Prediction of the Histidine-95 pK_a Perturbation in Triosephosphate Isomerase using an Electrostatically Trained Neural Network (SONNIC)

Howard B. Broughton,[a] Stuart M. Green[b] and Henry S. Rzepa[b]

[a] *Merck, Sharp and Dohme Research Laboratories, Neuroscience Research Centre, Terlings Park, Essex CM20 2QR, UK, E-mail: hbro@msdrl.com.*

[b] *Department of Chemistry, Imperial College of Science, Technology and Medicine, London SW7 2AY, UK, E-mail: rzepa@cc.ic.ac.uk, fgreen@cc.ic.ac.uk*

A backpropagation neural-network trained with AM1 SCF-MO derived molecular electrostatic potentials (MEPs) for a set of 30 substituted imidazoles of known ionisation constant can be used to predict the pK_a of imidazole under the influence of the α-helix dipole found for histidine-95 in triosephosphate isomerase; the results agree quantitatively with experimental evidence that this residue has an anomalous value such that its catalytic nature is unconventional.

The catalytic nature of histidine-95 in triosephosphate isomerase is a topic widely debated in the literature.[1] Recent experimental[2] and theoretical[3] investigations side with the view that the general acid corresponds to neutral rather than to protonated imidazole. The reduction in pK_a to *ca.* 4.5 required to achieve this is attributed to the charge-stabilisation effects of the α-helix dipole moment resulting from the alignment of dipoles in the peptide bonds.[4] A survey compiled by Hol[4] suggested that amino acids within three residues of the N-terminus of an α-helix may experience ionization constant reduction owing to such local electrostatic effects. Likewise, residues at the C-terminus have their pK_a correspondingly raised. Presently, there are several elegant examples of pK_a shifts of histidine residues lying at either end of helices.[5] We

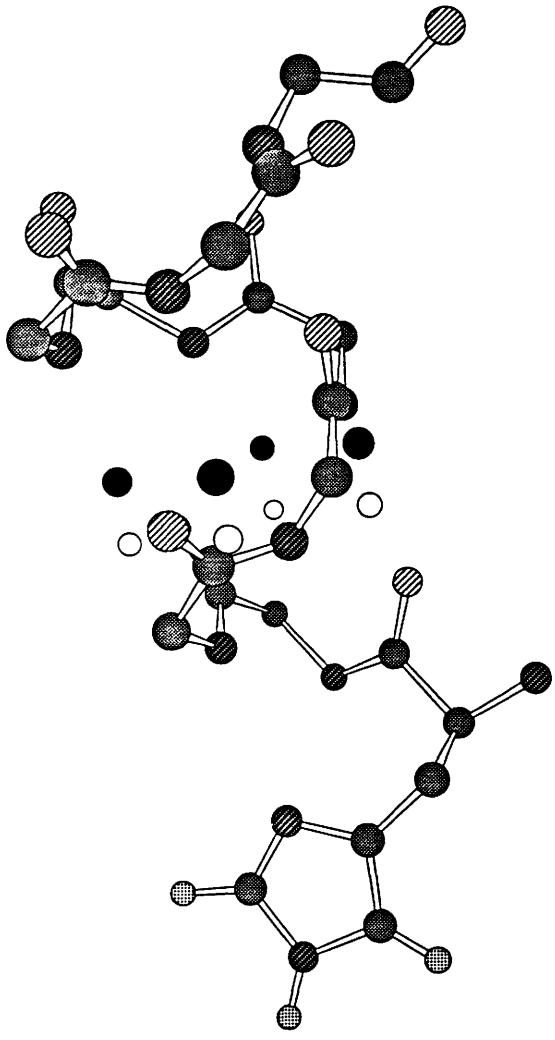

Fig. 1 α-Helical fragment his95-ser-glu-arg-arg-ser-tyr-phe102 of triosephosphate isomerase (side-chains removed), superimposed with the point charge model of the α-helix dipole of 20 D, 6 Å from the centre of the imidazole. The dipole alignment is similar to that of an α-helix if the unfilled spheres are point positive charges and solid black spheres are point negative charges.

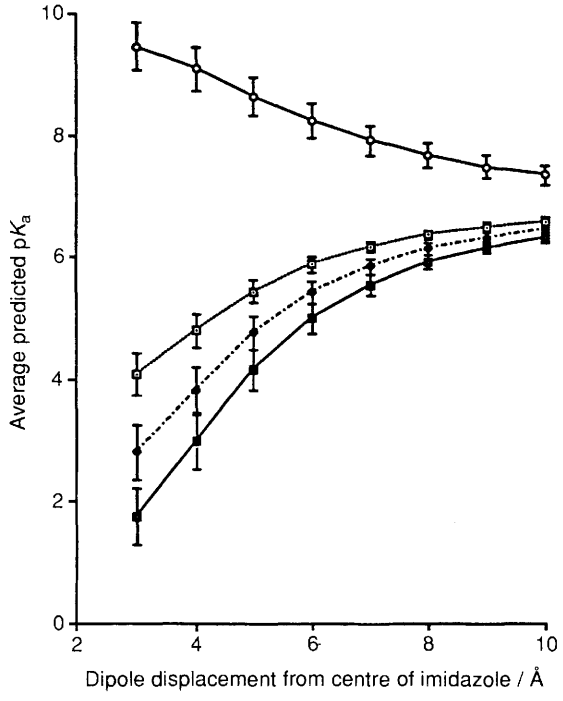

Fig. 2 Predicted pK_a values of imidazole obtained from a neural network as a function of the displacement and magnitude of a dipole vector, orientated as shown in Fig. 1. For dipole 10 D (dotted line), 15 D (dot-dashed line), 20 D (solid line, squares) and -20 D (solid line, open circles). Error bars are the standard deviation of average predicted pK_a for five differently seeded networks.

