Journal of Industrial Microbiology, 3 (1988) 127–137
Elsevier

SIM 00117

Analysis of the biomacromolecular architecture of eukaryotic and prokaryotic serine proteases

Michael N. Liebman

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

Received 14 September 1987
Revised 18 December 1987
Accepted 21 December 1987

Key words: Serine protease; Limited proteolysis; Biomacromolecular architecture; Genetic engineering

SUMMARY

We have been developing computational approaches to increase our ability to analyze the growing body of three-dimensional structural data with applications centered on the serine proteases and their natural inhibitors and substrates. It is essential that these approaches emphasize the comparison of these macromolecules at the separate levels of secondary, tertiary and quaternary structure. We assume in our analysis that in functionally related macromolecules (i.e., a family of evolutionarily related enzymes), regions of structural and/or physicochemical similarity will exhibit functional similarity; regions that are different in structure and/or physicochemical properties will function differently and, therefore, be the source of observed specificity. It is the intent of our research to encapsulate such 'knowledge' in a form which is capable of observing patterns which may serve as generalizable rules for macrostructural analysis (Liebman, M.N. 1986. Enzyme 36: 150–163), and to serve as the essential 'tools' for the rational design of modified serine proteases and/or their natural inhibitors by the methods available through genetic engineering.

INTRODUCTION

Recent advances in biotechnology have led its practitioners to return to an examination of the fundamental questions concerning the relationship among amino acid and nucleic acid sequences, and structure and function at the macromolecular level. These developments in the technology of genetic engineering make it increasingly possible to isolate genes, perform site-directed mutagenesis and have the 'synthetic' protein expressed in significant quantities, yet the essential information as to which changes are needed to yield desired results (e.g., enhanced protein stability, modification of specificity, etc.) remains elusive. Our long-standing interest in both the theoretical and applied aspects of bio-macromolecular architectural analysis have led to our examination of the serine proteases, with particular emphasis on their role in physiological control through limited proteolysis. This family of enzymes

Correspondence: M.N. Liebman, Department of Physiology and Biophysics, Mount Sinai School of Medicine of the City University of New York, New York, NY 10029, U.S.A.

is accepted to be evolutionarily related, possibly through gene duplication, and yet its members exhibit substrate and inhibitor specificity directed towards macromolecules, at a level of regulation which makes these enzymes essential components of such highly refined cascades as blood coagulation, fibrinolysis, complement activation, and milk clotting, as well as processes such as fertilization and hormone production. The importance of this enzyme family is also indicated by their suggested role in emphysema, arthritis and septicemia as caused by bacterial toxin production [8,22,23].

The process of limited proteolysis is characterized by the ability of the enzyme to complex with a specific macromolecular substrate, cleaving only one or two predetermined peptide bonds (as distinguished from the complete digestion of a peptide substrate) and releasing the product for subsequent activity within the physiological system. This specificity appears to be highly refined among the eukaryotes where the occurrences of zymogen activation following their production and transport are examples of such processes. Such a level of specificity exists among the eukaryotic serine proteases, in contrast with the structural and mechanistic homology which they exhibit independent of organ or organism source. Thus a key to the design of specific inhibitors or the modification of known enzymes or inhibitors of this family is the determination of the relationship between the architecture of these molecules and their specificity and reactivity determinants. An important aspect of this analysis comes from our identification of the source of specificity towards macromolecular substrates and inhibitors as a topographical feature, the macromolecular recognition surface (MMRS), and the differences in macromolecular specificity observed between the eukaryotic and prokaryotic proteases [16].

We have been developing computational approaches to increase our ability to analyze the growing body of three-dimensional structural data with applications centered on the serine proteases and their natural inhibitors and substrates. It is essential that these approaches emphasize the comparison of these macromolecules at the separate levels of secondary, tertiary and quaternary structure. We assume in our analysis that in functionally related macromolecules (i.e., a family of evolutionarily related enzymes), regions of structural and/or physicochemical similarity will exhibit functional similarity; regions that are different in structure and/or physicochemical properties will function differently and, therefore, be the source of observed specificity. It is the intent of our research to encapsulate such 'knowledge' in a form which is capable of observing patterns which may serve as generalizable rules for macrostructural analysis [16], and to serve as the essential 'tools' for the rational design of modified serine proteases and/or their natural inhibitors by the methods available through genetic engineering.

