TEXT MINING, ONTOLOGIES AND DATABASES

Chairs: Alfonso Valencia and Guy Cochrane

F-1. Resource of Asian Primary Immunodeficiency Diseases (RAPID) update: an open web-based integrated molecular database on primary immunodeficiencies

Kumar Ramadoss S (1), Keerthikumar S (4), Raju R (4), Kandasamy K (4,5), Balakrishnan L (4), Dhevi Nagarajha Selvan L (4), Raja Sekhar N (4), Mohan S (4), Bhattacharjee M (4), Hijikata A (2), Imai K (6), Kanegane H (7), Miyawaki T (7), Nonoyama S (6), Pa*

Primary immunodeficiency diseases (PIDs) are a genetically heterogeneous group of disorders that affect distinct components of the innate and adaptive immune system. Despite rapid developments in the science of PID, early diagnosis and effective treatment are posed with greater challenges to all physicians. In this milieu, a development of freely accessible dynamic web-based integrated tool on PID has become a paramount need for different target groups. Thus, we have developed a web-based compendium of molecular alterations in PID, Resource of Asian Primary Immunodeficiency diseases (RAPID).

Materials and Methods

RAPID, is an object-oriented database, used three-tier architecture namely Zope, Python and MySQL. It hosts information on sequence variations of all reported PID genes along with expression profiling, mouse studies, interaction network and available DNA sequencing protocols. All mutation data have been linked with a graphical user interface (GUI) enabled tool, named Mutation@A Glance, to visualize and evaluate reported mutations. Using RAPID data, we have used Support Vector Machine (SVM) based parameter classification to predict candidate PID genes by scanning genes in the whole human genome

Results

At present, RAPID comprises total of 195 genes that are involved in PID, out of which 183 PID genes are reported with over 4300 unique mutation data referred from about 1500 public citations. For prediction of novel PID gene candidates, we trained SVM with 69 features for both positive and negative gene datasets and obtained 1,442 candidate PID genes from the human genome - these can be prioritized further and experimentally validated for its role in PID pathogenesis. Apart from these, RAPID has been updated with standardized DNA sequencing protocols for selected PID genes and PID expert page.

Discussion

Our next focus will be on the analysis of annotated PID data such as correlation of mutation types and effects, influence of mutation occurrence in its functional domains, identification of hotspot mutations, disease-causing mutations distribution and frequency in various ethnic groups and then interpretation of all these analyzed information to a greater knowledgebase of PID. Also, we are in the process of introducing RAPID gateway to closed network PID clinical consultation forum for registered users, using in-built tool developed and maintained by RIKEN Scientists Networking System (SciNeS)

URL

http://rapid.rcai.riken.jp/

Presenting Author

Suresh Kumar Ramadoss (suresh@rcai.riken.jp)
RCAI, RIKEN Yokohama Institute

Author Affiliations

1. Research Unit for Immunoinformatics, RIKEN Research Center for Allergy and Immunology, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. 2. Immunogenomics Research Group, RIKEN Research Center for Allergy and Immunology, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. 3. Department of Human Genome Research, Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan. 4. Institute of Bioinformatics, International Tech Park, Bangalore 560 066, India. 5. McKusick-Nathans Institute of Genetic Medicine and the Departments of Pathology and Oncology, Johns Hopkins University, Baltimore, Maryland 21205, USA. 6. Department of Medical Informatics & Department of Pediatrics, National Defense Medical College, Saitama 359-8513, Japan. 7. Department of Pediatrics, Graduate School of Medicine and Pharmaceutical Sciences, University of Toyama, Toyama 930-0194, Japan

Acknowledgements

This work was supported by a Grant-in-aid for Special Coordination Funds for Promoting Science and Technology, from the Ministry of Education, Culture, Sports, Science and Technology of Japan through an International collaboration with Institute of Bioinformatics (IOB), India.

F-2. Measuring the functional similarity of drugs

Götz S (1,), Behrens S (3), Tarazona S (1), Marba M (1), Dopazo J (1,2,4), Conesa A (1)*

As functional profiling/enrichment methods become the standard procedure for interpreting functional genomic studies it becomes more and more necessary to develop bioinformatic methodologies that establish connections within the functional profiling world. We presents a novel strategy to assessing the functional similarity of profiles based on the GO topology. We applied this strategy to an extensive dataset of drug profiles derived from genome-wide expression data. The dominant biological functions of drugs were compared to identify and detect similar functional characteristics among them.

Materials and Methods

Drug expression data from the Connectivity Map project was analysed by 2 GSEA methods. The resulting functional profile of 1500 drugs were compared against each other by a 3-step approach: Pairwise “node-based” semantic similarity for each pair of GO terms, the “best-match average” to combine the similarity of profiles and a random based normalization to account for the profile's size bias. The resulting matrix of drug similarities was evaluated by comparing it against the ATC drug classification scheme to show that similar compounds obtained high semantic similarity scores.

Results

The strategy, based on an edge-based metric to considers the topology of the GO hierarchy was used to measure the functional distances of drug based on a large set of transcriptomics data. A classification scheme for these compounds made it possible to contrast our results with an external criterion of similarity. The ATC system codifies drugs at different levels of specificity and we confirmed that the proposed method captures this specificity and observed that the mean similarity value increased when compared drugs belong to more specific ATC level.

Discussion

The proposed strategy to assess the semantic similarity of functional profiles combines a edge-based similarity metric with random null-model comparison to assess absolute distances of functional profiles. The analysis includes the evaluation and preprocessing of a large gene-expression drug dataset. To evaluate the functional similarity, drugs had been grouped by the hierarchical ATC classification scheme. The strategy proposed for measuring the semantic similarity of functional profiles has proved to be a valid methodology to study functional proximity between biological samples.

Presenting Author

Stefan Götz (sgoetz@cipf.es)
Centro de Investigacion Principe Felipe (CIPF)

Author Affiliations

(1) Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain (2) CIBER de Enfermedades Raras (CIBERER), Valencia, Spain (3) Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, Freising, Germany (4) Functional Genomics Node (National Institute for Bioinformatics, INB), Centro de Investigaciones Príncipe Felipe (CIPF), Valencia, Spain

Acknowledgements

Ministerio de Ciencia e Innovación, Gobierno de España

F-3. G-language bookmarklet: a gateway for semantic web, linked data, and web services

*Arakawa K(1, *), Kido N(1), Oshita K(1), Tomita M(1)*

In order to efficiently navigate and query through the huge masses of biological information, concepts of Linked Data and Semantic Web are gaining momentum as promising means for data integration. In light of the advent of these new data representation models, we here present a bookmarklet that provides an intuitive user interface for accessing the Linked Data and for querying the resources of Semantic Web, through a graphical ring-shaped menu on any webpage that the user is browsing. In this way, users can take advantage of the Semantic Web and Linked Data coherently with existing data.

Materials and Methods

G-language Bookmarklet is implemented as a bookmarklet to seamlessly work with regular web browsing, runs on any modern browsers, and can be invoked from any websites, without the need for installation of software or any specialized browser plugins. It is implemented with Javascript and HTML5 with Asynchronous Javascript and XML (AJAX) programing paradigm, and the source code is available with MIT License. Supported data sources are Google, PubMed, and Wikipedia for standard web, NCBI Entrez, EBI EB-eye, KEGG LinkDB for Linked Data, and Bio2RDF for Semantic Web.

Results

By selecting keywords of interest within any webpage and by opening the G-language Bookmarklet, an array of icons in the shape of a ring appears with animation on top of the webpage that the user is currently browsing. Here the users can select the database to search with the selected keyword, such as NCBI Entrez, Pubmed, KEGG, and Bio2RDF. Results of the queries are readily shown as another ring of icons representing the top hits of the query. Users can also access web services such as BLAST and G-language. The bookmarklet is freely available at: http://www.g-language.org/wiki/bookmarklet.

