Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2018, Section Knowledge Representation and Management

Boudellioua I, Mahamad Razali RB, Kulmanov M, Hashish Y, Bajic VB, Goncalves-Serra E, Schoenmakers N, Gkoutos GV, Schofield PN, Hoehndorf R

Semantic prioritization of novel causative genomic variants

PLoS Comput Biol 2017;13(4):e1005500

This paper addresses the problem of how to distinguish which of the thousands of DNA sequence variants carried by an individual with a rare disease is (or are) responsible for the disease phenotype. This can help clinicians to reach a diagnosis, but also can be instrumental in improving the understanding of the underlying physiopathology of the disease. Many methods are currently available to help in identifying causative variant, *e.g.* by using information about species evolution and variant conservation or by the prediction of functional consequences of the variant DNA sequence. The authors has developed the PhenomeNET Variant Predictor (PVP) system that leverages semantic technologies and automated reasoning over genotype-phenotype relations to filter and prioritize variants in whole exome and whole genome sequencing datasets. In the heart of the system, a novel algorithm uses the patients' phenotype similarity to phenotypes in databases with known phenotype-genotype correlations to further rank potential candidate genes. In a retrospective study, they applied PVP to the interpretation of sequencing data in patients suffering from congenital hypothyroidism and showed that PVP accurately identified causative variants in whole exome and whole genome sequencing datasets. PVP provides a potentially significant resource for the discovery of causal variants.

The keys to the computational integration and comparison of phenotypes were (1) the systematic application of the PATO framework which provides a uniform way of describing phenotypes, and (2) the use of the UBERON ontology which can be used to systematically describe and relate anatomical structures between species. The PhenomeNET ontology also includes (as imports) GO (Gene Ontology), ZFA (Zebrafish Model Organism Database), CL (ontology for cell types), NBO (Neuro Behavior Ontology), ChEBI (database and ontology for chemical entities of biological interest), MPATH (the mouse pathology ontology), and others.

Galeota E, Pelizzola M

Ontology-based annotations and semantic relations in large-scale (epi)genomics data

Brief Bioinform 2017;18(3):403-12

This work investigates the possible use of public repositories for assembling data sets including chromatin immunoprecipitation assays with massive parallel sequencing (ChIP-seq) data that were widely under exploited.

The hypothesis is that using semantic annotation of the metadata of public data sets with concepts from biomedical ontologies allows for a common description and interoperability of these data. The authors demonstrated that these annotations efficiently support the retrieval of samples for a given condition of interest over several large repositories. Additionally, a process of clustering based on semantic similarity metrics resulted in large groups coherent samples. The comparison of tools based on the UMLS (Metamap on version 2014AA) with tools that use topic-specific OBO ontologies (Concept mapper on BRENDA Tissue Ontology (BTO) and Disease Ontology (DO)) showed that the latter outperforms the former both in the annotation process and in the computation of semantic similarity measures. But given the dates of the resources, it should be noted that the 2014AA version of UMLS did not include the HPO ontology.

This approach is positively assessed by a case-study on a set of semantically homogeneous ChIP-seq samples targeting a specific transcription factor (Myc) and expanded with semantically similar epigenetic samples. The semantic information proved to be coherent with the ChIP-seq signal and the current knowledge about this transcription factor.

Khan Y, Saleem M, Mehdi M, Hogan A, Mehmood Q, Rebholz-Schuhmann D, Sahay R

SAFE: SPARQL Federation over RDF Data Cubes with Access Control

J Biomed Semantics 2017;8(1):5

In this paper, the authors propose SAFE, a query federation engine that enables policy-aware access to sensitive statistical data sets represented as RDF data cubes. SAFE is designed specifically to query statistical RDF data cubes in a distributed setting, where access control is coupled with source selection, and user profiles and their access rights. SAFE proposes a join-aware source selection method that avoids wasteful requests to irrelevant and unauthorized data sources. In order to preserve anonymity and enforce stricter access control, SAFE's indexing system does not hold any data instances—it stores only predicates and endpoints.

