# The matching of protein sequences using color intrasequence homology displays

**Garrett M. Morris**

Physical Chemistry Laboratory, South Parks Road, Oxford OX1 3QZ, UK

*The program presented here enables one to see at a glance important regions of similarity within two protein sequences — both position and extent of homology are depicted. Essential to this representation is a technique of coloring the sequences according to the identity of the amino acid that allows one to see why the segments are similar. Coloring is governed by the color grouping file, or cgf. The design of useful cgf's is explained. These can be based on a variety of properties, qualitative and quantitative; examples based on structural, chemical, physicochemical and statistical parameters are given. The resulting intrasequence homology display (IHD) can be manipulated in real time and zoomed into to study interesting regions in greater detail. The overall package represents a very flexible and powerful tool for homology modeling.*

*Keywords: intrasequence homology display (IHD), color grouping file (cgf), primary sequence, secondary structure, homology modeling, Cytochromes P-450, reverse transcriptases*

Received 5 May 1988
Accepted 17 May 1988

## INTRODUCTION

DNA sequences are pouring into the literature at a faster rate than any other form of scientific data. It is a simple matter to translate such gene sequences into protein sequences by using the universal genetic code. What the intrasequence homology display (IHD) presented here enables one to do is see at a glance exactly where, why and to what extent two protein sequences are similar.

Alignment score histograms provide the "where" and "to what extent," while color-coded sequence-axes and connectors give the "why." Each amino acid is represented as a colored bar in the sequence. Colors are assigned to groups of like amino acids, and these groupings can be based on:

(1) A physical property of the amino acid or model side-chain compound, like:
  1.1 hydrophobicity or hydropathy[1-3]
  1.2 charge/polarity
  1.3 acidity/basicity
  1.4 steric bulk
  1.5 free energy of transfer from vapour to aqueous phase[4]
(2) A statistically assigned secondary structure score, such as the relative frequency of an amino acid occurring in an α-helix, β-sheet or β-turn[5]
(3) Interesting or infrequently occurring amino acids
(4) The evolutionary distance score between two amino acids[6]

The design of color grouping files (cgf's) is explained, and specific applications of the so-called intrasequence homology display are given.

The first example compares Cytochromes P-450 camphor 5-exo-hydroxylase (P-450$_{cam}$) from the soil bacterium *Pseudomonas putida*[7] and lanosterol 14α-demethylase (P-450$_{14DM}$) from the yeast *Saccharomyces cerevisiae*.[8] The resulting display reveals the regions of primary sequence similarity in P-450$_{14DM}$ that coincide with the following P-450$_{cam}$ features:

(1) The distal helix and $O_2$-binding pocket
(2) The haem-binding domain HR2/cysteine ligand loop, and the proximal helix
(3) Several α-helices of the helix-rich domain

Furthermore, coloring the P-450 sequences according to hydrophobicity and charge reveals the fungal enzyme to be more hydrophobic overall, with an especially hydrophobic region in the first hundred or so residues. The bacterial enzyme, however, tends to have predominantly either amphiphilic or hydrophilic stretches. The hydrophobic segment in P-450$_{14DM}$ seems to be the N-terminal transmembrane segment that anchors the enzyme to the phospholipid membrane of the endoplasmic reticulum. An IHD colored according to alpha-helix score[5] is also able to reproduce the observed α-helix secondary structure in cytochrome P-450$_{cam}$.

The second example is a comparison between two pol polyproteins that contain reverse transcriptases, taken from the National Biomedical Research Foundation's[9] Protein Sequence Database. The sequences were taken from T-cell leukemia viruses, HTLV-II[10] and HTLV-III.[11] The resulting IHD demonstrates very clearly segments of strong similarity that imply regions of functional similarity in the two enzymes. It is proposed that secondary and tertiary structure prediction in these regions could lead to a model of the active site of reverse transcriptase; hence the design of inhibitors of these enzymes, present only in HIV-infected cells and important in the replication of retroviruses, is brought one step closer.

