Sequence similarity scores and the inference of structure – function relationships

Michael S. Chapman

Abstract

Improved methods are described for the interpretation of two or more aligned protein or nucleic acid sequences. These methods can be used to interpret the possible biological importance of regions within a known three-dimensional structure, or, even without a structure, to correlate sequence similarity with the known function of particular amino acids and to associate sequence similarity with properties predicted from the sequences. Improvements include the calculation of a position-dependent, gap-penalized similarity score; computer-assisted graphical association of sequence similarity with structural, functional or chemical properties of the sequences; and statistical comparisons of the sequence conservation or variability of different groups of residues. An application is described to analyze the sequences of piconarviral capsid proteins.

Introduction

Over 40 000 protein sequences are currently accessible through databases such as the Protein Identification Resource (PIR; George *et al.*, 1986). As, by some estimates, there may be as few as 1000 families of distinct protein folds (Chothia, 1992), many of the 10 000 sequences that are currently added annually mostly increase the number of sequences within each group of related proteins. It is sometimes possible to associate functional significance with conserved residues by examining the alignment of related sequences. There has been much progress towards the recognition of sequences that are similar to each other (Gribskov *et al.*, 1990; Schuler and Altschul, 1991; Karlin and Brendel, 1992) or similar to known three-dimensional structures (Bowie *et al.*, 1991), and in the alignment of multiple sequences (Gribskov *et al.*, 1987; Vingron and Argos, 1989; Bacon and Anderson, 1990; Barton, 1990; Vihinen, 1990; Taylor, 1990). Some of these methods are implemented in comprehensive sequence analysis packages such as that of Genetics Computer Group (GCG) Inc. (1991; Devereux *et al.*, 1984). Methods for the interpretation of sequence alignments have received less attention, but are considered here and implemented in a program that is compatible with the GCG package.

Improvements have been made in the calculation of a position-dependent sequence similarity score; in the correlation of sequence conservation with structure or functional sites; and in statistical tests of significance upon the relative sequence similarity of different groups of residues. The sequence similarity score is calculated at each position in the sequence from the average of residue comparison scores of all possible pairs of aligned residues and gap penalties if there are insertions or deletions. Prior statistical treatments (Schuler and Altschul, 1991; Karlin and Brendel, 1992) have focused upon the significance of patterns of amino acids or nucleotides that are found repeatedly through the sequence database. The advantage of these methods is that they require no knowledge of the molecular structure or the function of its constituent residues. A different approach may be appropriate if, as is increasingly the case, the three-dimensional structure is known for one of the aligned sequences, or the functional importance of specific residues has been determined by experimental techniques. The methods discussed here present sequence similarity, structure, functional and chemical properties in ways that are conducive to recognition of possible correlations, and statistically compare the sequence similarity or properties of selected groups of residues with suitably chosen control groups. The methods are therefore particularly useful when combining the information from sequence alignment and three-dimensional structure.

Recognition of sequence – function correlations is facilitated through interactive graphics that are an extension of the capabilities of the program *Plotsimilarity* (Devereux *et al.*, 1984; Genetics Computer Group Inc., 1991). *Plotsimilarity* plots the mean similarity of previously aligned residues as a function of position in a multiple sequence file and using a residue comparison table, often based on mutational distance (Dayhoff *et al.*, 1979). Extensions incorporated within the program discussed here, *Similarity*, fall into four categories. (i) Scoring: the similarity score incorporates gap penalties and corrects for potential bias towards families of similar sequences. (ii) Display: notes showing functional sites or secondary structure can be aligned with the sequences, as can plots of sequence similarity, properties calculated from residue-type (such as hydrophobicity), or properties calculated from a known structure (such as surface accessibility). Plots can be numbered according to any of the constituent sequences. (iii) The sequence similarity may be displayed on the surface of a known structure using the program *Roadmap* (Chapman, 1993), or on a three-dimensional model using 'O' (Jones *et al.*, 1991). These displays help in the

Department of Chemistry and Institute of Molecular Biophysics, Florida State University, Tallahassee, FL 32306, USA

recognition of conserved regions comprised of several distinct regions of the primary structure. These interfaces to molecular graphics programs are useful if the sequences of interest are related to that of a known structure. Structures are known for ~200 of the estimated 1000 protein folds (Chothia, 1992). (iv) Statistical tests of significance are included.

The statistical methods are used to test hypotheses that residues associated with a particular function, region, structure, or sharing similar chemical properties (for example) have particularly high or low sequence similarity. By comparing two groups of residues from the same alignment, the difficulties of assessing sequence similarity on an absolute scale (Collins and Coulson, 1990) can be avoided. For example, comparisons within the same alignment will be much less sensitive to the overall sequence diversity and to the actual gap penalties used during alignment, because these factors will be the same for both groups of residues. Relative similarity scores can be calculated well within the low precision required for statistical tests. The statistical tests of hypotheses could be used, for example, to determine: (i) whether particular domains or secondary structures are more conserved than others within the molecule, (ii) whether particular surfaces are more or less hydrophobic than others or (iii) whether experimentally determined active site amino acids are conserved.

Picornaviruses were selected to test these methods, because their sequences have been examined previously (Rossmann and Palmenberg, 1988; Palmenberg, 1989) and because three-dimensional structures were known for several members: human rhinoviruses (Rossmann *et al.*, 1985; Kim *et al.*, 1989), polioviruses (Hogle *et al.*, 1985; Filman *et al.*, 1989), Mengo virus (Luo *et al.*, 1987), foot-and-mouth disease virus (Acharya *et al.*, 1989), and Theiler's virus (Luo *et al.*, 1992; Grant *et al.*, 1992). The extensive structural and experimental database could be used to cross-check conclusions reached from comparison of the sequences.

Systems and methods

The methods were developed on a cluster of Digital Equipment Corporation VAXstation 3100s running VMS v. 5.4, using VMS FORTRAN v. 5.6 and the subroutine library of the sequence analysis software (Genetics Computer Group Inc., 1991), v. 7.0, available from Genetics Computer Group Inc., 575 Science Drive, Madison, WI 53711.

