A2-D1

State-of-the-art in Bioinformatics

Project number:	IST-2004-506779
Project title:	Reasoning on the Web with Rules and Semantics
Project acronym:	REWERSE
Document type:	D (deliverable)
Nature of document	R (report)
Dissemination level:	PU (public)
Document number:	IST506779/Dresden/A2-D1/D/PU/b1
Responsible editor(s):	Michael Schroeder
Reviewer(s):	Patrick Lambrix and Tim Furche
Contributing participants:	Dresden, Edinburgh, Jena, Manchester, Linköping, Paris
Contributing workpackages:	A2
Contractual date of delivery:	31 August 2004

Abstract

With the explosion of online accessible bioinformatics data and tools, systems integration has become very important for further progress. Currently, bioinformatics relies heavily on the Web. But the Web is geared towards human interaction rather than automated processing. The vision of a Semantic Web facilitates this automation by annotating web content and by providing adequate reasoning languages as developed in the REWERSE project.

The report summarises the state-of-the-art in bioinformatics relevant to REWERSE. First, we give a motivation by sketching the growth of data in biology, the types of data, and their distribution and availability. We introduce two fundamental classes of bioinformatics problems: alignments (of sequences or structures) and structure prediction. Next, we discuss the most important databases and approaches towards their integration. A fundamental ingredient of a Semantic Web, which transparently integrates data sources are ontologies. We discuss a number of different bioinformatics ontologies including GeneOntology, the most widely used one with some 19.000 terms. Finally, we show three application areas in which rules are used: Rules to implement computation tree logic to represent and query metabolic pathways; Rules as constraints for structure prediction; and rules learned by inductive logic programming to

declaratively capture biological knowledge.

Keyword List
Bioinformatics, rules, reasoning, ontologies, sequence, structure, networks

State-of-the-art in Bioinformatics

Rolf Backofen[1], Mike Badea[2], Albert Burger[3], François Fages[4], Patrick Lambrix[5], Werner Nutt[3], Michael Schroeder[6], Sylvain Soliman[3], Sebastian Will[1]

[1] Friedrich-Schiller-Universität Jena, Germany, [2] Victoria University of Manchester, UK, [3] Harriot-Watt University, Edinburgh, UK, [4] INRIA Rocquencourt, France, [5] Linköpings universitet, Sweden, [6] Technical University of Dresden, Germany

2 August 2004

Abstract

With the explosion of online accessible bioinformatics data and tools, systems integration has become very important for further progress. Currently, bioinformatics relies heavily on the Web. But the Web is geared towards human interaction rather than automated processing. The vision of a Semantic Web facilitates this automation by annotating web content and by providing adequate reasoning languages as developed in the REWERSE project.

The report summarises the state-of-the-art in bioinformatics relevant to REWERSE. First, we give a motivation by sketching the growth of data in biology, the types of data, and their distribution and availability. We introduce two fundamental classes of bioinformatics problems: alignments (of sequences or structures) and structure prediction. Next, we discuss the most important databases and approaches towards their integration. A fundamental ingredient of a Semantic Web, which transparently integrates data sources are ontologies. We discuss a number of different bioinformatics ontologies including GeneOntology, the most widely used one with some 19.000 terms. Finally, we show three application areas in which rules are used: Rules to implement computation tree logic to represent and query metabolic pathways; Rules as constraints for structure prediction; and rules learned by inductive logic programming to declaratively capture biological knowledge.

Keyword List

Bioinformatics, rules, reasoning, ontologies, sequence, structure, networks

Contents

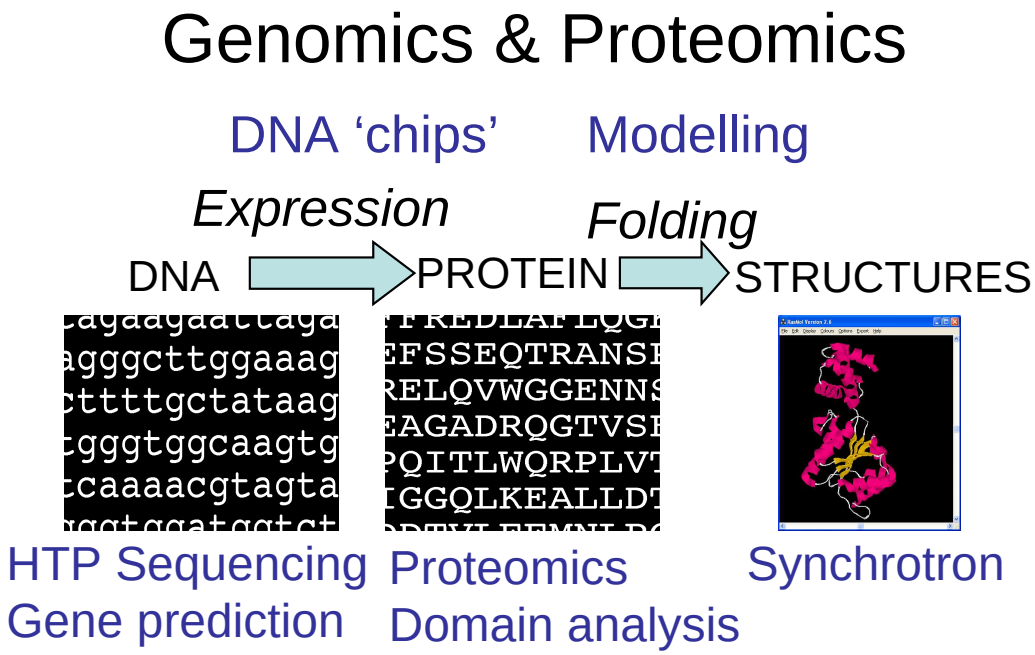

Figure 1: The central dogma of biology: DNA is expressed and transcribed into RNA, which is translated into an amino acid sequence, which then folds into its 3D structure. Biology has changed dramatically and high-throughput experimental methods such as produce masses of data.

Reviewers: Patrick Lambrix, Tim Furche

1 Introduction

1.1 Data, data, data,...

The central dogma of biology states that DNA is expressed and transcribed into RNA, which is translated into an amino acid sequence, which then folds into its 3D structure (see Fig. 1). High-throughput experimental methods produce masses of data, so that the whole of biology has changed from a data-light science into a data-driven science. To give an impression of the sheer size of data, consider the numbers below:

- DNA sequences:
  - 16.000.000.000 bases (= 16 Gbp (Giga base pairs))
  - Human genome = 3.2 Gbp (equivalent size to 6 complete years of the New York Times)
- Literature:

  - PubMed 14.000.000 abstracts

- Protein sequences:
  - SWISSPROT: 130.000 annotated protein sequences
  - TrEMBL: 850.000 protein sequences

- Protein structures:
  - PDB: > 25.000 protein structures with an average of ca. 400 residues

The above databases grow superlinearly!

1.2 Problems

The masses of data led to numerous challenging bioinformatics problems. Consider the following scenario [Lesk, 2002], where the numbers in brackets indicate the difficulty: (< 30: solution exists already; > 30: we cannot solve this (yet))

- A new virus occurs (e.g. SARS) and scientists want to develop a treatment
- Scientists isolate the genetic material of virus
- They screen the genome for relationships with previously studied viruses [10]
- From virus' DNA they compute the proteins it produces [1]
- Compute proteins' three-dimensional structure and thereby obtain clues about their functions
- Screen for similar proteins sequences with known structure [15]
- If any are found
  - Then interpret difference (homology modelling) [25]
  - look for known inhibitors of such proteins in metabolic networks and signaling pathways [35]
  - Else predict structure from sequence [55]
- Identify or design small molecule blocking relevant active sites of the protein [50]
- Design antibodies to neutralize the virus [50]

Three fundamental problems in this scenario are structure prediction, alignment of sequences and structures and querying of networks. Let us consider these problems in more detail.

1.2.1 Structure Prediction

The protein structure prediction problems aims to predict the 3D structure of a protein from its protein seqeunce. It is one of the most important unsolved problems of computational biology, which was shown to be at least NP-complete [Berger and Leighton, 1998, Crescenzi et al., 1998]. For this reason, simplified models have been successfully used by several groups in hierarchical approaches for protein folding [Xia et al., 2000].

The most important class of simplified models are the lattice models, where monomers are represented using a unified size having positions in a regular lattice. They are used for structure prediction [Xia et al., 2000] as well as a method for investigating general properties of protein folding (since they constitute a genotype versus phenotype mapping, e.g. [Bornberg-Bauer and Chan, 1999]).

A discussion of lattice proteins can be found in [Dill et al., 1995]. There is a bunch of groups working with lattice proteins. Examples of how lattice proteins can be used for predicting the native structure or for investigating principles of protein folding are [Šali et al., 1994, Abkevich et al., 1995, Dinner et al., 1996, Unger and Moult, 1996, Hinds and Levitt, 1996, Govindarajan and Goldstein, 1997, Abkevich et al., 1997, Ortiz et al., 1998]. Most of them use heuristic methods, ranging from Monte-Carlo simulated annealing (e.g. [MacDonald et al., 2000, Dinner et al., 1996]) to genetic algorithms (e.g. [Unger and Moult, 1996]), purely heuristic methods like hydrophobic zipper [Dill et al., 1993] and the chain growth algorithm [Bornberg-Bauer, 1997], as well as complete enumeration (often restricted to subset of all conformations, e.g. [Šali et al., 1994, Xia et al., 2000]).

