Searching for discriminating degenerated patterns between two populations of sequences

Nicola Cannata, Claudio Forcato, Giorgio Fabbro, Anna Pasin, Julie Balen and Giorgio Valle

CRIBI Biotechnology Centre, Università di Padova, Viale G. Colombo 3,
35121 Padova – ITALIA
nicola@cribi.unipd.it

Keywords. Annotation, Pattern discovery, Simplified Alphabets

Introduction

In this work we present the development of a bioinformatics tool aiming at the individuation of discriminating sequence patterns between two populations of sequences. Some examples in which it could be used are easy to find in genomics and proteomics: introns/exons in gene sequences, coding/non-coding in transcript sequences, proteins that are transported in some subcellular localization and those that are not. Once the patterns are detected they could be searched over non-annotated sequences from some program especially developed to find degenerated patterns. We expect that such a method, used jointly with other more traditional methods could lead to a better predictive power in annotation processes.

Methods

What is innovative here is the use of simplified alphabets that permit to "see" sequence patterns that are not normally visible using the conventional nucleotididic and aminoacidic symbols [1]. The grouping of letters (e.g. the simplest alphabets with only two simbols, Strong/Weak in nucleotides, Polar/Hydrophobic in aminoacids) permits to take in account intrinsic chemical-physical properties of the sequence and have been already used for example in finding consensus sequences in multiple alignment or in protein folding studies.
With the ALPHASIMP program [1] is possible to generate sistematically all the possible alphabets from a set of letters, choosing those that satisfy some given properties. The analyzed input sequences are then rewritten using the so-generated alphabets and searched for the presence of statistically surprising patterns, that are represented much more frequently than expected. This is done with an enhanced version of the pattern discovery program DISCOVER 1 [2], which is able to individuate patterns, with a suggested number of defined and undefined positions, that are found much more (or much less) frequently than expected.
The two instruments permit to build up a list of "typical patterns", expressed in some simplified alphabets for a particular population of sequences. As shown in fig. 1, by comparing the typical patterns of two sets of sequences is possible to detect some candidates pattern (present only in one of the two populations) to use as discriminating tool to classify the not yet annotated sequences.
We already applied the program to intron/exon discrimination taking as input the annotated sequences from the Exon-Intron Database [3]. It is easy (exon sequences are written in capital letters) to automatically separate the two sequence populations of introns and exons of some model organisms (H. *sapiens*, M. *musculus*, C. *elegans*, D. *melanogaster*). For both the sequence sets we detected the typical patterns and by comparison we found some possible intron/exon discriminating characteristic pattern for that species.

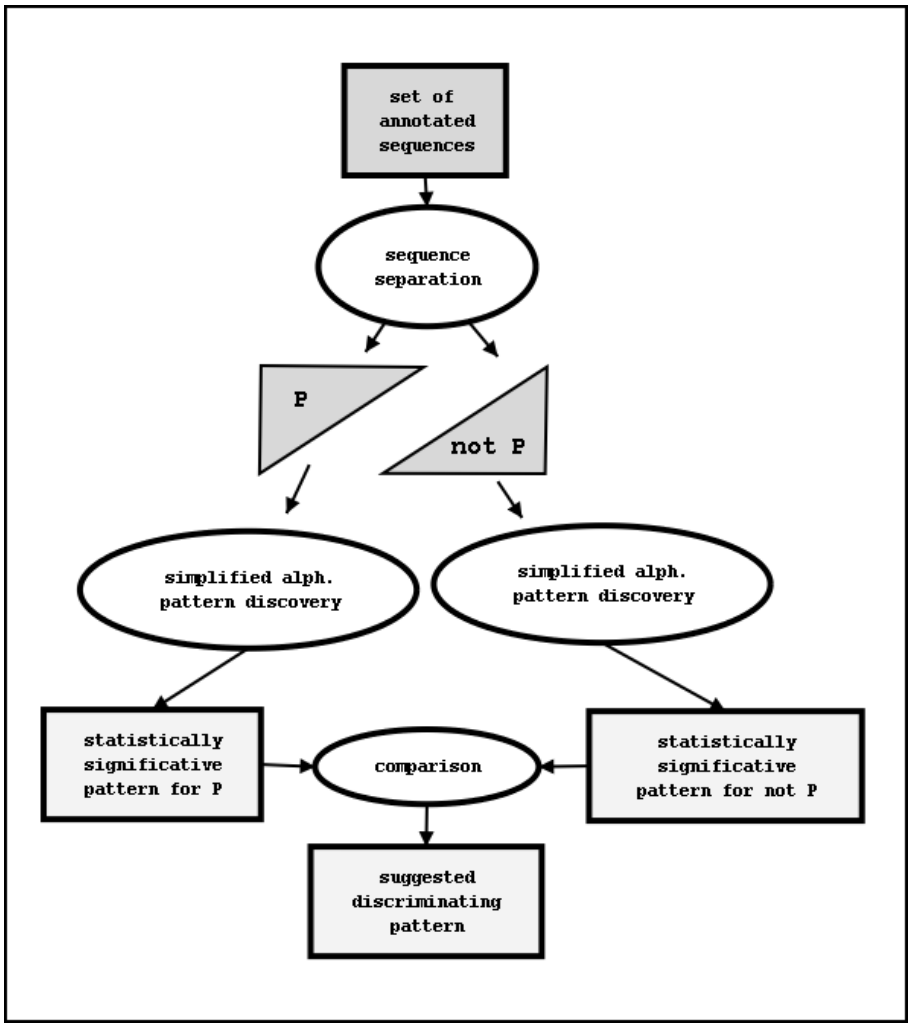

Fig. 1 The strategy adopted in our program for finding degenerated pattern suggested to discriminate between two disjoint sets of sequences (in one is valid a "P" property but not in the other set , indicated as "not P")

References

[1] Cannata, N., Toppo, S., Romualdi, C. and Valle, G. Simplifying amino acid alphabets by means of a branch and bound algorithm and substitution matrices. *Bioinformatics* 18 (8): 1102-1108 (2002)

[2] Valle, G. Discover 1: a new program to search for unusually represented DNA motifs. *Nucleic Acids Res* 21 (22): 5152-6 (1993).

[3] Serge Saxonov, Iraj Daizadeh, Alexei Fedorov, and Walter Gilbert. The Exon-Intron Database; An exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res*. 28(1):185-190 (2000).