WHEN GOING BACKWARDS MEANS PROGRESS: ON THE SOLUTION OF BIOCHEMICAL INVERSE PROBLEMS USING ARTIFICIAL NEURAL NETWORKS

Douglas B. Kell, Chris L. Davey,
Royston Goodacre, and Herbert M. Sauro

Dept. of Biological Sciences
University of Wales
Aberystwyth, Dyfed SY23 3DA, U.K.

1. Introduction

Many (perhaps most) biochemical problems are actually "inverse" problems, or problems of system identification. In general, if we know the parameters of a system of interest, we can set up a model (or actual) experiment, run the model (or experiment), and observe the behavior or time evolution of the system. However, especially in a complex biological system, the things which are normally easiest to measure are the variables, not the parameters, and it is the variables which depend on the parameters, not *vice versa*.

In the case of metabolism, the usual parameters of interest are the enzymatic rate constants, which are difficult to measure accurately *in vitro*[1] and virtually impossible so to do *in vivo*. Yet to describe, understand, and simulate the system of interest we need knowledge of the parameters. In other words, we need somehow to go backwards from variables such as the steady-state fluxes and metabolite concentrations, which are relatively easy to measure, to the parameters, which are not.

A similar situation exists in biological spectroscopy. If we have chemical standards, whether pure or mixed, it is easy to obtain their spectra. The inverse problem then involves obtaining the concentrations of metabolites, or the overall spectral parameters, from the spectra observed in a complex system.

The purpose of this article is to describe our successful exploitation of artificial neural networks (ANNs) in the solution of such biochemical (and other) inverse problems. We apply the approach specifically to:

Modern Trends in Biothermokinetics, Edited by
S. Schuster *et al.*, Plenum Press, New York, 1993

(i) model metabolic pathways studied by computer simulations,
(ii) parameter identification in biological dielectric spectroscopy, and
(iii) the extraction of chemical information from pyrolysis mass spectra of intact microbial cells.

2. Artificial Neural Networks

ANNs are collections of very simple "computational units" which can take a numerical input and transform it (usually via summation) into an output (see e.g. Refs 2-12). The inputs and outputs may be to and from the "external world" or to other units within the network. The way in which each unit transforms its input depends on the so-called "connection weight" (or "connection strength") and "bias" of the unit, which are modifiable. The output of each unit to another unit or to the external world then depends on both its strength and bias and on the weighted sum of all its inputs, which are transformed by a (normally) nonlinear weighting function refered to as its activation function.

The great power of neural networks stems from the fact that it is possible to present ("train") them with known inputs (and outputs) and provide some form of learning rule which may be used, iteratively, to modify the strengths and biases until the outputs of the network as a function of the inputs correspond to the desired ("true") outputs. The trained network may then be exposed to "unknown" inputs and will then provide its view of the "true" output(s).

A neural network therefore consists of at least three layers, representing the inputs and outputs and one or more so-called "hidden" layers. It is, in particular, the weights and biases of the interactions between inputs and outputs and the hidden layer(s) which reflect the underlying dynamics of the system of interest, even if its actual (physical) structure is not known. By training up a neural network with known data, then, it is possible to obtain outputs that can accurately predict things such as the (continuing) evolution of a time series, even if it is (deterministically) chaotic[13].

Other successful uses of neural networks include speech recognition, DNA sequence analysis, the correction of errors in optical astronomy, and the analysis of vapors by arrays of artificial sensors. One may also perhaps mention the successful use of simple neural nets in the analysis of chemical engineering systems[14].

3. Analysis of Metabolic Systems Using Neural Networks

As described in several other contributions in this volume, it is possible by computer simulation to determine steady-state variables such as fluxes and metabolite concentrations as a function of parameters such as the enzymatic rate constants and external metabolite concentrations. It is obviously then possible to change one or more of the parameters and to determine another set of associated variables, and so on. The idea is that having acquired related sets of parameters and variables, we would then be in a position to train neural networks *in which the (known) variables were the inputs and the parameters were the outputs*. When the nets had successfully learned to reflect the correct parameters when presented with the variables, we would have solved our problem. We could then present the net with "random" (experimental) variables and ask it for the parameters.

The correctness of the network's predictions could obviously be checked by running a simulation with the parameters provided by the network and seeing if they generated the variables used as the input to the net. The result would be that we *could* in fact obtain the (enzymatic) parameters of a metabolic network (*and hence the control coefficients and*

elasticities) by measuring the variables alone. We have now implemented this strategy, as outlined in what follows.

For computational simplicity we concentrate here on a simple, three-step linear pathway, as shown in Scheme 1, with each enzyme (one substrate / one product) possessing reversible Michaelis-Menten kinetics, and with the steps having equilibrium constants of 567, 13.4 and 2.3 respectively.

$$X_0 \longleftrightarrow S_1 \longleftrightarrow S_2 \longleftrightarrow X_1 \qquad \text{Scheme 1}$$

The dataset was obtained by varying the V_{max} values of each enzyme and the first forward K_m value. All other K_ms were held constant. The concentration of S_1, S_2 and the steady-state flux were recorded for each parameter set. In addition, each parameter set was applied using three different concentrations (1, 10, 100) of the starting metabolite X_0 so that we could obtain three sets of variables for the same set of parameters (in an experimentally realizable fashion), and so aid the training process.

