Neural Nets Catch the ABCs of DNA

By STEFI WEISBURD

Today's computers are exemplary number crunchers. But when it comes to performing some of the brain's repertoire of feats — such as recognizing patterns or extracting a general rule from a set of examples — computers are sadly lacking.

To make computer thinking more like human thinking, scientists are developing "neural networks," or highly interconnected webs of processing units (SN: 6/6/87, p.362; 7/4/87, p.14) — a design loosely based on how scientists suspect arrays of nerve cells interact in the brain. In anticipation of the day when electronic versions of such arrays are built as hardware, most researchers are presently running neural-net simulations on conventional computers in order to sketch the range of problems that neural nets might be able to solve in the future.

But researchers at Los Alamos (N.M.) National Laboratory are not content to merely flex the neural-net muscle. They've put it to work. Physicist Alan Lapedes says he and computer scientist Robert Farber recently used a supercomputer to demonstrate that "there are problems relevant to the real world that neural nets can attack *now* without waiting for chips [to be made]." In particular, the researchers applied one kind of neural-net simulation, a learning algorithm called back propagation, to problems in genetics and signal processing. In each case, says Lapedes, "the neural net beat the conventional approach."

The genetic problem addressed by Lapedes, Farber and their co-workers deals with a central question in biology: how to determine whether a particular sequence of bases in a fragment of DNA is coding for the production of a protein. "DNA is basically a bunch of symbols, composed of A, C, T and G [for the bases adenine, cytosine, guanine and thymine]," says Lapedes, "and there's something about those symbols that tells the real biological mechanism whether to make a protein or not." Using a statistical approach, scientists can predict the biological activity of a DNA strand quite well, provided the strand contains at least 200 base pairs.

But for smaller segments with one-tenth the number of bases, says Lapedes, the conventional method is correct only half the time — only as good as flipping a coin. The ability to evaluate fragments of 30 or so base pairs is important, he adds, because ribosomes, the "factories" where proteins are constructed in accordance with the DNA blueprint, are thought to work on only about 30 base pairs' worth of information at a time.

On 30 base pair segments, the neural net did much better than the conventional approach. After the scientists "trained" the net by showing it 900 examples of active DNA fragments and 900 examples of inert stands, the net was able to predict, with an accuracy of 80 percent, whether or not a newly presented strand was active. Moreover, Lapedes notes that in the process, the neural net appears to have learned some fundamental rules about genetics that may possibly have eluded biologists. "At the moment, it is just a black box giving us a yes or no answer," he says. "But the next step is to try to understand how the net did what it did" and to extract these biological rules.

Lapedes also says his group has been asked to use the neural net to check some of the experimental data being collected for GenBank, a DNA library administered by Los Alamos. This is because in the course of their studies, the neural net uncovered an error in the GenBank data; it found that a particular base pair sequence had been listed as not coding for a protein when it in fact did.

While the neural net did well at manipulating symbols in the genetic problem, the researchers found that it did even better when handling numbers in a signal processing task. They fed the computer a series of numbers, which Lapedes says was so complicated it appeared almost random, and then asked the net to predict the value of the next number. "The neural net was able to find the distinguishing simplicity under those [seemingly] random data" and make predictions with an accuracy orders of magnitude improved over conventional techniques, he says.

The only other method that has achieved comparable accuracy was also developed at Los Alamos in the last few months. Although it's not a neural-net method, Lapedes and Doyne J. Farmer, who developed it, are combining forces in the hope that the two techniques combined will be more powerful than either one alone.

The researchers think they understand how the neural net solved the processing problem. In accordance with a mathematical theorem, the net seems to have found a continuous function that describes a given set of points and then used this function to predict the next point to fit the pattern. The net constructed this generating function in a method somewhat analogous to the mathematical technique called Fourier analysis, which enables scientists to approximate any curve by adding together sine waves of different frequency, phase and amplitude. Instead of sine waves, however, the net used another trigonometric function, hyperbolic tangents (tanh).

The net's choice of the tanh function is more serendipitous than surprising, because the researchers in essence "loaded the dice" by using tanh functions throughout their program. "The surprising thing is that the net is able to do so much with it," says Lapedes. "It turns out that tanh was just what you need to do a very accurate job for this problem." The researchers suggest that this may be a useful technique for other nonlinear applications as well.

Lapedes says he and Farber were inspired to go "casting about for real-world problems" by the work of biophysicist Terrence Sejnowski at Johns Hopkins University in Baltimore. Sejnowski and a colleague developed NETtalk, an algorithm similar to that used by Lapedes and Farber, which, when run on a conventional computer, teaches itself to read English text aloud (SN: 1/24/87, p.60).

Sejnowski says he and a student have now applied NETtalk to the biophysical problem (also being addressed by Lapedes and Farber) of predicting the two-dimensional structure of proteins from their amino acid sequences. "The results are better than any existing techniques, so we know that the technique we're using is a very powerful one," says Sejnowski, "but unfortunately, it still isn't good enough to be able to predict" how the protein folds in three-dimensional space, which is key to understanding the function of a protein. "Our guess is that there is a limit to any technique based on pattern recognition," he adds. "And in some cases we can show that more power techniques are needed."

Sejnowski and R. Paul Gorman of Bendix Aerospace Technology Center in Columbia, Md., have also applied NETtalk to a signal processing problem of clear interest to the Navy: the recognition of underwater targets from the sonar signals that bounce off them. Gorman says that "the performance was better than many existing techniques that have been used and is even better than trained human beings."

Sejnowski says it's becoming evident that neural nets — with their ability to recognize patterns that can be too subtle or complex for the human brain — are well suited to many signal processing and biotechnology problems. Moreover, he is impressed at how smoothly the neural-net algorithms are making the transition from academic laboratories to real-world problems. Only time will tell how well they will eventually do, he says, but "the early returns from the networks are that they seem to be doing quite well." □