Available online at www.sciencedirect.com
SCIENCE DIRECT®
Data & Knowledge Engineering 53 (2005) 31–43
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

ConsDiff: an algorithm for the detection of conserved differences between protein sequences

Saumil Mehta, Deendayal Dinakarpandian *

Division of Computer Science and Electrical Engineering, School of Computing and Engineering, University of Missouri, 5110 Rockhill Road RHFH550E, Kansas City, MO 64110, USA

Received 22 June 2004; accepted 22 June 2004
Available online 3 August 2004

Abstract

Proteins have been classified into families based on metrics of similarity such as sequence or structural similarity. However, there are significant differences in function even within families. Mapping these differences to individual amino-acid residues is typically done by an expert. This is a subjective and non-scalable approach. ConsDiff is an algorithm that automates this process. It is based on a set of parametric rules using amino-acid substitution matrices and a multiple sequence alignment. This allows the automated discovery of candidate residues that may be responsible for critical differences in function, which may then be experimentally verified.

Keywords: Algorithm; Knowledge discovery; Multiple sequence alignment; Protein function; Amino-acid substitution matrix; Sequence analysis

* Corresponding author. Tel.: +1 816 235 5942; fax: +1 816 235 5159.
E-mail address: dinakard@umkc.edu (D. Dinakarpandian).

0169-023X/$ - see front matter
doi:10.1016/j.datak.2004.06.012

1. Introduction

1.1. Organization of paper

We begin with a brief background on bioinformatics that sets the context for the theme of this paper. We then present key concepts that represent the underpinnings of this paper. This is followed by a detailed description of the problem that we address. We then present the proposed solution, describing the algorithm and the implementation with examples drawn from the real world.

1.2. Background

The work described here addresses the general issues of the synergism between bioinformatics and experimental biology, increasing reliance on the comparison of sets of entities, the use of conserved patterns to suggest functional importance, and the need to discover and implement algorithms that can simulate domain experts. We explain each of these general concepts in this section before giving a detailed presentation of the problem in the subsequent section.

1.2.1. Synergism between bioinformatics and experimental biology

Bioinformatics and experimental biology represent a mutually reinforcing relationship where knowledge from one drives the design and development of the other. For example, an important role of bioinformatics today is that of predicting experimental outcomes. In turn, the development of experimental techniques for the large-scale generation of data has spurred research and applications of many areas of computer science like data warehousing, data mining, probabilistic searches and knowledge discovery. It is important to note that, in a large proportion of cases, a prediction made by bioinformatics does not necessarily give a definitive answer. Rather, it merely, but importantly, narrows the search space for verification by subsequent trial-and-error experiments. In this context, too, it serves a very useful purpose, as not all experimental approaches are equally scalable. Many kinds of experimental techniques continue to be highly labor intensive and any computational prediction of limiting the number and/or type of experiments to be performed remains quite valuable. For example, while it is possible to automate the generation of copies of a gene, actual preparation and obtaining pure amounts of the corresponding protein is partially an art that continues to require highly skilled manual intervention.

1.2.2. From one-at-a-time to many-at-a-time paradigm. Corollary: From 1:1 to n:m comparisons

A general class of problems that is important in bioinformatics is that of comparing two objects using some metric of similarity. Depending on the kind of data and level of abstraction, the object may be atomic, a set, a permutation, or a bag (multi-set) of attributes. In biological terms, respective examples are an amino-acid, the proteome, a protein and the genotype (including duplicate alleles).

For example, the score in an amino-acid substitution matrix may be approximated to be an index of similarity between any two amino-acids, normalized to the random evolutionary interchangeability between pairs of amino-acids [10]. Alternatively, at a higher level of abstraction, two proteomes may be subjected to an exhaustive comparison of n versus m proteins using a heuristic pair-wise sequence alignment algorithm like BLAST [2]. A composite score of similarity between the two proteomes may be derived from the matrix of pair-wise scores.