# A program to find regions of similarity between homologous protein sequences using dot-matrix analysis

**Nicholas Gray**

*Physical Chemistry Laboratory, Oxford, UK*

*MATRIX is a program designed primarily to enable the user to visualize all regions of similarity between two proteins at a glance. The program helps the user to see where they are similar—at what relative positions in the amino acid sequences of the two proteins in question does the similarity exist; how they are similar—what functional characteristics the two similar sequences have in common; and to what extent they are similar—is the similarity significant, if so how significant relative to other similar sequences in the protein. This is achieved by constructing a diagram in which quantitative parameters of amino acids are used to compare every amino acid residue of the first protein with every amino acid residue of the second.*

*Another function of the program is, given two sets of atomic coordinates—either of different proteins or for the same protein (for self-comparison)—to demonstrate which residues of the two proteins, when the two proteins are superimposed upon each other, appear in the same space (or are close to each other).*

*Keywords: dot-matrix analysis, protein similarities, protein structure, homology modeling, Cytochromes P–450, $F_1$–ATPase β–subunit, adenylate kinase*

## INTRODUCTION

The program MATRIX uses the analytical method of dot-matrix analysis, or a 2-dimensional matrix display, to compare sequences of two homologous proteins. Comparison data are fed in separately as a table of comparison scores (see Table 1 for an example), so that, in effect, any property of amino acid side chains can be utilized to compare the two proteins. Examples of such properties include mutation substitution[1] (Table 1), hydrophobicity,[2,3] propensity to form a α-helix, β-sheet or β-turns,[4] fraction buried[5] and steric bulk. Another property which can be utilized in the comparison is the use of the 3-dimensional coordinates to generate distance values between residues (α-carbon atom, or every atom bar hydrogens).

Once the display has been set up, it can be manipulated to produce a clear picture to the preference of the user by zooming in on specific portions of the diagram, changing property used in the comparison, altering the contrast and brightness, removing background to any cutoff level, altering size of averaging window (see later) and choosing between color or monochromic displays.

Finally, the output of a program (RELATE[6]) which utilizes the method of local alignment using the Needleman–Wunch[7] algorithm to identify regions of protein similarity can be "laid over" the MATRIX display to compare regions of similarity identified by both algorithms.

The program has been written on a Silicon Graphics IRIS 3120 Workstation, using SVS FORTRAN–77.

## PROTEIN SEQUENCE SIMILARITIES

### Description

The primary function of MATRIX is to generate a visual means of comparing two amino acid sequences of two proteins, highlighting regions of similarity between the sequences, and demonstrating visually where, how and to what extent the regions are similar.

The format of the visual comparison is based on that suggested by Gibbs and McIntire,[8] that is, in the form of a diagram. The technique is used in many programs[9,10,11] for comparing protein and also nucleic acid sequences. The two-protein amino acid sequences form the two axes of a graph, forming a 2-dimensional coordinate image. Sample output of the program is shown in Color Plate 1. Each point on the graph corresponds to one amino acid residue on the vertical axis, in this example, a residue from the sequence of the yeast (*Saccharomyces cerevisiae*) enzyme cytochrome P450 lanosterol 14 α-demethylase ($P450_{14DM}$), and one amino acid residue on the horizontal axis, in this example, from

Address reprint requests to Dr. Gray at the Physical Chemistry Laboratory, South Parks Road, Oxford OX1 3QZ, UK.
Received 12 July 1989; accepted 19 September 1989

**Table 1. The mutation data matrix (Ref. 1). The values of this matrix represent likelihoods of each residue type being replaced by each other residue type by the process of evolution**

