Journal of Computer-Aided Molecular Design, 1 (1987) 323–341
ESCOM

J-CAMD 026

Molecular modeling of protein structure and function: A bioinformatic approach

Michael N. Liebman

Departments of Physiology and Biophysics, and of Pharmacology, Mount Sinai School of Medicine of the City University of New York, New York, NY 10029, U.S.A.

Key words: Information theory; Artificial intelligence; Serine proteases; Fourier transform-infrared spectroscopy; Macromolecular modeling; Structure-function analysis

SUMMARY

This paper reports on the data/information structure of macromolecules as it extends beyond the three-dimensional conformation to include functional descriptors of biochemical (in vitro) and biological (in vivo) characteristics and as it contrasts with the limitations imposed by the data reduction and data classification techniques of traditional molecular modeling. Methodologies for structure-function representation are presented which are being incorporated within a knowledge-acquisition expert system. Examples of the bioinformatic approach are presented concerning macromolecular recognition by serine proteases and the use of Fourier transform-infrared (FT-IR) spectroscopy for structural assignment and analysis by a novel structure-perturbation approach.

INTRODUCTION

Researchers, both computational and experimental, acknowledge that understanding the molecular mechanisms of action in biological systems requires the quantitative description of the relationship between the structure of a molecule, its resultant physicochemical properties and its observed biological reactivity and specificity. Molecular modeling has evolved to incorporate a variety of computational approaches — including computer graphics and molecular dynamics along with quantitative structure-activity analysis methods and quantum mechanics — as well as expanded its use of high resolution experimental methods (e.g., NMR, FT-IR), yet no set of methodologies has proven consistent in predicting the structure or activity of small molecules or macromolecules of biological interest. Thus no library of expert rules exists which is capable of yielding consistently accurate structure-activity predictions.

To explain this failure, fault is assigned to one of three conditions: (1) the experimental and computational data are inadequate to describe the quantitative aspects of the structure-function relationship; (2) the tools of molecular modeling are presently inadequate for analyzing and inte-

grating the experimental and computational data, which are themselves adequate; or (3) no rules exist which are generally applicable for defining the structure-function relationship. Most scientists discard the latter explanation because a specific amino acid sequence will yield a unique three-dimensional structure for the protein which it defines; even small molecules exhibit functions which differ significantly among simple congeners. We report our research efforts to examine the information content of a biomacromolecular system and to determine if other methods of data representation and analysis might extend access to potential structure-function-activity relationships.

We are presently developing computational algorithms for representation and analysis of molecular structure, biological function, biochemical function and physicochemical properties [1–14]. These are being inserted into an intelligent user-friendly environment to construct a computer-based system for gathering observations and to inferring and evaluating potential rules from the observations made by individual scientists and the system itself. In this manner, we are attempting to enable the computational environment to become the potential expert. To achieve this, we incorporate the principles of information theory and the methods associated with artificial intelligence to overcome some of the present limitations in molecular modeling approaches. These limitations include the use of 'data/information reduction' methods, the tendency to describe a molecule or function using restricted 'data/information classification' methods, and the limited terminology which has become the jargon of a particular field through its appropriate use and misuse. We have found this bioinformatic approach to problem-solving in molecular modeling places fewer limitations on the questions which we can address. In this manner, we are learning to identify and answer the questions pertinent to the specific research problems being studied, and to learn from the experience and observations of having done so.

We describe data/information reduction methods as the set of processes by which the multidimensional (i.e., greater than three dimensions) nature of the information which completely describes a molecule is reduced in dimensionality for purposes of representation, visualization and analysis. Data reduction methods may be applied to data specifically pertaining to molecular structure, in vitro and in vivo function and physicochemical properties, individually or in combination. Thus we would describe the representation of a macromolecule using an interactive computer graphics display as a reduction of the information content of a macromolecule, selecting only the information concerning the three-dimensional coordinates and the atom connectivity table. This graphical representation is a further reduction of the total information because it depicts a single conformation, possibly the average crystallographic structure or photograph of the dynamics trajectory. Other common forms of data/information reduction include the representation of complex processes (e.g., the coagulation cascade) as a two-dimensional pathway for purposes of publication and the use of single statistic descriptors (e.g., root-mean-squared deviation (RMS) for describing the structural comparison of two proteins. In the former example, the actual coagulation process is at least four-dimensional in nature due to several components which are required to enter in a particular order or sequence, or at multiple sites. In the latter example, a single-value description may be used to describe a distribution that actually exhibits spatial relationships closely linked to function, as well as the fact that the same single descriptor may describe two such comparisons which differ significantly with respect to their actual distributions. While we present examples of the limitations that result from data/information reduction, we readily acknowledge the necessity and usefulness of these techniques, and emphasize the need to be aware

of the specific biases or limitations which might be inherent in the use of a particular form of data reduction.

The emphasis of specific characteristics or descriptors by data/information reduction methods may bias their use in data/information analysis. When this emphasis is coupled with nomenclature or jargon which may vary significantly between definition and use, we encounter the limitations of data/information classification. We typically effect classification by designating a list of characteristics to a particular label and assigning such a label to the object being examined. Thus the classification may be carried out using a reduced set of descriptors or with labels whose definitions may prove generally applicable because of the absence of a more quantitative method of classification. Classification does prove useful in conveying the list of associated descriptors, but the lack of quantization — even at the level of defining the descriptors — reduces the value of the information being transferred. Thus, two investigators might use the same label to describe a protein where its list of descriptors is not commonly applicable (e.g., serine protease is used to describe both trypsin and subtilisin, whereas this descriptor is based on the apparent similarity in active site residues and mechanism, not similarity in conformation); other enzymes designated in this manner (e.g., trypsin and elastase) can share both descriptors. Similarly, two investigators could readily generate non-identical descriptions of the secondary structure for the same protein based on the same data. Such an observation could result from each using different methods for describing a particular conformation (e.g., hydrogen-bonding patterns, neighboring atom distances, helix parameters), a lack of exact definition of particular structural patterns, or the limitations imposed by classifying structures which are represented by reduced data sets. In either example, the inability to evaluate the accuracy of the structural assignments by independent observation limits the accuracy of the classification, but exact definition of the parameters used to generate the classification establishes the constraints under which the classification is suitable for use. It is essential that any such constraints be associated and carried along with this classification to eliminate ambiguities. We examine below the identification and classification of protein secondary structure in the development of our substructure library; we also examine the classification of enzyme activity as contrasted between in vitro and in vivo studies in our discussion of the serine proteases.