First steps have been made to improve the situation on the algorithmic part. The first improvement was the introduction of an exact algorithm for finding minimal energy conformations in the cubic HP-Model [Yue and Dill, 1995]. The algorithm is called CHCC for "**C**onstraint **H**ydrophobic **C**ore **C**onstruction) (albeit it doesn't use constraint-based methods). This approach works only for the cubic lattice, but not for the FCC-lattice which is much better suited for modeling real protein conformations. The second improvement is the appearance of a bunch of approximation algorithms [Hart and Istrail, 1996, Agarwala et al., 1997] for different lattice models. Although they are a very important first step in the right direction, their approximation ratio is still not good enough to be used in practice.

Another interesting approach is to combine secondary structure predictions with lattice models. The secondary structure consists of local structure motifs like α-helices and β-sheets. One searches for conformations in the lattice model with low energy having the predicted secondary structure elements. Here, major improvements have been achieved using again a constraint-based approach and the FCC lattice [Dovier et al., 2002].

1.2.2 Alignment Methods

The alignment problem is a very important problem in bioinformatics, where one searches for an alignment of two strings, optimising a certain score. The strings represent biological macromolecules as DNAs or proteins. The scoring scheme evaluates the aligned columns. Alignment is one of the core techniques, which are used to search for similar objects in biology. Initially, they were used only for sequences, but it was later extended to many other problems (protein structure, pathways etc.)

In the simplest case, alignment is identical to computing the edit distance of strings. This problem is usually solved by dynamic programming (DP) as e.g. by Needleman and Wunsch in [Needleman and Wunsch, 1970].

However, dynamic programming approaches suffer from their inflexibility. If the problem is slightly modified, one has to develop a new DP algorithm (if one exists at all). For example, it is an unsolved problem to align two sequences incorporating biological knowledge that tells us which sub-sequences/domains should be aligned.

In addition, many extensions of the basic alignment problem have been investigated (with the need of generating new algorithms for every variation of the problem). Examples of this kind of "extended basic alignment methods" is RNA-Sequence-Structure Alignment [Jiang et al., 2002, Höchsmann et al., 2003, Backofen and Will, 2004]. This kind of sequence/structure similarity has been extended to proteins in many different ways, where the problem becomes NP-complete. This are protein protein threading (e.g., [Akutsu and Miyano, 1997]), the contact map problem [Lancia et al., 2001] and the alignment of TOPS-diagrams [Visna and Gilbert, 2002]

1.2.3 Metabolic and signaling networks

In recent years, molecular biology has engaged in a large-scale effort to elucidate cellular processes in terms of their biochemical basis at the molecular level. Mass production of post genomic experimental results, such as mRNA expression data, protein expression or protein-protein interaction data, is following and completing the initial piecemeal catalog of elementary components – genes and proteins – of the sequencing and genomic analyses projects by progressively painting a global picture of the complex interactions that take place in a cell. Exploiting these experimental data to understand the underlying processes requires much more than database integration and storage: it calls for a strong parallel effort on the formal representation of biological *processes*.

Several formalisms have been proposed in recent years for the modeling of metabolic pathways, extracellular and intracellular signaling pathways, or gene regulatory networks: boolean networks [Thieffry and Thomas, 1998], ordinary differential equations [Schoeberl et al., 2002], and more recently hybrid Petri nets [Matsuno et al., 2000, Hofestädt and Thelen, 1998] and hybrid automata [Alur et al., 2001, Ghosh and Tomlin, 2001]. Formal concurrent languages were also considered, including hybrid concurrent constraint languages [Bockmayr and Courtois, 2002], or rewriting logics [Eker et al., 2002a]. Regev and Shapiro [Regev et al., 2001a] were the first to propose the use of the π-calculus [Milner et al., 1992].

Most formal approaches mentioned above proceed by wholesale importation of a language (eg Petri nets, the π-calculus) that emerged in answer to very specific design goals, some of which may be relevant to our present modeling task, and some of which may not. While the expected benefit is direct inheritance of preexisting methods and tools, this results in some contorted translations and the existence of useless constructs, and somewhat defeats the explanatory purpose of the formalisation.

The current state-of-the-art in modeling is mostly based on simulation and graphical display [Alur et al., 2001, Bockmayr and Courtois, 2002, Maimon and Browning, 2001, Matsuno et al., 2000], with some attempts towards stability and bifurcation analyses of dynamical behaviour on small systems [Thieffry and Thomas, 1998, de Jong, 2001] described either by differential equations or by discretisations thereof. A different approach promotes symbolic manipulation and exploration of the model by means of computational logics which are commonplace, in hardware verification for instance. Formal methods extend the ways one can play with a given model and thus may second simulation and even replace it when quantitative information is sparse or inacurate.

1.3 Bioinformatics and the web

As argued above, there are masses of bioinformatics data and algorithms working on it. Much of the data and tools are published online. As a consequence, a strategic research workshop of the NSF and EU found that bioinformatics could play the role for the Semantic Web, which physics played for the web. The reasons are manyfold:

- There are masses of information
- The data are public
- The data are online
- The data are (more and more often) published in XML
- Data standards are accepted and actively developed by the bioinformatics community
- Much valuable information is scattered (as production is cheap and hence often not centralised)
- Systems integration and interoperation are a prime concern
- One can predict that in the not too distant future many tools and databases will be accessible as web services

2 Bioinformatics Data

With much data online accessible, let us consider some of the data sources. In [Lambrix and Jakonienė, 2003] a number of the most used and best known databanks were studied: GenBank, EMBL, DDBJ, SWISS-PROT, PIR, ENZYME, PDB, MMDB, PROSITE, PRINTS and BLOCKS. For these systems we investigated the organisation of the data, the data content and the data retrieval possibilities. With respect to the organisation of the data we looked at the kind of data, the source, the data model, the update frequency and the location of the data. The chosen databanks store information about nucleotide sequences (GenBank, EMBL, DDBJ), protein sequences (SWISS-PROT, PIR, ENZYME), 3D macromolecular structures (PDB, MMDB) and protein families (PROSITE, PRINTS, BLOCKS). The source of the data can be researchers (that submit their data), literature (data from published articles) and other databanks. This is summarised in table 1. Data can usually be retrieved in different ways: via a web interface, ftp and e-mail. The underlying data models for these databanks are the flat file model, the relational model and the object-relational model. Also different formats may be used for the databank behind the web interface and the databank that can be loaded from ftp servers. A summary of the formats is given in table 2.

The content information can be grouped into the header, the annotation and the actual information. The header contains a unique identifier for the data item, one or more entry dates, one or more names, the source of the information and references. The annotation part contains comments and feature information. Finally, there is additional information that can contain a protein or DNA sequence, structure descriptions or experiment descriptions depending on the kind of databank. The level of detail of the data is also different.

With respect to retrieval capabilities most databanks allow for queries based on the occurrence of text within a data item (full-text search) and all databanks support queries based on

Databank	R	L	DB
GenBank	Y		EMBL, DDBJ
EMBL	Y		GenBank, DDBJ
DDBJ	Y		EMBL, GenBank
SWISS-PROT	Y	Y	GenBank, EMBL DDBJ
PIR	Y	Y	GenBank, EMBL DDBJ
ENZYME	Y	Y	IUBMB
PDB	Y		
MMDB			PDB
PROSITE	Y	Y	SWISS-PROT
PRINTS	Y		SWISS-PROT, TREMBL
BLOCKS			Interpro

Table 1: Source of the data. (Researchers (R), Literature (L), Other databanks (DB).)

the occurrence of a text string within certain predefined fields. The user is often guided by the retrieval interface of the systems as to which fields are searchable. Two of the databanks, ENZYME and PRINTS, also allow for browsing the branches of a predefined structure. Most systems support Boolean queries (using and, or, not) as well as wildcards in the text strings. In addition to a form-based query interface, most systems also support command-line querying using the systems' query languages. A summary is given in table 3. An explanation of the abbreviations used in table 3 is given in table 4. The result of a query is for most systems a list of the data entries that match the query. The actual result data is the complete data entry. In some systems it is possible to define views over the result and in that way one may retrieve only the interesting fields. Two systems, PDB and PIR, support reuse of query results in new queries.