## INPUT DATA

The IHD program operates on the output of a Protein Identification Resource program from the national Biomedical Research Foundation, called RELATE.[9] It is a sibling program of the whole-sequence alignment program called ALIGN, which is based on the Needleman-Wunsch algorithm.[12] RELATE takes as one of its input variables a segment length, *L*, which is a number of amino acids. It is designed to detect unusual similarity between sequences by comparing all possible segments of length *L* from one sequence, with all segments of the same length from a second sequence. It works by placing a "window" of length *L* amino acids over the first sequence, starting at the first residue. It then positions a similar window over the start of the second sequence, and calculates a score for the pairwise alignment of corresponding amino acids contained in the windows based on a user-specified matrix. The mutation data matrix due to Schwartz and Dayhoff[6] is often used for finding similarities between evolutionary distantly related proteins. It is based on the amino acid replacements between present-day sequences and those inferred as common ancestors on evolutionary trees. The residues that did not change and the relative exposure of the sequences to mutational change were accounted for. Every possible pairwise combination of amino acids is assigned a score in this 23 by 23 matrix. The segment score is the sum of all the pairwise scores of the matched amino acids in the two windows being compared. The window over the second sequence is progressively shunted one amino acid along and an alignment score calculated for that displacement of the two windows, until the second window has traversed the whole of the second sequence. The highest alignment score for that run is recorded as a top score, along with the corresponding displacement and the positions of the two windows. ("Position" in this context is synonymous with residue number.)

The window on the first sequence is now moved along one amino acid and the whole process begins again, until the entire set of comparisons of the two sequences has been completed. In order to provide a statistical analysis of the scores, the whole process is repeated with an arbitrary number of randomized sequences. These are obtained by scrambling each of the two real sequences. The final output of the program includes a table of the statistically significant "top segment pair scores," given in descending order from the maximum score. The information given on each line is:

| Segment pair score | Sequence 1 position | Sequence 2 position | Displacement |
|---|---|---|---|
| **S** | **n1** | **n2** | δ |

and is interpreted thus:

"There is a statistically significant degree of homology between the two segments beginning at residue positions **n1** (in the first sequence) and **n2** (in the second); they have a segment pair score of **S** and a sequence displacement of **δ** (where **δ** = **n2** − **n1**)."

## IHD — INTRASEQUENCE HOMOLOGY DISPLAYS

The graphical analysis program IHD scans the table of top segment pair scores output from RELATE to find the maximum and minimum alignment scores. Hence IHD finds the range of scores and a scaling factor for drawing the alignment score histogram. IHD presents a graphical display of the results from RELATE (Color Plate 1). The screen is divided into essentially two halves. The upper half is occupied by the scores pertaining to the first sequence of the comparison, the lower half to the second. There is a central horizontal band running between these two regions, which is initially left blank, but which later on contains "connectors" between similar segments in the two sequence axes.

Although unusual in appearance at first, the lower histogram is inverted. The display is designed so that the two color-coded sequences are as near as possible to one another, thus facilitating direct comparison, but not so close that the connector zone is too small and confusing. A narrow horizontal band at the top of the screen is reserved for a key showing which amino acids belong to which color group (as can be seen in Color Plate 1).

The program begins the graphical output by drawing two horizontal axes centered about $x = 512$, the midpoint of the screen's $x$-range, with markers at a separation of 10 amino acids, and numeric labels at intervals of 50 amino acids. Then the two sequences are read in. For each sequence a colored vertical bar is drawn for each amino acid in the appropriate position along it sequence axis. The color chosen for the amino acid is governed by the color-grouping file (cgf).

The alignment scores are then plotted. For each line of top score data from RELATE, two white rectangles are plotted — one on each sequence-axis. The rectangle's height (or depth in the case of sequence 2) is proportional to the similarity score, **S**, and its width is proportional to the window length, **L**. The position of the rectangle along each "sequence axis" is determined by **n1** and **n2**.

The final stage of output involves drawing a connector between each pair of residues that belong to similar regions. So, for example, a line is drawn from residue **n1** in the first sequence to **n2** in the second. If the two amino acids connected belong to the same color group, then the connector is colored by the cgf according to the group identity of either amino acid. If, however, the two connected amino acids belong to different groups, then the connector is drawn as a grey dashed line. So "direct hits" — amino acids that belong to the *same group* — are easily spotted by looking at the connectors.

