BIOINFORMATICS

Vol. 15 no. 6 1999
Pages 510–520

An ontology for bioinformatics applications

Patricia G. Baker[1], Carole A. Goble[2], Sean Bechhofer[2], Norman W. Paton[2], Robert Stevens[2] and Andy Brass[1]

[1]School of Biological Sciences and [2]Department of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PT, UK

Received on September 17, 1998; revised and accepted on February 3, 1999

Abstract

Motivation: *An ontology of biological terminology provides a model of biological concepts that can be used to form a semantic framework for many data storage, retrieval and analysis tasks. Such a semantic framework could be used to underpin a range of important bioinformatics tasks, such as the querying of heterogeneous bioinformatics sources or the systematic annotation of experimental results.*

Results: *This paper provides an overview of an ontology [the Transparent Access to Multiple Biological Information Sources (TAMBIS) ontology or TaO] that describes a wide range of bioinformatics concepts. The present paper describes the mechanisms used for delivering the ontology and discusses the ontology's design and organization, which are crucial for maintaining the coherence of a large collection of concepts and their relationships.*

Availability: *The TAMBIS system, which uses a subset of the TaO described here, is accessible over the Web via http://img.cs.man.ac.uk/tambis (although in the first instance, we will use a password mechanism to limit the load on our server). The complete model is also available on the Web at the above URL.*

Contact: *tambis@cs.man.ac.uk*

Introduction

Biology is a knowledge-based discipline. Many predictions and interpretations of data in biology are made by comparing the data in hand against existing knowledge, e.g. the problem of predicting protein function from sequence. This is typically done by asking whether the unknown sequence resembles a well-characterized protein. The function of the unknown sequence can then be inferred from the type of similarities found. Similarly, it is often possible to predict the structure of a protein from its sequence using knowledge of known protein structures and asking which known protein structure, if any, could sensibly represent the structure of the unknown protein. The key difference, therefore, between 'knowledge-based' and 'axiomatic' disciplines is the role played by the knowledge base of past experience. The challenge and the skill in biology is often to make use of this knowledge in the most effective way.

Traditionally, the knowledge base in biology has resided within the heads of experienced biologists—scientists who have devoted much study to becoming experts in their particular domain of study. This approach worked well in the past when considerable effort was needed to tease new data out of biological experiments—the flow of data was not so great as to overwhelm the expert. However, this situation is changing rapidly—many complete genomes are appearing each year (Cole *et al*., 1998) and new experimental techniques are providing information on interactions. For example, a single experiment can now yield data on the transcription level of 100 000 different mRNA species from a given tissue (Winzeler *et al*., 1998). Therefore, not only is the rate of data acquisition growing exponentially, but also a single experiment can collect data on a huge range of molecules that would need an army of domain experts to be interpreted. This is proving to be a serious handicap to a knowledge-based discipline. Good predictions can only be made against a knowledge base, and the bigger the knowledge base, the better the predictions that can be made. However, the size of the existing knowledge base is too large for any human to assimilate. Therefore, predictions are only being made against a small subset of the available knowledge, and information is being neglected.

There is, therefore, a need to create systems that can apply the knowledge in the head of a domain expert to biological data. It is not envisaged that such systems could ever perform better than human experts; however, they could play a crucial role in filtering the flood of data to the point where human experts could again apply their knowledge sensibly. This then raises numerous questions, in particular regarding how concepts and their relationships can be captured in ways that make them computationally available and tractable.

An ontology is a system that describes concepts and the relationships between them. Therefore, what we would like to do is to build an ontology for the bioinformatics domain. It is important to point out that this will just be one of many possible ontologies for biology. A considerable body of research in the area of knowledge representation has shown that an ontology must necessarily reflect a specific view of the data (Gruber, 1995). Consider, for example, the concept of protein. From a bioinformatics perspective, it is clear that

the idea of an accession number should be associated with a protein—it is the key to retrieving information about a protein from sequence databases. However, it probably makes no sense to talk about accession number as an attribute of real proteins in an ontology built to describe the biochemistry of the cell.

In the present paper, we have investigated the use of a particular form of knowledge representation system, Description Logics (DLs), and argue that: (i) DLs are flexible and powerful enough to capture and classify biological concepts in a consistent and principled fashion; (ii) DLs can be used to construct ontologies that can be used for making inferences from biological data.

Description logics and ontologies

Ontologies have been developed in the artificial intelligence community to describe a variety of domains, and have been suggested as a mechanism to provide applications with domain knowledge and to facilitate the sharing of information. The importance of ontologies has been recognized within the bioinformatics community (Schulze-Kremer, 1998), and work has begun on developing and sharing biomolecular ontologies (ISMB Workshop, 1998).

In order to support these activities successfully, the representation used for the ontology must be rich enough in terms of the services it offers, and should have a consistent interpretation.

Traditionally, ontologies have been represented using static models (Schulze-Kremer, 1998). These can assist in the exchange of knowledge at a purely terminological or syntactic level, but can suffer due to the difficulties of interpretation—the relationships in the model rely solely on the perspective of the modeller. If we are to share knowledge, a clearer semantics is required. Full interaction with an ontology requires, in addition, a notion of the range of services, functionality or reasoning the ontology can provide.

