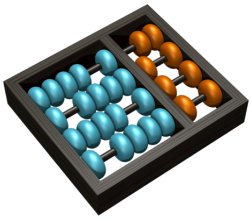
Proposta de Dissertação de Mestrado

Aluno: Lucas G. Farris
Orientador: Jacques Wainer
Coorientador: Guilherme P. Telles

Universidade Estadual de Campinas
Instituto de Computação



Resumo *Predição de Mapas De Contato* A predição de mapas de contato de proteínas é um problema muito estudado e de grande importância para a área de Bioinformática. Mapas de contato são utilizados como ferramenta para a predição da estrutura tridimensional de proteínas. O US National Institute of General Medical Sciences (NIH) realiza a cada dois anos uma avaliação (CASP) dos algoritmos que são considerados estado da arte. Os dados usados nos experimentos são então disponibilizados no site da instituição. A importância deste estudo é entender o *folding* de proteínas, e consequentemente entender como isso afeta seu comportamento bioquímico e físico. Neste trabalho estudamos como técnicas avançadas de *deep learning* afetam a predição de mapas de contato, e como isso se compara ao estado da arte.

Keywords: Aprendizado de Máquina, Bioinformática

1 Introdução

Este documento apresenta a proposta de um estudo de técnicas de predição de mapas de contato de proteínas a partir das sequências de aminoácidos que as compõem. A técnica em questão estima se dois aminoácidos, situados em determinadas posições da sequência, estão a uma certa distância mínima um do outro na estrutura fechada da proteína.

Buscamos entender as técnicas da literatura e desenvolver um modelo capaz de realizar predições com desempenho similar ou superior ao estado da arte. A melhora do desempenho visa tornar esse

tipo de predição uma técnica cotidiana nos estudos de proteômica, visto que hoje isso ainda não é possível.

O trabalho será realizado por meio da revisão bibliográfica dos artigos mais recentes (estado da arte), bem como de técnicas novas de *Machine Learning*. Buscamos avaliar o resultado da união de *deep learning* e da predição de mapas de contato.

No restante da proposta encontram-se a fundamentação teórica pertinente a este trabalho, a revisão bibliográfica descrevendo estudos relevantes, os detalhes da proposta, o cronograma e as referências bibliográficas.

1.1 Fundamentação Teórica

Um **resíduo** é um monômero específico dentro de uma cadeia polimérica. No caso das proteínas, um resíduo é um aminoácido (cada resíduo representa unicamente um aminoácido). Uma proteína por sua vez é uma cadeia com uma quantidade mínima de resíduos, não necessariamente diferentes entre si.

Proteínas, ao se formarem, se dobram, e denomina-se as subestruturas de acordo com o nível de escala e abstração. O primeiro nível é chamado de **estrutura primária**, representada pela sequência de aminoácidos presente na proteína na ordem em que são encadeados. A **estrutura secundária** de uma proteína consiste nos dobramentos locais dos aminoácidos (hélice-alfa, folha-beta, ...). A **estrutura terciária** é representada pelos dobramentos globais (após o *folding*, processo de compactação) em uma forma globular, gerada a partir de propriedades químicas e físicas das moléculas. Neste estudo, é interessante usar informações da estrutura primária, para descobrir informações sobre a secundária e a terciária. Na Figura 1 pode-se ver uma representação dessas estruturas.

Define-se dois resíduos de uma estrutura terciária como um **contato** se a menor distância euclidiana entre seus átomos C_β (C_α para glicinas) for no máximo 8Å ($8 \times 10^{-10} m$). Os contatos são classificados pela distância de aminoácidos entre os dois resíduos. Contatos com distância entre 6 e 11 aminoácidos são chamados de ***short-range***, entre 12 e 23 são chamados de ***medium-range*** e por fim os contatos separados por mais de 24 aminoácidos são chamados de ***long-range***. Ao analisar a sequência de uma proteína, não é tri-

