Toward Making Online Biological Data Machine Understandable

A Dissertation Proposal
Presented to the
Department of Computer Science
Brigham Young University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

by
Cui Tao
May 24, 2005

1 Introduction

Catalyzed by world-wide research communities producing publicly available data, the volume of biological data is increasing at a rapid pace. [MBD05], [BGM04], and [DBC04] all list hundreds of high-quality repositories of value to the biological community. The molecular biology database collection [MBD05], just to name one, includes 719 databases as of the beginning of 2005, an increase of 171 over 2004 [Gal05]. These repositories have a large amount of "bio-information" including, for example, DNA sequences, gene expressions, gene identifications, intermolecular interactions, metabolic pathways and cellular regulation, mutation databases, protein sequences, RNA sequences, and large molecule structures. To do activities such as performing background research for a field of study, gaining insights into relationships and interactions among different research discoveries, or building up research strategies inspired by other's hypotheses, biologists need a system that can efficiently locate, understand, and extract online bio-information.

Locating, understanding, and extracting online bio-information is challenging for several reasons. There are, or soon will be, thousands of available repositories of biological data, each with its own interface supporting different invocation protocols, processing methods, and data semantics. These repositories are highly diverse in both granularity and variety, and they use different terminologies, different ID systems, and different units of measurement to describe the same concepts. Most of these sources are unstable and unpredictable; independent developers modify their designs and schemas, remove or add data dynamically, and occasionally block access to sources for maintenance or other purposes. Many of the sources provide forms through which a user can query a source; but each individual source provides different forms, and each form only supports certain types of queries.

Sometimes the information a user needs spans multiple sources. A system that traverses only one source may not answer user queries completely. Therefore, a system that can automatically retrieve, understand, and extract online biological data independent of the source is needed.

Researchers have proposed solutions that address some of these issues.

- EMBL Harvester [Har04] searches in several different repositories and returns all the cached results to users. It, however, only covers ten resources. It is hard to add new resources to the system and the system is not flexible to any change of the sources because the mappings between each source and the system are hard-coded.

- Systems such as EnsEmbl [Hub02] and GenoMax [Gen04] allow users to submit queries through their interfaces, but they only focus on a small range of biological concepts and problems and can only deal with a specific set of online repositories. Like EMBL Harvester, these systems are not robust to changes in the sources, and it is hard to add new sources to them.

- It is straightforward to add new data sources into systems such as the Sequence Retrieval System (SRS) [EA96]. SRS, however, is only a keyword based retrieval system, rather than an information extraction system. The returned results are a simple aggregation of records that match the keyword query. SRS has the same problems as online search engines —it returns many records that contain the keywords and leave it to the user to traverse all of them to find the result needed. In addition, SRS has strict requirements for source documents, and it only works with relational databases.

- DiscoveryLink [Haa01] has greater generality than SRS, EnsEmbl, and GenoMax and allows simple user queries, but DiscoveryLink cannot deal with complex source data such as nested data. Most biological data, unfortunately, is highly nested. Therefore, there is a significant mismatch between most data sources and DiscoveryLink. Like, EMBL Harvester, it is hard to add a new data source to DiscoveryLink. In addition, DiscoverLink requires SQL as its query language, which is not easy for biologists to write. Another drawback of DiscoverlyLink is that it maps sources using only C++, which is not an ideal language for a web wrapper.

- TAMBIS (Transparent Access to Multiple Bioinformatics Information Sources) [SBB$^+$00] is a retrieval-based information integration system for biologists. Unlike DiscoveryLink, TAMBIS offers a global schema and data reconciliation. It also hides the sources from the users. Although TAMBIS is more of an upper level solution than DiscoveryLink, its core component, the mapping model, is implemented manually. Therefore it is not robust to changes in a source, and it is hard to add new sources into the system. In addition, its interface is complicated and requires a user to understand the TAMBIS query language.

- LIMBO [Phi04] is a light-weight approach to molecular biological database integration. The authors claim that compared with existing work in this area, the effort needed to integrate heterogeneous data sources is lowered. But LIMBO does not provide a global structure. It "integrates" the different source information by storing data in its original structure in a data warehouse. The data warehouse has a few tables: one for data which stores source attribute-value pairs; one for relations which stores relationships between attribute-value pairs; and one for meta data, which stores the source of each attribute-value pair. This approach provides a way to store heterogenous data together. It, however, does not resolve the heterogeneity and does not actually integrate data together.

- BIS (the Biological Integration System) [LBE03] tries to focus on data integration by addressing the integration of query capabilities available at the sources. BIS introduces the notion of derived wrappers that capture Inter-schema Correspondence Assertions (ICA) [NO95]. Unfortunately these ICA rules must be pre-defined manually. Therefore, the system is not flexible with respect to changes or new sources.

- BACIIS (Biological and Chemical Information Integration System) [MWL03] is an ontology driven information integration system for life science web databases. BACIIS works based on a domain ontology called BAO (BACIIS ontology) [MWN+02]. BAO captures biological domain knowledge. It provides the essential semantic knowledge that allows other components of BACIIS to accomplish the integration. BAO provides a global schema which is sour ce data independent. Each web database has a data source schema which describes the organization and content of the corresponding web database and defines the schema mapping from the source web database to BAO. Exactly how this source schema is created, however, is not clear. Presumably, both the schema and the mapping are manually specified since there is no description about how this major task is to be done. The data source schemas are the core of the system. If they cannot be generated automatically, the system is not robust to changes, and it is not easy to handle new sources.

- GenMapper (Genetic Mapper) [DR04] tries to integrate molecular biological annotation data. Like LIMBO, GenMapper does not have a global schema. It tries to physically integrate all data in a central database by using a generic data model called GAM (Generic Annotation Management), which represents annotation from different sources. GenMapper works in two phases: data import and view generation. In the first phase, the system downloads the data sources and then parses and imports the data into a central relational database according to the generic GAM representation. The system then needs to obtain annotation views to serve specific queries. The authors of [DR04] assume that the mappings between a target query and a source are given or can be derived as a composition of existing operations. Inferring the mappings in the biological domain, however, is non-trivial.

For my dissertation, I propose a system that can automatically locate and extract online biological data, and that can also automatically annotate biological data with respect to an ontology and thus make it machine understandable as semantic web data. I call this system BALLET (Biological dAta onLine: a Location and Extraction Tool).

2 Thesis Statement

There are huge and growing amounts of biological data that reside in different online repositories. Most of these web-based sources only focus on some specific areas or only allow limited types of user queries. To obtain needed information, biologists usually have to traverse different web sources and combine their data manually. In my dissertation research, I propose BALLET, an extraction-ontology-based system, that can help users to overcome these difficulties. I hypothesize that BALLET can achieve these goals:

1. Given an extraction ontology for molecular biology and a semi-structured source page in this domain, BALLET can automatically understand the structure of the source page, recognize attribute-value pairs in the source page, and match source attributes with ontology attributes.

2. Given a list of molecular biology resources and an extraction ontology for molecular biology, BALLET can index the resources with respect to both the meta data in the ontology and the data in the resources and this index can be used to locate the proper resources for a given query by machines.

3. Given an extraction ontology and a source page in this domain, BALLET can automatically convert the source page into a machine-understandable, semantic web page.

4. Given an extraction ontology for molecular biology and a semi-structured source page with structured information in this domain, BALLET can automatically do some interesting updates of the ontology, and thus can improve its ability to extract, locate, and make online biological data machine understandable.

Achieving these goals will make it possible to provide the basis for a query system that can automatically query target web resources, efficiently retrieve useful information, and return query results to users. To evaluate BALLET, various tests will be performed to verify the efficiency and effectiveness of the system.

3 Research Description

3.1 Extraction Ontologies

An application extraction ontology is at the core of BALLET. Before describing the proposed BALLET system, I need to introduce extraction ontologies and the ontology based extraction technology developed by the Data Extraction Group (DEG) at Brigham Young University.

An extraction ontology is a conceptual-model instance that serves as a wrapper for a domain of interest. For BALLET, the domain is molecular biology. The conceptual-model instance includes objects, relationships, constraints over these objects and relationships, descriptions of strings for lexical objects, and keywords denoting the presence of objects and relationships among objects. When we apply an extraction ontology to a web page, the ontology identifies objects and relationships and associates them with named object sets and relationship sets in the ontology's conceptual-model instance and thus wraps the recognized strings on a page and makes them "understandable" in terms of the schema specified in the conceptual-model instance. In previous research, the Data Extraction Group (DEG) at Brigham Young University has experimented with many ontologies for real-world applications such as car ads, apartment rentals, obituaries, pharmaceutical drugs, and many others. These experiments indicate that ontological conceptualization over recognized data items is a promising way to achieve the goal of semantic agreement over heterogeneous sources.

My dissertation research is based on the DEG ontology extraction technology, BYU Ontos. I will extend BYU Ontos and use it to deal with problems such as source-page understanding in the biological domain, indexing of biological resources, biological ontology evolution, and query answering in the biological domain. Although my research just focuses on molecular biology, the approach is likely to be general and applicable to other application domains.