report here the use of neural networks trained on quantum mechanically derived electrostatic potentials to predict the magnitude of such pK_a perturbations.

Traditional quantitative structure–activity relationships (QSARs) used to predict pK_a values require a set of, *e.g.* Hammett, parameters dependant on the position and nature of the substituent. This approach is inapplicable to the pK_a of a residue perturbed by a local environment since *e.g.* an α-helix dipole orientation and induced three-dimensional polarising effects cannot be expressed as a simple scalar parameter. We have recently shown[6] that a high level of correlation exists in chemically significant spatial regions between quantum mechanically derived AM1 and PM3 molecular electrostatic potentials (MEPs) expressed as a three-dimensional grid of values, and the measured pK_a of *para*-substituted benzoic acids and anilines. A similar correlation can be demonstrated at the AM1 level[7] for 30 substituted imidazoles, covering a pK_a range of > 10, and an improved correlation is seen if *ab initio* (3-21G) wavefunctions are employed. Thus, calculated MEPs could prove useful as general descriptors of the acidity of a molecule in cases in which the three-dimensional perturbing environment may not correspond to any simple linear free-energy relationship. Other existing statistical techniques such as CoMFA (comparative molecular field analysis)[8] have also been used to demonstrate the utility of MEPs in this context, but these methods all presuppose some form of a non-parametric[6] or more restrictively a parametric[8] relationship between individual grid values of the MEP and the observable quantity.

Neural networks provide an alternative analytical tool in which no assumption is made about any such relationship; instead this is derived from the process of training the network.[9] We adopted here the strategy of training a backpropagation neural network[10] using as inputs a relatively coarse (2.5 Å) grid of electrostatic potential values calculated for 30 imidazoles and employing the semiempirical AM1 Hamiltonian. The network was trained† with 29 of the

† A three-layer neural network was used in which the input layer had 144 input nodes for a 2.5 Å grid, with one output node layer. It has been suggested[10] that a suitable estimate of the number of nodes in the hidden (middle) layer is approximated by the square root of the sum of the inputs and outputs. We found nine nodes sufficient. Typically training for a single random seed took *ca.* 15 min on a ≈5 Mflop machine; prediction times are negligible. More hidden layer nodes and/or more input nodes (≈330 for a 2.0 Å) increased computing time and did not improve the results. The network weights were randomly initialised between ±0.3, training was then performed with a momentum parameter of 0.7 using a fast adaptive learning algorithm (superSAB[13]). This method was found to be a great improvement in learning times over the traditional generalised delta rule.[10] The MOPAC program (V 6.0)[14] was used to calculate electrostatic potentials. The dipole vector was simulated by using a box of various dimensions (Fig. 1), with unit charges placed at the vertices. The data used to train the network corresponding to the following imidazoles;[15] 2-NH_2-4,5-Me_2 (9.21), 2,4,5-Me_3 (8.92), 2-NH_2-1-Me (8.65), 2,4-Me_2 (8.50), 2-NH_2 (8.46), 1,2-Me_2 (7.85), 2-Me (7.85), 1,5-Me_2 (7.70), 4-Me (7.56), 1-Me (7.30), 1,4-Me_2 (7.20), imidazole (7.00), 5-Br-1-Me (5.26), 5-Cl-1-Me (4.75), 2-Br-1-Me (3.88), 4-Br (3.88), 2-Br (3.85), 5-F-1-Me (3.85), 2-C (3.55), 4-Cl-1-Me (3.10), 4-F (2.44), 2-F (2.40), 2-F-1-Me (2.30), 5-NO_2-1-Me (2.13), 4-F-1-Me (1.90), 4-NO_2-2-Me (0.50), 4-NO_2 (−0.05), 2-NO_2-1-Me (−0.48), 4-NO_2-1-Me (−0.53), 2-NO_2 (−0.81).

imidazole MEPs and used to predict, in turn, each that had been left out. Such cross-validation was carried out for five different network seeds, establishing that a pK_a can be predicted to an rms error of 0.8 based on a network trained using AM1 derived MEPs. Accordingly, we next calculated the MEP of imidazole itself perturbed by dipole vectors of various magnitudes orientated as shown in Fig. 1, with the positive end closest to the heterocyclic ring, and the pK_a values were predicted using a network trained on all 30 imidazoles (Fig. 2). The known crystal structure[11] of triosephosphate isomerase contains an α-helix with a dipole of ≈20 D, with the positive end placed 6 Å from the centre of the imidazole. The pK_a perturbation to ≈4.5 predicted for such a dipole (Fig. 2) is remarkably similar to that estimated from experimental data by Knowles for this enzyme.[2] As confirmation that the orientation of the dipole vector is critical, reversing its sign is predicted to increase[4] the imidazole pK_a (Fig. 2), an effect established both experimentally[5] and theoretically[12] for *e.g.* barnase.

