Adding Semantics to Genome Databases: Towards an Ontology for Molecular Biology

Steffen Schulze-Kremer
Max-Planck-Institute for Molecular Genetics
Ihnestraße 73, Dept. Lehrach, D-14195 Berlin, Germany
steffen@chemie.fu-berlin.de, http://igd.rz-berlin.mpg.de/~steffen

Abstract

Molecular biology has a communication problem. There are many databases using their own labels and categories for storing data objects and some using identical labels and categories but with a different meaning. Conversely, one concept is often found under different names. Prominent examples are the concepts "gene" and "protein sequence" which are used with different semantics by major international genomic and protein databases thereby making database integration difficult and strenuous.

This situation can only be improved by either defining individual semantic interfaces between each pair of databases (complexity of order n^2) or by implementing one agreeable, transparent and computationally tractable semantic repository and linking each database to it (complexity of order n).

Ontologies are one means to provide such semantic repository by explicitly specifying the meaning of and relation between the fundamental concepts in an application domain. Here, heuristics for building an ontology and the upper level and a database branch of a prospective Ontology for Molecular Biology are presented and compared to other ontologies with respect to suitability for molecular biology (http://igd.rz-berlin.mpg.de/~www/oe/mbo.html).

Introduction

There are a multitude of databases accessible over the Internet that cover genomic (Fasman *et al.* 1996), cellular (Jacobson & Anagnostopoulos 1996), structure (Sussman, Manning, & Abola 1996), phenotype (McKusick 1994) and other types of biologically relevant information (Bairoch 1993). Even for one type of information, e.g. DNA sequence data, there exist several databases of different scope and organisation (Fasman *et al.* 1996; Keen, Fields, & others 1996; Benson *et al.* 1997).

Unfortunately, naming conventions of data objects, object identifier codes and record labels differ between databases and do not follow a unified scheme. But worse, even the meaning of important high level concepts that are fundamental to many molecular biology databases is ambiguous. For example, for SwissProt (Bairoch & Apweiler 1996) and PIR (George, Barker, & others 1997) a *protein sequence* is the raw mRNA transcript before splicing, whereas PDB (Sussman, Manning, & Abola 1996) uses *protein sequence* to refer to mature and post-processed primary structure. The differences are syntactical (nucleotide vs. amino acid residue sequence) and semantical (gene transcript vs. folded sequence).

Another example is the concept *gene*. For GDB (Fasman *et al.* 1996), a *gene* is a DNA fragment that can be transcribed and translated into a protein; for Genbank (Benson *et al.* 1997) and GSDB (Keen, Fields, & others 1996), however, a *gene* is a "DNA region of biological interest with a name and that carries a genetic trait or phenotype" which includes non-structural coding DNA regions like intron, promoter and enhancer. There is a clear semantic difference between those two notions of *gene* but both continue to be used interchangeably causing misunderstanding and making the integration of databases non-trivial.

To eliminate semantic confusion in molecular biology, it will be therefore necessary to have a list of most important and frequently used concepts coherently defined so that database managers could use such set of definitions either to create a new database schemata or to provide an exact, semantic specification of the concepts used in an existing schema.

To become generally acceptable in the molecular biological community such semantic compendium must be accessible electronically and without licensing charges, preferably using a world wide web browser; be intuitively comprehensible without special computer programming background; be able to cope with natural language features as e.g. homonyms; be capable to perform logical inference over the set of concepts to provide for generalisation and explanation facilities; exhaustively cover the application domain; and be coordinated but open for input from the community. Also, special software to manage a semantic repository must be made available.

One way to consistently and transparently create such set of definitions for molecular biology is by using an ontology, as explained below. By adhering to a commonly agreeable ontology, uncertainties and misunderstandings about the semantic relations between database entries from different databases can be eliminated.

When all relevant concepts of an application domain will

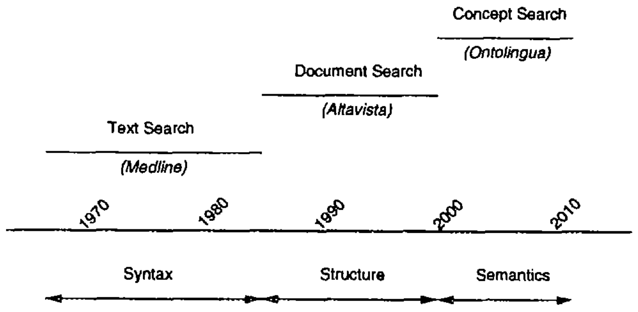

Figure 1: Semantic stages in information retrieval evolving from pure text search through document identification to concept search (Schatz 1997).

have been specified in an ontology, a computer program can search for *concepts* instead of words in a set of heterogeneous, autonomous databases (Figure 1); carry out semantic consistency checks; and detect ill-formed statements and interpret well-formed ones (Schulze-Kremer 1996).