Discussion

Semantic Web and Linked Data are generally accepted as the promising means for data integration in biology, lead by initiatives such as Concept Web Alliance, Banff Manifesto, SADI, and DBCLS BioHackathon 2010. G-language Bookmarklet aims to provide a gateway to these new technologies for the end-users, by providing an intuitive interface with icons and animations that works on any web pages without the need for installation. We therefore provide a unique and consistent intuitive user interface for web search engines, linked data, semantic web, and web services.

URL

http://www.g-language.org/wiki/bookmarklet

Presenting Author

Kazuharu Arakawa (gaou@sfc.keio.ac.jp)
Institute for Advanced Biosciences, Keio University

Author Affiliations

1. Institute for Advanced Biosciences, Keio University

Acknowledgements

This research was supported by the Grant-in-Aid for Young Scientists No.222681029 from the Japan Society for the Promotion of Science (JSPS), and by funds from the Yamagata Prefectural Government and Tsuruoka City.

F-4. A graphical view of the world of protein-protein interaction databases

Klingström T (1,2), Plewczynski D (1,)*

The amount of information regarding protein-protein interactions (PPI) at a proteomic scale is constantly increasing. This is paralleled with an increase of databases making information available. Consequently there are diverse ways of delivering information about not only PPIs but also regarding the databases themselves. This creates a time consuming obstacle for many researchers working in the field.

Materials and Methods

The initial selection of investigated databases was made from the index of databases kept at Pathguide (http://www.pathguide.org/) and based on their popularity ranking. Most databases have at least one article describing them in the annual database issue of Nucleic Acids Research (NAR) journal. Articles published in this annual issue are obligated to contain a section comparing it to other similar databases. Relevant databases mentioned in those articles have also been added to complement the initial selection from Path-guide.

Results

Our survey provides a valuable tool for researchers to reduce the time necessary to gain a broad overview of PPI-databases. The graphical representation of data exchange also provides a novel way for researchers to quickly assess the origins of their data.

Discussion

The graphical representantion of data exchange between databases will be made available in cooperation with http://www.pathguide.org/interactions.php in a new Cytoscape web implementation.

URL

http://www.pathguide.org/interactions.php

Presenting Author

Dariusz Plewczynski (darman@icm.edu.pl)
Interdyscyplinary Centre for Mathematical and Computational Modelling, University of Warsaw

Author Affiliations

(1) Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Pawinskiego 5a Street, 02-106 Warsaw, Poland, E-mail: darman@icm.edu.pl (2) Biology Education Centre, Uppsala University, Norbyvägen 14, 752 36 Uppsala, Sweden.

Acknowledgements

We would like to thank all the researchers who have taken the time to personally answer our questions. We would also like to thank Hagen Blankenburg for taking the time to extensively explain DAS to us, and Jonas Klingström for his linguistic advice. Funding: This work was supported by Linnéska stipendiestiftelsen, EC OxyGreen (KBBE-2007-212281) 6FP projects as well as the Polish Ministry of Education and Science (N301 159735, PBZ-MNiI-2/1/2005 and others).

F-5. StrainInfo: your microbiological stepping stone

Verslyppe B (1,2), De Smet W (2,), Gillis W (2), De Vos P (2,3), De Baets B (4), Dawyndt P (1)*

Microbiology traditionally builds upon actual material which can be deposited in Biological Resource Centers. BRCs assign so-called strain numbers to identify the accessioned material and are responsible for long-term preservation and world-wide distribution of the material. As BRCs mutually exchange these cultured microorganisms, a plethora of equivalent strain numbers has come into use. This implies that to find all information on a given strain, all equivalent strain numbers should be used when searching. Moreover, the information is not suited for electronical processing.

Materials and Methods

StrainInfo is a virtual, global catalog integrating BRC catalogs. It is envisioned as a platform giving uniform access to microbiological specimen information. It integrates the specimen information with genomic information, taxonomic resources and literature databases. It builds upon the proposed Microbiological Common Language (MCL), aimed at standardizing the electronic exchange of meta-information about microorganisms. It provides mechanisms to the community to curate the vast amount of microbiological information readily available.

Results

StrainInfo is accessible through its web application which is based on the concepts of "passport pages" and "browsers". All equivalent numbers, together with additional information, are listed on so-called "strain passports". The corresponding "strain browser" enables direct access to the underlying BRC catalogs and features an overlay panel allowing easy browsing to equivalent catalog entries. Taxon, sequence and literature passports and browsers are also available. StrainInfo also offers web services for advanced users that want to automate queries or build powerful workflows.

Discussion

StrainInfo is an open, publicly available platform which is accessed online through a web application and web services. Its integrated view allows to handle microbiological information on a new scale and prepares microbiology for an era where new insights can be obtained based on existing information. StrainInfo provides true globally unique identifiers for microbiological material, forming the basis for electronical processing of microbiological information. Integration results are made available in electronically processable MCL files, allowing consumption by downstream applications.

URL

http://www.straininfo.net

Presenting Author

Wim De Smet (Wim.DeSmet@Ugent.be)
Ghent University

Author Affiliations

1. Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281, 9000 Ghent, Belgium. 2. Laboratory of Microbiology, Ghent University, K.L. Ledeganckstraat 35, 9000 Ghent, Belgium. 3. BCCM/LMG Bacteria Collection, Ghent University, K.L. Ledeganckstraat 35, 9000 Ghent, Belgium. 4. Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, 9000 Ghent, Belgium.

Acknowledgements

The development the StrainInfo platform is funded by the Belgian Science Policy (contract IUAP VI/06)

F-6. Semantic integration of isolation habitat and location in StrainInfo

Verslyppe B (1,2), De Smet W (2,), De Vos P (2,3), De Baets B (4), Dawyndt P (1)*

The adoption of Microbiological Common Language (MCL) XML synchronization quickly increased the volume of semantic data in StrainInfo. However, the effective data values of the different semantic fields still are raw textual entries and therefore are of varying detail, can have different forms or languages, and sometimes contain inconsistencies. By consequence, in order to generate a strain level consensus value for each field, a specialized semantic integration of this data needs to be developed. As a case study, the focus was put on the isolation habitat and location information fields.

Materials and Methods

To integrate geographical information, NER is performed to annotate all geographic names with features from the GeoNames ontology. This yields a multitude of annotations, each annotation matching a name with one or more geographical features. Multiple heuristics are used to cope with ambiguous and broad geographical names to select the most specific geographical feature. This is the integration result; multiple remaining annotations or features being too distant indicate inconsistent data. The habitat fields can be integrated using a similar algorithm (based on other ontologies).

Results

The consensus values are made available to end-users by listing them on the corresponding strain passports. Geographical names can be visualized on a map. Advanced search functionality is made available to allow users to perform true semantic search (e.g. find all strains isolated from dairy products in Europe). The integration results are also available from the MCL XML exports.

Discussion

In order to increase the information content of StrainInfo, it is necessary to add fine-grained semantic information. This information enters StrainInfo on the culture level (synchronization with BRC catalogs), but must be integrated on the strain level (i.e. the set of equivalent cultures) in order to be presented on strain passports. These results allow to use ontological knowledge when searching and therefore increase precision and recall compared to full text search. In addition, the ontologies enrich the data by providing or linking additional information (e.g. GPS coordinates).

