e hidrogênio das duas vizinhanças. Inicialmente todas possíveis combinações, formadas por um receptor oxigênio da vizinhança A e um receptor oxigênio da vizinhança B, e a distância entre esse par de receptores são calculadas. Escolhe-se sucessivamente os pares com menor distância. Para a seleção do par sucessivo são ignoradas as combinações que contenham receptores contidos em pares que já foram selecionados anteriormente para formar uma correspondêndia. O mesmo é feito com os receptores do tipo nitrogênio.

Um novo algoritmo de clusterização que é uma variação do algoritmo k-means (DUDA et al., 2001), denominado de algoritmo de fusão de clusters, foi desenvolvido para agrupar as vizinhanças. No algoritmo k-means tradicional, inicialmente um conjunto de n células é iniciado com um vetor de pesos aleatoriamente. Em cada iteração, é anotado para cada padrão de entrada apresentado qual a célula que tem o vetor de pesos mais parecido com o vetor de pesos do padrão de entrada. Ao fim da iteração, o vetor de pesos de cada célula é atualizado com a média dos vetores de pesos dos padrões de entrada que foram associados àquela célula. O processo é repetido até haver uma convergência.

A variação proposta privilegia a alocação de um cluster (célula) cada vez que aparece um novo padrão muito diferente dos já apresentados, fazendo a fusão dos dois clusters mais próximos em um dos dois clusters, liberando o outro para abrigar o novo padrão apresentado. A quantidade de clusters é fixa na versão atualmente implementada. Outra diferença é que o vetor de pesos da célula é atualizado com o valor da média dos padrões após a apresentação de cada padrão individual, e não apenas ao final da iteração, como ocorre no algoritmo k-means (DUDA et al., 2001). Cada célula tem associado dois vetores de pesos: um com o somatório dos vetores de pesos dos padrões associados à célula e um com o valor médio dos vetores de peso dos padrões. Também são anotadas para cada célula a quantidade de padrões atribuídos a ela, a célula mais próxima (com o vetor de pesos mais parecido) e a distância à célula mais próxima. Inicialmente, cada célula tem o vetor de pesos iniciado com o vetor de peso de um dos padrões de entrada (um padrão distinto para cada célula). Então para cada célula é calculada qual é a outra célula mais próxima. A partir desse momento, para cada novo padrão apresentado é calculada qual a célula que tem o vetor de pesos mais próximo do padrão e a distância entre o padrão e essa célula. Se a distância for maior que a distância entre as duas células mais próximas, os dados de uma das duas células mais próximas são fundidos com os dados da outra célula mais próxima, e uma das células é liberada para abrigar os dados do novo padrão apresentado. Caso contrário, o vetor de pesos do padrão apresentado é adicionado aos dados do vetor de pesos da célula mais próxima do padrão. O termo fusão de dados compreende:

- a soma do vetor de pesos do padrão apresentado ou do vetor somatório de pesos da célula a ser fundida, com o vetor somatório de pesos da célula escolhida;
- o incremento da quantidade de padrões da célula escolhida com a quantidade de padrões associada à célula a ser fundida ou o incremento de um caso seja incluído apenas o padrão apresentado;
- a atualização do vetor média de pesos da célula escolhida dividindo-se o vetor somatório de pesos pela quantidade de padrões associada à célula escolhida.

Após a apresentação de cada padrão, como o vetor média de pesos da célula escolhida é atualizado, é necessário também recalcular para cada célula qual a outra célula mais próxima. O processo é repetido até haver convergência, isto é, uma interação em que não ocorra fusão de células, ou até que uma quantidade máxima de iterações seja executada se o processo

não convergir. Este algoritmo tem por objetivo priorizar o isolamento de padrões muito diferentes em clusters individuais, evitando assim que padrões bizarros sejam agrupados com padrões medianos.

O novo algoritmo de clusterização das vizinhanças é executado em duas fases:

- na primeira fase as vizinhanças são agrupadas em função da geometria do comprimento das ligações entre os átomos antecessor e doador, doador e hidrogênio;
- na segunda fase as vizinhanças previamente agrupadas em um dado cluster da primeira fase são usadas como dados de entrada e particionadas em novos clusters em função do agrupamento formado a partir das distâncias dos átomos receptores da vizinhança tomados dois a dois entre si, e das distâncias de cada átomo receptor aos átomos antecessor, doador e hidrogênio.

