Desenvolvimento de Métodos e Estratégias Computacionais para a Identificação de Padrões Conformacionais em Proteínas

Mariana Oliveira‡ e Márcio Dorn†

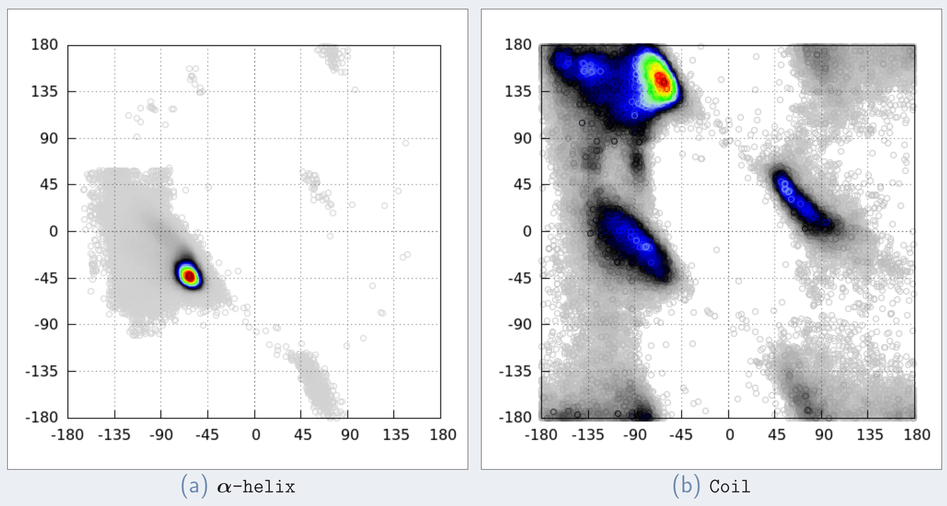
†Instituto de Informática, Universidade Federal do Rio Grande do Sul, Brasil.
‡Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Brasil.

Introdução

Proteínas são longas sequências de resíduos de aminoácidos que adotam uma estrutura tridimensional única quando em meio fisiológico, a qual está associada a sua função na célula [1]. A determinação dessa estrutura 3D pode ser obtida de maneira experimental, através das técnicas de cristalografia por difração de raio-x ou de ressonância magnética nuclear (RMN). Entretanto, ambas as técnicas são de alto custo e podem levar muito tempo, assim, o desenvolvimento e a utilização de métodos computacionais para a predição de estruturas 3D é fundamental. Atualmente é possível encontrar diversos métodos propostos para resolver os problemas de predição, os quais podem ser divididos em métodos baseados de primeiros princípios (*ab initio* e predição de novo) e métodos baseados em conhecimentos (alinhavamento e modelagem comparativa) [2]. Os métodos baseados em conhecimento são dependentes de informações estruturais determinadas experimentalmente. Porém, ainda há muito a ser obtido das bases de dados e, além disso, novos padrões ainda podem ser descobertos. Desta maneira, há a necessidade da construção de estratégias computacionais para a identificação e extração de padrões conformacionais a fim de melhorar a predição de métodos baseados no conhecimento.

Materiais e Métodos

Database: o sucesso da predição de estruturas 3D de uma proteína em métodos baseados em conhecimento está diretamente relacionado a qualidade das informações obtidas do PDB (Berman). Foram selecionadas ∼11000 estruturas proteícas contendo as informações de estrutura secundária e preferência conformacional dos residuos de aminoácidos. Todas as estruturas selecionadas foram experimentalmente determinadas através das técnicas de difração de raio-x com uma resolução ≤ 2.5 $\mathring{A}$. Somente foram consideradas estruturas depositadas até dezembro de 2014. Foram removidas todas estruturas com R-observado > 0.2 e selecionadas apenas aquelas com identidade de pelo menos 30%. Para os átomos da cadeia principal foram selecionados apenas resíduos com B-factor $\leq 30\mathring{A}$ e ocupância igual a 1.
Foi utilizado o programa STRIDE [3] para atribuir a estrutura secundária de cada resíduo de aminoácido, o qual possui oito estados de estrutura secundária: B, E, H, G, I, *b*, C e T. Entretando, devido a baixa ocorrência ($< 1\%$) dos modelos I e *b*, foram considerados apenas 6 modelos: H (α-helix), G ($\mathbf{3}_{10}$-helix), E (β-sheet), B (β-bridge), T (Turn) e C (Coil). Para cada estrutura foi gerado um arquivo contendo os aminoácidos, sua estrutura secundária e sua informação conformacional, ou seja, os ângulos de torção phi (ϕ) e psi (ψ) da cadeia peptídica.
Cada estrutura secundária tem sua preferência conformacional nos gráficos de Ramachandran, o qual representa ângulos phi (eixo x) *versus* ângulos psi (eixo y).



