Statistical method for surface pattern-matching between dissimilar molecules: electrostatic potentials and accessible surfaces

S Namasivayam and P M Dean*

Department of Pharmacology, University of Cambridge, Hills Road, Cambridge CB2 2QD, UK

This paper outlines a statistical method for pattern-matching between surfaces and is applicable to structural and energetic patterns found on molecular surfaces. Correlation coefficients generated for the pattern match are scale invariant. Regression analysis applied to the patterns reveals the scaling and displacement relationships. The method for measuring the similarities between molecular surfaces of two dissimilar molecules held in fixed orientations is given explicitly. Implicit in this procedure is a method for studying the inverse phenomenon, namely complementarity between surface parameters at a binding site and its ligand. The method has been used to assess surface differences in structural similarities generated by computer fitting and by visual comparison. Various pitfalls likely to be encountered in evaluating molecular structural similarities are noted.

Keywords: surface pattern-matching, accessible surface, electrostatic potential, rank correlation analysis, saxitoxin, tetrodotoxin

received 24 June 1985, revised 8 August 1985

In biomolecular interactions, studied within a quantitative-structure-activity relationship (QSAR), it is axiomatic that the 3-D geometries of the site-points determine the ligand-point interactions. The surface structure of the site forms a mould into which a ligand could fit. Site-points are atoms on the receptor surface that are thought to be involved in ligand binding; ligand-points are complementary positions on the ligand that can link the molecule to the receptor. Two ligands acting at the same site may have a very similar arrangement of atom point vertices in a topological frame; however, the edges of the topological network (bonds in the ligands) could be arranged very differently. In other words, the two ligands may be structurally unrelated but nevertheless bind, and have the same pharmacological action, at an identical set of site-points. Therefore the question arises: how can we assess similarity between unrelated molecular structures? The problem is simplified by the fact that in practice we may only be interested in that part of the ligand surface which fits into the binding site.

Consider the problem of visually comparing patterns contained within a patch on the accessible surfaces of two dissimilar molecules A and B. If the position of the patch on A is known, the orientation of A can be fixed while B is rotated round the Euler axes of the rigid body system; for each rotation step in B, the pattern presented by the patch on A has to be compared with an analogous region on B. If an orientation step in B is 10° then there are 36^3 comparisons to be made. Suppose now that the position of the patch on A is also unknown and A is incrementally rotated with the same step size, then approximately 2×10^9 comparisons would be necessary. The computing time, even with a large step size, would be formidable. An alternative approach is to carry out a rapid matching between putative ligand-points on the molecular surfaces to establish sets of feasible geometrical arrays of functionally corresponding ligand-points. In this paper the results of a search for corresponding hydrogen-bonding atoms, donors or acceptors, in the two ligands are used. This procedure has been chosen because the specificity in the interaction between ligand-points and a set of site-points is largely determined by directional hydrogen-bonding, although other interactions such as electrostatic and Van der Waals forces undoubtedly contribute to the energy of the molecular association. Geometrical searching of molecular structures has recently been examined in detail and its solution is similar to that of the archetypal travelling-salesman problem. The method employs an optimized branch-and-bound, recursive-descent, tree-search algorithm[1]. Once positional correspondences between two dissimilar molecules have been established, it is possible to superimpose the molecules and compare statistically the pattern match between the two accessible surfaces and any energetic parameter, such as the electrostatic potential, mapped

*Person to whom correspondence should be sent

onto the surface. The procedure used in this paper is scale invariant and investigates the similarity in pattern between two molecular structures held in a fixed relative orientation. Similarity is measured as a rank correlation coefficient between the surface parameters. If a linear relationship is assumed then differences in scale between the matched parameters can be scrutinized by regression analysis. The strength of the linear relationship can be determined from Pearson's product-moment correlation coefficient.

Statistical tests are applied to two models for matching; first a model based on matched hydrogen-bonding atoms only, and second a model for matching derived from visual assessment of similarities by experts. The methods outlined here could have widespread usage in medicinal chemistry since they break the deadlock in QSAR which relies heavily on comparing topologically related molecular structures within a congeneric series.