3 Integration of Databanks

There exist a number of systems that provide access to multiple biological databanks. These systems can be divided into two categories [Davidson et al., 2001]. The category of the link-driven federations contains most of the currently used websites that provide an interface to multiple biological resources, such as Entrez and SRS. These systems support a number of basic queries via a web interface. Often, they also allow to use alignment algorithms such as BLAST on the result of a query. Usually, the users need to explicitly state which resources should be used for retrieving the answers, requiring good knowledge of the underlying sources. The data source systems are often implemented using flat files and specialised retrieval packages. Most of the integration is link driven and is achieved by the creation of cross-reference indexes. For instance, the SRS language defines search in indexes of databanks (including string search, regular expressions, numeric ranges and dates), and combinations of queries using **and**, **or** and **andnot**. With respect to the combination of databanks the *link* construct is introduced. This allows for queries of the forms 'find all entries in databank A that are referenced in databank B', and 'find all entries in databank A that reference entries in databank B'. The advantage of such systems is that queries relating to knowledge in different databanks can be asked and that the query processing is fast. However, although this is a first step in integrating data sources, this solution does not handle the differences in terminology used in the underlying sources, is syntax based and only allows limited query functionality over multiple databanks. Also, adding

Databank	FTP server	user interface
GenBank	genbank flat file, asn.1	genbank flat file
EMBL	embl flat file	embl flat file, fasta, xml
DDBJ	ddbj flat file	ddbj flat file, embl flat file, xml, fasta
SWISS-PROT	swissprot flat file, fasta	user friendly view swissprot flat file
PIR	nbfr-pir, codata, fasta, xml	nbfr-pir, codata, fasta, xml
ENZYME	enzyme flat file, asn.1	user friendly view, enzyme flat file
PDB	pdb flat file, mmCIF	pdb flat file, mmCIF
MMDB	asn.1	mmdb flat file
PROSITE	prosite flat file	user friendly view, prosite flat file
PRINTS	prints flat file	prints flat file
BLOCKS	blocks flat file	blocks flat file

Table 2: Data formats.

a new resource requires cross-referencing with the other resources.

The BioKleisli [Davidson et al., 1997], K2 [Davidson et al., 2001], TINet [Eckman et al., 2001], P/FDM [Kemp et al., 2000], TAMBIS [Goble et al., 2001] and BioTRIFU [Lambrix and Jakonienė, 2003] systems use an approach based on view integration. Also IBM's DiscoveryLink can be placed into this category. In this approach the underlying schemas are integrated to form a global schema. The global schema is queried in a high-level language such as CPL (e.g. BioKleisli) or OQL (e.g. K2). The languages in TAMBIS [Stevens et al., 2001] and BioTRIFU [Lambrix and Jakonienė, 2003] have been inspired by the study of the use of current biological databanks. The other proposals seem to be based mainly on database technology. In general, the view approach allows for more complex querying and allows for support for integration on schema level. The advantages of such systems include the possibility of complex querying, the knowledge that is required of the end-user is not as large and the local conceptual models are used in the integration. Further, these view integration systems may also be used to create warehouses. As a possible solution of the terminology discrepancy problem in the local schemas as well as a step towards semantic querying, ontologies could be used.

4 Bioinformatics and Ontologies

In bioinformatics, the majority of data sources are databases and tools with a variety of user interfaces. The ontologies help manage the interoperation between these resources by providing a dynamic controlled vocabulary of concepts. One widely used definition of ontology is "a specification of a conceptualisation". Even with this definition, there is a spectrum of the structure of ontologies, ranging from flat lists of controlled vocabulary to very formal

Databank	UI	CL
GenBank	a,n,nr,k,o,g,au,c,spec	a,n
EMBL	a,n,nr	n,nr
DDBJ	a,n,nr,k	
SWISS-PROT	a,n,nr,d,o,g,au,c	a,n,nr,d,o,g,k,au
PIR	a,nr,k,g,o,au,c,f	n,nr
ENZYME	n,d,spec	n
PDB	a,n,d,k,au,c,spec	n
MMDB	a,n,o,au,c,spec	a,n
PROSITE	a,n,nr,d,au,c,spec	a,n,nr,d,au
PRINTS	n,nr,d,f,s,spec	n
BLOCKS	n,k,s	n,k

Table 3: Search within specific fields in the form-based user interface (UI) and the command-line interface (CL).

a: all text	n: entry name	nr: accession number
d: description	k: keyword	o: organism
g: gene name	au: author	c: citation
f: family	s: sequence	spec: specific fields

Table 4: Explanation of abbreviations in table 3.

logic-based implementations. (For an overview see e.g. [Stevens et al., 2000b, Lambrix, 2004, Guarino and Giaretta, 1995, Jasper and Uschold, 1999].) Pragmatically, an ontology typically includes a hierarchically arranged list of concepts (i.e., classes) of a given domain, relationships among these concepts, definitions of these concepts and relationships, and optional logical axioms that serve as further constraints among these entities. Apart from the high-level aim of representing the knowledge of a domain in a computationally amenable form, an ontology provides a controlled vocabulary in that a given term always has the same meaning. Thus, rather than attempting to parse the complex structure of natural language for embedded biological knowledge, computational agents can query an ontology and a knowledge base based on the ontology to specifically and reliably retrieve data.

4.1 Gene Ontology (GO)

The most successful ontology within the realm of bioinformatics is the Gene Ontology (GO) [GeneOntologyConsortium, 2004], a controlled vocabulary that details the molecular functions gene products may possess, the higher-level biological processes in which they may participate, and the cellular locations in which they may be active. Although GO was created for the express purpose of providing a common terminology for functional annotation of genes and gene products in biological databases towards the goal of database interoperability, it has since been widely used for a variety of purposes, including analyses of experimental data, predictions of experimental results, and document retrieval. GO is the flagship ontology of the Open Biological Ontologies (OBO), a collection of biological ontologies that are open in that they can be used by all without constraint so long as the sources are acknowledged and the ontologies are not edited and redistributed under the same names (http://obo.sourceforge.net/). In addition to the taxonomies of GO, the OBO ontologies deal with anatomies of humans and of various

model organisms, biochemical substances, and sequence types, among others. Over the past years GO has developed into the main ontology in molecular biology and it comprises over 19.000 terms organised in three subontologies for cellular location, function and process. The terms are linked by three relations: is-a, part-of, is-synonym.

GO was initially created to reflect Drosophila gene function via the Flybase database, but has expanded to encompass mouse, yeast and gene expression databases, and is expected to expand further. Proteins in UniProt and the Interpro databases are currently being assigned GO terms.

Designing a structured vocabulary of some 19.000 terms is far from easy and there are a couple of principles, an ontology should follow [Gruber, 1993], but which GO still violates:

- *Clarity*, an ontology should be objective and clearly describe its concepts. GO comprises terms, which appear like definitions such as e.g. the 29 word term 'oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors'.

- *Coherence*, definitions given to the concepts must not invalidate the logical relations between concepts. For example, the concept 'succinate (cytosol) to fumarate (mitochondrion) transporter' implicitly holds properties about location and orientation in the mitochondrial membrane and thus mixes functional and location concepts, which should be separate

- *Extendibility*, the basic concepts defined in the ontology need to allow a monotonic extension. The basic concepts need not be changed for such an extension. GO contains terms with *unknown location*. When the location becomes known the ontology would have to be changed. GO should only define positive knowledge.

- And *minimal encoding bias*, concepts should be defined on the knowledge level and not on the level of syntax. GO terms like 'structural protein of chorion (sensu Drosophila)' encode the information that the concept is to be seen in the context of a certain organism, Drosophila in this case. The name of the concept syntactically represents an attribute (the species).

However, all of the above problems are minor and the vast majority of GO terms are well-structured.

4.2 Anatomy Ontologies

While most biological databases contain information about findings at the molecular level, there is a growing need to link this information to concepts about the global structure of organisms, that is to their anatomy. This development is due to two main reasons.

A central question in genetics is which genes influence the development of which parts of an organism and which genetic mutations cause which deviations from the standard phenotype. Researchers tackle this question by exploring which genes are expressed at which stage of development in which tissues of an organism. To make such findings generally accessible, a standardised vocabulary about developmental stages and tissues is needed for annotations. A second reason is that biological image data are increasingly being published on the Web. To describe in a uniform way what tissue an image shows one has to resort to some anatomical vocabulary.

4.2.1 Anatomies for Model Organisms

Modern biology aims to identify fundamental principles that govern biological processes at the cellular and molecular level in arbitrary organisms. However, not all organisms are equally well suited for studying a specific question or for conducting a specific experiment. Fruit flies, for instance, have extremely short generation cycles and are inexpensive to raise, which makes them ideal for studying genetics and mutations. A drawback is that observations about flies are less likely to be transferable to humans than, for instance, observations about mammals. Among mammals, some small species like mice still have short generation cycles, which brings about advantages similar to those of flies. For some experiments it is crucial to influence or to directly observe the development of embryos. This is not possible for mammals, but it is for chickens, where the embryo develops in a bird's egg.

Thus, biologists will choose for their research a species whose characteristics are favourable for their research questions and their intended experiments. Such species are known as "model organisms." In practice, only relatively few species have been established as model organisms so far and practically all molecular and image data have been obtained from a small number of species.

Several collaborative groups of researchers have built up and are maintaining comprehensive web sites for such model organisms. For instance,

- the *Mouse Genome Informatics* (MGI) site at the Jackson Laboratory gives integrated access to various types of genetic and genomic data on the mouse [Ringwald et al., 2001];
- *Wormbase* does the same for the worm C. Elegans and other nematodes [Stein et al., 2001];
- *Flybase* collects genomic information on the fruit fly Drosophila [The FlyBase Consortium, 2003];
- the *Zebrafish Information Network* (ZFIN) makes gene expression, mutant, and other genomic data on the zebrafish available [Sprague et al., 2001].