We can begin to appreciate the impact of the potential biases imposed by data reduction and data classification methods by considering the various levels of comparison of two protein molecules. The quantitative comparison of two macromolecules can be based on a wide range of structure and/or functional descriptors. The most commonly used descriptors for proteins are the amino acid sequence data and the three-dimensional structure as observed by X-ray crystallographic analysis. We describe below some of the forms of comparisons which are typically carried out using such data, in light of the limitations outlined above.

An amino acid sequence represents the unique one-dimensional transform of the covalent, chemical organization of the polypeptide chain and relates to the actual process of synthesis on the ribosome. The relationship between two sequences can be evaluated from the separate sequences or after an alignment is carried out to optimize sequence identity, conservation of physicochemical properties, minimum base change per codon to accomplish sequence changes, probability of observed changes based on observed mutation rates, and locations and sizes of insertions and/or deletions needed to effect the alignment. The information content of the amino acid sequence differs from the amino acid composition (Table 1) only in the presence of the ordering of the successive amino acids in the chain. Thus a 20 amino acid peptide, containing one of each of

the 20 naturally occurring amino acids, is but one of 20! (i.e., 2.4×10^{18}) possible sequences, all of the same composition but exhibiting different ordering of the components. Thus the comparison of two specific amino acid sequences is distinct from that of their amino acid compositions and not commutative (i.e., similar protein sequences exhibit similar compositions but similar compositions may come from nonsimilar sequences). Comparison of the compositions, alone, can suggest a (dis)similarity which is not present when the sequence positions are also considered; also the presence of local (i.e., in sequence position) regions of sequence similarity may be obscured by large regions of nonsimilarity when a single statistic rather than the observed distribution is used to describe the data. Thus two structurally superimposed proteins (e.g., elastase and trypin in Table 2) can be evaluated as to the analogy of the structurally equivalenced sequence positions. Such a comparison may use minimum base change per codon as a measure of sequence similarity. This measure may not accurately reflect the relatedness of two sequences, as it does not represent the source of the relationship (i.e., the metric computes the minimum number of base changes needed to accomplish a sequence change, not the actual probability of the changes observed). In addition, the actual profile of individual residue comparisons may reveal localized regions of significant similarity or dissimilarity which differ from the average computed over the entire sequence (Table 3).

The availability of the gene and, therefore, the DNA coding sequence for some proteins additionally provides the intron/exon positions and sizes, codon selection statistics and information on leader sequences. The comparison of two proteins on the basis of their protein coding sequences more closely relates to the functional state involved in the initial folding of the protein at a level of the structure-function relationship that is not well characterized or understood. Correlation of gene structure with protein structural information is of great potential significance, but at pre-

TABLE I
COMPARISON OF THE AMINO ACID COMPOSITION OF SEVERAL SERINE PROTEASES

Elastase, porcine, 240 amino acids with composition																		
A	17*	C	8*	D	7*	E	4*	F	3*	G	25*	H	6*	I	10*	K	3*	L	18*
M	2*	N	17*	P	7*	Q	15*	R	12*	S	22*	T	19*	V	27*	W	7*	Y	11*
Trypsin, bovine, 223 amino acids with composition																		
A	14*	C	12*	D	6*	E	4*	F	3*	G	25*	H	3*	I	15*	K	14*	L	14*
M	2*	N	16*	P	8*	Q	10*	R	2*	S	34*	T	10*	V	17*	W	4*	Y	10*
Protease A, *Streptomyces griseus*, 181 amino acids with composition																		
A	19*	C	4*	D	4*	E	2*	F	5*	G	32*	H	3*	I	9*	K	0*	L	10*
M	1*	N	12*	P	4*	Q	6*	R	7*	S	21*	T	21*	V	12*	W	1*	Y	8*
Protease B, *Streptomyces griseus*, 185 amino acids with composition																		
A	15*	C	4*	D	7*	E	2*	F	5*	G	32*	H	2*	I	7*	K	1*	L	7*
M	2*	N	10*	P	5*	Q	2*	R	8*	S	21*	T	28*	V	15*	W	2*	Y	10*
Alpha lytic protease, Myxobacter 495, 198 amino acids with composition																		
A	24*	C	6*	D	2*	E	4*	F	6*	G	32*	H	1*	I	8*	K	2*	L	10*
M	2*	N	13*	P	4*	Q	9*	R	12*	S	20*	T	18*	V	19*	W	2*	Y	4*

sent remains an unknown [15]. Yet, without a mechanistic basis, correlation alone may not prove adequate for furthering our understanding in this complex area. Therefore, it is an important consideration that the correlation of properties that can be derived from an amino acid sequence with structural parameters exists as a correlation and not as a mechanistic statement of the source of that correlation.

The observations above concern the present lack of understanding of the structure-function relationship in proteins in general. In particular, observations on sequence (amino acid or nucleic acid) information can be further compounded by the various ways used to compare the three-dimensional structures of two proteins. Structural superposition by the commonly applied method of generation of a rotation/translation matrix, and as monitored by an RMS deviation of equivalenced protein residues, is limited in several respects: the need to initially identify equivalent residues; the potential effects of alignment of global structural features which may obscure localized perturbations; and the use of a single statistic for monitoring the quality of fit as opposed to the actual distribution (Table 2). To equate a specific RMS value from two different superpositions is therefore inappropriate (i.e., an RMS of 1.9 Å comparing parvalbumin and troponin, and comparing intestinal calcium binding protein and troponin, does not reveal details of the relationship between parvalbumin and intestinal calcium binding protein). The limitation of superposition derives from its initial development to reveal structural similarity between proteins, not to distinguish small differences [16]. Thus, to deal with the comparison of the three-dimensional structures of two proteins, we have developed or extended the use of other methods based on forms of structural representation that also serve for data/information storage of both structural and functional information.