Frame representations provide a precise, definitional framework in which to capture concepts and the relationships between them. The Frame formalism has been used to model biological data in the EcoCyc encyclopedia of *Escherichia coli* genes and metabolism (Karp *et al*., 1998). Specifications of interfaces describing the services offered by frame systems have been defined (Chaudhri *et al*., 1998). The representation is, however, static and all subsumption is asserted, in the sense that the kind-of hierarchy is asserted by the modeller, rather than deduced by the system from the descriptions of concepts.

Knowledge bases have also been used to retrieve information automatically from the literature on ribosome structure to provide constraints for predicting the organization of the ribosome complex (Chen *et al*., 1997).

DLs (Borgida, 1995) are a further example of a knowledge representation language. DLs provide a language for capturing declarative knowledge about a domain and a classifier that allows reasoning about that knowledge. Information captured using DLs is classified in a rich hierarchical lattice of concepts and their interrelationships. DLs are compositional and dynamic, relying heavily on the notion of services for classification, subsumption, consistency and retrieval or querying (KRSS, 1993). This means that new concepts can be constructed from existing concepts, and automatically and precisely placed in the lattice.

DLs have not, until now, been used to model the biological domain, although they have been used in a number of non-biological (Arens *et al*., 1993; Borgida, 1995) and medical applications, including the GALEN project (Rector *et al*., 1995; Rogers *et al*., 1997). The choice of a DL as the representation language was motivated partly by the success of these previous approaches, particularly the work of the GALEN project. The compositional nature and dynamic classification reasoning services are ideally suited to modelling aspects of the biological domain. In addition, the infrastructure required to support this effort in terms of implementations of terminological reasoners, modelling tools and user interfaces was present.

The GRAIL concept modelling language

The GRAIL language (Rector *et al*., 1997), used to describe biological concepts in this paper, is a DL in the KL-ONE family (Woods and Schmolze, 1992) that was originally developed to allow the modelling of medical terminology for a system to support clinical user interfaces. This section gives a brief description of GRAIL's major characteristics.

A DL models an application domain in terms of concepts (classes), roles (relations) and individuals (objects). The domain is a set of individuals, and a concept is a description of a group of individuals that share common characteristics. Roles model relationships between, or attributes of, individuals. Compositional concept descriptions can then be built up using recursive term constructors, where terms are concepts or roles. Individuals can be asserted to be instances of particular concepts, and pairs of individuals can be asserted to be instances of particular roles. All roles in GRAIL are bidirectional.

For example, Protein is a class of individuals—all proteins—and is thus modelled as a concept. An example of an instance of a protein is human α haemoglobin. Proteins can have components, for example Motifs, and we represent this through a binary role hasComponent. We can then form new concept descriptions, say Protein which hasComponent Motif, or Motif which isComponentOf Protein. An example instance of the latter is a haem binding site, which we know

is a Motif and is also a component of a protein, in this case human α haemoglobin.

A GRAIL model can be considered to consist of three parts: (i) assertions; (ii) concept-forming operations and reasoning services; (iii) sanctions.

Assertions

A model contains a collection of elementary concept definitions along with a collection of roles. Elementary concept definitions are simple, atomic concepts (such as Motif or Protein) which cannot be decomposed further.

Operations and reasoning services

GRAIL provides a collection of operations which allow the construction of compositions of concepts and roles. This composition is provided along with a collection of reasoning services which allow us to make inferences.

Central to the reasoning is the notion of classification, which infers the precise hierarchical position of a composite definition. Concept A is said to subsume concept B precisely when all instances of B are also instances of A. Concepts can be classified in a hierarchy based on this subsumption or kind-of relationship. Elementary concepts have their position in the concept hierarchy asserted by the modeller explicitly stating that it is a kind-of an existing concept. However, composite concepts are precisely classified automatically based on their definition.

For example, the elementary concepts Motif and Protein can be combined using the role isComponentOf to produce the complex concept Motif which isComponentOf Protein. The GRAIL classifier places this composite concept below Motif in the hierarchy. This contrasts with static representations, where the composite would need to be explicitly placed by the modeller if it appeared in the model at all. If this concept were made more specific by combination with further concepts, the GRAIL classifier would automatically reclassify it. If the specialization hasModification PostTranslationalModification was also applied to Motif, the complex concept would become:

Motif which < isComponentOf Protein hasModification PostTranslationalModification >

GRAIL supports multiple inheritance, allowing this concept to be classified as a kind-of Motif which isComponentOf Protein and a kind-of Motif which hasModification PostTranslationalModification. This property of concepts being classified with many parents makes classification in a DL very different from a more traditional taxonomic classification, in which concepts are organized in a tree-like structure and every concept can only have one parent. As a result, DLs are more flexible than taxonomic classifications and can naturally support multiple views of the same concept, as demonstrated in the example above.

The ability to create concepts by combining existing concepts is termed compositionality. The compositional nature of GRAIL allows an alternative and more powerful means of creating new concepts than by explicit subsumption, and means that a large number of concepts can be created from a relatively sparsely populated model. The use of such a model is inextricably bound up with notions of services and reasoning: a GRAIL model is not a static tree, but should be considered as a resource that can be queried by applications.

DLs have a well-defined semantics which allows the consistent interpretation of subsumption. When a composite definition is classified and placed in the hierarchy, we know that this position is based on well-founded reasoning. This contrasts with hand-crafted ontologies, where the position of a concept is purely dependent on the modeller (Schulze-Kremer, 1998). Of course, the assertional part of the model is still built by hand, should be based on sound underlying principles and requires verification. However, the composed definitions will have a coherent and consistent organization.