Algorithms

Pair-wise comparisons

At each position, p, of the aligned sequences, a score, S_p, can be assigned for the match of the residue type (T) of sequence k to that of sequence l by using the appropriate element of a comparison table, C:

$$S_{p,k,l} = C(T_{p,k}, T_{p,l}) \quad (\text{where } k,l \neq \text{ gap}) \tag{1}$$

C is usually one of the matrices derived from the mutational distance matrix of Dayhoff *et al.* (1979). The choice of C is not as important as in the alignment of distantly related proteins (Argos *et al.*, 1991), because large numbers of pairwise comparisons are used in the statistical tests (below). Here, the popular normalized matrix of Gribskov and Burgess (1986) has been used, but alternatives such as the PAM 120 matrix (Altschul *et al.*, 1990) can also be used.

C may also be based on the physical and chemical properties of the amino acids, so that a difference in charge, hydrophobicity or size is scored unfavorably. Several matrices are provided with the program. For example, a hydrophobicity comparison matrix has been formulated using the absolute values of the differences in the free energies of transfer (Eisenberg and McLachlan, 1986) derived from the data of Fauchere and Pliska (1983). The purpose of these alternative matrices is not to enable a more sensitive recognition of distantly related sequences—as in Kubota *et al.* (1991) and Argos (1987)—but to determine whether all of the sequences share some local attribute, such as hydrophobicity, charge or type (aromatic, hydrophobic, etc.).

Where a residue from one sequence is aligned to an internal gap of the second, a penalty score is evaluated. The calculation of the penalty is a slight modification of that used in several methods of sequence alignment (e.g. Smith and Waterman, 1981) that is appropriate for a multiple sequence alignment. If L is the last position with a residue in both sequences, N is the next, and B is the number of locations within the gap where there is a residue in neither sequence, then the score for each residue aligned with a gap is:

$$S = -G_l - \frac{G_i}{(N-L-1-B)} \tag{2}$$

where G_i and G_l are gap initiation and lengthening penalties respectively. Values of $G_i = 3.0$ and $G_l = 0.15$, slightly lower than usually used in sequence alignment, have proved effective. Where the position, p, falls within gaps in both sequences, then the score from these sequences at this position is ignored. Gaps at the end of sequences are optionally either ignored or treated as internal gaps.

Calculation of similarity scores

From each of the pairwise comparisons at each position, a mean similarity, $\bar{S}_p$, can be calculated:

$$\bar{S}_p = \frac{2}{N(N-1)} \sum_{k=2}^{N} \sum_{l=1}^{k-1} S_{p,k,l} \tag{3}$$

which is the mean of all possible pairwise comparison of N sequences.

An average $\bar{s}_{k,l}$ is also computed for each pair of sequences:

$$\bar{s}_{k,l} = \sum_{p} S_{p,k,l} \tag{4}$$

using all positions along the aligned pair of sequences. An $N \times N$ matrix with 'distance' elements of $\bar{s}_{k,l}$ can be used to compare each sequence to every other. The relative distances are similar to those calculated by the program *Distances* (Devereux *et al.*, 1984; Genetics Computer Group Inc., 1991) except that the distances are calculated from gap-penalized mutational distances rather than simply the number of matches.

Grouping similar sequences

Among a set of aligned sequences that are related by a phylogenetic tree, some branches are likely to contain several closely related species for which sequences are known while other branches will contain fewer. Highly populated branches increase the overall similarity and dominate the position-dependent scores. This can be avoided if only a single representative is used from each branch, but much of the information regarding sequence variability would be lost. Alternatively, one of several schemes (Sibbald and Argos, 1990, and references therein) could be used to calculate weights for all the sequences. Here, a simpler intuitive approach is taken in which the average similarity between families is determined. If the sequences are grouped into m families, then the mean similarity, $\bar{S}_p$, at position p is calculated from $m(m-1)/2$ comparisons of the different families:

$$\bar{S}_p = \frac{2}{m(m-1)} \sum_{i=2}^{m} \sum_{j=1}^{i-1} \frac{1}{N_i N_j} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} S_{p,k,l} \qquad (5)$$

where N_i and N_j are the numbers of sequences in the ith and jth families and $S_{p,k,l}$ is the similarity between sequences k and l at the pth position, calculated from the (mutational distance) comparison table or from gap penalties (equation 1) as appropriate. The $1/(N_iN_j)$ weighting gives each pair of families equal weight, irrespective of the numbers of sequences in each family. All sequences are used, but large numbers of similar sequence in one family do not bias $\bar{S}_p$, because sequences are never compared with other members of the same family (i.e. $i \neq j$). In this formulation, $\bar{S}_p$ corresponds to a comparison of the profiles (Gribskov *et al.*, 1987) or the 'average sequences' of different families. If families are not defined, $\bar{S}_p$ is calculated using $n = 1$ and with m equal to the total number of sequences.

Smoothed similarity scores

Smoothed similarity scores draw attention to general areas of the linear sequence of high or low similarity. Like the program *Plotsimilarity* (Devereux *et al.*, 1984; Genetics Computer Group Inc., 1991) an average similarity score is calculated for a window of residues. However, in the current implementation, gap penalties are used and the window-averaged similarity score, $< \bar{S}_p >$, is calculated using a weight that decreases linearly from unity at the center of the window to zero at the ends:

$$\left\langle \bar{S}_p \right\rangle = \frac{2}{w} \sum_{h=p-w/2}^{p+w/2} \left(1 - \frac{2|p-h|}{w}\right) \bar{S}_h \qquad (6)$$

where w is the (full) width of the window and $\bar{S}_h$ is the mean gap-penalized similarity score at position h. This weighted average ensures that the smeared score is a smoothly varying function.

