International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue (6) NCRTCST-2013, ISSN 2249–071X

National Conference on Recent Trends in Computer Science and Technology (NCRTCST)-2013

A Frame Work for Extraction, Integration and Analysis of Unified Medical Language System

I. Venkanna
Asst. Prof., Malla Reddy Institute of Technology & Science
Maisammaguda, Secunderabad-50010
Email: freefree4u@rocketmail.com

B. Sukumar
Asst. Professor
S. S. J. Engineering College, V. N. Pally
Email: sukumar512@gmail.com

B. Sivaiah
Assoc. Professor
C.M.R.C.E.T, Medchal
Email: sivabetld@gmail.com

***Abstract* – A Unified Medical Language System (UML) is a huge repository of biomedical literature. It helps us to compute the analysis of diverse data collections and to predict links between genes and diseases. We mainly focus on the analysis of biomedical literature for the identification of genes encoding DNA binding transcription factors in neurological diseases. We also use the biomedical literature and organized ontologies and vocabularies for both gene and diseases .To connect these data sources, we use both manually and automatically annotated linkages, such as the reviewed user-submitted Gene Reference into Function (GeneRIF) annotations in Entrez Gene and the computationally generated Related Articles from Pub Med.**

***Keywords* – UML, Transcription Factor, DNA, GeneRIF, Entrez Gene, PubMed, MeSH, Ontologies.**

I. INTRODUCTION

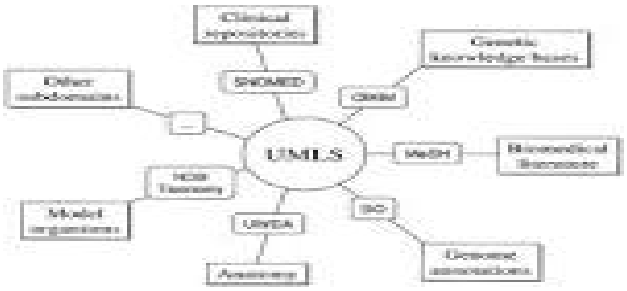

Fig.1. The various sub domains integrated in the UMLS.

We are interested in data integration, looking at discovering novel links between transcription factors and brain diseases. This will provide a framework to model existing relationships and prediction of novel relationships between genes and diseases. Biomedical motivation: As transcription factors are a key mechanism regulating gene expression in cells, we shall exploit their role and the availability of new transcription factor specific data and databases. We shall also focus the scope of the project on brain diseases, as many brain-related projects come online. Extending the existing techniques for analysis to the currently mostly manual evolution of linkages.

The paper can be subdivided into several parts. An ongoing review of existing data sources and methods relevant to the paper, extraction and integration of data for selected sources and relationships, the analysis and prediction of the relationships, and verification of our results. Ultimately, the results of the project will be made publicly available, as well as all the tools developed and their source will be made available for further First step, we plan to review existing literature relevant to this paper. This step ensures that the project remains relevant. We shall examine existing data sources for genes and diseases, and sources that link these to evidence. We shall also examine methods of integrating this data, and techniques for analyzing existing linkages and computing novel linkages.

Second step, will be extraction of data from the selected sources and integration of this data into a database. This will allow a unified point of access for analysis. The database will store the various data sources, such as Pub Med (medical literature), OMIM (human disease) and TF-CAT (transcription factors). It will also store direct links within each data source, such keywords in Pub Med or classes of transcription factors in TF-CAT, as well as direct links between data sources, such as Pub Med references in OMIM entries. We will also derive secondary links - for example, disease terms in Pub Med abstracts can be linked to the relevant OMIM entries, and transcription factor gene references in TF-CAT can be linked to Pub Med. We shall also look to interoperation between the databases, by extracting and mapping ontology terms and gene names from free text.