For BYU Ontos, each domain has its own extraction ontology. BALLET works based on a gene extraction ontology (GEO). Figure 1 shows a partial graphical version of our GEO. (It is partial because only contains concepts and relationship sets that we need in this proposal to illustrate how BALLET works.) As we can see, the GEO contains various concepts (object sets) in the molecular biology domain such as *Gene, Gene Name, Gene Location, DNA Sequence, Protein, Protein Name, Protein Activity,* and *Source Species/Organism.* The GEO also shows relationship sets among these concepts, e.g. *Protein has Protein Activity* and *Gene has Gene Location.* There are also aggregations and generalization/specializations among object sets. For example, *Gene Location* is an aggregate of *Chromosome Number, Map, Start*, and *End*, and *Protein Activity* is a generalization of *Enzyme*, *Binding*, and many more such specializations. There are two kinds of object sets: lexical and nonlexical. A lexical object set, represented by a dashed box, contains objects whose representation is considered indistinguishable from the object itself (*Gene Name* and *DNA Sequence* are examples in Figure 1). A nonlexical object set, represent by a solid box, contains objects represented by object identifiers (*Gene* and *Protein* are examples). For each object set,

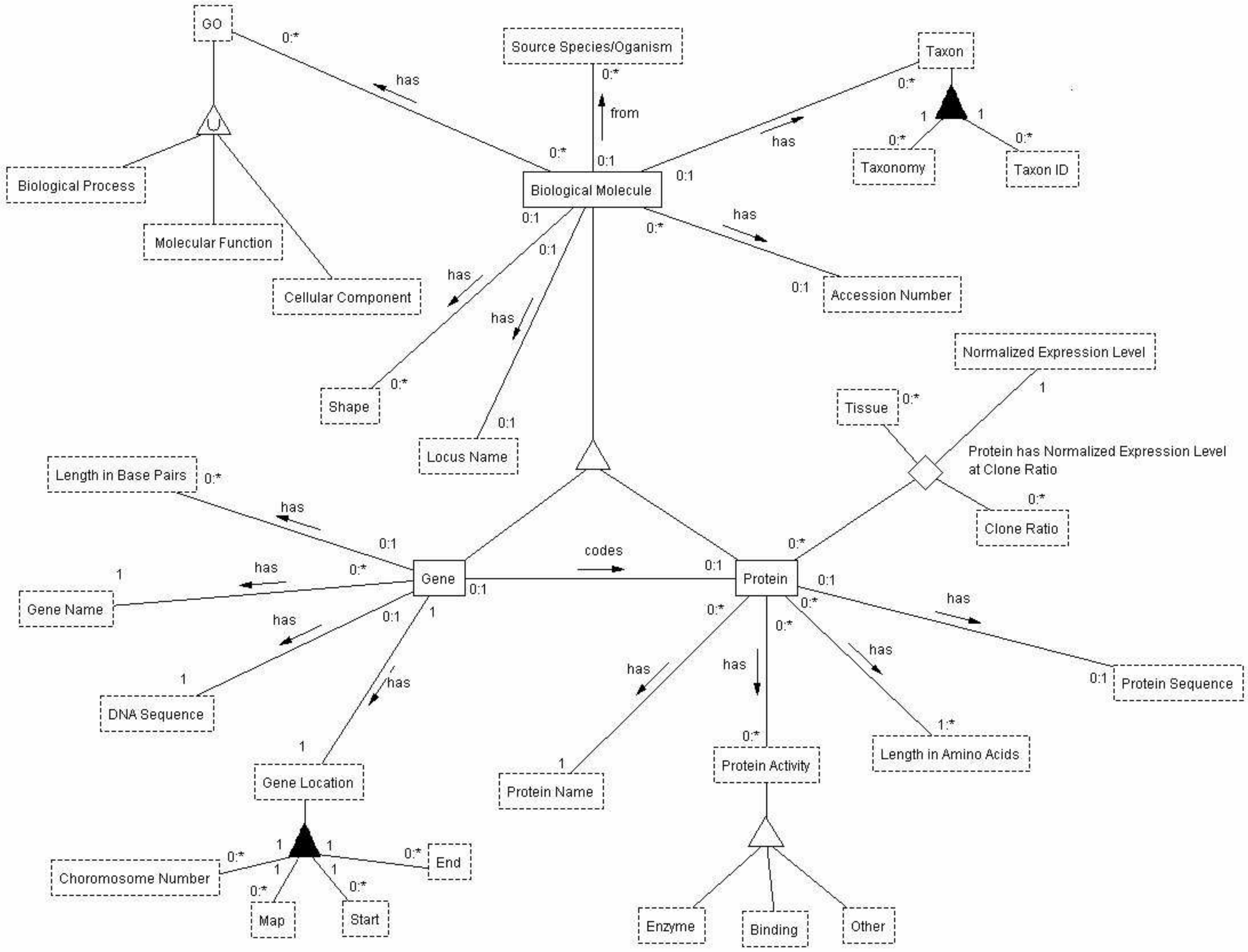


Figure 1: Graphical Version of GEO (partial)

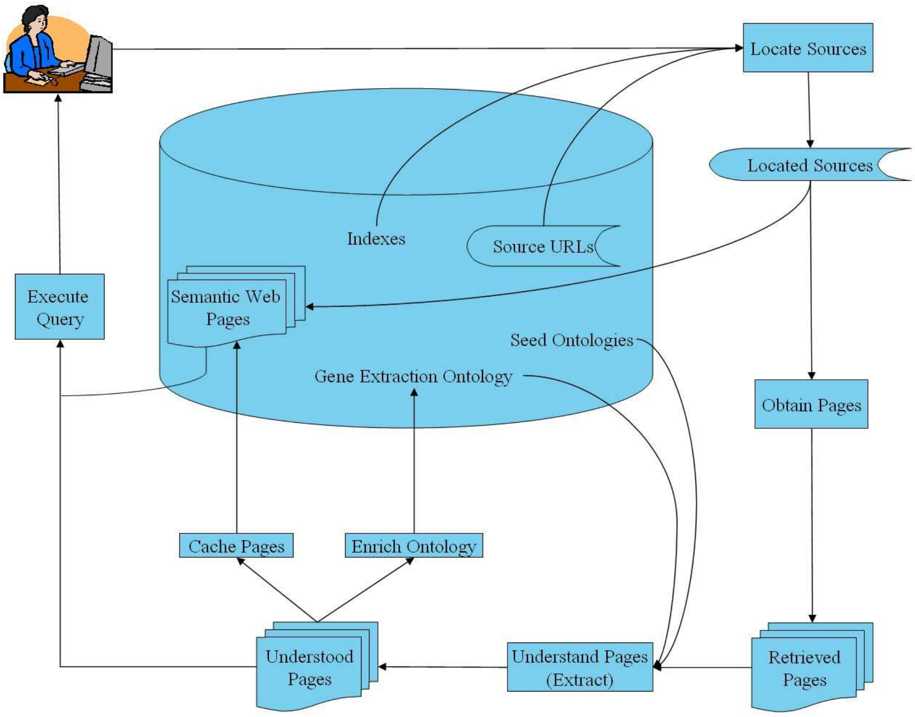


Figure 2: BALLET System Overview

there is a *data frame* that defines recognizers for value objects in the object set and keywords and keyword phrases that indicate the presence of a value of an object set. A data-frame recognizer consists of either regular expressions or lexicons of vocabulary terms or phrases that can help our system recognize values and concepts.

In order to build successful extraction ontologies for biological data, we need some trusted knowledge bases that the data frames can use for lexicons. Some of the lexicons currently available include the following. The Gene Ontology [GO04] is a generally acknowledged tool that provides a controlled vocabulary for the description of cellular components, molecular functions, and biological processes.[1] The Species Tool Kit [STK04] is a search engine that contains over 873,000 species, 120,336 common names, and 130,504 synonyms. Protein databases such as SCOP [SCO04], CATH [CAT04], and Protein Data Bank [PDB04] can provide the basis for a lexicon of protein names.

The domain of molecular biology is very broad. It is impractical to build an ontology that can cover every concept appears in different sources. We can, however, starts with an ontology that covers some concepts in the domain. We can then use this ontology to analyze different source pages and use these pages as resources to grow the ontology (will discuss in Section **??**.)

3.2 General Description

Figure 2 shows an overview of the BALLET system in which I will conduct my research. As the figure shows, a user can ask a query through the BALLET interface. Using given URLs and indexes it creates as it operates, BALLET can locate proper sources for each user query. BALLET first searches in the stored semantic web pages and checks whether these pages contain the answer to the query. If not, BALLET retrieves related pages from located sources, analyzes each retrieved page, understands the page by mapping source concepts to the concepts in the GEO, and extracts and returns the result to the user BALLET can also transform understood pages into semantic web pages and cache them for future use. In addition, BALLET can enrich the GEO according to understood source pages.

