VINÍCIO FRAGOSO MENDES

UMA ESTRATÉGIA HIERÁRQUICA E ESCALÁVEL PARA CLASSIFICAÇÃO ESTRUTURAL E FUNCIONAL DE PROTEÍNAS

Dissertação apresentada à Universidade Federal de Viçosa como parte das exigências do Programa de Pós-Graduação em Ciência da Computação para obtenção do título de *Magister Scientiae*.

Orientadora: Sabrina de Azevedo Silveira

Coorientador: Giovanni Ventorim Comarela

VIÇOSA - MINAS GERAIS
2019

Resumo

MENDES, Vinício Fragoso, M.Sc., Universidade Federal de Viçosa, setembro de 2019. **Uma estratégia hierárquica e escalável para classificação estrutural de proteínas.** Orientadora: Sabrina de Azevedo Silveira. Coorientador: Giovanni Ventorim Comarela.

A predição da classificação estrutural proteica é uma tarefa relevante, mas desafiadora e complexa, onde os dados estruturais das proteínas possuem grandes quantidades de informação a respeito de suas funções e relação entre proteína e seu gene codificante. Com o aumento dos dados biológicos publicamente disponíveis, há uma demanda por métodos computacionais para organizar, anotar e compreender os dados. Cada vez mais, são necessárias as tentativas de atribuir automaticamente a classificação estrutural ou da função proteica. Com o grande montante de dados reconhecidos e depositados, é difícil ou até mesmo impossível inferir manualmente a classificação proteica. Este trabalho propõe uma estratégia de aprendizado supervisionado para realizar a classificação estrutural de proteínas, com um interesse particular em modelos hierárquicos. Para avaliar a estratégia proposta, foram realizados três experimentos utilizando dados estruturais de proteínas disponíveis em bancos de dados biológicos (CATH, SCOPe e BRENDA). Cada conjunto de dados está associado a um esquema de classificação hierárquica bem conhecido (CATH, SCOP, EC *Number*). Primeiro os dados estruturais contendo a posição de cada átomo no espaço 3D foram modelados como uma matriz de distância (CSM - *Cutoff Scanning Matrix*). Em seguida, a quantidade de dados foi reduzida e parte do ruído removido, ambos a partir da aplicação do SVD (*Singular Value Decomposition*) à matriz. Em seguida, foi utilizada a matriz reduzida como entrada para o modelo, que é capaz de prever corretamente a classificação na maioria das vezes. Foi mostrado que a precisão do modelo varia de 86% a 95% ao prever a classificação de CATH, SCOP e EC *Number*, valores compatíveis ou superiores ao estado da arte em alguns casos.

Palavras-chave: Classificação hierárquica de proteínas. CATH. EC *number*. SCOP.

Abstract

MENDES, Vinício Fragoso, M.Sc., Universidade Federal de Viçosa, September, 2019. **A hierarchical and scalable strategy for protein structural classification.** Advisor: Sabrina de Azevedo Silveira. Co-advisor: Giovanni Ventorim Comarela.

The prediction of protein structural classification is a relevant but challenging and complex task, where the structural data of proteins have large amounts of information about protein function, data from the literature and the relationship between protein and its coding gene. With increasing publicly available biological data, there is a demand for computational methods to organize, annotate, and understand the data. With the large amount of data recognized and deposited, it is difficult or even impossible to manually infer protein classification. This work proposes a supervised learning strategy to perform the structural classification of proteins, with a particular interest in hierarchical models. To evaluate the proposed strategy, three experiments were performed using structural data of proteins available in biological databases (CATH, SCOPe and BRENDA). Each data set is associated with a well-known hierarchical classification scheme (CATH, SCOP, EC Number). First, the structural data, containing the position of each atom in the 3D space, were modeled as a distance matrix (CSM - Cutoff Scanning Matrix). Then the amount of data was reduced and some of the noise removed, both from the Singular Value Decomposition application (SVD) to the mentioned matrix. Then the reduced matrix was used as input to the model, which is able to classify protein structures according to different classification schemes. The accuracy of the model has been shown to range from 86% to 95% by predicting the levels of CATH, SCOP and EC Number. To the best of our knowledge, this work is the first one to achieve such high accuracy when dealing with large scale datasets.

Keywords: Protein hierarchical classification. CATH. EC number. SCOP.