METHODS

Atom alignments

The two marine neurotoxins, saxitoxin (STX) and tetrodotoxin A (TTX), were selected for comparison since they are believed to exert their action at the same binding site on cell membranes, namely the sodium channel[2]. The molecules were kept in the conformations found in the crystal structures[3,4]. It is not known how the molecules are oriented at the binding site. These two molecules serve only as a model for surface pattern matching; it is not important within the scope of this paper whether in fact the molecules have an identical binding mechanism. The objective is to develop methods for comparing the similarity between molecular surface patterns.

The surface positions of hydrogen-bonding correspondences have been elucidated for STX and TTX[1]. (The structures of STX and TTX are shown in Figure 1). Three matches were found to give the best fits from

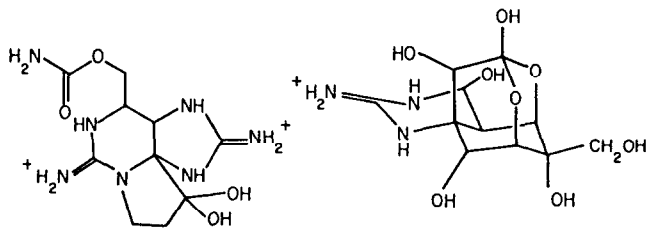

Figure 1. Structure of saxitoxin (STX) and tetrodotoxin (TTX)

all possible combinations of 4, 6, and 8 atoms. These matches, together with a visual match given by Hille[5] and Kao and Walker[6] (HKW match), have been used as starting points for pattern matching. Matched atoms in the two molecules were superimposed using the McLachlan algorithm[7]. The coordinates were rotated so that the mean plane of the matched atoms, where possible, was positioned parallel to the viewing plane and with the molecular centroids placed on the side of the plane distant from the observer. This manipulation allows the matched faces to be viewed graphically. The coordinate system of the molecules, in the superimposed positions, was translated along the z-axis to give $z = 0$ for the STX centroid. With the 4-atom match this procedure was not sufficient because the two centroids lie on opposite sides of the mean plane. In that case, the molecule with its centroid on the observer side of the viewing plane was rotated through 180° around the y-axis of the coordinate system.

Surface pattern matching

With the two molecules aligned in their positions for viewing, the accessible surface was calculated for each molecule as a set of z-values on a 0.5Å square grid projected on to the x–y plane. The probe radius employed for the surface calculation was 1.7Å[8]. The z-values were used to compare the two accessible surfaces. Molecular electrostatic potentials were then calculated at each z-value position from the computed CNDO charges [9] by the VSS method[10]. Therefore, at each position on the visible accessible surface there is a z-value and a value for the molecular electrostatic potential. Comparison of patterns on the accessible surfaces of the two molecules was studied only in those regions where overlap in the matched faces could be observed. Approximately 500 overlayed grid points were generated for each comparison.

The graphical display shown in Colour Plates 1 and 2 is taken from a Honeywell colour graphic recorder attached to a Sigma S5680 graphics terminal linked to an IBM 3081D computer. Blocks of colour representing the electrostatic potential were drawn first. The molecular skeleton, drawn in white, was computed next using the program of Beppu[11]. A surface shading program picks out the convex surfaces. This effect was achieved by placing the molecular structure in a block of space and projecting randomly spaced dots from the four edge faces of the block onto the accessible surface. A dot ratio of 3:1 from one face creates the illusion of shading[12].

