Gene 679 (2018) 341–351

Contents lists available at ScienceDirect

Gene

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

Methodological paper

Novel symmetry-based gene-gene dissimilarity measures utilizing Gene Ontology: Application in gene clustering

Sudipta Acharya[a,*], Sriparna Saha[a,1], Prasanna Pradhan[b]

[a] Department of Computer Science and Engineering, IIT Patna, India
[b] Department of Computer Applications, Sikkim Manipal Institute of Technology, India

ARTICLE INFO

Keywords:
Gene Ontology(GO)
Dissimilarity measure
Symmetry-based distance
Gene clustering
Gene-GO term annotation matrix
Multi-objective clustering

ABSTRACT

In recent years DNA microarray technology, leading to the generation of high-volume biological data, has gained significant attention. To analyze this high volume gene-expression data, one such powerful tool is *Clustering*. For any clustering algorithm, its efficiency majorly depends upon the underlying similarity/dissimilarity measure. During the analysis of such data often there is a need to further explore the similarity of genes not only with respect to their expression values but also with respect to their functional annotations, which can be obtained from Gene Ontology (GO) databases. In the existing literature, several novel clustering and bi-clustering approaches were proposed to identify co-regulated genes from gene-expression datasets. Identifying co-regulated genes from gene expression data misses some important biological information about functionalities of genes, which is necessary to identify semantically related genes. In this paper, we have proposed sixteen different semantic gene-gene dissimilarity measures utilizing biological information of genes retrieved from a global biological database namely Gene Ontology (GO). Four proximity measures, viz. Euclidean, Cosine, point symmetry and line symmetry are utilized along with four different representations of *gene-GO-term* annotation vectors to develop total sixteen gene-gene dissimilarity measures. In order to illustrate the profitability of developed dissimilarity measures, some multi-objective as well as single-objective clustering algorithms are applied utilizing proposed measures to identify functionally similar genes from *Mouse genome* and *Yeast* datasets. Furthermore, we have compared the performance of our proposed sixteen dissimilarity measures with three existing state-of-the-art semantic similarity and distance measures.

1. Introduction

The main aim of performing clustering (Eisen et al., 1998; Yeung and Bumgarner, 2003) on gene expression data is to determine co-regulated genes having similarity in expression values with respect to a particular underlying distance measure. Therefore, the underlying similarity measure used by a clustering algorithm is an important factor which determines the efficiency of corresponding clustering algorithm. Though, clustering on gene expression data helps in identifying co-expressed genes but it fails to find out semantically related genes. Motivated by this, in the existing literature (Chagoyen et al., 2006; Del Pozo et al., 2008; Lim et al., 2007), authors have proposed several gene-gene semantic similarity/dissimilarity measures to identify semantically related genes utilizing available biological knowledge.

One such genuine source of biological knowledge is Gene Ontology (GO)[2]. GO comprises of several *GO-terms* having direct or indirect relationships with each other.

For several organisms their genes are annotated with specific *GO-terms* and this information can be downloaded from the GO website. A snapshot of GO sub-tree retrieved from GO website namely Gene Ontology Consortium (http://www.geneontology.org/) is shown in Fig. 1. It is increasingly gaining interests in defining functional relatedness using 'semantic similarity' of genes based on GO annotations (Chagoyen et al., 2006; Del Pozo et al., 2008; Lim et al., 2007). According to the past literature, several works have been done in developing semantic similarity measures between genes. All of the existing

Abbreviations: MOO, Multi-Objective Optimization; SOO, Single-Objective Optimization; AMOSA, Archived Multi-Objective Simulated Annealing; PD, Point symmetry based Distance; LD, Line symmetry based Distance; ED, Euclidean Distance; GO, Gene Ontology; BP, Biological Process; MF, Molecular Function; CC, Cellular Component

* Corresponding author.
E-mail address: sudipta.pcs13@iitp.ac.in (S. Acharya).
[1] Member, IEEE.
[2] http://www.geneontology.org/.

https://doi.org/10.1016/j.gene.2018.08.062
Received 11 June 2018; Received in revised form 21 August 2018; Accepted 21 August 2018
Available online 02 September 2018
0378-1119/