TABLE 2
HISTOGRAM OF THE SUPERIMPOSITION OF ELASTASE ONTO TRYPSIN (RMS IS 0.72 Å/188 AMINO ACIDS)

Difference in position	Number	Fraction	Cum. fraction
$0.00 \leqslant DIF < 0.10$	0	0.000	0.000
$0.10 \leqslant DIF < 0.20$	10	0.054	0.054
$0.20 \leqslant DIF < 0.30$	12	0.065	0.120
$0.30 \leqslant DIF < 0.40$	29	0.158	0.277
$0.40 \leqslant DIF < 0.50$	27	0.147	0.424
$0.50 \leqslant DIF < 0.60$	27	0.147	0.571
$0.60 \leqslant DIF < 0.70$	13	0.071	0.641
$0.70 \leqslant DIF < 0.80$	17	0.092	0.734
$0.80 \leqslant DIF < 0.90$	12	0.065	0.799
$0.90 \leqslant DIF < 1.00$	8	0.043	0.842
$1.00 \leqslant DIF < 1.10$	6	0.033	0.875
$1.10 \leqslant DIF < 1.20$	9	0.049	0.924
$1.20 \leqslant DIF < 1.30$	3	0.016	0.940
$1.30 \leqslant DIF < 1.40$	6	0.033	0.973
$1.40 \leqslant DIF < 1.50$	2	0.011	0.984
$1.50 \leqslant DIF < 1.60$	2	0.011	0.995
$1.60 \leqslant DIF < 1.70$	1	0.005	1.000

As examples of our bioinformatic approach we present two studies. The first describes the ongoing research into the family of enzymes termed serine proteases [10,13]; the second describes how the type of knowledge gained in such studies can be extended to other macromolecular systems and, in particular, to the analysis of FT-IR [14].

METHODS

We consider the basic construct of the approach to be applicable to a wide variety of studies as it centers about the pathway:

Data or Information	→	Methods of Representation, Analysis and Comparison	→	Observations: (1) in a specific system (2) generally applicable (rules)

The data or information presented for study can be further organized by content: structural (i.e., three-dimensional conformation); sequence-based (both amino acid and nucleic acid); biochemical (e.g., enzyme activity/specificity as measured in the laboratory (in vitro conditions); biological (e.g., participation in physiological pathways or cascades); genetic (i.e., organization of the gene into exon and intron segments); and physicochemical (i.e., physical properties which can be measured such as spectroscopy and pH profiles). We include computationally-generated information within the methods of representation, analysis and comparison presented below.

Our interest in accessing the full information content of a macromolecule has led us to use and develop several methods of nontraditional structural representation, analysis and comparison [1–14]. We include only a brief summary of the approaches, as they have been described in detail elsewhere, and use this presentation to emphasize the relationships which exist among them. A certain degree of redundancy that appears within these methods is intentional, as it provides for consistency checks of the observations which can be made from the systematic analyses and reduces the potential for failing to make the observations. The methods can be categorized by their dimensionality (recognizing the data reduction component of the approach and its limitations) and the

TABLE 3
TOPOGRAPHICAL SUPERPOSITION OF SERINE PROTEASES AND EVALUATION OF SUPERIMPOSED AMINO ACID SEQUENCES

Start n1	Start n2	Number	MBC	MBC cum.	PAM	PAM cum.
TPO vs. EST, RMS = 0.72 Å/188 residues, MBC = 0.84, PAM = 14.28						
1	1	7	1.00	1.00	13.29	13.29
12	12	9	0.78	0.88	12.11	12.63
22	27	21	0.71	0.78	15.38	14.19
45	52	14	0.93	0.82	13.14	13.90
60	67	17	0.71	0.79	14.82	14.13
82	91	26	0.92	0.83	13.23	13.88
114	125	12	0.67	0.81	15.75	14.09

Start n1	Start n2	Number	MBC	MBC cum.	PAM	PAM cum.
TPO vs. EST, RMS = 0.72 Å/188 residues, MBC = 0.84, PAM = 14.28						
132	142	12	1.42	0.87	12.33	13.92
148	158	3	0.67	0.87	16.33	13.98
151	163	3	2.00	0.90	12.33	13.94
155	167	3	1.33	0.91	13.00	13.91
162	174	3	0.00	0.88	18.00	14.01
168	179	16	0.50	0.84	15.88	14.21
185	200	11	1.09	0.86	12.73	14.11
197	214	27	0.74	0.84	15.26	14.28
SGA vs. TPO, RMS = 1.48 Å/106 residues, MBC = 1.16, PAM = 12.49						
13	24	4	1.25	1.25	13.25	13.25
18	28	4	1.25	1.25	11.25	12.25
26	33	10	1.10	1.17	14.10	13.28
38	43	6	1.67	1.29	12.17	13.00
47	70	3	1.67	1.33	11.33	12.81
53	82	13	0.85	1.17	13.08	12.90
86	109	14	1.14	1.17	11.57	12.56
101	136	10	1.50	1.22	11.30	12.36
124	159	6	1.50	1.24	12.33	12.36
131	171	26	0.85	1.14	12.85	12.49
162	203	10	1.40	1.16	12.50	12.49
SGA vs. SGB, RMS = 0.37 Å/154 residues, MBC = 0.35, PAM = 15.13						
1	1	8	0.50	0.50	14.13	14.13
12	12	10	0.20	0.33	16.50	15.44
25	25	12	0.67	0.47	14.25	14.97
40	40	3	0.67	0.48	15.33	15.00
44	52	18	0.28	0.41	15.72	15.25
69	78	3	0.33	0.41	14.00	15.19
78	82	39	0.38	0.40	14.21	14.77
120	124	61	0.28	0.35	15.67	15.13
SGA vs. ALP, RMS = 0.60 Å/144 residues, MBC = 0.83, PAM = 13.62						
2	4	7	1.00	1.00	13.14	13.14
12	15	12	0.67	0.79	14.25	13.84
26	29	9	0.78	0.79	14.78	14.14
40	43	3	1.33	0.84	9.00	13.65
44	52	18	1.00	0.90	12.83	13.35
70	78	4	1.00	0.91	12.50	13.28
76	83	6	1.17	0.93	14.00	13.36
88	95	29	0.66	0.84	13.93	13.55
120	126	24	0.71	0.81	13.63	13.56
145	152	11	0.64	0.80	13.73	13.58
164	179	18	1.06	0.83	13.94	13.62

ability to extend the methods for structural representation to functional (e.g., physicochemical) descriptors within the same construct. We describe any representation which inherently assumes a reference to position along the amino acid sequence as one-dimensional; methods which describe pair-wise residue characteristics are two-dimensional; and those which represent physicochemical properties in Cartesian space or reduce four-dimensional results of molecular dynamics calculations (i.e., time is the fourth dimension) are three-dimensional.