In this study we extend our previous analysis [16] to include the available data on prokaryotic serine proteases (e.g., isolated from *Streptomyces griseus* and *Myxobacter* 495) [9,26,28], and to examine the relationship between the organization of the genes which code for the serine proteases and our observations concerning the MMRS. In addition we report on the application of a component electrostatic energy analysis which permits the correlation of the three-dimensional structure of a protein, as determined by X-ray crystallography, with spectroscopic measurements performed in solution studies (i.e., fluorescence), and which provides insight into the difference between sequence, structure and functional analogies.

DATA

The three-dimensional structures of the proteins used in this study are available as atomic coordinates provided by the Protein Data Bank (PDB) at Brookhaven National Laboratory [3]. The members of the serine protease family which previously have been studied have been tabulated [16] and we additionally include the alpha-lytic protease isolated from *Myxobacter* 495 (2ALP) which contains 198 amino acid residues and has had data collected to a resolution of 1.7 Å and has been solved to an *r*-factor of 0.131 [9].

METHODS

In this report we briefly summarize the method of component electrostatic energy analysis [17] and its correlation with fluoresence polarization studies on the proteases isolated from *Streptomyces griseus* termed A (SGA) and B (SGB) [18] which are described for the first time.

Several other methods used in this study have been described in detail elsewhere: distance matrix analysis [13–16,19,20,30], linear distance plot analysis [14,16,19], structural superposition [16,19], topographical mapping [16] and computation of the minimum base change per codon and point after mutation rate for topographically equivalenced structural features [5].

Deconvolution of component electrostatic energy (energy matrices)

Energy matrix analysis involves the use of the distance 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 a screening of the coordinates for completeness and then deconvolution into their constituent peptide and side-chain dipoles (*d*), using vector addition of the individual bond moments for each of the amino acids assigned from the values derived from spectroscopy [24]. Computing the dipole moment for the peptide bond in this manner yields an average value of 3.81 debyes while the dipole moment for the individual side-chains will vary with the observed conformational flexibility of that particular side-chain, e.g., glutamine ranges from 3.03 to 4.11 debyes in native trypsin (TPO). The conformation produced by the presence of a proline residue causes the peptide bond dipole moment to exhibit a computed dipole moment of 2.71 debyes, while *cis*-peptides may have peptide dipoles of 2.2–2.4 debyes. In this manner it is possible to compute the composite dipole moment for the peptide bonds and separately for the side-chain atoms as well as for the overall protein (see Table 1).

Table 1

Summary of composite dipoles in this study

	SGA	SGB	ALP
Peptide	51.2	59.6	56.3
Side-chain	15.8	15.1	25.0
Total	60.0	69.0	35.4

For the purpose of analyzing the organization of the electrostatic components of the potential energy of which the overall molecule is comprised, an algorithm is used which is analogous to that of distance matrix analysis [17,18]. In the representation of the energy profile of the protein and correlation with the three-dimensional structure of the enzyme, we have utilized the organizational format of the distance matrix. The symmetrical upper half of the distance matrix is replaced by a matrix which describes the potential energy calculated for each dipole-dipole (*d-d*), dipole-charge(formal) (*d-q*) and charge(formal)-charge(formal) (*q-q*) for each residue pair in the protein. In addition, because of the deconvolution of the protein into composite peptide dipoles and side-chain dipoles, separate calculations and representations are computed for each to differentiate between the energetics of the polypeptide backbones and the side-chains, thus permitting the separation of contributions to the electrostatic energetics from the homologous folding of the polypeptide backbones and the individual side-chains of each protein. This has been done to permit rapid assessment of the effects of site-directed mutations, or natural evolution, on the conformational pathways which may exist within the enzyme tertiary structure. Standard electrostatic potentials were computed using the dipole-dipole, charge-dipole and charge-charge potentials, and for the purposes of these computations, a single-valued dielectric was used ($e=2$). The limitations of this approximation are being analyzed in a separate study (Liebman and Prendergast, unpublished results). The intent of the present study is to examine the correlation of this computational approach with the experimental results to establish the potential for a quantitative description of computation-

ally accessible properties which may permit the successful prediction of selected protein environments from their three-dimensional structure.