The parameters of the system were varied as follows. Each parameter was first assigned a uniform random number u between 0 and 2.0, which was then used to generate a non-skewed distribution spanning two decades using the formula, 10^u. The reason why the parameters were generated in this manner was so that the distribution of values would not be confined to the upper decades of the range and thus there would be a roughly equal number of random values between 1 and 10 as there would be between 10 and 100. In this way 500 random sets of K_m and V_{max} values were generated, from which the steady-state concentrations of S_1, S_2 and the fluxes were obtained. A 12-18-4 net was then trained (with the stochastic backpropagation algorithm) using the 12 values of X_0, S_1, S_2 and J as the input and the four varied parameters as the output.

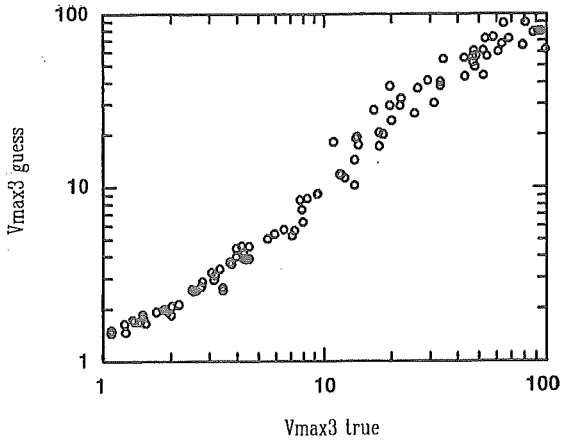

Figure 1. An Artificial Neural Network's estimate of unseen $V_{max,3}$ values.

After training to an RMS error of ca 0.08 (just under 4000 epochs), the net had successfully generalized, as illustrated in Fig. 1, which shows the network's estimate of the unknown values for $V_{max,3}$ against the "true" values. Other parameters had been learned to a more or less similar degree.

It is worth pointing out that it is only possible for ANNs to learn something that is actually learnable, so that if a variable is not significantly related to a particular parameter then the network will not learn it; this of course may be used to our advantage, since if the

parameter of interest does not significantly control the variable it may be assumed that one is not interested in studying it anyway. Because updating of the network is based on the *overall* RMS error, "bad" relationships interfere with the learning of "good" ones; our experience to date suggests that any linear correlation coefficient (of logarithmic parameters) below some 0.4 will inhibit the net from converging.

4. Solution of the Inverse Problem in Biological Dielectric Spectroscopy

In biological dielectric spectroscopy, where dispersions are substantially broader than that expected from a purely Debye-like process, it is not always possible, due to technical limitations, to obtain data over a wide enough range of frequencies to encompass the entire dispersion(s) of interest. Similarly, because of the breadth of the dispersions, it is common to seek to characterize the dielectric behavior of interest by means of the Cole-Cole function[15]. Whilst it is possible to fit dielectric data to this equation using appropriate nonlinear least-squares methods[16], these methods are computationally rather demanding, and must be undergone, iteratively, for each set of data.

We have found[17] that it is possible to train an artificial neural network with small sets of dielectric data (permittivities measured at various frequencies) as the inputs and the attendant parameters of the Cole-Cole equation as the outputs. The trained net can then give an essentially instantaneous output of the limiting permittivities at frequencies that are both high and low with respect to the characteristic frequency, and thus of their difference, a parameter which, for the so-called β-dispersion, scales with the biomass content of cell suspensions[16].

5. Solution of the Inverse Problem in Pyrolysis Mass Spectrometry of Microbial Cells

Pyrolysis is the thermal degradation of a material in an inert atmosphere, and leads to the production of volatile fragments (pyrolysate) from non-volatile material such as microorganisms. Curie-point pyrolysis is a particularly reproducible and straightforward version of this technique, in which the sample, dried onto an appropriate metal is rapidly heated (0.6s is typical) to the Curie point of the metal, which may itself be chosen and is commonly 530°C. The pyrolysate may then be separated and analyzed in a mass spectrometer[18], and the combined technique is then known as Pyrolysis Mass Spectrometry or PyMS.

Conventionally (within microbiology and biotechnology), PyMS has been used as a taxonomic aid in the *identification* and *discrimination* of different microorganisms[19]. To this end, the reduction of the multivariate data (150 normalized values in the range m/z 51-200) generated by the PyMS system is normally carried out using Principal Components Analysis (PCA), a well-known technique for reducing the dimensionality of multivariate data whilst preserving most of the variance. Whilst PCA does not take account of any groupings in the data, neither does it require that the populations be normally distributed, *i.e.* it is a non-parametric method.

