*Bulletin of Mathematical Biology* Vol. 51, No. 1, pp. 39–54, 1989.
Printed in Great Britain.

0092-8240/89$3.00+0.00
Pergamon Press plc


# ALGORITHMS FOR THE OPTIMAL IDENTIFICATION OF SEGMENT NEIGHBORHOODS

■ IVAN E. AUGER and CHARLES E. LAWRENCE
Laboratory of Biometrics,
Wadsworth Center for Laboratories and Research,
New York State Department of Health,
Albany, NY 12201, U.S.A.

Two algorithms for the efficient identification of segment neighborhoods are presented. A segment neighborhood is a set of contiguous residues that share common features. Two procedures are developed to efficiently find estimates for the parameters of the model that describe these features and for the residues that define the boundaries of each segment neighborhood. The algorithms can accept nearly any model of segment neighborhood, and can be applied with a broad class of best fit functions including least squares and maximum likelihood. The algorithms successively identify the most important features of the sequence. The application of one of these methods to the haemagglutinin protein of influenza virus reveals a possible mechanism for conformational change through the finding of a break in a strong heptad repeat structure.

*1. Introduction.* Many structural and functional features of proteins occur in segments of neighboring residues. Included in the many examples of these features are the following: polypeptide chains can be divided into segments which belong to distinct structural domains (Wetlaufer, 1972); functional domains tend to fold independently of each other (Sternberg and Thornton, 1977); beta strands and alpha helices represent segments of contiguous residues sharing a common structure; consecutive strands of beta sheets tend to be adjacent (Sternberg and Thornton, 1977). Furthermore, there is a kinetic advantage for the folding of proteins into segments of neighboring residues which stems from the associated high neighborhood correlation (Wetlaufer, 1972; Flory, 1956; Schultz and Schirmer, 1979).

The primary set of existing methods aimed at the identification of such segment neighborhoods are based on moving windows methods. Kyte and Doolittle (1982) developed a moving average method and a hydropathic index to aid in the identification of segments that span membranes or occupy the interior of globular proteins. Similarly, Hopps and Wood (1981) use a similar moving average method to detect antigenic sites. More recently, weighted moving averages have also been employed. DeLisi and Berzofsky (1985) have employed a moving window method to calculate the power spectrum of hydropathicity for the identification of amphipathic structures. Foulser and Karlin (1987) present methods for the identification of patterns in sequences

and assessing the significance of these patterns under null models of independent identically distributed (i.i.d.), and Markov dependence. They treat amino acid characteristics by employing multiple alphabets (Karlin and Ghandour, 1985). Their methods differ from those proposed here in two important ways. Their alphabets treat amino acid characteristics as unordered categorical data, while the methods presented here and moving window methods assign 20 ordered values for amino acid characteristics. Also, they treat errors as blocks of mismatches while we treat errors as a difference between predicted and observed values.

The use of fixed length overlapping windows in moving window methods brings to these methods a number of limitations. The requirement of the selection of a single window width is the most obvious. Since structural features occur in a wide variety of lengths no single width is best for an entire sequence. Consequently, the ability of these methods to detect features that occur in segments not compatible with the selected window size suffers. For example, von Heijne (1986) finds that the fixed window of size 11 of Eisenberg's hydrophobic moment method cannot find the segment of maximal total hydrophobic moment. He finds that the optimal window length varies between 17 and 26 for a sample of surface seeking peptides, and between 12 and 26 residues for a mitochondrial sample. Because existing methods provide no mechanism to deal with these varying window widths he chooses a window size of 18 residues for the analysis of mitochondrial sequences.

Proteins are complex molecules that exhibit different features in different segments of the primary structure of a protein. For example, some segments may have an amphipathic helical structure while others may be strongly hydrophobic. Because of the overlapping nature of moving window methods, it is not possible in a single analysis to identify multiple features in a single sequence. One's only recourse is to run multiple analyses and subsequently pick out regions of interest from the various plots. Since moving window methods provide no selection criteria, this selection process becomes at best an intuitive exploration. If the great complexity of protein structure is ever to be unraveled, it is clear that we have to go beyond this intuitive process.

To redress these limitations, we present two algorithms which allow for the simultaneous identification of several different features in different segments of the primary structure of a protein regardless of their location and length. Furthermore, the algorithms allow for the modeling of multivariate characteristics for each residue. For example, the algorithms are sufficiently general to allow for the simultaneous analysis of both charge characteristic and hydropathic characteristic of each residue. We begin by defining a segment neighborhood as a set of contiguous residues whose members share common features. Both algorithms have a dynamic programming (D.P.) step. Dynamic programming has been widely applied in sequence alignment algorithms for

biopolymers (Waterman, 1984; Sankoff and Kruskal, 1983). The first algorithm developed here is more closely related to algorithms used in the solution of the "shortest path" problem in operations research than it is to the above mentioned sequence alignment algorithms.

The first algorithm has two steps. First, for every possible segment candidate mathematical models are selected. The parameters of these models are estimated by best fit procedures. Simultaneously, the fit of each model to the data for each segment is assessed by statistical goodness of fit measures. We assign to a segment the model that best fits the data. Then, a dynamic programming procedure is employed to piece together a set of contiguous segment neighborhoods that provide the best fit to the total sequence.