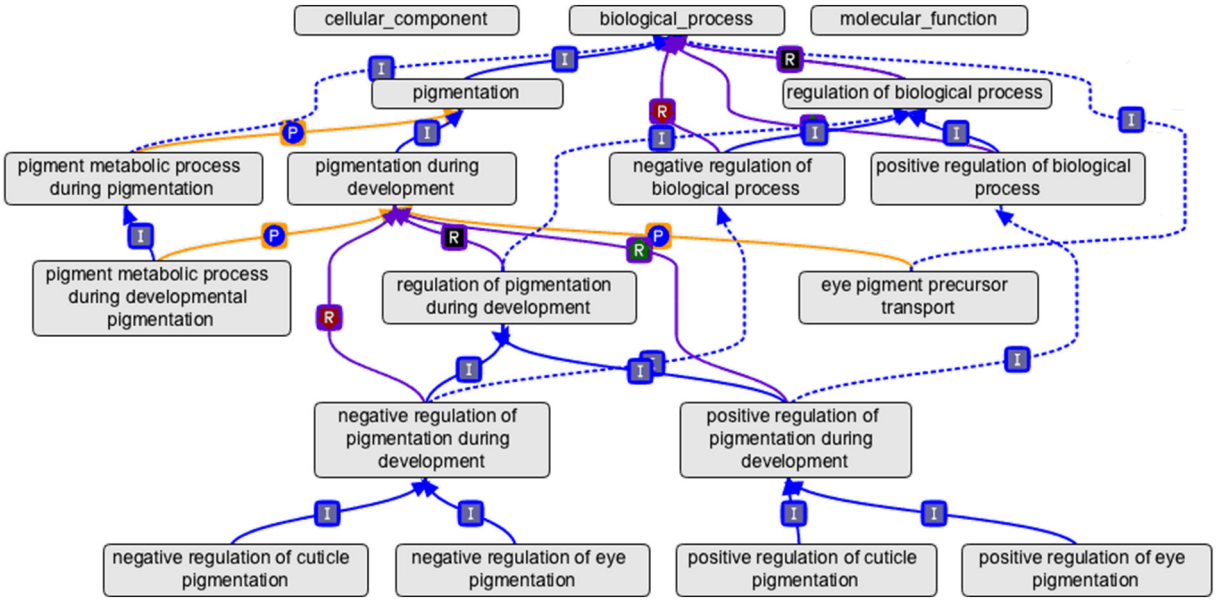

Fig. 1. Snapshot of GO retrieved from Gene-ontology consortium.

Table 1
Survey on existing literature of GO-based gene-gene similarity measures.

Literature	Proposed approaches
Jiang and Conrath (1997), Resnik (1995)	Pioneering works towards evaluating semantic similarity in a taxonomy; Concept of Information Content (IC) based similarity measure is proposed; The concept was conceived from the field of natural language processing where GO-based semantic similarity measures have been investigated by focusing on English language; Proposed measures were designed mainly for WordNet (Miller, 1995).
Lord et al. (2003)	First applied a measure of semantic similarity to GO by proposing a technique for calculating the semantic similarity of protein pairs based on Resnik's measure (Resnik, 1995).
Lin et al. (1998)	Proposed a IC-based measure considering ICs of common parent (lowest common ancestor or LCA) and term itself; relative distance of two terms from their LCA is taken into account.
Sevilla et al. (2005)	Analysed the correlation between gene expression and gene similarity computed by Resnik's, Jiang and Conraths' and Lin's measures of semantic similarity (Jiang and Conrath, 1997; Lin et al., 1998; Resnik, 1995).
Schlicker et al. (2006)	Introduced IC-based measure for computing the similarity between GO terms in GO based on a combination of Lin's and Resnik's techniques; It showed that proteins with highest sequence similarities tend to have similar molecular functions (Schlicker et al., 2006).
Martin et al. (2004)	Developed a set of methods for gene and GO term analysis; It uses counts of shared annotation terms among genes to compute similarities among genes.
Wang et al. (2007)	Proposed semantic similarity measure by considering topological information of GO graph; it not only considers LCA like previous literature but also all parent terms; No annotation information is utilized; Utilizing this measure (Du et al., 2009) G-SESAME web tool is developed.
Mistry and Pavlidis (2008)	Annotation set based semantic measure; It measures annotation set commonality between genes referred as Normalized term overlap (NTO) based method.
Pesaranghader et al. (2015)	Natural Language Processing (NLP) inspired definition based semantic similarity measure; Each GO term is represented as a definition vector; Cosine distance between definition vectors represents degree of similarity; Using proposed measure, SimDEF web tool is developed.
Shen et al. (2010)	Proposed a hybrid similarity measure that takes into account the path lengths between the terms as well as the IC of the ancestor terms.
Peng et al. (2015)	Proposed a hybrid similarity measure namely NETSIM (network-based similarity measure); It incorporates information from gene co-function networks in addition of using the GO structures and annotations; Here it is shown that incorporation of gene co-function network data clearly helps in improving the performance of *GO-term* similarity measures.
Peng et al. (2014)	Integrated approach to measure gene-gene similarity was proposed and the developed tool by the authors is named as *InteGO*; Gene-gene similarity is measured by considering three well-known existing similarity measures, namely, Yu et al. (2007), Wang et al. (2007) and Schlicker et al. (2006) and by integrating 'rank based gene-gene similarity' based on those measures; It automatically provides best integration policy (they named it as *seed measure integration*) to compute the similarity between given pair of genes.
Fröhlich et al. (2007)	Proposed GOSim package within R environment for similarity computations of genes and for gene clustering; Represent genes as feature vectors and during clustering the similarity between feature vectors x, y was considered as the normalized dot product between the vectors; In their developed R package for gene clustering, only hierarchical clustering can be applied on gene feature vectors.
Wolting et al. (2006)	Used a graph similarity measure (simUI implemented in Bioconductor in R) for computing gene similarity; PAM clustering algorithm is employed to group similar genes into clusters utilizing their proposed measures.
Mazandu et al. (2015), Mazandu and Mulder (2013)	Web tool A-DaGO-Fun (ADaptable Gene Ontology semantic similarity based Functional analysis) is developed which is repository of python modules; It provides flexibility to choose any semantic measure; Currently three clustering algorithms are incorporated: Hierarchical, K-means and community detecting model.