To keep the display clear, only one connector is drawn per window of *L* amino acids. Each connector is drawn from the midpoint of the upper window to the midpoint of the lower. This facilitates an immediate comparison of the first sequence window with that of the second. Furthermore, the segment of the second primary sequence that is similar is copied just below the first (and *vice versa*), also assisting the comparison.

Hence, regions of intrasequence similarity can be clearly picked out as features with tall peaks linked by bands of connectors of equal gradient. The broader this band, the higher the frequency of top scoring segments with the same displacement. For example, in Color Plate 1, the highest scoring region is between residues 234–290 and the band of connectors can be seen between the two sequence axes. Those gradients that are relatively small in magnitude correspond to large displacements ($|\delta| \geqslant 200$, say) and tend to be less meaningful or just coincidental similarities than connectors with larger magnitude gradients (N.B. These gradients are continuously variable when the two sequences are allowed to slide relative to one another.)

There is a second facility for drawing connectors, which draws them not just between the midpoints of windows, but between all *L* residues covered by the windows. Grey dashed lines between corresponding residues that join members of the dissimilar color groups are *not* drawn, as these would clutter up the display. Only those connectors which relate residues belonging to the *same* group are drawn. Now all direct hits in the windows clearly show up, not just the midpoint residue as before. This feature is particularly useful when studying a region of the sequences that has been zoomed up to.

Other facilities are interactive and allow the user to "interrogate" the display in real time about the identity of individual amino acids, sequences of *L* amino acids, and values of displacements between two similar segments. It is possible to slide one sequence parallel to the other so as to align particular segments, and to zoom up on any region to study its homology in greater detail. The program was written in the Window Manager, and so other jobs can be run simultaneously with the display. One particularly useful utility is the color editor, **cedit**, which permits easy adjustment of the (**R,G,B**) components of the color grouping currently in use (see the upper right-hand corner of Color Plate 1). By "attaching" to the color editor, the mouse can be used to select a group color in the key at the top of the display. Then the mouse can adjust the color's components manually by moving the (**R,G,B**) "sliders" in the **cedit** window. This facility is useful with a cgf that assigns each amino acid to a separate group. If one amino acid's color is selected with **cedit,** it is possible to highlight its occurrences in the sequences. The color editor also makes interactive experimentation possible in designing a clear color scheme.

There is one last facility that can be used to simplify a complex set of results. It involves interactively setting an alignment score "threshold" below which no histogram scores and no connectors are drawn. The level of the threshold is controlled with the mouse, each new threshold line being labelled with the corresponding score. So low scoring regions can be excluded and high scoring regions become obvious: important similarities can now be focused upon.

### The design of color grouping files

The most important factor to bear in mind is that the color grouping file (cgf) should be as simple as possible. Generally speaking, clearer patterns can be spotted if the number of amino acid groups is kept small. It is possible to give each amino acid a different color, but this would lead to a very confusing display. Differing intensities or saturation levels of colors may be given to similar groups of amino acids. For example, acidic residues could be red while slightly acidic could be pink.

A simple coloring scheme that proves to be very useful is that based on the hydrophobicity, charge and acidity of an amino acid — for example, phenylalanine is hydrophobic but arginine is basic. This example will be discussed in greater detail below.

### The structure of color grouping files

The cgf has the following FORTRAN-77 format:

| | |
|---|---|
| I2 | number of color groups, **N** (where **N** $\leqslant$ 23) |
| 3(I3, 1X), A23 | **R, G, B,** group member(s) |

Up to 23 different groups are allowed. Each color group can hold up to 23 members. No amino acid may appear in more than one group. Colors are described in **RGB** color-space. Given in the first three columns are the relative proportions of red, **R**, green, **G**, and blue, **B**, for each group's color. The brightest of each primary color corresponds to the value 255. Grey tones are generated by having equal amounts of the three primaries. Darker hues are obtained by decreasing the component values (**R, G, B**). No intensity corresponds to a component value of zero: so black would have the values (0, 0, 0) for (**R, G, B**). The fourth column contains the members of each color group. Each amino acid is referred to by its standard one-letter abbreviation, e.g., alanine would be "A." The list of amino acids forms one continuous string; there are no delimiters. For instance, if tryptophan, alanine and glycine were all members of the same group, then the corresponding group members would be "WAG." The abbreviations B (either asparagine or aspartic acid), Z (either glutamine or glutamic acid) and X (unknown amino acid) are also recognized.

