Abstract

The identification and characterization of important roles microRNAs (miRNAs) played in human cancer is an increasingly active area in medical informatics, and the prediction of miRNA target genes remains a challenging task to cancer researchers. We propose an innovative computing framework based on the Ontology for MicroRNA Target Prediction (OMIT) to facilitate knowledge acquisition from existing sources. The project aims to assist biologists in unraveling important roles of miRNAs in human cancer, and thus to help clinicians in making sound decisions when treating cancer patients.

Keywords

microRNA target prediction, ontology, semantic annotation, semantic integration, human cancer, bioinformatics, computational biology, biological computing

1. Introduction

Healthcare is a typical area where advances in computing have resulted in numerous improvements. In particular, the identification and characterization of important roles microRNAs (miRNAs) played in human cancer is an increasingly active area. MiRNAs are a class of small non-coding RNAs capable of regulating gene expression. They have been demonstrated to be involved in diverse biological functions [4, 6], and miRNAs' expression profiling has identified them associated with clinical diagnosis and prognosis of several major tumor types [3, 7, 5]. Unfortunately, the prediction of the relationship between miRNAs and their target genes still remains a challenging task [2, 1]. An example research scenario is as follows. Cancer patients' prognosis depends largely on their chemosensitivity (sensitivity to chemotherapy). Research has discovered that some specific genes increase the permeability of mitochondria (a cellular component) membrane, which in turn leads to apoptosis (cell death). As a result, the patient's chemosensitivity will increase and the chemotherapy will be more effective. Certain miRNAs can regulate the aforementioned genes and thus affect cancer patients' prognosis. If biologists were able to identify such miRNAs, a breakthrough on cancer treatment would have been made. However, this identification is very difficult: not only biologists need to extract a large number of candidate target genes from existing miRNA databases, but also they will need to manually search these genes' related information from resources other than miRNA databases for every one of hundreds of candidate target genes. The whole process is time-consuming, error-prone, and subject to biologists' prior knowledge.

On the other hand, ontologies are formal, declarative knowledge representation models, playing a key role in defining formal semantics in traditional knowledge engineering. The most successful example of applying ontological techniques into biological research is the Gene Ontology (GO) project [1]. Therefore, we propose a framework based on the Ontology for MicroRNA Target Prediction (OMIT) to handle the aforementioned challenge, and our overall objective is to explore a computing framework that will facilitate knowledge acquisition from existing sources, and assist biologists in unraveling important roles of miRNAs in human cancer. We aim to synthesize data from source miRNA databases into a comprehensive conceptual model that permits an emphasis on data semantics rather than on the forms in which the data was originally represented. Consequently, a more accurate, complete view of miRNAs' biological functions can be acquired. We thus provide users a single query engine that takes their needs in a nonprocedural specification format.

2. Methodology

The OMIT's overall structure is shown in Fig. 1. In order to develop a conceptual model that encompasses the structure and relationships described in the Gene Ontology (GO), the Gene Ontology Initiative (GOI) provides a computational framework that facilitates the creation of a consistent and comprehensive set of ontologies. The project aims to synthesize data from source miRNA databases into a comprehensive conceptual model that permits an emphasis on data semantics rather than on the forms in which the data was originally represented. Consequently, a more accurate, complete view of miRNAs' biological functions can be acquired.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM-BCB ’10 Niagara Falls, New York, USA
Copyright 2010 ACM 0-00000-00-0/02/10 ...$10.00.

1http://www.geneontology.org/index.shtml
required elements to properly describe medical informatics (especially in human cancer), it is essential to explore and abstract the miRNA data to the semantic level. The design of the OMIT will rely on two resources: existing miRNA databases and domain knowledge from cancer biologists. Besides cancer biology experts in the project team, there are six labs from around the world, (1) Yousef Lab in Israel, (2) DIANA Lab in Greece, (3) Sun Lab in Hong Kong, China, (4) Segal Lab in Israel, (5) Lin Lab in Taiwan, and (6) Wang Lab in St. Louis, MO, that have committed to actively participate in the project by providing original data sets and undertaking an in-depth analysis of integrated data and the query that follows. It is critical to present related gene information of miRNA targets to medical scientists in order for them to fully understand the biological functions of miRNAs of interest. We propose an innovative machine-learning algorithm to align the OMIT with the GO. Our approach is superior to most of state-of-the-art learning-based matching algorithms because we rely on ontology schema information alone, and do not require the assistance from instance data, which usually has constraints in either quality or quantity, or both. In particular, there are no instances at all in the GO (“GO, like most ontologies, does not use instances, and the terms in GO represent a class of entities or phenomena, rather than specific manifestations thereof”). Therefore, traditional learning-based ontology-matching algorithms, which depend heavily on instance data, are not appropriate in this scenario. In the OMIT system, an artificial neural network will be applied to learn weights for different semantic aspects. We design the hypothesis as a three-dimensional space consisting of three weights, and the learning objective is to find the vector that best fits the training examples. We adopt gradient descent as the training rule, and the searching strategy is to find the weight vector that minimizes the training error. Upon obtaining the learned weights for three semantic aspects, a similarity matrix will be calculated between the OMIT and the GO. An agglomerative clustering algorithm will then be adopted to find equivalent concepts, which in turn generate a set of mapping rules as the outcome. We will utilize W3C Rule Interchange Format-Production Rules Dialect (RIF-PRD), an XML language, to express such mapping rules.

