USING SINGULAR VALUE DECOMPOSITION AND LATENT SEMANTIC INDEXING TO SEARCH STRUCTURAL SIGNATURES IN SUBTLASE PROTEIN FAMILY

V.M.G ALMEIDA

Federal University of Minas Gerais - Bioinformatics PhD Program,
Av. Antonio Carlos, 6627 - Belo Horizonte, Brazil
E-mail: valdetemg@ufmg.br

S. A. SILVEIRA, M. A. CREPALDE AND M. A. SANTOS

Federal University of Minas Gerais - Department of Computer Science,
Av. Antonio Carlos, 6627 - Belo Horizonte, Brazil
E-mail: sabrinas@dcc.ufmg.br, mirlaine@dcc.ufmg.br and marcos@dcc.ufmg.br

R. M. MINARDI

Federal University of Minas Gerais - Bioinformatics PhD Program,
Av. Antonio Carlos, 6627 - Belo Horizonte, Brazil
E-mail: raquelcm@dcc.ufmg.br

C. SILVEIRA

Federal University of Itajuba - Department of Mathematics and Computer Science,
mailbox 50 - Itajuba, Brazil
E-mail: carlos.silveira@unifei.edu.br

J. C. D. LOPES

Federal University of Minas Gerais - Department of Chemistry,
Av. Antonio Carlos, 6627 - Belo Horizonte, Brazil
E-mail: jlopes@netuno.lcc.ufmg.br

M. M. SANTORO

Federal University of Minas Gerais - Department of Biochemistry-Immunology,
Av. Antonio Carlos, 6627 - Belo Horizonte, Brazil
E-mail: santoro@icb.ufmg.br

Abstract

The proteins are versatile macromolecules in living systems. They are important to crucial functions in many biologic processes. They work like catalysts, to carry and store others molecules, for exemple the oxygen, to give help and immune protection, to provide movement, to transmit nerve impulses, and to control growth and differentiation. All this diversity of biochemical functions is done by combination of 20 monomer units, called amino acids. In the polymerization process, the amino acids connect linearly to each other and built extensive peptide chains called primary structure.

The proteins are divided in families, by their biologic functions or standard mechanism used to their inhibitors. This situation can be observed, for exemple, on two subfamilies of non homologous Serine proteases, the Trypsin family and Subtilase family, that don't share similar tridimensional structure, but hydrolise their substrate and are inhibited by their inhibitors using same mechanism.

In this work we describe a method to analyze the interactions between proteins and protein-inhibitors of Subtilase family, with the objective of finding structural patterns. We based on the hypothesis that triplets of amino acids, forming motifs, can be a pattern found in protein families and these triplets could be identified with more frequency in residues that are part of the protein/inhibitor interface.

The LSI technique (Latent Semantic Indexing) based on the intelligent information Retrieval, contemplating the knowledge of the linear algebra associated in the SVD (Singular Values Decomposition), was used in an attempt of recovering a group of information, that is, patterns, based in triplets (motifs) in the protein Subtilase family.

The results were applied to proteins from PDB, with the objective to find other proteins that possess the same identified pattern for the Subtilase family. The proteins returned by our algorithm can possibly form complexes with Subtilases inhibitors.

Preliminary results indicate that the discrimination of Subtilases through the analysis of its protein/inhibitors interface was moderate. The accuracy obtained for the top 100 proteins which had the highest scores was 0.51. According to the similarity (defined as cosine between a query protein and a protein from our database), the Subtilases and their inhibitors took values with a minimum of 0.5, up to 0.95 and a median of 0.75. Among the top 100 proteins, some of them were not Subtilases nor had relationship

with its inhibitors and they obtained a median of 0.8.

We are analyzing why proteins not classified as Subtilases showed high similarities. This may be indicating that our algorithm was able to find similar interface patterns between seemingly unrelated proteins. If true, opens the way for the holding of new drugs and therapeutic targets. We are also examining ways to improve our ranking, trying to identify the best set of parameters that composes an ideal compromise between singular values and similarity.