URL

http://www.straininfo.net

Presenting Author

Wim De Smet (Wim.DeSmet@UGent.be)
Ghent University

Author Affiliations

1. Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281, 9000 Ghent, Belgium. 2. Laboratory of Microbiology, Ghent University, K.L. Ledeganckstraat 35, 9000 Ghent, Belgium. 3. BCCM/LMG Bacteria Collection, Ghent University, K.L. Ledeganckstraat 35, 9000 Ghent, Belgium. 4. Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, 9000 Ghent, Belgium.

Acknowledgements

The development the StrainInfo platform is funded by the Belgian Science Policy (contract IUAP VI/06)

F-7. Towards an ontology for farm animal reproduction traits

Hulsegge B (1,), Smits MA (1), te Pas MFW (1), Woelders H (1)*

Reproduction is an important characteristic of farm animals. Increasing amounts of information on reproduction traits in farm animals are becoming available in various formats and media, both electronically and otherwise. Accessibility and integration of this scattered information is an essential step not only in the understanding of the biology of reproduction traits but also for future optimizations. A prerequisite for this is to set-up an ontology to make relevant reproduction terms explicit and to define their relationships.

Materials and Methods

For the development of the Reproduction ontology we choose to start with the subdomain female fertility traits in cattle. Terms for this domain were collected from literature, from thesauruses, from existing ontologies and from experts in this field. The ontology is organized into a hierarchical structure. The relationships between terms is described as a class-subclass (“is_a”). relationship. Each term in the ontology has a name, a unique identifier (ID), a definition and may have synonyms. The ontology is constructed using Protégé and represented in the Web Ontology Language (OWL).

Results

A framework for the REProduction Ontology (REPO) was developed. A first version was provided which includes the major terms for the domain female fertility traits in cattle, together with their synonyms and, in most cases, text definitions. The top level terms of REPO include the following terms: reproductive behaviour, reproductive failure and reproductive performance. Currently REPO contains 94 terms.

Discussion

REPO can be used in databases, for literature analysis, for annotation (e.g. probes on microarrays) and linking high level traits to molecular information. The current version of REPO is not complete, its practical value will depend on future additions and corrections. An objective for future elaboration of REPO could be to include terms that relate to male fertility and to include other farm animals.

Presenting Author

Ina Hulsegge (ina.hulsegge@wur.nl)
Animal Breeding and Genomics Centre, Wageningen UR Livestock Research

Author Affiliations

Animal Breeding and Genomics Centre, Wageningen UR Livestock Research

Acknowledgements

This work was supported by the European Commission, within the Sixth Framework Programme, contract No. FOOD-CT-2006-016250 (SABRE) and from the Knowledge Base grant no KB-01-006-065 of the Dutch Ministry of Agriculture, Nature and Food security. The text represent the authors views and does not necessary represent a position of the Commission or Ministry who will not be liable for the use made of such information.

F-8. Human gene symbol validation at the HGNC

*Lush MJ *, Seal RL, Gordon SM, Wright MW, Bruford EB*

The aim of the HUGO Gene Nomenclature Committee (HGNC www.genenames.org) is to provide and promote unique and meaningful gene symbols and names for human genes. Standardisation of gene symbols is important as it allows researchers to refer to the same gene without ambiguity and facilitates electronic data retrieval. With the advent of microarraying and high throughput technologies, we have found there is an increasing demand to resolve large lists of gene symbols to their corresponding HGNC ID allowing gene data to be linked to many disparate sources.

Materials and Methods

The HGNC list search <http://www.genenames.org/list> is coded in perl 5.8, runs on a CentOS 5.1 server utilizing MySQL 5.0.41 and Apache version 2.0.

Results

Unfortunately this process is complicated by the use of symbol aliases (i.e. non HGNC-approved gene symbols) as a single alias may have been used to refer to multiple genes and may also clash with existing approved nomenclature. Additionally, approved nomenclature is not completely static; for example, gene records can be merged or split as new evidence becomes available.

Discussion

In response to demand we have developed a new "List Search" facility (www.genenames.org/list) which allows researchers to upload large lists of symbols for checking against the HGNC database. The results confirm which symbols are approved, and if the symbol is a known alias of a gene the corresponding approved symbol is displayed. Each approved symbol then leads the user to the gene entry in our database, which contains multiple links to other resources.

URL

http://www.genenames.org/list

Presenting Author

Michael J. Lush (mjlush@ebi.ac.uk)
HUGO Gene Nomenclature Committee

Author Affiliations

HUGO Gene Nomenclature Committee, EMBL Outstation - Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

Acknowledgements

The work of the HGNC is supported by National Human Genome Research Institute (NHGRI) grant P41 HG03345 and Wellcome Trust grant 081979/Z/07/Z.

F-9. miRSel: Automated extraction of associations between microRNAs and genes from the biomedical literature

Naeem H (1,), Küffner R (1), Csaba G (1), Zimmer R (1)*

microRNAs (miRNAs) have been discovered as important regulators of gene expression. To identify the target genes of miRNAs, several databases and prediction algorithms have been developed. Only few experimentally confirmed miRNA targets are available in databases. Many of the miRNAs targets stored in databases were derived from large scale experiments that are considered not very reliable.

Materials and Methods

The construction of a database of miRNA-gene co-occurrences via named entity recognition requires the compilation of miRNA and gene name dictionaries. The dictionaries for human, mouse and rat are compiled from several databases: HGNC, MGD, Entrez Gene, miRBase. In case of miRNAs we found that many miRNA names in the literature are not yet contained in databases. We detect miRNA names using a regular expression. Gene names are detected by string matching using the syngrep program. syngrep uses the Aho-Corasick algorithm for matching and context resolution techniques to resolve ambiguities.

Results

The miRNA-gene association database miRSel combines text mining results with existing databases and computational predictions. Text mining enables the reliable extraction of miRNA, gene and protein occurrences as well as their relationships from texts. Thereby, we increased the number of human, mouse and rat miRNA-gene associations by at least four-fold as compared to e.g. TarBase. miRSel is updated daily and can be queried using a web-based interface (http://services.bio.ifi.lmu.de/mirsel) via miRNA identifiers, gene and protein names, PubMed queries as well as gene ontology (GO) terms.

Discussion

miRNA related research depends on knowledge of miRNA target genes. In contrast to manual curation, we proposed a simple, automated approach for biological name identification that collects many potential targets for miRNAs not contained in current databases. Text mining of miRNA, gene or protein names results in good recall and precision for miRNA-gene associations detected in single sentences. We thereby extracted many pairs from human (2875 pairs), mouse (1466 pairs) and rat (595 pairs) abstracts. To keep the miRSel database up-to-date, newly available PubMed abstracts are included daily.

URL

http://services.bio.ifi.lmu.de/mirsel/

Presenting Author

Haroon Naeem (haroon.naeem@campus.lmu.de)
Institut für Informatik, Ludwig-Maximilians-Universität München, Amalienstr. 17 80333 München, Germany

Author Affiliations

Lehr- und Forschungseinheit Bioinformatik Institut für Informatik, Ludwig-Maximilians-Universität München, Germany

Acknowledgements

This work has been supported in part by the German Federal Ministry of Education and Research (via DLR, FKZ 01GS0801). Haroon Naeem is supported by the Deutscher Akademischer Austausch Dienst (Referat 442, Code A/07/96865).

F-10. Metarel: a relation metagraph for inferences in biomedical ontologies

Blondé W (1,), Antezana E (2), Venkatesan A (2), De Baets B (1), Kuiper M (2), Mironov V (2)*

Ontologies start to become ubiquitous in the life sciences for managing and integrating knowledge and data. They provide a reference framework where scientific terms are defined and interconnected with sound methods. Biologists and bioinformaticians need this framework in order to query the available knowledge and to fill the gaps where necessary. Gene Ontology Annotations (GOA), containing millions of annotations about proteins, makes use of this framework. Scalable inferencing tools are necessary for facilitating queries on the integrated Knowledge Base (KB).