STRUCTURE-FUNCTION REPRESENTATION

One-dimensional

(a) Linear distance plot (structure) [3–7,10,12–14] (Figs. 1 and 2)

Linear distance plot analysis involves generating a plot using each successive amino acid in the protein sequence as an origin for the computation of the sum of the series of distances from the origin alpha carbon to each of the four successive alpha carbons. The resultant plot yields a detailed profile of the local folding of a protein and has been used to identify new classes of local structure (i.e., secondary and supersecondary structure) a well as enable the structural comparison of two proteins using dynamic programming algorithms.

Advantages: independent of molecular orientation; useful for analyzing complex patterns of local folding, for defining and comparing secondary structure, and for topographically mapping proteins.

Disadvantages: lacks handedness as only distances are used; represents only contiguous structures (i.e., secondary and supersecondary structure, not tertiary).

(b) Linear property plot (function) [Williams and Liebman, submitted]

Linear property plot analysis represents the physicochemical properties that are associated with

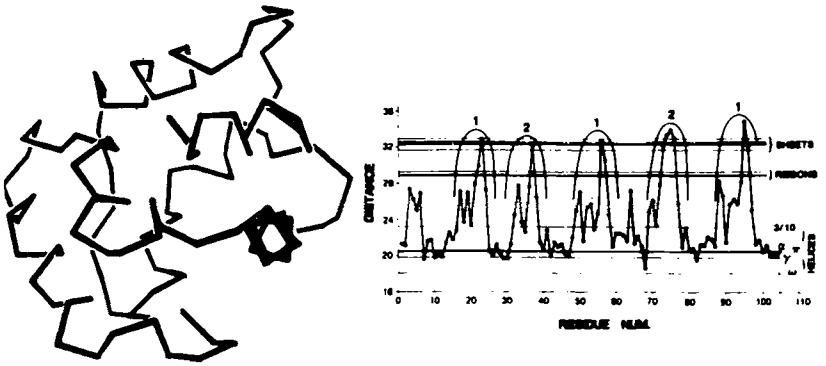

Fig. 1. Linear distance plot (LDP) of calcium-binding parvalbumin and the alpha carbon representation of the same protein. The neighborhood has been selected to include a sum of 4 distances [see Ref. 3] and reveals the apparent structural repeats termed 'E-F hand' which are responsible for calcium binding. Also noted are the two types of apparent structural repeats as exhibited from linear distance plot analysis. The horizontal lines represent similarly computed LDP values for ideal homopolymers of equal length to the protein shown.

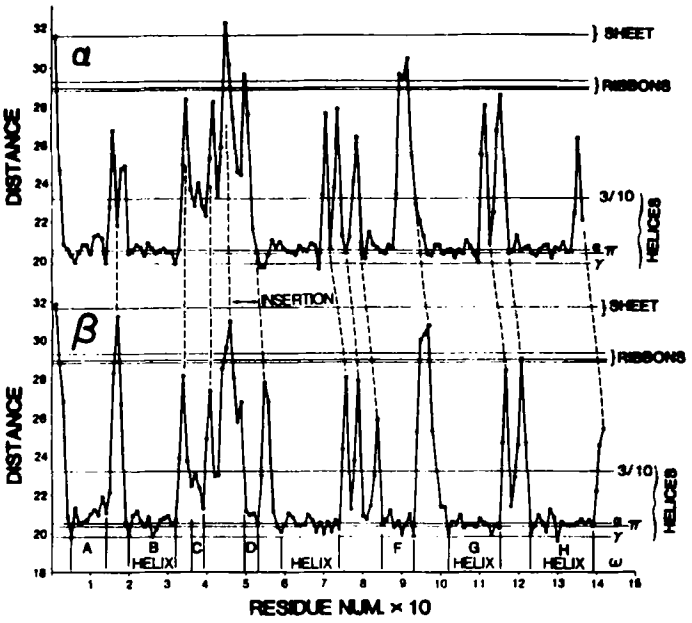

Fig. 2. Linear distance plot contrasting the alpha and beta chains of deoxyhemoglobin, computed as in Fig. 1. Indicated in this figure are the regions of structural analogy which have been matched independently of amino acid sequence comparison (i.e., on structure alone). Evident from the figure is the presence of considerable amounts of alpha helix in both forms, the site of the helix insertion which occurs and lengthens beta vs. alpha hemoglobin, and that the C-helix is not of the alpha helix conformation but rather the 3_{10} helix conformation.

independent amino acids in a manner which averages over a fixed number of consecutive amino acids and plotted as a single value for each residue in turn.

Advantages: can be generated from sequence alone for use in correlation with structure when available; neighborhood sizes and property scales can be modified or calibrated to assist in correlation with structural data; properties can include polarity hydrophobicity, bulk size, etc.

Disadvantages: structure-function correlation of these properties is not etablished but is an active area of research.

Two-dimensional

(a) Distance matrix (structure) [1–3,11–12] (Fig. 3)

Distance matrix analysis generates a square symmetric matrix of order n, where n is the number of amino acids in the protein to be represented. Each element of the matrix, i-j, contains the distance between the alpha carbon of residue i and of residue j, and the resultant matrix is invariant to rotation and/or translation of the protein structure. Shading of the matrix within preset distance ranges highlights secondary, tertiary and domain structure within the protein and also permits visual comparison of two proteins without requiring superposition.