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | B | Z | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 2 | −2 | 0 | 0 | −2 | 0 | 0 | 1 | −1 | −1 | −2 | −1 | −1 | −4 | 1 | 1 | 1 | −6 | −3 | 0 | 0 | 0 | 0 |
| R | −2 | 6 | 0 | −1 | −4 | 1 | −1 | −3 | 2 | −2 | −3 | 3 | 0 | −4 | 0 | 0 | −1 | 2 | −4 | −2 | −1 | 0 | 0 |
| N | 0 | 0 | 2 | 2 | −4 | 1 | 1 | 0 | 2 | −2 | −3 | 1 | −2 | −4 | −1 | 1 | 0 | −4 | −2 | −2 | 2 | 1 | 0 |
| D | 0 | −1 | 2 | 4 | −5 | 2 | 3 | 1 | 1 | −2 | −4 | 0 | −3 | −6 | −1 | 0 | 0 | −7 | −4 | −2 | 3 | 3 | 0 |
| C | −2 | −4 | −4 | −5 | 12 | −5 | −5 | −3 | −3 | −2 | −6 | −5 | −5 | −4 | −3 | 0 | −2 | −8 | 0 | −2 | −4 | −5 | 0 |
| Q | 0 | 1 | 1 | 2 | −5 | 4 | 2 | −1 | 3 | −2 | −2 | 1 | −1 | −5 | 0 | −1 | −1 | −5 | −4 | −2 | 1 | 3 | 0 |
| E | 0 | −1 | 1 | 3 | −5 | 2 | 4 | 0 | 1 | −2 | −3 | 0 | −2 | −5 | −1 | 0 | 0 | −7 | −4 | −2 | 2 | 3 | 0 |
| G | 1 | −3 | 0 | 1 | −3 | −1 | 0 | 5 | −2 | −3 | −4 | −2 | −3 | −5 | −1 | 1 | 0 | −7 | −5 | −1 | 0 | −1 | 0 |
| H | −1 | 2 | 2 | 1 | −3 | 3 | 1 | −2 | 6 | −2 | −2 | 0 | −2 | −2 | 0 | −1 | −1 | −3 | 0 | −2 | 1 | 2 | 0 |
| I | −1 | −2 | −2 | −2 | −2 | −2 | −2 | −3 | −2 | 5 | 2 | −2 | 2 | 1 | −2 | −1 | 0 | −5 | −1 | 4 | −2 | −2 | 0 |
| L | −2 | −3 | −3 | −4 | −6 | −2 | −3 | −4 | −2 | 2 | 6 | −3 | 4 | 2 | −3 | −3 | −2 | −2 | −1 | 2 | −3 | −3 | 0 |
| K | −1 | 3 | 1 | 0 | −5 | 1 | 0 | −2 | 0 | −2 | −3 | 5 | 0 | −5 | −1 | 0 | 0 | −3 | −4 | −2 | 1 | 0 | 0 |
| M | −1 | 0 | −2 | −3 | −5 | −1 | −2 | −3 | −2 | 2 | 4 | 0 | 6 | 0 | −2 | −2 | −1 | −4 | −2 | 2 | −2 | −2 | 0 |
| F | −4 | −4 | −4 | −6 | −4 | −5 | −5 | −5 | −2 | 1 | 2 | −5 | 0 | 9 | −5 | −3 | −3 | 0 | 7 | −1 | −5 | −5 | 0 |
| P | 1 | 0 | −1 | −1 | −3 | 0 | −1 | −1 | 0 | −2 | −3 | −1 | −2 | −5 | 6 | 1 | 0 | −6 | −5 | −1 | −1 | 0 | 0 |
| S | 1 | 0 | 1 | 0 | 0 | −1 | 0 | 1 | −1 | −1 | −3 | 0 | −2 | −3 | 1 | 2 | 1 | −2 | −3 | −1 | 0 | 0 | 0 |
| T | 1 | −1 | 0 | 0 | −2 | −1 | 0 | 0 | −1 | 0 | −2 | 0 | −1 | −3 | 0 | 1 | 3 | −5 | −3 | 0 | 0 | −1 | 0 |
| W | −6 | 2 | −4 | −7 | −8 | −5 | −7 | −7 | −3 | −5 | −2 | −3 | −4 | 0 | −6 | −2 | −5 | 17 | 0 | −6 | −5 | −6 | 0 |
| Y | −3 | −4 | −2 | −4 | 0 | −4 | −4 | −5 | 0 | −1 | −1 | −4 | −2 | 7 | −5 | −3 | −3 | 0 | 10 | −2 | −3 | −4 | 0 |
| V | 0 | −2 | −2 | −2 | −2 | −2 | −2 | −1 | −2 | 4 | 2 | −2 | 2 | −1 | −1 | −1 | 0 | −6 | −2 | 4 | −2 | −2 | 0 |
| B | 0 | −1 | 2 | 3 | −4 | 1 | 2 | 0 | 1 | −2 | −3 | 1 | −2 | −5 | −1 | 0 | 0 | −5 | −3 | −2 | 2 | 2 | 0 |
| Z | 0 | 0 | 1 | 3 | −5 | 3 | 3 | −1 | 2 | −2 | −3 | 0 | −2 | −5 | 0 | 0 | −1 | −6 | −4 | −2 | 2 | 3 | 0 |
| X | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