System and methods

What is an Ontology? Ontology was originally perceived by ancient philosophers as the study of *being*. They asked "What does the statement 'X *is*' mean?" and "Which things *are* ?" (Aristotle 350 BC). In modern times, computer science uses ontology in a narrower sense as a "specification of a conceptualization" (Gruber 1993) or, in other words, as a concise and unambiguous description of what principal entities are relevant in an application domain and how they can relate to each other. The entities can be objects, processes, functions, predicates, or of other type depending on the selected representation formalism.

An ontology is *not* a collection of facts that arise from an actual, specific situation but it defines and provides all semantic entities and their potential interactions that would be necessary to completely describe that situation. *Neither* is an ontology a model for an application domain (which would be a theory), but a compendium that holds all necessary "building blocks" with rules of how and which entities can relate to each other and which ones are semantically incompatible. For example, "transcription of a gene" is an ontologically valid expression whereas "transcription of a cell" is not because a cell cannot be transcribed.

Components of an Ontology

The core of an ontology consists of:

1. Components
   (a) the "is a subset of" relation;
   (b) the "is a member of" relation;
   (c) a knowledge representation formalism (e.g. first-order predicate logic);
   (d) a set of concepts containing all relevant types of entities in an application domain.
2. Constraints
   (a) each concept must be explicitly defined;
   (b) each concept must be unambiguously accessible within the ontology;
   (c) each concept must be connected to one another by one or more "is a subset of" or "is a member of" links or their inverses.

Additional modules are in practice essential for the implementation and use of a large ontological system:

3. Background Knowledge
   (a) a set of additional attributes and relations that capture domain specific properties of concepts (ideally those properties should all themselves be concepts in the ontology);
   (b) precise annotation of the concepts in an ontology using aforementioned additional properties to make the semantic scope of concepts computationally explicitly accessible;
   (c) links from the concepts of the ontology to natural language dictionaries and database keywords.
4. Software
   (a) an object-oriented DBMS to store the set of concepts and properties of an ontology;
   (b) an inference engine that can generalise across several layers of "is a member of", "is a subset of" and other types of relations;
   (c) a graph editor with tree formatting capabilities to visualise "is a subset of" and "is a member of" hierarchies in combination with other types of hierarchies (e.g. "is part of", "is made of");
   (d) an ontology browser and annotator that can be accessed over the Internet via a world wide web browser.

The graphical representation of an ontology is not a tree but a semantic net or conceptual graph (Sowa 1984) because there are two or more types of links ("is a member of" and "is a subset of" plus additional domain specific relations) which can give raise to circular loops in the graph. However, if only the "is a subset of" or "is a member of" relation is displayed the ontological graph becomes a tree.

Heuristics for Building an Ontology

There is no universal rule of how to build an ontology and hence no unique, provable ontology exists, mainly because of our incomplete knowledge in metaphysics and the essentially indeterministic choice of criteria for subclassifying a concept. It has therefore been argued that the scope, structure and refinement of an ontology is largely determined by its intended use (Mahesh & Nirenburg 1996; Uschold & Gruninger 1996). However, far from being totally arbitrary there are some rules one can follow when building an ontology.

An ontology can be built by answering the question "What is an X?" for each relevant concept X of the application domain and then linking X with either the "is a subset of" or "is a member of" relation to its parent concept, which should already occur in the ontology. If it does not, the question "What is an X?" is repeated for the parent concept, the parent of the parent and so on, until a most general concept is reached which becomes the root node of the ontology for that application.

In order to facilitate the addition of the bulk of concepts that are deemed relevant by experts in the application domain, it is helpful to first start top down from a most general concept, i.e. the tentative root node of the ontology, and subclassify it and its subnodes to the forth or fifth level into disjoint subclasses according to explicit and clearly defined discriminating criteria. These criteria should capture fundamental properties in the application domain and increase the information content of the "is a" hierarchy as much as possible. The most informative criteria in an ontology are properties that all concepts in the application domain express. The decision about a discriminating criterion should reflect the most significant subtypes of a concept with respect to the application domain. With the lack of an objective measure for the absolute utility of a discriminating criterion at a given position in the ontology different branching patterns are possible.