Advantages: independent of rotation and/or translation of molecule; inherent organization of

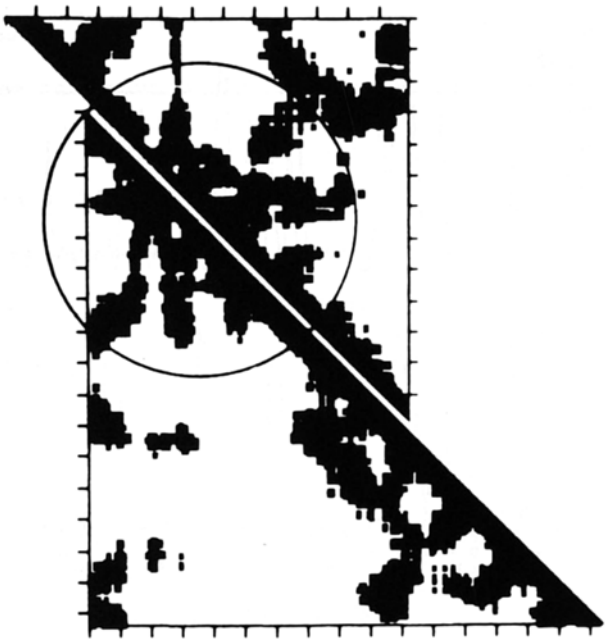

Fig. 3. Distance matrix comparison of lysozyme from hen egg white and from bacteriophage T4, with the separate half-matrices (hen egg white, upper; T4, lower) placed to emphasize the structural homology between the two enzymes [see Ref. 15]. These matrices have been contoured such that only those distance values less than 15.0 Å in magnitude are displayed as filled features. Note that the apparent similarity in the matrices extends beyond the near-diagonal elements, to the tertiary structure region, and beyond the encircled 80-residue segment which has been previously matched by rotation-translation search procedures [see Ref. 15].

secondary, tertiary and quaternary structure; identification of structural domains, palindromes; rapid visual comparison of proteins.

Disadvantages: loss of handedness; computational (vision) analysis less rapid than visual analysis.

(b) Energy matrix (function) [8,13]

Energy matrix analysis involves the use of the two-dimensional matrix representation form and an algorithm which separately examines the dipole-dipole, charge-charge and charge-dipole interactions within a protein whose three-dimensional structure is available from X-ray crystallographic study. This analysis first involves deconvoluting a protein into its constituent peptide and side-chain dipoles using the vector addition of the individual bond moments for each of the amino acids. The matrix is constructed to contain interaction energy terms which separately examine contributions from the main chain and side chains.

Advantages: can separately represent main-chain and side-chain contributions; correlate structure with function by comparison with actual distance matrix; predict regions of stability or instability, particularly with respect to evolution or site-directed mutagenesis, and organized into secondary, tertiary and quaternary structures; can represent stages along dynamics trajectory.

Disadvantages: relies on parameters such as internal dielectric constants; represents only single conformation of potentially dynamic state.

Three-dimensional

(a) van der Waals surface (structure) and molecular electrostatic potential (function) [3,6,8]

Computation of the physicochemical properties: (1) the molecular electrostatic potential surface is evaluated using the partial atomic charges assigned from quantum mechanical computations on a library of amino acid structures and computing the potential experienced by a point charge as it is moved about the region outside of the protein surface; and (2) the van der Waals surface is evaluated by examination of a three-dimensional cubic grid and assignment of each grid element as to its occupancy or vacancy.

Advantages: represents shape and/or properties of structure rather than simple Cartesian coordinate and bond set.

Disadvantages: selection of appropriate partial atomic charges and dielectric; only representative of single conformation.

(b) Cartesian representation of atoms and bonds (structure) (Fig. 1)

Our implementation of interactive, three-dimensional graphics involves its use primarily as a means of communication between the different scientist-users of the computational environment under development.

STRUCTURAL COMPARISON

The methods for structural comparison can be readily derived from the methods for structure-function representation presented above, and include:

(a) Difference linear distance plot analysis (structure) [3,5,7,10,13–14]

Difference linear distance plot analysis uses the representation of the linear distance plot of each of two related forms of a protein (e.g., zymogen and active enzyme) to reveal details of the regions of conformational difference. As this approach does not require structural superpositioning, it does not obscure the localized structural perturbations that relate to function and which may not be apparent from comparison averaging techniques.

(b) Partitioned distance matrix analysis (structure) [1–4]

Partitioned distance matrix analysis uses the parsed regions of secondary, supersecondary, tertiary and quaternary structure as typically derived from the linear distance plot, to form submatrices of the distance matrix as computed independently for each of two proteins. These partitions are compared between the two proteins by evaluation of both the signed sum of the distance pairs and the absolute sum of the distance pairs for the entire partition, as well as between partitions. Thus two molecules can be compared within only limited regions (e.g., nucleotide-folding domain), and with respect to the various levels of structure within that domain as well, rather than over the entire protein.

(c) Structural superposition (structure) [1–4,10,13]

Structural superposition incorporates a statistically refined structural equivalence refinement procedure and is useful in evaluating the global characteristics of structural similarity as is typically monitored by the RMS deviation.

OTHER MACROMOLECULAR DESCRIPTORS

In addition to the methods for structure-function representation and comparison, other aspects of analysis are of interest in studying macromolecular systems:

(a) Protein substructure library

A library is being generated which contains substructures of contiguous folding patterns of the proteins which are available from X-ray crystallographic analysis, and as computed from the linear distance plots. A dynamic programming algorithm (Williams and Liebman, submitted for publication) has been used to identify such substructures which exist and can be identified independent of any bias of a search for known structural templates. The statistical determination of the existence of a substructure, which includes the capability for expansion or contraction of a structural feature, is based on techniques which have been developed for analysis in areas independent of any knowledge of the existence of proteins.

This analysis is being continued to incorporate the linear property plots described above, for use in sequence/property-structure correlation and for potential incorporation within a structure prediction algorithm. In addition, these substructures are also being used to describe the tertiary structure of known proteins, in terms of the orientation between substructures which comprise the library. The potential for successful development of such libraries comes from the observation that the substructures generated in this manner, which contain the traditional conformations of alpha helices, beta sheets and turns, span a greater extent of the observed protein conformations than previously expressed by the limited state models that are presently in use.

(b) Other methods

Bulk property analysis [3,12]. The computation of bulk properties (i.e., properties which may be computed for a structurally-defined region of protein sequence, such as bulk (or weighted hydrophobicity) has yielded several interesting observations but remains a phenomenological observation rather than mechanistic correlate (see Table 4). Further evaluation and analysis of these metrics is underway.

Minimum base change (MBC) per codon and observed mutation rates (PAM) [11–13]. Segments of proteins which have been topographically aligned, using the methods described above, can be evaluated in terms of sequence relatedness in the regions of topographical equivalence (see Table 3).