The closely related Canonical Variates Analysis technique then separates the samples into groups on the basis of the principal components and some *a priori* knowledge of the appropriate number of groupings. Provided that the data set contains "standards" (*i.e.* type or centro-strains) it is evident that one can establish the closeness of any unknown samples to a known organism, and thus effect the identification of the former. An excellent example of the discriminatory power of the approach is the demonstration[20] that one can use it to

distinguish four strains of *Escherichia coli* which differ only in the presence or absence of a single plasmid.

We have found[21] that it is possible to train an ANN using the Pyrolysis Mass Spectra as the inputs and the known concentrations of target analytes in standards as the outputs. The trained net can then be tested with the Pyrolysis Mass Spectra of "unknowns", and then accurately outputs the concentration of the target analyte(s), in the case described here the concentration of tryptophan in the growth medium of indole-positive strains of *E. coli*.

We have also been able to effect a rapid distinction between extra virgin and adulterated olive oils using this approach[22]. It is obvious that this combination of PyMS and ANNs constitutes a powerful technology for the analysis of the concentration of appropriate substrates, metabolites and products in any biological process.

Acknowledgements

This work is supported by the Wellcome Trust and through SERC LINK schemes with Aber Instruments, FT Applikon, Horizon Instruments, ICI Biological Products and Neural Computer Sciences.

References

1. R.G. Duggleby, Analysis of biochemical data by nonlinear regression - is it a waste of time?, *Trends Biochem. Sci.* **16**:51-52 (1991).
2. D.E. Rumelhart, J.L. McClelland and the PDP Research Group. "Parallel Distributed Processing. Experiments in the Microstructure of Cognition," MIT Press, Cambridge (Mass.) (1986).
3. J.L. McClelland and D.E. Rumelhart. "Explorations in Parallel Distributed Processing; A Handbook of Models, Programs and Exercises," MIT Press, Cambridge (Mass.) (1988).
4. T. Kohonen. "Self-Organization and Associative Memory," 3rd Ed., Springer, Heidelberg (1989).
5. Y.-H. Pao. "Adaptive Pattern Recognition and Neural Networks," Addison-Wesley, Reading (Mass.) (1989).
6. P.D. Wassernan and R.M. Oetzel. "NeuralSource: the Bibliographic Guide to Artificial Neural Networks," Van Nostrand Reinhold, New York (1989).
7. P.D. Wasserman. "Neural Computing: Theory and Practice," Van Nostrand Reinhold, New York (1989).
8. R.C. Eberhart and R.W. Dobbins. "Neural Network PC Tools," Academic Press, London (1990).
9. P.K. Simpson. "Artificial Neural Systems," Pergamon Press, Oxford, (1990).
10. J.A. Freeman and D.M. Skapura. "Neural Networks," Addison-Wesley, Reading (Mass.) (1991).
11. J. Hertz, A. Krogh and R.G. Palmer. "Introduction to the Theory of Neural Computation," Addison-Wesley, Redwood City (1991).
12. J.M. Zurada. "Introduction to Artificial Neural Systems," West Publishing, St. Paul (1992).
13. D.M. Wolpert and R.C. Miall, Detecting chaos with neural networks, *Proc. Roy. Soc. Ser. B* **242**:82-86 (1990).
14. J.C. Hoskins and D.M. Himmelblau, Artificial neural network models of knowledge representation in chemical engineering, *Comput. Chem. Eng.* **12**:881-890 (1988).
15. R. Pethig and D.B. Kell, The passive electrical properties of biological systems: their significance in physiology, biophysics and biotechnology, *Phys. Med. Biol.* **32**:933-970 (1987).
16. C.L. Davey, H.M. Davey and D.B. Kell, On the dielectric properties of cell suspensions at high volume fractions, *Bioelectrochem. Bioenerg.* **28**:319-340 (1992).
17. D.B. Kell and C.L. Davey, On fitting dielectric spectra using artificial neural networks, *Bioelectrochem. Bioenerg.* **28**:425-434 (1992).
18. H.L.C. Meuzelaar, J. Haverkamp and F.D. Hileman. "Pyrolysis Mass Spectrometry of Recent and Fossil Biomaterials," Elsevier, Amsterdam (1982).
19. C.S. Gutteridge, Characterisation of microorganisms by pyrolysis mass spectrometry, *Meth. Microbiol.* **19**:227-272 (1987).

20. R. Goodacre and R.C.W. Berkeley, Detection of small genotypic changes in *Escherichia coli* by pyrolysis mass-spectrometry, *FEMS Microbiol. Lett.* **71**:133-138 (1990).
21. R. Goodacre and D.B. Kell, The rapid and quantitative analysis of indole production by bacteria, using pyrolysis mass spectrometry and neural networks, *J. Gen. Microbiol.*, in preparation.
22. R. Goodacre, D.B. Kell and G. Bianchi, Neural networks and olive oil, *Nature* **359**:594 (1992).