Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Theoretical Computer Science 335 (2005) 1–2
Theoretical Computer Science
www.elsevier.com/locate/tcs

Foreword

One of the ultimate goals of Genomic Research is to extract "structure" and "meaning" from biological sequences. Knowledge of "structure" and "meaning" may lead to better understanding of genetic diseases and to better methods for the design of proteins. Pattern Discovery has therefore assumed a fundamental role in this context. It deals with the development of automated methods for the inference of structure that presents itself in the form of regularities, subaggregates and associations thereof in Biosequences. The PROSITE Data Base is a good example. This is a database of protein families, where each family is described by a regular expression (the common "motif" or "pattern") highlighting important domains of "consensus" among the sequences of the given protein family. In PROSITE, motifs have been extracted semi automatically through the analysis of the alignment of multiple sequences: such a procedure is no longer adequate, given the amount of biological data being produced worldwide.

In general, the definition of motif/pattern emerges from the biological context in which these terms are used. Satellites and tandem repeats are repetitive structures, important to identify genetic markers and for their association to some genetic diseases; patterns such as subsequences common to a set of sequences indicate conservation, which in turn suggests that the pattern is implicated in an important (sometimes, still unknown) function; promoter sequences and regulatory regions have a specific "high level" structure which is believed to be essential to the machinery of cells; and so on.

Exploration of most of these issues has only begun: as websites and data banks accumulate known proteins, DNA sequences, and 3D structures in growing numbers, increasingly fast and sophisticated methods are sought. Pattern and Motif Discovery is perhaps the single most pervasive tool of Bioinformatics, ubiquitous to specialized (Prosite, Splash, Meme, Verbumculus, Teiresias, etc.) as well as more general, classical ones such as Blast, Bioaccelerator, etc. Antagonist trends represented by data explosion on the one hand, and a growing need for integration and cross-correlation on the other, make advances in motif and rule discovery a foremost need in this domain. Searches for relationships in the growing collection of data are carried out currently by manual annotation and experimentation. Capabilities akin to automatically cluster, classify, and annotate data across the traditional boundaries of individual databases is posing an increasing need for novel models and algorithms for the generation, analysis and cross-annotation of scattered biological data.

Pattern Discovery predates, on the one hand, over the rich arsenal of techniques and tools set up in the course of the development of Pattern Matching and, on the other, on some sophisticated probabilistic insights that accompanied the process. It differs from Pattern

doi:10.1016/j.tcs.2004.12.011

Matching in so far as the latter searches for finely prescribed patterns, whereas the former tries to unearth patterns that are loosely specified at best. The papers presented in this special issue offer a snapshot of the state of progress of this fascinating and challenging subject as applied to one of its most natural testbeds, the analysis of Biosequences. Because modelling is such an integral part of these studies, and often the only relevant validation for them, the papers feature as a rule a non-negligible experimental component.

The manuscripts were submitted in response to the Call for Papers for this Special Issue, and have been refereed according to the standards and procedures that TCS applies to its normal submissions.

Alberto Apostolico
Purdue University and Università di Padova
E-mail address: axa@cs.purdue.edu

Raffaele Giancarlo
Università di Palermo
E-mail address: raffaele@math.unipa.it