1. Ontologies in bioinformatics

Robert Stevens, Chris Wroe, Phillip Lord and Carole Goble

Department of Computer Science, University of Manchester, Oxford Road, Manchester UK, M13 9PL. email: robert.stevens|cwroe|carole@cs.man.ac.uk

Summary

Molecular biology offers a large, complex and volatile domain that tests knowledge representation techniques to the limit of their fidelity, precision, expressivity and adaptability. The discipline of molecular biology and bioinformatics relies greatly on the use of community knowledge, rather than laws and axioms, to further understanding, and knowledge generation. This knowledge has traditionally been kept as natural language. Given the exponential growth of already large quantities of data and associated knowledge, this is an unsustainable form of representation. This knowledge needs to be stored in a computationally amenable form and ontologies offer a mechanism for creating a shared understanding of a community for both humans and computers. Ontologies have been built and used for many domains and this chapter explores their role within bioinformatics. Structured classifications have a long history in biology; not least in the Linnean description of species. The explicit use of ontologies, however, is more recent. This chapter provides a survey of the need for ontologies; the nature of the domain and the knowledge tasks involved; and then an overview of ontology work in the discipline. The widest use of ontologies within biology is for conceptual annotation – a representation of stored knowledge more computationally amenable than natural language. An ontology also offers a means to create the illusion of a common query interface over diverse, distributed information sources – here an ontology creates a shared understanding for the user and also a means to computationally reconcile heterogeneities between the resources. Ontologies also provide a means for a schema definition suitable for the complexity and precision required for biology's knowledge bases. Coming right up to date, bioinformatics is well set as an exemplar of the Semantic Web, offering both web accessible content and services conceptually marked up as a means for computational exploitation of its resources – this theme is explored through the myGRID services ontology. Ontologies in bioinformatics cover a wide range of usages and representation styles. Bioinformatics offers an exciting application area in which the community can see a real need for ontology based technology to work and deliver its promise.

1.1 Introduction

This chapter gives an overview of the application of ontologies within bioinformatics. Bioinformatics is a discipline that uses computational and mathematical techniques to store, manage and analyse biological data, in order to answer and explore biological questions. Bioinformatics has received a great deal of attention in the past few years from the computer science community. This is largely due to the complexity, time and expense of performing bench experiments to discover new biological knowledge. In conjunction with traditional experimental procedures, a biologist will use computer based information repositories and computational analysis for investigating and testing a hypothesis. These have become known as *in silico* experiments.

Laboratory bench and *in silico* experiments form a symbiosis. The *in silico* representation of the knowledge that forms a core component of bioinformatics is the subject of this chapter.

The biological sciences, especially molecular biology, currently lack the laws and mathematical support of sciences such as physics and chemistry. This is not to say that the biological sciences lack principles and understanding that, for instance, in physics allows us to predict planetary orbits, behaviour of waves and particles etc. We cannot, however, yet take a protein sequence and from the amino acid residues present deduce the structure, molecular function, biological role or location of that protein. The biologist has two options: First, to perform many laboratory experiments, *in vitro* and *in vivo* to acquire knowledge about the protein; second, the biologist takes advantage of one of the principles of molecular biology, which is that sequence is related to molecular function and structure. Therefore, a biologist can compare the protein sequence to others that are already well characterised. If the uncharacterised sequence is sufficiently similar to a characterised sequence, then it is inferred that the characteristics of one can be transferred to the other. So a key tool of bioinformatics is the sequence similarity search [1.4]; the characterisation of single sequences lies at the heart of most bioinformatics, even the new high-throughput techniques that investigate the modes of action of thousands of proteins per experiment. As the first method is expensive, both in terms of time and money, the latter can reduce the time to characterise unknown biological entities. Thus, we often see a cycle between laboratory bench and the computer.

1.1.1 Describing and using Biological Data

It has been said that biology is a knowledge based discipline [1.7]. Much of the community's knowledge is contained within the community's data resources. A typical resource is the SWISS-PROT protein database [1.6]. The protein sequence data itself is a relatively small part of the entry. Most of the entry is taken up by what the bioinformatics community refers to as 'annotation' which describe: physico-chemical features of the protein; comments on

NiceProt View of SWISS-PROT: P08100

General information about the entry

Entry name	**OPSD_HUMAN**
Primary accession number	**P08100**
Secondary accession number	Q16414
Entered in SWISS-PROT in	Release 08, August 1988
Sequence was last modified in	Release 08, August 1988
Annotations were last modified in	Release 40, October 2001

Name and origin of the protein

Protein name	**Rhodopsin**
Synonym	**Opsin 2**
Gene name	**RHO** or **OPN2**
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

Comments

- ***FUNCTION***: VISUAL PIGMENTS ARE THE LIGHT-ABSORBING MOLECULES THAT MEDIATE VISION. THEY CONSIST OF AN APOPROTEIN, OPSIN, COVALENTLY LINKED TO CIS-RETINAL.
- ***SUBCELLULAR LOCATION***: Integral membrane protein.
- ***TISSUE SPECIFICITY***: ROD SHAPED PHOTORECEPTOR CELLS WHICH MEDIATES VISION IN DIM LIGHT.
- ***PTM***: SOME OR ALL OF THE CARBOXYL-TERMINAL SER OR THR RESIDUES MAY BE PHOSPHORYLATED.
- ***DISEASE***: DEFECTS IN RHO ARE ONE OF THE CAUSES OF AUTOSOMAL DOMINANT RETINITIS PIGMENTOSA (ADRP). PATIENTS TYPICALLY HAVE NIGHT VISION BLINDNESS AND LOSS OF MIDPERIPHERAL VISUAL FIELD; AS THEIR CONDITION PROGRESSES, THEY LOSE THEIR FAR PERIPHERAL VISUAL FIELD AND EVENTUALLY CENTRAL VISION AS WELL.
- ***DISEASE***: DEFECTS IN RHO ARE ONE OF THE CAUSES OF AUTOSOMAL RECESSIVE RETINITIS PIGMENTOSA (ARRP).

Fig. 1.1. An extract of the SWISS-PROT entry for Human Rhodopsin. Much of the information is held in the comment field.

the whole sequence, such as function, disease, regulation, expression; species; names and so on. All this can be considered as the knowledge component of the database. Figure 1.1 shows a typical annotation from SWISS-PROT; note that the knowledge is captured as textual terms describing the findings, not numeric data, making use of shared keywords and controlled vocabularies. Whilst this style of representation is suitable for human readers, the current representation of the knowledge component is difficult to process by machine. SWISS-PROT itself now has over 100 000 entries (and growing exponentially), so its size makes it no longer suitable for human analysis and computational support is needed.