An asserted hierarchy along with reasoning services concerning the classification of composite descriptions are standard to DLs, and provide what is often described as T-Box reasoning. DLs may also provide mechanisms for making assertions about particular individuals or instances along with corresponding reasoning services (e.g. retrieval). This is known as A-Box reasoning. In the example above, the T-Box would encompass reasoning about Proteins, Motifs and so on, while the A-Box would allow reasoning over the instances such as haemoglobin or phosphorylation site.

Sanctions

To restrict the construction of complex concepts to only those that are semantically meaningful, GRAIL provides rules or sanctions that dictate which roles may legitimately be applied to which concepts. Sanctioning is a mechanism unique to GRAIL; in other DLs, mechanisms such as role restriction are used to produce similar results. The philosophy is that a composition is not allowed unless it is explicitly sanctioned. However, sanctions are inherited, allowing the modeller to decorate the model at a high level, with the constraints filtering down. In order to provide greater flexibility and control, two levels of sanctioning are provided, known as grammatical and sensible. Grammatical sanctions express abstract or general relationships between classes of things, whereas sensible sanctions indicate that instantiable compositions can be built. A grammatical sanction must be in place before a sensible sanction can be made. Sanctioning relies on the classification, but is a separate operation that can be thought of as being layered on top of, and which uses, the classification hierarchy.

Figure 1 shows the sanctioning of the relationship hasComponent at the grammatical and sensible levels. The rela-

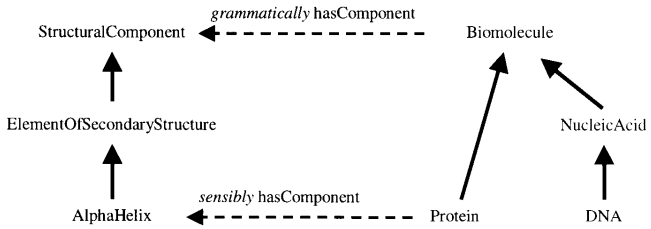

Fig. 1. An example of the use of two levels of sanctioning of a relationship between two concepts. The solid arrows indicate kinds-of relationships and dashed arrows indicate the non-subsumptive relationship, hasComponent. Relationships between concepts are bidirectional, so the reverse relationship isComponentOf is also sanctioned (although this is not shown in the figure).

tionship between the concepts Biomolecule and StructuralComponent is sanctioned at the grammatical level because it is grammatically permissible to speak of biomolecules having structural components, but not all kinds-of biomolecule can legitimately have any kind-of structural component. The solid arrows in Figure 1 show kind-of relationships. So Protein is a kind-of Biomolecule and an AlphaHelix is a kind-of StructuralComponent. The hasComponent relationship between the concepts Protein and AlphaHelix is sanctioned at the sensible level because any kind-of protein could legitimately have an α helix. However, not all proteins will have α helices—sanctioning is about representing the possibility of composition, not its necessity. This is a powerful mechanism to keep models sparse and compact, but which does require skill from the ontologist. Care has to be taken to apply the sensible sanction at the appropriate level; applying it to a relationship between concepts too high up in the hierarchy will allow the construction of biologically incorrect concepts. For example, Figure 1 shows that DNA is a kind-of Biomolecule and a AlphaHelix is a kind-of StructuralComponent. Sanctioning the hasComponent relationship between Biomolecule and StructuralComponent at the sensible level would allow the obviously incorrect concept DNA which hasComponent AlphaHelix to be built. Thus, the usual attendant verification and validation procedures required on all ontologies apply here (Guarino, 1998).

Although grammatical sanctions on their own do not permit the construction of instantiable composite definitions, they do represent valid queries that may be formed. In the above example, asking for all Biomolecules which have some StructuralComponent is a valid question.

Deciding on the appropriate position for sanctions or constraints is a challenging process. If sanctions are placed high up in the hierarchy, the effect may be to sanction compositions lower down which are in some way less meaningful. The concepts will be correctly classified, but the composition does not make sense (e.g. cDNA which hasComponent RibosomeBindingSite). This is a problem that occurs in many representations supporting composition.

The biological correctness of sanctions cannot be tested automatically, in the same way that the biological validity of the model is in many respects subjective—such models can only really be evaluated and verified through their use. As far as is possible, however, the model is checked to ensure that any concepts that can be built are biologically 'reasonable'. Tools are provided which assist in this process (Solomon, 1998), including a generation tool, which 'fills out' areas of the model based on the sanctions, allowing the modeller to see the ramifications of any new sanctions added.

The composition operation taken together with sanctioning provide a powerful mechanism which allows us to generate or infer concepts based on existing definitions without having to define everything pre hoc. For example, Motif has a child Site and this has many children, including Phosphorylation site and Methylation site. Site is sensibly sanctioned to be a component of Protein, so all its children are also allowed to be components of Protein. In this way, all the combinations of site and protein are available to be made and need not be made explicitly as part of the model, contrasting with a static hierarchical approach, where all combinations would have to be introduced explicitly.

Using GRAIL thus allows us to compose and extend a basic asserted hierarchical model in a coherent and well-founded way with the GRAIL classifier taking care of the maintenance of the conceptual hierarchy.