To keep the ontological graph connected, intermediate concepts have to be created that bridge the gap between specific, application derived concepts and between detailed and more general concepts. These intermediate concepts are also considered relevant to the application domain. Ontological concepts outside the application domain should not be included just because "they would fit in there" but only if needed to connect domain relevant concepts. Otherwise, the distinction between relevant and irrelevant concepts later on will be very difficult and the ontology becomes burdened with irrelevant information. The only exceptions should be concepts that are used to identify the position for future insertion of domain relevant concepts.

Although this is a straight forward procedure the greatest difficulties can arise as one tries to capture the "true" meaning (in analytical or empirical sense) of a concept. Here, one should refer to classical textbooks of the domain, dictionaries, reference handbooks, and resources of international standardisation organisations which implicitly and in an informal way contain ontological fragments, to find a sensible, representative and agreeable definition of a concept.

Thus, one establishes first the upper level of an ontology, where the scope and principal categories are identified (Figure 2). On the lower levels, more detailed distinctions between the categories are made until one arrives at the relevant concepts of the application domain. The granularity of the graph depends on the semantic range of the concepts to be covered.

Molecular Biology Ontology

An ontology for molecular biology should become a repository for all relevant concepts that are required to describe biological objects, experimental procedures and computational aspects of molecular biology.

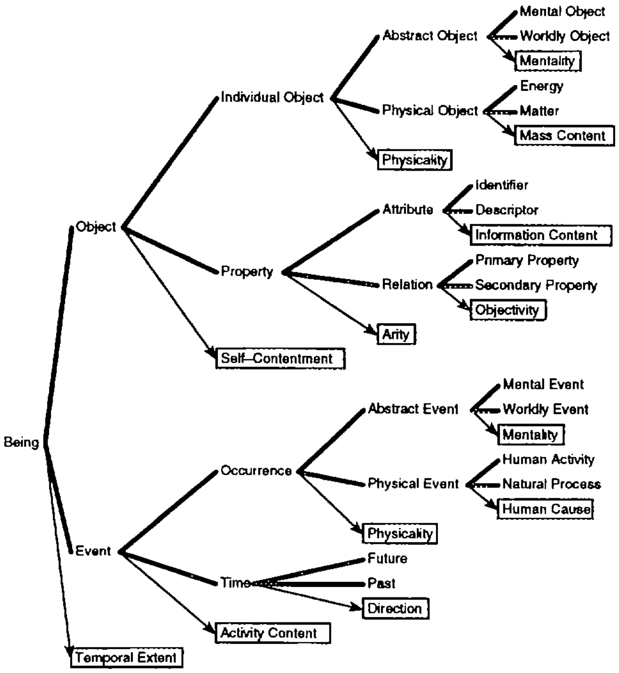

Figure 2: Upper Level of a prospective Molecular Biology Ontology. Links represent the "is a subclass of" relation. No instances are present; discriminating criteria have arrows and boxes; thick lines denote disjunct subclasses.

Although this looks like an impossible task at first sight it does not mean to compile all knowledge about molecular biology nor does it imply being able to explain every biological phenomena. It just means collecting all types of entities that a molecular biologist includes in her professional thinking and placing those concepts appropriately in a "is a subset of" and "is a member of" hierarchy plus annotating them with additional properties. By doing so in a consistent manner, where the discriminating criteria of subclassifying each concept are made explicit, the definition of a concept becomes the path from its own node to the root node of the ontology.

For example, the GDB database category *LinkageMap* is defined as a "database object class used to store maps based upon frequency of recombination between genomic segments, resulting in the ordering of markers along a chromosome backbone, usually measured in centiMorgans". Note that this is ontologically not the same as a linkage map itself which is an abstract concept with certain mathematical properties, nor is it an actual linkage map which is a concrete instance of the class of linkage maps for a particular organism and chromosome.

The complete path from root node *Being* to *LinkageMap* is summarised in Figure 3. The meaning of the database category LinkageMap is captured by a series of ontological specifications. This example shows how a semantic definition of a molecular biological concept can be extracted

Concept	Discriminator	Value
Being	Temporal Extent	instanteneous
Object	Self-Contentment	yes
Individual Object	Physicality	yes
Abstract Object	Mentality	no
Worldly Object	Domain Specific Usage	yes
Domain Specific Wordly Object	Subject Domain	mathematics
Mathematical Object	Complexity	high
Composed Mathematical Object	Application Specificity	yes
Applied Mathematical Object	Application Domain	computer science
Computer Science Object	Subject Matter	theory of data
Theory Of Data Object	Subject Matter	data structure
Data Structure Object	Representation Formalism	object-oriented
Object-Oriented Data Structure	Implementation	OPM
OPM Database Object	Database	GDB
GDB Database Object	Subject Matter	genomic
GDB Genome Object	Object Class	DBObject
DBObject	Subclass	MappingObject
MappingObject	Subclass	Map
Map	Subclass	LinkageMap