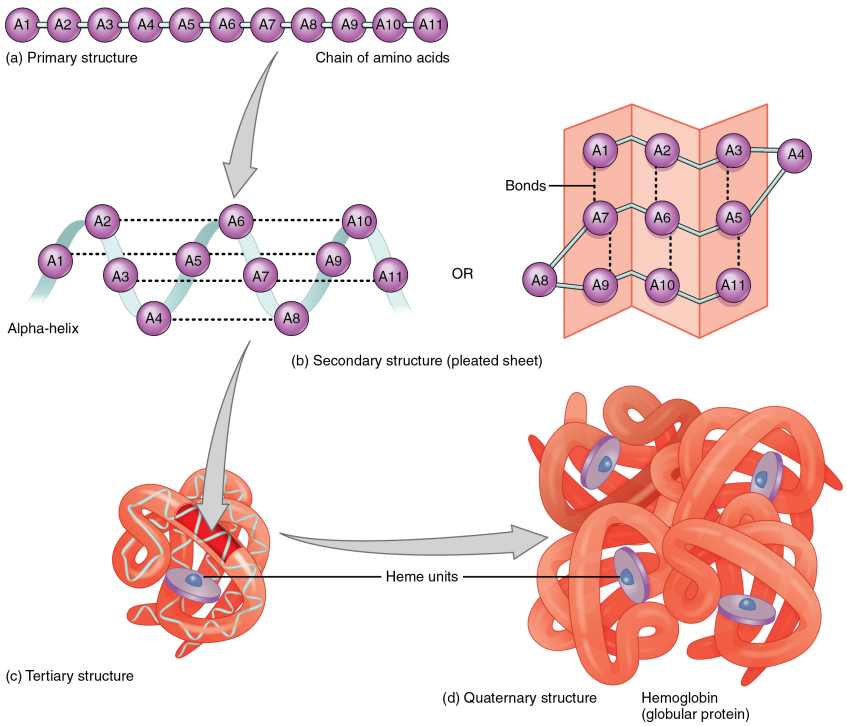


Figura 1. Representação gráfica das estruturas das proteínas. Fonte: [1]

vial obter os seus contatos, mas uma vez calculados eles fornecem informações importantes sobre a estrutura. Contatos *long-range* são mais difíceis de prever, mas fornecem informações mais importantes. Sabe-se que mesmo uma quantidade pequena de contatos previstos corretamente (35%) podem ser muito úteis na predição da estrutura terciária[2].

1.2 Mapa de Contatos

Mapa de contatos resíduo-resíduo é uma técnica que tenta inferir os contatos de uma determinada sequência. Por exemplo, seja S uma sequência de aminoácidos de tamanho N, e M uma matriz $N \times N$; a posição $M[x, y]$ será 1 se houver contato entre os resíduos S_x e S_y, ou 0 caso contrário. Mapas de contato são representações topológicas de uma proteína, invariantes quanto à rotação e translação. Na Figura 2 temos um exemplo de mapa criado a partir de uma sequência.