Ontologies should be seen not just as static hierarchies, but as resources providing services. This is particularly important given the current vogue for component-based technologies. GRAIL models are delivered through the use of a software component known as a terminology server (TeS) (Bechhofer *et al*., 1997). This is a component that provides a programming interface to the ontology such that applications can ask whether a concept classifies and ask questions about related concepts. For example, if the TeS is provided with a concept that classifies properly, it can return information about parent, child or sibling concepts and the attributes that can be attached to those concepts. This use of a service model for the delivery of conceptual models is in line with current thinking (Farquhar *et al*., 1996; Chaudri *et al*., 1998).

TAMBIS bioinformatics ontology

Transparent Access to Multiple Biological Information Sources (TAMBIS) (Baker *et al*., 1998) is a research project which aims to aid researchers in biological science by providing a single access point for biological information sources round the world. This is achieved through the use of a mediating ontology. Queries are phrased in terms of the

ontology and the TAMBIS system converts these to requests to the appropriate sources.

The aim of the TAMBIS ontology (TaO) is thus to capture biological and bioinformatics knowledge in a logical conceptual framework that is constrained in such a way that (i) only biologically sensible concepts classify correctly, (ii) it can encompass different user views and (iii) it makes biological concepts and their relationships computationally accessible.

The primary purpose of the TAMBIS system is to allow biologists to describe data they wish to recover from bioinformatics sources. Therefore, the model was designed to enable concepts to be described that cover the questions biologists wish to ask and those that can be asked of the sources. A survey of questions actually asked by biologists was used to aid in the construction of the TaO. The TaO is only one possible description; other, equally suitable models could be designed. This raises the issue of interoperation between ontologies. Our terminology server architecture provides access to the terms of the model in a consistent manner, easing the technological problems with interoperation. At the semantic level, however, the identification of relationships between terms is still a matter for human intervention. The consistent classification supported by the DL reasoning should facilitate this process.

DLs, and GRAIL in particular, have known limitations; for example, GRAIL offers limited support for cardinality constraints on roles. Similarly, DLs fail to support query expressions involving shared variables, although this is a topic of interest in the DL research community (Calvanese *et al.*, 1998). Concrete domains, such as numerical values, are another aspect which are poorly supported in DLs. When building applications, however, there is the opportunity to add mechanisms that deal with such things as user instantiation and ranges of numeric values. The appropriateness of any model can only be judged in pragmatic terms—the release of the TAMBIS system will allow us to judge the efficacy of this particular model and representation.

The principal role of the ontology is to describe biological concepts and their use in bioinformatics. This can be achieved by linking concepts together by their sanctioned relationships with other concepts. Therefore, the main considerations when building the model were:

- Which categories should a concept be placed in to cover all the ways in which it may be used?
- At what level should sanctions be applied to ensure that both generally applicable and biologically sensible relationships may be made in different 'is a kind of' hierarchies?

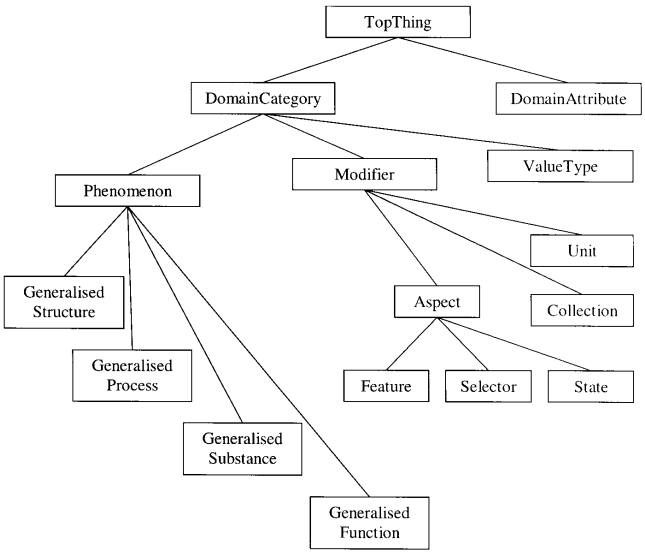

Fig. 2. High-level divisions in the biological concepts hierarchy. The joining lines denote kind-of relationships.

The TaO can be divided into two parts. The high-level divisions are taken from the models developed in the GALEN project and described in Rector *et al.* (1996). This general foundation has been extended in TAMBIS with the lower level concepts necessary to represent user's descriptions in the biological domain.

The high-level divisions

The ontology presented in Figure 2 represents the high-level, generic divisions of the model's elementary concepts. These high-level divisions are intended to group and organize any domain's concepts in an intuitive way. Below this point, domain-specific knowledge can be added.

The first division in the hierarchy is into DomainCategory and DomainAttribute, corresponding, lower in the model, to biological concepts (things) and roles (relationships). These divisions are considered in more detail in the following two sections.

The concepts hierarchy

The top-level divisions of Phenomenon, Modifier and ValueType are based on the GALEN high-level ontology. The major categories under Phenomenon and Modifier are summarized and explained in Table 1, together with examples of lower level concepts from the biological domain.