Figure 3: Semantic hierarchy for the GDB database category *LinkageMap.* A LinkageMap is a *Being* with instantaneous temporal extent, an *Object* which is self-contenting, etc. until the concept *LinkageMap* is reached at the bottom.

from its position in an ontological graph. Similarly, semantic differences and the least general common concept of a pair of concepts can be found by following the graph upwards along "is a subclass of" and "is a member of" links until both paths meet in one concept.

Discussion

An ontology is an explicit and hierarchical specification of the relevant concepts in an application domain and therefore one means to develop a semantic repository for molecular biology. Two ontologies from literature, mikroKosmos (Mahesh & Nirenburg 1996) and Cyc (Lenat 1995), have been reviewed with respect to suitability in molecular biology (data not shown here for spacial constraints). The mikroKosmos ontology is found not to be transparent and precise enough to collect and sort scientific concepts concerning molecular biology. Cyc contains a lot of knowledge about semantic distinctions in daily life but seems to be too complicated and overloaded with concepts not relevant for molecular biology to be of use here.

Here, heuristics for building an ontology are given and the upper level of a prospective Ontology for Molecular Biology. In contrast to other ontologies, the criterion used for subclassifying a concept is explicitly stated and therefore essential decisions and assumptions behind the ontology are made transparent. An interactive graphical representation of all public classes and instances of mikroKosmos, Cyc and a prospective Ontology for Molecular Biology was prepared by the author and is accessible at http://igd.rz-berlin.mpg.de/~www/oe/mbo.html.

Acknowledgements. This work was supported by the European research grant on IGD-GIS PL950263.

References

Aristotle. 350 BC. Categories. Translated by E. M. Edghill.

Bairoch, A., and Apweiler, R. 1996. The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. *Nucleic Acids Research* 24(1):21–25.

Bairoch, A. 1993. The ENZYME data bank. *Nucleic Acids Research* 21(13):3155–3156.

Benson, D. A.; Boguski, M. S.; Lipman, D. J.; and Ostell, J. 1997. GenBank. *Nucleic Acids Research* 25(1):1–6.

Fasman, K. H.; Letovsky, S. I.; Cottingham, R. W.; and Kingsbury, D. T. 1996. Improvements to the GDB(TM) Human Genome Data Base. *Nucleic Acids Research* 24(1):57–63.

George, D. G.; Barker, W. C.; et al. 1997. The Protein Information Resource (PIR) and the PIR-International Protein Sequence Database. *Nucleic Acids Research* 25(1):24–28.

Gruber, T. R. 1993. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* 5(2):199–220.

Jacobson, D., and Anagnostopoulos, A. 1996. Internet resources for transgenic or targeted mutation research. *Trends in Genetics* 12(3):117–118.

Keen, G.; Fields, C.; et al. 1996. The Genome Sequence DataBase (GSDB): meeting the challenge of genomic sequencing. *Nucleic Acids Research* 24(1):13–16.

Lenat, D. B. 1995. CYC: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM* 38(11).

Mahesh, K., and Nirenburg, S. 1996. Meaning Representation for Knowledge Sharing in Practical Machine Translation. In *Proceedings of the FLAIRS-96 Track on Information Interchange, Florida AI Research Symposium.*

McKusick, V. A. 1994. *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders.* Baltimore, MD: Johns Hopkins University Press, 11 edition.

Schatz, B. R. 1997. Information Retrieval in Digital Libraries: Bringing Search to the Net. *Science* 275(5298):327–334.

Schulze-Kremer, S. 1996. *Molecular Bioinformatics - Algorithms and Applications.* Walter de Gruyter, Berlin. 13–108.

Sowa, J. F. 1984. *Conceptual Structures: Information Processing in Mind and Machine.* Reading, MA: Addison-Wesley.

Sussman, J. L.; Manning, N.; and Abola, E. E. 1996. *Quarterly Newsletter 77.* Protein Data Bank, Brookhaven National Laboratory, P.O. Box 5000, Upton, NY 11973-5000, USA.

Uschold, M., and Gruninger, M. 1996. Ontologies: Principles, Methods and Applications. *Knowledge Engineering Review* 11(2).