Two statistical methods were used to compare the match (similarity) between surface parameters of STX and TTX. Spearman's rank correlation coefficient, R_{rank}, is independent of any assumption about the relationship between two sets of values; it is scale invariant. This fact is useful in comparing electrostatic potential surfaces derived from molecules possessing a different net charge. The coefficient is given by the equation

$$R_{rank} = \{(n^3 - n) - 6\sum_{k=1}^{n} d_k^2 - (T\vec{s} + T\vec{t})/2\}/\{[(n^3 - n) - T\vec{s}] \times [(n^3 - n) - T\vec{t}]\}^{1/2} \quad (1)$$

where d_k is the difference in ranks between corresponding members of a set of potentials or z-values, n is the number of observations to be compared, and $T^*_s = \Sigma(t_s^3 - ts)$ where ts is the number of ties (similarly for T^*_t). A value of $R_{rank} = 1$ would indicate that the two variables have a comparable fluctuation in xy-space although the members of one set may differ from the counterpart by a fixed amount and scale. This difference can then be assessed by regression analysis. Scattergrams of the data should be inspected to ensure that the correlation is monotonic.

For a linear regression, the relationship between variables for STX and TTX, Vs and Vt at corresponding spatial positions, k, is given by

$$V\mathrm{s}_k = \alpha V\mathrm{t}_k + \beta \quad (2)$$

where α is the regression coefficient and β is the regression constant. The goodness of fit between Vs, Vt and the regression equation (2) can be determined by Pearson's product-moment correlation coefficient, r, given by

$$r = \{\sum_{k=1}^{n}(Vt_k - \overline{Vt})(Vs_k - V_s)\}/$$
$$\{\sum_{k=1}^{n}(Vt_k - \overline{Vt})^2 \sum_{k=1}^{n}(Vs_k - \overline{Vs})^2\}^{1/2} \quad (3)$$

where r represents the fraction of the total variance that can be accounted for by the regression line of equation (2). The regression coefficient α and the constant β are obtained from

$$\alpha = \{\sum_{k=1}^{n}(Vt_k - Vt)(Vs_k - Vs)\}/\{\sum_{k=1}^{n}(Vt_k - Vt)^2\} \quad (4)$$

$$\beta = Vs - \alpha Vt \quad (5)$$

α is a scaling factor between Vs and Vt and β is the displacement between the two variables. If $R_{rank} = 1$, $r = 1$, $\alpha = 1$ and $\beta = 0$ the two surfaces are identical. The statistical methods for Spearman's rank correlation analysis and Pearson's product-moment analysis were programmed using the NAG library.

RESULTS

Of the three structural matches generated by Danziger and Dean[1], the 6-atom match gave a realistic best fit for corresponding acceptor/donor atoms of STX and TTX. The 4-atom match produced a fit that would be a mirror image of atom positions; the alignments did not allow the molecules to bind to the same set of site-points since the centroids were not found on the same side of the best fitting plane between the 4 atoms. The 6-atom match was therefore chosen to illustrate the statistical procedures. Colour Plates 1 and 2 illustrate the accessible surfaces and the superimposed molecular electrostatic potentials for STX and TTX for the 6-atom match with the matched faces oriented towards the observer. The accessible surfaces are shown by black shading and the molecular skeletons are drawn in white. Colour coded scales of the electrostatic potentials run from 324–683 kJmol^{-1} for STX and from 151–341 kJmol^{-1} for TTX. TTX and STX are singly and doubly charged cations respectively. Approximate similarities in the distribution of the potential can be observed visually. However, the regions of high potential on the two surfaces are not located at identical positions; for STX the locus is at $x = -2, y = 3$; for TTX the region is centred at $x = 2, y = 0$ (see Colour Plates 1 and 2). Furthermore, the region of high potential of TTX appears to spread over a greater area than for the analogous region in STX. These qualitative assessments of similarity are confirmed statistically.

