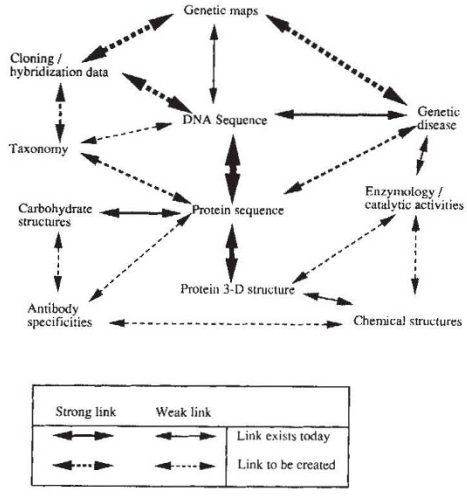
ther experiment. In short, bioinformatics makes the best use of the data we have.

LIMS and legs

Bioinformatics stands on three legs. The first is at the laboratory bench. Visionary suggestions for totally integrated Laboratory Information Management Systems (LIMS)[2,3] are starting to be implemented by large-scale cloning and sequencing groups, such as the MRC Laboratory of Molecular Biology (Cambridge, U.K.) and the Imperial Cancer Research Fund (London, U.K.). Several companies such as IntelliGenetics (Mountain View, CA) and Applied Biosystems (Foster City, CA) have also realised that the value of automated molecular genetics and DNA sequencing hardware can be improved by providing software that will take the huge amount of raw data—bytes, photon counts, hybridization signals—and turn them into information.

Figure 1. Links between data types.



LIMS will be able to "talk" intimately with the databases of biological information. Already all submissions to the Brookhaven Protein Database (PDB; Brookhaven, NY) of macromolecular three-dimensional structures—and nearly all submissions to the GenBank (Los Alamos, NM) and European Molecular Biology Laboratory (EMBL; Heidelberg, Germany) Data Library DNA sequence databases—are electronic. But they still require human intervention. Programmes such as GenBank's Authorin (which automatically takes a DNA sequence author through the steps necessary to produce a correct GenBank entry) are a small step towards integration of laboratory results with central databases. Ultimately, all the information, from the source literature through the experimental methods and results to the final write-up could be handled through a unified system.

Logic and links

The second leg of bioinformatics is the analysis and linkage of disparate data. Simple linkages, such as those between DNA and protein sequence or restriction maps, are easily understood and readily made by many software suites. More powerful packages, such as those developed by IntelliGenetics and GCG (Madison, WI), can take "your" sequence, interrogate the huge nucleotide and protein databases for similar ones, translate it into protein, find specific sequences within a sequence, design PCR primers that will recognise it, and so on. Other software packages, such as those of Oxford Molecular (Oxford, U.K.) and Biostructure (Strasbourg, France) explore more complex links, such as those between protein sequence and potential three-dimensional structure, but require interpretative skill from the user to know whether what the programme is suggesting is reasonable. And linking three-dimensional structure to potential enzymatic activity or to disease phenotypes, while possible, is still quite beyond the computer. It is the domain of the expert.

What is needed is a more extensive network of links between databases (Figure 1). The network is easy to depict, but can it be realised? As the ORF example illustrates, biologists are currently much better at making connections between sets of information than are computers. Biochemists understand the relationship between the representation of glycine used in molecular biology (a short string of letters—Gly), that used in chemistry ($HO_2CCH_2NH_2$), and that used in a structure optimization programme (which might treat the peptide chain as a ribbon with charged balls on it). But a computer will have to be taught how to translate between these very different views of the same thing.

Without these links, the central databases remain a resource for the specialist investigator. But it need not be that way. There may be over a million people in the developed world—life science researchers, product developers in healthcare and agriculture, university and high-school teachers—who would be interested in the genome sequence if it were not just information, but was linked up with other sequences to provide knowledge about people.

A new generation of database programmes is starting to address these links. The ICRF Prolog-based system aims to link the genetic map of an organism, the clones derived therefrom, and the DNA sequence data derived from those clones. The National Center for Biotechnology Information (NCBI) at the National Library of Medicine (Bethesda, MD) has created a database access programme called "Entrez," which can "jump" among the domains of DNA sequence, protein sequence, and literature reference: users can fol-

Data, Information, Knowledge.

Data is the raw output of experiments. Like the text of a bus timetable, it is a lot of letters and numbers, meaningless in themselves. Data can be readily transformed from one form to another—most obviously from "wet" biochemistry into computer bits and bytes. But that does not add anything to it. "Data processing" is the storage and handling of this material.

Information is data in the context of the real world. The context of the bus timetable is our experience of how clocks work and how buses do not. The mass of data now tells us something about the world that we did not know before—so it is information. "Information technology" is about handling data so that it retains its context.

Knowledge comes through linking information with other information. Thus we connect the bus timetable, the present time, and the sight of three buses vanishing down the hill: we are furnished with the knowledge that we are due a long wait or a long walk. Knowledge often needs far less data to represent it than does information: but knowledge is more powerful, forming the basis for decisions about what to do next. At the bus stop, a mass of information is refined into a single knowledge-based decision to wait, or not, for the next bus. "Knowledge engineering," the computerised handling of knowledge, is consequently an active artificial intelligence research topic. ///