

Integrated Biomedical Databases: Microarray, Molecular Biology Integration, Natural Language Processing, and Ontology Mapping

Jeremy L. Phillips, Carlos F. Santos, Jing Gao, Sirarat Sarntivijai, Alexander S. Ade, David J. States M.D. Ph.D.
The National Center for Integrative Biomedical Informatics and the University of Michigan Bioinformatics Program, Ann Arbor, Michigan

Abstract

The biomedical sciences are diverse and generate many different data types ranging from primary gene sequence to transcript expression, proteomics identifications and text. Publicly available biological data resources are expanding rapidly, but these resources are often weakly interconnected. For instance, a data set derived from a microarray experiment is unlikely to contain links to relevant contextual information in the biomedical literature. In addition, a single type of biological data is often distributed over several heterogeneous databases. Moreover, because web sites are most often the primary gateway into biological data repositories, users of these repositories are often unable to extract data with the same flexibility that is available in a standard query language such as SQL. Further, different data sources may overlap and often are incomplete or even contradictory. As a partial solution to these problems in three specific domains, we have created three closely interconnected in-house biological relational databases: The Molecular Biology Integration Database (MBI), the NCIBI Microarray Repository, and a Biological Natural Language Processing Database (BioNLP). Moreover, we are developing a method to intergrate information from free text and biomedical ontologies such as GO and MeSH to provide a more in-depth understanding of the context of biological problems.

MBI is a repository of human, mouse, and chimpanzee sequences and associated annotations from several public sequence databases. By joining sequence annotations from these disparate sources into single, unique sequence records, MBI serves as a crossing point between both public and local data sources. The Microarray Repository is a collection of both raw and normalized data from microarray experiments available in the public domain and from private contributors, covering several technologies and platforms. BioNLP contains a collection of biological named entities gleaned from the biomedical literature along with relationships between these entities, as an endpoint for the NCIBI natural language processing pipeline.

These databases are all built using standard relational database technology (Microsoft SQL Server), allowing for users to easily construct queries to gather data from all three databases. In addition, the databases are built to maximize query simplicity, allowing for easy and flexible access via standard SQL. Using MBI, the Microarray Repository, and BioNLP, NCIBI investigators will be able to easily cross reference biological named entities and expression data with biological sequences and sequence annotation.

Microarray Repository

Introduction

The NCIBI Microarray Repository collects Microarray data from different resources for researchers to analyze. The data in microarray database comes from NCBI GEO, CaArray, EBI ArrayExpress, DGAP (Diabetes Genome Anatomy Project) and NCIBI collaborators from University of Michigan and Johns Hopkins University. We have collected 26249 samples and 496 million observations from 14 species and 70 Affymetrix platforms. We will use this database for integration of expression data with external sources, for instance, the co-clustering of protein interaction and co-expression data.

We are in the process of renormalizing several raw data sets from Affymetrix GeneChip experiments using the Robust Multichip Analysis (RMA) procedure in order to standardize data for meta-analysis. We are normalizing data with both standard data definition (CDF) files from Affymetrix and custom CDF files developed by NCIBI investigator Dr. Fan Meng. These custom CDF files contain updated probe set mappings for several GeneChip platforms.

Current Database Contents

Source	Associated Database	Samples	Description
Gene Expression Omnibus (GEO)	Soft	23736	Microarray gene expression data in soft format
	Raw	3628	Microarray gene expression raw data
Diabetes Genome Anatomy Project (DGAP)	ChipperDB	432	Gene expression raw data for diabetes study
National Cancer Institute	caArray	60	Microarray repository with MIAME compliant and cel format data
European Bioinformatics Institute	ArrayExpress	20	Repository with MGED recommendation format and cel file data
NCIBI Collaborators	Microarray Data set	1502	Data from Drs. Angel Lee, Matthias Kretzler at U of M and Haiming Chen at Johns Hopkins University

MBI

Introduction

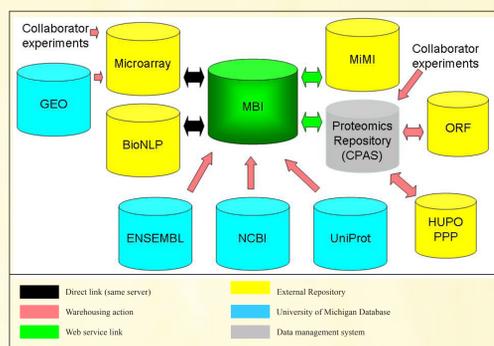
The MBI database is a large collection of biological sequences and sequence annotations collected from several public sequence repositories, including those provided by EBI and NCBI. MBI also warehouses both the NCBI (Entrez) and Ensembl gene models, containing links from gene to protein to sequence. Thus, for instance, given a gene name, a user can easily find all splice variant transcript and protein sequences for that gene, and all associated annotations for those sequences.

In addition to being a cross-reference point for sequence identifiers and sequences, MBI is a source of information for the chromosomal locations of various DNA features. The database is currently built to store the locations of genes, transcription factor binding sites, and CpG islands, but can be easily extended to anchor SNPs, microarray probes, and other items to genomic locations.