For each of the three proteins analyzed in this study, a summary of the energy terms, computed as described above, are presented below (see Results). Although these comparisons reveal the overall similarity of selected parameters of the energy descriptors of the enzymes being studied, the approach readily affords the ability to examine the energetic environments of each of the residues which is being probed experimentally, namely the sequence-homologous tryptophans of SGA and SGB in position 41, and the unique tryptophan in position 42 of SGB.

Other methods used in this study

(1) Distance analysis [13–20,30] 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 carbons 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 structures within the protein and also permits visual comparison of two proteins without requiring superposition. We have further extended the method of distance matrix analysis to study quaternary structure as well as the patterns of organization governing the interaction between more than one macromolecule [16]. This is achieved by extending the list of alpha carbons representing protein A from a list of length 'n', to 'N', where $N = n + n^*$, and 'n^*' is equal to the number of alpha carbons in protein B (the protein which is complexed with protein A). Thus, C-alpha(1) of protein B becomes element C-alpha($n+1$) in the augmented list, etc., and the newly constructed distance matrix is an $N \times N$ matrix. This resultant matrix can be partitioned into three major subpartitions containing, respectively, the distance matrix of protein A, the distance matrix of protein B, and the interaction subpartition which describes the relative orientation of protein A to B as observed within the complex A-B.

(2) Linear distance plot analysis [15,19,20] is based on a plot of the sum of the series of distances from the origin alpha carbon to each of four successive alpha carbons. The plot is generated by using each successive amino acid in the protein sequence as an origin for the computation. 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) as well as to enable the structural comparison of two proteins [16].

(3) Structural superposition [14,19] incorporates a statistically refined structural equivalencing procedure and is useful in evaluating the global characteristics of structural similarity as is typically monitored by the root-mean-square deviation (RMS).

(4) Computation of the minimum base change per codon (MBC) and point after mutation rate (PAM) proceeds as has been described in our analysis of pyruvate kinase [5], but is modified to account for the structural alignment which has been carried out between 2ALP, SGA and SGB. The first method evaluates the parameter MBC by comparison of each pair of amino acids, between two proteins, and extracting the appropriate element from a matrix comparing amino acids which contains the minimum number of bases which would have to be altered to transform any codon of amino acid from protein 1 to that amino acid found in protein 2. The sum and average of these values over a subsequence or the entire protein structure present an evaluation of the similarity of the sequences if they were to potentially undergo mutation/evolution by the simple process of such minimum base change, regardless of selection of codon(s) necessary to minimize this function. The second analysis is the Accepted Point Mutation Method of Dayhoff [5], which is based on the statistical frequency of amino acid mutations if a normal distribution of mutations is assumed. The table of values used is derived from the relatedness odds matrix for an evolutionary distance of 256 accepted point mutations per 100 amino acid links, and has been used primarily to search for similarity in distantly related proteins. We utilize it, in this study, as a representation of

the observed sequence substitution, potentially distinct from the MBC method described above, because the mechanism for the observations is not implicit in the derived values. As noted above, this study involves the comparison of proteins of non-equivalent numbers of amino acids, and the analysis of both the MBC and PAM values has been carried out for only those segments of two proteins whose three-dimensional structures have been established to be equivalent by the topographical mapping procedure.

(5) Topographical mapping of the three-dimensional structure of one protein onto another protein [16] identifies regions of structural similarity regardless of their relative positions within their respective amino acid sequences (i.e., independent of size and number of insertions and deletions in the sequence). This algorithm initially compares secondary structural matching using the linear distance plot, and then proceeds to examination of the distance matrix and partitioned distance matrix [14,19] of the two proteins under study. A non-crossover rule assures that regions being mapped appear in the same order within the polypeptide chains. Convergence is established by a statistical test which compares the distribution of the mapped coordinates with that of a statistically random sample of three-dimensional data points.

RESULTS

Topographical mapping

The results of the topographical mapping of SGA onto TPO, SGB onto SGA, and ALP onto SGA are presented in Table 2, which also includes the comparison of the amino acid sequences which have been brought into coincidence by the topographical mapping procedure. The analysis of the amino acid sequence comparison, based on those amino acids which are topographically equivalent only, has been computed as both the MBC and the point mutation values. The sequence numbering which appears for a particular protein uses the same convention of our earlier study of serine proteases, namely a sequential numbering from residue 1 to 'n', without the common bias imposed by a sequence relationship drawn on a comparison with chymotrypsinogen [16]. In addition, Fig. 1 shows a linear display of the topographically aligned structures with an additional alignment showing the relationship among the prokaryotic proteases SGA, SGB and ALP, and the mapping of SGA onto TPO.