This problem belongs to the class of segmented or piecewise regression problems. Bellman and Roth (1966) developed a D.P. procedure to fit a curve by segmented straight lines. The same fundamental D.P. step is used later by Hawkins (1976) who presents the algorithm in a statistical framework. The latter fits piecewise multiple regressions. Hawkins method yields maximum likelihood estimates, whereas Bellman and Roth's do not due to the fact that they use the maximum of the absolute value between a straight line and the curve as their goodness of fit measure instead of sum of squared errors. Hawkins also looks into which measure to use when the variances of the errors for each segmented regression are heteroscedastic. Lerman (1980) develops a grid search method to fit segmented regressions where the values of the fitted regressions have to be the same at the transition points. Segmented regression methods have been applied to many fields, but to our knowledge this is the first application to sequence analysis in molecular biology. As such, there are several requirements specific to this field. Because of the existence of turns in protein structures, there is no need to require continuity at the segment boundaries as has been the case in many other applications. Also, because proteins have several different structural features which may all be present in different segments of the same protein, we found it necessary to extend previous work to allow for the existence of multiple models in a single analysis. We have further extended the range of objective functions to include models other than regression models as illustrated in a previous application by Lawrence and Reilly (1985). In the latter, Markov chain mutation probabilities derived by Dayhoff *et al.* (1978) are used to estimate the conservation of a segment in two or more related protein. This estimation is done via maximum likelihood. Then the sequence is partitioned into segments with the objective of maximizing the total likelihood.

The second algorithm solves a related problem. It finds the $1, 2, \ldots, P$ non-overlapping segments over which the sum of a statistic of interest is minimized. These segments do not necessarily span the whole sequence. Bement and Waterman (1977) developed an algorithm that maximizes the sum of the variance of segments in an analysis of geologic data. We employ this algorithm

to identify segments in proteins which best fit a proposed model. The first step is the same as in the first algorithm except that we need a length independent measure of fit. Then a D.P. procedure is used to get the $1, 2, \ldots, P$ non-overlapping segment neighborhoods that minimize the sum of the measures of fit.

An application of these algorithms to influenza virus haemagglutinin is given.

*2. Description of Algorithms.* In this section we present the algorithms, and discuss some of the computational and statistical issues.

*2.1. Optimal segmentation algorithm.* Segment neighborhoods are identified by optimally partitioning a sequence into $Q$ contiguous segments based on the fit of the model to the data as follows: let $Y_i$ = vector of observed characteristics of the $i^{\text{th}}$ residue (data); $\theta_q$ = vector of unknown parameters of the $q^{\text{th}}$ segment neighborhood; $r_q$ = unknown index of last residue of the $q^{\text{th}}$ segment neighborhood (segment neighborhood boundaries); $F(Y_i \ldots Y_j, \theta_q)$ = model of the relationship between $Y_i \ldots Y_j$ and $\theta_q$ (model).

To find estimators for the segment neighborhood parameters $\theta_1, \ldots, \theta_Q$ and the segment neighborhood boundaries $r_1, \ldots, r_{Q-1}$ we employ methods of best fit as follows:

$$Z(\mathbf{Y}, \boldsymbol{\theta}, \mathbf{r}, Q) = \min_{\boldsymbol{\theta}, \mathbf{r}} \sum_{q=1}^{Q} C(F(Y_{r_{q-1}+1} \ldots Y_{r_q}, \theta_q)),$$

where $C$ = function that measures the fit of the model $F(\ )$ to the data $Y_{r_{q-1}+1} \ldots Y_{r_q}$ with the estimated parameters $\theta_1, \ldots, \theta_Q$ (objective function) and $(r_0, r_Q) = (0, n)$, where $n$ is the length of the sequence.

For example, in the case of the additive error model we can use a least square error fitting procedure as follows:

$$Z(\mathbf{Y}, \boldsymbol{\theta}, \mathbf{r}, Q) = \min_{\boldsymbol{\theta}, \mathbf{r}} \sum_{q=1}^{Q} \sum_{i=r_{q-1}+1}^{r_q} (Y_i - \theta_q)^2.$$

Define $c_{ij}^q$ to be $Z(Y_i \ldots Y_j, \boldsymbol{\theta}', \mathbf{r}', q)$. In other words, $c_{ij}^q$ represents the best partitioning of the sequence $Y_i \ldots Y_j$ into $q$ segments. It follows that $c_{1,n}^Q = Z(\mathbf{Y}, \boldsymbol{\theta}, \mathbf{r}, Q)$. Because there are $\binom{n}{Q}$ partitionings of an $n$ residue sequence into $Q$ segments, finding the optimum in the above minimization problem by exhaustive enumeration is extremely computationally intensive. A more efficient optimization method is presented below.

The algorithm is composed of the following two main steps as outlined below.

begin
(* Step 1: compute the measure of fit for each segment)
1. **for** all $(i, j)$, $i \leqslant j$, $i, j \varepsilon [1, n]$ **do**
2. $\quad c_{ij}^1 \leftarrow C(F(Y_i \ldots Y_j | \theta_{ij}))$
(* Step 2: compute optimal partition for 2, 3, . . . , $Q$ segments *)
3. **for** $q = 2$ **to** $Q$ **do**
4. $\quad$ **for** $j = 1$ **to** $n$ **do**
5. $\quad\quad c_{1,j}^q \leftarrow \min_{v \varepsilon [1,j]} (c_{1,v}^{q-1} + c_{v+1,j}^1)$

end.

THEOREM. *The above algorithm computes* $c_{1,n}^Q$.