Materials and Methods

The Semantic Web provides all the tools that are required for integration, querying and reasoning. We translated all the OBO ontologies to RDF and stored them in the BioGateway KB, powered by Virtuoso. We carefully separated all the information about the relations between terms, because these contain the rules about knowledge that is not explicitly present in the ontologies. We condensed all these rules into the relation metagraph of Metarel, which was created for task. It contains rules for reflexivity, transitivity, super-relation types, composite relations and more.

Results

By running SPARUL update queries over all the ontologies and the metagraph, we could infer hundreds of millions of new knowledge statements. All these statements together contain all the implied knowledge that follows from the knowledge that was stated explicitly. We call this a total relational closure. Any queries on the total closures (TC) return all the correct answers, which was not the case for queries on the original ontologies. We could create many new queries in BioGateway that are very useful from the perspective of biological scientists.

Discussion

This work proves that a sound ontological framework is necessary and desirable for Knowledge Management (KM), because it enables many intuitive queries within integrated systems. This sound framework should contain rules that allow to infer many knowledge statements from a small core that is well maintained. Metarel in RDF proves to be an excellent knowledge representation language for executing such inferences. By using the Semantic Web, the approach remains compatible with Semantic Web reasoning that is more advanced, though less scalable, like OWL reasoning.

URL

http://www.semantic-systems-biology.org/metarel

Presenting Author

Ward Blondé (ward.blonde@ugent.be)
Ghent University

Author Affiliations

1) Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Belgium 2) Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway

Acknowledgements

Blondé Engineering

F-11. Predicting disease causing genes using the information content surrounding the disease and gene in literature

van Haagen H (1,), Aten E (1), 't Hoen P-B (1), Roos M (1), Messemaker T (1), Mons B (3), van Ommen G-J (1), Schuemie M (2)*

Finding relevant information for putative disease causing genes is difficult since information almost never is explicitly available for the gene and the disease. Some bioinformatic approaches circumvent this problem with the use of alternative information already known for the disease such as previously found genes. However for the majority of the diseases recorded in OMIM this information in not available.

Materials and Methods

In this paper we describe a way to extract information for diseases to find putitave genes even if there is no explicit information available in text nor in any other database. The information generated is the surrounded literature content of the disease stored in a profile. We generated a test set where the disease was related with no gene before the landmark paper.

Results

Most prioritizers based on training genes fail, but our method still was able to prioritize the disease gene on average within the top 13 genes out of 200 genes in a linkage interval. In addition we evaluated our method on the test set used to evaluate Endeavour and showed that we perform twice as good in prioritizing genes for polygenic disease. Not only is the gene predicted, the information content shared between the gene and the disease explains the relationship.

Discussion

We have proven that our approach of indirect links in text works. Our text-mining system is implemented in the biosemantics tool Anni. Anni can be found on http://www.biosemantics.org/anni

Presenting Author

Herman H.H.B.M. van Haagen (hvanhaagen@lumc.nl)
LUMC

Author Affiliations

(1) LUMC, Leiden, Netherlands (2) Erasmus MC, Rotterdam. Netherlands (3) NBIC, Netherlands

F-12. Revealing heterogeneities and inconsistencies in protein functional annotations

Sanavia T (1,) Facchinetti A (1), Di Camillo B (1), Toffolo G (1), Lavezzo E (2), Toppo S (3), Fontana P (4)*

Gene Ontology (GO) is the most widely used annotation database to transfer biological knowledge to a new raw sequence. However, it is known that the existing annotations are incomplete and susceptible to several sources of errors (King, 2003); in this context, one of the most important challenges is to interpret the accuracy and consistency of the available annotations, in order to identify or correctly predict the functions of protein sequences.

Materials and Methods

We developed a new method to analyze heterogeneities of GO annotations. For groups of proteins sharing a high sequence similarity, the method performs two agglomerative hierarchical clustering based on semantic similarity between GO terms and protein annotations, respectively. Semantic similarity is computed using Lin's measure (Lin, 1998) and Best Match Average (Couto, 2007). The algorithm gives as output a color-coded matrix where each cell (i,j) is colored if the protein j is annotated with the GO term i, with different color intensities representing the information content of GO terms.

Results

Proteins with the same sequence length were first clustered in order to define groups of proteins at a high sequence similarity level (more than 95%) and their GO annotations were retrieved from GOA database. The method was able to reveal inconsistencies highlighting the presence of isolated cells representing GO terms and/or proteins outliers and indicating the specificity of the functional annotations through the information content. Several groups of proteins with high sequence similarity evidenced a lack of homogeneity among GOA annotations.

Discussion

The proposed method is able to efficiently organize information on GO annotations of groups of proteins through semantic similarity measures. Results in the GOA database evidenced unexpected and not easily traceable heterogeneities on protein annotations. The algorithm can be applied as a pre-processing step in order to identify incoherencies in biological information, thus avoiding propagations of potential annotation errors. The resulting output can be used as a useful guidance to predict functions of new protein sequences or to re-annotate the existing annotations of known proteins.

Presenting Author

Tiziana Sanavia (sanaviat@dei.unipd.it)
Department of Information Engineering, University of Padova

Author Affiliations

(1) Department of Information Engineering, University of Padova, via G. Gradenigo, 6/B, 35131 Padova, Italy. (2) Department of Histology, Microbiology and Medical Biotechnologies, University of Padova, via Gabelli 63, 35131 Padova, Italy. (3) Department of Biological Chemistry, University of Padova, viale Colombo, 3, 35131 Padova, Italy. (4) FEM-IASMA Research Center, Via E. Mach, 1, 38010 San Michele all'Adige(Trento), Italy.

Acknowledgements

This work was supported by CARIPARO 2008/2010 "Systems biology approaches to infer gene regulation from gene and protein time series data".

F-13. A formal framework for evaluating the significance of peptide-spectrum matches

Vandenbogaert M (1,)*

Tandem mass spectrometry (LC-MS/MS) experiments, as a high-throughput technology for protein identification, generate millions of observed spectra. To correctly match every spectrum to a theoretical spectrum, traditionally used target-decoy strategies reveal to be mostly imprecise. Exploiting all the peaks in the spectra, and using a more formal framework for assessing the relevance of assignments of peptides in an MS/MS proteomic dataset, allows to bypass the need of those expensive decoy searches, and in addition speeds up the decision process for matching an observed spectrum to a peptide.

Materials and Methods

A given MS/MS spectrum both gives the sequence of amino-acids and the fragment masses that constitute the spectrum-prefixes. We derive from a spectrum all possible explanations, (amino-acid sequences with or without post-translational modifications), that best explain the correspondence to putative peptides in a proteome database. For a given set of patterns one searches for multiple occurrences of the sequences from this set in a proteomic text. We combinatorially construct the set of sequences explaining a spectrum, and use the spectrum characteristics to assess a peptide's match propensity.

Results

A counting function is constructed that counts, given the length of a random text, the number of occurrences of a motif in this text. To consider the match score between a spectrum and a peptide, as well as the mass of the peptide, we consider the dot product as the match score between an observed spectrum and a theoretically derived spectrum. For the set of peptides that have an integer number of peaks in common with a spectrum, the set of possible masses are enumerable, which we integrate in a counting function with two variables, that is transformed into a probability generating function.