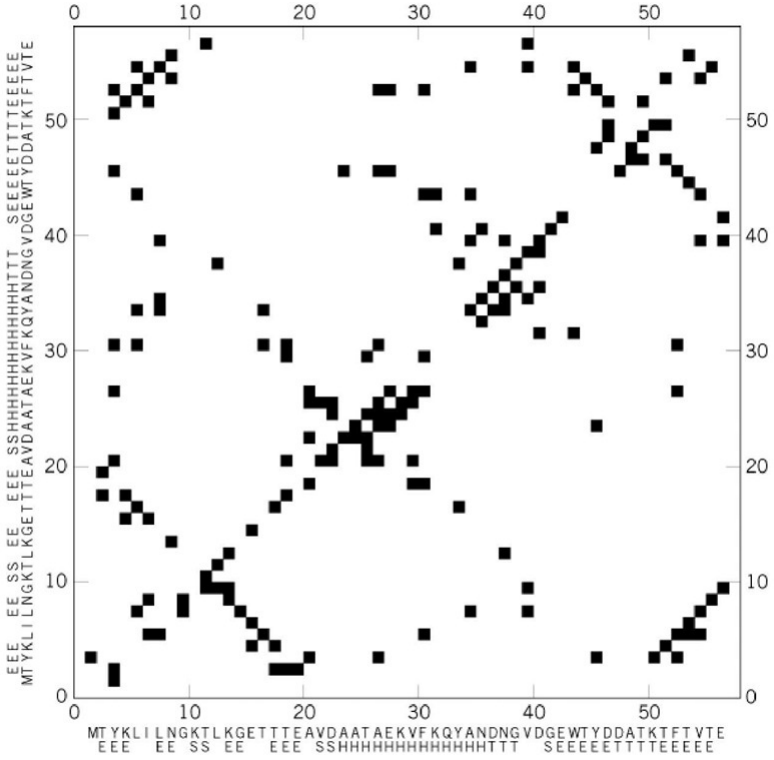


Figura 2. Exemplo de um mapa de contato resíduo-resíduo. Fonte: [3]

A performance da predição do mapa de contatos é avaliada pela a acurácia (*Acc*). A acurácia é a razão entre os contatos preditos corretamente (verdadeiro-positivos TP), e o total de contatos preditos (somando falso-positivos FP). Desta forma:

$$Acc = \frac{TP}{TP + FP}$$

Para realizar o cálculo da acurácia usa-se o conjunto dos $L/5$ melhores contatos preditos, onde L é o comprimento da sequência. Essa é a métrica mais aceita de acurácia [4], e usa apenas contatos *long-range*.

Existe um evento bienal chamado *Critical Assessment of protein Structure Prediction* (CASP) realizado pelo *US National Institute of General Medical Sciences*. Este evento tem o intuito de medir a performance dos melhores algoritmos existentes que predizem estru-

turas de proteínas a partir de sequências. O último evento foi o de número 11, realizado em 2014.

No CASP, as acurácias de contatos *long-range* têm repetidamente atingido a marca de aproximadamente 20% [4]. Isso sugere que predições desse tipo ainda não são suficientes para serem usadas no cotidiano.

1.3 Aprendizado de Máquina

Existem duas abordagens clássicas para o problema da classificação da estrutura terciária de proteínas. A primeira é um modelo comparativo, que usa estruturas conhecidas (*templates*) na literatura para inferir o modelo em estudo. A outra abordagem (*de novo* ou *ab initio*) usa a sequência para prever a estrutura final sem base em outras montagens.

Existem na literatura diversas técnicas *ab initio* que usam aprendizado de máquina. Essas técnicas aprendem as probabilidades dos contatos a partir de conjuntos de teste de estruturas de proteínas e dados obtidos experimentalmente. Os inputs para esses algoritmos geralmente são estruturas secundárias preditas, áreas de superfícies acessíveis (*solvent accessibility*, está relacionado com o quanto o aminoácido se "move"durante o folding [5]) preditas, e informações evolucionárias (perfis de frequência dos resíduos, calculados usando alinhamentos múltiplos).

Usar a sequência de aminoácidos para descobrir a estrutura secundária de uma proteína é, de certa maneira, um problema mais simples. A fim de obter o mapa de contatos a partir da estrutura secundária, precisamos saber como cada dobramento local se une ao próximo. Para armazenar essas informações usa-se um tipo de matriz chamada *coarse contact maps*, no qual as linhas e colunas são as estruturas secundárias, e cada posição da matriz recebe o tipo de ligamento delas. As opções são nenhum contato, contato paralelo ou contato anti-paralelo.

A orientação de duas estruturas é o ângulo entre os vetores de suas próprias orientações, que por sua vez é calculado unindo o centro de gravidade da primeira e segunda metades da estrutura. Um ângulo paralelo é menor que 90°, e o anti-paralelo é maior.