*Proof.* By induction on $q$: the basis $q = 1$ is trivial by line 2. For the induction step, note that the partition of the sequence from residues 1 to $j$ does not include any residue higher than $j$. Also the optimal partition must include a segment from residues $v+1$ to $j$ for some $v \leqslant j$. Moreover, the optimal partitioning of at most $q-1$ segment for every sequence from residues 1 to $v$ is known from the previous iteration. Consequently, the minimum across $v$ yields the optimal partitioning with at most $q$ segments. ■

Line 2 is executed $O(n^2)$ times. Assuming that it takes $T(n)$ operations to compute $c_{ij}$, then step 1 takes $O(n^2 * T(n))$ operations. Line 5 is executed $O(Q * n^2)$ times. Assuming that $Q < T(n)$, then we have a running time of $O(n^2 * T(n))$ operations.

Further improvements in computational time can be obtained by efficiently calculating $c_{ij}$ in step one. The computation of the $c_{ij}$ can be greatly facilitated when explicit solutions for the parameter estimates are available, or when the objective function is quasiconvex (or concave). A Newton–Raphson procedure can be used in the latter case. Also, the calculation of $c_{i,j+1}$ can be expedited when $c_{ij}$ can be used to iteratively calculate $c_{i,j+1}$.

2.2. *Optimal segment subset algorithm.* The previous algorithm fits a collection of models to the entire sequence. As an alternative, the method described below identifies a sub-set of $P$ non-overlapping segments, which do not necessarily span the entire sequence. The segments are chosen to minimize the sum of the length normalized measures of fit for the $P$ segments.

Let, $V(F(Y_i \ldots Y_j, \theta)$ be a length normalized measure of fit of the model $F()$ to the data $Y_i \ldots Y_j$ with the estimated vector of parameters $\theta$. Define $v(i, j)$ as $V(F(Y_i \ldots Y_j, \theta)$. For example, for least squares estimates, the unbiased estimates of residual error may be employed, i.e.

$$v(i, j) = \frac{1}{(j - i + 1 - l)} * \sum_{k=i}^{j} (Y_k - \hat{Y}_k(\theta))^2,$$

where $l$ is the number of estimated parameters.

The optimization problem then becomes:

$$\min \sum_{p=1}^{P} v(s_p, r_p),$$

where $r_{p-1} < s_p \leqslant r_p < s_{p+1}$, and $s_i \geqslant 1$, $r_i \leqslant j$.

Let this denote $S_p(j)$. This represents the optimal choice of $p$ segment neighborhoods from the first $j$ residues.

The algorithm is as follows:

$$S_i(1) = +\infty, \quad 1 < i < n$$

$$S_1(j) = \min\{S_1(j-1), \min_{1<k<j} v(k, j)\}, \quad 2 < j < n$$

$$\ldots$$

$$S_i(j) = \min[S_i(j-1), \min_{i \leqslant k < j} \{S_{i-1}(k-1) + v(k, j)\}], \quad 2 \leqslant i \leqslant P, \; i < j \leqslant n.$$

Bement and Waterman (1977) show that this algorithm computes $S_p(n)$.

*2.3. Statistical considerations.* Many of the problems of statistical inference for segmented regression remain intractable. Feder (1975b) has shown that in the identifiable case, i.e. when $r$ segments postulated by the model are known to exist and adjacent segments are distinct then the maximum likelihood parameter estimates are asymptotically normally distributed and the associated likelihood ratio is chi square distributed with the conventional number of degrees of freedom under the null postulated model. In the non-identified case the situation is more problematic (Feder, 1975a). It has been shown that the critical values appropriate to fixed partition points provide only a lower bound for the case when partition points are not fixed (Feder, 1975a).

The two segment regression model has been extensively studied. Even in this simple case, the exact distribution under the null of the likelihood ratio remain intractable except in special cases (Quandt, 1972), (Hinkley, 1971) and (Deshayes and Pichard, 1983). Approximations and bounds have been developed for the two segment problem. For example, Worsley (1983) has provided bounds based on an improved Bonferroni inequality for the two segment regression problem, and provides a simple power series approximation.

The problems of statistical inference for segmented regression are reminiscent of those found in the more extensively studied multiple subset regression problem. In the latter, the determination of the number of variables is a problem where as in the former the resolution of the number of segments is a

problem. In the case of subset regression, several stopping rule criteria have been developed including the adjusted coefficient of determination, the $F$ ratio for entry, Mallows' (1973) $C_p$, the MSEP statistic of Allen (1971), and the FPE and AIC statistics of Akaike (1970, 1974). Several of these statistics can be shown to minimize various measures of prediction errors in special cases when the future observations are taken from the same set of regressors. However, in the more general setting the inferences are more problematic. Consequently, as Seber (1977) points out, in practice the criterion choice is dependent on the planned use of the proposed regression model. In light of these difficulties, two approaches are commonly employed. Either the data analyst calculates several criteria and uses them to make a subjective judgement for stopping (Seber, 1977), or a single criteria is employed so that an objective stopping rule can be established. In stepwise regression the $F$ ratios are widely used. The stepping process continues until no variables meet pre-specified $F$ statistics for entry or removal.