(a) α-helix (b) Coil

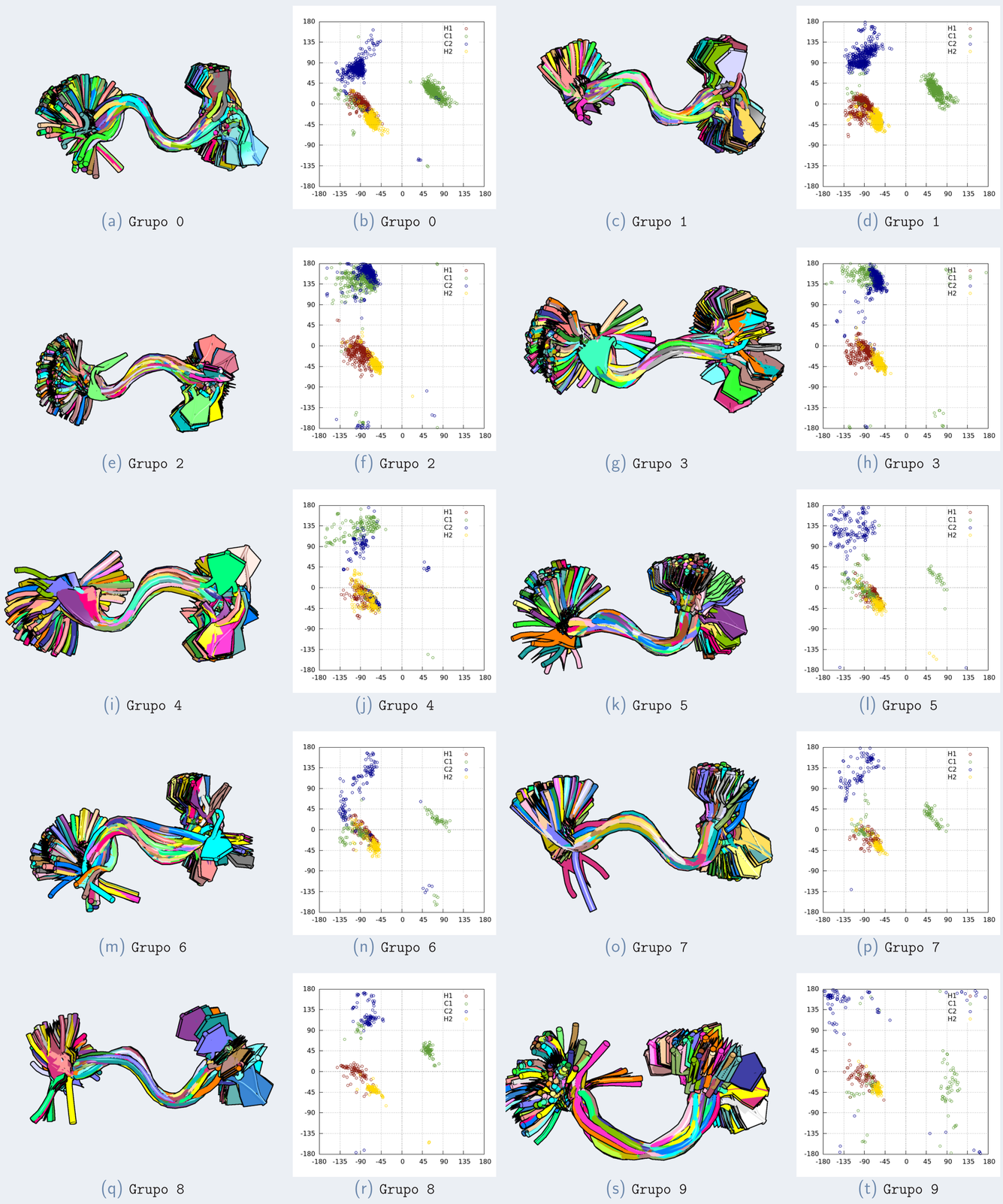
Figure 1 : Preferência Corfomacional das estruturas H e C.

A partir destas informações, foi feita uma fragmentação de estrutura secundária (SS) dos resíduos a fim de encontrar padrões mais específicos: *H - qualquer SS*, *E - qualquer SS*, *C - qualquer SS*, *G - qualquer SS*, *B - qualquer SS* ou *T - qualquer SS*. Ao final desta etapa foram encontrados ∼5000 diferentes padrões estruturais. Para as análises seguintes, foi selecionado o padrão **HCCH** contendo um total de 2978 ocorrências.
Clusterização: a partir dos ângulos phi/psi foi feita uma comparação par a par de todos os pedaços calculando o RMSD destes, com os dados sendo armazenados em uma matriz de distância. RMSD é a média das distâncias entre átomos de proteínas, a qual é utilizada para medir a semelhança entre estruturas 3D proteicas. Pares de estruturas em que o RMSD tem valor igual a 0 são consideradas iguais, e conforme vão se diferenciando será atribuido um valor positivo.
A fim de agrupar os padrões, foi aplicado um algoritmo de clusterização do pacote Scipy[4], cujo método utilizado é o de clusterização hierárquica. A clusterização hierárquica utiliza apenas critérios de similaridade para encontrar grupos nos quais as instâncias são mais semelhantes entre si. Dentre os diferentes métodos de agrupamento foi utilizado o método *complete-link* que considera a distância entre dois grupos como a maior distância entre todos os possiveis pares de elementos dos dois grupos. Como critério de semelhança se utilizou um cutoff de 1.0 [5], visto que pares de 4 resíduos são considerados semelhantes quando tiverem um RMSD de até 1.59 $\mathring{A}$, conforme a equação:

$$\text{RMSD}_{cutoff} = (N)^{1/3}$$, onde N representa o número de resíduos.

