Chapter 15

New Trends in Graph Mining: Structural and Node-Colored Network Motifs

Francesco Bruno
DEIS, Università della Calabria, Italy

Luigi Palopoli
DEIS, Università della Calabria, Italy

Simona E. Rombo
DEIS, Università della Calabria, Italy

ABSTRACT

Searching for repeated features characterizing biological data is fundamental in computational biology. When biological networks are under analysis, the presence of repeated modules across the same network (or several distinct ones) is shown to be very relevant. Indeed, several studies prove that biological networks can be often understood in terms of coalitions of basic repeated building blocks, often referred to as network motifs. This work provides a review of the main techniques proposed for motif extraction from biological networks. In particular, main intrinsic difficulties related to the problem are pointed out, along with solutions proposed in the literature to overcome them.

Open challenges and directions for future research are finally discussed.

INTRODUCTION

Recent years have witnessed the harvesting of enormous amounts of “rough” biological information which need to be interpreted. Such information include both string-shaped (e.g., nucleic and protein sequences) and complex-structured data, such as protein interaction networks and protein/RNA structures.

One of the main features characterizing biological data is the presence of regularities and repetitions, often associated to common biological functions and/or properties shared by the cellular components of different organisms. That is, biological data is characterized by the presence
of “syntactic” features modeling “semantic” properties common to the associated biological components. Such features become interesting when some particular conditions, related to the specific problem under analysis, are satisfied. For example, subsequences repeated more frequently than a fixed threshold in a set of protein sequences may be considered interesting, since this probably denotes that those subsequences embody some specific biological function common to the proteins in the given set.

When the repeated syntactic structures are worth being considered interesting, they are generally called motifs for the biological data in which they repeatedly occur, and it is usually important to look for them and single them out.

The problem of motif discovery has been thoroughly addressed during the last few years, specifically in the context of analyzing string-shaped data, and many important results have been obtained (see, e.g., (Brazma et al. 1998; Rombo et al., 2007) for surveys on this topic). Motif extraction from complex biological structures is relevant from the application viewpoint as well (Apostolico et al., 2008b; Apostolico et al. 2008; Cheng et al. 2008), as further discussed below. Indeed, following the huge efforts paid towards completing the genome coding of several organisms, a large deal of attention is now turning towards the analysis of how cellular components interact in order for biological functions to be realized (von Mering et al., 2002; Miller et al., 2005). Interactions among cellular components are most often modeled by biological networks (Alm et al., 2002), where the nodes represent components (e.g., proteins, genes, enzymes) and edges among nodes denote interactions among corresponding components. Different types of biological networks are currently available for analysis purposes, such as protein-protein interaction networks, gene regulatory networks and metabolic networks. Also, several studies (Cheng et al., 2008; Mangan et al., 2005; Milo et al., 2002; Shen-Orr et al., 2002) lead to the important observation that, similarly as with sequences, biological networks can be often understood in terms of basic repeated building blocks that are network motifs. And, for this reason, several techniques have been recently devised to single out motifs of biological networks. The aim of this paper is precisely that of providing a review on the main approaches to network motif extraction presented in the literature.

Preliminarily, it is worth noting that, technically speaking, the problem at hand presents several intrinsic difficulties such as, for instance, the need to deal with (variants of) the sub-graph isomorphism problem (a well known NP-hard problem (Garey et al., 1979)) when searching for identical or quasi-identical sub-graphs within several input graphs. In this respect, it will be interesting to point out how such difficulties have been dealt with in the context of the various proposed techniques. Also, different kinds of network motifs may be defined, referring to both network topology and node similarities, which is conducive to different approaches to motif extraction problems. This issue will also be highlighted in the following.

This is not the first review written on biological network motifs, rather, a few of them can be found in the literature (Alm et al., 2002; Ciriello et al., 2008; Willy, 2009). In more detail, (Alm et al., 2002) deals with transcriptional regulatory networks, discussing two types of network thereof, that are, sensory networks, modeling responses to signals such as stresses and nutrients, and developmental networks, that guide differentiation events. The author of (Alm et al., 2002) discusses four motif families that seem to cover most of the known interactions in the transcription networks of E. coli and S. cerevisiae, appearing to be the main building blocks of these sensory networks. In (Ciriello et al., 2008), a comprehensive survey on the main models and algorithms presented in the literature for motif discovery in protein-protein interaction networks is illustrated. In (Ciriello et al., 2008), the authors focus on both algorithmic issues and statistical aspects concerning the significance of repeated sub-graphs and the comparison between