The algorithms presented here provide a similar framework for model selection since differences or ratios between optimal values of the objective function are commonly used as test statistics. Consequently, several of the statistics mentioned above can be used as a stopping rule in this context. Those such as Mallows' $C_p$ which require an estimate of error with all variables included do not appear to be applicable to these problems. We have found the $F$ statistic useful for model identification when squared errors are minimized, and the likelihood ratio statistic when maximum likelihood is employed (Lawrence and Reilly, 1985). These statistics can be employed for both the parameter selection and for a stopping rule, in the same manner that such statistics are employed by traditional stepwise and optimal subset methods. We caution, however, that ". . . the critical values used in practice are purely conventional, being based on a sampling distribution when a fixed set of variables are included." (Kendall and Stuart, 1979).

The use of the $F$ statistic is strictly correct only when the residual from the model are normally distributed. This is not the case for indices defined on the 20 amino acid residues. However, the $F$ statistic appears to be fairly robust to deviations from normality (Pearson and Please, 1975). But the degree of robustness is dependent on the regressor variables. Box and Watson (1962) showed that when the distributional assumptions are relaxed, the $F$ ratio is distributed approximately as $F$ with an adjusted number of degrees of freedom.

Preliminary simulation studies (Reilly, personal communication) show that with the use of periodic models in segmented regressions, the calculated $F$ statistic follows the $F$ distribution with an adjusted number of degrees of freedom. Thus, in practice we employ such an $F$ criterion when least squares are employed or an equivalent likelihood ratio statistic criterion when more general maximum likelihood methods are used.

In some cases, it would be a mistake to be bound by fixed statistical criteria for model selection. In practice a biologic basis should be employed in modelling choices when prior knowledge warrants. An example is bacteriorhodopsin (Engelman and Zaccai, 1980) for which it is known that there are seven transmembrane segments.

*3. Alternative Hydropathic Structures.* Alternate structures for the two dimensional features of the polarity or hydropathicity of a protein are becoming recognized as important elements of protein structure. For example, DeLisi and Berzofsky (1985) have discussed the role of amphipathic structures (i.e. one with separate hydrophobic and hydrophilic surfaces) in T-cell antigenicity and Cohen and Parry (1986) have highlighted the importance of heptad repeat structures.

A coiled coil structure was proposed by Crick (1953) for the structure of the α-class of fibrous proteins. This structure is now believed to be a widespread motif in proteins, Cohen and Parry (1986). Crick showed that α-helices could mesh together with their axis inclined to one another at about 18°. In such a configuration the protein forms a regular pattern of "knobs into holes" employing a seven residue, hepted, repeat structure. Labeling the seven residues a, b, c, d, e, f and g, Crick recognized that the super coiled structure would be stabilized if residues a and d were apolar and fit into the holes of the opposing helix. To account for such coiled coil interactions we use a heptad repeat pattern in hydrophobic index as a possible model for each segment.

In the application presented below we allow three candidate model forms to represent any segment a constant level throughout the segment, a heptad repeat structure, and a binary repeat structure. Although we are looking for different features, we are using a single characteristic of a residue, namely hydropathy.

The classic heptad repeat pattern is one in which the residues exhibit the following alteration in hydropathic index: 1 hydrophobic, 2 hydrophilics, 1 hydrophobic and 3 hydrophilics. This pattern may continue for any number of residues beyond seven and it may exhibit any of seven possible phases. To model this pattern let:

$X_k$ = be a vector of the form $(1, -1, -1, 1, -1, -1, -1, 1, -1, -1, \ldots)$.

Then we model the heptad pattern in hydropathic index as follows:

$$\hat{Y}_{i+k} = \theta_0 + \theta_1 X_{k+\delta},\ k = 0 \ldots j;\ \delta = 1 \ldots 7.$$

In addition, we included a binary repeat pattern in an effort to identify sequences exhibiting an alternating pattern in hydropathic index. This would be exhibited by an amphipathic beta structure. To model this pattern let:

$$W_k = \text{be a vector of the form } (1, -1, 1, -1, 1, \ldots),$$

and:

$$\hat{Y}_{i+k} = \theta_0 + \theta_2 W_k,\ k = 0 \ldots j,$$

where $\hat{Y}_j$ is the predicted hydropathic index for residue $j$, $\theta_0$ is the unknown constant hydropathy for the segment, $\theta_1$ and $\theta_2$ are the amplitudes of their respective periodic patterns. We employ least squares to estimate the unknown parameters $\theta_0$, $\theta_1$, and $\theta_2$.

Due to the similarity between the heptad model and amphipathic helix model, we don't use the latter as a candidate model. However, we could model amphipathic helices as follows:

$$\hat{Y}_{i+k} = \theta_0 + \theta_3 \sin((2\pi/3.6)k + \theta_4),\ k = 1 \ldots j.$$

Cornette *et al.* (1987) show that least square estimation of the parameters give a more reliable estimate of the amphipathic periodicity than hydrophobic moment methods.

*3.1. Application to influenza virus haemagglutinin $HA_2$.* The viral membrane of influenza contains two integral membrane proteins, haemagglutinin and neuraminidase. The haemagglutinin of A/Hong Kong/1968 is a trimer. The X-ray crystal structure of haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution was presented by Wilson *et al.* (1981). Post translational processing divides this protein into two subunits, $HA_1$ and $HA_2$. $HA_2$ forms a triple stranded coiled coil of alpha helices which extends 76 Å from the membrane. Because of this extended form, $HA_2$ has a high ratio of surface area to volume. Its assembly as a trimer also leads to several regions of substantial contact between elements of the secondary structure of the monomers. A search for segment neighborhoods with a heptad amphipathic structure thus seems promising for the $HA_2$ subunit. Also, beta strands may exhibit an amphipathic pattern. Some segments are likely to follow neither of these patterns. These will be described as segments with random fluctuations in hydrophobicity around a constant mean value. Thus, for the haemagglutinin example, we search for multiple segments. Each segment follows one of these three candidate models. An $F$ ratio statistic for entry of either of these periodic structures into the model with a nominal $p$ value $<0.005$ is employed. The following analysis was done using the optimal segmentation algorithm.