As well as this knowledge component, biological data is characterised in the following ways:

– Large quantity of data – The genome sequencing projects now mean that data is being produced at increasing rates; a new sequence is deposited in

the public genome database EMBL every 10 seconds[1]. Microarray experiments measuring gene expression and other high-through-put techniques now mean that other data are also being produced in vast quantity at petabytes per year [1.40].
- Complexity of data – It is difficult to represent most biological data directly in numeric form. Bioinformatics resources need non-scalar data types such as collections and records [1.10, 1.22]. Bioinformatics does not have a convenient data model; much bioinformatics data is kept in a natural language text-based form, in either annotations or bibliographic databases. As well as the basic data-representation, a characteristic of biology's data are the many relationships held by each entity. For instance, any one protein has a sequence, function, a process in which it acts, a location, a structure, physical interactions it makes, diseases in which it may be implicated, and many more. Capturing this knowledge makes biological data an extreme example of complexity in representation.
- Volatility of data – Once gathered, biological data is not static. As knowledge about biological entities changes and increases, so the annotations within data resources change.
- Heterogeneity of data – Much biological data is both syntactically and semantically heterogeneous [1.12]. Individual concepts, such as that of a gene, have many different, but equally valid, interpretations. There is a widespread and deep issue of synomyny and homonymy in the labels used for concepts within biology and as well as those used for the names of individuals.
- Distribution of data – Bioinformatics uses over 500 data resources and analysis tools [1.13] found all over the Internet. They often have Web interfaces and biologists enter data for analysis; cut-and-paste results to new Web resources or explore results through rich annotation with cross-links [1.23].

As well as the large number of data resources there are many analytical tools that work over these data resources to generate new data and knowledge. These tools suffer from the problems of distribution, heterogeneity, discovery, choice of suitable tool, etc. Some investigations can be carried out in one resource, but increasingly, many resources have to be orchestrated in order to accomplish an investigation. Often data resources lack query facilities usual in DBMS. The semantic heterogeneity between the resources exists both in schema and the values held within those schema. The vocabulary used by biologists to name entities, functions, processes, species, etc. can vary widely.

This scene leaves both the curators of bioinformatics resources and their users with great difficulties. A typical user, as well as a bioinformatics tool builder, is left trying to deal with the following problems in order to attempt tasks:

- Knowing which resources to use in a task;

[1] `http://www.ebi.ac.uk/`

– Discovering instances of those resources;
– Knowing how to use each of those resources, and how to link their content;
– Understanding the content of the resources and interpreting results;
– Transferring data between resources and reconciling values;
– Recording all that occurred during the *in silico* experiment.

All these steps require knowledge on the part of the biologists. It is no longer tenable for an individual biologist to acquire and retain this range and complexity of knowledge. This means bioinformatics needs computational support for storing, exploring, representing and exploiting this knowledge. Buttler [1.11] gives a description of a bioinformatics task workflow.

Ontologies describe and classify knowledge. Though biologists may not have used the term 'ontology', the use of classification and description as a technique for collecting, representing and using biological knowledge has a long history in the field. For example, the Linnaean classification of species is ubiquitous[2] and the Enzyme Commission has a classification of enzymes by the reaction that they catalyse [1.18]. Families of proteins are also classified along axes such as function and structural architecture [1.16]. Over the past five years there has been a surge of interest in using ontologies to describe and share biological data reflecting the surge in size, range and diversity of data and the need to assemble it from a broad constituency of sources. The Gene Ontology Consortium has launched OBO (Open Biological Ontologies)[3] which offers an umbrella to facilitate collaboration and dissemination of bio-ontologies.

1.1.2 The Uses of Ontologies in Biology

Ontologies are used in a wide range of biology application scenarios [1.38]:

– A defining database schema or knowledge bases. Public examples include RiboWeb, EcoCyc and PharmGKB [1.2, 1.25, 1.36]. Commercial knowledge bases include Ingenuity[4].
– A common vocabulary for describing, sharing, linking, classifying querying and indexing database annotation. This is currently the most popular use of ontologies in bioinformatics, and among many examples we can count The Gene Ontology, MGED[5], as well as those originating from the medical community such as UMLS[6].
– A means of inter-operating between multiple resources. A number of forms appear, for example: indexing across databases by shared vocabularies of their content (domain maps in BIRN [1.9]), inter-database navigation in

[2] `http://www.ncbi.nlm.nih.gov/Taxonomy/`
[3] `http://obo.sourceforge.net`
[4] `http://www.ingenuity.com`
[5] `http://www.mged.org`
[6] `http://www.nlm.nih.gov/research/umls/`

Amigo using the Gene Ontology[7]; a global ontology as a virtual schema over a federation of databases and application (TAMBIS [1.15]); and a description of bioinformatics services inputs, outputs and purpose used to classify and find appropriate resources, and control the workflows linking them together. (myGRID [1.42]).
- A scaffold for intelligent search over databases (e.g. TAMBIS) or classifying results. For example, when searching databases for 'mitochondrial double stranded DNA binding proteins', all and only those proteins, as well as those kind of proteins, will be found, as the exact terms for searching can be used. Queries can be refined by following relationships within the ontologies, in particular the taxonomic relationships. Similarly, Fridman Noy and Hafner [1.28] use an ontology of experimental design in molecular biology to describe and generate forms to query a repository of papers containing experimental methods. The extensions to a typical frame based representation allow them to describe accurately the transformations that take place, the complexes that form within an experiment and then make queries about those features.
- Understanding database annotation and technical literature. The ontologies are designed to support natural language processing that link domain knowledge and linguistic structures.
- A community reference, where the ontology is neutrally authored in a single language and converted into different forms for use in multiple target systems. Generally, ontologies have been developed to serve one of the previous categories of use, and then adopted by others for new uses. For example, the Gene Ontology, which will be the first of our detailed case studies, was developed solely for database annotation but is now used for all the purposes outlined above. As we will discuss, this has had an impact on its form, representational language and content.

Not only do ontologies offer a means for biologists to improve representation of knowledge in their resources, but the very size, volitility and complexity of the domain has potential benefit for computer scientists involved in ontology research. If the technologies proposed by ontology researchers can deal with the biological domain, then it is most likely that it can cope with a wide range of other domains, both natural and human-made. Before we explore some these uses in more detail through a number of case studies, we should point out some of the difficulties in modelling biological knowledge

1.1.3 The Complexity of Biological Knowledge

One of the interesting aspects of the use of ontologies within bioinformatics is the complexity and difficulty of the modelling entailed. Compared to the modelling of man-made artefacts such as aeroplanes, some argue that natural

[7] `http://www.godatabase.org`

systems are difficult to describe [1.19]. Biology is riddled with exceptions and it is often difficult to find the *necessary* conditions for class membership, let alone the *sufficiency* conditions. Often, biologists will 'know' that x is a member of y, despite it not having some of the same characteristics as all the other members of y. There are several potential reasons for this, including:

- Membership claims are in fact incorrect;
- Current biological knowledge is not rich enough to have found the appropriate necessary and sufficiency conditions;
- In the natural world, the boundaries between classes may be blurred. Evolution is often gradual and the properties that distinguish one class from another may be only partially represented in some individuals.

Jones [1.19] gives the following examples and reasons for how difficult modelling biology can be:

1. **Atypical examples** – Where an example of the class differs from one of the defining features. For example, all eukaryote cells contain a nucleus, but red blood cells do not [1.1, p18].
2. **Multiple sibling instantiation** – Where a class instance is a member of multiple children of that class. For example, neuroendocrine cells behave like both endocrine and nerve cells (both kinds of remote signaling cells) [1.1, p26], but do not satisfy all the characteristics of either cell type.
3. **Context sensitive membership** – Some classes only exist in certain contexts. Chemists talk about a defined set of chemical bonds, but biochemists sometimes also include certain 'weak bonds', such as hydrophobic bonds, when talking about molecules [1.1, p88].
4. **Excluded instances** – 'Small organic molecules' are divided into four kinds, 'simple sugars', 'amino acids', 'fatty acids' and 'nucleotides' [1.1, p84]. The same source, however, then defines other kinds of molecules that do not fall into these classes.
5. **Non-instance similarity** – where individuals exhibit similar features to those defining a class, but are not close enough to be a member of that class. For instance, mitochondria and chloroplasts, parts of eukaryotic cells, are very similar to prokaryotic cells. These entities are thought to have arisen from prokaryotes, but have become symbiotic and divergent from their ancestors.