Finally, MBI is a reference point for groups of related or highly similar sequences. Groups of homologs and protein families are currently stored as MBI sequence sets containing high-level descriptive information.

Context and Database Interconnectivity

MBI is directly or indirectly linked to several previously existing biological databases at the University of Michigan. The Microarray and BioNLP databases described in this poster are stored on the same database server as MBI, allowing for the easiest possible generation of cross-database queries. The Michigan Molecular Interactions Database (MiMI) and others are linked to MBI via a publicly available SOAP web service, which will allow MiMI users to gather extensive annotation on proteins of interest in the future. In addition, web services will allow access to MBI from the Computational Proteomics Analysis System (CPAS), a web-based system for processing analyzing and storing tandem mass spectrometry experimental results. Allowing these disparate databases to easily access MBI will allow MBI to support cross-domain global analysis in molecular biology.



Methods

MBI is built by performing full length sequence comparisons between all sequences in each source database. Initially, annotations for fully identical sequences are grouped by a single unique sequence identifier. Sequence identifiers for homologs or sequences belonging to the same UniGene cluster are later combined into sequence sets. Annotations for DNA or Protein features are grouped based on matching coordinates on the host sequence, rather than by sequence comparisons.

Current Database Contents

Source Name	Content	Human	Mouse	Chimp
Ensembl	Gene, transcript, and peptide sequences and IDs. Gene to Transcript to Protein links.	X	X	X
Entrez Gene	Entrez Gene IDs. Gene to Transcript to Protein links, merged with Ensembl genes where possible.	X	X	
Homologene	Homologous sequence clusters	X	X	
RefSeq	Genomes, transcripts, and protein sequences	X	X	X
Swiss-Prot / trEMBL	Protein sequences and feature tables	X	X	
Swiss-Prot Varsplice	Swiss-Prot splice variants, linked back to originating transcripts	X	X	
GO	GO Identifiers function descriptions.	X	X	
PIR	PIRSF (PIR family classification resource).	X	X	
IPI	Protein sequences, and sequence identifier histories.	X	X	
UCSC	UCSC STS markers and synonyms	X		

BioNLP

Introduction

We have developed an automated natural language processing pipeline to assist in the linking of literature to biological sequence databases, allowing us to automatically generate from the literature up-to-date protein-interaction network and signal pathway models from biomedical literature. The pipeline processes articles retrieved from private full-text publishers (HighWire Press and Science Direct) and the public abstract repositories (NCIBI Pubmed/Pubmed Central) into a standardized format which identifies key biological entities within individual sentences and generates a structured knowledge base from those entities via a full text parser. The pipeline yields a collection of structured assertions that can then be cross-indexed against human, mouse, and chimpanzee and associated annotations from several public sequence databases in the NCIBI's Molecular Biology Integration (MBI) database. By joining sequence annotations with textual assertions from these disparate sources into single, unique sequence records, the pipeline allows the full merging of documents with high-throughput data sources like the the Microarray Repository and MBI.

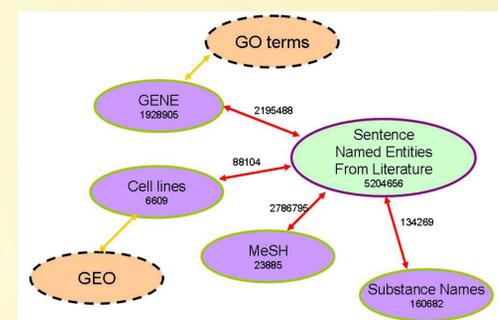
Current Database Contents

Database	Document	Individual Genes
dnabp	6987	9330
signal	5	69
signal2	15	276
regseq	1293	5212
c-myc	1841	7271
heme_regseq	135	1100
WntSignalingExpanded	2489	6757
polycomb.cit.xml	262	2687
bipolar	1	10
polycomb	322	3045
rtk	852	3347
prostatecancer	18819	9552
retinoblastoma	212	2330
androgenreceptor	861	3689
sig.cit.data.xml	2341	6269
WntSignalingCore	283	2040
ComparativeMH	2332	8668
sigtrans	151	1527
nfkB	1182	3278
NFKappaBRegulation	1239	3280
AutoimmuneDiabetes	1834	3336
dp2003	1603	6042
Hematopoiesis_core	525	1903
sleep	25	139
ComparativeMHBlood	1406	7375

Ontology Mapping

Integration of biomedical informatics knowledge has always been an important tool contributing to understanding driven-biology problem. Although all of these informatics studies focus on biomedical domain, they vary in the context of studies. We are developing a strategy to extract knowledge from the pool of available biological data by integrating ontologies and natural language processing.

There are two key points to this method. First, we designate UMLS as a central ontology, to which we intend to map all other major biological ontologies such as GO and FMA. Second, we add incorporate knowledge extracted from additional data sources such as PubMed, Entrez Gene, Hyper Cell line* DB, and GEO to assist in these mappings. As a proof of concept for this method, we have developed a mapping from GO to MeSH terms using NCIBI Gene and biological named entities extracted from the literature (PubMed) as intermediate data sources.



Acknowledgements

This work was completed with support from the NIH, grant numbers U54DA021519 and R01LM008106.