All these sites use anatomy ontologies for their model organisms to index their data, except the Zebrafish Network, which has published an anatomy, but is still in the process of linking it to other data.

Anatomies can also be an integral part of larger ontologies, like the MeSH term system (Medical Subject Headings), which contains mostly terms for human anatomy, but also some that relate to various mammal species. MeSH terms are used to annotate entries in large bibliographical databases.

The Edinburgh Mouse Atlas Project (EMAP) is creating a resource that combines an anatomy ontology with a 3-dimensional spatial model of the mouse embryo to give access to gene expression data [Baldock et al., 2003]. Anatomical terms are linked to regions in the spatial model and vice versa. The Mouse Atlas is based on the same anatomy as Jackson Lab's MGI, but has been enriched it to represent groupings between tissues such as the "skin" group, which comprises tissues in many different locations [Bard et al., 1998].

Anatomy ontologies can be sizable. The mouse anatomy, for instance, comprises more than 8,000 terms.

4.2.2 Structure of Anatomies

All anatomies have a natural hierarchical structure. Users can browse the hierarchy to find a tissue they are interested and then look up reports on genes that are expressed in these tissues. Similarly, it is possible to start with a gene and then to find where in the organism it is expressed. However, the anatomies mentioned above follow different approaches to structure as well as to the way in which terms are linked to other data.

For one, they differ with respect to the primitives by which they are expressed. While the mouse anatomy of MGI and the zebrafish anatomy of ZFIN are trees with unlabeled edges, the edges in the Flybase anatomy are labeled as *component part*, *instance of*, or *derived from* while those in the Wormbase anatomy are *component*, *instance*, *descendant*, and *develops from*. Moreover, the Flybase and Wormbase anatomies are directed acyclic graphs, but not trees. While in Wormbase almost exclusively leaf terms of the anatomical hierarchy are employed for annotations, in other systems arbitrary also non-leaf terms are used frequently.

Most anatomies, like the ones of mouse, zebrafish, drosophila, and worm mentioned above, are *developmental* in the sense that they model the organism at several stages of its development. However, only the fly and worm anatomy connect tissues in distinct stages by *descendant* or *develops from* links. The anatomy embedded in the MeSH terms is not developmental.

Although one may presume that anatomies using a rich set of labels support a richer set of queries, this is not the case. Typical queries are term based. The anatomical terms occurring in the query are then "exploded", that is, replaced by all terms in the hierarchy that are more specific. This means for instance, that a query for genes expressed in the mouse forelimb also returns genes that are only expressed in the paw. In anatomies where edges in the hierarchy are not labeled, usually all terms below the search terms are considered to be more specific, while in those with several labels it is common to include terms reachable via *component* and *instance* links. No existing biological data resource allows for queries that describe paths in the anatomies, like e.g., "genes expressed at developmental stage 4 or earlier in tissues that develop into the retina of the mouse eye."

4.2.3 Limitations of Current Formal Approaches to Anatomy

Anatomies, like many other ontologies are lacking a well-defined semantics and their current one is intuitive at best.

To give an example, the meaning of *component* links in the presence of *instance* links is not straightforward at all. Terms that have instances must stand for classes. However, what does it mean that class C_1 is a component of class C_2? There are at least two options: (1) every instance of C_1 is a component of some instance of C_2, or (2) every instance of C_2 has a component that is an instance of C_1. While both options may have arguments in their favour, only option 2 is compatible with a search strategy based on exploding terms. To see this, suppose a user is interested in genes expressed in tissues of class C_2. Then an explosion search will also return genes expressed in some tissue of class C_1. These are only answers to the intended query if every tissue in C_1 is part of some tissue of class C_2. To date, such semantic definitions of primitives are usually tacitly assumed by ontology designers, but have have not been formalised or otherwise made explicit. Only in rare cases have attempts at formalising anatomies been made (see e.g. [Burger et al., 2004]).

Often different anatomies follow fundamentally different approaches to capturing the structure of an organism. There is, for example, what might be called the butcher's view of an

organism, where the organism is virtually decomposed in space. Alternatively, there is a systemic approach, where the anatomical hierarchy follows organ systems like nervous system, blood circulation, skeleton, etc.

Ideally, for an anatomy that serves as a conceptual backbone for querying databases, one would like to see both approaches being realised. However, so far combinations have been only ad-hoc and give unsatisfactory results. In the Jackson Lab mouse anatomy, for instance, in embryonic stage 17, there are two nodes in the anatomy tree labeled with the term "central nervous system," one below the node "organ system," the other one below the node "tail". The semantics, as can be verified from the usage in the database, is that the first node represents the central nervous system in the body, but not the tail, while the second represents the part occurring in the tail.

Such limitations, like lack of formal semantics or ad-hoc mixture of incompatible approaches, are not serious as long as an anatomy is used to support access to a single database for a single species, and if the access mode is essentially browsing as opposed to querying. However, researchers in the area feel a need for more flexible access and for connecting anatomies for several species.

4.2.4 Integration of Anatomies

In the past few years various information resources for model organisms have been created, which are avaiable on the Web. In this situation biologists would like to combine both information on a single species residing in different resources and information across species.

Antomies are a key structure for accessing biological data and images. A necessary step therefore is to integrate anatomies within a single species and to link anatomies across species. Several initiatives are underway towards this goals. Within the "Standards and Ontologies for Functional Genomics" community, a working group has been established to integrate several anatomies of human and mouse and to link them with each other. The XSPAN project (www.xspan.org) aims at creating a web resource that holds several anatomies in a unified format together with cross species mappings between tissues.

4.3 EcoCyc

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 [Karp et al., 2004]. EcoCyc is presented to biologists using an encyclopaedia metaphor. It covers genes, metabolism, regulation and signal transduction, which a biologist can explore and use to visualise information. The knowledge base currently describes 695 enzymes encoded by a subset of these genes, 904 metabolic reactions and the organisation of these reactions into 129 metabolic pathways. EcoCyc uses the classification of gene product function from Riley as part of this description. Scientists can visualise the layout of genes within the chromosome, or of an individual biochemical reaction, or of a complete biochemical pathway (with compound structures displayed).

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. The user is not aware of this use of the ontology, except that the constraints expressed in the knowledge captured mean that the relationships on the data are captured precisely. In EcoCyc, for example, the concept of Gene is represented by a 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.

4.4 MBO

The Ontology for Molecular Biology (MBO) is an attempt to provide clarity and communication within the molecular biology database community [Schulze-Kremer,].

The MBO contains concepts and relationships that are required to describe biological objects, experimental procedures and computational aspects of molecular biology. It is very wide ranging and has over 1200 nodes representing both concepts and instances. In the conceptual part of the MBO, the primary relationship used is the 'is a kind of' relationship. The MBO has an organising, upper-level ontology. The root concept "Being' divides into 'object' and 'event'. 'Object', for instance, is subdivided into 'physical-' and 'abstract-' object. This helps give a precise classification for lower level concepts. The actual biological content of the MBO is currently relatively small, ending at quite large grained concepts such as Protein, Gene, and Chromosome. The framework, however, exists for extending the MBO much further into the biological domain.

4.5 RiboWeb

RiboWeb is a resource whose primary aim is to facilitate the construction of three-dimensional models of ribosomal components and compare the results to existing studies [Chen et al., 1997]. The knowledge that RiboWeb uses to perform these tasks is captured in four ontologies: The physical-thing ontology; the data ontology; the publication ontology and the methods ontology. The physical-thing ontology describes ribosomal components and associated 'physical things'. It has three principle conceptualisations: Molecules, Molecule-Ensembles and Molecule-Parts. The first describes covalently bonded molecules and includes the main biological macromolecules. Molecule-ensembles captures non-covalently bonded collections of molecules, such as enzyme complexes. The molecule-part ontology holds knowledge about regions of molecules that do not exist independently, but need to be talked about by biologists. These would include amino acid side chains and the 3' and 5' ends of nucleic acid molecules. The data ontology captures knowledge about experimental detail as well as data on the structure of physical-things. The methods ontology contains information about techniques for analysing data. It holds knowledge of which techniques can be applied to which data, as well as the input and outputs of each method.

Instances are added to RiboWeb that correspond to these concepts. For example, a publication in a peer-reviewed article describes the three-dimensional structure of the 30s ribosomal subunit. This means linked instances need to be created in the publication, data and physical-thing ontologies. A user may want to see if this structure is consistent with others captured within RiboWeb. The constraints described within RiboWeb can highlight conflicts with current knowledge to the biologist.

4.6 Tambis

TAMBIS (Transparent Access to Multiple Bioinformatics Information Sources) uses an ontology to enable biologists to ask questions over multiple external databases using a com-

mon query interface. The TAMBIS ontology (TaO) describes a wide range of bioinformatics tasks and resources, and has a central role within the TAMBIS system [Baker et al., 1999, Stevens et al., 2000a].