Jones *et al* give several such examples of the difficulties in modelling biology. It is not necessarily that modelling is more difficult in biology than other domains, but several of the commonly occurring factors come together in modelling biology. The sample of 'atypical examples' given above, bears some investigation. Jones *et al*'s examples are taken from an undergraduate text book; such books often give 'simplified truth' or 'staged revelation', thus it is dangerous to take defining criteria from such resources. Like all

modelling, the conceptualisation has to come from many sources and depends upon the task to which the ontology is to be used.

In the rest of this chapter the use, nature and representation of some exemplar bio-ontologies will be described. In Section 1.2 the need for a shared vocabulary for the annotation of database entries is described. The Gene Ontology is used as the exemplar for this topic – it can be seen as the driving force behind much of the ontology activity in bioinformatics. Section 1.3 continues the theme of *ontology as specification* when the knowledge bases RiboWeb and EcoCyc are explored. In Section 1.4 we move to the use of ontology for *query management across multiple databases* with TAMBIS. Finally, in Section 1.5 several of these uses come together in an ontology of bioinformatics services used for discovery in the myGRID project.

The ontologies we describe come in three representational forms:

- Structured hierarchies of concept names;
- Frames defining concepts asserted into an isa hierarchy. Slots on frames carry the properties of each concept and constrain their fillers. Both the structured hierarchies of terms and frames require all concepts to be comprehensively pre-enumerated;
- Description Logics whose concepts can be combined dynamically via relationships to form new, compositional concepts. These compositional concepts are automatically classified, using reasoning. Compositional concepts can be made in a post co-ordinated manner: That is, the ontology is not a static artefact, users can interact with the ontology to build new concepts, composed of those already in the ontology, and have them checked for consistency and placed at the correct position in the ontology's lattice of concepts.

Biology is naturally compositional and hard to pre-enumerate; however even simple hand-crafted hierarchies are extremely useful.

1.2 Annotation: the Gene Ontology

The need for annotation is the driving force behind much of the ontology activity within bioinformatics. Information about *model organisms* has grown at a tremendous rate, leading to the development of model organism databases. Each has been built by an independent community of scientists, but the driving aim is to unify the results to synthesize an overall understanding of biological processes. Their effective use therefore demands a shared understanding in order to combine results. The Gene Ontology Consortium[8] set out to provide 'a structured precisely defined common controlled vocabulary for describing the roles of genes and gene products in any organism' [1.40].

[8] `http://www.geneontology.org`

1.2.1 Features of The Gene Ontology

The GO is really a handcrafted ontology in which phrases are placed in a structure of only is-a and part-of relationships. For example 'GO:0019466 ornithine catabolism, via proline' is a phrase which informs the biologist that the term represents the concept of catabolism of the chemical ornithine with a particular intermediate chemical form proline. These phrases form the controlled vocabulary with which to annotate three specific aspects of a gene product:- its functions; its role in a biological process; and its localization within a cell. Instead of using scientific English, annotation can now take place with terms taken from GO. This leads to better precision and recall of information within one database and more effective integration of information *across* databases.

Concept Definitions- Appropriate and consistent use of GO concepts requires all annotators to have a common understanding of what each concept represents. Therefore the GO consortium (GOC) places a great deal of effort in providing a definition for each concept. Currently over 60% of GO concepts have a textual definition. The concepts are represented as strings descriptions of increasing detail coupled with a unique identifier that carries no semantics. This separates the labels as they are used in the databases from the current definition of the term.

Hierarchial organisation- It is impractical to deliver such a large vocabulary as a simple list. Therefore the concepts are organized into hierarchies. The semantics of the parent child link is stated explicitly as either subsumption or partomomy. Each concept can have any number of parents and so its place in the hierarchy is represented as a directed acyclic graph (DAG).

The hierarchical structure is used by users for a number of purposes:

- **Internal Navigation**. The hierarchy acts as a way of grouping similar concepts and so allowing annotators to find the concept they require quickly;
- **Database content browsing**. The hierarchy acts as a index into each database. GO Browsers, e.g. AmiGO[9] allow users to link directly from the hierarchical view of the ontology to database entries annotated with those concepts (see Figure 1.2).
- **Aggregate information**. A GO Slim is a non-overlapping subset of high-level GO concepts. Aggregating all entries annotated with hierarchical descendants of each GO Slim term can produce useful summary statistics. The 'GO summary' feature of the AmiGO browser demonstrates how this information is used to provide a high level view of GO annotation statistics.

At the time of writing the Gene Ontology stands at some 15,000 concepts and continues to rapidly expand. Its success is attributed to many factors, including:

[9] `http://www.godatabase.org`

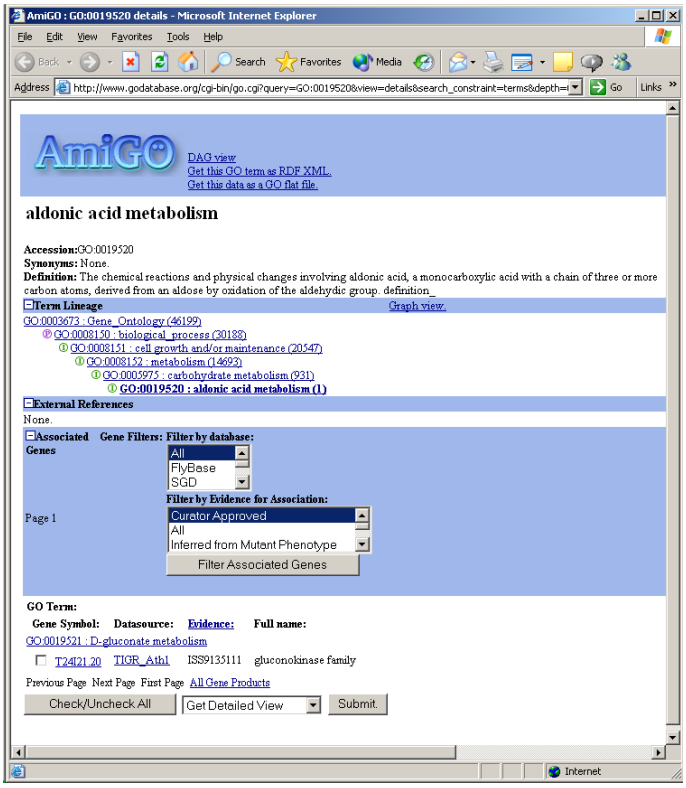

Fig. 1.2. Screenshot of the Amigo browser showing how a Gene Ontology concept 'aldonic acid metabolism' has been used to annotate a Gene product entry in the TIGR database.

– There was no attempt to try to model everything but instead to chose a narrow, but useful part of biology. Despite its narrow focus GO has already gained wide acceptance and it is already being used for purposes outside annotation;
– There was no attempt to wait for the ontology to be 'complete' or 'correct'; as soon as GO was useful, the GOC used it and put in place mechanisms to deal with changes and the depreciation of terms. The GO identifiers hold no semantic information and thus separate the labelling of database entries from the interpretation of the labels. Biological knowledge changes constantly as do ways of modelling that knowledge. As development is continuous they use CVS[10] to manage version control. The GO editorial team also annotate their terms with author date, definitions and provenance argumentation.

