SALÃO DE INICIAÇÃO CIENTÍFICA
XXVIII SIC
paz no plural

Predição da Flexibilidade Conformacional de Resíduos de Aminoácidos através de Neuroevolução

Bruno Iochins Grisci Márcio Dorn
Instituto de Informática, Universidade Federal do Rio Grande do Sul
bigrisci@inf.ufrgs.br, mdorn@inf.ufrgs.br

Resumo

Este trabalho aborda o desafio da predição da estrutura tridimensional de uma dada sequência de aminoácidos, o que foi relatado pertencer à classe dos problemas NP-Completos. É apresentado um novo método baseado na evolução de redes neurais artificiais através de NeuroEvolução de Topologias Crescentes e em agrupamento hierárquico para extração de características estruturais de proteínas determinadas experimentalmente e definir a flexibilidade conformacional de uma sequência de aminoácidos alvo. A técnica proposta manipula informação estrutural do Protein Data Bank para gerar intervalos de ângulos de torção com probabilidades associadas para cada aminoácido em uma sequência alvo, representando a sua flexibilidade conformacional. Essa informação pode ser usada para predizer a estrutura tridimensional de sequências proteicas desconhecidas e ajudar na redução do espaço de busca conformacional de moléculas de proteína em métodos de predição da estrutura de proteínas baseados em conhecimento. O método proposto foi testado com uma variedade de proteínas e os resultados indicam que ele de fato é uma opção funcional de representar a flexibilidade de aminoácidos.

1. Introdução

A Bioinformática Estrutural trata de problemas nos quais as regras que determinam processos e relações bioquímicas são apenas parcialmente conhecidos, o que torna difícil o projeto eficiente de estratégias computacionais para estes problemas. Um deles, o problema da predição da estrutura 3D de proteínas, é especialmente importante pois a estrutura tridimensional de um polipeptídeo permite a inferência da função de uma proteína no organismo. As proteínas são formadas por uma cadeia de aminoácidos. O método proposto neste trabalho utiliza informações estruturais do Protein Data Bank (PDB) para predizer a flexibilidade de aminoácidos em uma cadeia polipeptídica.

Este trabalho tem como objetivo o desenvolvimento de métodos e estratégias computacionais para o problema de predição in silico da estrutura tridimensional de proteínas. O método proposto, baseado em redes neurais coevolutivas, busca identificar padrões conformacionais em proteínas cuja estrutura foi experimentalmente determinada através de métodos de cristalografia por raio-X ou Ressonância Magnética Nuclear. A partir destes padrões, busca-se predizer a flexibilidade de aminoácidos de uma cadeia polipeptídica de proteínas cuja estrutura 3D ainda não é conhecida.

2. Materiais e Métodos

PROTEÍNAS são polímeros formados por uma sequência de 20 possíveis diferentes aminoácidos que sob condições fisiológicas enovelam-se em formas precisas conhecidas como seu estado nativo. Um peptídeo é uma molécula composta por dois ou mais aminoácidos ligados por ligação peptídica. A interação entre os aminoácidos em uma proteína fazem a cadeia polipetídica dobrar-se, normalmente em uma configuração própria, como uma hélice ou folha. Esses padrões de enovelamento descrevem a estrutura secundária. A topologia da proteína é dada pelo tipo de sucessão de estruturas secundárias conectadas e produzem a forma que essas estruturas se organizam no espaço 3D [3].

Uma maneira de representação da estrutura é pensarmos na forma tridimensional de uma proteína em termos das suas rotações internas. Nesse caso, a forma de dois aminoácidos vizinhos pode ser descrita pelos ângulos de torção ao redor do átomo de Cα, chamados de phi (ϕ) e psi (ψ), com valores que podem variar entre $-180°$ e $180°$. Estes ângulos internos são denidos por conjuntos de quatro átomos sucessivos na cadeia principal da proteína [1].

O método faz uso majoritário de dois algoritmos de aprendizado de máquina. O primeiro é o agrupamento hierárquico. Agrupamento é uma técnica computacional usada para agrupar dados pelas suas similaridades. O agrupamento hierárquico foi o selecionado dentre as diferentes opções por não ser necessário o conhecimento prévio do número final de grupos graças ao uso de um limiar de distância e um funcionamento aglutinador [2].

O segundo algoritmo fundamental ao método é NEAT (NeuroEvolution of Augmenting Topologies), que evolui redes neurais usando algoritmos genéticos. A vantagem de NEAT é não ser necessário determinar a topologia das redes neurais de antemão, permitindo que elas se adaptem ao problema e dados fornecidos, gerando uma arquitetura minimalista e especializada [4].

3. Método proposto

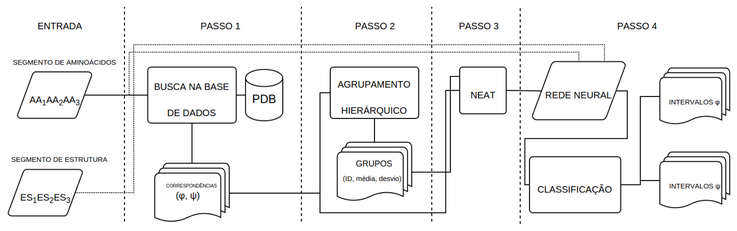

Figura 1: *Resumo esquemático do método.*

O método proposto é iniciado com a divisão de uma sequência de aminoácidos em segmentos menores que são buscados em proteínas presentes no PDB. A partir da lista de proteínas contendo os segmentos, são extraídos os pares de ângulos de torção correspondentes aos segmentos. Tais pares são agrupados com o algoritmo de agrupamento hierárquico, criando-se um conjunto de dados compostos de padrões contendo os aminoácidos do segmento juntamente com sua estrutura secundária e grupo atribuído pelos pares de ângulos de torção.