(c) Object-oriented descriptors of proteins

The goal of developing the representation of proteins in terms of an object-oriented construct, which permits greater access to the information content of these macromolecules, has prompted our incorporation of a wide range of descriptors within our database. These include, but are not limited to: protein family; organism; organ source; in vitro specificity; in vivo specificity; partici-

TABLE 4
ANALYSIS OF BULK HYDROPHOBICITY IN PYRUVATE KINASE, BY DOMAIN

Region	Sequence	Cat (m1) hyp[a]/hyd[b]	S[c]	Chick (m1) hyp[a]/hyd[b]	S[c]	Rat (m2) hyp[a]/hyd[b]	S[c]
n-domain	10–42	12/21	0.4	12/21	0.6	13/20	1.5
A1-domain	43–115	34/39	−10.0	36/37	−4.9	35/38	−10.7
A2-domain	224–387	85/79	24.5	82/82	21.3	77/87	−3.7
B-domain	116–223	57/51	−5.7	57/51	−4.4	59/49	−4.5
C-domain	388–530	63/80	−19.2	63/80	−11.2	66/77	−8.3
		64/79	−11.2				
TOTAL	10–530	251/270	−10.0	250/271	1.4	250/271	−25.7
		252/269	− 2.2				

[a]hyp: hydrophilic
[b]hyd: hydrophobic
[c]S: sum

pation in cascades and/or metabolic pathways; structural definitions including primary, secondary, supersecondary domains, tertiary, substructure library components, quaternary structure; physicochemical properties including spectroscopy (e.g., circular dichroism, NMR, FT-IR), van der Waals surfaces, molecular dipoles, properties as computed from averages over amino acid sequence neighborhoods; organization of the gene in terms of introns/exons; and other user-defined objects whereby a computational algorithm may be developed/utilized to generate the descriptor for incorporation within the knowledge base. These descriptors serve as the pointers for the examination of observations for potential incorporation as generalizable rules.

EXAMPLES OF THE ANALYSIS

1. Serine proteases

We have previously described results from our ongoing studies of the relationship between structure, function, biochemical activity and biological activity in the serine proteases [10,13], and present a summary of these results in terms of the concepts presented in this report. The problem we have been addressing in these studies is the role of the structure-function relationship in the serine proteases within the processes of physiological control that we term limited proteolysis.

The data/information contained in the system can be treated in terms of its 'classification', a term which is used here to denote the association of particular methods for representation, analysis and comparison that might be restricted to particular data/information content.

Data/information content

Structural data: This structural data would include all the available X-ray crystallographic structures as might be found in the Protein Data Bank [17]. Included in this data set would be the different proteases (e.g., trypsin, chymotrypsin, elastase, kallikrein); the zymogen and active enzyme forms; the inhibited enzyme forms, both in vivo and in vitro; the pH- and chemically-modi-

fied forms; the heavy atom binding sites from preliminary phase determiniations.

Sequence: Amino acid sequence data is available for a wide range of enzymes of the serine protease family, from both eukaryotic and prokaryotic sources. In addition, there are also the amino acid sequences for the naturally occurring protein inhibitors which have been isolated from both animal and non-animal sources. Several genes which code for the enzyme sequences of trypsin, elastase and chymotrypsin, as well as some of the coagulation factors for which no three-dimensional structure exists, have been determined [18].

Biochemical: Specificity data are available for many of the serine proteases as measured against small, synthetic substrates, but must be distinguished from the comparable assays that have been performed against natural, macromolecular substrates as is observed in limited proteolytic function.

Biological: Participation of the serine proteases in enzyme cascades (e.g., coagulation, fibrinolysis, complement activation, zymogen activation) is noted as a common theme of the biological activity termed limited proteolysis [19]. In series of limited proteolytic processing of macromolecular substrates, single bonds are cleaved to yield active products that, in turn, are involved in the cascade processes. Most notably, the interactions between enzyme and substrate/inhibitor proteins appear to be highly specific and reflective of advanced evolutionary development of these processes.

Genetic: The involvement of many of the serine proteases within proteolytic cascades is suggestive that the common origin of these molecules, and their resultant structure, are evolved to promote this activity. The changes which are observed in the amino acid sequence within either the eukaryotes or prokaryotes appear to be evidence of homologous origins, with the evolution of the structure-function relationship as yet unknown when one differentiates between biochemical (in vitro) and biological (in vivo) specificity.

Physicochemical: Measurement of pH profiles of enzyme activity, aggregation properties (i.e., dimerization and autolysis [10], spectroscopy (e.g., fluorescence polarization studies [8]), circular dichroism, and FT-IR [14] are readily available for representative enzymes of many of the cascade systems.

RESULTS OF THE ANALYSIS

The results of the application of the methods for representation, comparison and analysis, as described above, and as applied to understanding the role of the serine proteases in limited proteolysis, are summarized here [10,13]:

Identification of a macromolecular recognition surface

Identification of a macromolecular recognition surface (MMRS) as topographically constructed in the eukaryotic serine proteases and identified in terms of topographically analogous (85%) and nonanalogous (15%) regions of the enzymes. This is indicative of a pseudo-hypervariable region in the serine proteases which extends as approximately 10-loop regions protruding above the traditional active site. This region controls access by macromolecules to the active site [10,13].

Analysis of amino acid sequence insertions/deletions and structure

Correlation of the patterns of amino acid sequence insertions and deletions within the topo-

graphically nonanalogous regions (Table 2) of the serine proteases, suggests that these evolutionary differences are amplified by means of the tertiary fold of the protein [10,13].

Structure-function relationships beyond the active site

Identification of patterns of structural perturbation within the tertiary structure beyond the active site of the serine proteases appear to correlate with specific activity within the active site. Thus the conformational changes outside of the active site are more pronounced, yet predictably similar, when the enzyme interacts with natural, macromolecular inhibitors, and are less significant and with little similarity when the enzyme is challenged with small, active, site-directed inhibitors. This is potentially indicative of the source of the strong binding which is observed to differentiate between these two classes of natural and synthetic inhibitors [10].

Macromolecular symmetry in formation of enzyme complexes

Observation of the apparent pseudo-twofold symmetry within the MMRS which accompanies binding of inhibitor macromolecules may be a requisite component of recognition [10].