## APPLICATIONS

### Color-grouping according to hydrophobicity: Cytochromes P-450$_{\text{cam}}$ and P-450$_{\text{14DM}}$

This cgf shows color-grouping of amino acids according to their hydrophobicity, charge and acidity/basicity at physiological pH. All hydrophobic nonpolar residues are colored white, amphiphilic residues grey, hydrophilic

**Table 1. Color-grouping file based on hydrophobicity, charge and acidity**

| charge.cgf | Color | Group members |
|---|---|---|
| 7 | | (number of groups) |
| 255 255 255 FPMVLI | white | Phe, Pro, Met, Val, Leu, Ile |
| 150 150 150 WAGX | grey | Trp, Ala, Gly, Unk |
| 0 200 0 NQTSC | green | Asn, Gln, Thr, Ser, Cys |
| 255 0 50 DE | red | Asp, Glu |
| 255 150 150 Y | pink | Tyr |
| 50 0 200 KR | blue | Lys, Arg |
| 150 150 255 H | light blue | His |

neutral residues green, while charged, acidic residues are red and basic ones blue. Pink denotes slightly acidic while light blue denotes slightly basic. The color-grouping file (suffix ".cgf") for this is given in Table 1.

This color selection was applied to the primary sequences of the two enzymes cytochromes P-450 camphor 5-*exo*-hydroxylase[7] (P-450$_{cam}$) & lanosterol 14α-demethylase[8] (P-450$_{14DM}$). The RELATE program was run with a segment length of 40 residues. This showed the latter yeast enzyme to have more white stretches and therefore to be *more hydrophobic* than its bacterial analogue. These white regions predominate in the first 100–150 residues of the yeast enzyme, and seem to indicate the presence of a hydrophobic "tail" by which this membrane-bound protein is anchored to the cell wall. This is indeed borne out by experimental evidence because, although the bacterial enzyme P-450$_{cam}$ is water soluble, the yeast cytochrome P-450$_{14DM}$ is *lipid soluble.* (See Color Plate 1.)

## Parameterized color-grouping files

These color schemes are based on a parameter of an amino acid: the smaller the parameter, the darker the shade. The initial scheme is a range of greys, the color components being a linear mapping of the parameter range. Each amino acid belongs to a separate color group. This type of color scheme is more amenable to interactive analysis and experimentation, because individual amino acids can be addressed directly. Important amino acids — those with relatively high or low parameter values — can be given the same distinct color to enhance their significance. Using the interactive color editor, the number of distinctly colored residues can be varied so as to discover the extent of "important" regions. Of course, the initial appearance of the IHD is not very informative, as there is a bland expanse of grey shades on the sequence axes. It is only by experimenting with the color editor that interesting features can be revealed and emphasized.

Such parameterized color schemes can be based on the following amino acid parameters:

(1) Hydrophilicity[1]

(2) Hydrohilic score[2] These data were adjusted to highlight antigenic regions of proteins;

(3) α-helix score[5]

(4) β-sheet score[5] The values for each amino acid are the fraction found in the appropriate region divided by 5.

(5) β-turn score[5] The average fraction for all amino acids. 29 proteins were used to collect data from.

(6) Free energy[4] The values are the free energies of transfer from the vapor to the aqueous phase of model side-chain compounds. The value for Pro is not given: when estimated from the data of Kyte and Doolittle (1982) it is found to be −7.130.

(7) Fraction buried[13] The values are the fraction of each amino acid that is inaccessible to water (100% buried), as compiled from 12 calcium-binding proteins.