In order to achieve this vision, I need to resolve three main research problems.

- Source Page Understanding
  - How can the system discover source attribute-value pairs?
  - How can the system match the source attribute-value pairs to the concepts in the GEO?
  - In the case of different granularities or different views between source and target concepts, how can we detect complex mappings?
- Source Location through Semantic Indexing
  - What should the system index? Meta-data? Recognized instances? Unrecognized instances?
  - How do we create an index to make it most efficient?
  - How do we maintain the indexes? How should we update, insert, delete according to changeing a source page?
  - Which features in a query are important to the source location problem?
- Ontology Evolution
  - For the classified, but unmatched values in a source file, how can the system decide if it should add them into the lexicons for the ontology?
  - For the unmatched concepts in a source file, how can the system decide if we should add them into the ontology?
  - How can the system resolve the problems caused by different granularities and different versions of the same data?

We discuss each of these problems in the following subsections.

[1] The Gene Ontology is not an extraction ontology. It is a controlled vocabulary. As of September 20, 2004, it contained 1458 components, 7413 functions, and 8907 process terms.

3.2.1 Source Page Understanding

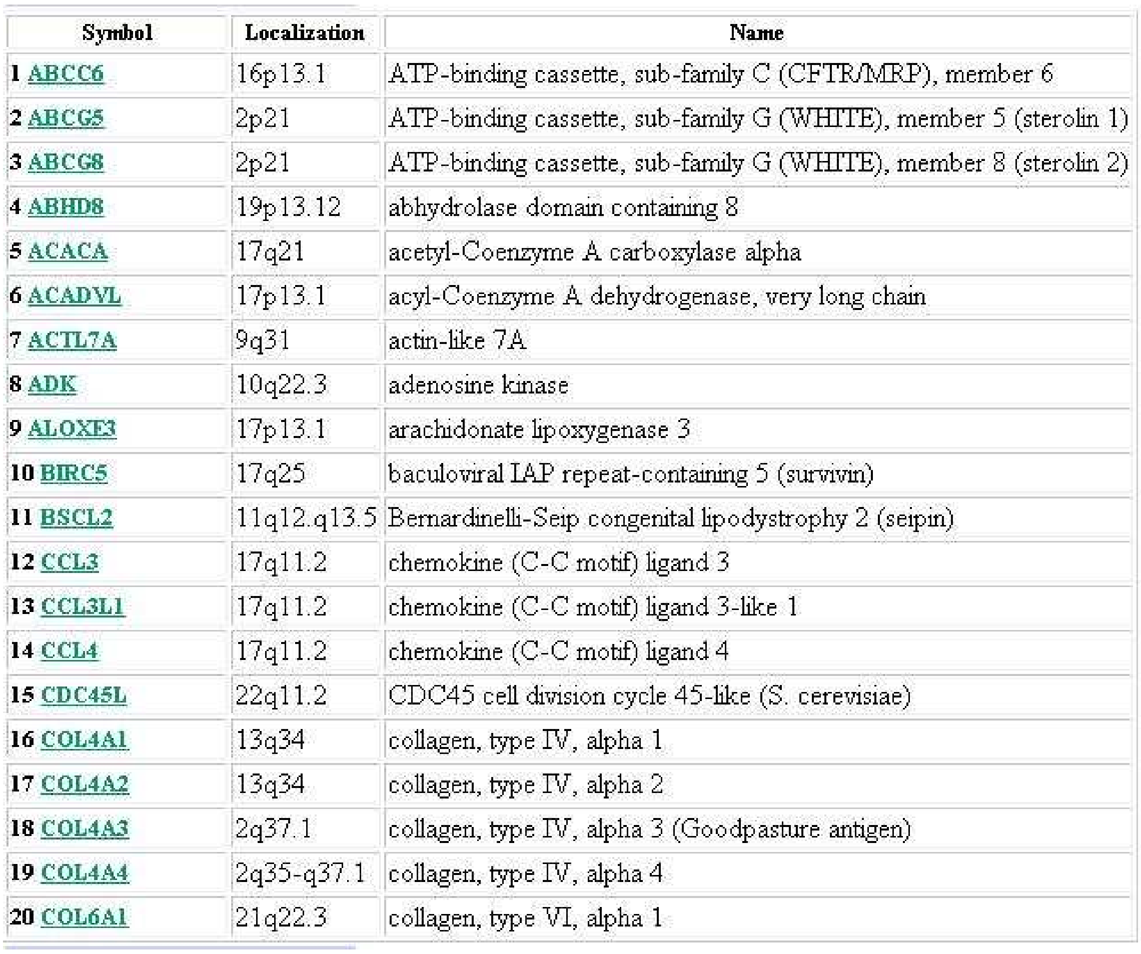
Source page understanding is one of the core components of BALLET. All other components are based on it. Understanding a source page means to recognize attribute-value pairs and to map these attribute-value pairs to the concepts in the gene extraction ontology (GEO). These requirements seem to be easy for a human expert, but they are non-trivial for automatic machine processes.

Traditional source page understanding in terms of table research, such as [PC97], [HD97], and [HKLW01], uses low-level geometric information to recognize attribute-value pairs [LN99b]. [LCC99] uses a lattice of field descriptions to automatically identify fields and facilities date extraction from business reports. Recent research on table understanding for web pages takes this research to a higher level [ETL05]. HTML tags provide helpful information for recognizing table structures, but poor HTML encoding, nontraditional use of HTML tags, and the presence of images all challenge the full exploitation of information contained in tables on the web [Hur01]. Existing approaches to determining the structure of an HTML page use source page pre-analysis [HGMN$^+$97, LKM01], HTML tag parsing [LN99a], and generic ontological knowledge base resolution [YTT01]. This research, however, only focuses on the structure of the page (locating attribute-value pairs), but not on the semantic meaning of the page (mapping the attribute-value pairs to concepts in the ontology). Molecular biology resources usually present data in downloadable relational databases, simple online tables, or semi-structured web pages. For the first two situations, the page understanding problem is relatively easy since we either know the structure already or we can recognize the structure using some techniques we already have. The third situation leads more challenging problems and is what I will focus on my research.

Relational Databases and Simple Tables

Several molecular biology databases are available for free download. Ensembl [ENS05] [Hub02], for example, is freely available to download in several different formats such as Fasta, EMBL, GenBank, and MySQL. If we download the database in the MySQL format, we already have attribute-value pairs, and thus we only need to match them with concepts. In this case, we can use our previous technique (introduced in [ETL02] and [ETL05]) to map source attribute-value pairs to target concepts.

Figure 3 shows a simple example of an online molecular biology table. By “simple”, we mean that the table is tagged by HTML tags, has one attribute row/column, and has a set of value attribute rows/columns. Slight variations, (e.g, multiple appearances of header rows) are also possible. [ETL02] and [ETL05] introduce our previous work on understanding both the structure and semantics of simple tables. This technique does not work with complicated structured data (such as nested tables), which are common in the biology domain. For this dissertation, we plan to incorporate and extend the work we have done and make it apply to more complicated data sources. In particular, we plan to use techniques we call *sibling page comparison* and *seed ontology*



Symbol	Localization	Name
1 ABCC6	16p13.1	ATP-binding cassette, sub-family C (CFTR/MRP), member 6
2 ABCG5	2p21	ATP-binding cassette, sub-family G (WHITE), member 5 (sterolin 1)
3 ABCG8	2p21	ATP-binding cassette, sub-family G (WHITE), member 8 (sterolin 2)
4 ABHD8	19p13.12	abhydrolase domain containing 8
5 ACACA	17q21	acetyl-Coenzyme A carboxylase alpha
6 ACADVL	17p13.1	acyl-Coenzyme A dehydrogenase, very long chain
7 ACTL7A	9q31	actin-like 7A
8 ADK	10q22.3	adenosine kinase
9 ALOXE3	17p13.1	arachidonate lipoxygenase 3
10 BIRC5	17q25	baculoviral IAP repeat-containing 5 (survivin)
11 BSCL2	11q12.q13.5	Bernardinelli-Seip congenital lipodystrophy 2 (seipin)
12 CCL3	17q11.2	chemokine (C-C motif) ligand 3
13 CCL3L1	17q11.2	chemokine (C-C motif) ligand 3-like 1
14 CCL4	17q11.2	chemokine (C-C motif) ligand 4
15 CDC45L	22q11.2	CDC45 cell division cycle 45-like (S. cerevisiae)
16 COL4A1	13q34	collagen, type IV, alpha 1
17 COL4A2	13q34	collagen, type IV, alpha 2
18 COL4A3	2q37.1	collagen, type IV, alpha 3 (Goodpasture antigen)
19 COL4A4	2q35-q37.1	collagen, type IV, alpha 4
20 COL6A1	21q22.3	collagen, type VI, alpha 1

Figure 3: An Example Simple Table from [gen05]

recognition to achieve this goal.