the sequence of the bacterial (*Pseudomonas putida*) enzyme cytochrome P450 camphor 5′-exo-hydroxylase ($P450_{CAM}$) sequence. The axes of the graph are labeled by MATRIX with the names of the two proteins and as much information concerning the sequences as room will allow. As an example, the point could correspond to that linking serine–82 in the $P450_{CAM}$ sequence with serine–397 in the $P450_{14DM}$ sequence. The point is then shaded or colored according to a comparison between these two amino acid residues.

The simplest system which can be used is to follow the rules: if the two amino acid residues represented at the point are identical, for example both are serine residues, the point would be highlighted white; otherwise, the point would be left black. If these rules are carried out for every point in the diagram, the image displaying the similarity information is built up. Similarity between sequences is characterized by diagonal runs of highlighted points; for instance, if the two sequences were identical, each point in the longest diagonal (corner to corner) of the diagram would correspond to an identity—methionine–28 (sequence 1) and methionine–28 (sequence 2); alanine–41 (sequence 1) and alanine–41 (sequence 2). The result would be that the salient feature of this diagram would be a white diagonal corresponding to the whole sequence of both proteins. (N.B., the diagram is unlikely to consist simply of a corner-to-corner diagonal against a pure-black background since each residue will probably be represented more than once, for instance methionine–28 and methionine–121; this will result in a speckled background to the diagonal line.)

Given, then, that diagonals indicate regions of similarity between two protein sequences (even if the region consists of the entire sequence), and that the contiguity of the diagonal represents the extent of the similarity between the sequences, the diagram can thus be used to compare non-identical, but very similar protein sequences. The diagram here would clearly pick out similarities between very similar proteins such as human and monkey cytochromes *c*, but a problem arises in that the diagram would need to be used to show similarities between far less similar proteins such as the cited example $P450_{CAM}$ and $P450_{14DM}$, which using these simplest rules would describe the two P450 cytochromes as totally dissimilar.

Two improvements can be made by MATRIX on this simplest diagram which can enable similarities of the less similar proteins to be more clearly picked out. The first improvement takes advantage of the regularities used by nature in maintaining structure and function of a protein whose amino acid residues are being changed by mutation—the basis of evolution. Residues which play a more important role in the structure and function of the protein will tend to evolve more slowly and remain functionally similar (e.g., glutamate can evolve to aspartate in a region of conserved negative charge, or it can evolve to either alanine or leucine in a region promoting $\alpha$-helix formation[4]).

If each amino acid type is given a score according to a particular property, such as charge, hydrophobicity, bulkiness, propensity to form $\alpha$ or $\beta$ secondary structures or the free energy of transfer from the vapor to the aqueous phase of the residue, the points on the diagram can be shaded according to the similarity as opposed to identity of the compared amino acid residues. For instance, if the residues are compared according to charge, a comparison between glutamate and glutamate would produce a white point on both this diagram and the simpler diagram described earlier.

