Pattern Recognition 41 (2008) 3547 – 3561

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Structure-based graph distance measures of high degree of precision

Yanghua Xiao[a,*], Hua Dong[b], Wentao Wu[a], Momiao Xiong[b,c], Wei Wang[a], Baile Shi[a]

[a]Department of Computing and Information Technology, Fudan University, P.O. Box 200433, Shanghai, China
[b]Theoretical Systems Biology Lab, School of Life Science, Fudan University, Shanghai, China
[c]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX 77225, USA

ARTICLE INFO

Article history:
Received 26 March 2008
Received in revised form 5 June 2008
Accepted 9 June 2008

Keywords:
Graph distance
Distance metric
Structure-based graph distance
SNP linkage disequilibrium

ABSTRACT

In recent years, evaluating graph distance has become more and more important in a variety of real applications and many graph distance measures have been proposed. Among all of those measures, structure-based graph distance measures have become the research focus due to their independence of the definition of cost functions. However, existing structure-based graph distance measures have low degree of precision because only node and edge information of graphs are employed in these measures. To improve the precision of graph distance measures, we define substructure abundance vector (SAV) to capture more substructure information of a graph. Furthermore, based on SAV, we propose unified graph distance measures which are generalization of the existing structure-based graph distance measures. In general, the unified graph distance measures can evaluate graph distance in much finer grain. We also show that unified graph distance measures based on occurrence mapping and some of their variants are metrics. Finally, we apply the unified graph distance metric and its variants to the population evolution analysis and construct distance graphs of marker networks in three populations, which reflect the single nucleotide polymorphism (SNP) linkage disequilibrium (LD) differences among these populations.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

As a universal data structure, graph has been widely used to model complex interaction relations among objects and define concepts. Compared to other data structures such as sequence, tree, graph is more sophisticated and more general, and consequently studies on graph have attracted research interest in various disciplines.

Many real applications [1–7] need to measure the similarity or distance between objects represented by graphs. For example, in computer vision and pattern recognition [2,3], similarity between unknown graph pattern and model graph pattern must be measured in the well-known graph matching process. In chemoinformatics [4–7], similarity searching based on 2D representation of molecular structure is one of the most common approaches to virtual screening, where some appropriate measure of inter-molecular structural similarity is the key of the success of the searching task.

Great efforts have been devoted to studying graph distance measures in different application domains over the past decades [8]. As a result, various graph distance measures have been proposed in the literatures [9–15]. These graph distance measures can be classified into three classes: *cost-based distance measures*, *structure-based distance measures* and *feature-based distance measures*. In Ref. [5], cost-based distance and structure-based distance are considered as one class, because it has been proved in Ref. [16] that given certain cost functions, the structure-based graph distance measures, such as graph distance measures based upon maximal common subgraph (MCS) [9],[1] are equivalent to corresponding edit distance measures with certain cost functions.

Considering error tolerance or error correcting, cost-based graph distances [17,18], e.g. graph edit distances, have been proposed, which are measured by the minimum edit cost to transform one graph into another. When defining graph edit distance, it is essential to define appropriate cost function for edit operations, which is usually based on the domain knowledge. Hence, cost-based graph distances give users opportunities to integrate domain knowledge into the definition of graph distance by parameterizing the cost

* Corresponding author. Tel.: +86 21 55075013.
E-mail addresses: shawyh@fudan.edu.cn (Y. Xiao), hdong0425@gmail.com (H. Dong).

0031-3203/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.06.008

[1] The term 'MCS' has been widely used, but it also has brought much confusion to the existing literatures. Strictly speaking, the graph distance metric proposed in Ref. [9] is based on *maximal common vertex induced subgraph*, abbreviated as MCIS, and some following graph distance metrics are based on *maximum common edge induced graph*, abbreviated as MCES. In this paper, to distinguish these two concepts, we will explicitly use MCIS or MCES, instead of MCS.

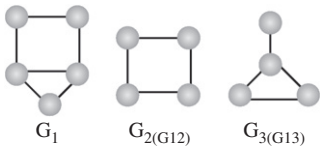

Fig. 1. Three graphs G_1, G_2, G_3 and two maximal common subgraphs G_{12}, G_{13}.

function. However, such flexibility also impose a severe limitation on graph edit distance in that it is difficult to define cost functions due to the variety of domain knowledge, despite the fact that great efforts have been dedicated to find the automatic procedures to infer edit operation costs [19,20]. Another class of graph distance measures is feature-based measures, which have also been widely studied in chemoinformatics and bioinformatics. In feature-based measures, distance or similarity has been measured according to the feature vectors derived from the chemical or biological structures. Hence, the effects of the feature-based measures heavily rely on the definition of the characteristic structures.

Compared to the other two classes of graph distance measures, structure-based distance measures do not rely on the cost functions and characteristic structures. In structure-based distance measures, the common substructure or superstructure has been considered as the measure of the degree of the similarity between graph patterns. Recently, some effective algorithms [4] to compute structure-based graph distance have been available, which further make structure-based measures, especially those measures based on MCS become the most popular graph distance measures in recent years.