Discussion

To assess the significance, a probability assumption is needed to express what is expected and some criteria to establish that the observation deviates significantly from the expected. Here we consider the probability that a random peptide matches a spectrum with a score above a threshold. We consider the Delta score as the match quality for peptides, i.e. the difference of score of the best PSM and the score of the second best PSM, for a specific spectrum. Considering specific neutral losses are a natural extension to this framework, and lead to the mining of MSn/MSA fragmentation spectra.

Presenting Author

Mathias L.C. Vandenbogaert (mathias.vandenbogaert@pasteur.fr)
Institut Pasteur, Paris, France.

Author Affiliations

(1) Institut Pasteur, Paris, France.

Acknowledgements

Benno Schwikowski's Systems Biology lab, Institut Pasteur, Paris, France.

F-14. Improving the execution of bioinformatical workflows

Subramanian S (*), Sztromwasser P, Puntervoll P

Most scientific workflow engines use centralized coordination as the choice of approach for executing bioinformatical workflows, requiring the coordinator to send and receive all the input and output data of component services. Such indirect data communication between the component services increases the data-traffic of the coordinator and weakens the performance of the workflow. To optimize this, we propose an approach where data-flow is dynamically delegated from the coordinator to the component services, with direct transportation of data between the component services.

Materials and Methods

The vanilla (e.g., BPEL) and the proposed data-flow delegation methods are considered for the performance analysis. The technologies used in the implementation are: (i) Python v2.6.4 language, for the development of a workflow coordinator and the three component web services, (ii) ZSI v2.1_a1 library, for the development an deployment of the web services, and (iii) the suds v 0.3.8 library, for sending and receiving messages between the web services.

Results

We have implemented a simple gene annotation workflow to compare the performance of the proposed and existing approaches. The gene annotation workflow integrates and coordinates different bioinformatical resources like UniProt/SwissProt database and BLAST tool, located in different places. To generalize the proposed approach, the execution semantics are formally defined, and an XML schema is developed to support the data-flow delegation.

Discussion

By comparing the obtained results, the proposed approach optimizes 32% of workflow runtime, 7 times of bandwidth usage, 36 times of memory usage, and 8.7 times of CPU usage. This shows that the proposed data-flow delegating approach improves the performance of the bioinformatical workflow and its coordinator considerably, compared to the non-delegating classical approach. We believe that this poster session would be a great opportunity to identify different data-centric workflows in the domain of bioinformatics for testing our approach further.

Presenting Author

Sattanathan Subramanian (sat@uni.no)
Uni BCCS

Author Affiliations

Uni BCCS, Bergen, Norway

Acknowledgements

This work was carried out as part of the eSysbio project (No: 178885) funded by the Research Council of Norway.

F-15. BridgeDb: standardized access to gene, protein and metabolite identifier mapping services

Van Iersel MP (1,), Pico AR (2), Kelder T (1), Gao J (3), Ho I (2), Hanspers K (2), Conklin BR (2), Evelo CT (1)*

Many interesting problems in bioinformatics require integration of data from various sources. For example when combining microarray data with a pathway database, or merging co-citation networks with protein-protein interaction networks. Invariably this leads to an identifier mapping problem, where different datasets are annotated with identifiers that originate from different databases. Several existing services (e.g. BioMart) can perform identifier mapping. Applications can be made to flexibly support multiple mapping services or mapping services could be combined to get broader coverage.

Materials and Methods

We implemented a development library that forms an interface that connects mapping services and bioinformatics tools. This interface is in the form of both a Java and REST API.

Results

BridgeDb has already been integrated into several popular bioinformatics applications, such as Cytoscape, WikiPathways, PathVisio, Vanted and Taverna. To encourage tool developers to start using BridgeDb, we've created code examples, online documentation, and a mailinglist to ask questions. The BridgeDb library is open source and available at The BridgeDb library is available at http://www.bridgedb.org.

Discussion

BridgeDb helps to improve the user-friendliness of existing bioinformatics applications, by making it easy to add important new capabilities. BridgeDb isn't just yet another mapping service: it tries to build further on existing work, and integrate existing partial solutions. The framework is intended for customization and adaptation to any identifier mapping service. BridgeDb was open source from the very beginning of the project. We believe the philosophy of open source is closely aligned to academic values, of building on top of the work of giants.

URL

http://www.bridgedb.org

Presenting Author

Martijn P. van Iersel (martijn.vaniersel@bigcat.unimaas.nl)
BiGCaT Bioinformatics, Maastricht University

Author Affiliations

(1) Department of Bioinformatics - BiGCaT, Maastricht University, Maastricht, the Netherlands (2) Gladstone Institute of Cardiovascular Disease, San Francisco, CA 94158, USA (3) Department of Computer Science, University of Missouri, Columbia, MO 65201, USA

Acknowledgements

This work was supported by the Google Summer of Code program, transnational University Limburg (tUL), the BioRange program of the Netherlands Bioinformatics Consortium (NBIC), the Netherlands Consortium for Systems Biology (NCSB), the National Institutes of Health [GM080223,HG003053], the European Nutrigenomics Organization (NuGO) and theDutch Scientific Organisation (NWO)

F-16. HGVbaseG2P: an advanced database for the integration and interrogation of genetic association datasets

Free R, Hastings R, Thorisson GA, Gollapudi VLS, Beck T (*), Lancaster O, Brookes AJ

Genetic association study data are rarely published in any comprehensive fashion in journals or databases, and in the case of negative findings such data are often not reported at all. Consequently, it is difficult to compare and contrast the results of different studies, and it is completely impossible to examine a full and unbiased picture of all the data that exist. To address this deficit, the Human Genome Variation Genotype to Phenotype database (HGVbaseG2P: http://www.hgvbaseg2p.org) has been constructed.

Materials and Methods

HGVbaseG2P is implemented as a traditional relational database using the MySQL platform. Data import/export tools and database middleware components are all implemented in the Perl programming language, leveraging a number of open-source software packages from the CPAN. The web application is built using the Perl-based Catalyst MVC-framework.

Results

HGVbaseG2P provides a simple interface and advanced search capabilities. Researchers are able to identify studies of interest based on chromosomal regions/genes and markers. Studies of interest can be selected for comparison at both genome-wide and region-specific levels. Studies can also be searched based on phenotype, via the Medical Subject Headings (MeSH) and Human Phenotype Ontology (HPO) hierarchies. We constantly and actively search for new data available in publications and multiple online public resources. The database now hosts >18,000,000 p-values and 352 studies.

Discussion

Researchers are encouraged to directly submit studies and association data into HGVbaseG2P. A data access control system enables data submitters to decide what studies and which data elements are to be shared and with whom (e.g. collaborators, own research group). By summer 2010 the complete HGVbaseG2P code base will be available for anyone to install on their own servers so they can run their own genetic association database.

URL

http://www.hgvbaseg2p.org

Presenting Author

Tim Beck (tb143@le.ac.uk)
University of Leicester

Author Affiliations

Department of Genetics, University of Leicester, UK

Acknowledgements

The European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 200754 – the GEN2PHEN project.

F-17. Extracting phytogeography information from species distribution data

Bakis Y (1,), Sezerman OU (2), Babaç MT(1)*

Today, it is possible to analyze large datasets as large as species distribution of Turkish flora. Availability of species distributions in digital format had motivated us to analyze this data. Thus, we have calculated pair wise similarities among the 30 grid squares by presence of species in the grid pairs. An algorithm has been de-signed for this purpose because of known limitations in current multivariate analyzing techniques. By using the similarity matrix, we constructed a phylogeny for the 30 grids squares.