An interesting difference between the TaO and some of the other ontologies presented in this report, is that the TaO does not contain any instances. The TaO only contains knowledge about bioinformatics and molecular biology concepts and their relationships - the instances they represent still reside in the external databases. As concepts represent instances, a concept can act as a question. The concept Receptor Protein represents the instances of proteins with a receptor function and gathering these instances is answering that question.

The TaO is a dynamic ontology, in that it can grow without the need for either conceptualising or encoding new knowledge. In contrast, the other ontologies described here are static - developers must intervene and encode new conceptualisation to form new concepts. The TaO uses rules within the ontology to govern what concepts can be joined to another concept via relationships, to form new concepts. Thus the TaO places great emphasis on relations. A user can form a complex, multi-source query, using relationships, in the following manner. Starting with the concept Protein, the TaO is consulted as to which relationships can be used to join Protein to other concepts. Amongst many, the following two are offered: is homologous to Protein and hasAccessionNumber AccessionNumber. Initially, the original Protein is extended to give a new concept Protein isHomologous to Protein (The concept Protein Protein homologue); then the second 'protein' is extended with hasAccessionNumber AccessionNumber. The resulting concept ('Protein homologue of Protein with Accession Number') describes proteins which are homologous to protein with a particular accession number. This concept can be used as a source independent query containing no information on how to answer such a query. The rest of the TAMBIS system takes this conceptual query and processes it to an executable program against the external sources.

The TaO is available in two forms - a small model that concentrates on proteins and a larger scale model that includes nucleic acids. The small TaO, with 250 concepts and 60 relationships, describes Proteins and enzymes, as well as their motifs, secondary and tertiary structure, functions and processes. There is also supporting material on subcellular structure and chemicals, including cofactors. Motifs extend to detail such as the principal modification sites; function and process to broad classifications such as Hormone and Receptor, and Apoptosis and Lactation; structure extends to detail such as gross architecture - for example, SevenPropellor. Important relationships include is component of, has name, has function and is homologous to, as well as many more. The larger model, with 1500 concepts, broadens these areas to include concepts pertinent to nucleic acid, its children and genes.

4.7 Editing, Browsing, and Mapping Ontologies

There are many tools to edit, browse and map ontologies. As GeneOntology is the most prominent bioinformatics ontology, we consider the following list of editors and browsers (for details see www.geneontology.org).

- DAG-Edit: This Java application provides an interface to browse, query and edit GO or any other vocabulary that has a DAG data structure.
- COBrA: COBrA is an ontology editing and mapping tool for GO and OBO ontologies developed by AIAI and the XSPAN project. COBrA displays two ontologies simultaneously for the purpose of defining mappings between terms, e.g. between tissues and cell

types. Export and import to/from GO flat file, GO RDF, plus RDFS and OWL formats are supported. Ontology merging and inference is also possible.

- AmiGO from BDGP: With AmiGO, you can search for a GO term and view all gene products annotated to it, or search for a gene product and view all its associations. You can also browse the ontologies to view relationships between terms as well as the number of gene products annotated to a given term.

- MGI GO Browser: With the MGI GO Browser, you can search for a GO term and view all mouse genes annotated to the term or any subterms. You can also browse the ontologies to view relationships between terms, term definitions, as well as the number of mouse genes annotated to a given term and its subterms.

- QuickGO at EBI: With QuickGO, a GO browser integrated into InterPro at the EBI, you can search for a GO term to see its relationships and definition, as well as any available mappings to SWISS-PROT keywords, to the Enzyme Classification or Transport Classification databases, or to InterPro entries.

- EP GO Browser: The EP:GO browser is built into EBI's Expression Profiler, a set of tools for clustering, analysis and visualisation of gene expression and other genomic data. With it, you can search for GO terms and identify gene associations for a node, with or without associated subnodes, for the organism of your choice.

- GoFish: The GoFish program, available as a Java applet, allows the user to construct arbitrary Boolean queries using GO attributes, and orders gene products according to the extent they satisfy such queries. GoFish also estimates, for each gene product, the probability that they satisfy the Boolean query.

- GenNav: GenNav is a GO browser developed at NLM. It searches GO terms and annotated gene products, and provides a graphical display of a term's position in the GO DAG.

- GeneOntology@RZPD: With the GeneOntology@RZPD tool at the Resource Center/Primary Database (RZPD) in Germany, you can search for GO identifiers associated with UniGene ClusterIds, Genes (Name/Symbol) and Clones provided by the RZPD. You can also search for UniGene Clusters, Genes and Clones annotated with a certain GO identifier or a combination of GO identifiers. So far, GO annotations for human and mouse genes/clones are linked.

- ProToGO: ProToGO, developed at the Hebrew University in Jerusalem, searches the GOA@EBI and Compugen annotation datasets. The output is a graphical view of the relevant sub-graph of GO, containing those GO terms assigned to the query proteins. Documentation is provided.

- CGAP GO Browser: With the GO browser at the The Cancer Genome Anatomy Project, you can browse through the GO vocabularies, and find human and mouse genes assigned to each term. The help documentation is at: http://cgap.nci.nih.gov/Genes/AllAboutGO.

- GOBrowser: GO terms are presented in an explorer-like browser, and behaviour can be configured by altering Perl scripts.

- TAIR: Keyword Browser: The TAIR Keyword Browser, developed at The Arabidopsis Information Resource searches and browses for Gene Ontology, TAIR Anatomy, and TAIR Developmental stage terms, and allows you to view term details and relationships among terms. It includes links to genes, publications, microarray experiments and annotations associated with the term or any children terms.

- PANDORA: With PANDORA, developed at The Hebrew University of Jerusalem you can search for any non-uniform sets of proteins and detect subsets of proteins that share unique biological properties and the intersections of such sets. PANDORA supports GO annotations as well as additional keywords (from UniProt Knowledgebase, InterPro, ENZYME, SCOP etc). It is also integrated into the ProtoNet system, thus allowing testing of thousands of automatically generated protein families.

A number of other ontology engineering tools have been developed and used in bioinformatics. However, only few evaluations of ontology tools using bio-ontologies have been performed. In [Lambrix et al., 2003] Protégé-2000 [Noy et al., 2001], Chimaera [McGuinness et al., 2000], OilEd [R et al., 2001] and DAG-Edit were evaluated as ontology development tools using GO ontologies as test ontologies. In [Lambrix and Edberg, 2003] PROMPT [Noy and Musen, 2000] (plug-in for Protégé-2000) and Chimaera were evaluated as ontology merging tools. An extension of this evaluation is found in [Lambrix and Tan, 2004]. First a number of merging algorithms are evaluated and then the SAMBO tool is compared to PROMPT and Chimaera. The test ontologies were GO ontologies, Signal-Ontology, MeSH and the Anatomical Dictionary for the Adult Mouse. In [Yeh et al., 2003] Protégé-2000 was assessed for its use in developing and maintaining GO ontologies.

Much research has been dedicated to mapping and comparing different ontologies. [Cantor et al., 2003] evaluate mappings between GeneOntology and UMLS, the Unified Medical Language System, which is focused on clinical medicine focusing on precision and recall, which range from 0.65 to 0.90 and 0.59 to 0.93, respectively. [McCray et al., 2002] found that in the order of 50% of GO-terms relating to function can be mapped to the ontologies MeSH and SNOMED, while GO-terms covering processes and localisation are largely absent from MeSH and SNOMED.

5 Rules in Bioinformatics

A natural extension of factual knowledge are rules and logic programming. Here we have three application areas in which rules are used: Rules to represent metabolic pathways and query them in computation tree logic; Rules as constraints for structure prediction; and rules learned by inductive logic programming to declaratively capture biological knowledge.

5.1 Modelling Networks with Computation Tree Logic

This idea of introducing formal methods was mentioned as a prime motivation by early efforts at formal modeling [Regev et al., 2001a], and the specific prospects of using computational logics were clearly articulated in [Eker et al., 2002a]. Substance was given to this idea by proposing the use of the Computation Tree Logic CTL as a query language for biomolecular networks [Chabrier et al., 2004].

In networks biology, the complexity of the systems at hand (metabolic networks, extracellular and intracellular networks, networks of gene regulation) clearly shows the necessity of software tools for reasoning globally about biological systems. Several formalisms have been proposed in recent years for modeling biochemical processes either qualitatively [Regev et al., 2001b, Nagasaki et al., 2000, Eker et al., 2002b] or quantitatively [Matsuno et al., 2000, Hofestädt and Thelen, 1998, Alur et al., 2001, Ghosh and Tomlin, 2001, Bockmayr and Courtois, 2002]. State-of-the-art tools integrate a graphical user interface and a simulator, yet few formal tools are available for reasoning about these processes and proving properties of them. The focus in Biocham [Chabrier-Rivier et al., 2004] has been on the design of a biochemical rule language and a query language of the model in temporal logic, that are intended to be used by biologists.

Biocham is a language and a programming environment for modeling biochemical systems, making simulations, and querying such models in temporal logic. Biocham is composed of :

1. a rule-based language for modeling biochemical systems,
2. a simple simulator,
3. a powerful query language based on Computation Tree Logic CTL,
4. an interface to the NuSMV [Cimatti et al., 2002] model checker for automatically evaluating CTL queries.