2 Revisão Bibliográfica

Os primeiros trabalhos que relacionam machine learning e estruturas de proteínas datam do final dos anos 90. Em 99, foi proposto um método [6] que usava redes neurais para gerar mapas de contato de proteínas. O algoritmo tinha acurácia melhor que os da época. Em 2001, uma continuação do mesmo trabalho foi proposta [7], na qual eram usadas mais informações como input do algoritmo.

Ainda usando redes neurais, um estudo de 2006 [8] usava também *Principal Component Analysis* (PCA) para melhorar a acurácia do algoritmo. O estudo divide o problema em obter os autovetores principais e em seguida obter o mapa a partir dos vetores, usando uma rede neural bidirecional de duas camadas. Em 2007, outra técnica[9] de redes neurais determinava a probabilidade de contato entre pares de resíduos usando alinhamento múltiplo como *feature*. Esse trabalho se saiu melhor que as demais redes neurais do CASP7.

Outra técnica importante de 2007 foi um trabalho [10] que usava *Support Vector Machines* (SVM) na predição de mapas de contato, bem como grande quantidade de informações como input do algoritmo. Foi um dos melhores classificados do CASP 7 e 8, e se destacou pela acurácia em contatos *medium-range* e *long-range*.

Nos últimos anos, com o destaque de técnicas de *deep learning* na área de inteligência artificial, alguns pesquisadores utilizaram essa técnica para predizer mapas de contatos. A seguir, há a descrição, de maneira geral, de dois deles.

2.1 Bidirectional recurrent neural network [11]

Para obter as estruturas secundárias a partir da sequência, pode-se usar uma rede neural bidirecional recorrente [12], que prediz a probabilidade de dois elementos estarem em contato paralelo, antiparalelo ou sem contato. O input da rede neural é, além da distância entre as duas estruturas:

- A distribuição média de aminoácidos entre as estruturas e seus vizinhos
- O comprimento, em aminoácidos, das estruturas
- A distância entre cada estrutura e seus vizinhos

– Flags para identificar se as estruturas estão entre as duas primeiras ou as duas últimas
– Dois vetores contendo a distribuição de aminoácidos, mas divididos entre os índices pares e ímpares (isso é necessário para classificar corretamente strands)

Para prever o mapa de contato, usa-se uma rede neural profunda, consistindo em uma pilha tridimensional de redes neurais. Cada nível feed-forward é uma rede de três camadas treinada com back-propagation, e todos os níveis tem o mesmo tamanho de input, de camadas ocultas e um output: um mapa de contato. Cada nível recebe como input o mapa do anterior, e refina a predição. As camadas desta arquitetura são importantes, pois simulam a ideia de que o folding não é instantâneo, mas sim um processo organizado com estágios de aperfeiçoamento. Usar o mapa anterior como input é importante, pois experimentalmente sabemos que um contato tem alta probabilidade de possuir outros contatos em sua vizinhança, ou seja, um ponto i,j do mapa tem grande chance de ser um contato se seus vizinhos também o forem. Os inputs para essa estrutura são:

– Informações evolucionárias
– Estruturas secundárias preditas
– Acessibilidade dos solventes (buried ou exposed)
– Contatos entre estruturas secundárias

2.2 Restricted Boltzmann machine and deep belief networks [13]

Neste artigo, os autores usam deep learning e boosting para a predição do mapa de contato das proteínas (medium-range e long-range), este autor alimentou as redes com informações dos resíduos vizinhos ao que está sendo encontrado. Usando como input para as redes uma janela de resíduos, centralizada no resíduo cujo contato queremos descobrir, com tamanho variando entre 7 e 19.

No trabalho, eles também realizam a predição de contatos short-range, usando janelas de 12 resíduos, para descobrir todos os contatos existentes dentro da janela. Em ambas as arquiteturas descritas, os autores fizeram o ajuste fino com back-propagation.

Os resultados obtidos se comparam ao estado da arte, demonstrando comparações com os resultados dos algoritmos $SVMCom$ e

*ProC_S*3. Os autores focam na performance, usando GPU's e computação paralela para melhorar o desemprenho.