[10] `http://www.cvshome.org/`

– The process and the ontology is open and involves the community. The development of GO is controlled by a small team of curators who manage the publication and versioning activity, with a wider team of active ontology developers who provide, update and correct the content. The GO developers will take all suggestions from the general community, process them and incorporate or reject with reasons in a timely fashion.
– The developers are biologists and experts who have been supported by knowledge management tools. Attempts by professional knowledge engineers to elicit knowledge from experts and do the modelling are doomed to failure: the GO curators are all post-doctoral biologists and the GO represents their and their communities distilled and accumulated knowledge.

By these procedures and principles GO has become a widely used and respected ontology within bioinformatics. The coverage of GO is narrow, but nonetheless important. Molecular biology is a vast domain and an attempt to cover the whole would have undoubtedly failed. GO was also created for a specific purpose, namely that of annotation – there are many task that GO does not support in its current representation, such as mappings to linguistic forms that would make generation of natural language annotations of databases easier. GO has, however, demonstrated to the community that even with a simple representation, a shared view on the three major attributes of gene products may be achieved.

1.2.2 Computationally Amenable Forms of GO

All the uses of GO described above revolve around human interpretation of the phrase's meaning. However, there is a growing need for applications to have access to a more explicit machine interpretable description of each phrase. For example, instead of relying on similarities of proteins by the similarity of their sequences, they could be clustered on the similarity of their function by grouping their Gene Ontology terms. This requires several measures of 'semantic similarity', for example those of [1.29, 1.30] which exploit both the DAG structure of GO, and the usage of GO terms within the various databases now annotated with GO. This uses the notion of 'information content', which says commonly occurring terms, like 'receptor' are not likely to be very discriminatory [1.32].

The definition of a metric for 'semantic similarity' between GO terms, allows us to exploit the machine interpretable semantics of GO for large datasets. By comparing these metrics to sequence similarity measures we managed to isolate a number of errors in either GO, or the use of GO within the annotated databases [1.29]. We have also investigated the use of these metrics as the basis for a search tool, to allow querying within a database[11].

Perhaps the most pressing need is that of maintaining the structure of the Gene Ontology itself. The growing size and complexity of GO is forcing

[11] `http://gosst.man.ac.uk`

its curators to spend more and more time on the mundane task of maintaining the logical consistency and completeness of its internal structure. Within GO many concepts have multiple parents. The maintenance of these links is a manual process. Experience from the medical domain has shown that numerous parent-child links are omitted in such hand crafted controlled vocabularies [1.35]. While of less importance to manual interpretation, machine interpretation will falter in the face of such inconsistencies.

The Gene Ontology Next Generation project (GONG)[12] aims to demonstrate that, in principle, migrating to a finer grained formal conceptualization in DAML+OIL [1.17] will allow computation techniques, such as description logics, to ensure logical consistency freeing the highly trained curators to focus on capturing biological knowledge [1.43]. GO is large so GONG takes a staged approach in which progressively more semantic information is added *insitu*. Description logic reasoning is used early and often, and suggested amendments sent to the GO editorial team.

To use the description logic to maintain the links automatically, the concepts are dissected, explicitly stating the concepts definition in a formal representation. This provides the substrate for description logic reasoners to infer new is-a links and remove redundant links.

Within a large phrased based ontology such as GO, which contains many concepts within a narrow semantic range, it is possible to use automated techniques to construct candidate dissections by simply parsing the term name. For example many metabolism terms in GO follow the pattern 'chemical name' followed by either 'metabolism','catabolism' or 'biosynthesis'. If a term name fits this pattern a dissection can be created from the relevant phrase constituents as shown in Figure 1.3. These patterns have to be spotted by a developer and the scripts that generate the DL representation targeted at the appropriate regions of the GO. This provides a semi-automated, targeted approach, which avoids patterns being too general: For example, confusing 'Protein Expression' and 'Gene Expression', which may fit a general pattern, but where the former describes a 'target' and the latter a 'source'.

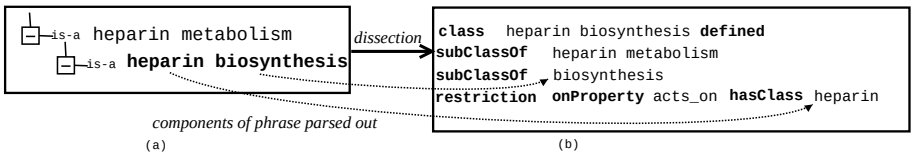

Fig. 1.3. Diagram showing the dissection of (a) the GO concept heparin biosynthesis in its original DAG into (b) a DAML+OIL like definition with additional semantic information.

[12] `http://gong.man.ac.uk`

The process of dissection breaks down the existing concept into more elemental concepts related together in a formal semantic manner. These elemental concepts are then placed in orthogonal taxonomies. Taxonomic information such as the classification of chemical substances which was previously implicit and repeated in many sections of the GO ontology is now made explicit in an independent chemical ontology. The reasoner combines the information in these independent taxonomies to produce a complete and consistent multi-axial classification. The changes reported by the DL reasoner represent mostly additional relationships hard to spot by human eye, and not errors in biological knowledge. The effect of adding descriptions and using the reasoner can be seen in Figure 1.4.

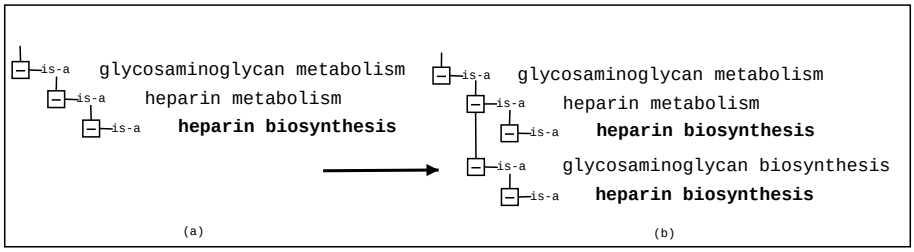

Fig. 1.4. Directed acyclic graph showing additional parent for heparin biosynthesis found using the reasoner.

For example, the reasoner reported that 'heparin biosynthesis' has a new is-a parent 'glycosaminoglycan biosynthesis'. These reports can then be sent to the editorial team for comment and action if necessary. Even at this early stage of the GONG project, the utility of the approach can be recognised. Many missing and redundant is-a relationships have been spotted, making GO more complete and robust. Members of the GO editorial team have recognised the potential of using such a logic based approach to automatically place concepts in the correct location – a task seen as difficult by the team in GO's current hand-crafted form.

1.3 Schema Definition: EcoCyc

The complexity of biological systems means that relational databases do not make good management systems for biological data and their associated knowledge [1.22]. It is possible to develop relational schemata for such complex material, but it is hard work. Major sequence repositories are stored in relational form, but using highly complex, less than intuitive schema. Such repositories are managed by skilled bioinformaticians and database administrators. For biologists investigating the data different presentations are re-

quired. An object style approach, with its complex data types (especially collections and user defined classes as domains for attributes) makes ontological modelling of the data much easier. Object databases have not reached the same level of technical reliability as the relational form, but frame based knowledge bases provide an object like view of the world, but can store and retrieve large amounts of data efficiently. While many bioinformatics resources have simply used a flat-file system to hold these data, others have explored the use of ontologies to describe the data contained within that resource.