Table 2

Topographical superposition of serine proteases and evaluation of superimposed amino acid sequences

*n*1 start	*n*2 start	No.	MBC	MBC cum.	PAM	PAM cum.
SGA vs. TPO, RMS = 1.48 Å/106 residues						
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						
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. 2ALP, RMS = 0.60 Å/144 residues						
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

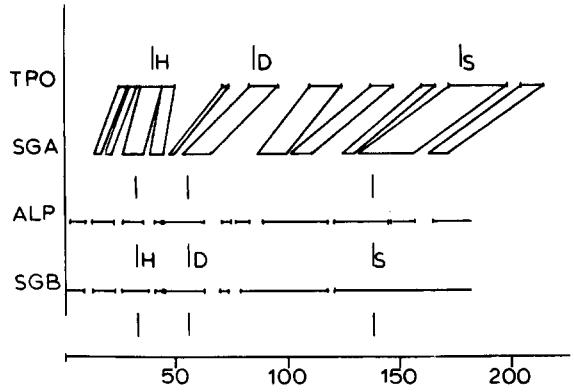

Fig. 1. Topographical mapping of prokaryotic serine proteases onto eukaryotic serine proteases. An alignment of the residues of TPO which have been made topographically equivalent with residues on SGA, revealing regions of non-alignment and the positions of the active site residues, His-40 (H), Asp-85 (D) and Ser-177 (S). The reverse alignment of TPO onto SGA is also shown to permit the additional comparison of the alignments of SGB onto SGA and ALP onto SGA.

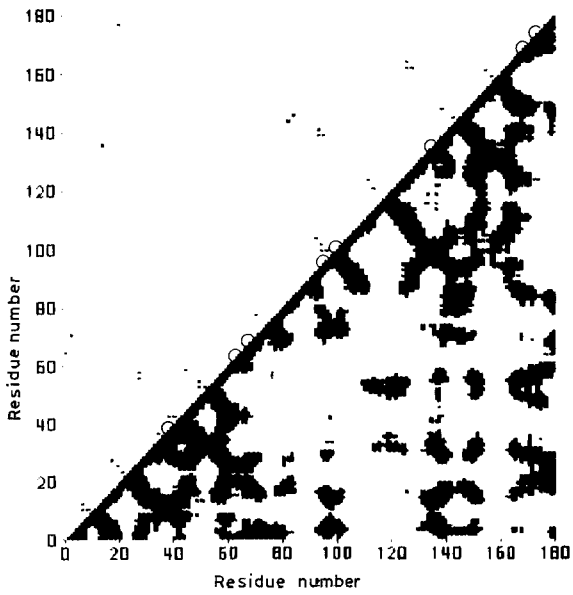

Fig. 2. Energy matrix for SGA superimposed on the distance matrix of the protein. The energy matrix represents the dipole (peptide)-dipole(peptide) interaction energy in the upper half-matrix, contoured to exhibit both stabilizing and destabilizing (encircled) interactions. The distance matrix is shown in the lower half-matrix and is contoured at the single level to indicate those distance pair elements which are less than 15.0 Å.

Energy matrix analysis

The energy matrices have been computed for each of the deconvoluted energy terms, for each of the proteases in this study, and a representative plot of the dipole(peptide)-dipole(peptide) energy map for SGA is shown in comparison with its distance matrix in Fig. 2. A summary of the deconvolved energy terms, as summed through the interactions throughout each of the proteases, is given in Table 3. In addition, Fig. 3 reveals a plot of several of the energy parameters for the sequence and topographically equivalent tryptophan of SGA and SGB, represented as interactions with each residue of the remainder of the protein, to permit comparison of the sequence homology, structural homology and

Table 3

Summary of energy deconvolutions (unscaled energy values)

	SGA	SGB	ALP
d(pep)-*d*(pep)			
min	−0.41	−0.47	−0.50
max	0.40	0.38	0.38
total	−15.1	−14.0	−18.5
d(sc)-*d*(sc)			
min	−0.21	−0.18	−0.15
max	0.14	0.18	0.28
total	0.50	0.46	−0.77
d(sc)-*d*(pep)			
min	−0.41	−0.41	−0.39
max	0.37	0.37	0.31
total	−8.71	−3.94	−9.7
q(sc)-*q*(sc)			
min	−0.21	−0.20	−0.33
max	0.35	0.34	0.33
total	5.43	9.34	7.70
q(sc)-*d*(pep)			
min	−0.34	−0.78	−1.02
max	0.32	0.32	0.33
total	3.12	1.84	0.11
q(sc)-*d*(sc)			
min	−0.20	−0.33	−0.19
max	0.15	0.31	0.33
total	−1.29	0.31	−0.27

pep = peptide; sc = side-chain.