Quando usamos todos os átomos disponíveis da vizinhança em uma única fase, a similaridade fica diluída em todos os parâmetros usados para o cálculo. A justificativa para esse procedimento em duas fases é a obtenção de clusters contendo comprimento da ligação doador–hidrogênio e ângulo antecessor – doador -hidrogênio mais similares possíveis para posteriormente alimentarem redes neurais especializadas nesses dados. Cada cluster é então usado para treinar uma rede neural específica para o cluster.

Resultados e discussão

As arquiteturas das redes neurais MLP utilizadas nos experimentos é mostrada na Tabela 1. Os três neurônios da camada de saída referem-se às coordenadas x, y e z do átomo de hidrogênio a ser previsto. Por uma questão de limitação de espaço apenas os experimentos da arquitetura composta por 48 neurônios na primeira camada intermediária e 24 neurônios na segunda camada intermediária são apresentados neste trabalho.

Tabela 1. Arquiteturas de redes neurais MLP utilizadas nos experimentos.

Quantidade de neurônios na primeira camada intermediária	Quantidade de neurônios na Segunda camada intermediária	Quantidade de neurônios na camada de saída
12	12	3
24	24	3
48	12	3
48	24	3

Inicialmente, imaginou-se que a baixa taxa de acertos era decorrência de problemas com o treinamento da rede neural. Nesse estágio inicial da pesquisa, treinava-se várias arquiteturas da rede neural com centenas de milhares de iterações e cada experimento consumia horas de processamento. Como esse treinamento exaustivo não estava atingindo um nível de acerto desejável, restringiu-se o número máximo de iterações para 10.000, para acelerar a execução dos experimentos.

Nas Tabelas 2, 3 e 4 com dados de experimentos (mostradas na seção Anexo), os campos da linha cabeçalho (em negrito) e seus valores são:

- Cluster_id – identificação do cluster onde *i_j* é a célula (*i*,*j*) de uma matriz quadrada de tamanho 10 (total de 100 clusters) para as tabelas 2 e 3, para vizinhanças contendo apenas um receptor do tipo oxigênio. Para a tabela 4 a identificação é (*k_l*)*i_j* onde k é o numero de receptores oxigênio, l é o número de receptores nitrogênio, seguido do i_ésimo cluster escolhido da primeira fase e do j_ésimo cluster escolhido da segunda fase. Assim, se na primeira fase foram criados *m* clusters, $0 \leq i \leq m-1$. Na segunda fase o cluster *i* é dividido em *n* clusters, e $0 \leq j \leq n-1$. Portanto, o número total de cluster é *m*n*. Os experimentos foram feitos com $m \leq 10$ e $n \leq 30$ (total de 300 clusters);
- Tamanho - quantidade de padrões contidas no cluster;
- Iterações - número de iterações que a rede neural foi treinada;
- Melhor Iter. - iteração que apresentou melhores resultados para os dados de validação;
- %dados treino - porcentagem de dados do cluster usados para treino da rede neural;
- %dados val. - porcentagem de dados do cluster usados para validação da rede neural;
- %dados teste - porcentagem de dados do cluster usados para teste da rede neural.

Nas linhas seguintes da tabela são quantificados o número de padrões (Treino Qt, Val Qt, Teste Qt) e a porcentagem (Treino %, Val %, Teste %) em relação aos subconjuntos de dados de treino, validação e teste, por faixas de classificação do resultado previsto pela rede (ótimo, bom, ruim).

A utilização de nomeação de receptores por projeção escalar e agrupamento por SOM (self organizing map) apresentaram altas taxas de resultados não satisfatórios após o treinamento das redes neurais MLP (multilayer perceptron), para vizinhanças no formato esférico (Tabela 2) e em toro (Tabela 3). Para a configuração evoluída da metodologia, usando nomeação de receptores por paridade e agrupamento por fusão de clusters, obtém-se um grau de acerto muito maior, conforme pode ser verificado nos experimentos listados na Tabela 4.