3 Objetivos

Pretende-se, neste estudo, cobrir os seguintes tópicos:

- Encontrar e formatar um banco de dados para ser usado como teste e validação do algoritmo
- Desenvolver um modelo de *Deep Learning* capaz de gerar um mapa de contato a partir de uma sequência
- Realizar ajustes finos e alterações necessárias para maximizar sua acurácia
- Realizar medições e comparações para a avaliação dos resultados

4 Etapas previstas e Cronograma

A Tabela 1 apresenta as atividades a serem desenvolvidas durante o período deste trabalho:

1. Obtenção dos créditos obrigatórios em disciplinas
2. Revisão Bibliográfica
3. Escrita da proposta de mestrado
4. Exame de Qualificação do Mestrado
5. Desenvolvimento do trabalho
6. Escrita da dissertação
7. Revisão da dissertação
8. Defesa da dissertação

É importante ressaltar que o cronograma apresentado pode ser adaptado à medida que os avanços ocorram, pois poderão nos levar a resultados mais promissores do que outros, o que nos faria dedicar mais tempo em alguma(s) atividade(s), em detrimento de outras.

5 Metodologia

Os resultados serão validados utilizando *cross-validation*, e comparados com outros reultados da literatura usando métricas de acurácia.

Tabela 1. Cronograma de atividades

Etapa	2015										2016												2017	
	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
1	*	*	*	*		*	*	*	*															
2	*	*	*	*	*																			
3	*	*	*	*	*																			
4								*																
5						*	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*		
6									*	*				*	*			*	*		*	*		
7																							*	
8																								*

Referências

1. OpenStax College - Anatomy & Physiology, "Connexions," 2013. [Online; acessado 28 de Março, 2015].
2. M. Tress and A. Valencia, "Predicted residue-residue contacts can help the scoring of 3d models," *Proteins*, vol. 78, p. 1980–1991, 2010.
3. What When How, "Contact maps - molecular biology." [Online; acessado 28 de Março, 2015].
4. A. Kryshtafovych *et al.*, "Casp9 results compared to those of previous casp experiments," *Proteins*, vol. 79, pp. 196–207, 2011.
5. R. Adamczak, A. Porollo, and J. Meller, "Combining prediction of secondary structure and solvent accessibility in proteins," *Proteins*, vol. 59, pp. 15–21, 2005.
6. P. Fariselli and R. Casadio, "A neural network based predictor of residue contacts in proteins," *Protein Engineering*, vol. 12, no. 1, pp. 467–475, 1999.
7. P. Fariselli, O. Olmea, A. Valencia, and R. Casadio, "Prediction of contact maps with neural networks and correlated mutations," *Protein Engineering*, vol. 14, no. 11, pp. 835–843, 2001.
8. A. Vullo, I. Walsh, and G. Pollastri, "A two-stage approach for improved prediction of residue contact maps," *BMC Bioinformatics*, vol. 7, no. 180, 2006.
9. G. Shackelford and K. Karplus, "Contact prediction using mutual information and neural nets," *Proteins: Structure, Function and Genetics*, vol. 69, no. 8, pp. 159–164, 2007.
10. J. Cheng and P. Baldi, "Improved residue contact prediction using support vector machines and a large feature set," *BMC Bioinformatics*, vol. 8, no. 113, 2007.
11. P. Di Lena, K. Nagata, and P. Baldi, "Deep architectures for protein contact map prediction," *Bioinformatics (Oxford, England)*, vol. 28, no. 19, pp. 2449–2457, 2012.
12. P. Baldi and G. Pollastri, "The principled design of large-scale recursive neural network architectures dag-rnns and the protein structure," *Journal of Machine Learning Research*, vol. 4, pp. 575–602, 2003.
13. J. Eickholt and J. Cheng, "Predicting protein residue-residue contacts using deep networks and boosting," *Bioinformatics (Oxford, England)*, vol. 28, no. 23, pp. 3066–3072, 2012.