Macromolecular recognition, zymogens, transport and limited proteolysis

The major difference between the smaller, prokaryotic serine proteases and the eukaryotic proteases, which differ in size by approximately 20%, occurs within the region we have defined as the MMRS. This is significant since the MMRS would appear to be directly responsible for zymogen activation following transport, as well as limited proteolytic specificity in physiological processes which occur predominantly in the eukaryotes. It is also noteworthy that the evolution of natural inhibitors in the prokaryotic systems appears to utilize this difference in specificity (i.e., toward the P2-P1-S1-S2). Small peptide inhibitors in the prokaryotes are directed toward this subsite rather than toward the macromolecular recognition surface [10].

Directional component of macromolecular recognition at long range

Observation of the apparent MMRS-directing force as an electrostatic component which is expressed in the eukaryotes as an anomalous distribution of charged amino acids over the MMRS region, thus providing a long-range directing force towards the MMRS. This is further evidenced by the analysis of the positions of the metal ions used in the X-ray crystallographic phase determination and which serve as suitable and independent probes of the effective electrostatic nature of the protein and its surface. Analyses of these metal binding sites reveal that throughout all of the eukaryotic proteases they occur predominantly within the MMRS [10].

Non-active, site-directed, inhibitor design

We have further observed that the macromolecular recognition capabilities of trypsin can be modulated by binding of a compound in a region of the MMRS, approximately 20Å distant from the traditional active site (Liebman, Kumosinski and Brown; Buono and Liebman; unpublished results). This inhibition occurs solely at the level of macromolecular recognition, preventing the process of trypsin-trypsin recognition and autolysis, while not blocking the active site from being functional toward small, synthetic inhibitors. This observation serves as the basis for the design of serine protease-specific, non-active, site-directed inhibitors which should exhibit no cross-reactivity with other common serine proteases (i.e., side effects) and should prove potentially suitable for rational drug design (Buono and Liebman, unpublished results).

Prokaryotic versus eukaryotic serine proteases

The serine proteases from prokaryotes bear significant topographical homology with those from eukaryotes, but exhibit similarities among themselves which differ from those which correlate with the eukaryotes. Thus regions of large insertions and deletions signify the differences between the two classes of serine proteases, most notably in that regions of the MMRS of the eukaryotes are located within the large regions deleted in the prokaryote enzymes. By contrast, several regions where these gaps occur in the prokaryotic proteases bear strongly conserved similarity among the SGA, SGB and ALP residues. Most gaps indicate that the eukaryotic proteases are larger, containing more amino acids than the proteases found in prokaryotes, except for a region bordered by residues 70-90 in the prokaryotes which is larger than that found in the eukaryotes and is also apparently conserved [13] (Table 3).

Difference between sequence analogy and structural analogy

Analysis of the data in Table 3 reveals that a difference exists between an evaluation of structural analogy and amino acid sequence analogy. This is particularly evident in that regions of topographically mapped residues may show great variability in amino acid sequence as monitored by either the MBC or PAM methods. Perhaps most significant is the observation that the PAM and MBC evaluations of a particular sequence comparison do not always indicate the same relative goodness of fit (Table 3). This discrepancy bears on the attempts to build analogous proteins from amino acid sequence alignment of potentially homologous proteins whose three-dimensional structure is known. This is of particular relevance in the analysis of proteins with little sequence identity, as we have recently shown in pyruvate kinase that sequence identity of as high as 80-90% (Table 4) may not be sufficient to detail the source of functional differences [11,12]. We have proceeded further into the area of mapping one protein onto another in terms of physicochemical properties derived from the amino acid sequence rather than the identity of the amino acids alone (Williams and Liebman, submitted for publication).

Relationship of gene organization to enzyme function

Examination of the organization of the genes which code for the serine proteases (e.g., Ref. 18] reveals that the sites of introns occur in regions immediately adjacent to those which comprise the macromolecular recognition surface in the eukaryotic serine proteases. This observation is not totally generalizable, but suggests that the occurrence of these introns (i.e., noncoding regions of the DNA) are functionally related, either at the level of protein folding or definition of protein function in the folded macromolecule [10,13].

Analysis of energetic organization of related enzymes

Examination of the energy matrices of the serine proteases reveals that certain component energy terms are absolutely conserved within the analogous segments of the topographically equivalenced serine proteases, while other regions differ considerably depending on amino acid identity. This suggests that the ability to discern the potential effects of site-directed mutagenesis will require a further evaluation of the conformational linkages within these macromolecules, such as the combination of the energetic analysis with the observations of the conformational perturbation throughout the tertiary structure of the eukaryotic proteases, as noted previously [13].

Correlation of structure-sequence-spectra-energetics

Examination of the component energy profile of Trp41 of SGA and SGB [8,13] reveals that these two amino acid sequence analogous residues do not, in fact, experience the same energetic environment because of variations within their interactions with neighboring amino acids, both those adjacent in sequence and those which occur in the local environment as a result of the tertiary structure of the enzyme. These data correlate well with the observed differences in the fluorescence spectra measured for these two enzymes [8] and suggest that such analysis may provide the necessary insight for understanding the source of experimental observations and the means by which such measurements may be incorporated into structure prediction (i.e., folding) algorithms.

As a result of the serine protease analysis, it is readily apparent that not only has the original question been addressed, but many of the observations made in this system exhibit characteristics which might prove to be generalizable to other enzyme systems. This is the goal of the incorporation of these approaches within the artificial intelligence-based environment.

2. Application of the methodologies to analysis of Fourier-transform infrared spectroscopy (FT-IR) [14]

Because of our interest in developing the use of FT-IR for the analysis of the structure of proteins in solution, we have initiated a series of studies using the methodologies described here and integrated the results with the experimental observations carried out in collaboration with Dr. Alan Lipkus of Batelle Laboratories, Columbus, Ohio, and Drs. Michael Byler, Heino Susi and Thomas F. Kumosinski of the Eastern Regional Research Laboratories of the U.S. Department of Agriculture, Philadelphia, Pennsylvania (Prestrelski and Liebman, unpublished results).