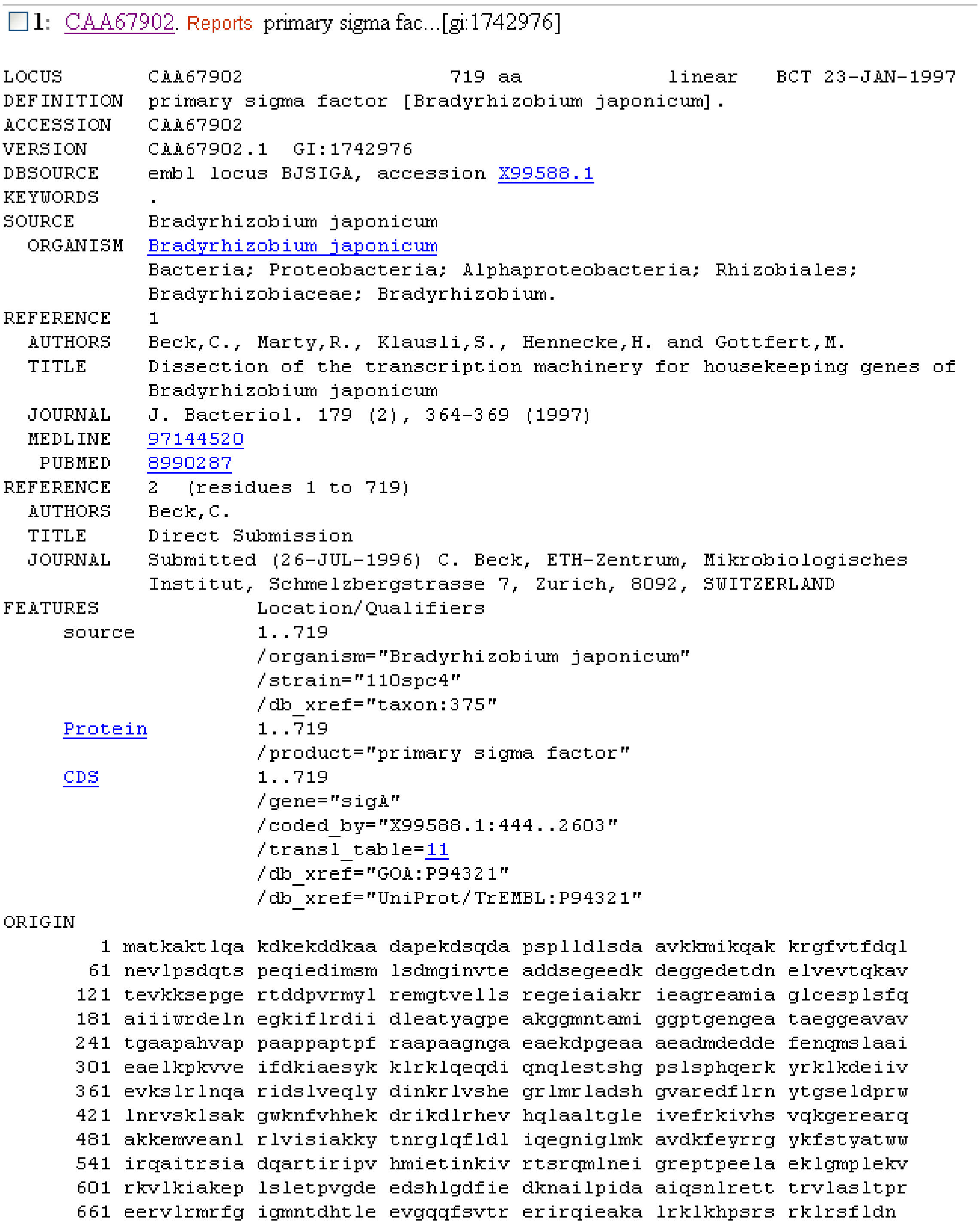
Sibling Page Comparison

Molecular biology web resources usually generate output pages after receiving a user query by placing the results into a predefined page structure. Thus, pages from the same web site are usually structured in the same way. We call pages that are from the same web site and have similar structures *sibling pages.* Figures 4 and 5 show a pair of sibling pages from the National Center for Biotechnology Information (NCBI) [NCB05]. We use them as an example to show how BALLET compares the two pages, searches for similarities, and detects patterns.

There are several levels of structure in pages from this source. BALLET detects that the two sibling pages in Figures 4 and 5 share a similar header row (the first row of each page starts with a check mark, number 1, and some information specific for each page). This is the first-level structure. Under it, there are two columns, and the two sibling pages share identical or similar information for the first column. So the system can assume that these first-column names are second-level attributes (e.g. *LOCUS*, *DEFINITION*, *ACCESSION*, etc.). Some of these attributes have third-level structures such as *source*, *Protein*, and *CDS* under *FEATURES*. We have no guarantee, of course, that two sibling pages always have exactly the same structure. For example, Figure 4 has two references while Figure 5 has only one reference and only one of the three references has a *MEDLINE* entry. Nevertheless, this information is good enough for us to detect the structure pattern we need to determine attribute-value pairs.

After the system detects the positions of attributes in a source page, it tries to find structured patterns for attribute-values pairs. For example, if the system recognizes that most of the cells in one column/row contain attributes (such as the first column in our example), and the cells of the column/row next to it (such as the second column in our example) do not contain any attributes or only contains a few attributes, we can assume that the first column/row is for attributes and the second column/row is for values. We can thus pair a value together with its attribute. Further, if the page has a table-like structure, we can pair all values under each attribute in a row or column with the attribute. We use several heuristics to detect attribute-value-pair structure patterns according to the attribute positions we find using the sibling page comparison technique.

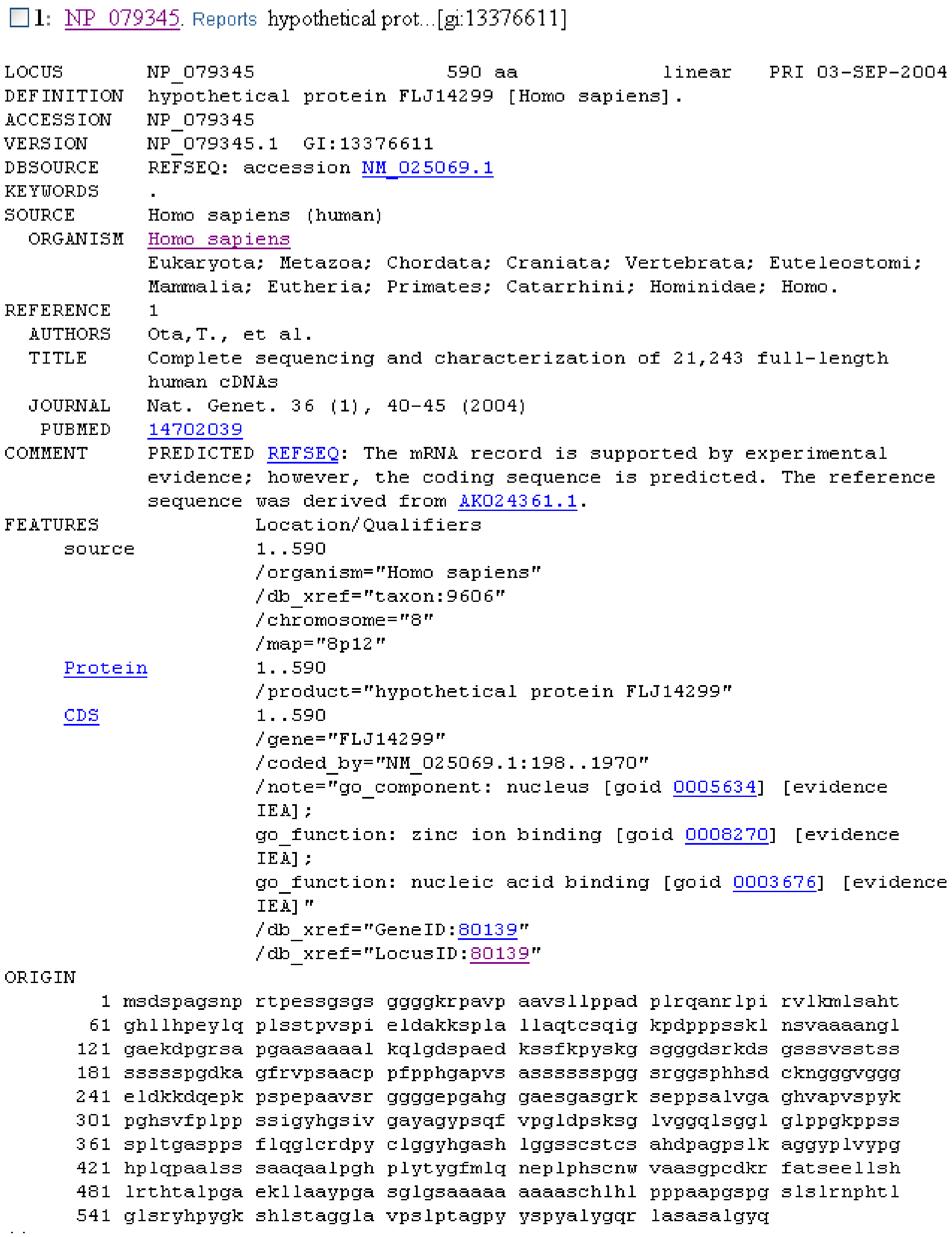
After we detect the structure of a page, we also need to understand the semantics of the page. Given that the pattern of attribute-value pairs has been recognized, BALLET can collect values on several sibling pages for the same concept (same attribute), and see if the GEO can recognize the values or the attribute. If it can recognize enough of them and especially, also, if it can recognize the attribute, we can be confident that the system has found a correct mapping. Since the values associated with one attribute from all sibling pages should map to the same target attribute (or attributes, in the case of a one-many mapping), we can classify all the values as belonging to one concept, even those the system does not recognize.