The elements within the ontology describe the data held in the resource and these descriptions are used to gather and represent the facts described by the ontology. These knowledge bases form one of the earlier uses of ontology within bioinformatics. Indeed, the development of the EcoCyc [1.25] KB necessitated the description of a classification of the function of gene products [1.34]; an early forerunner of GO. EcoCyc uses frames as a knowledge representation formalism; using slots to gather all the attributes that describe, for instance, a protein.

1.3.1 EcoCyc: Encyclopaedia of E.coli

EcoCyc uses an ontology to describe the richness and complexity of a domain and the constraints acting within that domain, to specify a database schema [1.24]. Classes within the ontology form a schema; instances of classes, with values for the attributes, form the facts that with the ontology form the knowledge base. EcoCyc is presented to biologists using an encyclopaedia metaphor. It covers *E. coli.* genes, metabolism, regulation and signal transduction, which a biologist can explore and use to visualise information [1.26].

The instances in the knowledge base currently include 165 Pathways, involving 2604 Reactions, catalysed by 905 Enzymes and supported by 162 Transporters and other proteins expressed by 4393 Genes [1.26]. EcoCyc uses the classification of gene product function from Riley [1.34] as part of this description. Scientists can visualise the layout of genes within the *E. coli.* chromosome, or of an individual biochemical reaction, or of a complete biochemical pathway (with compound structures displayed).

EcoCyc uses the frame-based language Ocelot, whose capabilities are similar to those of HyperTHEO [1.24], to describe its ontology. The core classes that describe the *E. coli* genome, metabolism, etc. include a simple taxonomy of chemicals, so that DNA, RNA, polypeptides and proteins may be described. `Chromosomes` are made of `DNA` and `Genes` are segments of `DNA`, located on a `Chromosome`. Pathways are collections of `Reactions`, that act upon `Chemicals`. All *E. coli* genes are instances of the class gene and consequently share the properties or attributes of that class. Each EcoCyc frame or class contains slots that describe either attributes of the biological object that the frame represents, or that encode a relationship between that object and other objects. For example, the slots of a polypeptide frame encode the

molecular weight of the polypeptide, the gene that encodes it, and its cellular location.

EcoCyc's use of an ontology to define a database schema has the advantages of its expressivity and ability to evolve quickly to account for the rapid schema changes needed for biological information [1.24]. The user is not aware of this use of an ontology, except that the constraints expressed in the knowledge captured mean that the complexity of the data held is captured precisely. In EcoCyc, for example, the concept of `Gene` is represented by a concept or class with various attributes, that link through to other concepts: `Polypeptide product`, `Gene name`, `synonyms` and `identifiers` used in other databases etc. The representation system can be used to impose constraints on those concepts and instances which may appear in the places described within the system. EcoCyc's ontology has now been used to form a generic schema MetaCyc, that is used to form the basis for a host of genomic knowledge bases [1.27]. These ontologies are used to drive pathway prediction tools based upon the genomic information stored in the knowledge base. From the presence of genes and knowledge of their function, knowledge can be inferred about the metabolomes of the species in question [1.21]. Such computations are not only possible with the use of ontology, but EcoCyc's developers would argue that their ontology based system and the software it supports makes such a complex task easier.

The rich, structured and constrained nature of these knowledge bases mean that they form a better founded platform for bioinformatics software than would be usual with, for instance, the community's reliance upon flat-file storage. Ecocyc uses the knowledge base to generate pathways, perform cross-genome comparisons and generate sophisticated visualisations. Similarly, RiboWeb [1.2] uses the constraints in its ontological model to guide a user through the analysis of structural data: it captures knowledge of which methods are appropriate for which data and can use knowledge to perform validations of results. Ontologies as bioinformatics database schema prove their worth in capturing knowledge with high fidelity and managing the modelling of complex and volatile data and associated knowledge.

1.4 Query Formulation: TAMBIS

This section presents an approach to solving the problems of querying distributed bioinformatics resources called TAMBIS (Transparent Access to Multiple Bioinformatics Information Sources) [1.15]. The TAMBIS approach attempts to avoid the problems of using multiple resources by using an ontology of molecular biology and bioinformatics to manage the presentation and usage of the sources. The ontology allows TAMBIS: to provide a homogenising layer over the numerous databases and analysis tools; to manage the heterogeneities between the data sources; and to provide a common, con-

sistent query-forming user interface that allows queries across sources to be precisely expressed and progressively refined.

A concept is a description of a set of instances, so a concept can also be viewed as a query. The TAMBIS system is used for retrieving instances described by concepts in the model. This contrasts with queries phrased in terms of the structures used to store the data, as in conventional database query environments. This approach allows a biologist to ask complex questions that access and combine data from different sources. However, in TAMBIS, the user does not have to choose the sources, identify the location of the sources, express requests in the language of the source, or transfer data items between sources.

The steps in the processing of a TAMBIS query are as follows:

1. A query is formulated in terms of the concepts and relationships in the ontology using the visual *Conceptual Query Formulation Interface*. This interface allows the ontology to be browsed by users, and supports the construction of complex concept descriptions that serve as queries. The output of the query formulation process is a *source independent conceptual query*. The query formulation interface makes extensive use of the TAMBIS *Ontology Server* which supports various reasoning services over the ontology, to ensure that the queries constructed are biologically meaningful.
2. Given a query, TAMBIS must identify the sources that can be used to answer the query, and construct valid and efficient source independent query plans for evaluating the query given the facilities provided by the relevant sources. Concepts and relationships from the Ontology are associated with the services provided by the sources.
3. The *Query Plan Execution* process takes the plan provided by the planner and executes that plan over the *Wrapped Sources* to yield an answer to the query. Sources are wrapped so that they can be accessed in a syntactically consistent manner.

The TAMBIS ontology describes both molecular biology and bioinformatics tasks. Concepts such as `Protein` and `Nucleic acid` are part of the world of molecular biology. An `Accession number`, which acts as a unique identifier for an entry in an information source, lies outside this domain, but is essential for describing bioinformatics tasks in molecular biology. The TAMBIS ontology has been designed to cover the standard range of bioinformatics retrieval and analysis tasks [1.39]. This means that a broad range of biology has been described. The model is quite shallow, although the detail present is sufficient to allow most retrieval tasks supportable using the integrated bioinformatics sources to be described. In addition, precision can arise from the ability to combine concepts to create more specialised concepts. The model is described

in more detail in [1.7] and can be browsed via an applet on the TAMBIS Web site[13].

The TAMBIS ontology is described using an early Description Logic called GRAIL [1.31]. The GRAIL representation has a useful extra property in its ability to describe constraints about when relationships are allowed to be formed. For example, it is true that a `Motif` is a component of a `Biopolymer`, but not all motifs are components of all biopolymers. For example, a `PhosphorylationSite` can be a component of a `Protein`, but not a component of a `Nucleic acid`, both of which are `Biopolymers`. The constraint mechanism allows the TAMBIS model to capture this distinction, and thus only allow the description of concepts that are described as being biologically meaningful, in terms of the model from which they are built.

