Screening and Functional Prediction of Conserved Hypothetical Proteins from Escherichia coli

William F Porto¹, Simone Maria-Neto¹, Diego O Nolasco¹, Octavio L Franco¹

¹Centro de Análises Proteômicas e Bioquímicas, Pós-Graduação em Ciências Genômicas e Biotecnologia, Universidade Católica de Brasília, Brasília - DF, Brazil, 70790-160

Introduction: Thousands of proteins have been annotated as conserved hypothetical proteins. Protein structures may provide some evidence of function, helping in protein annotation when identity is less than 40% to a known protein. Efforts to identify the functions of hypothetical proteins could bring novel advances in our understanding of biological systems. To this end, a new strategy to mine the NCBI’s non-redundant protein database in the search for candidates for function prediction was here described. Material and Methods: The data mining strategy was applied to conserved hypothetical proteins from Escherichia coli. Briefly, data mining selects small conserved hypothetical proteins (30-100 amino acid residues) without significant templates on Protein Data Bank, without transmembrane regions and with similarity to Eukaryote proteins. Through this strategy, 13 protein sequences were selected from a total of 11,505,486 sequences deposited in the NCBI’s non-redundant protein database. Results and Discussion: From these, only two sequences could be modelled and have their functions predicted. GI 26251020 is related to cupredoxins, which can be found in Bacteria and Eukarya domains, while GI 227886634 seems to be able to bind to lipids, showing structural similarities to phospholipases from Eukaryotes. Conclusion: The method described in this paper can be relevant to mining databases or genomes, looking for hypothetical sequences that can be targets for in vitro and/or in vivo functional characterization.

Keywords: Data Mining, Structural Genomics, Molecular Modelling, Molecular Dynamics