Com as melhorias inseridas na metodologia com novos algoritmos de nomeação dos receptores por paridade e fusão de clusters em 2 fases, obteve-se clusters contendo maior similaridade entre os padrões. Como consequência dessa maior similaridade, a rede neural conseguiu generalizar melhor os dados e aumentou significativamente a precisão da

previsão. Um bom ajuste dos pesos da rede neural foi possível com uma quantidade menor de iterações, reduzindo o tempo de processamento. Após as melhorias introduzidas, a quantidade máxima de iterações de treinamento foi limitada entre 500 e 1.000 iterações.

A melhor iteração em um experimento é o número da que apresentou melhores resultados de previsão para os dados de validação, entre todas as iterações de treinamento. A rede apresentou a iteração 321, no experimento do cluster (2_2)3_9, como a maior melhor iteração entre os experimentos, isto é, no experimento com pior treinamento foram necessárias 321 ciclos para atingir os melhores resultados de previsão da rede. Nos outros experimentos atingiu-se o melhor resultado de ajuste de pesos da rede com menos iterações. Antes da melhoria na metodologia, a menor das melhores iterações para vizinhança esférica foi 516 para o cluster 2_2, ou seja, uma quantidade superior a 516 iterações foi necessária nos outros experimentos. Para a vizinhança tórica, a menor das melhores iterações foi a de número 4.717 para o cluster 1_0.

Podemos observar que antes das melhorias da metodologia os experimentos com uso de 50% dos dados para treino apresentam em geral apresentam melhor desempenho de treinamento. Uma possível explicação para essa situação é que os parâmetros livres estão sendo usados para decorar os exemplos de treino, e, portanto, quanto menos exemplos, mais memória para decorá-los a rede tem à sua disposição na fase de treinamento.

Outro resultado observado é que antes das melhorias os experimentos que usam 50% dos dados para treino tem também melhor desempenho para os dados de validação e teste. A possível explicação é que uma vez que a rede não está conseguindo generalizar bem os dados, quanto maior a quantidade de dados de validação em cima dos quais é testada a capacidade de generalização da rede, maior a possibilidade de que esses dados possam ter alguma redundância e se benefiarem do pouco de generalização que a rede conseguiu obter. De forma similar, os resultados dos dados de teste acompanham esse raciocínio, uma vez que foi usada a mesma porcentagem dos dados para validação e teste. Após as melhorias, como é conseguido um aumento de similaridade para os padrões contidos nos clusters, a rede neural consegue generalizar os dados, e o desempenho para os dados de treino, validação e teste ficam semelhantes quando se faz o treino com 50 ou 80% dos dados para treino. O desempenho dos dados de treino é próximo ao desempenho dos dados de validação atestando a capacidade de generalização da rede neural.

Os bons resultados obtidos na primeira fase da pesquisa, que consistia em estabelecer uma nova metodologia baseada em métodos computacionais de reconhecimento de padrões, e verificar se era possível fazer a previsão da posição de átomos de hidrogênio usando esse novo método, sinalizam positivamente para o potencial do novo método. A continuidade da pesquisa para tornar a nova metodologia completamente viável envolve os seguintes problemas: dado um padrão de ponte de hidrogênio sem informações sobre o átomo de hidrogênio, identificar qual dos clusters previamente agrupados

contém padrões mais similares ao padrão apresentado para se escolher a rede neural mais adequada para fazer a previsão do hidrogênio; estender o método para cadeias laterais contendo dois, três ou quatro hidrogênios.

Conclusão

Uma metodologia inédita na literatura está em desenvolvimento com resultados iniciais promissores para a previsão da posição de hidrogênios em proteínas. Os experimentos demonstraram alto grau de acerto para o problema sendo investigado. Quando a pesquisa estiver completa, uma técnica de previsão alternativa rápida e com alto grau de acerto é esperada. A rapidez do novo método decorre que o custo computacional ocorre na fase de preprocessamento dos dados, que pode ser feito apenas uma vez periodicamente, enquanto que a previsão feita pelas redes neurais já treinadas é um processo muito rápido. Também foram obtidos como resultados da pesquisa um novo algoritmo de clusterização, o algoritmo fusão de clusters, desenvolvido especificamente para situações em que se deseja isolar dados com baixa ocorrência.