Materials and Methods

The distribution of each taxa had been indicated in Davis' "Flora of Turkey and East Aegean Islands" (Davis, et al., 1965). Recently, Babaç has digitized the species distributions into a database of Turkish Plants Data Service (TUBIVES) (Babaç, 2003). All occu-rences were coded as "1"s and non-occurences as "0"s as in figure 2. By using the species distribution records from the Turkish Plant Data Service (M. Tekin Babaç, 2010) of all taxa we tried to cluster the Davis' grids. Some taxa have been represented by 1 or no dis-tribution record or show a global distribution.

Results

Cladogram constructed by the cluster analysis has shown that spe-cies distributions over the grid squares were highly correlated with the phytogeographic regions. There were 3 main clusters in the cladogram topology serving as Euro-Siberian, Irano-Turanian and Mediterranean regions. The grids squares were all correct within these clusters with some exceptions. A2 is assigned to be a member of Mediterranean region not Euro-Siberian which is an important result since more than half of the sea shores of the region is included within the Mediterranean Region.

Discussion

Species distribution data gas given accurate results with the phytogeographical regions. The two very important distinction were pointing already defined problems in phytogeography of Turkey.

Presenting Author

Yasin Bakis (bakis_y@ibu.edu.tr)
Abant Izzet Baysal University

Author Affiliations

1 Department of Biology, Abant Izzet Baysal University, Bolu, 14280 Turkey 2 Biological Sciences and Bioengineering, Sabanci University, Tuzla, Istanbul, 34956 Turkey

F-18. D2K - data to knowledge: data integration for biological reasoning

Wienecke-Baldacchino AK (*), Heinäniemi M , Carlberg C

Hypothesis driven data mining faces problems, like a high amount of public databases to evaluate and quality control, in order to extract information in a computationally processable format and its automating via interfaces. Additionally, for new types of high throughput data, e.g. ChIP-seq, no public databases are established yet, which would provide a certain standardized data format in particular concerning annotation or analysis aspects. Therefore, we decided to set up a database system integrating published, in silico generated and experimentally achieved data.

Materials and Methods

Our tool is based on a relational database system. Processing or insert functions are implemented in Python. The selection of public data sources and ontology was done by hand. Here aspects like data quality, data structure, formats, automation possibilities and provided interfaces were considered. Data consistency especially for published data is provided by inter-database-confirmation, only cross-references confirmed over different public databases are inserted. We also integrated in silico generated genome-wide transcription factor binding sites and respective experimental data.

Results

The present version of the database consists of 98 tables and 209 relations. Several public databases, e.g. dbSNP, miRBase, GAD, HGNC, Entrez and Ensembl, are integrated. Further data sources are under revision. For 1058 position weight matrices genome-wide data are inserted and the integration of experimental data is ongoing. Selected current applications are: Decision making if to evaluate published, non-coding, significantly associated SNPs; ID mappings for automated text mining; selection of candidate SNPs for association studies and selection of candidate miRNAs by ChIP-seq data matching.

Discussion

The main intention of the presented system is to take advantage of the huge amount of already available data and to transfer data to knowledge namely a reasonable biological hypothesis, so that the wet-lab serves only for validation. Main challenges are to curate the system and keep it up-to-date, especially in context of genome versions, coordinates and IDs.

Presenting Author

Anke K. Wienecke-Baldacchino (Anke.Wienecke@uni.lu)
Life Sciences Research Unit, University of Luxembourg, Luxembourg

Author Affiliations

Life Sciences Research Unit, University of Luxembourg, Luxembourg

F-19. BRENDA text-mining: new developments for obtaining enzyme-related disease information from scientific literature

Soehngen C (1,), Scheer M (1), Schomburg I (2), Chang A (2), Grote A (1), Schomburg D (1)*

BRENDA (BRaunschweig ENzyme DAtabase) provides manually annotated data and additionally information extracted from the scientific literature by a text mining approach. The part of the text mining component, which focuses on the search for enzyme and disease related references, has been revised and augmented with a subsequent classification of mining results to get an insight on the stressed issues of the reference. The F1 score for classification, evaluated in a 5fold cross validation, ranges between 0.802+/-0.032 - 0.738+/-0.033 depending on the text class and preprocessing procedures.

Materials and Methods

In a text mining approach PubMed references with enzyme hits in title or abstract are searched for co-occurring disease terms. The dictionaries are BRENDA enzyme names or synonyms and MeSH disease terms. The classification is processed by SVMlight, a Support Vector Machine (SVM) implementation. For training purposes a corpus has been manually annotated, containing 5,033 sentences derived from PubMed abstract and titles. The aim is the classification of mining results which may belong to one or more of the four pre-defined categories or none of them.

Results

In the summer release 2010.2 910,897 disease entries (522,720 distinct PubMed references) are added to BRENDA (co-occurrence related F1 score 0.89). The further SVM classification was evaluated in a 5 fold cross validation: Categories (F1 scores +/- std. dev.) The enzyme is a therapeutic agent or target (0.802+/-0.032) The intention of research about the connection of the enzyme and the disease is stated (0.744+/-0.020) The enzyme is causal for the disease or part of the disease process (0.743+/-0.009) The characteristic values of the enzyme are used for diagnostic purpose (0.738+/-0.033)

Discussion

BRENDA provides a large number of disease related enzyme entries which are half yearly updated in the course of every new release of the database. However, the categorisation through SVM classification will be a supplemental information to disease entries in future BRENDA releases. A high value on precision will be set on the choice of parameter settings for the SVM. Thus, the classification will be a reliable indicator on the stressed issues of every reference found by disease mining.

URL

http://www.brenda-enzymes.org

Presenting Author

Carola Söhngen (c.soehngen@tu-bs.de)
Technische Universitaet Braunschweig

Author Affiliations

(1) Technische Universitaet Braunschweig (2) Enzymeta GmbH, Erftstadt

Acknowledgements

European Union Projects: FELICS (Free European Life-Science Information and Computational Services) SLING (Serving Life-science Information for the Next Generation)

F-20. Beegle: a new search engine for discovering novel genes

Brohée S (1,), Gonçalves JP (1,2,3), Nitsch D (1), Moreau Y (1)*

Observing the high rate of publication and the complexity of many fields in genetics, it becomes increasingly difficult to keep this knowledge up-to-date considering a given biological process, to have a quick overview of the underlying genetic mechanisms or to initiate new studies to discover novel genes implied in this process. To solve this, we designed a new user-friendly strategy allowing the easy retrieval of known genes and the prediction of novel genes related to any query or search term.

Materials and Methods

Our approach combines two methods. Firstly, starting from any text query, our tool retrieves the set of linked PUBMED abstracts and builds a vector whose components reflect the weight of given terms found in these abstracts. This query vector is then compared to a set of reference vectors, each one corresponding to a given gene, built using the GeneRIF resource (which lists references for each gene). In the second step, wes discover novel genes by applying a network search strategy with the genes whose word weight vector was the most similar to the query vector.

Results

The first results delivered by our approach are very encouraging. We are currently benchmarking and statistically assessing our approach for its ability to detect relevant known and novel genes.

Discussion

Starting from a simple free text query, our user-friendly method is able to return relevant genes linked to this query. We are confident that our method will be of interest: on the one hand, to the wet lab scientist willing an overview of the underlying genetic processes or to discover new genes concerning a process of interest; on the other hand, to the bioinformatician in need of training data sets of genes to run an application (e.g. prioritization tool).

URL

http://homes.esat.kuleuven.be/~bioiuser/beegle/

Presenting Author

Sylvain Brohée (sbrohee@esat.kuleuven.be)
Bioinformatics group. ESAT-SCD. Katholieke Universiteit Leuven.