Although various structure-based graph distance or similarity measures have been available, many graph pairs in some application domains cannot be correctly measured using these measures. For example, as shown in Fig. 1, given three graphs G_1, G_2 and G_3, we need to evaluate the similarity or distance among these graphs. If MCES-based distance metric, a widely used graph distance metric, is used, the MCS G_{12} (between G_1 and G_2) and the maximum common subgraph G_{13} (between G_1 and G_3) will have the same number of nodes and edges. Consequently, we can reach the conclusion that G_2 is similar to G_1 to the same extent as G_3 similar to G_1.

However, in the following sections, we will show that G_{13} contains much richer substructure information than G_{12}. As shown in Fig. 2, G_{13} contains some unique substructures, such as *triangle* and *star*, which do not appear in G_{12}. Hence, from such *substructure abundance* perspective, G_{13} is intuitively of more significance than G_{12}; and consequently, G_3 should be evaluated to be more similar to G_1 than G_2 to G_1. Therefore, the *richness of the unique substructures* occurring in a graph can contribute to the evaluation of graph distance, which is the basic principle underlying the measures we proposed in this paper.

Since nodes and edges are elementary constituents of a graph, size about nodes or edges in MCS will be a significant indication of the similarity between graphs, which is the fundamental idea of existing structure-based graph distance measures. For example, two representatives of them, MCIS-based graph distance [9] and MCES-based graph distance [4] use the number of *nodes* of MCIS, and the number of the *edges* of MCES, respectively, to evaluate the similarity between two graphs. However, in our studies, besides node or edge information in MCS, information about more complex and larger substructures in MCS will be utilized to evaluate distance between a graph pair.

In the following parts of the paper, we will show that structural differences between graphs can be amplified when considering information of larger substructures. Thus, if we evaluate graph distance in terms of certain larger or more complex substructures instead of some trivial substructures, such as nodes or edges, we can

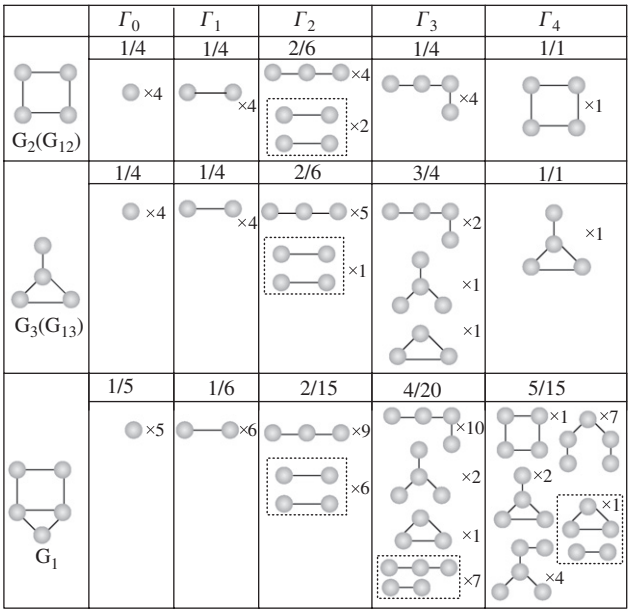

Fig. 2. Substructures in G_1, G_2 (G_{12}), and G_3 (G_{13}).

evaluate graph distance with higher degree of precision or in much finer grain than graph distance measures based on MCIS or MCES.

Evaluating graph distance according to *richness of the unique substructures* is also practically meaningful in many real applications. For example, in the analysis of protein interaction networks, such as protein–protein interaction network, protein–DNA and gene–gene interaction networks, it has been widely believed that substructures of these networks represent certain *functional modules* of cells or organisms. Thus, in Fig. 1, if *triangle* and *star* appearing in G_{13} are considered as functional modules of biological networks, then G_{13} will contain more functional modules than G_{12}. Consequently, we can naturally come to the conclusion that G_3 is more similar to G_1 than G_2 to G_1. Hence, comparing protein networks in terms of substructure information is biologically meaningful.

To accurately quantify graph distance is in great demand for many applications, especially for researches on evolution of biology networks. For example, we could use Bayesian Networks [21] to study SNPs [22] LD structure and their evolutions among different populations [23]. In such studies, how to measure similarity or distance among the constructed networks is an interesting but challenging problem. One of the great challenges is that traditional MCS-based graph distance metric can only evaluate the graph distance in much coarser grain, which cannot satisfy the requirement of identifying the minute difference between different population structures. Hence, it is of great need to devise new graph distance measures that can evaluate graph distances precisely.

2. Preliminaries

We begin this section with some basic notations. Let $G=(V,E,L,l)$ be a *labeled graph*, where V is the set of vertices, E is the set of edges and $E \subseteq V \times V$, L is the set of labels, and $l : V \cup E \to L$ is a labeling function that assigns a label to an edge or a vertex. Note that graph labeling is one of key issues in problems related to graph isomorphism. However, in some contexts, where graph isomorphism is not significant, *G also can be denoted as a* 2*-tuple* (V,E).

The vertex set of G is referred to as $V(G)$, and its edge set as $E(G)$. A *path* P in a graph is a sequence of vertices $v_1, v_2, \ldots, v_k$, where $v_i \in V$ and $v_i v_{i+1} \in E$. The vertices v_1 and v_k are linked by P and are called the *ends* of path P. The number of edges of a path is its *length*, and the

Download English Version:

https://daneshyari.com/en/article/533703

Download Persian Version:

https://daneshyari.com/article/533703

Daneshyari.com