Database contamination

SIR — We have examined sequence data from vertebrate and bacterial species available in update 27 of the EMBL sequence database, using the program FASTA with the complete M13mp18 sequence as well as its polylinker only as search strings, and found extensive homologies to these to be present in several submitted sequences. We found 78 sequences with a total of 81 occurrences of vector sequences not documented in the feature table. Of these 81 overlaps, 55 were found in the 5′ or 3′ end of the sequence, while 23 were found internally (and presumably used in the assembly of the sequence) and 3 covered essentially the whole submitted sequence. In 19 of the 35 instances where the vector-containing parts were included in the feature table they were described as part of introns, whereas 7 listed them as part of the coding sequence. The largest overlap reported by FASTA contained 599 nucleotides.

Because FASTA will report only the one area in a sequence with the highest similarity to the search sequence, some of the vector-containing sequences reported were investigated further using WORDSEARCH, which will unravel occurrences of several patches of vector sequence. In one case, a sequence of 2,537 nucleotides where FASTA reported an overlap of 562 nucleotides with M13mp18, WORDSEARCH revealed the presence of four regions (125–412, 691–1,183, 1,538–1,700 and 2,155–2,533) with a total of 1,319 bases that clearly (assuming a rather high frequency of reading mistakes) are M13-derived.

Although M13-based vectors have been the most widely used vectors for DNA sequencing, we assume that searches with different polylinkers, lambda and other vector-based sequences as search strings would give several new overlaps. Also, there are probably entries in the databases with vector components too small to be assigned as such by mere homology considerations. The real number of vector-derived sequences in the databases thus is considerably larger than the number reported here.

The abundance of sequencing errors in these vector sequences is also of interest. In some instances, of course, this may be due to a heavily mutated vector being used, but in most cases we expect these errors to be a result of sloppy sequencing. If this reflects the quality of the total sequence information available in the databases, it must influence the reliability of these data as basis for biologically significant conclusions.

Finally, looking at the date of submission of the sequences, there are no indications that vector sequence contamination is a thing of the past, as 30 of the sequences we found were submitted in 1990 and 11 in 1991.

How can we avoid errors like these finding their way into the databases? Most sequencing software packages contain some sort of vector screening. Thus a simple screening of raw sequence data before assembly, as well as screening of assembled sequences against vector databases before submission, should be a simple task. Another possibility is that database administrators should screen all submitted sequences against the vector database. But is it part of their job to function in this way? One thing they could do is to add a question to the sequence submission forms: "Has this sequence been checked for the presence of vector sequences ?". Then the submitters would at least have been made aware of the possibility of contamination of the data. Also, it probably would be useful if submitters were encouraged to include the vector used in the feature table.

What should be done to contaminated sequences? First, such sequences constitute a very small percentage of the total database. In many cases, it would be enough to include the presence of vector data in the feature table of the sequence. In several instances, we presume that complete removal of the sequences from the database would be the best line of action. Again, it is understandable if the database administrators are reluctant to perform this cleaning-up: preferably, the submitters themselves should be made aware of the doubtful nature of their submissions and given the opportunity to rectify the data or withdraw their submission. (A list of the contaminated sequences we found is available on request from us, and has been submitted to the EMBL Data Library.)

RODRIGO LOPEZ*
TOM KRISTENSEN
HANS PRYDZ

*The Biotechnology Centre of Oslo,
University of Oslo,
P.O. Box 1125,
Blindern N-0316 Oslo, Norway
The Norwegian EMBnet node

Death by superantigen

SIR — Cohen *et al.*[1], commenting on our report[2] on superantigen-induced peripheral T-cell tolerance, contended that the T-cell death observed in SEB-primed mice may not reflect *in vivo* activation-triggered cell death but rather the *in vitro* death of SEB-activated T cells consequent to lymphokine "starvation". To support this contention, they cite their detection of DNA fragmentation and subsequent cell lysis in interleukin-2 dependent T-cell clones 12 hours after interleukin-2 withdrawal *in vitro*[3].

Cohen *et al.* suggested that in our system, SEB remaining in the circulation continues to stimulate Vβ^+ T cells, such that at day 2 post-priming T-cell activation is still in progress and accordingly lymphokines are sufficiently abundant that the cells do not undergo DNA fragmentation, even after 3 hours of culture *in vitro*. They postulate that instead, by day 4 post-priming, SEB is cleared from the system and the lack of ongoing SEB stimulation leaves Vβ8$^+$ T cells relatively lymphokine-"starved", a situation that is exacerbated in culture resulting in programmed cell death. Therefore, they ascribe the DNA fragmentation we observed in Vβ8$^+$ T cells at day 4 post-priming to a lack of growth factors necessary for cell survival rather than to activation-induced programmed cell death.

Scientific Correspondence

SCIENTIFIC Correspondence is a relatively informal section of *Nature* in which matters of general scientific interest, not necessarily those arising from papers appearing in *Nature*, are published. Due to space limitations priority is usually given according to general interest and topicality, to contributions of fewer than 500 words, and to contributions using simple language. Contributions may be sent to referees and, in the case of matters arising from material published in *Nature*, are often sent to the author of that article for comment.

Scientific Correspondence submitted for publication should be typed double-spaced and three copies sent. □

The explanation proposed by Cohen *et al.* is interesting, but we believe that their interpretation of our results is not supported in the context of the following observations: (1) We found[2] that at 2 days after SEB priming, the percentage of Vβ8$^+$ T cells in the periphery was elevated from 20 to 40% both before and after 20 hours' incubation at 37 °C. If as suggested by Cohen *et al.*, *in vitro* cell "starvation" were the primary cause of cell death, one would have expected to see a reduction in the Vβ8$^+$ T-cell number after 20 hours of culture. (2) Anti-interleukin-2 antibodies added to the 37 °C 20-hour culture of spleen cells obtained at 2 days post SEB-priming to block the possible effect of autocrine interleukin-2 do not induce programmed cell death of Vβ8$^+$ T cells. (3) Splenic cells and serum obtained 36 hours after SEB priming did not induce proliferation of fresh spleen cells. Thus, the absence