Author Affiliations

(1) Bioinformatics group, ESAT-SCD, Katholieke Universiteit Leuven. (2) Knowledge Discovery and Bioinformatics group, INESC-ID. (3) Instituto Superior Técnico, Technical University of Lisbon. (+) Equal contribution

Acknowledgements

Sylvain Brohée is the recipient of a post-doctoral grant supported by the EU FP7 CHeartED project (Grant HEALTH-F2-2008-223040). Bioinformatics group is supported by Research Council KUL: ProMeta, GOA Ambiorics, GOA MaNet, CoE EF/05/007 SymBioSys, START 1, several PhD/postdoc & fellow grants. Joana P. Gonçalves is the recipient of a doctoral grant supported by FCT (SFRH/BD/36586/2007). INESC-ID multiannual funding is supported by FCT through the PIDDAC Program funds.

F-21. An overview of the pathogen-host interactions database, PHI-base

Janowska-Sejda E (1,), Defoin-Platel M (1), Hammond-Kosack K (2), Urban M (2), Tsoka S (3), Saqi M (1)*

Recent advances in sequencing technologies resulted in a wealth of full-genome information becoming available for many pathogenic species. Over the years, the number of publications describing a functional characterisation of pathogenicity genes, virulence and effector genes has grown significantly. PHI-base is a database, manually curated by scientists, which systematically catalogues phenotypic information extracted from peer-reviewed publications for genes involved in pathogen–host interactions.

Materials and Methods

The sequences in PHI-base version 3.2 were clustered on the basis of sequence similarity and represented graphically to give a global overview of the database content in terms of phenotype and experimental hosts.

Results

General analysis of PHI-base content revealed the presence of 807 non-redundant genes for which nucleotide and protein sequences are available in the database. These genes belong to 75 different pathogenic species. The total number of interactions within the PHI-base equals 1314. Overall 111 clusters with two or more genes were identified. The clustering identified 52 plant specific gene clusters and 25 animal specific gene clusters. Pathogenicity genes in 33 mixed clusters identified genes important for host invasion both in animal and plant pathogens.

Discussion

Our comparative gene set analysis of genes involved in pathogen-host interactions across diverse species identified orthologous genes conserved in plants and/or animal pathogens. We next plan to use HMM3 profiling to confirm these findings and identify the molecular themes required for host specificity.

Presenting Author

Elzbieta I Janowska-Sejda (elzbieta.janowska-sejda@bbsrc.ac.uk)
Rothamsted Research

Author Affiliations

1 Department of Biomathematics and Bioinformatics, Rothamsted Research 2 Department of Plant Pathology and Microbiology, Rothamsted Research 3 Department of Computer Science, King's College London

Acknowledgements

This work is fully funded by BBSRC

F-22. Finding disease related genes by GeneRank algorithm using co-occurrence based network structures

Lee H-M(1), Shin M-Y(1,), Hong M-P(2)*

To find significant genes related to specific diseases, microarray gene expression profiles have been widely used for gene ranking. Lately, however, the GeneRank algorithm was proposed which employs gene annotation data from Gene Ontology to construct gene networks and use them for the gene ranking along with gene expression profiles. To make use of relevant information from bio-literature, this work presents a method to decide the ranks of genes by using text-mining based method to construct gene networks for the GeneRank algorithm.

Materials and Methods

We assume that two or more genes are likely to be related, if they co-occur in a sentence. In this approach the correct recognition of gene entities is of great importance. For this purpose, we employ a dictionary-based approach and a machine learning approach. Once co-occurring genes are found in a sentence, they are taken to be related and a new link is added to the gene network. For the experiment, we used text data related to prostate cancer from the PubMed. The prostate cancer gene expression profiles were obtained from the Gene Expression Omnibus.

Results

Based on the gene networks and gene expression profiles, we compute the gene ranking scores with GeneRank algorithm. Once the ranking scores are computed, the genes are sorted in a decreasing order. For biological validation, we obtained gene lists already known to be related to prostate cancer. Using these gene lists, we find out how many genes in the list are identified by our method. The experiments show that our method outperforms the previous works based on the Gene Ontology. It also shows that the machine learning approach produces better results than the dictionary based approach.

Discussion

The gene networks based on text mining produced a better result than the GO-based gene networks in computing gene ranking scores. Our experiment shows that it is of great worth to use gene relations found in bio-literature for disease related gene identification. As a future work, we are planning to employ a syntactic parser customized for a medical domain to extract relations between genes in a more sophisticated manner, so that a directed gene network can be built.

Presenting Author

Miyoung Shin (shinmy@knu.ac.kr)
Kyungpook National University

Author Affiliations

(1) Kyungpook National University (2) Sungkyunkwan University

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation grant (No. R01-2008-000-11089-0)

F-23. Identifying variability of human splicing forms in their interactions by using literature mining

Kafkas S (1,2,), Varoglu E (1), Rebholz-Schuhmann D (2), Taneri B (3)*

Alternative splicing (AS) is a prominent mechanism contributing to transcriptome and proteome diversity by generating multiple protein isoforms from a single gene. Isoforms exhibiting differences in their structures and functions can be expected to exhibit variations in their interactions. Although there are several PPI databases providing structured information on PPIs, they cover only a portion of the interactome and interaction information on isoforms is sparse. Literature has rich information content. Our work provides a novel extension both to AS and PPIs including variability analysis.

Materials and Methods

We analyze the transcript data from HumanSDB3 for 16,826 different genes and collect relevant abstracts from PubMed. Then we employ an SVM trained on BioCreative-II IAS corpus with a novel and high performing feature set for selecting interaction abstracts. Another SVM trained on AIMed corpus by using syntactic features is utilized to extract interacting proteins from the selected abstracts. We evaluate our DB against PINA. We utilize it to select the genes having multiple interacting isoforms and identify their unique and shared interactions to determine the variation in their interactions.

Results

We process around 4 million abstracts and find a total of 31,819 distinct interactions belonging to 5,615 isoform from HumanSDB3. 69.04% of proteins and 9.00% of interactions overlap with PINA. Variability analysis shows that all isoforms from the majority of genes having multiple interacting isoforms (81.82%) have unique interaction partners only. Isoforms from 17.04% of such genes have both; shared and unique interaction partners while isoforms from the remaining 1.14% genes exhibit shared interactions only.

Discussion

Evaluation of our PPI DB against PINA yields low interaction overlap due to different interaction extraction methods used to build PPI databases and low number of isoform interactions in such databases. Interaction variability analysis expose that the majority of protein isoforms exhibit differences in their interaction partners. Our analysis demonstrates that alternative splicing significantly contributes to the variability in the isoform interactions. Our findings will be publicly available through a web interface for further usage.

Presenting Author

Senay Kafkas (kafkas@ebi.ac.uk)
European Bioinformatics Institute

Author Affiliations

1-Department of Computer Engineering, Eastern Mediterranean University, Famagusta, North Cyprus 2-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK 3-Faculty of Arts and Sciences, Eastern Mediterranean University, Famagusta, North Cyprus

F-24. The meaning behind the choice of words: a case in cancer metastasis.

*Divoli A (1,2, *), Mendonca EA (3,5), Rzhetsky A (1,2,3,4)*

We interviewed 28 cancer metastasis experts to investigate how uniform their views are in a number of topics in the research field. While studying the audio files and transcripts we noticed several linguistic phenomena: hesitation, (un)certainty, hedging, use of anecdotes, metaphors and so on. Based on our initial observations we decided to study the texts more systematically to detect correlations between specific language use and topics of discussion and/or types of experts.