Biocham shares several similarities with the Pathway Logic system implemented in Maude. Both systems rely on an algebraic syntax and are rule-based languages. One difference is the use in Biocham of CTL logic which allows us to express a wide variety of biological queries, and the use of a state-of-the-art symbolic model checker for handling the complexity of highly non-deterministic models.

5.2 Bioinformatics and Constraint

The protein structure prediction is one of the most important unsolved problems of computational biology. It can be specified as follows: Given a protein by its sequence of amino acids, what is its native structure? NP-completeness of the problem has been proven for many different models (including lattice and off-lattice models) [Berger and Leighton, 1998, Crescenzi et al., 1998]. These results strongly suggest that the protein folding problem is NP-hard in general. Therefore, it is unlikely that a general, efficient algorithm for solving this problem can be given. Actually, the situation is even worse, since the general principles why natural proteins fold into a native structure are unknown. This is cumbersome since rational design is commonly viewed to be of paramount importance e.g. for drug design, where one faces the difficulty to design proteins that have a unique and stable native structure.

To tackle structure prediction and related problems simplified models have been introduced. These simplified models have been successfully used by several groups in the international contest on automated structure prediction (see the meeting review of CASP3 [Koehl and Levitt, 1999]). They are used in hierarchical approaches for protein folding [Xia et al., 2000]. Given a protein sequence, then one first enumerates all low (or minimal) energy conformations in a simplified model. In the simplified model, only some aspects of the protein structure are modeled (for which reason they are also called low-resolution or coarse-grained protein models). Then, the say 10 000 best structures are taken and fine-tuned

using other methods. These methods usually incorporate biological knowledge and simulation of protein folding on full atomic detail (i.e. molecular dynamics simulation).

Constraint-based approaches to structure prediction express distance constraints on consecutive and adjacent monomers. E.g. to express that the distance between two successive monomers is 1, we introduce for every monomer i the three variables $\texttt{Xdiff}_i$, $\texttt{Ydiff}_i$ and $\texttt{Zdiff}_i$. The value range of these variables is [0..1]. Then we can express the unit-vector distance constraint by

$$\begin{aligned} \texttt{Xdiff}_i &=: \texttt{Abs}(\texttt{X}_i - \texttt{X}_{i+1}) \\ \texttt{Ydiff}_i &=: \texttt{Abs}(\texttt{Y}_i - \texttt{Y}_{i+1}) \\ \texttt{Zdiff}_i &=: \texttt{Abs}(\texttt{Z}_i - \texttt{Z}_{i+1}) \\ \texttt{Sum}[\texttt{Xdiff}_i, \texttt{Ydiff}_i, \texttt{Zdiff}_i] &=: 1 \end{aligned}$$

On some models such a constraint-based approach can outperform other existing approaches such as Monte-Carlo simulated annealing (e.g. [MacDonald et al., 2000, Dinner et al., 1996]), genetic algorithms (e.g. [Unger and Moult, 1996]), purely heuristic methods like hydrophobic zipper [Dill et al., 1993] and the chain growth algorithm [Bornberg-Bauer, 1997], as well as complete enumeration (often restricted to subset of all conformations, e.g. [Šali et al., 1994, Xia et al., 2000]).

5.3 Bioinformatics and Inductive Logic Programming (ILP)

Inductive logic programming uses examples and background knowledge to construct hypotheses, which explain the examples using the background knowledge. Muggleton et al. have applied inductive logic programming to a wide variety of biological problems such as the following:[1]

- Structure-activity prediction: In [King et al., 1992] it was shown that ILP system Golem [Muggleton and Feng, 1992] was capable of constructing rules which accurately predict the activity of untried drugs. Rules were constructed from examples of drugs with known medicinal activity. The accuracy of the rules was found to be slightly higher than traditional statistical methods. More importantly the easily understandable rules provided insights which were directly comparable to the relevant literature concerning the binding site of dihydrofolate reductase.

- Mutagenesis: In [King et al., 1996, Srinivasan et al., 1996] ILP system Progol [Muggleton, 1995] was used to predict the mutagenicity of chemical compounds taken from a previous study in which linear regression had been applied. Progol's predictive accuracy was equivalent to regression on the main set of 188 compounds and significantly higher (85.7% as opposed to 66.7%) on 44 compounds which had been discarded by the previous authors as unpredictable using regression. Progol's single clause solution for the 44 compounds was judged by the domain experts to be a new structural alert for mutagenesis.

- Pharmacophores: In a series of "blind tests" in collaboration with the pharmaceutical company Pfizer UK, Progol was shown [Finn et al., 1998] capable of re-discovering a 3D description of the binding sites (or pharmacophores) of ACE inhibitors (a hypertension drug) and an HIV-protease inhibitor (an anti-AIDS drug).

[1] This survey is taken from www.doc.ic.ac.uk/~muggleton

- Carcinogenicity: Progol was entered into a world-wide carcinogenicity prediction competition run by the National Toxicology Program (NTP) in the USA. Progol was trained on around 300 available compounds, and made use of its earlier rules relating to mutagenicity. In the first round of the competition Progol produced the highest predictive accuracy of any automatic system entered [Srinivasan et al., 1997].

- Proteins: Protein secondary structure prediction. In [Muggleton et al., 1992] Golem was applied to one of the hardest open problems in molecular biology. The problem is as follows: given a sequence of amino acid residues, predict the placement of the main three dimensional sub-structures of the protein. The problem is of great interest to pharmaceutical companies involved with drug design. For this reason, over the last 20 years many attempts have been made to apply methods ranging from statistical regression to decision tree and neural net learning to this problem. Published accuracy results for the general prediction problem have ranged between 50 and 60%, very close to majority-class prediction rates. In an investigation it was found that the ability to make use of background knowledge from molecular biology, together with the ability to describe structural relations boosted the predictivity for a restricted sub-problem to around 80% on an independently chosen test set.

- Discovery of fold descriptions: Protein shape is usually described at various levels of abstraction. At the lower levels each family of proteins contains members with high sequence similarity. At the most abstract level folds describe proteins which have similar overall shape but are very different at the sequence level. The lack of understanding of shape determination has made protein fold prediction particularly hard. However, it is intriguing that although there are around 300 known folds, around half of all known proteins are members of the 20 most populated folds. Progol was applied to discover rules governing these 20 most populated protein folds. Average in class cross-validated prediction was around 70% and many of the rules were judged to be good characterisations of the fold classes.

References

[Abkevich et al., 1997] Abkevich, V., Gutin, A., and Shakhnovich, E. (1997). Computer simulations of prebiotic evolution. In Altman, R. B., Dunker, A. K., Hunter, L., and Klein, T. E., editors, *PSB'97*, pages 27–38.

[Abkevich et al., 1995] Abkevich, V. I., Gutin, A. M., and Shakhnovich, E. I. (1995). Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *Journal of Molecular Biology*, 252:460–471.

[Agarwala et al., 1997] Agarwala, R., Batzoglou, S., Dancik, V., Decatur, S. E., Hannenhalli, S., Farach, M., Muthukrishnan, S., and Skiena, S. (1997). Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the hp model. *Journal of Computational Biology*, 4(3):275–96.

[Akutsu and Miyano, 1997] Akutsu, T. and Miyano, S. (1997). On the approximation of protein threading. In *Proc. of the First Annual International Conferences on Compututational Molecular Biology (RECOMB97)*. ACM Press.

[Alur et al., 2001] Alur, R., Belta, C., Ivancic, F., Kumar, V., Mintz, M., Pappas, G. J., Rubin, H., and Schug, J. (2001). Hybrid modeling and simulation of biomolecular networks. In Springer, editor, *Hybrid Systems: Computation and Control*, LNCS 2034, pages 19–32, Rome, Italy.

[Backofen and Will, 2004] Backofen, R. and Will, S. (2004). Local sequence-structure motifs in RNA. *Journal of Bioinformatics and Computational Biology (JBCB)*. accepted for publication, preprint.

[Baker et al., 1999] Baker, P., Goble, C., Bechhofer, S., Paton, N., Stevens, R., and Brass, A. (1999). An ontology for bioinformatics applications. *Bioinformtaics*, 15(6):510–20.

[Baldock et al., 2003] Baldock, R., Bard, J., Burger, A., Burton, N., Christiansen, J., Feng, G., Hill, B., Houghton, D., Kaufman, M., Rao, J., Sharpe, J., Ross, A., Stevenson, P., Venkataraman, S., Waterhouse, A., Yang, Y., and Davidson, D. (2003). EMAP and EMAGE: A framework for understanding spatially organised data. *Neuroinformatics*, 1:309–325.

[Bard et al., 1998] Bard, J., Kaufman, M., Dubreuil, C., Brune, R., Burger, A., Baldock, R., and Davidson, D. (1998). An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. *Mechanisms of Development*, 74:111–120.

[Berger and Leighton, 1998] Berger, B. and Leighton, T. (1998). Protein folding in the hydrophobic-hydrophilic (HP) modell is NP-complete. In *Proc. of the RECOMB'98*, pages 30–39.