(8) Hydropathy index[3] The data of Wolfenden *et al.*[4] and Chothia[13] were correlated to give a unified hydropathy scale.

## Parameterized color-grouping according to α-helix score

Table 2 gives the α-helix score for each amino acid, which is the fraction found in α-helix regions divided by the average fraction found for all amino acids of

**Table 2. α-Helix scores of the amino acids**

| Amino acid | | α-Helix score |
|---|---|---|
| Ala | A | 1.420 |
| Arg | R | 0.980 |
| Asn | N | 0.670 |
| Asp | D | 1.010 |
| Cys | C | 0.700 |
| Gln | Q | 1.110 |
| Glu | E | 1.510 |
| Gly | G | 0.570 |
| His | H | 1.000 |
| Ile | I | 1.080 |
| Leu | L | 1.210 |
| Lys | K | 1.160 |
| Met | M | 1.450 |
| Phe | F | 1.130 |
| Pro | P | 0.570 |
| Ser | S | 0.770 |
| Thr | T | 0.830 |
| Trp | W | 1.080 |
| Tyr | Y | 0.690 |
| Val | V | 1.060 |
| Asx | B | 0.840 |
| Glx | Z | 1.310 |
| Unk | X | 1.000 |

the sequence.[5] The data were collected from 29 proteins. The corresponding color-grouping file based on the linear-mapped grey scale is given in Table 3.

The values for the primary color components were calculated as follows: the maximum α-helix score, **αmax**, and the minimum α-helix score, **αmin**, were found. The range of scores, **δα**, was calculated (**δα = αmax − αmin**). Then each score, **α**, was taken in turn and its red, green and blue components calculated:

$$\mathbf{R} = \mathbf{G} = \mathbf{B} = 255 \times (\boldsymbol{\alpha} - \boldsymbol{\alpha}\mathbf{min})/\boldsymbol{\delta\alpha}$$

Hence this generates a grey scale of intensity, in which the smallest α-helix score appears black and the largest as white.

Those amino acids whose score is greater than 1.00 can be colored magenta, say (by setting **G** = 0) because these are more likely than the average to belong to an α-helix. Very high scoring amino acids, i.e., scores greater than 1.40, are colored red (by setting **B** and **G** = 0). The relative shade of the residue is left unchanged by setting one or two components to zero.

The greater the α-helix score, the more likely the amino acid at that residue position belongs to an α-helix. Thus the red-shaded amino acids will help to clarify regions of likely α-helices, and if similar regions in the two sequences both show up as red, then it is probable that there is a common structural element, namely, an α-helix.

This technique of coloring does not have to give individual amino acids different colors or shades of grey. Indeed, this can lead to a more complicated coloring scheme, and the real power of the technique is lost. The point about this technique is that amino acids can be *grouped* together so as to give them a common color.

**Table 3. Color-grouping file for α-helix scores: linear-mapped grey scale**

| **alphah.cgf** | | | | |
|---|---|---|---|---|
| 21 | | | | (number of groups) |
| 230 | 230 | 230 | A | |
| 112 | 112 | 112 | R | |
| 28 | 28 | 28 | N | |
| 120 | 120 | 120 | D | |
| 36 | 36 | 36 | C | |
| 145 | 145 | 145 | Q | |
| 255 | 255 | 255 | E | |
| 0 | 0 | 0 | G | |
| 117 | 117 | 117 | H | |
| 138 | 138 | 138 | I | |
| 173 | 173 | 173 | L | |
| 161 | 161 | 161 | K | |
| 240 | 240 | 240 | M | |
| 153 | 153 | 153 | F | |
| 0 | 0 | 0 | P | |
| 54 | 54 | 54 | S | |
| 71 | 71 | 71 | T | |
| 138 | 138 | 138 | W | |
| 33 | 33 | 33 | Y | |
| 133 | 133 | 133 | V | |
| 117 | 117 | 117 | X | |

## Three-zoned color-grouping based on α-helix score: Cytochromes P-450$_{\mathrm{cam}}$ and P-450$_{\mathrm{14DM}}$

The alternative method is to decide beforehand on the cutoff values for different "zones" of α-helix scores. The best cutoffs fall naturally where there is a relatively large jump in the series of values. If this procedure is followed to divide the α-helix scores into three zones, cutoffs are found between valine (1.06) and aspartic acid (1.01) and also between arginine (0.98) and Asx (0.84). This color-grouping file is shown in Table 4 (the assigned colors are arbitrary, but fall in a spectral order).