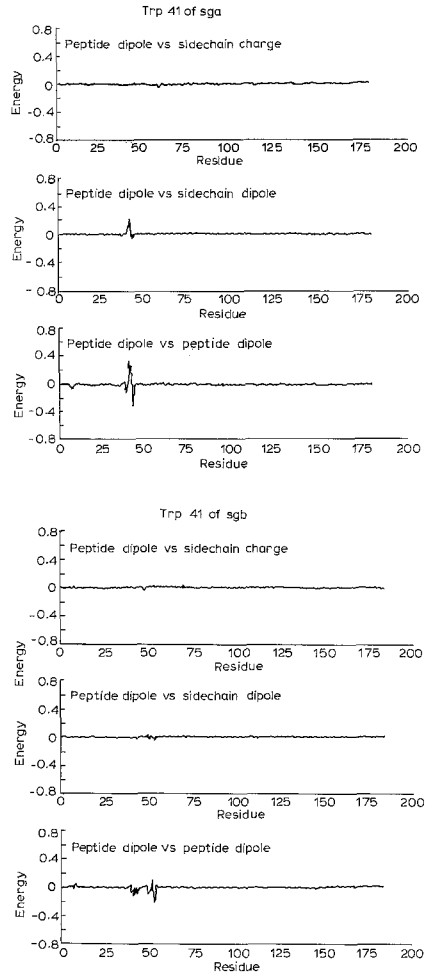

Fig. 3. Energy profiles of Trp 41 from SGA and SGB which have been represented as their respective interactions with all other amino acids in the respective protein structures. The energy profiles shown are for the dipole(peptide)-charge(side-chain); dipole(peptide)-dipole(side-chain); and dipole(peptide)-dipole(peptide) interactions for both of the sequence homologous tryptophans.

functional relationships. The summary of these analyses for the tryptophans of SGA and SGB has been provided elsewhere [18].

DISCUSSION

Our analysis of the relationship between structure and function in the serine proteases [16] concentrated on the specificity directed towards macromolecular recognition among the eukaryotic proteases and yielded the following conclusions which may bear on any attempts to alter the structure of a naturally occurring serine protease or macromolecular inhibitor:

(1) Identification of a macromolecular recognition surface (MMRS) as topographically constructed in the eukaryotic serine proteases and identified in terms of topographically homologous (85%) and non-homologous (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 macromolecule to the active site.

(2) Correlation of the patterns of amino acid sequence insertions and deletions within the topographically non-homologous regions of the serine proteases, suggesting that these evolutionary differences are amplified by means of the tertiary folding of the protein.

(3) Identification of patterns of structural perturbation within the tertiary structure beyond the active site of the serine proteases which 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.

(4) Observation of the apparent pseudo-twofold symmetry within the MMRS which accompanies binding of inhibitor macromolecules and may be a requisite component of recognition.

(5) Identification of the differences between the smaller, prokaryotic serine proteases and the eukaryotic proteases, which differ in size by approximately 20%; the major difference occurs within the region we have defined as the MMRS. This bears significance since the MMRS would appear to be directly responsible for zymogen activation follow-

ing transport, as well as limited proteolytic specificity in physiological processes which occur predominantly in the eukaryotes. It is also of note that the evolution of natural inhibitors in the prokaryotic systems appears to utilize this difference in specificity (i.e., towards the P2-P1-S1-S2). Small peptide inhibitors in the prokaryotes are directed toward this subsite rather than towards the macromolecular recognition surface.

(6) 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.

(7) 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 that is approximately 20 Å 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 towards small, synthetic inhibitors. This observation is serving 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).

What we have determined from our extension of the serine protease studies to encompass the prokaryotic proteases, as reported in this work, are the following:

(1) 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. 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.