Referências Bibliográficas

TURNER, P. C.; MCLENNAN, A. G.; BATES, A. D.; WHITE, M. R. H. *Molecular biology*. 2nd. ed. London: BIOS Scientific Publishers, 2000. 346 p. (The Instant Notes).

MCDONALD, I. K.; THORNTON, J. M. *Satisfying hydrogen bonding potential in proteins*. Journal of Molecular Biology, v. 338, p. 777-793, 1994.

GANCIA, E.; MONTANA, J. G.; MANALLACK, D. T. Theoretical hydrogen bonding parameters for drug design. J. Mol. Graphics Modell., v. 19, n. 3, p. 349-362, 2001.

NIKOLIĆ, A.; JOVIĆ, B.; KRSTIĆ, V.; TRIČKOVIĆ, J. N-H…*O hydrogen bonding*. FT-IR, NIR and ^{1}H NMR study of N-methylpropionamide - Dialkyl ether systems. Journal of Molecular Structure, v. 889, n. 1-3, p. 328-331, Oct. 2008.

TORSHIN, I. Y.; WEBER, I. T.; HARRISON, R. W. Geometric criteria of hydrogen atoms by neutron crystallography on fully deuterated myoglobin. PNAS, v. 97, n. 8, p. 3872-3877, 2000.

ENGLER, N.; PARAK, F. G. *Hydrogen atoms in proteins: position and dynamics*. PNAS, v. 100, n. 18, p. 10243-10248, 2003.

GEORGESCU, R. E.; ALEXOV, E. G.; GUNNER, M. R. *Combining conformational flexibility and continuum electrostatics for calculating pKas in proteins*. Biophys. J., v. 83, p. 1731-1748, 2002.

BROOKS, B. R.; BRUCCOLERI, R. E.; OLAFSON, B. D.; STATES, D. J.; SWAMINATHAN, S.; KARPLUS, M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem*., v. 4, p. 187-217, 983.

BRÜNGER, A. T.; ADAMS, P.; CLORE, G. M.; DELANO, W. L.; GROS, P.; GROSSE-KUNSTLEVE, R. W.; JIANG, J. S.; KUSZEWSKI, J.; NILGES, M.; PANNU, N. S.; READ, R. J.; RICE, L. M.; SIMONSON, T.; WARREN, G. L. *Crystallography and NMR system: a new software for macromolecular structure determination. Acta Cryst*. D Biol. Crystallogr., v. 54, p. 905-921, 1998.

LINDAHL, E.; HESS, B.; VAN DER SPOEL, D. *GROMACS 3.0: a package for molecular simulation and trajectory analysis*. J. Mol. Model., v. 7, p. 306-317, 2001.

WORD, J. M.; LOVELL, S. C.; RICHARDSON, J. S.; RICHARDSON, D. C. Asparagine and Glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol*., v. 285, p. 1735-1747, 1999.

VRIEND, G. WHAT IF: a molecular modelling and drug design program. *J. Mol. Graph*., v. 8, p. 52-56, 1990.

BRÜNGER, A. T. X-*PLOR version 3.1: a system for X-ray crystallography and NMR*. New Haven, CT: Yale University Press, 1992.

FORREST, L. R.; HONIG, B. *An assessment of the accuraccy of methods for predicting hydrogen positions in protein structures*. Proteins: Structure, Function and Bioinformatics, v. 61, p. 296-309, 2005.

FABIOLA, F.; BERTRAM, R.; KOROSTELEV, A.; CHAPMAN, M. S. *An improved hydrogen bond potential: impact on medium resolution protein structures*. Protein Science, v. 11, p. 1415-1423, 2002.

KORTEMME, T.; MOROZOV, A. V.; BAKER, D. *An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein complexes*. J. Mol. Biol, n. 326, p. 1239-1259, 2003.

KOHONEN, T. *Self-organization and associative memory*. 3rd ed. NewYork: Springer-Verlag, 1988.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. 2nd edition. New York: Wiley- Interscience, 2001. 654 p.

Embrapa

Informática Agropecuária

Ministério da
Agricultura, Pecuária
e Abastecimento

BRASIL
UM PAÍS DE TODOS
GOVERNO FEDERAL