Materials and Methods

We split our data into 14 topics and grouped our experts in categories based on their doctoral training (PhD and/or MD), their gender, the number of years since their first doctorate, and if the interviews were conducted in person or over the phone. We then manually analyzed our data to pick up language idioms such as metaphors and proceeded with several automatic methods. We exploited word and N-gram frequencies, applied commercial linguistic inquiry tools, and performed extensive studies employing crowd-sourcing.

Results

Our results exhibited several interesting points, such as that some discussion topics like “tropism” provoked more strong opinions and inspired the use of metaphors, and certain expert groups overall used a more assured tone of voice. The linguistic analysis revealed high use of language associated with cognitive processes (certainty and insight in particular). When this language is known to be under-represented in scientific text, in our data it was over-represented, often, more than one standard deviation over any control text, including oral speech and even novels.

Discussion

Unlike the view one might get by looking at textbooks and the scientific literature, our results show that, in reality, the knowledge is not as crisp. There is speculation, uncertainty and difference in opinion. Beyond the value of these data for the generation of hypothesis, researchers that analyze biomedical text and are concerned with knowledge representation should examine the effect that this uncertainty has on their work.

Presenting Author

Anna Divoli (divoli@uchicago.edu)
The University of Chicago

Author Affiliations

The University of Chicago, (1) Department of Medicine, (2) Institute for Genomics and Systems Biology, (3) Computation Institute (4) Department of Human Genetics (5) Department of Pediatrics

Acknowledgements

Funded by the Ludwig Institute of Cancer Research

F-25. Linked open experimentation

Roos M (1,2,), Van Haagen HHHBM (1), Singh B (3), Marshall MS (1,2), Mons BM (1)*

The amount of data and knowledge and the number of methods for exploring biological data towards a new hypothesis are large and increasing each day. We investigate the application of a combination of information technologies that can help us explore knowledge beyond our own locally produced data.

Materials and Methods

Our approach combines (i) Linked Data and the Semantic Web (RDF/OWL) to capture the relations between data from several experiments and the biological hypotheses related to them; (ii) the Concept Web as a reference for disambiguated biological concepts and ‘nano-publications’; (iii) ‘Concept Mining’ methods to mine knowledge from text and other sources in terms of the Concept Web; (iv) Taverna, myExperiment.org, BioCatalogue.org, and the 'AIDA' plugin for Taverna to capture the design and provenance of a bioinformatics analysis workflow in relation to its results and biological interpretation.

Results

We present a workflow that helps interpret the pre-processed data from genome wide association studies in the context of two case studies: (i) elucidating the role of epigenetic factors in the progression of Huntington’s Disease and (ii) finding relations between BioBanks (a pilot by the Free University of Amsterdam, the Erasmus Medical Centre Rotterdam, and the Leiden University Medical Centre). We present semantic (SPARQL) queries that retrieve evidential relations, and putatively novel relations across experiments.

Discussion

We conclude that adoption of web standards to expose (in a controlled manner) all elements of bioinformatics experiments, including the biological concepts that make a hypothesis, has the potential to enhance biological research by enabling computational support for reasoning over the multitude of data and information that may be relevant for a better hypothesis.

Presenting Author

Marco Roos (m.roos@lumc.nl)
Leiden University Medical Centre (LUMC) & University of Amsterdam (UvA)

Author Affiliations

1) Leiden University Medical Center 2) University of Amsterdam 3) Erasmus Medical Center Rotterdam

Acknowledgements

This work was supported through the Netherlands BioInformatics Centre (NBIC; BioRange program), the Virtual Laboratory e-Science Project (Dutch Ministry of Economic Affairs) and the Centre for Medical Systems Biology (CMSB).

F-26. First CALBC challenge: first results

Rebholz-Schuhmann D (1,), Jimeno Yepes A (1), Li C (1), van Mulligen E (2), Kors J (2), Milward D (3), Hahn U (4)*

CALBC is a support action project that brings together the researchers from international biomedical text mining groups to address the difficult issue of annotating large text corpora with a large set of semantic types [1, 2]. CALBC introduced a collaborative approach to this annotation task in the form of an open challenge to the biomedical text mining community. The task is the annotation of named entities in a large biomedical corpus, for a variety of semantic categories [3]. CALBC will deliver a large, collaboratively annotated corpus, marked with the mentions of biomedical entities.

Materials and Methods

The corpus was a selection of 100k Medline abstracts. The documents have been automatically annotated by the 16 participants of the first CALBC challenge round (22 submissions overall). The harmonization of these annotations will be made available to the research community in Juli 2010 and will serve as a training corpus for the CALBC challenge II (second half of 2010).

Results

From the 22 submitted annotations to the first CALBC challenge, 9 were trained on the 50k abstracts training corpus. 14 of these submissions annotated disorders, 19 annotated proteins and genes, 13 annotated species, and 14 annotated chemicals. The best performing solutions for the annotation of the boundaries in the corpus reached an F-measure above 80% across all semantic groups.

Discussion

Submissions of annotations to the CALBC annotation server contain positional information on the phrases that have been recognized as biomedical entities and information on the medical entity type. For the biomedical entity types we used a classification that is a variant of McCray's classification [4]. Optionally, the text mining systems could provide annotations for term from additional lexical resources (MeSH, GO, etc.).

URL

http://www.calbc.eu

Presenting Author

Dietrich RG Rebholz-Schuhmann (rebholz@ebi.ac.uk)
European Bioinformatics Institute

Author Affiliations

1 EMBL-EBI, Hinxton, U.K. 2 Erasmus Medical Center, The Netherlands 3 Linguamatics, U.K. 4 Friedrich-Schiller-Universität Jena, Germany

Acknowledgements

This work was funded by the EU Support Action grant 231727 under the 7th EU Framework Programme within Theme “Intelligent Content and Semantics” (ICT 2007.4.2).

F-27. EuroPhenome: a repository for high-throughput mouse phenotyping data

Morgan H (*), Hassan A, Blake A, Hancock JM, Mallon A-M

The broad aim of biomedical science in the postgenomic era is to link genomic and phenotype information to allow deeper understanding of the processes leading from genomic changes to altered phenotype and disease. Essential to developing such a linkage are databases which contain information on inbred mouse strain and mutant phenotypes. EUMODIC is gathering data from the EMPReSSslim pipeline which is performed on inbred mouse strains and on knock-out lines arising from the EUCOMM project. The EuroPhenome interface allows the user to access the data via the phenotype or genotype.

Materials and Methods

The EuroPhenome database is implemented in MySQL running on Solaris. The Phenome Data Viewer is implemented as a Java Servlet and the PhenoMap and Ontology Tree are implemented in PHP. All the web site is styled using css and uses DHTML and AJAX to provide the relevant functionality. The images are generated by JFreeChart. Memcached is used to improve performance. Particular pheodeviants are annotated with terms from the Mammalian Phenotype Ontology.

Results

EuroPhenome allows users access the raw data and the annotated data described above through three new integrated web tools. The ‘Phenome Data Viewer’ allows access to the raw data, presented graphically. All lines within Europhenome are compared to relevant baseline animals and significant differences are recorded and are viewable by the ‘PhenoMap’ and the ‘Mine for a Mutant’ Tool.

Discussion

Europhenome has data for 151 mutant lines, including a total of over 2 million data points. This will increase to 500 mutant lines over the next 18 months. This provides useful data for the examination of individual genes, finding mutant lines if interest to a particular field of research and analysis of the data set as a whole to find interesting biological phenomena.

URL

http://www.europhenome.org

Presenting Author

Hugh L Morgan (h.morgan@har.mrc.ac.uk)
MRC Harwell, Mammalian Genetics Unit

Author Affiliations

MRC Harwell, Mammalian Genetics Unit

Acknowledgements

This work was supported by the EUMODIC project and the MRC.

AUTHOR INDEX