(2) Examination of Table 2 reveals that a difference exists between an evaluation of structural homology and amino acid sequence homology. 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. This discrepancy bears on the attempts to build homologous proteins from amino acid sequence alignments 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% may not be sufficient to detail the source of functional differences [5]. 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 acid alone (Williams and Liebman, submitted for publication).

(3) Examination of the organization of the genes which code for the serine proteases [5,6,29] reveals that the sites of introns occur in regions immediately adjacent to those which comprise the macromolecular recognition surface in the eukaryotic ser-

ine proteases. This observation is not totally generalizable but suggests that the occurrence of these introns (i.e., non-coding regions of the DNA) are functionally related, either at the level of protein folding or definition of protein function in the folded macromolecule.

(4) Examination of the energy matrices of the serine proteases, an example of which is shown in Fig. 2, reveals that certain component energy terms are absolutely conserved within the homologous 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 observation of the conformational perturbation throughout the tertiary structure of the eukaryotic proteases as noted previously [16].

(5) Examination of the component energy profile of tryptophan 41 of SGA and SGB (Fig. 3) reveals that these two amino acid sequence homologous 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 measurements for these two enzymes [18] 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.

CONCLUSIONS

Our analysis of the serine proteases from eukaryotic and prokaryotic sources reveal that this family of enzymes, which has most often been thought of as simple, digestive macromolecules, are quite complex in their architectural organization. We have established the occurrence of a macromolecular recognition surface which appears to operate under evolutionary control to amplify small insertions and deletions in the amino acid sequence in terms of their function. This feature is essential for the macromolecular specificity observed in limited proteolysis and distinguishes the eukaryotic and prokaryotic proteases presumably because of the evolved differences in enzyme transport and activation. It is interesting to note that the natural inhibitors of these proteases also appear to have an evolutionary distinction. We have further shown some of the details of the conformational perturbations which accompany binding of natural versus synthetic inhibitors to these enzymes, and their potential involvement in producing the high affinities that are observed in the naturally occurring complexes. The method of energy deconvolution that we have described permits the observation of functional differences between localized regions of these homologous enzymes as well as insight into the correlation of solution physicochemical measurements with the three-dimensional structure available from X-ray crystallographic analysis. This has suggested that these 'primitive' digestive enzymes may exhibit conformational linkages within their macromolecular framework which may be analogous to the pathways necessary for allosteric modulation. In addition, it has been shown that amino acid sequence homology is not strictly the same as structural homology, and the variation in the methods of assessing sequence homology for those regions which are topographically equivalent suggests that differences in PAM and MBC indicators reflect on the importance in understanding the mechanism for amino acid substitution that occurs in nature. An interesting observation is the apparent correlation of intron positions in the genes identified as coding for certain forms of chymotrypsin, elastase and trypsin, with the functional region of these enzymes that we have identified as associated with macromolecular recognition in limited proteolysis.

All of these results from the analysis of the biomacromolecular architecture of the serine proteases reveals how little is absolutely understood about their integrated functionality. It has established the

need for continued computation and experimental probing of this important enzyme family, both for use in rational drug design and biotechnology as applied to the serine proteases, and in terms of the understanding of the structure-function relationship which is basic to all protein systems.

ACKNOWLEDGEMENTS

The author would like to acknowledge the programming and data analysis contributions of Dr. A.L. Williams, Mr. R. Buono and Mr. S. Prestrelski. A generous grant from ImClone Systems, Inc., is also gratefully acknowledged.

REFERENCES

1 Bell, G.I., C. Quinto, M. Quiroga, P. Valenzuela, C.S. Craik and W.J. Rutter. 1984. Structure of the rat pancreatic chymotrypsin B gene. J. Biol. Chem. 259: 14265–14270.

2 Bender, M.L. and F.J. Kezdy. 1964. The current status of the alpha-chymotrypsin mechanism. J. Am. Chem. Soc. 86: 3704–3714.

3 Bernstein, F.C., T.F. Koetzle, G.J.B. Williams, E.F. Meyer, Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi. 1977. The protein data bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. 112: 535–542.

4 Chothia, C.H. and J. Janin. 1976. Stability and specificity of protein-protein interactions: the case of trypsin-trypsin inhibitor complex. J. Mol. Biol. 100: 197–212.