[Bockmayr and Courtois, 2002] Bockmayr, A. and Courtois, A. (2002). Using hybrid concurrent constraint programming to model dynamic biological systems. In Springer, editor, *18th International Conference on Logic Programming*, pages 85–99, Copenhagen.

[Bornberg-Bauer, 1997] Bornberg-Bauer, E. (1997). Chain growth algorithms for HP-type lattice proteins. In *Proc. of the* 1^{st} *Annual International Conference on Computational Molecular Biology (RECOMB)*, pages 47 – 55. ACM Press.

[Bornberg-Bauer and Chan, 1999] Bornberg-Bauer, E. and Chan, H. (1999). Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc Natl Acad Sci U S A*, 96(19):10689–94.

[Burger et al., 2004] Burger, A., Davidson, D., and Baldock, R. (2004). Formalization of mouse embryo anatomy. *Bioinformatics*, 20(2):259–267.

[Cantor et al., 2003] Cantor, M., Sarakr, I., Gelman, R., Hartel, F., Bodenreider, O., and Lussier, Y. (2003). An evaluation of hybrid methods for matching biomedical terminologies: Mapping the GeneOntology to the UMLS. *Stud Health Technol Inform.*, 95:62–7.

[Chabrier et al., 2004] Chabrier, N., Chiaverini, M., Danos, V., Fages, F., and Schächter, V. (2004). Modeling and querying biochemical networks. *Theoretical Computer Science*, To appear.

[Chabrier-Rivier et al., 2004] Chabrier-Rivier, N., Fages, F., and Soliman, S. (2004). The biochemical abstract machine BIOCHAM. In Danos, V. and Schächter, V., editors, *CMSB'04: Proceedings of the second Workshop on Computational Methods in Systems Biology*, Lecture Notes in Computer Science. Springer-Verlag.

[Chen et al., 1997] Chen, R., Felciano, R., and Altman., R. (1997). RIBOWEB: linking structural computations to a knowledge base of published experimental data. In *Proc Int Conf Intell Syst Mol Biol*, volume 5, pages 84–7.

[Cimatti et al., 2002] Cimatti, A., Clarke, E., Enrico Giunchiglia, F. G., Pistore, M., Roveri, M., Sebastiani, R., and Tacchella, A. (2002). Nusmv 2: An opensource tool for symbolic model checking. In *Proceedings of the International Conference on Computer-Aided Verification, CAV'02*, Copenhagen, Danmark.

[Crescenzi et al., 1998] Crescenzi, P., Goldman, D., Papadimitriou, C., Piccolboni, A., and Yannakakis., M. (1998). On the complexity of protein folding. In *Proc. of STOC*. To appear. Short version in *Proc. of RECOMB'98*, pages 61–62.

[Davidson et al., 2001] Davidson, S., Crabtree, J., Brunk, B., Schug, J., Tannen, V., Overton, C., and Stoeckert, C. (2001). K2/kleisli and gus: Experiments in integrated access to genomic data sources. *IBM Systems Journal, Issue on Deep computing for the life sciences*, 40(2):512–531.

[Davidson et al., 1997] Davidson, S., Overton, C., Tannen, V., and Wong, L. (1997). Biokleisli: A digital library for biomedical researchers. *Journal of Digital Libraries*, 1(1):36–53.

[de Jong, 2001] de Jong, H. (2001). Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):69–105.

[Dill et al., 1995] Dill, K., Bromberg, S., Yue, K., Fiebig, K., Yee, D., Thomas, P., and Chan, H. (1995). Principles of protein folding – a perspective of simple exact models. *Protein Science*, 4:561–602.

[Dill et al., 1993] Dill, K. A., Fiebig, K. M., and Chan, H. S. (1993). Cooperativity in protein-folding kinetics. *Proc. Natl. Acad. Sci. USA*, 90:1942 – 1946.

[Dinner et al., 1996] Dinner, A. R., Sali, A., and Karplus, M. (1996). The folding mechanism of larger model proteins: role of native structure. *Proc. Natl. Acad. Sci. USA*, 93(16):8356–61.

[Dovier et al., 2002] Dovier, A., Burato, M., and Fogolari, F. (2002). Using secondary structure information for protein folding in clp(fd). In *Proc. of Workshop on Functional and Constraint Logic Programming*, volume ENTCS vol. 76.

[Eckman et al., 2001] Eckman, B., Kosky, A., and Laroco, L. (2001). Extending traditional query-based integration approaches for functional characterization of post-genomic data. *Bioinformatics*, 17(7):579–670.

[Eker et al., 2002a] Eker, S., Knapp, M., Laderoute, K., Lincoln, P., Meseguer, J., and Sonmez, K. (2002a). Pathway logic: Symbolic analysis of biological signaling. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 400–412.

[Eker et al., 2002b] Eker, S., Knapp, M., Laderoute, K., Lincoln, P., Meseguer, J., and Sönmez, M. K. (2002b). Pathway logic: Symbolic analysis of biological signaling. In *Proceedings of the seventh Pacific Symposium on Biocomputing*, pages 400–412.

[Finn et al., 1998] Finn, P., Muggleton, S., Page, D., and Srinivasan, A. (1998). Pharmacophore discovery using the inductive logic programming system progol. *Machine Learning*, 30:241–71.

[GeneOntologyConsortium, 2004] GeneOntologyConsortium (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, 1(32):D258–61.

[Ghosh and Tomlin, 2001] Ghosh, R. and Tomlin, C. (2001). Lateral inhibition through delta-notch signaling: A piecewise affine hybrid model. In Springer, editor, *Hybrid Systems: Computation and Control*, LNCS 2034, pages 232–246, Rome, Italy.

[Goble et al., 2001] Goble, C., Stevens, R., Ng, G., Bechhofer, S., Paton, N., Baker, P., Peim, M., and Brass, A. (2001). Transparent access to multiple bioinformatics information sources. *IBM Systems Journal, Issue on Deep computing for the life sciences*, 40(2):532–551.

[Govindarajan and Goldstein, 1997] Govindarajan, S. and Goldstein, R. A. (1997). The foldability landscape of model proteins. *Biopolymers*, 42(4):427–438.

[Gruber, 1993] Gruber, T. R. (1993). Toward principles for the design of ontologies used for knowledge sharing. *Knowledge Acquisition*, 5:199–220.

[Guarino and Giaretta, 1995] Guarino, N. and Giaretta, P. (1995). Ontologies and knowledge bases: Towards a terminological clarification. In Mars, N., editor, *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, pages 25–32. IOS Press.

[Hart and Istrail, 1996] Hart, W. E. and Istrail, S. C. (1996). Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal. *Journal of Computational Biology*, 3(1):53–96.

[Hinds and Levitt, 1996] Hinds, D. A. and Levitt, M. (1996). From structure to sequence and back again. *Journal of Molecular Biology*, 258:201–209.

[Höchsmann et al., 2003] Höchsmann, M., Töller, T., Giegerich, R., and Kurtz, S. (2003). Local similarity in rna secondary structures. In *Proceedings of Computational Systems Bioinformatics (CSB 2003)*.

[Hofestädt and Thelen, 1998] Hofestädt, R. and Thelen, S. (1998). Quantitative modeling of biochemical networks. In *In Silico Biology*, volume 1, pages 39–53.

[Jasper and Uschold, 1999] Jasper, R. and Uschold, M. (1999). A framework for understanding and classifying ontology applications. In *Proc. 12th Workshop on Knowledge Acquisition, Modeling and Management*, Banff, Canada.

[Jiang et al., 2002] Jiang, T., Lin, G., Ma, B., and Zhang, K. (2002). A general edit distance between RNA structures. *Journal of Computational Biology*, 9(2):371–88.

[Karp et al., 2004] Karp, P., Arnaud, M., Collado-Vides, J., Ingraham, J., Paulsen, I., and Jr, M. S. (2004). The E. coli EcoCyc database: No longer just a metabolic pathway database. *ASM News*, 70(1):25–30.

[Kemp et al., 2000] Kemp, G., Angelopoulos, N., and Gray, P. (2000). A schema-based approach to building a bioinformatics database federation. In *Proceedings of the IEEE International Symposium on Bioinformatics and Biomedical Engineering*, pages 13–20.

[King et al., 1992] King, R., Muggleton, S., Lewis, R., and Sternberg, M. (1992). Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proceedings of the National Academy of Sciences*, 23(11322-26).

[King et al., 1996] King, R., Muggleton, S., Srinivasan, A., and Sternberg, M. (1996). Structure-activity relationships derived by machine learning: the use of atoms and their bond connectives to predict mutagenicity by inductive logic programming. *Proceedings of the National Academy of Sciences*, 93:438–442.

[Koehl and Levitt, 1999] Koehl, P. and Levitt, M. (1999). A brighter future for protein structure prediction. *Nature Structural Biology*, 6:108–111.

[Lambrix, 2004] Lambrix, P. (2004). Ontologies in bioinformatics and systems biology. In Dubitzky, W. and Azuaje, F., editors, *Artificial Intelligence Methods and Tools for Systems Biology*. Kluwer Academic Press. to appear.

