Designing databases for molecular biology

SIR—I would like to endorse the approach recently proposed by Pongor[1] to the development of higher-order databases for molecular biology[2] and the view that artificial intelligence (AI) might be used in the study of macromolecular structure. This is, in fact, the motivation for some of our research in the ICRF Biomedical Computing Unit[3].

One of the contributions that AI techniques make to our work[4] is to combine data retrieval and logical deduction within a common framework. It is my contention that extensive deductive databases, as such systems are called, will be a minimum requirement for the development of large scale knowledge-based systems in molecular biology. Knowledge-based systems are programs that can reason using representations of knowledge about the world and the ways in which knowledge is used to solve problems. Furthermore, the higher-order databases discussed by Pongor[1] and Pabo[2] will need to be knowledge-based if they are to achieve the required flexibility and extensibility.

In my experience, however, the development of extensive AI applications in molecular biology stretches present AI programming tools beyond their limits, in part because they are ineffective at complex deductive reasoning with large bodies of data. Research into knowledge-based management systems[5,6] is now being undertaken to alleviate this problem by combining the reasoning capability of AI programs with the efficient retrieval and management of shared data provided by database management systems. Another problem is that the AI techniques for representing some of the knowledge required for applications in molecular biology are relatively immature and require further research.

In addition to the problems of representational adequacy in AI techniques, there are significant issues that must be resolved in biology before a new generation of knowledge-based information systems for molecular biology can be developed. In order to represent a set of concepts in a computer, there must be an agreed syntax and semantics that are both consistent and complete. Agreements on the syntax and semantics of biological and biochemical concepts have previously been reached by nomenclature committees and workshops of the scientific unions (IUPAC and IUB, for example). But agreeing a nomenclature can often require considerable time, and clearly one that encompasses a signficant part of modern biology is likely to require an enormous effort. Nevertheless, the development of a systematic ontology and epistemology of biology will be an important part of designing an integrated and comprehensive model for information systems in molecular biology. These concerns make me disagree with Pongor's assertion that the concepts used to describe molecular biology can now readily be defined in unequivocal terms.

A next-generation molecular biology information resource should describe in a conceptual model the relationships among all the entities with data in the molecular sequence and structure data libraries as well as the higher-order relationships among sequences that were suggested by Pabo[2]. The model should be based on a systematic model of modern biology and should accommodate the views of data held by the different biological research communities — such as protein structure prediction, gene expression and molecular evolution. It must also accommodate uncertain or partial information and be easy to modify in the light of changes to our knowledge. Consideration should also be given to incorporating or referencing data collections of other kinds such as those listed in refs 7 and 8. It might also be feasible to integrate knowledge bases of experimental methods and computer software[9], and it will be important that the scientific literature be available through abstract or bibliographic databases[10].

This integrated information or knowledge source would be a powerful resource for biomedical research and its interdisciplinary nature demands that it be a collaborative venture; not just among biological scientists but also between the computer sciences and biology.

C.J. RAWLINGS

Biomedical Computing Unit,
Imperial Cancer Research Fund,
PO Box 123, Lincoln's Inn Fields,
London WC2A 3PX, UK

1. Pongor, S. *Nature* **332**, 24 (1988).
2. Pabo, C.O. *Nature* **327**, 467 (1987).
3. Fox, J. & Rawlings, C.J. *Knowledge Based Systems in Biology and Medicine: A Progress Report* (Imperial Cancer Research Fund Biomedical Computing Unit, 1987).
4. Rawlings, C.J., Taylor, W.R. Nyakairu, J., Fox, J. & Sternberg, M.J.E. *J. molec. Graphics* **3**, 151–157 (1985).
5. Brodie, M.L. & Myulopoulos, J. (eds) *On Knowledge Base Management Systems. Integrating Artificial Intelligence and Database Technologies* (Springer, Heidelberg, 1986).
6. Rawlings, C.J. *et al. Tech. Rep.* no. 54 (ICRF Biomedical Computing Unit, London 1987).
7. Lawton, J. *Listing of Molecular Biology Databases* (Los Alamos National Laboratory, Los Alamos, 1988).
8. von Heijne, G. *Sequence Analysis in Molecular Biology*. 153–160 (Academic Press, New York, 1987).
9. Rawlings, C. *Software Directory for Molecular Biologists* (Macmillan, London and Stockton, New York, 1986).
10. Bicknell, E.J. Rada, R., Davidson, S. & Stander, R. *Nucleic Acids Res.* **16**, 1667–1680 (1988).

Lod score *Redivivus*

SIR—The proposal of Edwards[1] that 'log-likelihood' be used as a synonym for 'lod score' is analogous to suggesting that 'cat' be now known as 'monkey'. The advantages of such substitution are controversial, whereas the disadvantages are clear.

Wald[2] introduced the log-probability ratio without naming it, and Barnard[3] coined the term lod for this statistic. Neither specified the logarithmic base in the definition, although Barnard assumed natural lods in his mathematical development. Several other synonyms have been proposed including log-likelihood ratio (consistent with the definition of likelihood but originally with a different meaning[4]). Edwards himself has previously suggested 'support' which is seldom used, perhaps because he equates it to log-likelihood when considering the probability under the null hypothesis as an arbitrary constant, and otherwise to the log-likelihood ratio[5]. In this he is unique, because the two statistics have different properties: by itself a log-likelihood ratio gives a test significance and a measure of information, whereas a log-likelihood does not. Of all the uncommon synonyms for lod (credibility, plausibility, information, weight of evidence and support), log-likelihood is the most idiosyncratic and the only one that is inadmissible.

Logic apart, use of lods is well established. The term is recognized in statistical dictionaries[6] and *Index Medicus*. Thousands of publications report lods, invariably to the base 10, a convention that was introduced by a member of Barnard's audience[7]. The advantage of common lods is that significance levels and conservative confidence limits ('support intervals') are easy to remember. For example, a lod of 3 corresponds to a significance level of 0.001, and a 1-lod interval to at least 90% confidence. In natural lods this translates to a lod of 6.9 and a 2.3-lod interval, respectively. A philosophical preference for natural lods hardly justifies a change from common lods or, what is worse, an ambiguous mixture of the two bases.

Edwards also objects to the word 'score', which entered the literature as tables for mating and ascertainment types (z_1, z_2 and so forth)[8]. Computer programs now make lod and lod score equivalent. By the phrase "score is now reserved for the first derivative of the log-likelihood" Edwards implies a consensus. On the contrary, eminent statisticians have termed this use "regrettable"[6], and score continues to be applied in contexts as diverse as log-ranks and cricket.

Terminological quibbles never led to scientific advance. As Kendall remarked in discussion of the paper that sparked this controversy[3], "all these people were really saying the same thing in rather different

1. Edwards, A.W.F. *Nature* **333**, 308 (1988).
2. Wald, A. *Sequential Analysis* (Wiley, New York, 1947).
3. Barnard, G.A. *J.R. statist. Soc.* **B11**, 115–149 (1949).
4. Neyman, J. & Pearson, E.S. *Biometrika* **20A**, 175–240 (1928).
5. Edwards, A.W.F. *Likelihood* **12**, 31 (Cambridge University Press, 1972).
6. Kendall, M.G. & Buckland, W.R. 4th edn *Dictionary of Statistical Terms* (Longman, London, 1982).
7. Smith, C.A.B. *J.R. statist. Soc.* **B15**, 153–192 (1953).
8. Morton, N.E. *Am.J.hum. Genet.* **7**, 277–318 (1955).