Revealing Distinct Association Patterns in Disease-Gene-Drug Based on Coupling Network and Subject-Action-Object-Triples

Abstract. A huge amount of associations among different biological entities (*e.g.*, disease, drug, and gene) are scattered in millions of bio-medical articles. Systematic analysis of such heterogeneous data can infer novel associations among different biological entities and make further efforts to propose novel therapeutic targets or decipher disease mechanisms. However, little research has been devoted to investigating associations among drugs, diseases, and genes in an integrative manner. In this paper, we use MEDLINE/PubMed data and extract biological entities and their associations by applying SAO (Subject-Action-Object) method. Further, we construct a three-layered coupling network to describe the connection between them and use community-detection algorithm to cluster the entities on different layers respectively. Our results investigate whether the community-detection algorithm could help prioritize disease genes based on the associations between diseases and their surrounding genes, and it will help researchers generate testable hypothesis of possible roles of genes in specific disease research. In addition, diseases in the same community are likely to be associated than the other diseases on the basis of "guilt by association" rule, which can also help researchers to infer new disease relationships. Second, this paper also suggests association pattern between disease and drug, in which two disease associated with each other are targets for the same drug, which support the hypothesis that similar disease can be treated by same drugs, allowing the opportunity for drug repositioning purpose.

Keywords: Subject-Action-Object (SAO); Three-Layered Coupling Network; Drug-Disease-Gene Linkage; Community-detection Algorithm

1 Introduction

A large amount of associations among bio-medical entities are scattered in bio-medical literature. Systematic analysis of such heterogeneous data provides bio-medical researchers with unprecedented opportunities to infer novel associations among different biological entities in the context of personalized medicine and translational research studies. Applications of investigating associations among drugs, diseases, and genes include disease gene prioritization (Kohler *et al*., 2008; Chen *et al*., 2009; Piro & Di, 2012), identification of disease relationships (Goh *et al*., 2007; Suthram *et al*., 2010) and drug repositioning (Arrell & Terzic ,2010; Dudley *et al*., 2011). However, majority of these approaches focus on relationships between only two kinds of entities (*e.g.*, association between gene and disease) and the associations among different entities are very sparse (Zhang *et al*., 2013). In this paper, we extract the biological entities and their associations on the basis of SAO method and propose a three-layered coupling network to analyze the complex associations between them. In addition, we apply community-detection algorithm to forcast the new associations among different entities.

2 Method

2.1 Subject-Action-Object method for entity extraction

The first step is selection of bio-medical articles. In this paper, we focus on top 10 published journals of the papers cited by marketed drugs approved by United States Food and Drug Administration (USFDA), which include *Journal of Medical Chemistry*, *Journal of Pharmaceutical Sciences*, *International Journal of Pharmaceutics*, *Proceedings of the National Academy of Sciences of the United States of America*, *Journal of Biological Chemistry*, *New England Journal of Medicine*, *Journal of Organic Chemistry*, *Pharmaceutical Research*, *Antimicrobial Agents and Chemotherapy and Cancer Research* (Du *et al*., 2018).

After collecting relevant papers, biological entities and associations are semantically annotated by using SAO method and concepts in the Unified Medical Language System (UMLS) from title and abstract section, which are precise and have been regarded as the most meaningful parts (Chen et al., 2003). The semantic information defined in the UMLS can be further leveraged to extract association through advanced methods.

2.2 Construction of three-layered Coupling network

We construct a three-layered coupling network (drug-disease-gene) with the following two steps:

First, after extracting biological entities and their associations from the above section, we obtain six different types of associations among them, such as *Disease-Disease*, *Disease-Gene*, *Disease-Drug*, *Drug-Gene*, *Drug-Drug*, *Gene-Gene*. Different entities are represented by different nodes and the associations among them are regarded as edges in the network.

Second, nodes representing disease, drug, and gene entities are placed on different network layers as shown in figure 1 and we establish correspondence between them for downstream analysis.

2.3 Community-detection algorithm

In this section, we divide the community structure in different network layers. For example, we set the community number of node v_i as σ_i (node v_j as σ_j), when $\sigma_i=\sigma_j$, δ (σ_i, σ_j)=1. Community density function Q can be defined as

$$Q = \frac{1}{2M}\sum_{i,j}[(a_{ij} - \frac{k_i k_j}{2M})\delta(\sigma_i, \sigma_j)] \tag{1}$$

In the above equation, a_{ij} represents elements in network adjacency matrix, $M=\sum a_{ij}/2$ represents the number of edges in community. k_i and k_j represent the degree of v_i and v_j respectively. A higher value of Q indicates a closer connections between nodes within the community

The general research framework is designed as follows (Figure 1).

3 Preliminary Results

Our results investigate whether the community-detection algorithm could help prioritize disease genes based on the associations between diseases and their surrounding genes, and it will help researchers generate testable hypothesis of possible roles of genes in specific disease research. In addition, diseases in the same community are likely to be associated than the other diseases on the basis of “guilt by association” rule, which can also help researchers to infer new disease relationships. Second, this paper also suggests association pattern between disease and drug, in which two disease associated with each other are targets for the same drug, which support the hypothesis that similar disease can be treated by same drugs, allowing the opportunity fro drug repositioning purpose.