Brain is a very heterogeneous organ, containing an extremely large number of cell types. Diseases in the brain have some known associations to genes, yet most diseases are complex, involving multiple interacting genes. Phenotype characterization can be complex, based on biased, subjective analyses. While some animal models exist, mapping the results of animal studies to human diseases is often difficult due to the apparent dissimilarity between animal and human brain behavior.

Both development and function of brain cells are carefully mediated by transcription factors (TFs). By binding directly to the DNA, changing the structure of chromatin and recruiting other transcriptional units to modify the transcription of genes. Although much success has been had implicating single genes as the cause "simple" diseases, many disease have far more complex causal relationships. Genes may be sufficient but not necessary, or necessary but not necessary. Other external environmental factors may be involved. Even the so-called simple genes are often modified by genetic factors, resulting the broad spectrum of phenotypes we see.

All copyrights Reserved by NCRTCST-2013, Departments of CSE
CMR College of Engineering and Technology, Kandlakoya(V), Medchal Road, Hyderabad, India.

II. OBJECTIVES

Open Access: The only freely available data sources will be used. Tools and analysis results will be made publicly available as free software and published in open access journals.

Integrated, Unified Repository: The system will be designed to provide a complete storage solution for genes, diseases and evidence from disparate databases, as well as existing and computed annotations and relationships. A consistent interface will allow straightforward access to all the data.

Efficient Programmatic Framework:

A comprehensive toolkit for analysis will be developed, using scalable algorithms to handle the large, expanding datasets involved.

Exploration of Gene-Disease Relationship

The tools will be used to examine existing annotations of gene-disease relationships in literature. From this, we aim to develop a method to infer novel relationships and evaluate the accuracy of such inferences quantitatively using statistical techniques such as overrepresentation.

Example Output

- A list of Pub Med IDs of papers implicating gene X with disease Y
- A list of genes highly correlated with disease Y, with p-values
- All diseases that gene X plays a role, with p-values

III. LITERATURE SURVEY

Gene Seeker: Given a chromosomal region and expression locations, such as particular limbs or organs, Gene Seeker searches multiple databases in parallel to find candidate human genes expressing in the selected tissues that lie in the specified chromosomal location. It will also search for expression of orthologous (via MGD) mouse genes.

Disease Gene Prediction (DGP): They investigate the use of phylogenetic features for distinguishing disease genes, using a decision tree-based model[3].

PROSPECTR: By looking at sequence-based features (e.g. gene length, coding sequence length, GC content, conservation with mouse homolog), an automatic alternating decision tree classifier was trained to rank genes for their likelihood of involvement in disease [4].They that gene length and protein length alone were fairly effective predictors of potential involvement in disease.

SUSPECTS: They extend PROSPECTR, combining sequence-based approach with annotation-based information. Scoring looks at sequence features via PROSPECTR, but also rare shared protein domains (via InterPro), semantic similarity (via GO terms), co expression with a training set (GNF expression data). Scores are integrated, and weighted depending on the amount of information available, to avoid bias introduced by a line of evidence with little support.

CAESAR: Searches all genes using text and data mining, by initially specifying some input corpus of text on the disease in question (e.g. OMIM entry). The input text is analyzed for occurrences of genes and ontology terms. Similarity between a term and the corpus is done by creating a vector of word occurrences for the term and description, and for the corpus, then computing the cosine of the angle between the vectors (dot product divided by magnitude) – larger cosine equates to greater similarity. Eight sources of gene-centric information are used to map the various ontology terms to gene annotations. For example, Mammalian Phenotype terms are mapped to MGD, eVOC terms are used to query UniProt, GO terms matched to GOA, and extracted gene names used to query BIND, HPRD, KEGG and InterPro. Scores are integrated into a combined score, using methods such as sum, mean or maximum, as well as a fourth method “int4”, which is done both at the gene source level, and then to integrate the scores from the multiple sources.

IV. OTHER RELEVANT WORK

Alison Meynert's Common Evidence Network investigated quantitative measures of similarity between the GO terms for two genes. Other tools such as GO Toolbox look at the overrepresentation of GO terms in a set of genes, compared to a background set of genes, using various statistical tests (hyper geometric distribution, Chi-Square) with correction for multiple hypothesis testing.