□ 1: CAA67902. Reports primary sigma fac...[gi:1742976]

```
LOCUS       CAA67902                 719 aa            linear   BCT 23-JAN-1997
DEFINITION  primary sigma factor [Bradyrhizobium japonicum].
ACCESSION   CAA67902
VERSION     CAA67902.1  GI:1742976
DBSOURCE    embl locus BJSIGA, accession X99588.1
KEYWORDS    .
SOURCE      Bradyrhizobium japonicum
  ORGANISM  Bradyrhizobium japonicum
            Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales;
            Bradyrhizobiaceae; Bradyrhizobium.
REFERENCE   1
  AUTHORS   Beck,C., Marty,R., Klausli,S., Hennecke,H. and Gottfert,M.
  TITLE     Dissection of the transcription machinery for housekeeping genes of
            Bradyrhizobium japonicum
  JOURNAL   J. Bacteriol. 179 (2), 364-369 (1997)
  MEDLINE   97144520
   PUBMED   8990287
REFERENCE   2  (residues 1 to 719)
  AUTHORS   Beck,C.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-JUL-1996) C. Beck, ETH-Zentrum, Mikrobiologisches
            Institut, Schmelzbergstrasse 7, Zurich, 8092, SWITZERLAND
FEATURES             Location/Qualifiers
     source          1..719
                     /organism="Bradyrhizobium japonicum"
                     /strain="110spc4"
                     /db_xref="taxon:375"
     Protein         1..719
                     /product="primary sigma factor"
     CDS             1..719
                     /gene="sigA"
                     /coded_by="X99588.1:444..2603"
                     /transl_table=11
                     /db_xref="GOA:P94321"
                     /db_xref="UniProt/TrEMBL:P94321"
ORIGIN
        1 matkaktlqa kdkekddkaa dapekdsqda psplldlsda avkkmikqak krgfvtfdql
       61 nevlpsdqts peqiedimsm lsdmginvte addsegeedk deggedetdn elvevtqkav
      121 tevkksepge rtddpvrmyl remgtvells regeiaiakr ieagreamia glcesplsfq
      181 aiiiwrdeln egkiflrdii dleatyagpe akggmntami ggptgengea taeggeavav
      241 tgaapahvap paappaptpf raapaagnga eaekdpgeaa aeadmdedde fenqmslaai
      301 eaelkpkvve ifdkiaesyk klrklqeqdi qnqlestshg pslsphqerk yrklkdeiiv
      361 evkslrlnqa ridslveqly dinkrlvshe grlmrladsh gvaredflrn ytgseldprw
      421 lnrvsklsak gwknfvhhek drikdlrhev hqlaaltgle ivefrkivhs vqkgerearq
      481 akkemveanl rlvisiakky tnrglqfldl iqegniglmk avdkfeyrrg ykfstyatww
      541 irqaitrsia dqartiripv hmietinkiv rtsrqmlnei greptpeela eklgmplekv
      601 rkvlkiakep lsletpvgde edshlgdfie dknailpida aiqsnlrett trvlasltpr
      661 eervlrmrfg igmntdhtle evgqqfsvtr erirqieaka lrklkhpsrs rklrsfldn
```

Figure 4: Sample Sibling Page 1 from NCBI [NCB05]



□ 1: NP_079345. Reports hypothetical prot...[gi:13376611]

```
LOCUS       NP_079345                590 aa            linear   PRI 03-SEP-2004
DEFINITION  hypothetical protein FLJ14299 [Homo sapiens].
ACCESSION   NP_079345
VERSION     NP_079345.1  GI:13376611
DBSOURCE    REFSEQ: accession NM_025069.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1
  AUTHORS   Ota,T., et al.
  TITLE     Complete sequencing and characterization of 21,243 full-length
            human cDNAs
  JOURNAL   Nat. Genet. 36 (1), 40-45 (2004)
   PUBMED   14702039
COMMENT     PREDICTED REFSEQ: The mRNA record is supported by experimental
            evidence; however, the coding sequence is predicted. The reference
            sequence was derived from AK024361.1.
FEATURES             Location/Qualifiers
     source          1..590
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="8"
                     /map="8p12"
     Protein         1..590
                     /product="hypothetical protein FLJ14299"
     CDS             1..590
                     /gene="FLJ14299"
                     /coded_by="NM_025069.1:198..1970"
                     /note="go_component: nucleus [goid 0005634] [evidence
                     IEA];
                     go_function: zinc ion binding [goid 0008270] [evidence
                     IEA];
                     go_function: nucleic acid binding [goid 0003676] [evidence
                     IEA]"
                     /db_xref="GeneID:80139"
                     /db_xref="LocusID:80139"
ORIGIN
        1 msdspagsnp rtpessgsgs ggggkrpavp aavsllppad plrqanrlpi rvlkmlsaht
       61 ghllhpeylq plsstpvspi eldakkspla llaqtcsqig kpdpppsskl nsvaaaangl
      121 gaekdpgrsa pgaasaaaal kqlgdspaed kssfkpyskg sgggdsrkds gsssvsstss
      181 sssspgdka gfrvpsaacp pfpphgapvs asssssspgg srggsphhsd ckngggvggg
      241 eldkkdqepk pspepaavsr ggggepgahg gaesgasgrk seppsalvga ghvapvspyk
      301 pghsvfplpp ssigyhgsiv gayagypsqf vpgldpsksg lvggqlsggl glppgkppss
      361 spltgaspps flqglcrdpy clggyhgash lggsscstcs ahdpagpslk aggyplvypg
      421 hplqpaalss saaqaalpgh plytygfmlq neplphscnw vaasgpcdkr fatseellsh
      481 lrthtalpga ekllaaypga sglgsaaaaa aaaaschlhl pppaapgspg slslrnphtl
      541 glsryhpygk shlstaggla vpslptagpy yspyalygqr lasasalgyq
```

Figure 5: Sample Sibling Page 2 from NCBI [NCB05]

Seed Ontology Recognition

An alternate way to discover attribute-value pairs and map them to concepts in the ontology is through the use of a seed ontology. A *seed ontology* contains as much information as we can collect for one object in a specified application domain with respect to the extraction ontology, e.g. all the information about one protein or gene. For a seed ontology to be useful, it must commonly appear in many sites. We therefore select objects we expect to find in many repositories. Figure 6 shows a partial seed ontology for the gene named *FLJ14299* for our GEO. The shaded boxes contain the values associated with concepts for *FLJ14299*.

Instead of attributes, our seed ontology technique depends on values to detect structural patterns. If the system can find a page that matches a seed ontology, it can be confident about where many of the values are on the page, and it can be confident about the classification of those values since the classification is part of the given seed ontology. From this information, the system can try to infer the structure pattern of a page by observing the layout of the page with respect to the seed ontology.

Using Figure 5 as an example, we can see that by matching the seed ontology in Figure 6 with the page, the system recognizes many values in the right column, but no value in the left column. The system can thus assume that the left column is for attributes and the right column is for values. It can then attempt to match the attributes with the attributes it expects, given the values it has found. By way of comparison with our sibling-page technique, instead of looking for values to the right or bottom, our recognition technique using a seed ontology searches for attributes to the left and top and then tries to pair attributes and values together according to the detected pattern.

After the system recognizes attribute-value pairs, the next step is to map these attribute-value pairs to the concepts in the GEO. Since the seed ontology recognizes values in a source page, the system already "knows" the target concepts to which these values belong. Thus, since we have already detected attribute-value pairs, we can map these attribute-value pairs to the target GEO concepts easily. For example, the seed ontology recognizes "Hypothetical protein FLJ14299" as *Protein Name* and "Homo sapiens (Human)" as *Source Species/Organism.* The system then can map the attribute-value pair "DEFINITION: Hypothetical protein FLJ14299 [Homo sapiens]" via a one-many mapping to the target concepts *Protein Name*, and *Source Species/Organism.*

Although the two techniques—sibling pages and seed ontologies—can work independently, we can also combine them. The sibling page comparison technique focuses on attributes, and the seed ontology recognition technique focuses on values. Combining them together, when they both yield the same result, we can have even greater confidence that the system has properly detected the right structure pattern. Each technique has its own advantages and disadvantages. Using them together, we are likely to obtain better mapping results.