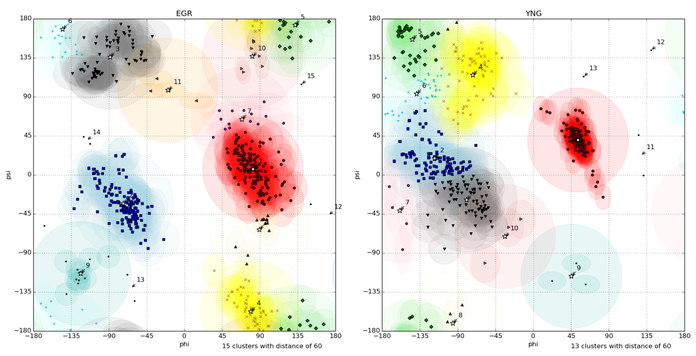

Figura 2: *Exemplo de grupos criados.*

Com estes dados são treinadas redes neurais para a aprendizagem da classificação de aminoácidos nos grupos obtidos anteriormente, permitindo generalização para novas sequências de aminoácidos. As redes neurais são treinadas com NEAT, um algoritmo de neuroevolução.

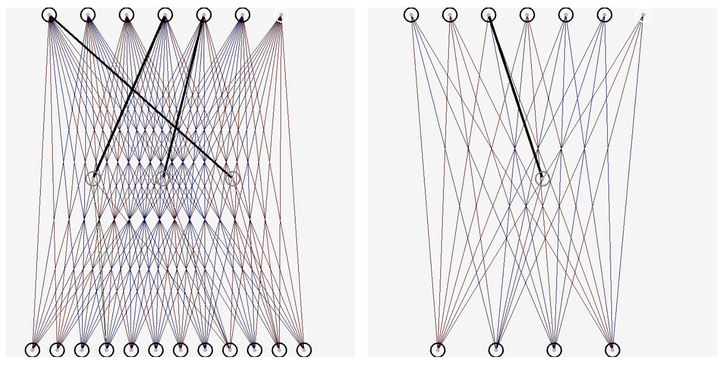

Figura 3: *Exemplos de redes neurais evoluídas com* `NEAT`.

A sequência de aminoácidos original é submetida às redes neurais, que retornam as probabilidades de cada aminoácido pertencer a um dos grupos criados. Com esta informação são criados intervalos de valores para os ângulos de torção de cada aminoácido centrados nos valores médios dos grupos e com as probabilidades associadas obtidas com as redes neurais.

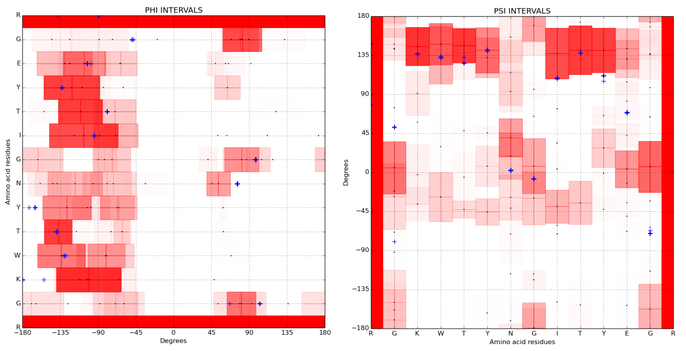

Figura 4: *Exemplo de intervalos gerados.*

4. Experimentos e resultados

O método foi testado com 25 proteínas de estrutura conhecida. Os resultados revelam que os intervalos gerados contêm informações estruturais compatíveis com os dados experimentais, e podem ser utilizados para a visualização da flexibilidade estrutural de cadeias de aminoácidos e redução do espaço dos dados em estratégias de busca, auxiliando na obtenção de métodos de predição mais eficientes e precisos para a obtenção de conformações aproximadas às experimentais.

5. Conclusão

O método proposto foi capaz de predizer a flexibilidade de aminoácidos, podendo ser incorporado em estratégias de busca para o problema da predição da estrutura 3D de proteínas. Uma versão preliminar deste trabalho foi aceita em forma de artigo e apresentação oral, com o nome ”Predicting Protein Structural Features with NeuroEvolution of Augmenting Topologies”, em co-autoria de Bruno Iochins Grisci e Márcio Dorn, no IEEE World Congress on Computational Intelligence realizado entre 24 e 29 de julho de 2016 em Vancouver, Canadá.

6. Agradecimentos

ESTE trabalho foi parcialmente financiado por recursos da FAPERGS (00202125.51/13), MCT/CNPq (473692/2013-9) e CNPq (311022/2014-4), Brasil.

Referências

[1] Márcio Dorn, Luciana S. Buriol, and Luis C. Lamb. Moirae: A computational strategy to extract and represent structural information from experimental protein templates. *Soft Computing*, 18(4):773–795, 2013.

[2] SC Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(2):241–254, 1966.

[3] A. M. Lesk. *Introduction to Protein Science*. Oxford University Press, New York, 2 edition, 2010.

[4] Kenneth O. Stanley and Risto Miikkulainen. Evolving neural networks through augmenting topologies. *Evolutionary Computation*, 10(2):99–127, 2002.