Table 1. A summary of high-level concepts in the Domain Category of the TAMBIS biological model. The left-hand column shows the five major divisions and their immediate subdivisions. The right-hand column shows biological examples from lower in the hierarchy

Entity	Description or example
Generalized Structure	*Discrete abstract or physical things independent of time*
Abstract Structure	
protocol	
method	alignment method, comparison method
classification	taxonomic classiciation (NCBI), structural classification
reaction	enzymic reaction
pathway	metabolic pathway
information source	database, user
Physical Structure	
solid structure	cell, organelle, liver
biomolecular structure	secondary structure, tertiary structure
component of biomolecular structure	motif, domain, site
chemical	protein, nucleic acid, sodium, drug
Generalized Substance	*Continuous abstract or physical things independent of time abstract or physical things independent of time*
Body Substance	
body fluid	blood
tissue	muscle, liver
Cellular Substance	cytoplasm
Generalized Process	*Transformations which occur over time*
Biological Process	
body process	disease process, clotting
biomolecular process	transcription, cleavage, metabolism
Generalized Function	*Roles or purposes independent of time*
Biological Function	
MolecularModification	cleavage, glycosylation
Binding	nucleic acid binding
DNA Replication and Repair	
Transport	transmembrane transport, electron transport
Maintenance of Structure	
Cellular Growth and Profileration	
Signal Transduction	
Enzymic Function	transferase
Receptor	
Hormone	
Toxin	
Inhibitor	
Targetting	
Stress	
Modifier Concept	
Aspect	*adjectival expressions which refine the meaning of other concepts*, e.g. internal/external selector, physicochemical property
Collection	*concepts which take their meaning from another category but which are fundamentally different from that category*, e.g. alignment, complex

Structures and substances. The division of Generalized Structure into Physical and Abstract Structures represents the distinction between discrete physical phenomena and the structured representations of those phenomena. For example, a protein is considered to be a physical entity, but its structural classification is an abstract entity. Generalized Substance is a relatively small category with divisions that require no further explanation. The category Body Substance contains many concepts which are also classified under Physical Structure, namely body organs, which can

be viewed as being solid, bounded objects or as types of tissue. Similarly, chemicals can be viewed as discrete 'things' (e.g. a protein molecule) or as continuous 'stuff' (e.g. some protein). However, in the context of bioinformatics, it is unlikely that the user would want to use the latter interpretation. Concepts from the Structures and Substances categories allow concepts to be constructed concerning, for example, the structure of proteins, the cellular or molecular site of action of enzymes, and reactions occurring in a given metabolic pathway.

Processes and functions. There is a large degree of overlap between members of the GeneralizedFunction and GeneralizedProcess categories. For example, the concept of 'transcription' can be viewed as (i) a function, taking DNA and some amino acids as input and producing a protein as output, or (ii) as a process, occurring over time, with the transcriptional machinery moving along the DNA, and so on.

The question, therefore, is: 'Do we care that in real, physical terms transcription is a process?'. In situations where the classification is unclear, modelling is guided by the manner in which the terms may be used. For example, one may reasonably make the queries 'find all proteins that have the function transcription' or 'find all proteins that function in the process transcription', and interpret both to have the same meaning. In this situation, where a concept may be used in different ways, it is placed into multiple categories. This example demonstrates the ability of the DL representation to support a concept being viewed in different ways.

Modifiers. The Modifier category contains adjectives or adjectival expressions used to describe phenomena. Classification in this category is straightforward and unambiguous. The two high-level divisions under Modifier are Aspect and Collection. Concepts in the Aspect category allow us to describe properties of other concepts such as the MolecularWeight of a protein or the Length of a DNA sequence. Also included under Aspect is Selector, a category containing concepts such as Internal and External, Male and Female. These are mutually exclusive adjectives that can be used to refine a concept's description. Concepts in the Collection category allow us to describe concepts in their collective state, e.g. an Alignment of nucleic acid sequences or a Complex of proteins.

The attributes hierarchy

GRAIL's expressivity with respect to cardinality is limited compared to other DLs. GRAIL allows a specification that a role can be filled by one or any number of concepts. Other DLs have more sophisticated number restrictions (Borgida, 1995). Attributes are arranged in a hierarchy which provides both a means of specializing or generalizing concepts and a means of creating equivalent relationships with different cardinalities, e.g. the relationship isComponentOf which has a manyMany cardinality is the parent of isSpecificComponentOf which has a manyOne cardinality.

The top-level divisions of DomainAttribute are:

- ConstructiveAttribute—relationships which exist between abstract or physical things or processes.
- ModifierAttribute—relationships that exist between abstract or physical things or processes and the modifier concepts that refine their meaning.

The major categories under ConstructiveAttribute are explained below:

- CollectionAttribute—relationships between concepts and their parts or multiples (e.g. Alignment which isAlignmentOf Protein).
- FunctionalAttribute—relationships between either process/function and physical/abstract things (e.g. Metabolism which isMetabolismOf Thymine) or between physical/abstract things where the relationship itself is a process/function (e.g. Enzyme which cleaves CleavageSite).
- LocativeAttribute—relationships between things or processes and their physical location (e.g. Gene which isExpressedIn Liver).
- StructuralAttribute—relationships between physical structural concepts and biomolecules (e.g. Protein which hasStructuralClassification AllAlpha, TertiaryStructure which isStructureOf Protein).

The one category under ModifierAttribute is explained below:

- SelectorAttribute—relationship between physical/abstract things or processes and the 'selector' modifiers applied to them (e.g. Domain which hasInternalExternalSelector External).

The high-level relationships within the categories listed above are summarized in Table 2 with more examples of lower level relationships from the biological domain.