Statistical tests of significance

Through inspection of the similarity scores (see below) or for other *a priori* reasons, it may be suspected that the similarity (or physical/chemical properties) of different regions or classes of residues might differ. For testing, two groups of residues would be compiled. For example, group 1 might contain all active site residues, and group 2 contain all other buried residues. Alternatively, group 1 might contain all residues of one domain/subunit and group 2 all residues of another. For a transmembrane protein, group 1 might contain all surface residues on the inner surface and group 2 residues of the outer surface. In each case, for both groups, the mean is calculated of either the similarity score or some other parameter, such as hydrophobicity.

The significance of the difference between the mean similarities (hydrophobicities ...) of two groups of residues are tested in three ways: (i) the Z-test using the normal distribution (Spiegel, 1975); (ii) a modified Wilcoxonian test (Walpole and Myers, 1972); and (iii) through Monte Carlo simulations. All of these tests depend on the number of independent evaluations of the similarity score within each set of residues being compared. Here, the simple assumption is made that the scores from all residues within each set are independent measurements. This may not be true for the scores of neighboring residues. With only a small chance of a gap between a pair of neighboring residues, certainly the alignments at consecutive positions are highly linked. If, for example, the type of residue at one position depends on the type of residue at neighboring positions, then the similarity scores at neighboring positions are not completely independent, because following a good alignment at one position, there is a better than random chance of finding a good alignment at the neighboring position. Note that this effect is not caused by the similarity of the sequences themselves. Note also that this is a potential problem for sets of consecutive residues, but not for sets well distributed throughout the sequence. If the similarity scores are not as independent as assumed, then the variation in score will be underestimated and the significance of any difference exaggerated. Thus the statistical tests should be used as a 'necessary, but insufficient' criterion for establishing a significant difference.

Of the three tests, the Z-test is preferred, but only appropriate, when there are large numbers of residues in the smaller of the two groups. It is a standard statistical test that assumes that sample means drawn from the population are normally distributed. For sample sizes exceeding 30 (Walpole and Myers, 1972), as a consequence of the central limit theorem, this is

approximately true irrespective of the actual distribution of values (Hamilton, 1964). The distribution of similarity scores from aligned sequences is usually far from normal, with frequent exact matches and gap penalties leading to peaks in the distribution at the highest and lowest scores, respectively. The Wilcoxonian non-parametric test depends only on the rank of values when the two groups are combined, is independent of the distribution of their values, and is therefore more appropriate for small samples (Walpole and Myers, 1972). For groups of size n_1 and n_2 with mean values of μ_1 and μ_2, where w_1 is the sum of the ranks of group 1, ranked in ascending order for $\mu_1 \leq \mu_2$ (and vice versa), the probability of the hypothesis H_0, that $\mu_1 = \mu_2$ is given by Walpole and Myers, 1972):

$$\Pr(W_1 \leq w_1 | H_0 \text{ is true}) = \frac{n(W_1 \leq w_1)}{\binom{n_1+n_2}{n_1}} \quad (7)$$

where $n(W_1 \leq w_1)$ signifies the total number of different samples of size n_1 with sum of ranks not exceeding w_1 that could be drawn from a population of $n_1 + n_1$ values. In the current implementation, the right-hand side is evaluated *ab initio*, rather than from tabulated value. Thus, the test can still be used when one of the samples is large. The complexity of the computation rises quickly with sample size, due to the total number of combinations of ways that $n_1 + n_2$ can be split into groups of size n_1 and n_2:

$$\binom{n_1+n_2}{n_1} = {}_{n1+n2}C_{n1} = \frac{\prod_{i=1}^{n_1+n_2}(i)}{\prod_{j=1}^{n_1}(j)\prod_{k=1}^{n_2}(k)} \quad (8)$$

For example, if $n_1 = 10$ and $n_2 = 14$, there are nearly 2 million combinations. As demonstrated in Table I, a good approximation to the probability can be obtained using a subset of combinations, selected at random. In our implementation, this modification is invoked if a full combinatorial search would involve an excessive number of trials. The data of Table I and other tests (not shown) indicate that, with typical distributions of similarity scores, 1000–10 000 combinations are required to estimate the probability with an error of $< 1\%$. As the Wilcoxonian test is dependent only on the rank of the scores and not on their actual values, it is less dependent than the *Z*-test on the values arbitrarily assigned to gap penalties and is also, therefore, a useful check on the results of *Z*-tests for larger samples.

The Monte Carlo simulations similarly use a combinatorial search that is either full or sampled randomly. By redrawing samples of sizes n_1 and n_2 from the pooled groups, the frequency is calculated with which a difference in means exceeds that of the original two groups. This crude, rank-independent Monte Carlo test is used primarily to check the stability of the Wilcoxonian test in case the ranking is hypersensitive to small changes of similarity score. In fact, all three tests usually agree well.

Table I. Approximations to the Wilcoxonian test

Combinations tested	Estimate of $\Pr(W_1 \leq w_1 \mid H_0 \text{ is true})$
10	0.500
32	0.156
100	0.110
316	0.139
1000	0.162
3162	0.164
10 000	0.165
31 662	0.168
100 000	0.169
All 1 961 256	0.170

Samples of $n_1 = 10$ and $n_2 = 14$ residues were drawn from an alignment of 10 human rhinovirus (HRV) sequences (Palmenberg, 1989, and references therein). The first group contained those residues that interact with the 'WIN' antiviral agents in both HVR14 and HRV1A, and the second contained those that interact in the complexes of only one of the serotypes (Kim *et al.*, 1993). Although those residues interacting in both serotypes are more conserved than those interacting in only one, the Wilcoxonian test showed that the difference was not significant. The table shows the estimate of the Wilcoxonian statistic (Walpole and Myers, 1972), calculated either systematically from all possible combinations, or from smaller numbers of randomly drawn combinations. Convergence upon the actual value is smooth and indicates that 1000–10 000 combinations are required.

For all tests, the groups are defined using ranges of numbers from a reference sequence (which can be that of a known structure). Within a range, positions without a residue in the reference sequence (insertions) usually need to be included. Half of any insertion immediately before or after a range is also included so that there is not an inherent bias linking the average gap penalty assessed to the length of the range. This is especially important if one is comparing a number of short segments to all other residues.

