Pattern Discovery: a web based interface for exhaustive analyses on multiple biological sequences

Raimondo E (1,2), Chiusano ML (2)

(1) PhD fellow in Computational Biology, Interdepartmental Research Center for Computational and Biotechnological Sciences, Second University of Naples, Naples, Italy
(2) Department of Structural and Functional Biology, University 'Federico II', Naples, Italy

Motivation

The search for common motifs in biological data is fundamental to find structural correlation that could be informative of functional and/or evolutionary relationships. In its simplest form, the Generic Pattern Discovery Problem on biological sequences can be formulated as the problem of finding all the patterns that occur in at least K sequences in a given sample of n elements. However, the relevant discovery problem is NP-hard. As a consequence, existing algorithms are commonly based on two main approaches: either they settle for incomplete results in order to achieve reasonable performance (approximation algorithms), or their execution time is suitable only for medium-sized inputs. We discuss here on a web based methodology to support pattern discovery in biological sequences. The algorithm proposed is based on a variant of the deterministic approach by Rigoutsos and Floratos (1). The web based approach aims to provide useful tools for suitable mining on multiple sequences. The novel algorithm is designed to overcome computational limits of an exhaustive pattern discovery approach.

Methods

We designed a novel algorithm for pattern discovery based on the TEIRESIAS method described in (1). TEIRESIAS solves the problem of finding all the maximal patterns occurring in at least K sequences in a set of n elements. Considering "maximal patterns" reduces the output redundancy as well as the computational complexity typical of an exhaustive search, while the "density" of the patterns can be user-driven by the input parameters L and W, where L indicates the minimum number of defined characters in every sub-pattern which length is at most W. A preliminary "scanning" phase determines all of the elementary patterns, i.e. patterns which length is at most W, with exactly L defined characters. Then, a combinatorial search for more specific patterns, starting from the set of elementary ones, follows. This step of the algorithm is termed the "convolution" phase. In our strategy, the scanning phase is the same as in (1), while the convolution step is rather different. Indeed, the main drawback in TEIRESIAS algorithm is the need of memory space when stacking all the possible extensions of elementary patterns in a representative set of long sequences. We re-designed the algorithm by splitting the convolution phase into a number of different steps, which, in their turn, are based on several convolve-send-receive cycles. Patterns generated at each step are used, during the following step, to generate new patterns, and then removed, as they won't be used any more. This leads to a more efficient data management and memory usage. A friendly PHP-based interface has been built to exploit the algorithm for different purpose in biological sequence analyses and to allow the software usage through web access.

Results

Our efforts have been focused on two principal aspects: a novel algorithm strategy and a suitable PHP-based interface that supports specific analyses on both nucleic acids and protein data. Our strategy is based on the re-organization of an exhaustive algorithm (1) to provide a more efficient implementation in terms of space requirements. Moreover, our method results suitable for a parallel implementation, improving time computational costs. The web based interface to the algorithm was designed to exploit the algorithm for solving varied specific analyses on biological sequence data as well as to allow user-driven mining on the reported results.

Contact email: enrico.raimondo@unina2.it

References

1. Rigoutsos, I., and Floratos, A. (1998) Bioinformatics 14(1):55-67.

Supplementary informations

This work is supported by the Agronanotech Project (MIPAF, ITALY)