Current status of the bioinformatics ontology

The TAMBIS ontology was designed to support descriptions of both retrieval and analysis tasks. To achieve this, the ontology has been made broad and shallow. The breadth, as indicated in the discussion of the GALEN ontology and Tables 1 and 2, means that a wide variety of descriptions or queries can be formed. Obviously, most bioinformatics tasks centre about proteins and nucleic acids (and their various children: DNA, RNA, Gene, Enzyme, etc.) and things that can be said about those core concepts in the bioinformatics sources. Table 2 gives examples of many of the kinds of roles that can be used to link concepts together in descriptions. For example, most of the attributes annotated in a SWISS-PROT entry can be described in the TAMBIS ontology. Deciding on the depth of the model is somewhat

more problematic. At present, the model is quite shallow (see below), reflecting both the ability of a source to answer detailed queries and the difficulty in modelling ill-defined sources. For example, the concept of 'biological function' is only specialized to the level of 'receptor' or 'secretion'. This is because the PROSITE documentation (Bairoch *et al*., 1997) only describes function to this level, rather than the belief that users will not wish to ask more specific questions. CATH (Orengo *et al*., 1997) has a detailed classification of protein structure, but some of the lower classes have labels derived from representative examples of the class. Thus, mapping to an abstract class name is difficult. More detail, however, will be added as users demand detail in particular areas. Table 3 shows some of the major concepts in the model along with examples of the leaves of the subsumption hierarchies beneath them.

Table 2. A summary of high-level concepts in the Domain Attributes hierarchy of the TAMBIS biological model. The left-hand column shows the five major divisions and their immediate subdivisions. The right-hand column shows biological examples from lower in the hierarchy

Entity	Example
CollectionAttribute	
MultipleAttribute	isComplexOf, isAlignmentOf
PartitiveAttribute	isComponentOf, isSubProcessOf
FunctionalAttribute	cleaves, isCatalyzedBy, hasFunction
TransformationAttribute	isTranslatedTo, codesFor
MethodAttribute	hasComparisonMethod, hasAlignmentMethod
LocativeAttribute	
PhysicalLocativeAttribute	hasCellularLocation, hasMembraneLocation
ProcessLocativeAttribute	operatesInTissueType, operatesInOrganismType
StructuralAttribute	isCompositionOf
ArchitecturalAttribute	isSequenceOf, isSecondaryStructureOf
ClassificationalAttribute	isStructuralClassificationOf, hasSourceSpecies
SelectorAttribute	hasInternalExternalSelector, hasDoubleSingleStrandSelector, hasUpperLowerSelector

Table 3. The level of detail in the TAMBIS ontology shown by some core bioinformatics concepts and a selection of examples of concepts at the terminus of the asserted 'is a kind of' hierarchy under those core concepts. There will be many other leaves not shown in this table

Concept	Leaf Child(ren)
Motif	Phosphorylation Site
Protein	Enzyme
Biological Function	Receptor, Secretion
Enzymic Function	Oxidoreductase, Transferase
Biological Process	Lactation, Endocytosis
Domain	Propellor, Prism, Barrel

The ontology currently contains around 1800 asserted concepts. The concepts covered and the sources with which they are associated are shown below, along with examples of GRAIL constructs in which the concepts are used.

- Protein and protein sequence [from SWISS-PROT (Bairoch *et al*., 1996)], protein component motifs [from PROSITE (Bairoch *et al*., 1997)], protein structure [as classified by CATH (Orengo *et al*., 1997)] and enzyme function [as defined in Prosite, and the Enzymes and Metabolic Pathways database—EMP (Selkov *et al*., 1996)]. We can therefore build concepts such as the 'tertiary structures of proteins which contain motifs that are involved in hydrolase activity':
  TertiaryStructure which isStructureOf (Protein which hasComponent (Motif which indicatesFunction Hydrolase))
- Enzymes and metabolic pathways [as defined in the Enzyme database (Bairoch, 1996)]. This allows the construction of queries regarding enzymes and their reactions, e.g. enzymes which catalyse reactions which occur in the metabolism of thymine.
  Enzyme which catalyses (Reaction which occursIn (Metabolism which isMetabolismOf Thymine))
- Expressed sequence tags (ESTs) [as defined by dbEST (Boguski *et al*., 1993)]. We can therefore create the concept of ESTs that code for proteins that contain glycosylation sites.
  EST which codesFor (Protein which hasComponent GlycosylationSite)
- Nucleic acids, their component motifs, gene function and expression (Stoesser *et al*., 1997, 1998). The concept given below should be relatively self-explanatory.
  Gene which codesFor (Protein which hasFunction TransmembraneTransport)
- Sequence homology [BLAST (Altschul *et al*., 1990)]. Using ideas of homology, we can create concepts linked to specific bioinformatics processes, e.g. the concept of the set of proteins homologous to a protein with a specific accession number.
  Protein which isHomologousTo (Protein which hasAccessionNumber P12345)
- Taxonomy [as defined at the NCBI web site (NCBI) http://www.ncbi.nlm.nih.gov/Taxonomy].
  TaxonomicRank which < isRankOf PoeciliaReticulata isRankOf AmoebaProteus >
  i.e. the taxonomic rank common to both *Poecilia reticulata* and *Amoeba proteus*.

Applications of a bioinformatics ontology

The aim of this work was to provide an ontology that could help underpin the development of systems that perform at least some of the functions of a domain expert. In general

terms, these functions amount to knowing (i) what things are in the domain and (ii) when and how these things are related. An ontology by itself is not very useful, so two software components have been created that allow the ontology to be queried, explored and used as a component by other programs. The first is a graphical user interface that allows users to explore the ontology and construct ad hoc concepts (Bechhofer and Goble, 1997). The second is the terminology server discussed above. The server can be accessed both locally and in a distributed fashion, opening up the possibility of use by third parties.