In Figure 2 the regression line is drawn for the 6-atom matched accessible surfaces of STX and TTX. The scatter of points shows a good surface match with an $R_{rank} = 0.783$; Pearson's product-moment correlation coefficient, r, is 0.8 for the data on the regression line and this indicates a linear trend. In marked contrast, the back faces of STX and TTX, at the opposite side from the matched faces, show a wide scatter (see Figure 3); the surface correlation is poor, $R_{rank} = 0.036$ and

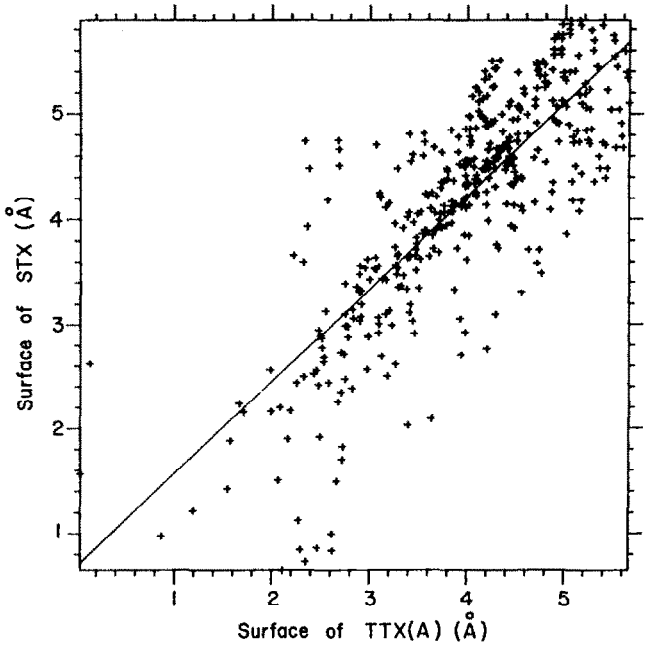

Figure 2. Scattergram for points on the accessible surface of STX plotted against corresponding points on TTX. The points are taken from the matched faces of the 6-atom match and the regression line is drawn. Axes denote the z-coordinate value for each point

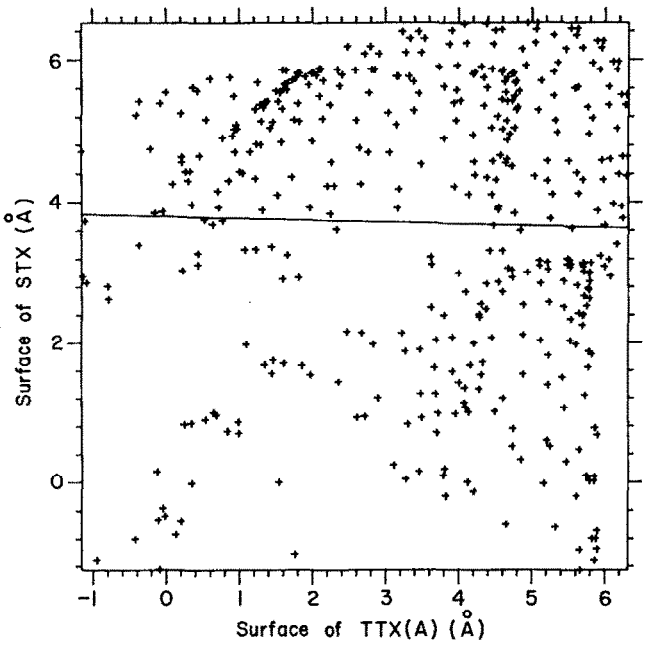

Figure 3. Scattergram for points on the accessible surface of STX plotted against corresponding points on TTX. The points are taken from the back faces of the 6-atom match and the regression line is drawn

$r = 0.025$. The statistical parameters for the accessible surfaces of all the matched faces between STX and TTX are given in Table 1. The 4-atom match shows the worst surface correlation; this is not surprising since the matched surfaces are an approximate mirror image of each other. Of the three hydrogen-bonded matches, the 6-atom match gives the best surface fit. However, of all the matches studied, the HKW match, obtained by visual comparison of surface groups, surpassed the matching based on hydrogen-bonding atoms; both correlation coefficients were about 0.9 with $\alpha = 1.05$ and

Table 1. Accessible surface matching between STX and TTX

Match	R_{rank}	r	α ± SEM	β ± SEM (Å)
4-atom	0.087	0.046	0.04 ± 0.05	4.1 ± 0.08
6-atom	0.783	0.800	0.88 ± 0.03	0.7 ± 0.13
8-atom	0.581	0.565	0.58 ± 0.04	1.9 ± 0.14
HKW	0.904	0.897	1.05 ± 0.02	0.5 ± 0.11

β = 0.5Å. These values for the HKW surfaces in the overlapped regions show the surface fluctuations between STX and TTX as having an identical scale, and also reveal that the surfaces are displaced by a separation of only 0.5Å. The surfaces of the HKW match therefore resemble each other closely.