While the algorithm can be employed to describe multivariate characteristics of each residue, in this analysis we consider only a single characteristic, the hydrophobic index. No other features of the amino acids are employed in the analysis. For example $\alpha$-helix, $\beta$-sheet and turn tendencies are not used. Thus,

$\alpha$-helices or $\beta$-sheets not located at a polar to apolar interface are not expected to exhibit amphipathic structure, i.e. periodicity in hydropathic index. Furthermore, as pointed out by Cohen and Parry (1986), such periodic structures can occur even when the corresponding secondary structure is not believed to be present. Whenever a segment lies between an aqueous environment and a more polar environment periodicity in hydropathicity appears plausible. The finding of such an amphipathic structure which does not correspond to the secondary structure may indicate a change in secondary structure which occurs during conformational change.

The observed and predicted hydropathicities for the 221 residues of $HA_2$ are given in Figs 1a and b. The major structural features of the protein are also indicated on this figure. The amino terminus forms an ". . . uncommon series of four reverse turns . . ." (Wilson *et al.*, 1981). This terminus is highly similar with the $F_1$ component of the Sendai virus fusion protein. This peptide is believed to be the fusion peptide of influenza virus haemagglutinin (see Wilson *et al.*, 1981). The first ten residues have previously been described by Skehel *et al.* (1982) as an uncharged hydrophobic sequence. We also find this region to be hydrophobic, $\theta_0 = 1.88$. In addition, as indicated in Fig. 1, the algorithm finds a periodic structure for the first ten residues. Thus, this segment exhibits a helical hydropathic structure in addition to it's reverse helical physical structure. In this case, there is a good correspondence between the secondary structure of contiguous reverse turns and the hydrophobic structure. The amphipathic nature of this segment neighborhood may play a role in the membrane fusion.

As indicated in Figs 1a and b, no binary repeat patterns are predicted but four heptads helices are predicted between residues 45 and 150. While there is substantial agreement between the predicted heptad helices and the alpha helices observed in the crystal structure, there is a striking departure from this agreement in the central portion of the long alpha helix. The question remains why did the heptad structure break down in the middle of this long helix?. The tertiary structure gives a clue. A stem from the $HA_1$ sub-unit formed from residues $HA_1$ (25–34) intertwines with $HA_2$ at residue 101 of $HA_2$ (17). Thus, the residues that make up the clearly hydrophobic segment $HA_2$ (98–102) are in close proximity with this stem from $HA_1$. The fusion peptide sequence at the *N*-terminus of $HA_2$ is one helix turn below this intersection, Skehel *et al.* (1982).

Furthermore, Skehel *et al.* (1982) have shown that the influenza haemagglutinin undergoes a conformational change in the pH range 4.9–5.2. Trypsin treatment of this induced conformation at a pH of 5.0 results in a cleavage of $HA_1$ at residue 27. A trypsin cleavage at this site does not occur in the untreated protein. The fusion peptide sequence lies 35 Å away from the viral membrane and 100 Å away from the cell membrane. A conformational change which allows a closer approach of the fusion peptide to these membranes is thus indicated. Marginal stability is often important to allow conformational

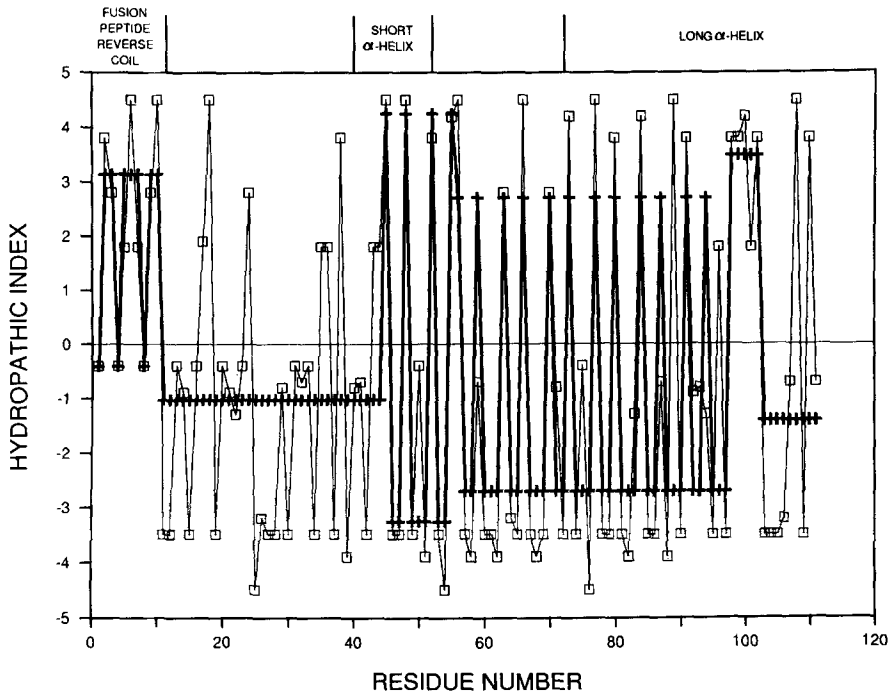


Figure 1(a).

change. Breaks in the heptad repeat structure can result in marginal stability of coiled coil structures (Cohen and Parry, 1986). Thus, the indicated interruption in the heptad repeat structure may provide the necessary instability in the coiled coil to allow for the observed conformational change.