In order to evaluate the effectiveness of the ontology, we need to assess how it can be used to support a range of tasks that we might expect a domain expert to undertake. We have therefore explored the task of describing the information stored in different biological data repositories and allowing complex queries to be posed against this distributed data set.

The ontology described here has been constructed as part of the TAMBIS project and provides the TAMBIS user with the concepts necessary to construct complex queries. The ontology is used to facilitate integration of heterogeneous data sources, acting as a broker between them and the user. Much of the content has been derived from the sources' schemas, leading to a broad and shallow ontology. Complex queries are phrased against the ontology rather than against individual sources. The ontology mediates between the underlying sources, reconciling mismatches such as semantic differences and differences in the levels of abstraction to which data are held. The TAMBIS system has been described more fully elsewhere (Baker *et al*., 1998). However, within TAMBIS, the ontology plays a key role in guiding the user to create sensible queries, and then in providing information to other parts of the system to help find and instantiate examples of the concept created. GRAIL does not support an A-Box; in the TAMBIS application, retrieval is through a rewriting process rather than through A-box reasoning. In its current status, the TAMBIS prototype contains rewrite rules covering a small subset of the complete TAMBIS ontology.

The TAMBIS system has been released as an application on the Web for selected users as part of an evaluation exercise since March 1999.

In an additional application of the ontology, a simple test has been made to check the taxonomic information contained within the SWISS-PROT database. The ontology was used to generate the full taxonomic lineage for the species in a SWISS-PROT entry and this structure was checked against the taxonomic structure reported in the database annotation files. Any discrepancies identified between the official SWISS-PROT taxonomy (as captured within a version of the ontology) and a database entry were reported.

Shown below is a part of the annotation from a SWISS-PROT file describing *Mus musculus*. Not only is the species name given, but so are all the nodes going up the taxonomic tree to its root (Eukaryota).

```
OS MUS MUSCULUS (MOUSE).
OC EUKARYOTA; METAZOA; CHORDATA;
   VERTEBRATA; TETRAPODA; MAMMALIA;
OC EUTHERIA; RODENTIA.
```

To test these entries, an application was written which took the last term from the OC line (in this case Rodentia). This term was then sent to the TeS and checked to see that it properly classified. The TeS was then asked to generate all parent nodes of the concept. Asking for such a lineage, by repeatedly asking for a parent of each successive node, is a key terminology service, and so is already present in the system. These generated lineages were then compared against the list of terms in the OC line. The program uncovered a small number of errors, typically where the sequence entry was using an older version of the published SWISS-PROT classification scheme and the OC line had not been updated to reflect the most recent version. This example, though simple, indicates that the services available within a DL TeS mean that it is possible and beneficial to re-use an ontology (originally built for TAMBIS) in another application.

Future work

In this paper, we have shown that it is possible to use DLs to produce a rich ontology of the bioinformatics domain. Examples have been given to show such an ontology in use. However, there are other ways in which an ontology could be used to make bioinformatics resources more effective.

As another example, sequence database annotation is currently provided at the textual/keyword level. Although this information is convenient for human readers, it does not lend itself to being interpreted computationally. Ontologies could provide a semantic framework for sequence annotation which would allow more effective data submission. Using the ontology as a means of describing new sequences would provide a rigorous and consistent means of sequence annotation. A newly submitted sequence is described in terms from the ontology and is, hence, classified in the hierarchy. Such an annotation would be sensible, consistent and, using the TeS, would be machine interpretable. Ontologies would, therefore, also allow for more effective information retrieval and analysis.

The functional similarity between hits from a similarity search is often not apparent from the output of the search. By reference back to a structured representation of the annotation of those sequences, any common features can be seen. As a simple example, consider a BLAST search that produced top hits that were either calcium, magnesium or iron binding proteins. Reference to a terminological model would show that the common characteristic of these proteins is that they are all metal binding. Other, more subtle relationships

could easily be missed without reference to a conceptual model. The ontology can also be used to cluster sequence data based on a variety of characteristics (e.g. source organism classification or tissue expression). Ontologies provide a powerful mechanism for making conceptual information about biology computationally available. Ontologies therefore provide one mechanism by which conceptual information can be attached to the current flood of biological data and thereby help turn data into useful biological knowledge.

Acknowledgements

The work was funded under the BBSRC/EPSRC Bioinformatics Initiative and by Zeneca Pharmaceuticals, whose support we are pleased to acknowledge. We would also like to thank the anonymous reviewers for their valuable comments on an earlier draft of this paper.

References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Arens,Y., Chee,C.Y., Hsu,C.-N. and Knoblock,C.A. (1993) Retrieving and integrating data from multiple information sources. *Int. J. Coop. Inf. Syst*., **2**, 127–158.

Baader,F., Bürckert,H-J. Heinsohn,J., Hollunder,B., Müleer,J., Nebel,B., Nutt,W. and Profitlich,H.-J. (1991) Terminological knowledge representation: a proposal for a terminological logic. Deutches Forschungszentrum für Künstliche Intelligenz (DFKI) Technical Memo TM-90-04.

Bairoch,A. (1996) The Enzyme data bank in 1995. *Nucleic Acids Res*., **24**, 221–222.