The tests may depend on the gap penalties whose values are somewhat arbitrary. To check for such a dependence, a test can be rerun, excluding all residues where there is a gap in one of the aligned sequences. When the tests agree, it can be concluded that the difference in similarity is not an artefact of the gap penalties. Disagreement might have several causes:

- With differing numbers of gaps and if excessive penalties have been used, an artificial difference in similarity scores may have been created.
- There may be a real difference in the distribution of gaps that might be biologically relevant.
- The residues of highest sequence variability are likely also to be the sites of gaps. By excluding 'gapped' residues, particularly if the alignment contains many sequences, only a small number of the most conserved residues of each group may remain.

If a difference in similarity is only significant when 'gapped'

residues are included, the first of these possibilities must be eliminated for the difference to have any biological importance. It can be eliminated if the difference remains significant when the test is repeated, including gaps, but with penalties to values that could not be considered excessive. Penalties of zero are a very conservative choice when using the Dayhoff mutational distance matrix normalized by Gribskov and Burgess (1986), which has a mean matching score of zero.

Implementation

Alignment graphics

To facilitate the identification and interpretation of regions of the sequence with significantly high or low similarity, sophisticated graphical output can collate sequence and structural information in a way that incorporates and extends greatly the functions of the GCG programs *Pretty* and *Plotsimilarity* (Devereux *et al.*, 1984; Genetics Computer Group Inc., 1991). The following information can be presented, aligned one above another, according to the position within the aligned sequences.

1. Sequences, including a consensus sequence, numbered at the beginning and end of each line.
2. Sequence numbers for every 10th residue of a reference sequence of choice.
3. Labels for individual residues (surface residues or the sites of mutations, for example), or ranges of residues (secondary structural elements, for example). As the positions of these annotations are defined with reference to a sequence of choice, they are automatically updated if the sequence alignment is changed.
4. Arbitrary functions that vary according to the sequence position may be plotted. These will usually include the individual residue similarity scores and smeared similarity scores. These scores can be overlayed upon horizontal lines of standard error about the mean of the scores (Spiegel, 1975) to give an approximate guide to the significance of individual scores. Other externally calculated functions may also be plotted. Some functions might be calculated from a sequence, such as the propensity to form secondary structures (Chou and Fasman, 1978), or the hydrophobic moment (Eisenberg *et al.*, 1984). Others, such as solvent accessibility (Richards, 1985) might be calculated from a known structure. Graphs may be superimposed with different line types and ordinate scales or stacked one above another, all aligned with the sequences.
5. Numerical values for the similarity scores, tabulated to align with the sequences.

The contents and ordering of the display are flexible as are the page layout, magnification and fonts. Each sequence is split up into a number of panels that is dependent on the page or screen and font sizes.

The graphics output is viewable on a great variety of devices including several graphics terminals, with and without Tektronix 4014 emulation, personal computers, workstations and several printers including those with PostScript® (Adobe Systems Inc., 1990) support.

Correlation with structure

Especially conserved or variable regions formed by the juxtaposition of different parts of the linear sequence or at the interfaces between subunits become obvious when the similarity scores are mapped to the three-dimensional structure of one of the sequences, if known. Residues may be colored according to sequence similarity to display the molecular surface using the program *Roadmap* (Chapman, 1993 and example therein) or the full three-dimensional structure using 'O' (Jones *et al.*, 1991). Residues of the reference structure that are deleted in other sequences have similarity scores that are lowered by gap penalties. Residues of sequences that form insertions relative to the reference structure cannot, by definition, be viewed directly. To highlight the locations of such insertions, the scores of the residues of the reference structure that are immediately before and after an insertion are averaged with the lower gap-penalized scores of the inserted residues.

Program input

The program can be run in a batch queue or interactively on any ASCII terminal. Program control is through hierarchical character-based menus and free-format input. Control may be redirected to nested input files to run frequently used option sequences. Defaults are used for most parameters unless they are changed through command-line control or through the running of specific options.

Relative to the length of one of the constituent sequences, the length of a multiple sequence alignment is extended by insertions in other sequences. The position of a particular residue within the alignment depends on the locations and sizes of preceding gaps. Consequently, the position depends on which other sequences are contained in the alignment and how they have been aligned. In the current implementation, look-up tables are maintained so that input and output residue positions can be specified according to a reference sequence of choice. This facilitates the annotation of important residues and secondary structure and the definition of groups of residues for hypothesis testing. Furthermore, residue-specifying commands do not need to be updated when the alignment is changed.

Discussion

As an example, the sequence similarity for a part of the N-terminal region of viral protein (VP)1 of picornaviruses is displayed in Figure 1. The sequence similarity was calculated from 51 sequences aligned by Palmenberg (1989), grouped into 11 families of 2–13 sequences with each family corresponding to a sub-genera as proposed by Palmenberg (1989). Gap initiation and lengthening parameters of 3.0 and 0.15 respec-