Red-colored amino acids are likely to be members of an α-helix, while blue ones are not. Green amino acids are intermediate in frequency. This is more promising than the linear-mapped grey scale, because the human eye is better at "blocking" when there is a small number of groups.

This cgf was first tested with Cytochromes P-450$_{\mathrm{cam}}$ from the soil bacterium *Pseudomonas putida*[7] (whose tertiary structure was already known) and P-450$_{\mathrm{14DM}}$ from the yeast *Saccaromyces cerevisiae*[8] (for which only the primary structure was known). Like the earlier example, RELATE was run with a segment length of 40 residues. A particularly telling region runs from residues 234–292. This contains the α-helices I (distal helix), J and K. It will be seen from Color Plate 2 that there are mainly red stretches in this region, which also happens to be very similar to the stretch 300–356 in P-450$_{\mathrm{14DM}}$. There is an antihelical blue connector between residues 268(cam) and 334(14DM): this is where α-helix I ends and α-helix J begins in P-450$_{\mathrm{cam}}$. There is a type I β-turn in P-450$_{\mathrm{cam}}$ from 277–281, between α-helices J and K. This segment shows up with **alpha3.cgf** as green-blue-red-green in the bacterial sequence: indeterminate α-helix. Then α-helix J follows this β-turn, with four red connectors showing up at a lower score threshold. Furthermore, the blue connectors at 248, 249 and 252 and the green at 251 (in P-450$_{\mathrm{cam}}$) coincide exactly with the residues of the distal helix that do not take part in the regular α-helix hydrogen-bonding pattern; it is these residues that form a "kink" in the distal helix, creating the $0_2$-binding pocket of cytochrome P-450$_{\mathrm{cam}}$.

This cgf was then applied to two pol polyprotein sequences; the first one is a 109.9 kDa protein with 982 residues, containing the reverse transcriptase and endonuclease of HTLV-II (human T-cell leukemia virus II) due to Shimotohno *et al.*[10]; and the second is a 115.0 kDa protein of 1015 residues, containing the protease, reverse

**Table 4. Three-zoned color-grouping file based on α-helix score**

| **alpha3.cgf** | | | | Color |
|---|---|---|---|---|
| 3 | | | | (number of groups) |
| 255 | 0 | 0 | EMAZLKFQIWV | Red |
| 0 | 255 | 0 | DHXR | Green |
| 0 | 0 | 255 | BTSCYNGP | Blue |

**Table 5. Regions of intrasequence homology in HTLV-II and HTLV-III**

| HTLV-II residues | HTLV-III residues | Displacement, δ |
|---|---|---|
| 194–280 | 272–358 | 78 |
| 333–363 | 410–440 | 77 |
| 704–727 | 751–774 | 47 |
| 824–854 | 860–890 | 36 |

transcriptase and endonuclease of HTLV-III (human T-cell leukemia virus, BH10) due to Ratner *et al.*[10]

### Color-grouping according to α-helix score: reverse transcriptases

Color Plate 3 is obtained by applying **alpha3.cgf** to the sequences of two pol polyproteins that contain reverse transcriptases, taken from two human T-cell leukemia viruses HTLV-II (upper sequence) and HTLV-III (lower sequence). RELATE was run with a segment length of 20. There are clearly prominent regions of similarity in both sequences (see Table 5).

The above residue ranges were obtained by setting the score threshold at 286.0, thus excluding the lower scoring regions. The regions for similarity seem to be α-helices at 248–257, 346–351 and 711–718 (in HTLV-II), corresponding to stretches 326–335, 423–428 and 759–766 (in HTLV-III), respectively, because helical red connectors predominate here. The similarity of two regions may, of course, arise because of other structural features: such cases where an α-helix seems not to be present are 222–233 and 833–840 (in HTLV-II) matching 300–311 and 869–876 (in HTLV-III), respectively, where antihelical blue connectors cluster together.