Traditional attempts to utilize spectroscopic data for analysis of the secondary and tertiary structure of proteins originate with circular dichroism spectroscopy. Such analysis was limited in the sensitivity of the experimental method and the analytical description of the proteins being studied. The interpretation of such spectra in terms of helical, sheet, turn, and 'random coil' structures appeared adequate for the resolution of the spectra and was most closely related to the consideration of percent composition of each structural component. Even in this form of analysis, the collation of several typical proteins with their structural description and measured absorptive spectra did not provide for an easily assignable set of characteristic spectral bands.

The use of FT-IR has extended the ability to resolve bands within the Amide I, II and III regions, due to high signal-to-noise ratios. This higher resolution holds the promise for correlation with the extended structural descriptors that are available from the enlarged data base of protein structures from X-ray crystallographic study, particularly when substructure libraries are created as described above. The limiting process remains the assignment of (sub)structure to spectral band(s), particularly at this heightened degree of resolvability for each.

To approach this problem, informationally, we have reduced the scope of the assignment from a global to a local situation which makes the solution more tenable. We are now able to quantitatively determine conformational changes within a protein structure upon small changes in environment (e.g., pH, temperature) or upon complexation with other molecules (e.g., inhibitors) through the application of difference linear distance plot analysis and partitioned distance matrix analysis, as described above. The observations of localized regions of conformational change within the protein backbone, as exhibited in the trypsin analysis [10], and the ability to assign

these regions of change to particular substructures within the protein, provide the necessary transition for spectra-structure assignment. The computations which are performed on the available X-ray structures are matched with experimental studies of the FT-IR spectra [14]. The differences between the pertubed spectra and the native spectra are indicative of which bands are affected by the same conformational changes which have been observed crystallographically. Assignments for these particular substructures are then made with the spectral bands that are perturbed. These assignments can be tested against the related serine proteases which exhibit homologous structures to establish the confidence of the assignment. Further extension to nonhomologous proteins which exhibit the same substructures permits an examination of the additional influence of non-analogous tertiary structural interactions of analogous substructures on the assignments. This methodology is recursively applied, successively eliminating correlated substructures and bands, to extend the assignable features to other components of the substructure library.

The application of the information in this manner is developed to successively incorporate directly related and correlatable observations through an integrated use of data reduction methods. This approach is highly dependent on the recognition of interrelationships between existing data/information as is provided by the structure of the data storage as outlined above. Extension of the approach to other forms of experimental data analysis are presently being developed.

CONCLUSIONS

We can conclude from these studies, as addressed in terms of the information content of the problem, that although some observations might prove unique to the particular system under study, a number of these might be potentially generalizable to other protein systems. This generalization could arise because of relationships among the proteins that may be at the structural, functional, evolutionary or organizational level, not solely because of a homologous origin. Therefore it has become an important tool in our analysis to utilize the conceptual descriptors (i.e., object-oriented descriptors available in the realm of artificial intelligence analysis) for initiating relationships which are probed by the computational environment to search for potential rules.

The methodologies and concepts which we have reported here appear to be of general applicability in a wide variety of systems. The computational environment which supports their integration is being developed to continue the potential for interaction of both the applications-user and tool-builder within a host which supports both, and which is capable of learning and applying the knowledge gained from both. Development continues within a Fortran/LISP VAX/SUN-oriented system, with an early prototype being developed in collaboration with ImClone Systems Inc. and Bolt, Beranek and Newman Inc. on a Symbolics processor.

ACKNOWLEDGEMENTS

The author would like to acknowledge the partial financial support of ImClone Systems Inc., and the scientific contributions of Dr. A.L. Williams Jr., Dr. D. Gibson, Mr. S.V. Amato, Mr. R. Buono and Mr. S. Prestrelski. He regretfully acknowledges the untimely passing of one of his collaborators, Dr. Heino Susi.

REFERENCES

1 Liebman, M.N., In Balaban, M., Sussman, J. and Yonath, A. (Eds.) Structural Aspects of Recognition and Assembly in Biological Macromolecules, Balaban ISS, Brooklyn, 1981, pp. 147–149.
2 Liebman, M.N., In Griffin, J. (Ed.) Molecular Structure: Biological Activity, Elsevier, New York, 1982, pp. 193–214.
3 Liebman, M.N., Venanzi, C.A. and Weinstein, H., Biopolymers, 24 (1985) 1721–1758.
4 Liebman, M.N., Prog. Clin. Biol. Res., 172B (1985) 285–299.
5 Liebman, M.N. and Weinstein, H., In Clementi, E., Corongiu, G., Sarma, M.H. and Sarma, R.H. (Eds.) Adenine Press, Schenectady, 1985, pp. 339–359.
6 Weinstein, H., Liebman, M.N. and Venanzi, C.A., In Makriyannis, A. (Ed.) New Methods in Drug Research I, J.R. Prous, Ft. Lauderdale, 1984, pp. 233–246.
7 Liebman, M.N., J. Cell. Biochem., 9B (1985) 132.
8 Liebman, M.N. and Prendergast F.G., Biochemistry, 24 (1985) 3384.
9 Pozsgay, M., Michaud, C., Liebman, M.N. and Orlowski, M., Biochemistry, 259 (1986) 1292–1299.
10 Liebman, M.N., Enzyme, 36 (1986) 115–140.
11 Conselor, T.G., Uberacher, E.C., Bunick, G.J., Liebman, M.N. and Lee, J.C., J. Biol. Chem., 263 (1988) 2794–2801.
12 Conselor, T.G., Liebman, M.N. and Lee, J.C., J. Biol. Chem., submitted.
13 Liebman, M.N., J. Ind. Microbiol. (1988) in press.
14 Prestrelski, S., Lipkus, A.H. and Liebman, M.N., Biophys. J., 53 (1988) 299A.
15 Bajaj, M. and Blundell, T., Ann. Rev. Biophys. Bioeng., 13 (1984) 453–492.
16 Cox, J.M., J. Mol. Biol., 28 (1967) 117–156.
17 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535–542.
18 Yoshitake, S., Schach, B.G., Foster, D.C., Davie, E.W. and Kurachi, K., Biochemistry, 24 (1985) 3736–3750.
19 Neurath, H., In Reich, E., Rifkin, D.B. and Shaw, E. (Eds.) Proteases and Biological Control, Cold Spring Harbor Conferences on Cell Proliferation, Vol 2, Cold Spring Harbor, 1975, pp. 51–64.