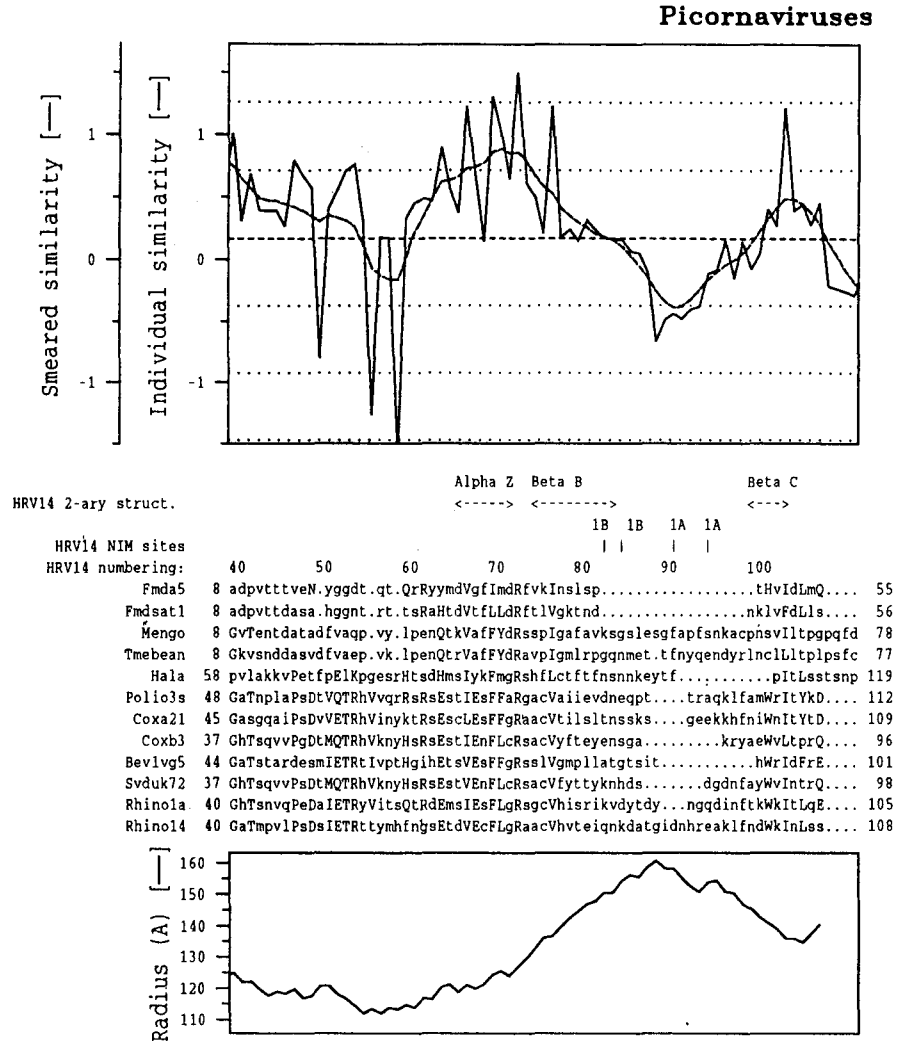

Fig. 1. Picornaviral sequence similarity. A short fragment of the N-terminal region of VP1 of the picornaviruses is shown. Describing the figure from top to bottom: (i) the solid line shows the similarity scores, calculated for each residue, using gap initiation and lengthening penalities of 3.0 and 0.15, and a normalized Dayhoff mutational distance matrix (Dayhoff *et al.*, 1979). Fifty-one sequences were used, as aligned and grouped into 11 sub-genera by Palmenberg (1989). The slowly undulating dashed line shows the scores, smeared over a window of 10 residues. Horizontal lines are drawn at the mean similarity (dashed) for all VP1 residues and at unit multiples of their standard deviation (dotted). (ii) Labels show the secondary structure of HRV14 (top; Arnold and Rossmann, 1990) and the locations of mutations that escape neutralizing immunogenic responses (bottom; Rossman *et al.*, 1985; Sherry *et al.*, 1986). (iii) Every 10th residue of the reference sequence (HRV14) is numbered with the first digit aligned above the relevant residue. (iv) For brevity, most of the 51 sequences used in the calculation of the similarity scores were deactivated, leaving only a representative subset to be shown in the figure. The sequence identifiers, left of each sequence, and the numbers are taken from the input sequence alignment. (v) The graph shows the distance of each residue from the center of the virus, calculated from the atomic coordinates of HRV14 (Arnold and Rossmann, 1990).

tively were used. Through the labeling of the secondary structural elements of human rhinovirus (HRV)14 (Arnold and Rossmann, 1990) on a display of the similarity of VP1 of picornaviruses (part of which is shown in Figure 1), it appeared that many of the secondary structural elements were more conserved than the loops between them. The similarity scores of these residues were compared to all other picornaviral residues in *Z*-tests of significance (Table II). They were very significantly more conserved than other residues (99.8% level of significance) when gap penalities were used, but the difference was not significant (to the 90% level) when all positions containing a gap in any sequence were excluded. This disparity indicated that the significance test should be checked by re-running it with low gap penalities, in this case zero. These confirmed that the difference was significant. When these tests were repeated using only rhinoviral sequences (Table II), the significance tests were not dependent upon the treatment of gaps. This suggested that the disparity in the picornaviral tests might be due to the high proportion (60%) of residues, of presumably relatively low conservation, that were rejected through the exclusion of all the positions containing gaps. The high conservation of the secondary structure of picornaviruses that forms

the core of the capsid protein might be the result of selective pressure to maintain structural elements, or it might be because buried residues are more likely that surface residues to encounter unfavorable interactions upon mutation. This possibility is easier to test with parvoviruses than picornaviruses, because with a capsid protein that is twice as large, there are more residues that are buried, but not part of the secondary structure. Significance tests, using the aligned sequences of parvoviruses (Chapman and Rossmann, 1993) (Table II),

Table II. Sequence conservation of the β-barrel fold

Virus	Residues	No.	Gaps	<Similarity>	σ	Confidence (%)
Picorna	'virus' motif	111	+	0.44	0.60	
	other residues	213		0.08	0.56	≥99.8
Picorna	'virus' motif	82	−	0.65	0.36	
	other residues	86		0.60	0.40	≤90.0
Picorna	'virus' motif	111	0	0.59	0.36	
	other residues	204		0.49	0.42	≥99.0
Rhino	'virus' motif	106	+	1.24	0.32	
	other residues	186		0.94	0.61	≥99.8
Rhino	'virus' motif	106	−	1.24	0.32	
	other residues	163		1.08	0.40	≥99.8
Rhino	'virus' motif	106	0	1.24	0.32	
	other residues	173		1.07	0.41	≥99.8
Parvo	'virus' motif	65	+	0.75	0.32	
	buried residues	694		0.22	0.64	≥99.8
Parvo	'virus' motif	64	−	0.76	0.32	
	buried residues	224		0.66	0.33	≥95.0
Parvo	'virus' motif	65	0	0.77	0.36	
	buried residues	560		0.60	0.41	≥99.8

