Precision Medicine Knowledge Base

Chair by Lei Liu, Lei Wang

Proposed format (times are approximate) with confirmed speakers:

（1）Building Knowledge Representation Model towards Precision Medicine

Speaker: Li Hou (15 minutes)

Abstract: With the rapid development of biomedical technology, great amounts of data in the field of precision medicine are growing exponentially. Valuable knowledge has been implicated in scattered data in which meaningful biomedical entities and their semantic relationships were buried. Therefore, it is necessary to develop a knowledge representation model to explicate relationships among diseases, phenotypes, genes, gene mutations, drugs and etc. Ontologies are recognized as an effective approach to achieve a consensus for knowledge representation. In biomedical domain, efforts have been made such as DO(Disease Ontology) for disease & syndrome, HPO(Human Phenotype Ontology) for phenotype and genetic disease, the UMLS (Unified Medical Language System) for medical vocabulary integration. On basis of existing work, our study focus on solving the following issues: (i) defining the top classes of precision medicine concepts; (ii) collecting and integrating biomedical vocabularies related to the above precision medicine concepts; (iii) defining and normalizing semantic relationships between different biomedical entities. In this study, we built a Precision Medicine Vocabulary(PMV), containing 4.53 million biomedical terms collected from 53 medical vocabularies, and developed a Precision Medicine Ontology(PMO), including ten top classes of precision medicine concepts such as Disease, Chemical and Drug, Phenotype and Gene, Mutation, Cell described by 93 semantic relationships between them. The PMO can be accessible in formats of RDF, OWL and XML. Our PMO and PMV can contribute the unified framework for describing the data and defining the semantic relationships in the field of precision medicine, and provide support for precision medicine studies and clinical applications.

（2）An Universal Framework for Multiple Biomedical Entity Relation Extraction from Biomedical Literature

Speaker：Ling Luo (15 minutes)

Abstract: With the rapid growth of biomedical literature, a large amount of the relations between important biomedical concepts (e.g. diseases, drugs, genes and proteins) are hidden in the literature. Biomedical entity relation extraction, aiming to automatically discover these relations with high efficiency and accuracy, is becoming an increasingly well understood alternative to manual knowledge discovery. However, most of the previous work on relation extraction from biomedical literature focuses on specific or predefined types of relations, such as protein-protein interactions, protein-gene interactions, drug-drug interactions, drug-disease treatment, and biomolecular events. In this work, we present a universal framework which can extract the relations between multiple biomedical concepts including diseases, drugs,

proteins, RNAs, DNAs, cells and phenotypes.

First, we propose a neural network method, i.e. attention-based bidirectional Long Short-Term Memory with a conditional random field layer (Att-BiLSTM-CRF), to document level Named Entity Recognition (NER). The method leverages document-level global information obtained by attention mechanism to enforce tagging consistency across multiple instances of the same token in a document. It achieves better performances with little feature engineering than other state-of-the-art methods on the BioCreative IV chemical compound and drug name recognition (CHEMDNER) corpus and the BioCreative V chemical-disease relation (CDR) task corpus.

Second, we propose a hybrid multiple biomedical entity relation extraction method. In this method, firstly, machine learning techniques are used to recognize binary entity relation. Then, the syntactic patterns and a dictionary are employed to find corresponding relation words that represent the relationships between two entities. This method obtains a much higher F-score than the rule based Stanford open information extraction method on the AImed corpus.

A DEMO of our method is available at http://202.118.75.18:8893/precision_medicine/index.html

（3）The construction of a Precision Medicine knowledgebAse Platform (PMap)

Speaker：Fan Zhong (15 minutes)

Abstract: Precision medicine is raised to prevent and treat diseases precisely depending on individual genome and other specific characteristics. A knowledgebase that integrated multiple knowledge from molecular mechanisms, clinical Information and cases, can help us to make a right decision on precision medical research and application. For this purpose, we have developed a Precision Medicine knowledgebAse Platform (PMap).

The PMap has overall integrated 42 databases in the form of a knowledge map, and its main frame comprises 4 parts: (1) genes and their products, (2) biological pathways and molecular networks, (3) disease-causing variations, and (4) drugs. The PMap includes annotations of 20,656 coding and 38,943 noncoding human genes, 178,562 RNAs, and 111,716 proteins. These genes and their products constitute the principal part of the entity repository in the PMap. So far, the PMap has integrated 21 pathway/network databases, which contains 13 main interaction categories, 22 biological effects, 28 modifications, and 1 experimental annotation. The whole network including canonical pathways involves 31,264 biological entities (nodes) and 1,804,000 interaction pairs (edges). The PMap has collected 5,738,719 disease-causing variations, which were from 18,022 genes and corresponding to 10,725 diseases. The information of 9,746 drugs and their 78,664 targets was stored in the PMap, which contains 561,180 drug-drug, 1,191 drug-food, 5,118 drug-enzyme, and 1,839 drug-transporter interactions. Further, the pMap will also integrate TCGA

(profiling matrix and clinic information) and ENCODE data.

（4）Combining biomedical literature annotation with manual curation

Speaker: Wan LIU; Yunchao LING (15 minutes)

Abstract: The main purpose of manual curation is taking advantage of human understanding or human intelligence, to correct mistakes in the process of extracting and constructing structured knowledge by computational automation, or artificial intelligence. However, structured information extraction in interdisciplinary science, such as precision medicine, is an iterative process of automatic annotation, manual curation, training set correction and recognition. Here, we developed platforms, such as distributed dictionary management system and manual curation platform, to collect and manage the massive scale resources of biomedicine, omics and clinical medicine, and to take advantage the power of manual curation to complement to DIKW (data, information, knowledge, wisdom) system, which is including bio-entity recognition, knowledge inference discovery and knowledge network construction.

(5) ProDiGy: towards the omics datasets analyses of precision medicine based on the PMap

Speaker: Dong Li (15 minutes)

Abstract: ProDiGy is a professional omics datasets knowledge discovery gateway, which is designed for exploring, visualizing, and analyzing omics profile datasets from organ, tissue or cell lines based on PMap (precision medicine knowledge map).

A friendly interface enables researchers to interactively search, filter and download datasets across samples. And several tools are integrated to provide data visualization, including CAPER, UbiBrowser, Pathview, SuperHeatmap, Network-view and Hierarchical network. And a galaxy system is used to integrate the database, data analyses and data presentation seamlessly. ProDiGy also established a MeSH ontology-based gene annotation and enRichment analysis systEm (MORE). For the gene list submitted by users, MORE will search co-occurrence gene-MeSH term relations in PubMed, and further utilize hypergeometric distribution to identify significantly enriched MeSH terms.

The intuitive Web interface of the portal not only makes omics datasets analyses accessible to researchers and clinicians freely and handily, and also thus facilitate medical and biological discoveries in precision medicine.

Free Discussion（15 minutes）