The task of query formulation involves the user in constructing a concept that describes the information of interest. By using a post-co-ordinated ontology, TAMBIS is able to provide a variety of complex queries over a range of diverse bioinformatics resources. Mappings from concepts to resource specific calls or values allows TAMBIS to deal with the heterogeneity present in the resources and give the illusion of a common query interface. A small sample of such queries are: *'Find the active sites of hydrolase enzymes, with protein substrates and metal cofactors'* and *'Find all chimpanze proteins similar to human apoptosis proteins'.*

1.5 Service Discovery: the myGRID Service Ontology

Both data and analytical resources provide services to bioinformaticians. A characteristic of bioinformatics is the *discovery* of suitable resources and the marshalling of those resources to work together to perform a task. However, the 'craft-based' practice of a biologist undertaking the discovery, interoperation and management of the resources by hand is unsupportable, as described in Section 1.1. These difficulties mean that the discovery and assembly of resources or services on those resources must be at least semi-automated.

Users will typically have in mind a task they want to perform on a particular kind of data. They must match this task against available services taking into account the function of the service, the data it accepts and produces and the resources it uses to accomplish its goal. In addition, they must select, from the candidates that can fulfill their task, the one that is best able to achieve the result within the required constraints. This choice depends on metadata concerning function, cost, quality of service, geographical location, and who published it. The discovery process as a whole requires a much more conceptual description of a service than the metadata usually associated with a web service which focuses on its low level syntactic interface.

[13] `http://img.cs.man.ac.uk/tambis`

The process of narrowing down a selection into the appropriate set is currently supported by simple conceptual classifications rather than sets of individual conceptual descriptions, in a manner analogous to using the Yellow Pages^{TM}. This classification of services based on the functionality they provide has been widely adopted by diverse communities as an efficient way of finding suitable services. For example, the EMBOSS suite [1.33] of bioinformatics applications and repositories has a coarse classification of the 200 or so tools it contains, and free text documentation for each tool. The bioinformatics integration platforms ISYS [1.37] and BioMOBY (http://www.biomoby.org) use taxonomies for classifying services. The Universal Description, Discovery, and Integration specification (UDDI) [1.41] supports web service discovery by using a service classification such as UNSPSC [1.14] or RosettaNet [1.20].

The advent of the Semantic Web has meant that there is increasing interest not only in the semantic description of content, but in the semantic description of the services provided through the Web [1.8]. As with EcoCyc described earlier, ontologies have been used as a schema for the description of web services. DAML-S [1.3] offers an upper level ontology for the description of Web Services. Within ^{my}GRID (see below) ontologies can also provide the vocabulary of concepts with which to compose these descriptions. Working with a formal representation such as DAML+OIL also allows classifications to be validated/ constructed from these description as has been described with the GONG project.

^{my}GRID [14] is a UK e-Science pilot project specifically targeted at developing open source high-level middleware to support personalised semantics-rich *in-silico* experiments in biology. The emphasis is on database integration, workflow, personalisation and provenance, with a primary focus on the use of rich ontology based semantics to aid in the discovery and orchestration of services. ^{my}GRID uses a suite of ontologies expressed in DAML+OIL [1.5], to provide: (a) a schema for describing services based on DAML-S; (b) a vocabulary for expressing service descriptions and (c) a reasoning process to both manage the coherency of the classifications and the descriptions when they are *created*, and the service discovery, matching and composition when they are *deployed*.

1.5.1 Extending DAML-S in terms of properties

A key bottleneck in the utilisation of services is the discovery from the myriad available those that will fulfil the requirements of the task at hand. This discovery involves matching the users requirements against functional descriptions of the available services.

DAML-S provides a high level schema in DAML+OIL with which to capture some of these functional attributes together with additional attributes

[14] `http://www.mygrid.org.uk`

describing authorship, cost etc. From our experience in writing over 100 descriptions, during the development of myGRID, for preexisting bioinformatics services we have found DAML-S defined attributes describing the inputs and outputs to the service the most discriminatory. In addition, we felt it necessary to add a set of attributes to the service profile to capture common ways of describing bioinformatics service. These include a generic description of the overall *task*; associated *resources* used to fulfil the task; software *tools* and *algorithms* with which the task is performed.

myGRID has additionally built a suite of DAML+OIL ontologies specific to bioinformatics and molecular biology which provides the vocabulary for the services to be described. Figure 1.5 shows how these ontologies are interrelated.

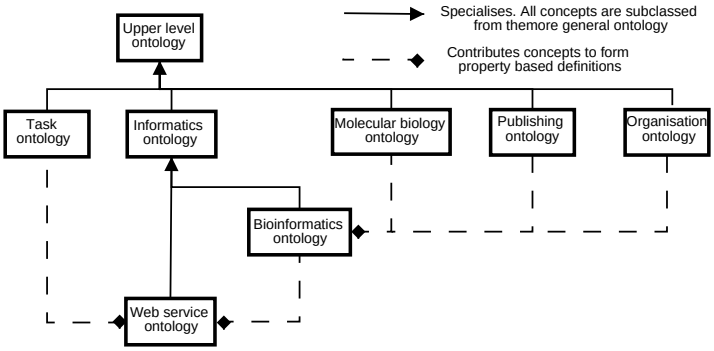

Fig. 1.5. Suite of ontologies used in myGRID and their inter-relationships.

A *standard upper level ontology* forms the foundation for the suite of ontologies. An *informatics ontology* captures the key concepts of data, data structures, databases, metadata and so forth. As the DAML-S service ontology is designed specifically to support web services it becomes an extension of the informatics ontology. A **bioinformatics ontology** builds on the informatics ontology adding specific types of bioinformatics resource such as `SWISS-PROT database`, `BLAST application`, and specific bioinformatics data such as `protein sequence`. By explicitly separating general informatics concepts from more specific concepts applicable only to bioinformatics, we hope to reuse as much as possible of the ontology suite for other domains. A *molecular biology* ontology with which to describe the content of data passed into and out of bioinformatics services. Examples of concepts include `protein`, `nucleic acid`, and `sequence`. These concepts tend to be much more general than found in existing ontologies such as the Gene Ontology. Small *publishing*, *organisation* and *task* ontologies have also been constructed to provide the necessary vocabulary for service descriptions.

Figure 1.6 gives an example of the formal definitions for one of the operation BLAST-n which compares a nucleotide sequence against a nucleotide sequence database using alignment.

```
class  BLAST-n service operation defined
  subClassOf atomic service operation
  restriction onProperty performs_task   hasClass
     aligning  restriction  onProperty  has_feature   hasClass local
               restriction  onProperty  has_feature   hasClass pairwise
  restriction onProperty produces_result hasClass
     report    restriction  onProperty  is_report_of  hasClass sequence alignment
  restriction onProperty uses_resource   hasClass
     database  restriction  onProperty  contains      hasClass
          data  restriction   onProperty   encodes   hasClass
             sequence    restriction  onProperty is_sequence_of hasClass
                   nucleic acid molecule
```

Fig. 1.6. Fully expanded formal description of the BLAST-n service operation written in a human-readable pseudo version of DAML+OIL.