However, a comparison between glutamate and aspartate, which would produce a black (nonidentical) point on the simpler diagram, would produce an almost white point (similar charge) on the diagram using charge as a comparison. (N.B., a color range of red through to blue can be employed instead of monochrome, to clarify the difference between intensities.) Thus similarity between two compared sequences which had conservation of charge, but not conservation of specific residue type, would show up more clearly.

Dayhoff[1] and coworkers have produced a means of comparing amino acid residues by examining regions of proteins (which have evolved from a common ancestral protein) which perform similar roles. The result is the Dayhoff Mutational Substitution Matrix (Table 1) which contains comparisons between any two amino acid types. The value of the comparison is proportional to the probability that during evolution, one amino acid residue which performs some role in the structure or function of its host protein, will evolve into the other.

The Dayhoff matrix can thus be used as an initial means of comparing two sequences in a diagram. The advantage of this is that an overall evolutionary comparison between two protein sequences is made, providing the information: *where* the sequences are similar—the position of diagonals; *how* the sequences are similar—the probability of amino acids of conserved function evolving from one type to another; and *to what extent* the two sequences are similar—the lightness of the diagonal (lighter regions indicate closer similarity). More information on *how* the sequences are similar can then be produced by generating diagrams using more specific properties such as charge, hydrophobicity and size.

Thus, by using means to compare amino acid residues from two different protein sequences other than identity (simply deciding whether or not the two residues are identical), a diagram with clearer diagonals and containing more information, can be constructed.

The second improvement upon the simplest diagram tackles the problem that as the similarity between two sequences reduces, so more and more gaps in the corresponding diagonal are inserted until a stage is reached where the diagonal is sufficiently hidden by the speckled background to be undetectable, even though the similarities between the two sequences are still very significant. There are two techniques that can be employed to clarify these hidden diagonals:

(1) to increase the clarity of the diagonals with respect to the background;
(2) to remove the background.

MATRIX is able, in effect, to "smear" the points along the diagonals, thus "filling in" the gaps which would otherwise hide the diagonal line. The algorithm used to effect this smearing is:

$$\text{average intensity} = \sum_{a=-\frac{\text{window}}{2}}^{a=+\frac{\text{window}}{2}} \frac{i(x+a,\, y+a)}{\text{window}}$$

in which *window* is the number of points averaged (window size), $i(p, q)$ is the comparison score of point $[p, q]$ ($p$ and $q$ referring to amino acid positions in the two sequences), and $x$ and $y$ are the coordinates around which points are being averaged, and at which the final intensity will be plotted.

Having taken each point in turn, generated an average and plotted the new point, the resulting diagram contains all of the points "smeared" along their diagonals. The diagram is initially confusing due to the smearing of background points "cluttering" the display. MATRIX allows the diagram to be simplified by employing a cutoff value—a value representative of the minimum intensity displayed. Since this cutoff value is variable, it can be adjusted to remove unwanted background diagonals, and to highlight important similarity diagonals.

## Applications

When comparing two homologous proteins with evolutionary relationships, MATRIX can be utilized to identify the evolutionary characteristics, such as the degree of similarity between the two proteins, and so how closely they are related. Such analysis can be useful when creating "evolutionary trees" of proteins.

A major application for analyzing similarities between homologous proteins sequences is in the process of homology modeling.

The recent genetic revolution, amongst its advancements, has provided rapid DNA sequencing methods which has resulted in gene sequences being published at a faster rate than any other form of scientific data. It is tertiary structure, however, that is of more use to the scientist studying functional properties of proteins. This structure is derived from many proteins by means of X-ray crystallography, though this creates problems for proteins which cannot as yet be crystallized. Nuclear magnetic resonance offers little structural information due to the limitations in the size of the protein analyzed.