## HARDWARE AND SOFTWARE

The program presented here was written on a Silicon Graphics IRIS 3120 Workstation, with 20 bitplanes and double buffering. However, the most important requirements of the hardware are that it has a reasonably high resolution monitor and that it is able to display a reasonably wide range of colors — sixteen would suffice.

The version used to generate these results was written in SVS FORTRAN-77, an implementation of the full ANSI FORTRAN-77 computer programming language for Motorola MC68000 and MC68020. The color monitor has a screen resolution of $(0 \leqslant x \leqslant 1023)$ by $(0 \leqslant y \leqslant 767)$ pixels. Up to 4096 colors may be displayed simultaneously, being available from a palette of $2^{24}$ colors.

## CONCLUSIONS

The intrasequence homology display (IHD) presents a simple, direct and visual way of analyzing the results of sequence homology studies. Its power lies in its flexibility. Groupings of amino acids can be chosen to highlight a wide variety of properties: structural, chemical, physicochemical and statistical.

Important regions of primary sequence similarity can be identified between biofunctionally similar enzymes. If just one of the sequences compared has a known tertiary structure, then it is possible to model the region of similarity in the second sequence. Using the techniques of molecular modeling, the residues at the "similarity site" in the known tertiary structure can be mutated into those of the second, unknown sequence, pending a considered examination of the steric compatibility of the two segments. Then the reiterative process of calculating semi-empirical energy refined structures commensurate with the known substrate and inhibitor binding properties begins, until a valid model of the active site is obtained.

Thus the IHD lies at the root of all such "homology modeling," advising which residues should be mutated and into what. The IHD provides invaluable facilities that assist in the interpretation of sequence homologies. It identifies those salient regions of similarity that may be the cause of the enzymes' biofunctional similarity. And it ultimately ensures that subsequent modeling rests on a more solid and reliable foundation.

## ACKNOWLEDGEMENTS

I thank Dr. Neil Barclay for his assistance in accessing the Protein Identification Resource programs; the Protein Identification Resource itself (address as in reference 9) for allowing the use of their programs; and Dr. Federico Gago for his helpful discussions.

## REFERENCES

1 Hopp, T. P., and Woods, K. R. *Proc. Nat. Acad. Sci. USA* 1981, **78**, 3824–3828

2 Levitt, M. *J. Mol. Biol.* 1976, **104**, 59–107

3 Kyte, J., and Doolittle, R. F. *J. Mol. Biol.* 1982, **157**, 105–132

4 Wolfenden, R. *et al. Biochemistry* 1981, **20**, 849–855

5 Chou, P. Y., and Fasman, G. D. *Annu. Rev. Biochem.* 1978, **47**, 251–276

6 Schwartz, R. M., and Dayhoff, M. O. in *Atlas of Protein Sequences and Structure* (M. O. Dayhoff, ed.), NBRF, Washington, D.C., 1979, **5**, suppl. 3, pp. 353–358

7 Private communication of coordinates; Poulos, T. L., Finzel, B. C., and Howard, A. J. *J. Mol. Biol.* 1987, **195**, 687–700

8 Kalb, V. F. *et al. DNA* 1987, **6**, no. 6, 529–537

9 George, D. G. *et al.* PIR Report REL-0185, 1985. Protein Identification Resource, National Biomedical Research Foundation, Georgetown University Medical Center, Washington, D.C. 20007, USA

10 Shimotohno, K. *et al. Proc. Nat. Acad. Sci. USA* 1985, **82**, 3101–3105. (Sequence translated from the DNA sequence. The authors translated the codon TCC for residue 637 as Ala.)

11 Ratner, L. *et al. Nature* 1985, **313**, 277–284. (Sequence translated from the DNA sequence.)

12 Needleman, S. B., and Wunsch, C. D. *J. Mol. Biol.* 1970, **48**, 443–453

13 Chothia, C. *J. Mol. Biol.* 1976, **105**, 1–14