Within myGRID, this ontology of services and its contributory ontologies have provided the vocabulary for about a hundred bioinformatics service descriptions. These descriptions have been linked to entries within a UDDI service registry allowing users to search and find appropriate registered services via a myGRID 'web portal'. The use of a reasoner and the consequent post co-ordinated nature of the ontology means that a flexible variety of views or queries by the user can be provided. As well as searching for services by the descriptions already asserted in the ontology, new 'partial descriptions' can be created, that provide more general descriptions of classes. It is easy for instance, to create a new class 'all services provided by the European Bioinformatics Institute' or 'all services that take a protein sequence as input'.

Concepts from the bioinformatics ontology can be used to give semantic descriptions of data, both inputs and outputs, stored in a bioinformatician's personal storage. This annotation would allow services to be sought by the kind of data in hand. Such an activity could also work backwards. Given a particular 'analytical goal', workflows could be composed backwards to suggest protocols to users. A bioinformatician could ask the question 'how do I generate a phylogenetic tree?'. Starting with the concept 'Phylogenetic tree', an inverse of the 'output' relation would be followed to find the service that generates such a tree. Continuing this process would generate a range of possible paths by which that output could be derived. Similarly, decoration of all these data with semantic annotations allows a variety of views to be taken of those data. They can be organised along multiple axes, including experiment, experimenter, genes, proteins, species, etc. Such flexible semantic views allow a personalisation of science that is traditionally difficult to achieve.

In the myGRID service ontology many themes of this chapter come together. The integrated ontology itself, provides a global schema, giving a

common view over all the services it includes. Like TAMBIS, it allows 'query concepts' to be built to retrieve services suited to the query. Heterogeneity in the services are reconciled to the myGRID ontology to give a common view. Fragments of the ontology are also used for annotation of data and results (that may also form data for input in their own turn) can be queried and assembled using those semantic descriptions. Here, annotation, schema definition and query formulation can be seen at one time in a bioinformatics ontology.

1.6 Discussion

Ontologies have become increasingly widely used in biology because of need. Science is all about increase in the understanding of the world about us; so, the communities within a scientific discipline need to have a shared understanding. The Gene Ontology's principle purpose is to provide a shared understanding between different model organism communities. The use of ontologies to deliver terminologies for annotation of data is undoubtedly the area of greatest use of ontologies within biology. The need for confidence in the use of terms when curating and querying resources is a strong driving force behind this effort. The GO is without doubt the largest of these efforts, but many others exist within the domain. To accommodate these efforts, the Gene Ontology Consortium has launched OBO (Open Biological Ontologies)[15], which offers an unbrella to facilitate collaboration and dissemination of bio-ontologies and offers a set of rules for inclusion. One ontology will not cover the whole of biology, so a range of ontologies will have to work together; moreover, ontologies need to be exchanged and preferably represented using the same formalism. The community originated the XOL exchange markup language, that was one of the influences on the OIL ontology language, later to become DAML+OIL. OBO has enthusiastically embraced DAML+OIL as a common language.

The original need to provide a shared understanding mainly for humans, is now leading towards an increased emphasis on shared understanding within and between humans and computers. The GONG project (Section 1.2) shows how modern Description Logic representation in the form of DAML+OIL can be used to manage GO to give a more complete and robust GO. This is a good demonstration of the computer science ontology community aiding domain experts in building an ontology and a domain offering a superb test bed for a new language and technology. Bioinformaticians have a role to act as intermediates between biologists and the knowledge engineering community.

Knowledge models are not simply created as instances of truth and beauty – they need to work and be useful. Knowledge bases such as EcoCyc provide

[15] `http://obo.sourceforge.net`

complex visualisation and prediction systems based upon their knowledge and the representations have to work in order for this to happen.

Biology provides real world examples of interesting, useful problems for computer scientists to explore and solve. Technology should be able to free the scientists to do his or her science. If knowledge engineers believe ontologies to be useful, then they should be able to be useful in biology. Are we able to express the range and complexity of the biological world with high-fidelity in our knowledge representation languages? Are our technologies, such as reasoning services, scalable to the size and complexity of the domain? Are we able to cope with the volatility of scientific knowledge? Trying to cope with all these aspects will push at the boundaries of our technologies.

This interplay can be seen within the ontologies discussed in this chapter: The GO is relatively simple, but very widely used, with a huge community. It is also an on-going effort, being updated and released continually, as the domain knowledge itself grows. EcoCyc uses an ontology in a standard knowledge representation language to create a large knowledge base of instances that can drive sophisticated visualisation and querying tasks. Again, this ontology evolves with the community knowledge and has a large user base. The other ontologies described lie more within the computer science research comunity and use bioinformatics as a rigorous test domain. GONG demonstrates that description logics can aid such a community in building and maintaining large, complex ontologies. TAMBIS and myGRID again show that complex domains can be represented and managed with modern DL technology. These projects currently lie within the research domain and will become more widely used as the bioinformatics community itself starts enlarging and using the ontologies.

Classification is an old, tried and tested scientific tool. The computer scientists' understanding of the meaning of ontology is often wider than just classification, but it is no surprise that biologists take to the technology. Classification has formed an underpinning of science from the periodic table of elements to the linnaean taxonomy of species. From organising data and classes of data, new scientific insights may arise – the most prominent example of this is the periodic table of elements; the taxonomy of species also reflects evolutionary change. New fields of scientific investigation, like genomics and the wider field of bioinformatics, mean vast new fields of data now need to be organised. Ontologies offer a good, flexible way of organising these data and what we know about these data. The ultimate dream of those who model knowledge is that their modelling will lead to new scientific insights. Maybe this will happen with bio-ontologies.

Acknowledgements: Chris Wroe and Phillip Lord are funded by EPSRC GR/R67743 eScience project myGRID and the GONG DARPA DAML Stanford subcontract PY-1149.