Experimental evidence[12] indicates that the native conformation of a protein is coded in its amino acid sequence, thus many efforts have been made to predict protein structure from the sequence data. An example of this is the secondary structure prediction technique of Chou and Fasman.[4]

The method of homology modeling involves the utilization of the known tertiary structure of a homologous protein in the 3-D modeling of the protein for which there is little or no tertiary structural information. Firstly, the regions of sequence similarity are identified, and these regions are used as a starting point in modeling the unknown 3-D structure of the protein. In the case of an enzyme, it is likely that the most significant similarities correlate with the enzyme active site, since the two homologous enzymes will be biofunctionally similar. The next stage is to mutate the side-chains at the positions of similarity from the enzyme of known structure to the enzyme of unknown structure. Finally the conformational energy can be minimized and theoretical binding energies can be calculated until the structure is fully refined. The technique can be applied to any protein which has a related homologous protein whose structure is already known.

$P450_{14DM}$ is a membrane-bound protein, hence is difficult to crystallize, so there is no X-ray data for the 3-D structure. However, the 3-D structure of the homologous $P450_{CAM}$ has been derived[13] to 1.63Å. $P450_{CAM}$ is a cytoplasmic

protein which forms good crystals. If the tertiary structure of $P450_{14DM}$ is derived, blocking agents may be discovered by modeling proposed agents near the cytochrome active site. Since $P450_{14DM}$ is an enzyme vital in the construction of a complete functional cell wall of yeast, and also of many fungi, a blocking agent would constitute a novel fungicide.

Table 2 is taken from Color Plate 1, which uses the Mutation Data Matrix to compare $P450_{CAM}$ with $P450_{14DM}$. It shows the six most significant visible diagonals (being all such diagonals with a length greater than 20 residues), arranged in order of significance. The two most significant diagonals from Table 2 coincide with the $O_2$-binding pocket (residues 250–253) and HR2 heme-binding domain (residues 351–360) moieties respectively. From this, it can be inferred (though at this stage, not concluded) that part of the $O_2$-binding pocket appears in the sequence 314–337, and that part of the HR2 heme-binding domain appears in the sequence 456–483 in the cytochrome $P450_{14DM}$. This information can thus be utilized in the first step of the homology modeling of $P450_{14DM}$ tertiary structure.

MATRIX can also be utilized to test the viability of 3-D structures modeled by alternative means. For instance, the tertiary structure of the $\beta$-subunit of the enzyme $F_1$-ATPase from *Escherichia coli* has been modeled[14] (residues 141 to 321 only, since this is the homologous region) by secondary structural prediction and from comparison with the structure of the homologous protein adenylate kinase (whose tertiary structure has been determined by X-ray crystallography). A MATRIX diagram to identify sequence similarities between these two proteins is shown in Color Plate 2. Examination of the diagram, considering that sequence homologies are represented as diagonals from top-left to bottom-right of the diagram, shows that the two sequences could be homologous in one of two ways (with $F_1$-ATPase residues in parentheses):

(1) homologous residues 0(136) to 35(170); 85(170) to 119(204); and 163(278) to 194(308);
(2) homologous residues 0(136) to 35(170); 33(208) to 52(227); 55(258) to 89(292) (split into two regions in this picture); 126(341) to 147(362); and 174(402) to 194(422).

Other diagrams generated for the two proteins (e.g., using $\alpha$-helix probability to replace the mutation data matrix in Color Plate 2) do not support the similarities found for Case 2, so Case 1 is the more likely. Case 1 also supports the model proposed for $F_1$-ATPase.

**Table 2. The most significant visible diagonals from Color Plate 1, arranged in order of significance. This order was decided by judgement after studying the color (bluer diagonals are more significant), length (longer diagonals are more significant) and contiguity (diagonals with less breaks are more significant) of the diagonals**

| CAM | 14DM | |
|---|---|---|
| 248 — 271 | 314 — 337 | $O_2$bp |
| 343 — 370 | 456 — 483 | HR2 |
| 173 — 199 | 221 — 247 | |
| 184 — 218 | 158 — 192 | |
| 31 — 51 | 314 — 334 | |
| 181 — 213 | 400 — 432 | |

The first region of similarity in Case 1 (residues 0–35 in adenylate kinase) appears at the position of the first $\alpha$-helix. It also appears at the predicted position of the homologous $\alpha$-helix of $F_1$-ATPase. This correlation is also supported by a diagram constructed using $\alpha$-helix probability in the sequence comparisons, which highlights this similarity.