The 'virus' motif contains all residues that are aligned to secondary structural elements of HRV14 (Arnold and Rossmann, 1990) that are found in picorna-like viral capsid structures (Harrison, 1990). Significance tests using 51 picornaviral sequences and 10 rhinoviral sequences were calculated using the alignment of Palmenberg (1989) and as described in the text. They compared the viral capsid motif to all other residues. The tests on parvoviruses were calculated from 11 sequences, aligned by Chapman and Rossmann (1993), and compared the corresponding elements of the CPV structure (Tsao *et al.*, 1991) to only those other residues that are buried. Comparisons with '−' in the 'Gaps' column excluded all positions with one or more gaps in any of the sequences, those with '+' were calculated with gap initiation and lengthening penalties of 3.0 and 0.15 respectively, and those with '0' were calculated with gap penalties of zero. For each group is given the mean similarity, its standard deviation and the number of residues (which, including gaps, can exceed the number of residues in any one sequence). For each comparison is given the confidence with which the means can be thought to differ according to one-tailed *Z*-tests, appropriate for such large numbers of residues.

yielded results similar to the tests for picornaviruses, indicating that the structural motif is conserved relative to other residues that are buried.

Experience with many tests, such as those shown in Tables II and III, suggests that 95% confidence limits are usually appropriate. This is less stringent than the 3σ criterion suggested by Feng *et al.* (1985) for determining the significance of alignment scores. Feng *et al.* (1985) tested whether the alignment score was significantly above that expected for randomized sequences. Here, to avoid the problems of determining scores on an absolute scale (Collins and Coulson, 1990), the alignment score of one set of residues is compared to a scored 'control' set from the same alignment. Clearly the score from a control set of aligned residues will exceed that from random sequences, so the differences calculated here will be smaller than those calculated by Feng *et al.* (1985).

Smeared sequence similarity, distance from the center of the virus, and the locations of mutations that escape *n*eutralizing *im*munogenic (NIm) response are superposed in Figure 1. The NIm sites are within external hypervariable loops to which antibodies bind (Rossmann *et al.*, 1985; Sherry *et al.*, 1986; Smith *et al.*, 1993). The variability of one of these loops is evident in Figure 1. The results of tests of significance (Table III) confirm the variability of rhinoviral NIm loops. The failure of the test when 'gapped' residues were excluded may be ignored, because this included only one-quarter of the residues, and because the test calculated with no penalties confirmed the initial gap-penalized test.

Only 15 residues remain when 'gapped' residues are excluded from NIm loops. When their similarity was compared to that of other external loops, the Wilcoxonian test was used (Table III). The one-tailed probabilities that the samples are drawn from different populations, calculated using the *Z*, Wilcoxonian and Monte Carlo tests are in excellent agreement: 0.64, 0.67 and 0.65 respectively. Good agreement is usually obtained when there are close to 30 residues in the smaller group, but the agreement is sometimes worse with smaller samples or with a large proportion of scores that are gap-penalized. For example, for the test used in Table I, with 10 and 14 residues in each group, the *Z* and Wilcoxonian tests estimate the one-tailed probabilities as 0.94 and 0.83 respectively.

The structures of polioviruses and rhinovirus 1A (Kim

Table III. Sequence variability of NIm loops

Virus	Residues	Number	Gaps	< Similarity >	σ	Confidence (%)
Rhino	NIm loops	108	+	0.68	0.82	
	other exterior loops	52		0.94	0.48	≥99.5
Rhino	NIm loops	33	−	0.91	0.41	
	other exterior loops	15		0.96	0.47	≤90.0
Rhino	NIm loops	108	0	0.84	0.49	
	other exterior loops	52		097	0.44	≥95.0

Residues of NIm loops are compared with those of all other loops on the outside surface of rhinoviruses. Significance tests were calculated as described for Table II, except that the Wilcoxonian test was used when one of the groups contained <30 residues.

et al., 1989, 1993; Filman *et al.*, 1989; Yeates *et al.*, 1991) have revealed a lipid or fatty acid bound in the same pocket to which antipicornaviral agents bind (Smith *et al.*, 1986), suggesting that the antiviral agents mimic natural cofactors. It might be expected that the residues of the cofactor binding site might be conserved. Unexpectedly, the mean similarity score for the 24 residues that interact with antiviral agents or cofactor in HRV14 or HRV1A (Kim *et al.*, 1993) is 1.25; this is lower than the 1.29 found for all other buried residues, but the Wilcoxonian test shows that this difference is insignificant. In retrospect, perhaps the lack of conservation of the pocket is consistent with the finding of different cofactors in different picornaviruses.

Care should be taken to avoid misleading statistical tests. If, through manual alignment, or through judicious setting of gap penalties in automatic alignment, the alignment is forced to match well in particular regions, then these regions may show high similarity even with completely unrelated sequences. If particular residues are of interest (e.g. residues thought to be part of an active site), then the initial alignment should not be constrained to yield a good alignment to these residues. Only when it has been verified that these residues are highly conserved should the alignment be repeated, adding constraints as appropriate.

Ideally, the residues of interest should be compared with those that are in a similar environment. For example, the residues of the drug-binding pocket of rhinoviruses appears highly conserved when compared to all other residues, but not when compared only to buried residues. As it is a hydrophobic pocket mostly inside the β-barrel of the capsid protein, it is inappropriate to compare its residues with those of the surface, which are inherently more variable.

Statistics can, of course, be misleading. As discussed earlier, interdependence of the scores of neighboring residues may exaggerate the difference between sets containing consecutive residues. Even in this case, a 'necessary, but insufficient' test of significance has advantages over a subjective analysis. A 95% confidence limit is commonly used, so that even with randomly distributed sequence similarity, statistical significance is expected for one in 20 tests. False positives will occur less frequently if the tests are used sparingly, and when the selection of residues is constrained tightly by the null hypothesis. Even with high statistical significance, the biological cause may be ambiguous (Karlin and Brendel, 1992). Some of the residues under consideration may be members of another group that has different sequence similarity for an unrelated reason.

