of air travel, and will become involved in studies such as those that have been mentioned.

D.S. MINORS
J.M. WATERHOUSE
Chronobiology Group,
Department of Physiological Sciences,
University of Manchester,
Manchester M13 9PT, UK

1. Winfree, A.T. *Nature* **330**, 311–312 (1987).
2. Woodruff, M. *Nature* **331**, 217 (1988).
3. Ford, B.J. *Nature* **331**, 309 (1988).
4. Holley, D.C. *et al. Effects of Circadian Rhythms Phase Alteration on Physiological and Psychological Variables: Implications to Pilot Performance* NASA Tech. Mem. No. 82177 (1981).
5. Aschoff, J. *et al. Chronobiologia* **2**, 23–78 (1975).
6. Klein, K.E. & Wegmann, H.M. *Significance of Circadian Rhythms in Aerospace Operations*. AGARDograph No.247 (NATO, 1980).
7. Minors, D.S. & Waterhouse, J.M. *Aviat. Med. Quart.* 1, 9–26 (1987).
8. Mrosovsky, N. & Salmon, P.A. *Nature* **330**, 372–373 (1987).

Roaring and oestrus

SIR—Karen McComb's account of her interesting experiment on the effect of roaring on oestrus in hinds (*Nature* **330**, 648; 1987) is marred by the analysis of her results. To remove 'skewness' in the distribution of calving dates by log transformation (a liberty, at best); to 'adjust' them to take account of two other significant variance components and then plot as a cumulative percentage is over-egging the statistical pudding. In reports such as this where the amount of data is quite small, could it not be given raw, so those with a particular interest can transform it as they wish? After deducing from the plot some idea of what the actual calving dates might have been, I could suggest that the case McComb makes out is not as convincing as is implied by the statistical contortions she has used.

J. C. BIGNALL
The Health Centre,
Newport, Dyfed SA42 0TJ, UK

MCCOMB REPLIES—Bignall's criticisms of my recent paper are mistaken. In that paper I plot cumulative percentage calving in three treatment groups to show that red-deer hinds exposed either to playback of recorded roaring, or to a vasectomized stag, calve earlier than control hinds (see my Fig. 1). I found these effects to be statistically significant in an analysis of variance. Bignall seems to believe that the plot of cumulative percentage calving (my Fig. 1) shows log-transformed values. This is clearly not the case, for a plot of the log-transformed (normalized) data would produce a symmetrical logistic (sigmoidal) curve. If Bignall had, as he suggests, "deduced from the plot some idea of what the actual calving dates might have been" by anti-logging the values, he would have obtained some very odd results indeed. The last calves would then have been born $> 2.7 \times 10^{67}$ years after the first.

Bignall implies that log transformation is a misleading contortion of the data. But real data do commonly exhibit some skewedness, and under such circumstances it is usual to normalize data by log transformation before analysis of variance is performed to ensure that the assumptions of the analysis of variance model are not violated (Snedecor, G. W. & Cochran, W. G. *Statistical Methods*; Iowa State Univ. Press, Annes, 1967). Failure to use normalized data is, under such circumstances, far more likely to give misleading results.

Bignall also complains that I have statistically controlled for sire and sex effects, implying that the treatment effects shown in Fig. 1 might disappear if I did not do so. Analysis of variance reveals that both sire stag and a sire stag × sex-of-calf interaction have significant effects on the variance in calving date. As these confounding effects fell disproportionately across treatment groups (some hinds not completing the experiment) it was appropriate to control statistically for them when examining the effect of treatment. However, as I describe in my paper, the effect of treatment was significant before the confounding effects were removed.

KAREN MCCOMB
Large Animal Research Group,
Department of Zoology,
Cambridge CB3 0DT, UK

Novel databases for molecular biology

SIR—Pabo[1] has highlighted the need for second-generation databases for molecular biology. One important task would be to organize non-structural data into accessible databases, which in itself would justify a major collaborative effort[2]. Furthermore, structural information should be accessible in the same conceptual framework as non-structural data. Development of a uniform and coherent conceptual framework could be based on a systematic study of terms, concepts and cognitive structures we use to perceive, to interpret and to memorize macromolecular data.

The approach suggested here is based on the analysis of scientific communications and statements pertinent to structural data. A preliminary — and by no means impartial — survey of current papers on protein and DNA structure suggests that the conceptual machinery we use to describe macromolecular data is in fact simple and uniform.

For example, macromolecular structures, which are too complex to be memorized, are recoded as a simplified set of substructures related to each other by a few relationships. The constituent elements are in turn characterized in terms of size, composition and similarity to other known substructures. The relationships of elements are usually described as symmetries and vectorial distances. Primary protein and DNA structures, for example, seem to be translated into higher-order sequences of complex elements (such as cleavage sites or ligand-binding sites). These segments can be described in different ways, as there are several symbolic and parametric methods to represent sequences.

Structures of DNA and protein can best be visualized as hierarchical and colinear data-structures (represented by sequences of symbols, numbers or sub-files), on which the same types of operations (such as insertion, deletion, exchange, symmetry operations) can be defined. Primary sequence similarity is by far the most frequent tool for detecting and characterizing structural similarities, even though it is becoming apparent that distant homologies are more easily understood in terms of parametric representations[3].

Even these simple concepts can be highly efficient if used in a generalized sense. For example, notions like a repetitive pattern of beta strands or amphiphilic helices can both be described as translational symmetries (in secondary structure and in hydrophobicity, respectively). Similarly, the statement 'a hydrophobic helix flanked by ionic residues' can be considered as a specific sequence of three elements which are defined in terms of composition.

A generalized framework of concepts and relationships abstracted from human thinking could enormously increase the performance of molecular-biology software. New categories identified by computer-based methods could be introduced to complement the known substructures. For example, linguistic methods have been used for the automatic definition of characteristic subwords and syntactic rules in DNA[4]. Moreover, artificial intelligence methods could be built up using these extended concepts, for automatic analysis of the available structural data and for building up new structural databases. Novel database structures could thus be the final result, rather than the beginning of this process.

Interpretation of macromolecular data seems to be ripe for artificial intelligence methods. Highly accurate data are available in abundance; concepts and terms used to describe these data are efficient and can be unequivocally defined in scientific terms; and our obvious inability to obtain an adequate overview of the ever growing body of information leaves us no other alternative.

SÁNDOR PONGOR
Agricultural Biotechnology Centre,
2101 Gödöllö, PO Box 170, Hungary

1. Pabo, C.O. *Nature* **327**, 467 (1987).
2. Davison, D. *Nature* **329**, 194 (1987).
3. Pongor, S. *Meth. Enzymol.* **154**, 450–473 (1987).
4. Brendall, V. *J. biomol. Struct* **4**, 11–22 (1985).
5. Bocksteele, F. *Biochemie* **67**, 509–516 (1985).