The algorithm identifies a distinctive pattern of 5 hydrophobic residues ($HA_2$ 98–102). The hydrophilic segment $HA_2$(103–115) contains a histidine amino acid 3 residues down stream of the disjunction at residue 103. The Histidine R-group has a pK of 6. At a neutral pH of 7, only about 10% of these R-groups are expected to be charged. But at a pH of 5, at which the conformational change occurs, 90% of these residues are expected to be charged. This indicated change in charge is consistent with the observed pH induced conformational change. Also, the hydrophobic segment neighborhood $HA_2$(98–102) may assist in making the approach of the fusion peptide to a membrane more energetically favorable. Thus, both the hydrophilic and the hydrophobic segments which are identified by the algorithm in the interruption of the heptad pattern have features consistent with the conformational change that occurs with a pH change. Because of the overlapping windows of fixed size that are inherent in moving average methods this pattern which emerges from the algorithm as a distinctive feature would be completely averaged out by moving average methods.

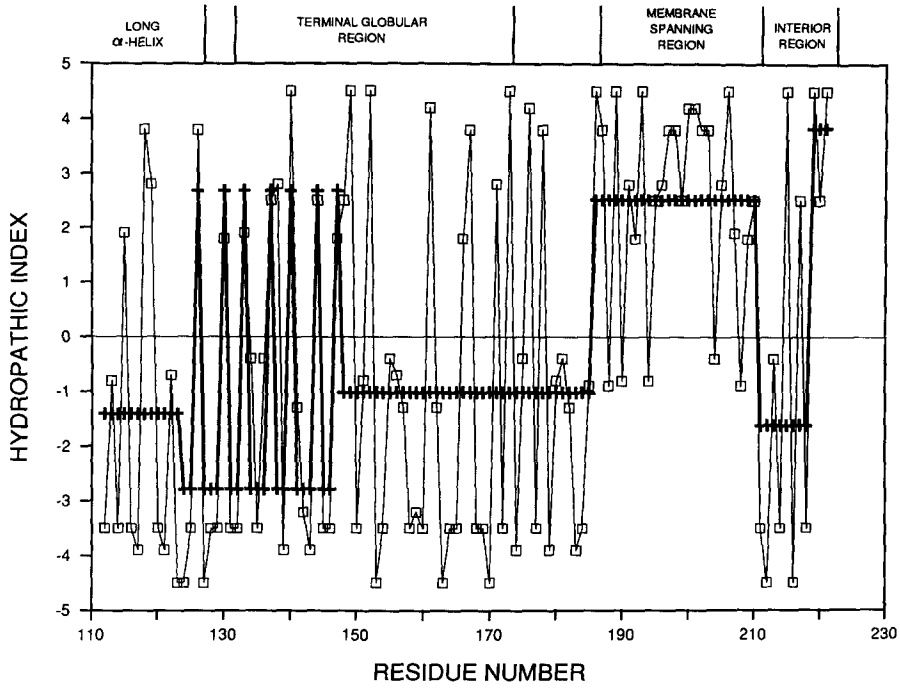


Figure 1(b).

Figure 1. Predicted (thick line with +) and observed (thin line with squares) hydropathic indices for the $HA_2$ sub-unit of haemagglutinin protein from influenza virus. Major features of the X-ray crystal structure are indicated by the captions above the graphs. Four heptad segments are predicted by the algorithm. They are (1–10), (45–55), (56–101) and (124–147). The algorithm also identifies a distinctive pattern of 5 hydrophobic residues (98–102). The membrane spanning region is identified by the algorithm as a hydrophobic segment at (186–210). The interior region has a predicted hydrophilic segment (211–218) and a hydrophobic segment (219–221). The number of segments $Q$ is 10, and was obtained when the $F$ statistics first exceeded the nominal value for the $F$ distribution of 0.025. Tl e minimum segment length for heptads is 7.

We also applied the optimal subset algorithm to $HA_2$. The results are shown in Table I. We should note that the results are similar to the optimal segmentation analysis presented above. It finds heptads in subsegments of the regions identified by the other algorithm. however, it misses the one from 1–10, and finds one at 196–202. Identified segments tend to be shorter. The measure used for this analysis was mean squared residual error.

*4. Conclusion.* We have presented two algorithms to identify segment neighborhoods and to estimate the parameters of the models that characterize these segment neighborhoods. These algorithms will be useful when analyzing polymers characterized by segments composed of a contiguous set of residues

TABLE I
Segments with Minimum Total Mean Residual Squared Errors (MRSE) for $HA_2$

| Number of segments | Total MRSE | Boundaries | | | | |
|---|---|---|---|---|---|---|
| 1 | 0.06 | 196–202 | | | | |
| 2 | 0.26 | 127–133 | 196–202 | | | |
| 3 | 0.51 | 76–82 | 127–133 | 196–202 | | |
| 4 | 0.77 | 60–69 | 76–82 | 127–133 | 196–202 | |
| 5 | 1.88 | 60–69 | 76–82 | 127–133 | 141–147 | 196–202 |

Minimum segment length is 7. First 5 segments only. We note that segment 141–147 contributes as much to the total MRSE as the other ones combined.

sharing common structural or functional characteristics. The principal remaining difficulties lies in the identification of the classes of models that describe structural characteristics in native proteins, and solution of the problems in statistial inference common to segmented regression.

