Cleaning up gene databases

SIR—We have discovered errors in the EMBL nucleotide sequence databank in the course of training of artificial neural networks to recognize pre-messenger RNA splicing signals¹ in human genes. In training on thirty-three human genes, the seven genes listed in the table appeared to disturb the learning process. Further

randomly lower the degree of regularity, which is also a prerequisite for the function of a specific biological recognition mechanism.

The problem of errors in the sequences incorporated in databanks could be amplified in the future, as a consequence of the increasing rate at which sequences are

Errors in the EMBL nucleotide sequence database

Databank sequence (Accession no.)	Published sequence	Type of error
HSGLIA (X00492)	ref. 5	Splice donor cited at base pair (bp) 328 in feature table (FT). Correct position: 329 (there is a typographical ambiguity in the article)
HSBGL3(V00499)	ref. 6	Splice donor cited at bp 2971 in FT. Correct position 2981
HSERPG (X02158)	ref. 7	Presumed -1 bp error in splicing frame of intron I in article* Nucleotides CG missing after bp 578 All following FT entries -2 bp wrong until last acceptor site which is +3 bp wrong
HSDGL1(V00505)	ref. 8	Presumed +1 bp error in splicing frame of intron I in article*
HSEGL1(V00510)	ref. 9	Presumed +1 bp error in splicing frame of intron I in article*
HSHSP27(X03900)	ref. 10	Presumed -1 bp error in splicing frame of intron I in article*
HSGROW2(V00520)	ref. 11	Splice donor cited at position 1081 in FT. Correct position: 1091

Discrepancies between published and databank entered sequences, or presumed errors in the published sequences.

* The occurrence of repeated bases around splice donors and splice acceptors means that the splicing frame of an intron cannot always be unambiguously assigned on the basis of the messenger RNA sequence and amino-acid sequence. A neural network can, however, point out the frame by recognition of the splice signals present in the sequence. Splicing always occurs at a position that is fixed with respect to the sequence signal¹².

investigation revealed discrepancies between the databank sequences and the original published papers for three of the genes. In the other four examples, we found wrongly assigned splicing frames of introns.

The subset of thirty-three human genes from the EMBL databank carried information about the location of splicing signals. On the basis of the feature table information, we assigned each nucleotide to one of two categories, splicing donor sites or otherwise. We used these data to train a neural network of the perceptron type² to classify nucleotides, on the basis of their context, as belonging to one of the two categories. The network was similar to that used for predicting secondary structures of proteins^{3,4}.

Neural networks are now being applied to many classification tasks; they have in common the ability, when trained, of dealing with non-linearities in the association between objects and categories. But in the presence of strong non-linearities, as when many similar objects must be put into widely different categories while many dissimilar objects are destined for the same category, the classification can usually be learned by the network only with difficulty. Errors introduced

determined and by the initiative to sequence the human genome. We therefore propose that computerized proof reading should be incorporated in the databanks, so as to take advantage of the ability of neural networks to reveal non-linearities in a dataset, as we have demonstrated.

SØREN BRUNAK

Department of Structural Properties of Materials,
The Technical University of Denmark,
DK-2800 Lyngby,
Denmark

JACOB ENGELBRECHT

Department of Dairy Science,
Royal Veterinary and Agricultural University,
Bülowsvej 13,
DK-1870, Frederiksberg C,
Denmark

STEEN KNUDSEN

University Institute of Microbiology,
Øster Farimagsgade 2A,
DK-1353 Copenhagen C,
Denmark

1. Padgett, R.A., Grabowski, P.J., Konarska, M.M., Sheiler, S. & Sharp, P.A. *A. Rev. Biochem.* **55**, 1119–1150 (1986).

2. Rumelhart, D.E., Hinton, G.E. & Williams, R.J. in *Parallel Distributed Processing* (eds McClelland, J.L., Rumelhart, D.E. and the PDP Research Group) 318–362 (MIT, Cambridge, Massachusetts, 1987).

3. Quian, N. & Sejnowski, T.J. *J. molec. Biol.* **202**, 865–884 (1988).
4. Bohr, H. *et al. FEBS Lett.* **241**, 223–228 (1988).
5. Strachan, T., Sodyer, R., Damotte, M. & Jordan, B.R. *EMBO J.* **3**, 887–894 (1984).
6. Lawn, R.M., Efstratiadis, A., O'Connell, C. & Maniatis, T. *Cell* **21**, 647–651 (1980).
7. Jacobs, K. *et al. Nature* **313**, 806–810 (1985).
8. Spritz, R.A., DeRiel, J.K., Forget, B.G. & Weissman, S.M. *Cell* **21**, 639–646 (1980).
9. Baralle, F.E., Shoulders, C.C. & Proudfoot, N.J. *Cell* **21**, 621–626 (1980).
10. Hickey, E. *et al. Nucleic Acids Res.* **14**, 4127–4145 (1986).
11. DeNoto, F.M., Moore, D.D. & Goodman, H.W. *Nucleic Acids Res.* **9**, 3719–3730 (1981).
12. Jacob, M. & Gallinero, H. *Nucleic Acids Res.* **17**, 2159–2175 (1989).

Neutron irradiation of superconductors

SIR—We were interested to see additional technical work¹ concerning the effect of neutron irradiation on critical currents in high-temperature superconductors. Our first paper on single-crystal neutron irradiation in $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$ demonstrated the improvements that were possible², and suggested additional work needed for the technique to reach its full potential³⁻⁵. Our most recent work⁶ showed an enhancement of the critical current due to neutron irradiation which increases strongly with temperature and is still effective in magnetic fields up to 8 T.

The enhancement factors are comparable to those reported in ref. 1. However, these factors are not in themselves very meaningful: if the critical current density of a very high-quality ($J_c=0$) single crystal is compared to the irradiated state, the enhancement factor will be infinity!

If comparisons of this kind are to be made, it is necessary to measure the critical currents before and after neutron irradiation on the same single crystal, as we have been careful to do in our work. In addition, significant flux motion effects must be included in the discussion, especially when results obtained by different measuring techniques are compared. These effects were ignored in ref. 1.

GEORGE W. CRABTREE

Argonne National Laboratory,
Argonne,
Illinois 60439, USA

HAROLD W. WEBER

Atomic Institute of the Austrian Universities,
Schüttelstrasse 115,
A-1020 Vienna, Austria

1. van Dover, R.B. *et al. Nature* **342**, 55–57 (1989).
2. Umezawa, A. *et al. Phys. Rev. B* **36**, 7151–7154 (1987).
3. Weber, H.W., Crabtree, G.W., Umezawa, A., Liu, J.Z. & Nunez, L.H. in *High-T_c Superconductors* (ed. Weber, H.W.) 273–278 (Plenum, New York, 1988).
4. Weber, H.W. *et al. in MRS International Meeting on Advanced Materials* Vol. 6 (eds Tachikawa, K. & Kitazawa, K.) 365–369 (Materials Research Society, 1989).
5. Weber, H.W. in *Studies of High Temperature Superconductors* Vol. 3 (ed. Narlikar, A.V.) 197–227 (Nova, New York, 1989).
6. Sauerzopf, F.M., Wiesinger, H., Weber, H.W., Crabtree, G.W. & Liu, J.Z. *Poster presented at Conf. on Materials and Mechanisms of Superconductivity/High-Temperature Superconductors*, Stanford, 23–28 July (1989); *Physica C* (in the press).