Although care is necessary, statistical methods should be superior to purely empirical analyses of sequence similarity, because (i) they suggest a limit as to how much may reasonably be concluded from a sequence alignment; (ii) they may reveal relations not obvious to the eye; and (iii) they provide a measure of the relative robustness of different proposed correlations.

In addition to the above examples, the methods have also been applied to parvoviral sequences, to study potential receptor binding sites and antigenic determinants (Chapman and Rossmann, 1993a). They have also been used to study the sequence similarities of picornaviral surfaces with the receptor footprint of the major receptor group rhinoviruses (Olson *et al.*, 1993; Chapman and Rossmann, 1993b). The methods are not restricted to the study of viral sequences, and it is expected that they will find wide application in many fields of study.

Acknowledgements

I am grateful to Ann Palmenberg for sending her alignment of picornaviral sequences with which the program was tested, and from which the examples were drawn. The methods were developed using computers of the AIDS Center Laboratory for Computational Biochemistry at Purdue University (NIH AI27713). The program, *Similarity*, will be distributed under license from Purdue Research Foundation to licensees of the GCG software at no additional cost for the purpose of non-commercial research. Those interested should contact the author. I gratefully acknowledge the help and support of Michael G.Rossmann in all aspects of this work, and in whose laboratory this work was performed, with the assistance of grants from the National Science Foundation, National Institutes of Health and the Medical Research Council.

References

Acharya,R., Fry,E., Stuart,D., Fox,G., Rowlands,D. and Brown,F. (1989) The three-dimensional structure of foot-and-mouth disease virus at 2.9 Å resolution. *Nature*, **337**, 709–716.

Adobe Systems Inc. (1990) *PostScript Language Reference Manual*, 2nd edn. Addison-Wesley, Reading, MA.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Argos,P. (1987) A sensitive procedure to compare amino acid sequences. *J. Mol. Biol.*, **193**, 385–396.

Argos,P., Vingron,M. and Vogt,G. (1991) Protein sequence comparison: methods and significance. *Prot. Engng*, **4**, 375–383.

Arnold,E. and Rossmann,M.G. (1990) Analysis of the structure of a common cold virus, human rhinovirus 14, refined at a resolution of 3.0 Å. *J. Mol. Biol.*, **211**, 763–801.

Bacon,D.J. and Anderson,W.F. (1990) Multiple sequence comparison. *Methods Enzymol.*, **183**, 438–447.

Barton,G.J. (1990) Protein multiple sequence alignment and flexible pattern matching. *Methods Enzymol.*, **183**, 403–428.

Bowie,J.U., Lüthy,R. and Eisenberg,D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Chapman,M.S. (1993) Mapping the surface properties of macromolecules. *Prot. Sci.*, **2**, 459–469.

Chapman,M.S. and Rossmann,M.G. (1993a) Structure, sequence and function correlations among parvoviruses. *Virology*, **194**, 491–508.

Chapman,M.S. and Rossmann,M.G. (1993b) Comparison of surface properties of picornaviruses: strategies for hiding the receptor from immune surveillance. *Virology*, **195**, 745–764.

Chothia,C. (1992) One thousand families for the molecular biologist. *Nature*, **357**, 543–544.

Chou,P.Y. and Fasman,G.D. (1978) Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.*, **47**, 45–147.

Collins,J.F. and Coulson,A.F.W. (1990) Significance of protein sequence similarities. *Methods Enzymol.*, **183**, 474–487.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1979) A model of evolutionary change in proteins. Detecting distant relationships: computer methods and results. In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, Vol. 5, Suppl. 3, pp. 353–358.

Devereux,J., Haerberli,P. and Smithies,O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.*, **12**, 387–395.

Eisenberg,D. and McLachlan,A.D. (1986) Solvation energy in protein folding and binding. *Nature*, **319**, 199–203.

Eisenberg,D., Weiss,R.M. and Terwilliger,T.C. (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA*, **81**, 140–144.

Fauchere,J.-L. and Pliska,V. (1983) Hydrophobic parameters π of amino-acid side chains from the partitioning of *N*-acetyl-amino-acid amides. *Eur. J. Med. Chem.-Chim. Ther.*, **18**, 369–375.

Feng,D.F., Johnson,M.S. and Doolittle,R.F. (1985) Aligning amino acid sequences: comparison of commonly used methods. *J. Mol. Evol.*, **21**, 112–125.

Filman,D.J., Syed,R., Chow,M., Macadam,A.J., Minor,P.D. and Hogle,J.M. (1989) Structural factors that control conformational transitions and serotype specificity in type 3 poliovirus. *EMBO J.*, **8**, 1567–1579.

Genetics Computer Group Inc. (1991) Sequence Analysis Software Package. Madison, WI.

George,D.G., Barker,W.C. and Hunt,L.T. (1986) The protein identification resource (PIR). *Nucleic Acids Res.*, **14**, 11–15.

Grant,R.A., Filman,D.J., Fujinami,R.S., Icenogle,J.P. and Hogle,J.M. (1992) Three-dimensional structure of Theiler virus. *Proc. Natl. Acad. Sci. USA*, **89**, 2061–2065.

Gribskov,M. and Burgess,R.R. (1986) Sigma factors from *E.coli*, *B.subtilis*, phage SP01, and phage T4 are homologous proteins. *Nucleic Acids Res.*, **14**, 6745–6763.

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA*, **84**, 4355–4358.

Gribskov,M., Lüthy,R. and Eisenberg,D.S. (1990) Profile analysis. *Methods Enzymol.*, **183**, 146–159.

Hamilton,W.C. (1964) *Statistics in Physical Science*. Ronald Press, New York.

Harrison,S.C. (1990) Principles of virus structure. In Fields,B.N. and Knipe,D.M. (eds), *Virology*, Raven Press, New York, Chap. 3, pp. 37–61.