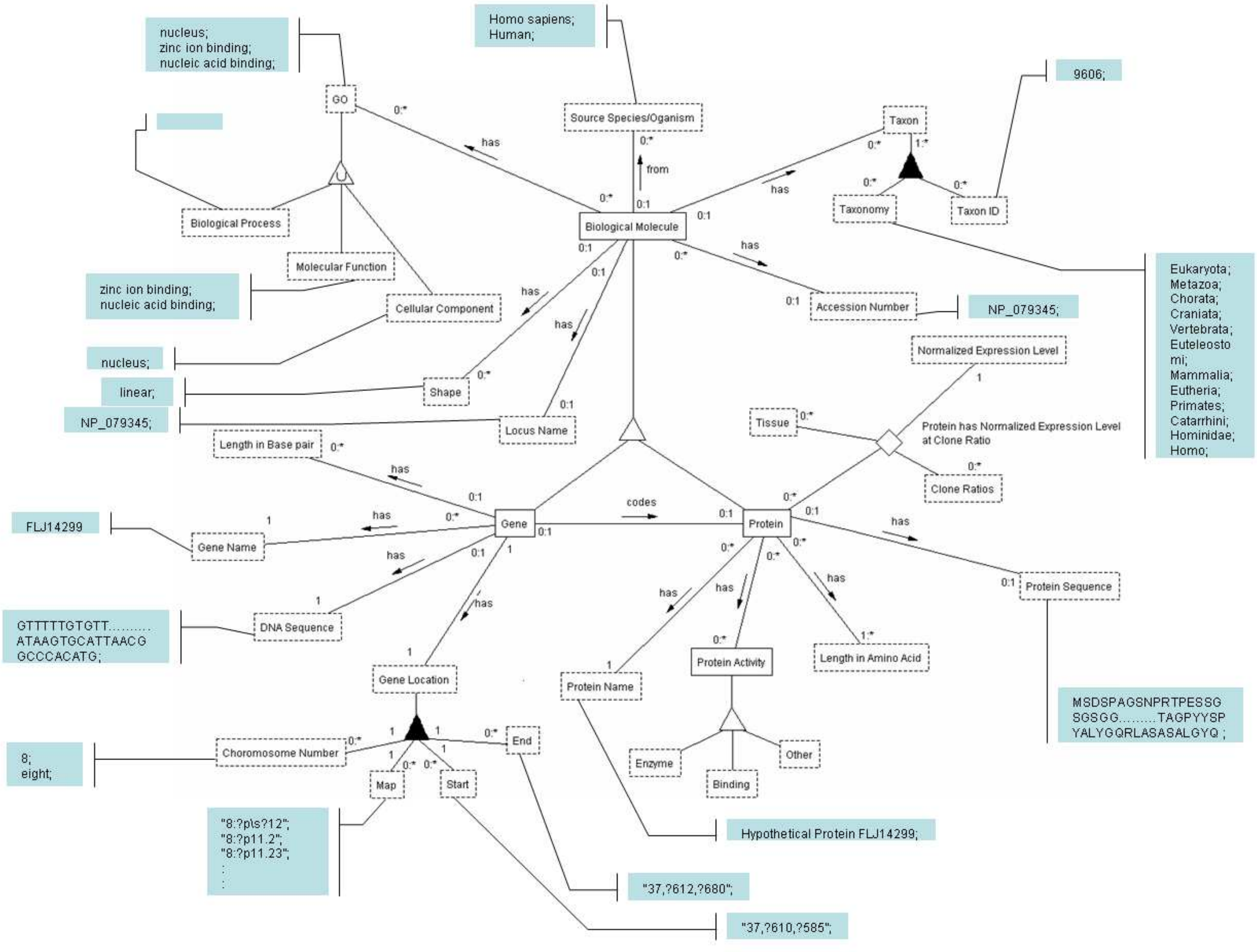


Figure 6: Example Seed Ontology for Gene FLJ14299 (Partial)

3.2.2 Source Location through Semantic Indexing

As mentioned in the introduction, there are hundreds of available repositories of biological data that are valuable to the biological community. The biology domain is broad. One source repository, however, usually only focuses on a narrow sub-domain. Therefore, it is not likely that all repositories contain data to answer a specific user query, and thus query answering can be more efficient if we can determine which sources to ignore and which are needed for a particular query. In this section, I will discuss how I plan to locate proper sources according to a given query.

In my research, I plan to collect and keep a list of source repositories, and then to index each source repository automatically. The indexing can be done either at the meta data level or at the data level. Then, given a query, the system can try to locate the proper resources according to these indexes.

For meta-data indexing, we can sample enough pages from each source until we are statistically confident that we have covered most of the meta data for the source. For each page in the sample set, we pre-run the source-page understanding process described in Section 3.2.1. As a result, since we obtain a list of source-to-target concept mappings, each source repository has a list of target concepts associated with it, and each target concept has a list of source repositories associated with it. When BALLET receives a user's query, it lists all the concepts in the query. It then only tries to retrieve information from those source repositories that contain concepts included in the list.

To illustrate how the indexing works, suppose a biologist is interested in the human gene called "FLJ14299," wanting to know the protein sequence and amino acid length for the gene and wanting to know its normalized expression level on stomach tissue. The target concepts mentioned in this query include: *Gene Name, Source Species/Organism, Protein Sequence, Length in Amino Acids, Normalized Expression Level*, and *Tissue*. As we can see, Figures 4 and 5, which are from the NCBI site [NCB05], contain the concepts *Gene Name, Protein Sequence*, and *Length in Amino Acid*. Thus, these target concepts have the resource NCBI associated with each of them. When the system encounters this query, the source NCBI is among those we locate. In another situation, if a biologist only wants to find the DNA sequences for all genes that have a normalized expression level greater than 5% on stomach tissue, the query concepts are *DNA Sequence* and *Tissue*. In this case, the system can ignore the NCBI site (of which the pages in Figure 4 and 5 are samples) because the pages in this site cannot provide answers to this query.

It is also possible to index source documents by each value. If we can understand a source page and have already matched values in the source page to concepts in the ontology, it is not hard to index the source. Alternatively, it is also easy to semantically annotate values for each page in the site using the GEO as the annotation ontology since the machine has already "understood" it. This means that we can transform a source page to a semantic web page, which is machine-understandable [BLHL01]. There are few well-known ontology and knowledge representation languages aiming at semantic interoperability such as RDF (Resource Description Framework) [RDF03], DAML+OIL (The DARPA Agent Markup Language + Ontology Inference Layer) [DAM01], and OWL (Web Ontology Language) [OWL04]. Exactly how we should represent the annotation is not clear since the semantic web community itself has not decided, but since our system has all the information it needs once it has "understood" a page, it can generate the annotation in any representation deemed desirable.

Although the system can, in principle, index every recognized value in every page, we must recognize that it is not easy to index all source pages from all resources. [LDEY02] introduces a method (1) to retrieve all the pages behind a site form, when this is possible, (2) to statistically retrieve most of the data, when retrieving all data is too time consuming, and (3) to know when it is not possible to automatically retrieve any of the data. For biological sites, we expect it to be difficult to retrieve the data behind forms. Forms found in biological sites usually do not allow an empty query that results in retrieving all data. Instead, they typically require a user to type in a word or phrase to retrieve a source page. Some of these words or phrases may be found in our lexicons and seed ontologies, but it is highly unlikely that we will have all the words and phrases needed to obtain all the pages from a site. Therefore, we only plan to index pages at the data level as we see them. This may not be helpful for the source location problem in the beginning. As we continue, however, we will index increasingly more pages and have more semantic web pages. We can search for these pages first when we have a query, or exclude them altogether if the index tells us they have no data for the query.

Most biology web sources are unstable. Independent developers modify their designs and schemas and remove or add data dynamically. How to dynamically update the indexes is still an open question. Since BALLET generates a semantic index automatically, we can always re-run the procedure to update the index. This process, however, is time consuming, and it is not clear how often we should update the index.

3.2.3 Ontology Enrichment

After we have mapped a source concept or a set of source concepts to a concept in the ontology, we can use these concepts as a source for ontology enrichment. There are two ways to enrich an ontology: (1) enrich the data frames and (2) enrich the object sets and relationship sets.

Data Frame Enrichment

A data frame for a lexical object set defines the lexical appearance of constant objects for the object set and establishes appropriate keywords that are likely to appear in a document when objects in the object set are mentioned. A data frame for a lexical object set either contains a set of regular expressions that describes the lexical patterns for the concept, or a lexicon file, listing concept values. It is not easy to update regular expressions for a concept automatically; however, it is straightforward to add new values to a lexicon.