10

1.1 B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J.D. Watson. *Molecular Biology of the Cell*. Garland, New York, 1989.
1.2 R. Altman, M. Bada, X.J. Chai, M. Whirl Carillo, R.O. Chen, and N.F. Abernethy. RiboWeb: An Ontology-Based System for Collaborative Molecular Biology. *IEEE Intelligent Systems*, 14(5):68–76, 1999.
1.3 A. Ankolekar, M. Burstein, J. Hobbs, O. Lassila, D. Martin, S. McIlraith, S. Narayanan, M. Paolucci, T. Payne, K. Sycara, and H. Zeng. DAML-S: Semantic Markup for Web Services. In *Proceedings of the International Semantic Web Working Symposium (SWWS)*, 2001.
1.4 T.K. Attwood and D.J. Parry-Smith. *Introduction to bioinformatics*. Addison Wesley Longman, 1999.
1.5 F. Baader, D. McGuinness, D. Nardi, and P. P. Schneider, editors. *The Description Logic Handbook Theory, Implementation and Applications* . Cambridge University Press, 2003.
1.6 A. Bairoch and R. Apweiler. The SWISS-PROT Protein Sequence Data Bank and its Supplement TrEMBL in 1999. *Nucleic Acids Research*, 27:49–5, 1999.
1.7 P.G. Baker, C.A. Goble, S. Bechhofer, N.W. Paton, R. Stevens, and A Brass. An Ontology for Bioinformatics Applications. *Bioinformatics*, 15(6):510–520, 1999.
1.8 T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, pages 28–37, May 2001.
1.9 Maryann E. Martone Bertram Ludscher, Amarnath Gupta. Model-based mediation with domain maps. In *Conference on Data Engineering (ICDE), Heidelberg, Germany*. IEEE Computer Society, 2001.
1.10 P. Buneman, S.B. Davidson, K. Hart, C. Overton, and L. Wong. A Data Transformation System for Biological Data Sources. In *Proceedings of VLDB*, pages 158–169. Morgan Kaufmann, 1995.
1.11 David Buttler, Matthew Coleman1, Terence Critchlow1, Renato Fileto, Wei Han, Ling Liu, Calton Pu, Daniel Rocco, and Li Xiong. Querying multiple bioinformatics data sources: Can semantic web research help? *SIGMOD Record*, 2002. Special Issue.
1.12 I.A. Chen and V.M. Markowitz. An Overview of the Object-Protocol Model (OPM) and the OPM Data Management Tools. *Information Systems*, 20(5):393–418, 1995.
1.13 C. Discala, X. Benigni, E. Barillot, and G. Vaysseix. DBcat: A Catalog of 500 Biological Databases. *Nucleic Acids Research*, 28(1):8–9, 2000.
1.14 Electronic Commerce Code Management Association Technical Secretariat. Universal Products and Services Classification Implementation Guide , June 2001. Available: `http://eccma.org/unspsc`.
1.15 C.A. Goble, R. Stevens, G. Ng, S. Bechhofer, N.W. Paton, P.G. Baker, M. Peim, and A. Brass. Transparent Access to Multiple Bioinformatics Information Sources. *IBM Systems Journal Special issue on deep computing for the life sciences*, 40(2):532 – 552, 2001.
1.16 Caroline Hadley and David T. Jones. A Systematic Comparison of Protein Structure Classifications: SCOP, CATH and FSSP. *Structure*, 7(9):1099– 1112, 1999.
1.17 I. Horrocks. DAML+OIL: a reason-able web ontology language. In *Proc. of EDBT 2002*, pages 2–13. Lecture Notes in Computer Science, 2002.
1.18 International Union of Biochemistry . *Enzyme Nomenclature 1984 : Recommendations of the Nomenclature Committee of the International Union of*

Biochemistry on the Nomenclature and Classification of Enzyme-Catalyzed Reactions. Academic Press (for The International Union of Biochemistry by), Orlando, FL, 1984.

1.19 D.M. Jones, P.R.S. Visser, and R.C. Paton. Addressing Biological Complexity to Enable Knowledge Sharing. In *AAAI'98 Workshop on Knowledge Sharing Across Biological and Medical Knowledge-based Systems*, 1998.

1.20 R. Kak and D. Sotero. Implementing RosettaNet E-Business Standards for Greater Supply Chain Collaboration and Efficiency, 2002. RosettaNet White Paper Available: `http://www.rosettanet.org`.

1.21 M. Karp, P. amd Krummenacker, S. Paley, and J. Wagg. Integrated pathway/genome databases and their role in drug discovery. *Trends in Biotechnology*, 17:275–281, 1999.

1.22 P. Karp. Frame representation and relational data bases: Alternative information management technologies for systematics. In R. Fortuner, editor, *Advanced Computer Methods for Systematic Biology: Artificial Intelligence, Database Systems, Computer Vision*. The Johns Hopkins University Press, 1993.

1.23 P. Karp. A Strategy for Database Interoperation. *Journal of Computational Biology*, 2(4):573–586, 1995.

1.24 P. Karp and S. Paley. Integrated Access to Metabolic and Genomic Data. *Journal of Computational Biology*, 3(1):191–212, 1996.

1.25 P.D. Karp, M. Riley, M. Saier, I.T. Paulsen, S.M. Paley, and A. Pellegrini-Toole. The EcoCyc and MetaCyc Databases. *Nucleic Acids Research*, 28:56–59, 2000.

1.26 Peter D. Karp, Monica Riley, Milton Saier amd Ian T. Paulsen amd Julio Collado-Vides, Suzanne M. Paley, Alida Pellegrini-Toole, and esar Bonavides amd Socorro Gama-Castro. The EcoCyc Database. *Nucleic Acids Research*, 30(1):56–58, 2002.

1.27 Peter D. Karp, Monica Riley, Suzanne M. Paley, and Alida Pellegrini-Toole. The MetaCyc Database. *Nucleic Acids Research*, 30(1):59–61, 2002.

1.28 Natalya Fridman Noy and Carole D. Hafner. Representing scientific experiments: Implications for ontology design and knowledge sharing. In *AAAI/IAAI*, pages 615–622, 1998.

1.29 P.W.Lord, R.D. Stevens, A. Brass, and C.A.Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–83, 2003.

1.30 P.W.Lord, R.D. Stevens, A. Brass, and C.A.Goble. Semantic similarity measures as tools for exploring the Gene Ontology. In *Pacific Symposium on Biocomputing*, pages 601–612, 2003.

1.31 A.L. Rector, S.K. Bechhofer, C.A. Goble, I. Horrocks, W.A. Nowlan, and W.D. Solomon. The GRAIL Concept Modelling Language for Medical Terminology. *Artificial Intelligence in Medicine*, 9:139–171, 1997.

1.32 P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

1.33 P. Rice, I. Longde, and A. Bleasby. EMBOSS: The European Molecular Biology Open Software Suite . *Trends in Genetics*, 16(6):276–277, 2000.

1.34 M. Riley. Functions of the gene products of Escherichia coli. *Microbiological Reviews*, 57:862–952, 1993.

1.35 J.E. Rogers, C. Price, A.L. Rector, W.D. Solomon, and N. Smejko. Validating Clinical Terminology Structures: Integration and Cross-Validation of Read Thesaurus and GALEN. In *AMIA Fall Symposium*, 1998.

1.36 Daniel L. Rubin, Farhad Shafa, Diane E. Oliver, Micheal Hewett, and Russ B. Altman. Representing genetic sequence data for pharmacogenomics: an evolutionary approach using ontological and relational models. In Chris Sander, editor, *Proceedings of Tenth International Conference on Intelligent Systems for Molecular Biology*, volume 18 Supplement 1, pages 207–215, 2002.

1.37 A.C. Siepel, A.N. Tolopko, A.D. Farmer, P.A. Steadman, F.D. Schilkey, B.D. Perry, and W.D. Beavis. An integration platform for heterogenous bioinformatics software components. *IBM Systems Journal*, 40(2):570–591, 2001.

1.38 R. Stevens, C.A. Goble, and S. Bechhofer. Ontology-based Knowledge Representation for Bioinformatics. *Briefings in Bioinformatics*, 1(4):398–416, November 2000.

1.39 R.D. Stevens, C.A. Goble, P. Baker, and A. Brass. A Classification of Tasks in Bioinformatics. *Bioinformatics*, 17(2):180–188, 2001.

1.40 The Gene Ontology Consortium. Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*, 25:25–29, 2000.

1.41 UDDI. UDDI Technical White Paper, September 2000. Available: `http://www.uddi.org`.

1.42 Chris Wroe, Robert Stevens, Carole Goble, Angus Roberts, and Mark Greenwood. A Suite of DAML+OIL Ontologies to Describe Bioinformatics Web Services and Data. Accepted for publication in the International Journal of Cooperative Information Systems, 2003.

1.43 C.J. Wroe, R.D. Stevens, C.A. Goble, and M. Ashburner. A Methodology to Migrate the Gene Ontology to a Description Logic Environment Using DAML+OIL. 8th Pacific Symposium on biocomputing (PSB), 2003.