The electrostatic potentials mapped onto the accessible surfaces and illustrated in Colour Plates 1 and 2 are analysed in more detail in Figure 4. Spearman's

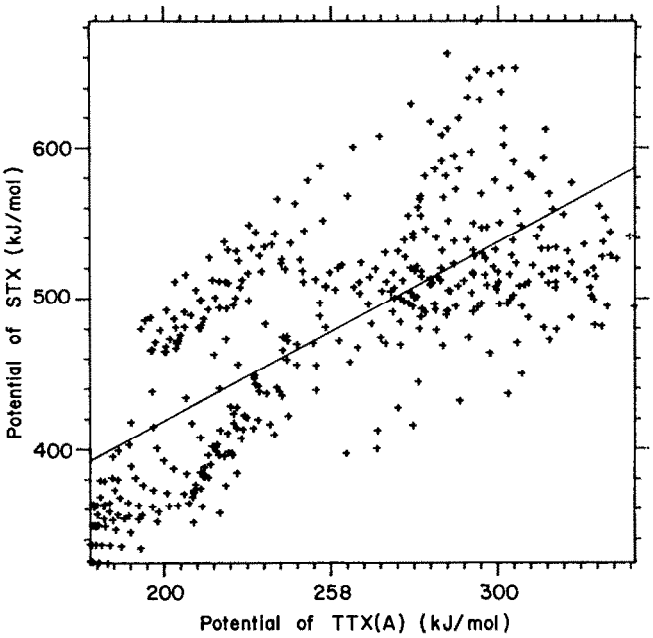

Figure 4. Regression data for the electrostatic potential computed at corresponding points on the accessible surfaces of STX and TTX. The points and the regression line are drawn for the matched faces of the 6-atom match

coefficient, R_{rank}, has a value of 0.723 for the 6-atom match and the linear regression has a correlation coefficient $r = 0.72$. This linear relationship between the two potential distributions shows the same scale, $\alpha = 1.18$, and a displacement of 181 kJmol^{-1} between the distributions. The 8-atom match gives a better fit than the 6-atom match; the 4-atom match exhibits poor correlations in surface potential. Once more, the back face of the 6-atom match shows virtually no correlation in fitting parameters and Figure 5 illustrates the wide scatter of points. Correlation coefficients for the back face are $R_{rank} = 0.167$ and $r = 0.233$. Statistical parameters for the potential at the matched faces are given in Table 2. Apart from the 4-atom match, all other matches show R_{rank} and r with values greater than 0.72, α values are close to 1 and the displacement values, β, are approximately 200 kJmol^{-1}. For all these matches the values are reasonable, but once again the HKW match gives the best fit.