Lexicons are common in biology. In the GEO in Figure 1, we would use lexicons to describe concepts such as *Gene Name, Protein Name*, and *Protein Activity.* It is impractical, and unnecessary, to include all values for these concepts in the lexicon file. We can, however, collect as many values as possible for each concept. Since BALLET can understand structured/semi-structured source documents and infer mappings from source concepts to GEO object sets, we can enrich our ontology by adding all the values we find in the various site pages we process to the data frame lexicons. For example, assume that the lexicon file for *Protein Name* in the GEO does not contain "14-3-3 protein beta/alpha" which appears in Figure 4, but we know that the source attribute *Protein name* maps to the target concept *Protein Name.* Therefore, we know that "14-3-3 protein beta/alpha" is also a protein name, and we can add it to the lexicon file for the target concept *Protein Name.* If later we encounter this phrase in other files, the ontology can recognize it directly.

Object Set and Relationship Set Enrichment

In addition to data frame enrichment, it is also possible to enrich the set of object sets and relationship sets in the GEO. As already mentioned, different data sources use different structures, granularities, or semantics. As a consequence, there exist complex mappings such as 1-n, n-1, and n-m mappings. As indicated in an earlier example about protein names and source Species/Organism, the system can detect these complex mappings. As another example, Figure 7 shows two different ways to describe the location of the gene "FLJ14299." If we want to map the first to the second, we have an n-1 mapping; and if we want to map the second to the first, we have a 1-n mapping. As special kinds of complex mappings, we can recognize some of them as union/selection or merge/split mappings. The example in Figure 7 is an example of union/selection, and the example in the seed-ontology discussion in Section 3.2.1 is an example of merge/split.

Union/selection and merge/split mappings correspond respectively to generalization/specialization or aggregation in an ontology, and these are the only kinds of complex mapping I intend to investigate in my research. When the system discovers a merge/split of source values during extraction, it either adds an aggregation for recognized components or adds components for a recognized aggregation. When the system discovers a union/selection of source values during extraction, it appropriately adds either the missing generalization or the missing specialization. For example, assume that the upper table in Figure 7 shows the target view and the lower table in Figure 7 shows the source view. Here we have a 1-n mapping, and the values in the source need to be placed together in the same set in the target view. In this case, the system adds a generalization/specialization to the target ontology. Figure 8 shows of how to enrich the ontology.

Since the ontology enrichment process may be error-prone, we intend to have this part of the system be interactive. The system can find and suggest possibilities as we have discussed, but the system also lets users or human experts make the final decision.

Biological Process	**Molecular Function**	**Cellular Component**
calcium-mediated signaling	nucleus	transcription factor activity
circulation		
signal transduction		
central nervous system development		

GO Terms
calcium-mediated signaling
circulation
signal transduction
central nervous system development
nucleus
transcription factor activity

Figure 7: Example of Information in Different Granularities

We treat each source repository as a knowledge base. When the system detects a new concept, even if we do not know how to describe the relationships and constraints for this concept, or we decide not to include this concept to the GEO, we try to build a data frame for this concept depends on the source repository. By doing so, we can build a data frame library in the molecular biology domain.

3.3 Research Plan

BALLET has three components. Although their work is interdependent with respect to the overall vision of BALLET, each component, by itself, can make a contribution to the field of information technology and bioinformatics. Here I explain how I will evaluate the system.

- ***Source Page Understanding.***

  We will run several experiments to determine the effectiveness of this component. First, we need to choose training data and test data for the evaluation. We would like to select web sites that represent a rich set of cases, so that we can evaluate whether the approach is reliable and has high performance over the whole domain. I plan to use the Molecular Biology Database Collection (MBDC) as a repository. This collection has over 700 databases as of November, 2004. We plan to use a statistical sampling method to select training/test sites.

  Because this is a proof-of-concept research, the gene extraction ontology will not contain all concepts that appear in the whole collection of databases. Thus only a subset of the 700 databases may apply. I will pre-select the databases that fit our GEO using some document classify methods such as VSM [SAY75] and onto-based VSM [ENX01]. I will then choose the

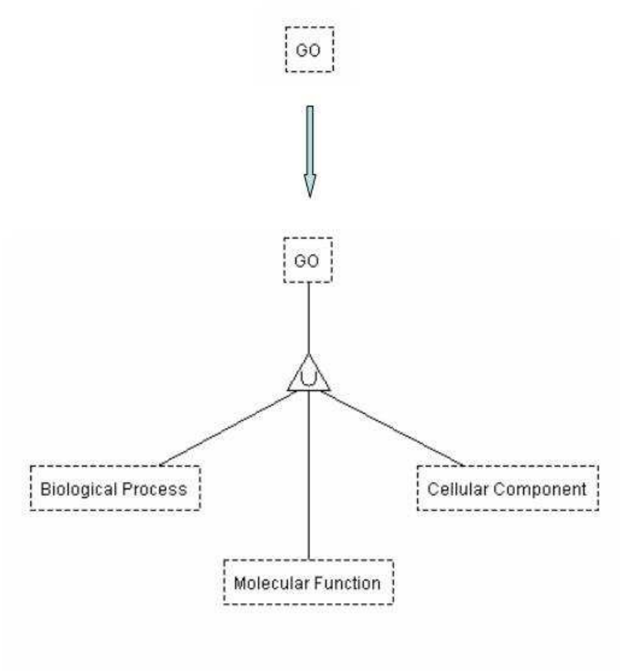


Figure 8: Ontology Enrichment according to Figure 7

training set [2] and test set from this pre-selected list.

I will establish page understanding heuristics based on the training set. For example, I will determine how to choose thresholds for attribute column/row heuristics, how to set up rules for recognizing attribute-value pair patterns, and how to combine results of different pairwise comparisons. I plan to use the training set to refine the seed ontologies and then establish rules to do page recognition based on seed ontologies. I also want to use this training set to establish rules about how combine these two technologies. To choose the training set, I will begin with a small number of training databases, and use them to train the system. I will test the trained system using some new databases and see if the result is acceptable. If not, I will add more training data and re-train the system.

For choosing the test set, I plan to use some statistical method to make sure that the test set covers the cases and can provide statistically significant. The page understanding component has three steps: (1) detect attributes and values; (2) pair attribute-value pairs, and (3) recognize mappings between source attribute-value pairs to target concepts. We will test them one by one.

[2] The "training set" is a set of examples to be used to fine-tune the source the source-page-understanding procedures (not a training set in machine learning

For the attribute or value detection, I plan to do the following tests. Of all the attribute or attributes or values in the source test pages, how many of them does the system recognize correctly? What are the accuracy, precision, and recall values? The evaluation for the attribute-value pair recognition is similar, except that I will test attribute-value pairs instead of attribute and value individually.

For the mapping recognition process that maps the recognized attribute on a source page to the GEO, I will test how the system maps source-page attributes to ontology concepts. I will count how many concepts the system can recognize completely correctly, which means that for one source attribute, the system recognizes all the ontology concepts to which the source attribute should map. I will also count how many concepts the system can recognize partially correctly, how many concepts are missed, and how many concepts are mapped incorrectly. Based on these numbers, I will then calculate precision and recall values.

- ***Source Location through Semantic Indexing.***

  The results for semantic indexing depend on the results of source page understanding. So I will not test for accuracy. The efficiency of using semantic indexing is obvious. It can prune away unrelated URLs to a query or even find the answer of a query in the cache. Therefore, no test will be done for this component.

- ***Ontology Enrichment.***

  The performance of ontology enrichment on the value level depends on the performance of the source page understanding component. So there will not be a separate test for it.

  For ontology concept/structure enrichment, I plan to implement an interactive interface so that a user can decide if a system-proposed new concept should be added or if a system-proposed change to current concept/relationship set should be made. The user would also have the choice to declare constraints of new relationship sets or to accept the default constraints the system provides. Unfortunately, it will be difficult (probably impossible) to conduct controlled experiments because ontology evolution depends on finding actual cases in real data that can cause evolution. These appear arbitrarily, and it is not clear how many will be found — likely, only a few. I do, however, plan to document the situations I see and list the cases the system can handle and those the system cannot handle.

3.4 Artifacts to be Produced

I plan to implement a demo based on the data extraction and annotation tools provided by other people in our group. This demo allows a user to submit queries in the ontology schema of the GEO and then locate needed resources, understand the sources, and extract information from different sources. The system will return the results back to the user in terms of the ontology. I also plan to implement a tool that can automatically transform an understood page to a semantic web page. And I also plan to implement a tool to update an ontology semi-automatically given a set of understood pages.

3.5 Limitations of the Dissertation

The prototype system is to be built for research purposes. It will use an available front-end query interface and will not do any integration beyond synchronization with the target gene extraction ontology. The gene extraction ontology will not cover all the concepts, relationships, and values in the molecular biology domain. Its aim is only for testing whether the research ideas can work well. The resources I will use in my research are only those generated based on online databases. I will not use hand-crafted semi-structured pages or unstructured sources such as literature. The data frame enrichment will not do automatic regular expression enrichment. The object set and relationship set enrichment will be limited to enriching ISA and Part-Of hierarchies and, possibly, simple attribute additions.