As Table I shows, there is a tendency of the optimal segment sub-set algorithm using the minimum sum of mean residual error to select segments that are near the minimum specified length. This stems from the fact that there is higher variability in the measure for short sequences. One alternative measure that has shown promise in our preliminary work employs the probability of type one error under a null model of regression coefficients equal to 0 and a specified error distribution as a statistic. However, a statistic of this kind makes the difficult inference problems discussed above central to the algorithm. This is an area that clearly requires further research.

During our experience in the analysis of numerous proteins with the optimal segment algorithm, we have observed that the apparent biologic features that are identified by the analysis are not particularly sensitive to the number of segments included in the analysis. For example, in the the haemagglutinin application presented here all of the biologic features described are identified in the results with $Q$ ranging from 10 to 20.

Native proteins are undoubtedly composed of several segments which exhibit many different structural features. The forms of some of these have been characterized, such as membrane spanning domains and amphipathic helices. New forms remain to be described. For example, Leszczynski and Rose (1986) have recently described a new structural feature, the loop. Existing moving window methods do not provide a means for the delineation of multiple structural characteristics in a single analysis and do not provide any selection criteria for the combination of such features from multiple analyses. The simultaneous identification of multiple structural features using a common criteria is a fundamental feature of the algorithms presented here. The example

illustrates this capability. Because of the general form of the algorithms, as the characteristics of other structural features, such as loops, become more thoroughly described they can be simply and routinely incorporated into the analysis. Model selection is aided by the use of test statistics, in a manner analogous to widely used procedures for model selection in multivariate statistical analysis.

The optimal segment algorithm considers all possible partitionings of a sequence into $Q$ segment neighborhoods. This is advantageous over moving average methods when the sequence under consideration has segment neighborhoods of varying lengths and in unknown locations. This algorithm enjoys its greatest advantage in this respect over existing methods when the features of interest are contained in segments whose lengths do not correspond to the selected window size. Features that are longer than the selected window size will tend to be overlooked by moving window methods because no segments of the given window size is in itself striking. On the other hand, segments shorter than the window width will be averaged over by moving window methods. These difficulties are overcome by the algorithm since it identifies the best model for every possible segment regardless of length, and then successively chooses the 2, 3, 4, . . . , $Q$ most significant features.

The prediction of features of higher order structure or function from primary structure data remains a challenging problem. While the algorithm presented here is broadly applicable to this class of problems, it is important to keep in mind that it can only be expected to be useful when the characteristics of interest are believed to be shared by contiguous residues.

We thank Randy Weinberg, Larry Sturman, Leo Grady, Bob Rej, and Marlene Belfort for suggestions and discussions on the biological examples. We also thank Dan Davison for the DeLisi reference. We acknowledge Andrew Reilly for suggesting the statistical test for the least square examples, as well as for several informal discussions. Finally, we thank the editor, Prof. M. S. Waterman, for bringing to our attention the optimal segment sub-set algorithm, as well as the work by Hawkins (1976) and Bellman and Roth (1966) on segmented regression. Computer resources were generously provided by the New York State Department of Health and by BIONET's NIH grant # 1 U41 RR-01685-03. The algorithm has been implemented in VAX/VMS FORTRAN. This code will be made available through the BIONET computer resource.

## LITERATURE

Akaike, H. 1970. "Statistical Predictor Identification." *Ann. Inst. Statist. Math.* **22**, 203–217.

———. 1974. "A New Look At Statistical Model Identification." *IEEE Trans. Auto. Control* **19**, 716–723.

Allen, D. M. 1971. "Mean Square Error of Prediction as a Criterion for Selecting Variables." *Technometrics* **16**, 469–475.

Bellman, R. and R. Roth. 1966. "Curve Fitting by Segmented Straight Lines." *J. Am. Statist. Assoc* **64**, 1079–1084.

Bement, T. R. and M. S. Waterman. 1977. "Locating Maximum Variance Segments in Sequential Data." *Math. Geol.* **9**, 55–61.

Box, G. E. P. and S. Watson. 1962. "Robustness to non-normality of Regression Tests." *Biometrika* **17**, 83–91.

Cohen C. and D. A. D. Parry. 1986. "α-Helical Coiled Coils—A Widespread Motif in Proteins." *Trends in Biochemical Sciences* **11**, 245–248.

Crick, F. H. 1953. "The packing of α-helices: Simple Coiled Coil." *Acta Cryst.* **6**, 689–697.

Cornette, J. L., K. B. Cease, H. Margalit, J. L. Spouge, J. A. Berzofsky and C. DeLisi. 1987. "Hydrophobicity Scales and Computational Techniques for Detecting Amphipathic Structures in Proteins." *J. Molec. Biol.* **195**, 659–685.

Dayhoff, M. O., R. N. Schwartz and B. C. Orcutt. 1978. *Atlas of Protein Sequence and Structure*, Vol 3, pp. 345–352. Silver Spring, MD: National Biomedical Research Foundation.

DeLisi C. and J. A. Berzofsky. 1985. "T-cell Antigenic Sites Tend to be Amphipathic Structures." *Proc. Natn. Acad. Sci. U.S.A.* **82**, 7048–7052.