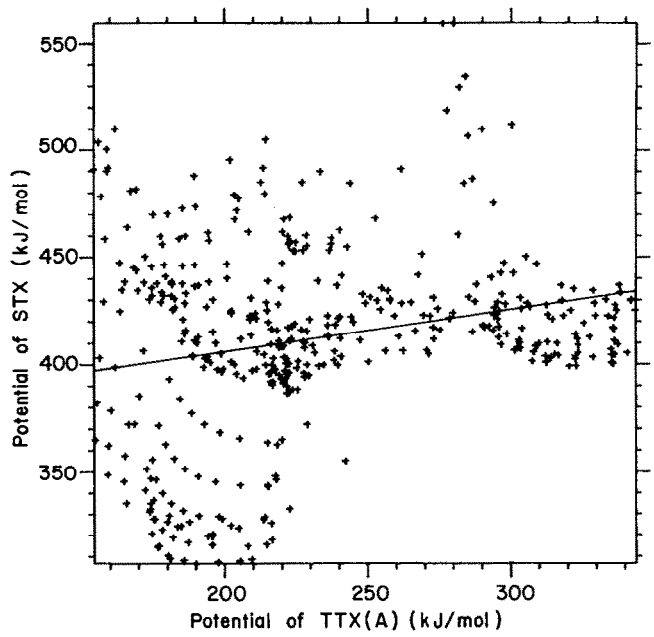

Figure 5. Regression data for the electrostatic potential computed at corresponding points on the accessible surfaces of STX and TTX. The points and the regression line are drawn for the back faces of the 6-atom match

Table 2. Molecular electrostatic potential matching on the accessible surface of STX and TTX

Match	R_{rank}	r	α ± SEM	β ± SEM(kJmol^{-1})
4-atom	0.532	0.525	0.71 ± 0.05	323.4 ± 14.1
6-atom	0.723	0.720	1.18 ± 0.05	181.0 ± 14.2
8-atom	0.814	0.790	1.10 ± 0.04	216.0 ± 9.1
HKW	0.912	0.871	1.23 ± 0.03	182.3 ± 7.5

DISCUSSION

The general problem of extracting surface features of similarity between dissimilar molecular structures is a challenging one. Selecting the best alignment for comparison is a formidable task; a brute force method, at Euler angles of 10°, gives too many sets of comparisons ($> 10^9$) to be assessed conveniently by current computing technology. However, once test surfaces have been defined, they can be evaluated readily by the statistical methods described in this paper. These standard methods can then be used to distinguish unambiguously between various models for molecular matching.

Four parameters for structure matching between STX and TTX can now be compared (see Table 3). The r.m.s.

Table 3. Ordering of structural similarities between STX and TTX expressed in descending order within each column of the table. Δd and Δr are taken from Danziger and Dean[1]

Δd	Δr	R_{rank} (surface)	R_{rank} (potential)
4	4	HKW	HKW
6	6	6	8
8	HKW	8	6
HKW	8	4	4

values for the difference distance matrix, Δd, and for the rotation of position vectors, Δr, are not equivalent but both measure the similarity in atom point positions[1]. From Table 3 it can be seen that all four methods for studying structural similarity give a different ordering for the quality of the atom matches. Atom positional matches are not necessarily correlated with accessible surface matches. These findings suggest strongly that caution should be exercised when interpreting results of structural comparisons. One pitfall to be aware of is the matching of approximate mirror images of atom arrangements; this potential hazard is well illustrated by the 4-atom match.

Two models for structural comparisons between the dissimilar molecules STX and TTX are examined in this paper. First a set of three structural matches (4, 6, 8-atom matches), developed by searching for correspondences in the positions of hydrogen-bonding atoms, were considered. Second, the molecules were visually aligned by experts (HKW match) identifying positions of similar groups in the two molecules. These models for matching are conceptually quite different; the first model compares similarities only between points in space, whilst the second compares bonded points and could not be extended strictly to the general case for totally dissimilar molecular structures. Notwithstanding this difficulty inherent in the second model, the results indicate that the first model, comparing only points in space, is less efficient in generating a surface match between STX and TTX. Moreover, it may be expected that after aligning equivalent intramolecular groupings there will be direct consequences for similarities between the electrostatic potentials; identical groups should produce closely related potentials at comparable positions nearby on the accessible surfaces.