Foram selecionados os primeiros 10 grupos mais populosos, para os quais foram gerados gráficos de Ramachandran, alinhamento dos pedaços via Pymol e atributos (aminoácidos, polaridade e ângulos chi (χ)).
Classificação: a partir dos atributos de cada grupo, foram feitos testes utilizando uma metodologia genérica, em que se observou resultados preliminares interessantes. Foi usado o pacote de ferramentas Weka [6] o qual contem técnicas de mineração de dados. Foi aplicado o método de classificação Random Forest, o qual é capaz de classificar uma grande quantidade de dados. A eficácia do modelo é calculada através de uma matriz de confusão, a qual mostra o número de classificações reais contra o número de classificações preditas em cada grupo. A diagonal da matriz representa as classificações corretas, enquanto que o restante representa as classificações incorretas.

Resultados e Discussão



(a) Grupo 0 (b) Grupo 0 (c) Grupo 1 (d) Grupo 1
(e) Grupo 2 (f) Grupo 2 (g) Grupo 3 (h) Grupo 3
(i) Grupo 4 (j) Grupo 4 (k) Grupo 5 (l) Grupo 5
(m) Grupo 6 (n) Grupo 6 (o) Grupo 7 (p) Grupo 7
(q) Grupo 8 (r) Grupo 8 (s) Grupo 9 (t) Grupo 9

Figure 2 : Estruturas e Gráficos de Ramachandran dos 10 grupos mais densos.

Os resultados dos gráficos de Ramachandran da Fig. 1 demonstram uma preferência das estruturas Coils por 3 determinadas áreas (ângulos), ilustradas de maneira colorida no gráfico. Estas diferentes áreas levam a uma maior chance de erro nas predições visto que há variação da preferência conformacional. Analisando os gráficos da Fig. 2, é possível ver uma preferência conformacional das estruturas Coil em uma menor área (ângulos). Esta menor área leva a diminuição da região de busca em métodos de predição baseados em conhecimento para casos de estruturas Coil, reduzindo as chances de erros. O alinhamento das estruturas de cada grupo mostra a semelhança destas, o que representa os padrões/preferências de ângulos para cada grupo gerado.
Com a formação de cada grupo e os atributos dos mesmos gerados foi feita uma análise de classificação afim de medir a eficiência do método de clusterização e se os grupos formados e suas respectivas estruturas estavam classificadas corretamente. No pacote de ferramentas Weka, foi utilizado esses dados de atributos como input e foi gerado uma árvore de classificação a partir da metodologia Random Forest. A análise da árvore foi demonstrada através de uma matriz de confusão, a qual apresentou uma eficiência de 71.03% para o método.

Conclusão

Os resultados preliminares corroboram para podermos afirmar que um método de classificação pode ser aplicado nesses dados possibilitando, assim, um conhecimento a mais para ser aplicado em métodos de conhecimentos de predição de estruturas 3D. Como perspectivas futuras, temos a implementação de um método de classficação específico para o problema, aplicação da metodologia realizada no trabalho nos outros padrões estruturais. Por fim, aplicar este conhecimento em um método de predição para estruturas 3D, visando sua melhoria.

Referências

1 Lehninger, A.L., Nelson, D.L., Cox, M.M. *Principles of Biochemistry.* Freeman, 2005.
2 Dorn, M., Barbachan e Silva, M., Buriol, L.S., Lamb, L.C., **Three-dimensional protein structure prediction: methods and computational strategies.**, *Computational biology and chemistry* 53 (2014): 251-276
3 Heinig, M., Frishman, D., **STRIDE: a Web server for secondary structure assignment from known atomic coordinates of proteins.**, *Nucl. Acids Res.* 32 (2004): W500-2.
4 http://www.scipy.org/
5 Skolnick, J., Reva, B.A., Finkelstein, A.V., **What is the probability of a chance prediction of a protein structure with an rmsd of 6 $\mathring{A}$?**, *Folding and Design*, Vol. 3, (1998): 141-147.
6 Witten, I.H., Eibe, F., Hall, M.A., *Data Mining: practical machine learning tools and techniques.*, 3^{rd} edition.

Acknowledgement:

FAPERGS CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico propesq Pró-Reitoria de Pesquisa CAPES