SAFE is motivated by the needs of three clinical organizations in the context of a European Union (EU) project which aims at enabling controlled federation over statistical clinical data – such as data from clinical trials – owned and hosted in situ by multiple clinical sites, represented in the form of data cubes. However, the methods proposed by SAFE can be used in other settings involving data cubes outside of the Health Care and Life Sciences domain (even for open data).

The resulting data summary has a significantly lower index generation time and size compared to existing engines, which allows for faster updates when sources change. Moreover, the authors show that SAFE enables granular graph-level access control over distributed clinical RDF data cubes and efficiently reduces the source selection and overall query execution time when compared with general-purpose SPARQL query federation engines in the targeted setting.

Notaro M, Schubach M, Robinson PN, Valentini G

Prediction of Human Phenotype Ontology terms by means of hierarchical ensemble methods

BMC Bioinformatics 2017;18(1):449

The prediction of human gene–abnormal phenotype associations is a fundamental

step toward the discovery of novel genes associated with human disorders, especially when no gene is known to be associated with a specific disease. In this context, the Human Phenotype Ontology (HPO) provides a standard categorization of the abnormalities (phenotypes) associated with human diseases.

In this paper, the authors tackle the problem of learning associations when the annotation source is HPO, i.e. a formal ontology, per se. HPO, as a formal ontology, is structured as a direct acyclic graph, where more general classes are found at the top levels of the hierarchy and the class specificity increases moving towards the lower levels of the hierarchy, *i.e.* from root to leaves. As a consequence, each class may have more than one parent and such an ontology is governed by the annotation propagation rule: if a gene is annotated with a given functional class, then it is annotated with all the "parent" classes, and with all its ancestors in a recursive way. On the contrary if a gene is not annotated to a class, it cannot be annotated to its offspring.

The authors present two hierarchical ensemble methods that they formally prove to provide biologically consistent predictions according to the hierarchical structure of HPO. The modular structure of the proposed methods, that consists in a "flat" learning first step and a hierarchical combination of the predictions in the second step, allows the predictions of virtually any flat learning method to be enhanced. The experimental results show that hierarchical ensemble methods are able to predict novel associations between genes and abnormal phenotypes with results that are competitive with state-of-the-art algorithms and with a significant reduction of the computational complexity. The implementation of the proposed methods is available as an R package from the CRAN repository. This result is important because it enhances prediction algorithms (for gene–abnormal phenotype associations) when association annotations are described with formal ontologies.

Petegrosso R, Park S, Hwang TH, Kuang R Transfer learning across ontologies for phenome-genome association prediction Bioinformatics 2017;33(4):529-36

In this paper, the authors tackle the problem of learning associations when the Knowledge Organization System (KOS) of annotations is HPO, i.e. a formal ontology, per se. Moreover, they take into account the nature of the KOS describing genome, the Gene Ontology (GO). They note that there are only few known associations available for training. For example, in HPO, more than half of the phenotypes are annotated at best with only one gene association, and this sparsity makes prediction impossible or much less reliable even if gene–gene interactions can be introduced as additional training information.

The authors introduce Dual Label Propagation (DLP) to impose consistent associations with the entire phenotype paths in predicting phenotype–gene associations in HPO. DLP is then used as the base model in a transfer learning framework (tlDLP) to incorporate functional annotations in GO. By simultaneously reconstructing GO, term–gene associations and HPO phenotype–gene associations for all the genes in a protein–protein interaction network, tlDLP benefits from the enriched training associations indirectly through relations with GO terms.

In their experiments to predict the associations between human genes and phenotypes in HPO based on human protein–protein interaction network, both DLP and tlDLP improved the prediction of gene associations with phenotype paths in HPO in cross-validation and the prediction of the most recent associations added after the snapshot of the training data. Moreover, the transfer learning through GO term–gene associations significantly improved association predictions for the phenotypes with no more specific known associations by a large margin.

Finally, the paper suggests that transfer learning can fulfill prediction with missing training information by the relation among GO and HPO. The results on predicting GO gene functions further support the conclusion that transfer learning across the two domains is beneficial to both learning tasks.