- Automatic annotation
- interactions (protein-protein), concept mapping, location mapping
- protein, gene and sentence structure extraction
- text analysis, machine learning

Existing Data Sources: Here I will review existing resources as they pertain to the project. These are composed of the data sources, as well as tools for analysis and discovery of linkages between this data. Here I shall summarize existing sources of data, as well as the tools currently available and analytical techniques used in similar problems.

Entrez Gene: From the National Center for Biotechnology Information (NCBI), Entrez Gene extends the NCBI Locus Link project, providing a gene-centric view of the information at the NCBI. This database tracks genes annotated in genomes, from protein coding regions (e.g. in viruses), to predicted genes, in addition to known genes.

A unique gene identifier is assigned for each gene in each species. Data in Entrez Gene comes from both curated and automatically generated sources. This includes information from and links to sequences in NCBI Reference Sequence (RefSeq). Gene Ontology (GO) annotations are provided by the Gene Ontology Annotation (GOA) Database.

All copyrights Reserved by NCRTCST-2013, Departments of CSE
CMR College of Engineering and Technology, Kandlakoya(V), Medchal Road, Hyderabad, India.

GeneRIF: Gene Reference into Function (GeneRIF) annotations are provided by the public and the National Library of Medicine, describing references to gene function. Gene function in this case defined very broadly, referring to not only biological function, but also information about the gene's role in disease, its discovery and mapping. In addition to these basic GeneRIFs, there are also two other headings of GeneRIFs: information from HIV-1, the Human Protein Interaction Database, and information about general interactions from BIND, BioGRID, EcoCyc and HPRD. All GeneRIFs include a reference to at least one Pub Med article as evidence.

Gene Ontology: GO terms for a collaborative effort to provide a consistent nomenclature for gene annotations and for indicating the strength of the evidence supporting such annotations. In addition to the three original members, the model organism databases FlyBase, Saccharomyces Genome Database external link (SGD) and the Mouse Genome Database (MGD), there are now over ten full members, including GOA, and several associate members. GO is composed of three main ontologies - biological processes, cellular components and molecular functions. Annotations are described by a three-letter controlled vocabulary of evidence codes, allowing for the circumstantial inferred by electronic annotation (IEA) to the more concrete traceable author statements (TAS). However, GO does not describe "abnormal" features, such as mutant or disease-specific traits. The Gene Ontology Annotation Database is responsible for annotations to proteins in the human, chicken and cow genomes in UniProtKB, and is supplemented by annotations from other groups. Priority is given to proteins without annotation, those with disease relevance and those relevant to high-throughput analyses.

The GO ontology as well as the GOA collected annotations is also available directly from the Gene Ontology Website.

Statistics:

Total Genes	: 2631524
Transcription Factor (TF) Genes	: 7866
Human Genes	: 38624
Human TFs	: 1209
Total Genes with GeneRIFs	: 33216
Human TF Genes with GeneRIF	: 798
Human TF Genes with GeneRIF (including Interactions and HIV Interactions)	: 914

Access:

Data from Entrez Gene (and other databases at the NCBI, including PubMed) can be queried using the NCBI EUtils interface, allowing users to craft HTTP queries and retrieve XML formatted results. GeneRIFs can also be downloaded as gzip compressed tab-delimited files from [ftp://ftp.ncbi.nih.gov/gene/GeneRIF].

Other Sources for Transcription Factors

I shall also examine the integration of other data sources to increase both the coverage of transcription factors as well as providing more direct links to literature. For example, curated sequence databases, such as UniProtKB /Swiss-Prot, may have a different set of annotations (e.g. UniProt Keywords). The locally developed TF-Cat database provides specialized, annotated resources for transcription factors, and the PAZAR database identifies regulatory elements associated with genes, potentially revealing interconnected regulatory programs.