An analogous concept to the assessment of similarity between two ligands is that of determining the complementarity between a binding site and its ligand. This problem is identical to measuring the fit between a mould and its cast. Nakamura *et al.*[13] have studied complementarity in electrostatic potentials between enzyme binding sites and their ligands or cofactors. The Van der Waals surface of the ligand is approximated to a polyhedron, each polygonal face, i, adjacent to the binding site is assigned two values of electrostatic potential, e^{G}_{i} and e^{H}_{i}, respectively for the potential generated at i by the guest (ligand) and the host (enzyme). Complementarity C^{GH}_{i} is then given by

$$C^{GH}_{i} = \text{sign}(e^{G}_{i}e^{H}_{i}) \times |e^{G}_{i}e^{H}_{i}|^{1/2} \quad (6)$$

and has units of energy. A large negative value for C^{GH}_{i} indicates good complementarity. However, this equation is an inadequate description of complementarity if one of the pair of potentials is strongly negative or positive and the other is close to zero.

The statistical measurement of similarity described in our paper has complementarity as a corollary and does not have the limitations associated with equation (6). A negative value for Spearman's rank correlation coefficient, R_{rank}, in equation (1) measures the inverse of similarity, that is, the complementarity of the pattern match between the two surfaces; regression analysis then takes account of the scaling and parameter displacement differences between host and guest molecules. Values for $R_{rank} = -1$, $r = -1$, $\alpha = -1$ and $\beta = 0$ would be produced by two surface patterns showing exact complementarity. Thus a measure of complementarity is straightforward. This statistical method should, therefore, have wide applicability in medicinal chemistry since one can compute similarity between surface variables in a set of drug molecules, and if the structure of the binding site is known, complementarity can also be calculated by the same statistical routines.

ACKNOWLEDGEMENT

P M Dean wishes to thank the Wellcome Trust for continued financial support.

REFERENCES

1 **Danziger, D J and Dean, P M** 'The search for functional correspondences in molecular structure between two dissimilar molecules' *J. Theor. Biol.* Vol 116 pp 215–224 (1985)

2 **Ritchie, J M and Rogart, R B** 'The binding of tetrodotoxin and saxitoxin to excitable tissue *Rev. Physiol. Biochem. Pharmac.* Vol 79 (1977) pp 2–50

3 **Furusaki, A, Tomiie, Y and Nitta, I** 'The crystal and molecular structure of tetrodotoxin hydrobromide' *Bull. Chem. Soc. Japan* Vol 43 (1970) pp 3332–3341

4 **Bordner, J, Thiessen, W E, Bates, H A and Rapoport, H** 'The structure of a crystalline derivative of saxitoxin. The structure of saxitoxin' *J. Am. Chem. Soc.* Vol 97 (1975) pp 6008–6012

5 **Hille, B** 'The receptor for tetrodotoxin and saxitoxin' *Biophys. J.* Vol 15 (1975) pp 615–619

6 **Kao, C Y and Walker, S E** 'Active groups of saxitoxin and tetrodotoxin as deduced from actions of saxitoxin analogues on frog muscle and squid axons' *J. Physiol. Lond.* Vol 323 (1982) pp 619–637

7 **McLachlan, A D** 'Rapid comparison of protein structures' *Acta Cryst. A* Vol 38 (1982) pp 871–873

8 **Richards, F M** 'Areas, volumes, packing, and protein structure' *Ann. Rev. Biophys. Bioeng.* Vol 6 (1977) pp 151–176

9 **Dobosh, P A** 'CNDO and INDO molecular orbital program CNINDO' *Quantum Chem. Program Exchange* Vol 9 (1968) program no. 141

10 **Giessner-Prettre, C** 'VSS' *Quantum Chem. Program Exchange* Vol 10 (1974) program no 249

11 **Beppu, Y** 'NAMOD: a computer program for drawing perspective diagrams of molecules' *Quantum Chem. Program Exchange* Vol 11 (1978) program no 370

12 **Dean, P M** 'Graphical methods for the analysis and display of the molecular electrostatic potential surrounding a drug or its binding site'. In *QSAR in Design of Bioactive Compounds* (Symposium chairman: Kuchar, M) Barcelona: J R Prous (1984) pp 253–264

13 **Nakamura, H, Komatsu, K, Nakagawa, S and Umeyama, H** 'Visualization of electrostatic recognition by enzymes for their ligands and cofactors' *J. Mol. Graph.* Vol 3 (1985) pp 2–11