Semantic annotation is the process of tagging source files with predefined metadata, which usually consists of a set of ontological concepts. We adopt a “deep” annotation that takes two steps. (1) To annotate the source database schemas, resulting in a set of mapping rules (specified in the RIF-PRD format) between OMIT concepts and elements from source database schemas. (2) The next step is to annotate data sets from each source, and the annotated data sets will be published in the resource description framework (RDF). Being a structure based on the directed acyclic graph model, the RDF defines statements about resources and their relationships in triplets. Such generic structure allows structured and semi-structured data to be mixed, exposed, and shared across different applications, and the data interoperability is thus made easier to handle. Based on the annotation outcomes, we will create a centralized RDF data warehouse, which better fits the project objective than a traditional relational data warehouse. We propose to adopt a “Globale-As-View (GAV)-like” approach to specify the correspondence between the source databases and the global schema, i.e., the OMIT ontology. Our approach differs from the traditional GAV approach in that we include aggregated, global data sets as well. As a result, user query will be composed according to OMIT concepts, and the query answering will be based on the centralized data sets with an unfolding strategy over the original query.

3. PRELIMINARY RESEARCH OUTCOMES

A first version of the OMIT ontology was designed using

---

2 http://www.geneontology.org/GO.ontology.relations.shtml
4 http://www.w3.org/RDF
Figure 3: Concept “MiRNA”

There are 117 concepts in total, including 13 top-level ones: “Disease,” “ExperimentValidation,” “Gene-Expression,” “GeneSequence,” “HarmfulAgent,” “MiRNA,” “OrganSystem,” “Organism,” “PathologyEvent,” “Protein,” “TargetGene,” “TargetPrediction,” and “Treatment.” Fig. 2 demonstrates a portion of OMIT concepts, while Fig. 3 shows the detailed design for the concept “MiRNA” that contains a set of properties and relationships: sub/superClassOf (also known as is-a), hasTarget, hasPrediction, regulateProtein, and inOrgan, etc. As exhibited in Fig. 1, the data flow of a typical knowledge acquisition is envisioned as:

1. The user sends a query to the OMIT system.
2. OMIT recognizes the miRNA concept in the OMIT is used to query the centralized RDF data warehouse.
3. OMIT retrieves miRNA targets.
4. The obtained targets are used to acquire more general information about the target gene.
5. Related gene information is returned.
6. The returned information is presented to the user.

When presenting a miRNA of interest, its potential targets can be retrieved by our system from existing miRNA databases. We then present such candidate targets to biologists for further validation. Besides obtaining miRNA targets for the user, additional information will be acquired from the GO. Based on the established alignment between the OMIT and the GO, related gene information can be easily retrieved, which is critical to fully understand the biological functions of the miRNA of interest. For example, suppose a cancer biologist is interested in investigating the chemosensitivity of breast cancer cells. By comparing chemosensitive and chemoresistant cancer cells, it is demonstrated that miR-125b, a specific miRNA, may confer the increased chemosensitivity of cancer cells. After the OMIT system obtains candidate targets for miR-125b, the gene information of these targets will be further acquired, including cellular localization (e.g., in mitochondria) and biological processes (e.g., apoptosis). The availability of such integrated knowledge will make it much easier for the cancer biologist to deduce the actual targets for miR-125b, and a breakthrough in breast cancer treatment may be granted. The corresponding RDF-based query is shown as follows:6

SELECT DISTINCT OMIT:targetGene
FROM OMIT:miRNA, GO-CC:cellComponent, GO-BP:bioProcess
WHERE OMIT:miRNA_ID = "miR-125b"
AND OMIT:miRNA:targetID = GO-CC:cellComponent:geneID

5http://protege.stanford.edu/
6"GO-CC" and "GO-BP" refer to the cellular component ontology in the GO and the biological process ontology in the GO, respectively.

We aim to provide users (biologists) a single query engine that takes their needs in a nonprocedural specification format. Such query is unified: although source miRNA databases are geographically distributed and usually heterogeneous among each other, the OMIT system presents biologists a uniform view of such heterogeneous data, along with integrated information from the GO.

4. CONCLUSIONS

In this work-in-progress paper, we propose the OMIT framework to handle the challenge of predicting target genes of miRNAs. This research will assist biologists in unraveling important roles of miRNAs in human cancer, and thus help clinicians in making sound decisions when treating cancer patients. The methodology has been discussed in detail. Our continuing progress will be updated in the project website.

5. ADDITIONAL AUTHORS

Additional author: Robert Matt Rudnick (School of Computer and Information Sciences, University of South Alabama, email: rmr501@jaguar1.usouthal.edu).

6. REFERENCES


