Estimando Domínio de Proteínas por Abordagem Estatística

Daniel Ferreira de Souza Reis, Reinaldo Viana Alvares
Centro Universitário Augusto Motta – Ciência da Computação
21041-020, Bonsucesso, Rio de Janeiro, RJ
E-mail: danielfsr18@gmail.com, reinaldoviana@gmail.com

Resumo: *O objeto de estudo abordado nesse trabalho compreende investigações em bancos de dados de proteínas, tendo como foco o uso de uma estratégia computacional, baseada em regras estatísticas, visando a identificação de domínios proteicos, os quais representam uma relevante região das proteínas, normalmente associadas com suas funções.*

Palavras-chave: Domínios proteicos, Bioinformática, Estatística

1 Introdução

Este trabalho envolve estudo de proteínas armazenadas no banco de dados PFAM [1], o qual se encontra disponível para livre acesso e uso. Proteínas são as estruturas mais complexas e sofisticadas funcionalmente que se conhece. O corpo humano produz aproximadamente cem mil proteínas, tendo em cada uma, centenas ou mesmo milhares de aminoácidos.

Há quatro tipos de estruturas de proteínas: primária, secundária, terciária e quaternária, de acordo com os aminoácidos que possui, do tamanho da cadeia e da configuração espacial da cadeia polipeptídica. Foi escolhido o tipo primário que é o nível estrutural mais simples e mais importante. Proteínas são classificadas em famílias e *clans*¸os quais representam superfamílias.

Proteínas geralmente possuem um domínio funcional. Diferentes combinações de domínios podem dar origem a uma diversificada gama de proteínas encontradas na natureza. A identificação dos domínios de uma proteína, pode portanto, fornecer informações sobre sua função.

2 Abordagem Baseada em Regra

A estrutura baseada em regras visa estimar o domínio de uma sequência primária tendo como base o tamanho (número de aminoácidos) da sequência. Para um melhor entendimento sobre uma sequência e um domínio de uma proteína, vide modelo a seguir, na Figura 1, em que:

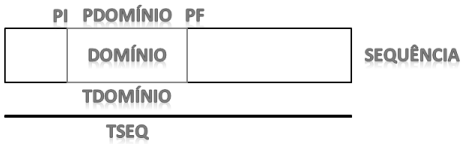

Figura 1. Representação de uma sequencia proteica e seu domínio

Sequência: representa a sequência de aminoácidos da proteína.
TSeq: representa o tamanho ou comprimento da sequência.
Domínio: corresponde ao domínio da proteína.
TDomínio: indica o tamanho ou comprimento do domínio.
PDomínio: representa o tamanho relativo do domínio em relação a TSeq.
PI: corresponde à posição inicial do domínio, em percentual, em relação a TSeq.
PF: corresponde à posição final do domínio, em percentual, em relação a TSeq.

Foi realizado um estudo estatístico no banco PFAM, visando detalhar o parâmetro PI. Percebeu-se que em 42% dos casos o domínio inicia em região inferior aos primeiros 10% do tamanho da sequência. Observou-se que em geral, o domínio ocupa 50% do tamanho da sequência. Então, formulou-se hipótese, por meio de uma regra, visando estimar o domínio de uma sequência qualquer, da seguinte forma:

DOI: 10.5540/03.2015.003.01.0080

DomínioCandidato = (**Inícioc**, **PDomínioc**):

Inícioc indica em percentual, a posição inicial do domínio candidato em relação a TSeq. **PDomínioc** indica em percentual, o tamanho do domínio candidato em relação a Tseq.

A regra escolhida foi DomínioCandidato = (**2%Tseq**, **50%TSeq**):

3 Estudo de Caso

Foram escolhidas aleatoriamente sequencias que pertencessem a *clans* com a condição de que estes deveriam ter agrupado a si entre 8 e 10 famílias. Então escolheu-se em média 30% das sequencias de cada superfamília, totalizando 500 sequências. A assertividade expressa em forma de percentual para avaliação do domínio foi:

$$\frac{\text{EST}=(\text{DomReal} \cap \text{DomCandidato})*100}{\text{DomReal}}, \text{ onde}$$

EST = Estatística percentual de acerto entre domínio real e domínio candidato.
DomReal = Domínio real da sequência.
DomCandidato = Domínio candidato da sequência proposto em regra.

Foram gerados três histogramas (figura 2), que permitem validar a aplicação da regra de maneira gráfica por meio da assertividade das sequências dentro do seu respectivo *clan*.

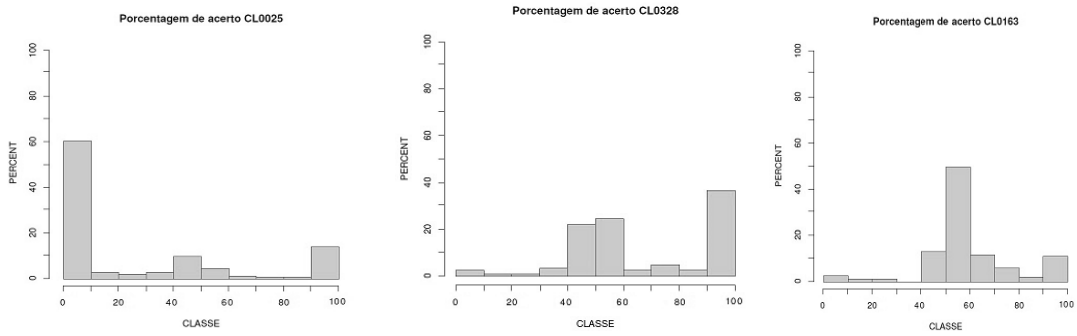

Figura 2. Histograma dos *clans* CL0025, CL 0328 e CL0163

3 Conclusões

Depois de aplicados todos os conceitos aqui retratados , verifica-se que a regra apresenta mais de 50,9% de acerto entre as médias dos três *clans* e suas sequências utilizadas como amostra. Testes em amostra maior são recomendados como trabalhos vindouros. O desempenho médio de métodos puramente estatísticos é cerca de 60% [2]. regras mais refinadas também representam motivação para continuidade do trabalho.

Referências

[1] Punta M., Coggill P.C., Eberhardt R.Y., Mistry J., Tate J., Boursnell C., Pang N., Forslund ., Ceric G., Clements J., Heger A., Holm L., Sonnhammer E.L.L, Eddy S.R., Baeman R.D. The Pfam protein families database. 2012. Nucleic Acids Research. Database Issue 40:D290-D301.

[2] Veretnik S., Gu J., Wodak S. Identifying Structural Domains in Proteins. 2009. Structural Bioinformatics, Second Edition. John Wiley & Sons, Inc.