PubMed: PubMed is a searchable citation database at the NCBI, indexing biomedical literature. Bibliographical citation information is taken primarily from the National Library of Medicine (NLM) MEDLINE database, although some journals indexed for their biomedical articles have all their articles indexed, and there are also legacy articles from OLDMEDLINE, as well as other initiatives that experimented with indexing other scientific literature.

MeSH: PubMed/MEDLINE entries are continually being indexed using Medical Subject Headings (MeSH), a controlled vocabulary thesaurus of descriptors, arranged in a hierarchical structure. Sixteen main categories (e.g. Anatomy, Disease) at the top are divided into subcategories, and then the descriptors are placed into the tree, with more general terms near the top to the most specific, with a descriptor potentially occurring more than once in the tree. Of particular interest is tree number C10.228.140, "Brain Diseases". An article is indexed by one or more MeSH terms, each of which may also have one of 83 topical qualifier subheadings (e.g. analysis, education or therapy) to potentially indicate a more specific topic.

Access: PubMed and MeSH can be accessed via the Entrez EUtils interface. They can also be licensed from NLM (free for research).

Statistics

PubMedarticles	: 16,120,074
PubMedarticleswithMeSHheadings	: 15,806,221
PubMedarticleswithBrainDiseaseMeSH(ormorespecific)terms	: 660538
MeSHterms	: 47143
UniqueMeSH terms	: 24355
MeSH terms under Brain Diseases	: 312

V. UNIFIED MEDICAL LANGUAGE SYSTEM

The UMLS Met thesaurus contains database of medical terminology, provided by the National Library of Medicine. It provides a mapping to concepts (with unique concept identifiers) from vocabularies such as Medical Subject Headings (MeSH - used to index MEDLINE), the International Classification of Diseases (ICD – used for reporting clinical diagnoses in healthcare) and SnoMed CT (an emerging international standard for medical terms).

Integration: Our project will focus on unifying information on genes, literature evidence and disease into a single controlled resource, which we shall use to examine both the existing relationships between genes as

All copyrights Reserved by NCRTCST-2013, Departments of CSE
CMR College of Engineering and Technology, Kandlakoya(V), Medchal Road, Hyderabad, India.

well as infer novel relationships. I intend to develop a single, consistent framework for accessing and manipulating this data efficiently. I shall focus on making the results reproducible as well as provide tools for further analysis.

Preliminary statistics and the prototype currently built indicate that the amount of information involved leads itself naturally towards a databases solution.

Programming Tools: The Atlas Data Warehouse provides a local repository for sequence information, and of particular interest, also provides a source for several protein-protein interaction databases.

VI. Preliminary Analysis and Proposed Methods

Integrated Database Back-End: Several reasons necessitate the creation of a separate database for this project, rather than accessing the online versions of the databases. By creating our own database for this project, we can optimize the form and interface to the data stored. Validation will also be simplified, as freezing the current state of the data is simply keeping an archived copy of the database at a particular point in time.

Tools/Access: We aim to ultimately provide access to researchers in general, as well as more direct access for bioinformaticians

- *Met thesaurus:* Terms and codes from many vocabularies, including CPT®, ICD-10-CM, LOINC®, MeSH®, RxNorm, and SNOMED CT®
- *Semantic Network:* Broad categories (semantic types) and their relationships (semantic relations)
- *SPECIALIST Lexicon and Lexical Tools:* Natural language processing tools

To this end, I will design a web-accessible interface to allow researchers to perform straightforward queries on the data, via both an interactive interface as well as a batch querying interface (cite example?). I will also provide the means for any researcher to generate their own version of the database by releasing the source code under an Open Source license and publishing the methods and results in Open Access journals (further motivate?).

We hope to integrate various statistical models to measure strength of relationships between genes and diseases. I shall design the system to be easily extensible, to allow integration of several data sources, and to ease the addition of new data sources or for analysis of new classes of disease and genes.