Bairoch,A. and Apweiler,R. (1996) The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res*., **24**, 21–25.

Bairoch,A., Bucher,P. and Hofmann,K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res*., **25**, 217–221.

Baker,P.G., Brass,A., Bechhofer,S., Goble,C.A., Paton,N.W. and Stevens,S. (1998) TAMBIS—Transparent Access to Multiple Bioinformatics Information Sources. In *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, pp. 25–34.

Bechhofer,S. and Goble, C.A. (1997) Using description logics to drive query interfaces. DL97, International Workshop on Description Logics, Gif sur Yvette.

Bechhofer,S., Goble,C.A., Rector,A.L., Solomon,W.D. and Nowlan,W.A. (1997) Terminologies and terminology servers for information environments.In *Proceedings of the IEEE 8th International Conference on Software Technology and Engineering Practice*. London, pp. 35–42.

Borgida,A. (1995) Description logics in data management. *IEEE Trans. Knowledge Data Eng*., **7**, 671–682.

Buguski,M.S., Lowe,T.M.J. and Tolstoshev,C.M. (1993) dbEST—database for expressed sequence tags. *Nature Genet*., **4**, 332–333.

Calvanese,D., De Giacomo,G., Lenzerini, M., Nardi,D. and Rosati,R. (1998) Description logic framework for information integration. In *Proceedings of the 6th International Conference. on the Principles of Knowledge Representation and Reasoning (KR-98)*, pp. 2–13.

Chaudhri,V.K., Farquhar,A., Fikes,R., Karp,P.D. and Rice,J.P. (1998) Open knowledge base connectivity 2.0.3. Stanford Knowledge Systems Laboratory Report: KSL-98-06.

Chen,R.O., Felciano,R. and Altman,R.B. (1997) RiboWeb: linking structural computations to a knowledge base of published experimental data. In *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, pp. 84–87.

Cole,S.T. *et al*. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.

Farquhar,A., Fikes,R. and Rice,J. (1996) The Ontolingua Server: a tool for collaborative ontology construction. Stanford Knowledge Systems Laboratory Report: KSL-96-26.

Gruber,T.R. (1995) Towards principles for the design of ontologies used for knowledge sharing.

Guarino,N. (ed.) (1998) *Formal Ontology in Information Systems*. IOS Press.

Guha,R.V., Lenat,D.B., Pittman,K., Pratt,D. and Shepherd,M. (1990) CYC: A Midterm Report. *Commun. ACM*, **33**.

ISMB Workshop (1998) *Semantic Foundations for Molecular Biology Schemata, Controlled Vocabularies and Ontologies*. Workshop at the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB-98).

Karp,P., Riley,M., Paley,S. and Pellegrini-Toole,A. (1998) Ecocyc: Electronic Encyclopedia of *E.coli* genes and metabolism. *Nucleic Acids Res*., **26**, 50.

KRSS Group of the ARPA Knowledge Sharing Effort (1993) Description-logic knowledge representation system specification. Available from http://www-db.research.bell-labs.com/user/pfps/papers/krss-spec.ps

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.S. and Thornton,J.M (1997) CATH: a hierarchic classification of protein domain structure. *Structure*, **5**, 617–634.

Rector,A.L., Zanstra,P., Solomon,W.D. and the GALEN Consortium (1995) GALEN: terminology services for clinical information systems health in the new communications age. In Laires,M.F., Ladeira,M.J. and Christensen,J.P. (eds), *Health Technology and Informatics Vol. 24*. IOS Press.

Rector,A.L., Rogers,J.E. and Pole,P. (1996) The Galen high level ontology. *Stud. Health Technol. Informatics*, **34**, 174–178.

Rector,A.L., Bechhofer,S., Goble,C.A., Horrocks,I., Nowlan,W.A. and Solomon,W.D. (1997) The GALEN modelling language for medical terminology. *AI Med*., **9**, 139–171.

Rogers,J.E., Solomon,W.D., Rector,A.L., Pole,P.M., Zanstra,P. and van der Haring,E. (1997) Rubrics to dissections to GRAIL to classifications. MIE 97, Thessalonika.

Schulze-Kremer,S. (1998) *Ontologies for Molecular Biology. Proceedings of the Third Pacific Symposium on Biocomputing*. AAAI Press, Hawaii, pp. 693–704.

Selkov,E. *et al*. (1996) The metabolic pathway collection from EMP: The enzymes and metabolic pathways database. *Nucleic Acids Res*., **24**, 26–28.

Solomon,W.D. (1998) GALEN-IN-USE Project Deliverable 9.2 The GRAIL KnoME Knowledge Modelling Environment. http://www.galen-organisation.com

Stoesser,G., Sterk,P., Tuli,M.A., Stoehr,P. and Cameron,G. (1997) The EMBL nucleotide sequence database. *Nucleic Acids Res*., **25**, 7–13.

Stoesser,G., Moseley,M.A., Sleep,J., McGowran,M., Garcia-Pastor,M. and Sterk,P. (1998) The EMBL nucleotide sequence database. *Nucleic Acids Res*., **26**, 8–15.

Woods,W.A. and Schmolze,J.G. (1992) The KL-One family. *Comput. Math. Applic*., **23**, 133–177.

Winzeler,E.A. *et al*. (1998) Direct allelic variation scanning of the yeast genome. *Science*, **281**, 1194–1197.