The accuracy of our technique may be improved both by the use of more accurate wavefunctions, and by modelling the full protein environment using a lattice of partial charges rather than as a simple dipole vector, such that higher moment perturbations would be included. Even at the current level of approximation, however, we believe that the use of electrostatically trained neural networks will have useful application to the prediction of environmental pK_a perturbations in enzymes. Indeed, the analysis and prediction of a wide range of non-linearly related properties of molecules may be possible using such spatially orientated neutral networks in chemistry (SONNIC).

We thank Merck Sharp and Dohme and the SERC for financial support (to S. M. G.).

Received, 8th June 1992; Com. 2/03021G

References

1 J. G. Belasco and J. R. Knowles, *Biochemistry*, 1980, **19**, 472; E. A. Komives, L. C. Chang, E. Lolis, R. F. Tilton, G. A. Petsko and J. R. Knowles, *Biochemistry*, 1991, **30**, 3011; E. B. Nickbarg, R. C. Davenport, G. A. Petsko and J. R. Knowles, *Biochemistry*, 1991, **27**, 5948.
2 P. J. Lodi and J. R. Knowles, *Biochemistry*, 1991, **30**, 6948.
3 R. C. Davenport, P. A. Bash, B. A. Seaton, M. Karplus, G. A. Petsko and D. Ringe, *Biochemistry*, 1991, **30**, 5821; P. A. Bash, M. J. Field, R. C. Davenport, G. A. Petsko, D. Ringe and M. Karplus, *Biochemistry*, 1991, **30**, 5826.
4 W. G. J. Hol, P. T. van Duijnen and H. J. C. Berendsen, *Nature*, 1978, **273**, 443; W. G. J. Hol, *Prog. Biophys. Mol. Biol.*, 1985, **45**, 149.
5 D. Sali, M. Bycroft and A. R. Fersht, *Nature*, 1988, **335**, 740; M. F. Perutz, A. M. Gronenborn, G. M. Clore, J. H. Fogg and D. T. Shih, *J. Mol. Biol.*, 1985, **185**, 491.
6 H. B. Broughton, S. M. Green and H. S. Rzepa, *J. Chem. Soc., Chem. Commun.*, 1992, 37.
7 For system with two or more nitrogen atoms, AM1 is superior to the alternative PM3 parameters in this type of calculation; H. S. Rzepa and M. Yi, *J. Chem. Soc., Perkin Trans. 2*, 1990, 943; H. S. Rzepa and M. Yi, *J. Chem. Soc., Perkin Trans. 2*, 1991, 531.
8 G. R. Marshall and R. D. Cramer, III, *Trends Pharmacol. Sci.*, 1988, **9**, 285; M. Clark, R. D. Cramer, III, D. M. Jones, D. E. Patterson and P. E. Simeroth, *Tetrahedron Comput. Methodol.*, 1990, **3**, 47; R. D. Cramer, III, M. Clark, P. E. Simeroth and D. E. Patterson, *Pharmacochem. Libr.*, 1991, **16**, 239; K. H. Kim and Y. C. Martin, *Pharmacochem. Libr.*, 1991, **16**, 151.
9 M. E. Lacy, *Tetrahedron Comput. Methodol.*, 1990, **3**, 119.
10 D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Nature*, 1986, **323**, 533; D. E. Rumelhart, G. E. Hinton and R. J. Williams, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1: *Foundations*, ed. D. E. Rumelhart and J. L. McClelland, pp. 318–362, MIT, Cambridge, 1986.
11 E. Lolis, and G. A. Petsko, *Biochem.*, 1990, **29**, 6619.
12 J. Aaquist, H. Luecke, F. A. Quiocho and A. Warshel, *Proc. Natl. Acad. Sci. USA*, 1991, **88**, 2026.
13 T. Tollenaere, *Neural Networks*, 1990, **3**, 561.
14 J. J. P. Stewart, MOPAC, Program 455, QCPE, University of Indiana, Bloomington, USA; F. J. Luque, F. Illas and M. Orozco, *J. Comp. Chem.*, 1990, **11**, 416; F. J. Luque and M. Orozco, *J. Comp. Chem.*, 1990, **11**, 909, G. G. Ferenczy, C. A. Reynolds and W. G. Richards, *J. Comp. Chem.*, 1990, **11**, 159; B. H. Besler, K. M. Merz and P. A. Kollman, *J. Comp. Chem.*, 1990, **11**, 431.
15 M. R. Grimmett, in *Comprehensive Heterocyclic Chemistry*, ed. A. R. Katritzky and C. W. Rees, Pergamon Press, 1984, vol. 5, p. 384.