The gap between the first and second regions of similarity in adenylate kinase (residues 35–85) appears to have been deleted in $F_1$-ATPase (residue 170 appears in both regions). This conforms with the proposed model in which a loop distal from the proposed active site of adenylate kinase is absent in the proposed $F_1$-ATPase structure.

The second region of similarity spans a long region of random coil (residues 85–119) in adenylate kinase believed to be very close to the active site, and includes a short region of $\alpha$-helix. The similar region in $F_1$-ATPase (residues 170–204) matches the equivalent region of coil in the proposed structure, lending further support to the model.

The third region of similarity (residues 163–194 in adenylate kinase) appears at the C-terminal $\alpha$-helix in the kinase. The similar region in $F_1$-ATPase (residues 278–308) covers the first half of the equivalent helix in the proposed model. The second half of the helix showed similarity with the adenylate kinase region (163–194) in another diagram using secondary structure probabilities to indicate similarity. This could be explained in two possible ways:

(1) the diagram using secondary structure probabilities showed the similarity out of phase (i.e., one sequence matched with the other slightly up or down the chain to the true homology) due to similarities of any short sequence of one $\alpha$-helix to any other;
(2) the $\alpha$-helix of $F_1$-ATPase has been lengthened by gene duplication.

In further support of both this work and of the proposed model, sequence similarities have been identified by computer searches andd by painstaking visual inspections (see Ref. 15 for a review) which correlate to the three regions of similarity described here.

## 3-DIMENSIONAL DISTANCE DIFFERENCE COMPARISONS

Using two sets of atomic coordinates—either of different proteins or for the same protein (for self-comparison)—MATRIX can demonstrate which residues of the two proteins, when the two proteins are superimposed upon each other, appear in the same space (or are close to each other). Consider for instance, two proteins which are homologous enzymes, have been superimposed with the same orientation, and with residues known to perform the same function occupying the same space (or very similar) in the superimposition. Other residues performing suspectedly the same function, such as parts of the active site—catalytic or binding sites, or conserved regions of the protein framework—vital for creating the necessary shape of the active site would be expected to be close in space. Comparison of two proteins in this manner may pick up similarities between the two structures which could be missed by simply comparing properties of residues (such as hydrophobicity, charge, pro-

pensity to form $\alpha$ or $\beta$ secondary, size, etc.). If one set of coordinates is from the cited example $P450_{CAM}$ and the other is a set predicted for $P450_{14DM}$ (e.g., by homology modeling), this function of the program can be used to check that the regions of the two proteins predicted to be similar using the first function of MATRIX, and hence believed to perform a similar role, actually appear in similar space. If two regions of similarity which were believed to be parts of the catalytic sites of the two enzymes were being studied, and the two enzymes were superimposed according to one of these regions, or if the other region proved to be in greatly differing positions, one or more of the following must be true:

(a) the model for $P450_{CAM}$ is incorrect;
(b) the assumption that the related sequence is part of the active site of both enzymes is incorrect;
(c) the similarity of at least one region in the two proteins is coincidence;
(d) divergent evolution has conserved the same sequence in the two proteins, but has given this sequence two different functions (cf., the immunoglobulin fold in IgG and $\beta$-2 microglobulin).

