New generation databases for molecular biology

SIR—Databases, particularly those of DNA and protein sequences, play an important part in modern molecular biology. When a new protein is sequenced, these databases allow a rapid search for homologous proteins. Molecular biology has advanced so rapidly that it may now be time to develop 'second generation' databases. Instead of storing linear sequence information, these would emphasize concepts and relationships, for example:

(1) A database of protein–DNA interactions might contain the amino-acid sequences of DNA-binding proteins and the DNA sequences of their affined binding sites. This information could be organized at different levels of detail, so that information about the precise size of the binding site, about mutations that affect binding, and even atomic details of the interaction could be added.

(2) A database that, organized with a system of 'pointers' for cross-referencing, might keep track of all known sequence homologies, along with information about the significance of each sequence match. It might also compare protein and DNA homologies.

(3) Databases of structural motifs (similar to the one started by Blundell *et al.*, *Nature* **326**, 347-352; 1987) might contain all the structural units that had been observed in more than one protein or all sequences that are homologous to proteins of known structure.

There is also a need to develop libraries of subroutines or programs that are written in different laboratories and yet can be used interchangeably. Programs using common protocols or data structures might provide the best way to handle information from a project to sequence the human genome. Agreeing on programming protocols would help to make software units compatible, just as standardization in hardware design may allow different computers to work together.

What makes a database useful? In a very general sense, the use of a database may be determined by the amount of information, the accuracy (how many mistakes were made in sequencing the protein and entering the data), and the relevance of the data to a particular problem. Sequence databases clearly satisfy these criteria — they contain a large amount of highly reliable information that is relevant to problems involving protein structure, function and evolution.

Versatility is another central issue in database design. Databases can be used in many different ways because data and analysis are clearly separated. The sequence data are permanently stored on disk or tape; many different programs can access the same database, using the sequence information in completely different ways. Relatively little interpretation and judgement are needed as the database is developed. The more critical problems of interpretation and judgement are deferred until the sequence data are analysed with a particular program — the appropriate strategy will depend on the goals of the project, and criteria for matches are a matter of scientific judgement.

How would higher order databases differ from existing databases? Sequence data are relatively easy to organize because the natural linear structure of the information simplifies information storage. New databases, however, may be highly branched. In a database of protein–DNA interactions there might be a protein that binds at many sites; in a database of protein homologies, a protein could be related to many other proteins. Such branched structures will be inherently more difficult to organize and search than linear sequence information. (Presumably this is one of the reasons that a database of carbohydrate structures — see *Nature* **324**, 208; 1986 — has only recently been organized.)

There are additional, more fundamental problems involved with second generation databases. Establishing them will require many difficult scientific judgements. One must choose the most important relationships and concepts to include in the database, and prescient decisions will be needed if one is to develop a useful tool for scientific research. Data entry will also be difficult. Data and analysis will not be clearly separated, and constant decisions will be required as information is entered. Data will have to be evaluated with the same care that is used when reviewing a manuscript. Finally, the versatility of these new databases must be a major concern. Will they be able to incorporate new types of information and to be used in unanticipated ways?

Is it feasible to establish higher order databases? Which will be most useful? Should they be started as collaborative ventures or should they be started in individual research laboratories and be used by other groups only after they have proven their utility? Can one really anticipate the needs for particular databases and foresee the best ways to organize them? Unlike physics, which moved to an era of 'big science' because of the costs of equipment for high energy research, it may be information that drives molecular biology into big science, and leads to a cooperative style of research. Organizing data at high levels of abstraction may be a step towards the more widespread use of artificial intelligence programming methods in molecular biology, and these higher order databases should be useful 'knowledge sources' for future programs.

CARL O. PABO
Howard Hughes Medical Institute,
Department of Molecular Biology and Genetics,
Johns Hopkins University School of Medicine,
Baltimore, Maryland 21205, USA

Our mistake

SIR—Bourne[1] is correct in suggesting that a problem exists with data presented in our paper[2]. The data as presented in Figs 1 and 2 were compromised by a serious error in calculating the specific activity of GTPγS. The correct values for GTPγS binding should be 1,000-fold lower than shown in our paper; that is, the *y* axis of the left-hand side of Fig. 1 should be fmol, not pmol, and the *x* axis of Fig. 2 should be pmol mg^{-1}, not nmol mg^{-1}.

We apologize if the data presented in our paper misled any investigators and accept full responsibility for this error. We are confident that the observations reported are real and that they demonstrate a role for a GTP binding protein in IL-2 mediated signal transduction.

STUART W. EVANS
SUZANNE K. BECKNER
WILLIAM L. FARRAR
National Cancer Institute,
Division of Cancer Treatment,
Frederick Cancer Research Facility,
Frederick, Maryland 21701, USA

1. Bourne, H.R. *Nature* **326**, 833-834 (1987).
2. Evans, S.W., Beckner, S.K. & Farrar, W.L. *Nature* **325**, 166-168 (1987).

Chromaffin cell synapsin?

SIR—In two recent issues of *Nature* there were reports with accompanying News and Views items on the role non-erythroid brain spectrin (fodrin) and associated cytoskeletal proteins play in the mechanisms of chromaffin cell exocytosis in the adrenal medulla[1,2] and neurotransmitter release in the brain[3,4]. Data discussed in these articles suggest the possibility that the two systems, both of which are derived from the neural crest, may share a common mechanism.

In the case of synaptic transmission, it was suggested[4] that neurotransmitter-containing synaptic vesicles are restrained in a fodrin/actin network in the presynaptic terminal until nerve depolarization elevates free cytosolic calcium. This activates a calcium/calmodulin-dependent protein kinase that phosphorylates synapsin I, a synaptic vesicle-associated protein that shares a number of properties with erythrocyte protein 4.1 (refs 5,6) and protein 4.9 (refs 3,7). On phosphorylation, synapsin I, which can enhance spectrin/actin interactions only in its dephosphorylated form[8], would lose its ability to mediate that interaction thus leading to