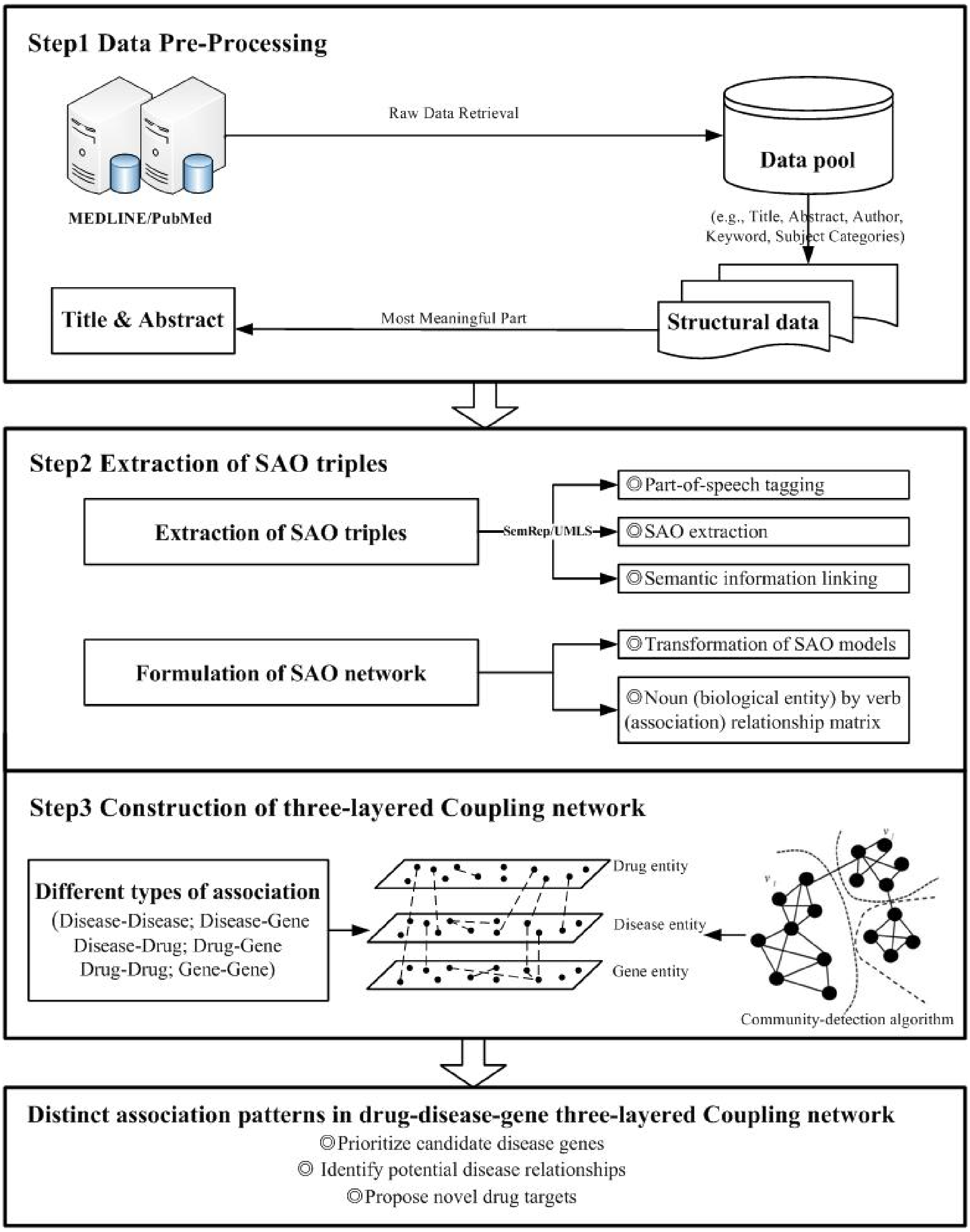
Step1 Data Pre-Processing
MEDLINE/PubMed
Raw Data Retrieval
Data pool
(e.g., Title, Abstract, Author, Keyword, Subject Categories)
Structural data
Most Meaningful Part
Title & Abstract
Step2 Extraction of SAO triples
Extraction of SAO triples
SemRep/UMLS
◎Part-of-speech tagging
◎SAO extraction
◎Semantic information linking
Formulation of SAO network
◎Transformation of SAO models
◎Noun (biological entity) by verb (association) relationship matrix
Step3 Construction of three-layered Coupling network
Different types of association
(Disease-Disease; Disease-Gene Disease-Drug; Drug-Gene Drug-Drug; Gene-Gene)
Drug entity
Disease entity
Gene entity
Community-detection algorithm
Distinct association patterns in drug-disease-gene three-layered Coupling network
◎Prioritize candidate disease genes
◎ Identify potential disease relationships
◎Propose novel drug targets

Figure 1. Framework for revealing distinct association patterns in disease-gene-drug