Hogle,J.M., Chow,M. and Filman,D.J. (1985) Three-dimensional structure of poliovirus at 2.9 Å resolution. *Science*, **229**, 1358–1365.

Jones,T.A., Zou,J.-Y., Cowan,S.W. and Kjeldgaard,M. (1991) Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr.*, **A47**, 110–119.

Karlin,S. and Brendel,V. (1992) Chance and statistical significance in protein and DNA sequence analysis. *Science*, **257**, 39–49.

Kim,S., Smith,T.J., Chapman,M.S., Rossmann,M.G., Pevear,D.C., Dutko,F.J., Felock,P.J., Diana,G.D. and McKinlay,M.A. (1989) The crystal structure of human rhinovirus serotype 1A (HRV1A). *J. Mol. Biol.*, **210**, 91–111.

Kim,K.H., Willingmann,P., Gong,Z.X., Kremer,M.J., Chapman,M.S., Minor,I., Oliveira,M.A., Rossmann,M.G., Andries,K., Diana,G.D., Dutko,F.J., McKinlay,M.A. and Pevear,D.C. (1993) A comparison of the anti-rhinoviral drug binding pocket in HRV14 and HRV1A. *J. Mol. Biol.*, **230**, 206–227.

Kubota,Y., Nishikawa,K., Takahashi,S. and Ooi,T. (1982) Correspondence of homologies in amino acid sequence and tertiary sequence of protein molecules. *Biochim. Biophys. Acta*, **701**, 242–252.

Luo,M., Vriend,G., Kamer,G., Minor,I., Arnold,E., Rossmann,M.G., Boege,U., Scraba,D.G., Duke,G.M. and Palmenberg,A.C. (1987) The atomic structure of Mengo virus at 3.0 Å resolution. *Science*, **235**, 182–191.

Luo,M., He,C., Toth,K.S., Zhang,C.X. and Lipton,H.L. (1992) Three-dimensional structure of: Theiler murine encephalomyelitis virus (BeAn strain). *Proc. Natl. Acad. Sci. USA*, **89**, 2409–2413.

Olson,N.H., Kolatkar,P.R., Oliveira,M.A., Cheng,R.H., Greve,J.M., McClelland,A., Baker,T.S. and Rossmann,M.G. (1993) Structure of a human rhinovirus complexed with its receptor molecule. *Proc. Natl. Acad. Sci. USA*, **90**, 507–511.

Palmenberg,A.C. (1989) Sequence alignments of picornaviral capsid proteins. In Semler,B.L. and Ehrenfeld,E. (eds), *Molecular Aspects of Picornavirus Infection and Detection*. American Society for Microbiology, Washington, DC, pp. 211–241.

Richards,F.M. (1985) Calculation of molecular volumes and areas for structures of known geometry. *Methods Enzymol.*, **115**, 440–464.

Rossmann,M.G. and Palmenberg,A.C. (1988) Conservation of the putative receptor attachment site in picornaviruses. *Virology*, **164**, 373–383.

Rossmann,M.G., Arnold,E., Erickson,J.W., Frankenberger,E.A., Griffith,J.P., Hecht,H.J., Johnson,J.E., Kamer,G., Luo,M., Mosser,A.G., Rueckert,R.R., Sherry,B. and Vriend,G. (1985) Structure of human common cold virus and functional relationship to other picornaviruses. *Nature*, **317**, 145–153.

Schuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) A workbench for multiple alignment construction and analysis. *Proteins*, **9**, 180–190.

Sherry,B., Mosser,A.G., Colonno,R.J. and Rueckert,R.R. (1986) Use of monoclonal antibodies to identify four neutralization immunogens on a common cold picornavirus, human rhinovirus 14. *J. Virol.*, **57**, 246–257.

Sibbald,P.R. and Argos,P. (1990) Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J. Mol. Biol.*, **216**, 813–818.

Smith,T.F. and Waterman,M.S. (1981) Comparison of biosequences. *Adv. Appl. Math.*, **2**, 482–489.

Smith,T.J., Kremer,M.J., Luo,M., Vriend,G., Arnold,E., Kamer,G., Rossmann,M.G., McKinlay,M.A., Diana,G.D. and Otto,M.J. (1986) The site of attachment in human rhinovirus 14 for antiviral agents that inhibit uncoating. *Science*, **233**, 1286–1293.

Smith,T.J., Olson,N., Chase,E., Cheng,R.H., Liu,H., Lee,W.-M., Mosser,A., Rueckert,R. and Baker,T.S. (1993) Structure of human rhinovirus complexed with Fab fragments from a neutralizing antibody. *J. Virol.*, **67**, 1148–1158.

Spiegel,M.R. (1975) *Probability and Statistics*. McGraw-Hill, New York.

Taylor,W.R. (1990) Hierarchical method to align large numbers of biological sequences. *Methods Enzymol.*, **183**, 456–473.

Tsao,J., Chapman,M.S., Agbandje,M., Keller,W., Smith,K., Wu,H., Luo,M., Smith,T.J., Rossmann,M.G., Compans,R.W. and Parrish,C.R. (1991) The three-dimensional structure of canine parvovirus and its functional implications. *Science*, **251**, 1456–1464.

Vihinen,M. (1990) Simultaneous comparison of several sequences. *Methods Enzymol.*, **183**, 447–455.

Vingron,M. and Argos,P. (1989) A fast and sensitive multiple sequence alignment algorithm. *Comput. Applic. Biosci.*, **5**, 115–121.

Walpole,R.E. and Myers,R.H. (1972) *Probability and Statistics for Engineers and Scientists*. Macmillan, New York.

Yeates,T.O., Jacobson,D.H., Martin,A., Wychowski,C., Girard,M., Filman,D.J. and Hogle,J.M. (1991) Three-dimensional structure of a mouse-adapted type 2/type 1 poliovirus chimera. *EMBO J.*, **10**, 2331–2341.

Received April 1, 1993; accepted September 9, 1993