Eisenberg, D., E. Schwarz, M. Komaromy and R. Wall. 1984. "Analysis of Membrane and Surface Protein Sequences with the Hydrophobic Moment Plot." *J. Molec. Biol.* **179**, 125–142.

Engelman, D. M. and G. Zaccai. 1980. "Bacteriorhodopsin is an Inside-Out Protein." *Proc. Natn. Acad. Sci. U.S.A.* **77**, 5894–5898.

Eventoff, W., M. G. Rossmann, S. S. Taylor, H. J. Torff, H. Meyer, W. Keil and H. H. Kiltz. 1977. "Structural Adaptation of Lactate Dehydrogenase Isozymes." *Proc. Natn. Acad. Sci. U.S.A.* **74**, 2677–2681.

Feder, P. I. 1975a. "On Asymptotic Distribution Theory in Segmented Regression Problems—Identified Cases." *Ann. Statistics* **3**, 49–83.

———. 1975b. "The log Likelihood Ratio in Segmented Regression." *Ann. Statistics* **3**, 84–97.

Flory, P. J. 1956. "Theory of Elastic Mechanisms in Fibrous Proteins." *J. Am. Chem. Soc.* **78**, 5222–5235.

Fousler, D. E. and S. Karlin. 1987. "Maximal Success Duration for A Semi-Markov Process." *Stochastic Processes Applic.* **24**, 203–224.

Hawkins, D. M. 1976. "Point Estimation of the Parameters of Piecewise Regression Models." *Appl. Statistics* **25**, 51–57.

Heijne, G. von, 1986. "Mitochondrial Targeting Sequences May Form Amphiphilic Helices." *EMBO J.* **5**, 1335–1342.

Hinkley, D. V. 1971. "Inference in Two-phase Regression." *J. Am. Statis. Assoc.* **66**, 736–743.

Hopps, T. P. and K. P. Woods. 1981. "Prediction of Protein Antigenic Determinants from Amino Acid Sequences." *Proc. Natn. Acad. Sci. U.S.A.* **78**, 3824–3828.

Karlin, S. and G. Ghandour. 1985. "Multiple Alphabet Amino Acid Sequence Comparisons of the Immunoglobulin Kappa-gene." *Proc. Natn. Acad. Sci. U.S.A.* **82**, 8597–8601.

Kendall, M. and A. Stuart. 1979. *The Advanced Theory of Statistics*. New York: Macmillan.

Kirschner, K. and H. Bisswanger. 1976. "Multifunctional Proteins." *A. Rev. Biochem.* **45**, 143–166.

Kyte, J. and R. P. Doolittle. 1982. "A Simple Method for Displaying the Hydropathic Character of a Protein." *J. Molec. Biol.* **157**, 105–132.

Lawrence, C. E. and A. A. Reilly. 1985. "Maximum Likelihood Estimation of Subsequence Conservation." *J. Theor. Biol.* **113**, 425–439.

Lerman, P. M. 1980. "Fitting Segmented Regression Models by Grid Search." *Appl. Statistics* **29**, 77–84.

Leszczynski, J. F. and G. D. Rose. 1986. "Loops in Globular Proteins: A Novel Category of Secondary Structure." *Science* **234**, 849–855.

Mallows, C. L. 1973. "Some Comments on $C_p$." *Technometrics* **15**, 661–675.

Pearson, E. S. and N. W. Please. 1975. "Relation Between the Shape of Population Distributions and the Robustness of Four Simple Statistical Tests." *Biometrika* **62**, 223–241.

Quandt, R. E. 1972. "New Approaches to Estimating Switching Regressions." *J. Am. Statist. Assoc.* **67**, 306–330.

Rose, G. D. 1979. "Hierarchic Organization of Domains in Globular Proteins." *J. Molec. Biol.* **134**, 447–470.

Sankoff, D. and J. B. Kruskal (Eds). 1983. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.* Reading, MA: Addison-Wesley.

Schulz, G. E. and R. H. Schirmer. 1979. *Principles of Protein Structure.* New York: Springer.

Seber, G. A. F. 1977. *Linear Regression Analysis.* Wiley: New York.

Skehel, J. J., P. M. Baylel, E. B. Brown, S. R. Martin, M. D. Waterfield, J. M. White, I. A. Wilson and D. C. Wiley. 1982. "Changes in the Conformation of Influenza Virus Hemagglutinin at the pH Optimum of Virus-mediated Membrane Fusion." *Proc. Natn. Acad. Sci. U.S.A.* **79**, 968–972.

Sternberg, M. J. and E. Thornton. 1977. "On the Conformation of Proteins: An Analysis of $\beta$-Pleated Sheets." *J. Molec. Biol.* **110**, 285–296.

Waterman, M. S. 1984. "General Methods of Sequence Comparison." *Bull. Math. Biol.* **46**, 473–500.

Wetlaufer, D. E. 1972. "Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins." *Proc. Natn. Acad. Sci. U.S.A.* **70**, 697–701.

Wilson, I. A., J. J. Skehel and D. C. Wiley. 1981. "Structure of the Haemagglutinin Membrane Glycoprotein of Influenza Virus at 3 Å Resolution." *Nature* **289**, 366–373.

Worsley, K. J. 1983. "Testing for Two-phase Multiple Regression." *Technometrics* **25**, 35–42.

Received for publication 16 June 1988