4 Research Papers

- Automatically Understanding Web Documents by Sibling Page Comparison
- Automatically Understanding Biological Web Documents
- Automatic Biological Ontology Enrichment
- Semantic Indexing for Molecular Biology Repositories
- BALLET: Biological dAta onLine Location and Extraction Tool

5 Contribution to Computer Science

This dissertation will contribute in both information technology and bioinformatics. With the overwhelming amount of biological date available online, biologists need a tool that can automatically locate, understand, and extract information of interest. Our prototype uniquely address these issues. The system will also transform understood source pages into semantic web pages. Although I will implement and test the system in the molecular biology domain, this approach will likely be general to all application domains that have similar characteristics—a strong, coherent domain ontology with no more than a few dozen concepts and plentiful structured or semi-structured data pages retrieved from databases hidden behind web forms.

6 Dissertation Schedule

- Create Ontology — 08/2005

- Source Page Understanding
  - Generate seed ontologies — 10/2005
  - Detect attribute-value pairs — 12/2005
  - Schema Matching — 04/2006
- Semantic Web Page Generation — 07/2006
- Source Location through Semantic Indexing — 09/2006
- Ontology Enrichment — 10/2006
- Evaluation and Writing — 05/2007

References

[BGM04] Biology/genetics/microbiology databases. http://www.edae.gr/bio-databases.html, 2004.

[BLHL01] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, pages 28–37, May 2001.

[CAT04] Cath database. http://www.biochem.ucl.ac.uk/bsm/cath/index.html, 2004.

[DAM01] W3C Annotated DAML+OIL Ontology Markup. http://www.w3.org/TR/daml+oil-walkthru/, 2001.

[DBC04] The public catalog of databases. http://www.infobiogen.fr/services/dbcat/, 2004.

[DR04] H. Do and E. Rahm. Flexible integration of molecular-biological annotation data: The genmapper approach. In *9th International Conference on Extending Database Technology (EDBT 04)*, pages 811–822, Crete, Greece, March 2004.

[EA96] T. Etzold and P. Argos. SRS: Information retrieval system for molecular biology data banks. *Methods Enzymol*, 266:114–128, 1996.

[ENS05] ENSEMBL. http://www.ensembl.org/Download/, 2005.

[ENX01] D. W. Embley, Y.-K. Ng, and L. Xu. Recognizing ontology-applicable multiple-record web documents. In *Proceedings of the 20th International Conference on Conceptual Modeling (ER '01)*, pages 555–570, Yokohama, Japan, November 2001. Springer-Verlag.

[ETL02] D.W. Embley, C. Tao, and S.W. Liddle. Automatically extracting ontologically specified data from HTML tables with unknown structure. In *Proceedings of the 21st International Conference on Conceptual Modeling (ER'02)*, pages 322–327, Tampere, Finland, October 2002.

[ETL05] D.W. Embley, C. Tao, and S.W. Liddle. Automating the extraction of data from html tables with unknown structure. *Data and Knowledge Engineering*, 2005. To Appear.

[Gal05] M. Y. Galperin. The molecular biology database collection: 2005 update. *Nucleic Acids Research*, 33:5–24, 2005.

[Gen04] GenoMax. http://www.informaxinc.com/solutions/genomax, 2004.

[gen05] GENATLAS. http://www.dsi.univ-paris5.fr/genatlas/, 2005.

[GO04] Gene ontology (go) consortium. http://www.geneontology.org/, 2004.

[Haa01] L.M. Haas. DiscoveryLink: A system for integrated access to life sciences data sources. *IBM Systems Journal*, 40(2):489–511, 2001.

[Har04] Embl (european molecular biology laboratory) bioinformatic harvester. http://harvester.embl.de/, may, 2004.

[HD97] M. Hurst and S. Douglas. Layout and language: Preliminary investigations in recognizing the structure of tables. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'97)*, pages 1043–1047, Ulm, Germany, August 1997.

[HGMN+97] J. Hammer, H. Garcia-Molina, S. Nestorov, R. Yerneni, M. Breunig, and V. Vassalos. Template-based wrappers in the TSIMMIS system. In *Proceedings of 1997 ACM SIGMOD International Conference on Management of Data*, pages 532–535, Tucson, Arizona, May 1997.

[HKLW01] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong. Table structure recognition and its evaluation. In P.B. Kantor, D.P. Lopresti, and J. Zhou, editors, *Proceedings of Document Recognition and Retrieval VIII*, volume SPIE-4307, pages 44–55, San Jose, California, January 2001.

[Hub02] T. Hubbard. The ensembl genome database project. *Nucleic Acids Research*, 30(1):38–41, 2002.

[Hur01] M. Hurst. Layout and language: challenges for table understanding on the web. In *Proceedings of the First International Workshop on Web Document Analysis (WDA2001)*, pages 27–30, Seattle, Washington, September 2001.

[LBE03] Z. Lacroix, O. Boucelma, and M. Essid. The biological integration system. In *The 5th ACM international workshop on Web information and data management (WIDM 03)*, pages 45–49, New Orleans, LA, November 2003. ACM Press.

[LCC99] S. W. Liddle, D. M. Campbell, and C. Crawford. Automatically extracting structure and data from business reports. In *Proceedings of the eighth international conference on Information and knowledge management (CIKM '99)*, pages 86–93, Kansas City, Missouri, November 1999. ACM Press.

[LDEY02] S.W. Liddle, D.T. Scott D.W. Embley, and S.H. Yau. Extracting data behind web forms. In *Proceedings of the Joint Workshop on Conceptual Modeling Approaches for E-business: A Web Service Perspective (eCOMO 2002), Lecture Notes in Computer Science (LNCS 2784)*, pages 402–413, Tampere, Finland, October 2002.

[LKM01] K. Lerman, C.A. Knoblock, and S. Minton. Automatic data extraction from lists and tables in web source. In *Proceedings of Automatic Text Extraction and Mining Workshop (ATEM-01)*, Seattle, Washington, August 2001.

[LN99a] S. Lim and Y. Ng. An automated approach for retrieving heirarchical data from HTML tables. In *Proceedings of the Eighth International Conference on Informaiton and Knowledge management (CIKM'99)*, pages 466–474, Kansas City, Missouri, November 1999.

[LN99b] D. Lopresti and G. Nagy. Automated table processing: An (opinionated) survey. In *Proceedings of the Third IAPR Workshop on Graphics Recognition*, pages 109–134, Jaipur, India, September 1999.

[MBD05] The molecular biology database collection. http://www3.oup.co.uk/nar/database/, 2005.

[MWL03] Z. B. Miled, Y. W. Webster, and Y. Liu. An ontology for semantic integration of life science web databases. *International Journal of Cooperative Information Systems*, 12(2):275–294, 2003.

[MWN$^+$02] Z. B. Miled, Y. W. Webster, L. Nianhua, O. Bukhres, A. K. Nayar, J. Martin, and R. Oppelt. BAO, a biological and chemical ontology for information integration. *Online Journal of Bioinformatics*, 1:60–73, 2002.

[NCB05] Ncbi protein sequence. http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein, 2005.

[NO95] C. Naiman and A.M. Ouksel. A classification of semantic conflicts. *Journal of Organizational Computing*, 5(2):167–193, 1995.

[OWL04] W3C Web Ontology Language. http://www.w3.org/2004/OWL, 2004.

[PC97] P. Pyreddy and W.B. Croft. TINTIN: A system for retrieval in text tables. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 193–200, Philadelphia, Pennsylvania, July 1997.

[PDB04] Protein data bank. ftp://ftp.rcsb.org/pub/pdb/, 2004.

[Phi04] S. Philippi. Light-weight integration of molecular biological databases. *Bioinformatics*, 20(1):51–57, 2004.

[RDF03] W3C Resource Description Framework (RDF). http://www.w3.org/RDF/, 2003.

[SAY75] G. Salton, A.Wong, and C. S. Yang. A vector space model for information retrieval. 18(11):613620, November 1975.

[SBB$^+$00] R. Stevens, P. Baker, S. Bechhofer, G. Ng, A. Jacoby, N. W. Paton, C. A. Goble, and A. Brass. TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformatics*, 16(2):184–185, 2000.

[SCO04] Scop database. http://scop.mrc-lmb.cam.ac.uk/scop/, 2004.

[STK04] The species tool kit. http://www.speciestoolkit.org/, 2004.

[YTT01] M. Yoshida, K. Torisawa, and J. Tsujii. A method to integrate tables of the world wide web. In *Proceedings of the International Workshop on Web Document Analysis (WDA 2001)*, pages 31–34, Washington, DC, September 2001.

This dissertation proposal by Cui Tao is accepted in its present form by the Department of Computer Science of Brigham Young University as satisfying the dissertation proposal requirement in the Department of Computer Science for the degree of Doctor of Philosophy.

David W. Embley, Committee Chair

Stephen W. Liddle, Committee Member

Deryle W. Lonsdale, Committee Member

Dan R. Olsen, Committee Member

Kevin D. Seppi, Committee Member

David W. Embley, Graduate Coordinator