Ontologies in Bioinformatics

Robert Stevens, Chris Wroe, Phillip Lord and Carole Goble

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

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

32.1 Introduction

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

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

The biological sciences, especially molecular biology, currently lack the laws and mathematical support of sciences such as physics and chemistry. This is not to say that the biological sciences lack principles and understanding that, for instance, in physics allows us to predict planetary orbits, behaviour of waves and particles etc. We cannot, however, yet take a protein sequence and from the amino acid residues present deduce the structure, molecular function, biological role or location of that protein. The biologist has two options: First, to perform many laboratory experiments, in vitro and in vivo to acquire knowledge about the protein; second, the biologist takes advantage of one of the principles of molecular biology, which is that sequence is related to molecular function and structure. Therefore, a biologist can compare the protein sequence to others that are already well characterised. If the uncharacterised sequence is sufficiently similar to a characterised sequence, then it is inferred that the characteristics of one can be transferred to the other. So a key tool of bioinformatics is the sequence similarity search [4]; the characterisation of single sequences lies at the heart of most bioinformatics, even the new high-throughput techniques that investigate the modes of action of thousands of proteins per experiment. As the first method is expensive, both in terms of time and money, the latter can reduce the time to characterise unknown biological entities. Thus, we often see a cycle between laboratory bench and the computer.

32.1.1 Describing and Using Biological Data

It has been said that biology is a knowledge based discipline [7]. Much of the community’s knowledge is contained within the community’s data resources. A typical resource is the SWISS-PROT protein database [6]. The protein sequence data itself is a relatively small part of the entry. Most of the entry is taken up by what the bioinformatics community refers to as ‘annotation’ which describe: physico-chemical features of the protein; comments on the whole sequence, such as function, disease, regulation, expression; species; names and so on. All this can be considered as the knowledge component of the database. Figure 32.1 shows a typical annotation from SWISS-PROT; note that the knowledge is captured as textual terms describing the findings, not numeric data, making use of shared keywords and controlled vocabularies. Whilst this style of representation is suitable for human readers, the current representation of the knowledge component is difficult to process by machine. SWISS-PROT itself now has over 100,000 entries (and growing exponentially), so