The other use for this function of MATRIX is to compare a protein 3-D structure with itself. Color Plate 3 shows a comparison of $P450_{CAM}$ 3-D structure, superimposed with itself. The long blue diagonal line shows where each residue is in exactly the same position as its superimposed counterpart. Diagonals running parallel to this line indicate sequences running parallel to each other in 3-D space (such as strands of parallel $\beta$-sheet), and diagonals perpendicular to this line represent sequences running antiparallel. If the line crosses the long diagonal line, this indicates that the sequence folds back upon itself, the "point" of the loop being at the residue corresponding to intersection. Secondary structure, such as $\alpha$-helices, can be recognized as lines running parallel to, and close to the main diagonal. This is due to the coil of the helix bringing every third residue into close juxtaposition. Thus, in this way, simple 3-D structural motifs can be recognized. One use of this type of diagram is to use it in the attempt to predict epitopes on proteins.[16,17] The philosophy is that discontinuous epitopes (antigenic determinants on the surface of a protein consisting of amino acids close in space, but not in residue position) can be picked out from the diagram.

## CONCLUSIONS

The main purpose of the program MATRIX is in the process of homology modeling to provide a means of visualizing at a glance the similarities between two homologous proteins as a starting point of such modeling. The program can then be utilized to study the progress of 3-D structural similarity between the protein of known structure and the proposed model of the protein of unknown structure.

Homology modeling aside, MATRIX can be used in the study of protein structure, providing an alternative viewpoint to the standard 3-D model style of portrayal of structure; and also in general comparisons of protein primary, secondary and tertiary structure. In short, MATRIX could prove to be an invaluable tool in any analytical process which involves the study and comparison of protein structures.

## ACKNOWLEDGEMENTS

I would like to thank Garrett Morris for his helpful discussions and also Dr. D. A. Harris for his help with the work on the two kinase proteins. Details of the program and its availability can be obtained from the author.

## REFERENCES

1 Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (M. O. Dayhoff, Ed.) NBRF, Washington, D.C., 1979, **5**, suppl 3, 345–362
2 Nozaki, Y. and Tanford, C. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions: establishment of a hydropathy scale. *J. Biol. Chem.* 1971, **246**, 2211–2217
3 Kyte, J. and Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 1982, **157**, 105–132
4 Chou, P. Y. and Fasman, G. D. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymology* 1978, **47**, 45–148
5 Chothia, C. The nature of the accessible and buried surfaces of proteins. *J. Mol. Biol.* 1976, **105**, 1–14.
6 George, D. G., Barker, W. C. and Hunt, L. T. The protein identification resource (PIR). *Nucleic Acid Res.* 1986, **14**, 11–15
7 Needleman, S. B. and Wunch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 1970, **48**, 443–453
8 Gibbs, A. J. and McIntire, G. A. The diagram, a method for comparing sequences. *Eur. J. Biochem.* 1970, **16**, 1–11
9 Staden, R. An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. *Nucleic Acids Res.* 1982, **10**, 2951–2961
10 Devereux, J., Haeberli, P. and Smithies, O. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* 1984, **12**, 387–395
11 Reisner, A. H. and Bucholtx, C. A. The use of various properties of amino acids in color and monochrome dot-matrix analyses for protein homologies. *CABIOS* 1988, **4**, 395–402
12 Antinsen, C. B., Huber, E., Sela, and White Jr., F. H. *Proc. Natl. Acad. Sci. (USA)* 1961, **47**, 1309
13 Poulos, T. L., Finzel, B. C. and Howard, A. J. High resolution crystal structure of cytochrome $P\text{-}450_{CAM}$. *J. Mol. Biol.* 1987, **195**, 687–700
14 Duncan, T. M., Parsonage, D. and Senior, A. E. Structure of the nucleotide-binding domain in the $\beta$-subunit of *Escherichia coli* $F_1$-ATPase. *FEBS Lett.* 1986, **208**, 1–6
15 Fry, D. C., Kuby, S. A. and Mildvan, A. S. ATP-binding site of adenylate kinase: Mechanistic implications of its homology with *ras*-encoded p21, $F_1$-ATPase, and other nucleotide-binding proteins. *Proc. Natl. Acad. Sci.* 1986, **83**, 907–911
16 Durant, J. C. The prediction of antigenic peptides. unpublished
17 Thomas, D. G., unpublished