[Lambrix and Edberg, 2003] Lambrix, P. and Edberg, A. (2003). Evaluation of ontology merging tools in bioinformatics. In *Proc. Pacific Symposium on Biocomputing*, pages 589–600, Kauai, Hawaii, USA.

[Lambrix et al., 2003] Lambrix, P., Habbouche, M., and Pérez, M. (2003). Evaluation of ontology development tools for bioinformatics. *Bioinformatics*, 19(12):1564–1571.

[Lambrix and Jakonienė, 2003] Lambrix, P. and Jakonienė, V. (2003). Towards transparent access to multiple biological databanks. In *Proc. 1st Asia-Pacific Bioinformatics Conference*, pages 53–60, Adelaide, Australia.

[Lambrix and Tan, 2004] Lambrix, P. and Tan, H. (2004). Merging daml+oil ontologies. In *Proceedings of the 6th International Baltic Conference on Databases and Information Systems*, Riga, Latvia.

[Lancia et al., 2001] Lancia, G., Carr, R., Walenz, B., and Istrail, S. (2001). 101 optimal PDB structure alignments: a branch-and-cut algorithm for the maximum contact map overlap problem. In *Proc. of the Fifth Annual International Conferences on Compututational Molecular Biology (RECOMB01)*. ACM Press.

[Lesk, 2002] Lesk, A. (2002). *Introdcuton to Bioinformatics*. OUP.

[MacDonald et al., 2000] MacDonald, D., Joseph, S., Hunter, D. L., Moseley, L. L., Jan, N., and Guttmann, A. J. (2000). Self-avoiding walks on the simple cubic lattice. *J. Phys. A: Math. Gen.*, 33:5973–5983.

[Maimon and Browning, 2001] Maimon, R. and Browning, S. (2001). Diagrammatic notation and computational structure of gene networks. In *Proceedings of the 2nd International Conference on Systems Biology*.

[Matsuno et al., 2000] Matsuno, H., Doi, A., Nagasaki, M., and Miyano, S. (2000). Hybrid Petri net representation of gene regulatory network. In *Pacific Symposium on Biocomputing (5)*, pages 338–349.

[McCray et al., 2002] McCray, A. T., Browne, A., and Bodenreider, O. (2002). The lexical properties of the Gene Ontology (GO). In *Proc. of AMIA Symp.*, pages 504–8.

[McGuinness et al., 2000] McGuinness, D., Fikes, R., Rice, J., and Wilder, S. (2000). The chimaera ontology environment. In *Proc. 17th National Conference on Artificial Intelligence*, pages 1123–1124, Austin, TX, USA.

[Milner et al., 1992] Milner, R., Parrow, J., and Walker, D. (1992). A calculus of mobile processes I and II. *Information and Computation*, 100:1–41, 42–78.

[Muggleton, 1995] Muggleton, S. (1995). Inverse entailment and progol. *New Generation Computing*, 13:245–86.

[Muggleton and Feng, 1992] Muggleton, S. and Feng, C. (1992). Efficient induction of logic programs. In *Inductive Logic Programming*, pages 281–98. Academic Press.

[Muggleton et al., 1992] Muggleton, S., King, R., and Sternberg, M. (1992). Protein secondary structure prediction using logic-based machine learning. *Protein Engineering*, 5(7):647–57.

[Nagasaki et al., 2000] Nagasaki, M., Onami, S., Miyano, S., and Kitano, H. (2000). Bio-calculus: Its concept, and an application for molecular interaction. In *Currents in Computational Molecular Biology.*, volume 30 of *Frontiers Science Series*.

[Needleman and Wunsch, 1970] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–53.

[Noy and Musen, 2000] Noy, N. and Musen, M. (2000). Prompt: Algorithm and tool for automated ontology merging and alignment. In *Proc. 17th National Conference on Artificial Intelligence*, pages 450–455, Austin, Texas, USA.

[Noy et al., 2001] Noy, N., Sintek, M., Decker, S., Crubézy, M., Fergerson, R., and Musen, M. (2001). Creating semantic web contents with protégé-2000. *IEEE Intelligent Systems*, March/April:60–71.

[Ortiz et al., 1998] Ortiz, A. R., Kolinski, A., and Skolnick, J. (1998). Combined multiple sequence reduced protein model approach to predict the tertiary structure of small proteins. In Altman, R. B., Dunker, A. K., Hunter, L., and Klein, T. E., editors, *PSB'98*, volume 3, pages 375–386.

[R et al., 2001] R, R. S., Horrocks, I., Goble, C., and Bechhofer, S. (2001). Building a reasonable bioinformatics ontology using oil. In *Proc. IJCAI Workshop on Ontologies and Information Sharing*.

[Regev et al., 2001a] Regev, A., Silverman, W., and Shapiro, E. (2001a). Representation and simulation of biochemical processes using the pi-calculus process algebra. In *Proceedings of the Pacific Symposium of Biocomputing*, pages 6:459–470.

[Regev et al., 2001b] Regev, A., Silverman, W., and Shapiro, E. Y. (2001b). Representation and simulation of biochemical processes using the pi-calculus process algebra. In *Proceedings of the sixth Pacific Symposium of Biocomputing*, pages 459–470.

[Ringwald et al., 2001] Ringwald, M., Eppig, J., Begley, D., Corradi, J., McCright, I., Hayamizu, T., Hill, D., JA, J. K., and Richardson, J. (2001). The mouse gene expression database. *Nucleic Acids Research*, 29:98–101.

[Schoeberl et al., 2002] Schoeberl, B., Eichler-Jonsson, C., Gilles, E., and Müller, G. (2002). Computational modeling of the dynamics of the map kinase cascade activated by surface and internalized EGF receptors. *Nature Biotechnology*, 20:370–375.

[Schulze-Kremer,] Schulze-Kremer, S. Ontology for molecular biology (mbo). http://igd.rz-berlin.mpg.de/ www/oe/mbo.html.

[Sprague et al., 2001] Sprague, J., Doerry, E., Douglas, S., and Westerfield, M. (2001). The Zebrafish Information Network (ZFIN): a resource for genetic, genomic and developmental research. *Nucleic Acids Research*, 29:87–90.

[Srinivasan et al., 1997] Srinivasan, A., Muggleton, R. K. S., and Sternberg, M. (1997). Carcinogenesis predictions using ilp. In *Proceedings of the Seventh International Workshop on Inductive Logic Programming*. LNAI 1297 Springer.

[Srinivasan et al., 1996] Srinivasan, A., Muggleton, S., King, R., and Sternberg, M. (1996). Theories for mutagenicity: a study of first-order and feature based induction. *Artificial Intelligence*, 85(1,2):277–99.

[Stein et al., 2001] Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., and Spieth, J. (2001). WormBase: network access to the genome and biology of caenorhabditis elegans. *Nucleic Acids Research*, 29:82–86.

[Stevens et al., 2000a] Stevens, R., Baker, P., Bechhofer, S., Ng, G., Jacoby, A., Paton, N., Goble, C., and Brass, A. (2000a). TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformtaics*, 16(2):184–5.

[Stevens et al., 2001] Stevens, R., Goble, C., Baker, P., and Brass, A. (2001). A classification of tasks in bioinformatics. *Bioinformatics*, 17(2):180–188.

[Stevens et al., 2000b] Stevens, R., Goble, C., and Bechhofer, S. (2000b). Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics*, 1(4):398–414.

[The FlyBase Consortium, 2003] The FlyBase Consortium (2003). The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Research*, 31:172–175.

[Thieffry and Thomas, 1998] Thieffry, D. and Thomas, R. (1998). Qualitative analysis of gene networks. In Altman, R. B., Dunker, A. K., Hunter, L., and Klein, T. E., editors, *Pacific Symposium on Biocomputing*, volume 3, pages 77–88, Singapore. World Scientific Press.

[Unger and Moult, 1996] Unger, R. and Moult, J. (1996). Local interactions dominate folding in a simple protein model. *Journal of Molecular Biology*, 259:988–994.

[Visna and Gilbert, 2002] Visna, J. and Gilbert, D. (2002). Pattern matching and pattern discovery algorithms for protein topologies. Paper as Print Copy.

[Šali et al., 1994] Šali, A., Shakhnovich, E., and Karplus, M. (1994). Kinetics of protein folding. *Journal of Molecular Biology*, 235:1614–1636.

[Xia et al., 2000] Xia, Y., Huang, E. S., Levitt, M., and Samudrala, R. (2000). Ab initio construction of protein tertiary structures using a hierarchical approach. *Journal of Molecular Biology*, 300:171 – 185.

[Yeh et al., 2003] Yeh, I., Karp, P., Noy, N., and Altman, R. (2003). Knowledge acquisition, consistency checking and concurrency control for gene ontology (go). *Bioinformatics*, 19(12):241–248.

[Yue and Dill, 1995] Yue, K. and Dill, K. A. (1995). Forces of tertiary structural organization in globular proteins. *Proc. Natl. Acad. Sci. USA*, 92:146 – 150.