Predição de classes de enzimas usando método de agrupamento super-paramagnético

Marcelo Boareto, Vitor B P Leite
Ibilce - Unesp - SP - Brasil

Michel Eduardo Beleza Yamagishi
Embrapa - SP - Brasil

Nestor Caticha
IF - USP - SP - Brasil

As proteínas com atividade enzimática podem ser divididas em seis grandes classes de acordo com suas funções específicas, que são: oxirredutases, hidrolases, transferases, lyases, isomerases e ligases [1]. Utilizando algoritmo de agrupamento não-paramétrico, este trabalho tem como objetivo predizer diferentes classes enzimáticas usando parâmetros estruturais e físico-químicos. Esta abordagem é justificada uma vez que a função das proteínas está relacionada com a sua estrutura e com as suas características físicas e químicas. O algoritmo não-paramétrico, conhecido como *Super Paramagnetic Clustering* (SPC) [2], é baseado no modelo físico de Potts e não assume nenhuma hipótese sobre a distribuição dos dados. Cada elemento recebe um valor de variável spin, de forma que se dois elementos tem o mesmo valor de spin, eles são considerados pertencentes à um mesmo grupo. Um conjunto de interações entre vizinhos próximos é introduzida alterando a configuração das variáveis spin a cada incremento na temperatura do sistema. Assim, através de correlações spin-spin observamos a divisão dos dados de acordo com classes naturais presentes nos mesmos, associadas às diferentes magnetizações. Como informação relevante para tal predição utilizamos a frequência de cada aminoácido nas proteínas bem como os seus paramêtros físico-químicos, hidrofobicidade, cargas na superfície, área acessível ao solvente, energia de interação, volume molecular, porcentagem de alfa-hélices, porcentagem de folhas-beta e porcentagem de aminoácidos na superfície. Estes parâmetros foram extraídos da base de dados STING-DB [3]. Em um primeiro passo o objetivo foi classificar as proteínas entre enzimas e não enzimas, em um conjunto não-redundante formado por 1044 enzimas e 474 não-enzimas. Na análise subsequente as enzimas foram classificadas segundo o SPC buscando uma concordância com a classificação usual da literatura [1]. Os resultados preliminares indicam uma eficiência superior a 55 % (uma classificação aleatória baseada na frequência de cada classe conseguiria cerca de 30%). Apesar desta abordagem ainda estar sujeita à otimização, o desempenho de classificacao já é superior aos métodos existentes. Formas de otimização na performance de classificação também são discutidas.

[1] Dobson PD, Doig AJ, *J. Mol. Biol.* **345**: 187-199 (2005).
[2] Blatt M, Wiseman S, Domany E, *Phys. Rev. Lett.* **76**: 3251-3254 (1996).
[3] Neshich G, Togawa RC, Mancini AL, Kuser PR, et al, *Nucleic Acids Res.*, **31**:3386-3392 (2003).