- Free text indexing (Lucerne?)
- Extensibility to other data sources
- Similarity Measures
- free text similarity
- Sequence level similarity
- keyword overlap

Analysis

- Numerical Scoring methods
- statistical comparison
- A better model for relatedness?
- model of information and processes
- Adaptation of existing methods for overrepresented terminology integration with quantitative information.

VII. Validation

We plan to use two methods to validate the results obtained: comparing results obtained using a frozen snapshot of the databases against the most recent versions of the database, and checking the results against an external data source. Both of these will involve holding some data out from the analysis set, which will then be used to compare against the results.

Due to the interconnected nature of the data sources involved, cross-validation and other simple test strategies are difficult to employ. These techniques require that the data sources can be separated into independent entities.
The first method will involve archiving a copy of the databases to be used as a snapshot dataset. The results obtained by analyzing this dataset can be compared against the up-to-date versions of the databases and other data sources. We shall look for instances where predictions made are confirmed by more recent experimental evidence. In addition to storing a duplicate copy of the original data, we can also use publicly available archives of the data, such as the yearly MEDLINE Baseline Repositories, or the update-dates of time-stamped fields in the XML files, such as for Entrez Gene GeneRIFs.

Another method of verifying the predictions would be to compare results against an existing curated reference collection. The Online Mendelian Inheritance in Man (OMIM) database provides free-text, curated information about genetic disease, listed by gene and disease. By not providing the information in OMIM to the system, we can later verify our results by examining the relevant OMIM entries.

Finally, it would also be possible to correlate the results with the results against large-scale gene-disease correlation studies, where analysis was done using only the data generated within the experiment. This would have to combined with the a snapshot of the databases before the publication of the data, as of course once the data is incorporated in public databases, it could be referenced and subsequently used or added to annotations of the system.

Also to be noted is that due to the nature of research, these methods can only minimize the contamination — none of the techniques compensate for the influence of private sharing of results between researchers before publication.

All copyrights Reserved by NCRTCST-2013, Departments of CSE
CMR College of Engineering and Technology, Kandlakoya(V), Medchal Road, Hyderabad, India.

REFERENCES

[1] Lindberg, DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Meth Inf Med. 1993[PubMed]
[2] Humphreys, B. L., Lindberg, D. A., Schoolman, H. M., & Barnett, G. O. (1998). The Unified Medical Language System: an informatics research collaboration System. J Am Med Inform Assoc,
[3] Aronson, A.R. (2001) Effective mapping of biomedical text to the UMLS Metthesaurus: the MetaMap program.
[4] Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology System terminology..Nucleic Acids Res 32 (Database issue).
[5] Current descriptions, documentation, and information about obtaining the UMLS Knowledge Sources are available from NLM's Web site: www.nlm.nih.gov
[6] [ftp://ftp.ncbi.nih.gov/gene/GeneRIF]
[7] "GeneSeeker"vanDrielet.al.2005 (Netherlands).
[8] "PROSPECTR:"Adie et al., 2005 (Edinburgh, UK)
[9] "SUSPECTS" Adie et al, 2005
[10] "CAESAR" Gaulton et al 2007 (Chapel Hill, NC)
[11] "Genes, Behavior, and the Social Environment**"** Lyla M Hernandez and Dan G Blazer.
[12] "Genetic Engineering" Smita Rastogi, Neelam Pathak.
[13] "Data Mining: Concepts and Techniques", Jiawei Han and Michelin Kamber
[14] "Genetic Algorithms", David E Goldberg
[15] "Information Storage and Retrieval Systems ",Gerald J Kowalski and Mark T Bury
[16] "New Trends in Software Methodologies, Tools and Techniques", Hamido Fujita and Paul Johansson
[17] "Computer Medical Databases: The First Six Decades (1950–2010)**"** By Morris F. Collen.

All copyrights Reserved by NCRTCST-2013, Departments of CSE
CMR College of Engineering and Technology, Kandlakoya(V), Medchal Road, Hyderabad, India.