5 Conselor, T., M.N. Liebman and J.C. Lee 1988. Domain interaction in rabbit muscle pyruvate kinase: intersubunit contacts and allosteric switch. J. Biol. Chem. 263: 2794–2801.

6 Craik, C.S., Q. Choo, G.H. Swift, C. Quinto, R.J. MacDonald and W.J. Rutter. 1984. Structure of two related rat pancreatic trypsin genes. J. Biol. Chem. 259: 14255–14264.

7 Craik, C.S., T. Fletcher, S. Roczniak, P.J. Barr, R. Fletterick and W.J. Rutter. 1985. Redesigning trypsin: alteration of substrate specificity. Science 228: 291–297.

8 De Haen, C., H. Neurath and D.C. Teller. 1975. The phylogeny of trypsin-related serine proteases and their zymogens: new methods for the investigation of distant evolutionary relationships. J. Mol. Biol. 92: 225–259.

9 Fujinaga, M., L.T.J. Delbaere, G.D. Brayer and M.N.G. James. 1985. Refined structure of alpha-lytic protease at 1.7A resolution: analysis of hydrogen bonding and solvent structure. J. Mol. Biol. 184: 479–494.

10 Greer, J. 1980. Model for haptoglobin heavy chain based on structural homology. Proc. Natl Acad. Sci. USA 77: 3393–3397.

11 Hartley, B.S. 1970. Homologies in serine proteases. Phil. Trans. R. Soc. Lond. Ser. B 257: 77–87.

12 Ingles, D.W. and J.R. Knowles. 1967. Structure and stereospecificity of alpha-chymotrypsin. Biochem. J. 104: 369–377.

13 Liebman, M.N. 1980. Quantitative analysis of structural domains in proteins. Biophys. J. 32: 213–217.

14 Liebman, M.N. 1985. Topographical analysis of specificity in chemotherapeutic systems. Prog. Clin. Biol. Res. 172B: 285–299.

15 Liebman, M.N. 1985. Distance approaches to protein structural analysis and prediction. J. Cell. Biochem. 9B: 132.

16 Liebman, M.N. 1986. Structural organization in the serine proteases: macromolecular recognition in limited proteolysis. Enzyme 36: 150–163.

17 Liebman, M.N. and T.F. Kumosinski. 1987. Analysis of the structure-function relationship in proteins using two-dimensional energy profiles. Biophys. J. 51: 450a.

18 Liebman, M.N. and F.G. Prendergast. 1987. Electrostatic interactions of the tryptophan residue in *Streptomyces griseus* proteinase A and ribonuclease T1. Biophys. J. 51: 276a.

19 Liebman, M.N., C.A. Venanzi and H. Weinstein. 1985. Structural analysis of carboxypeptidase A and its complexes with inhibitors as a basis for modeling enzyme specificity. Biopolymers 24: 1721–1758.

20 Liebman, M.N. and H. Weinstein. 1985. Heuristic studies of structure-function relationships in enzymes – carboxypeptidase A and thermolysin. In: Structure and Motion: Membranes, Nucleic Acids and Proteins. (Clementi, E., G. Corongiu, M.H. Sarma and R.H. Sarma, eds.), pp. 339–359, Adenine Press,

21 McLachlan, A.D. 1972. Repeating sequences and gene duplication in proteins. J. Mol. Biol. 64: 417–437.

22 Neurath, H. and K.A. Walsh. 1976. Role of proteolytic enzymes in biological regulation. Proc. Natl. Acad. Sci. USA 73: 3825–3832.

23 Ottesen, M. 1967. Induction of biological activity by limited proteolysis. Annu. Rev. Biochem. 36: 55–76.

24 Pethig, R. 1979. Dielectric and Electronic Properties of Biological Materials, John Wiley and Sons, New York.

25 Phillips, D.C., D.M. Blow, B.S. Hartley and G. Lowe. 1970. A discussion on the structure and function of proteolytic enzymes. Phil. Trans. R. Soc. Lond. Ser. B 257: 65–266.

26 Read, R.J., M. Fujinaga, A.R. Sielecki and M.N.G. James. 1983. Structure of the complex of *Streptomyces griseus* protease and the third domain of the turkey ovomucoid inhibitor at 1.8A resolution. Biochemistry 22: 4420–4428.

27 Rose, Z., S.V. Amato and M.N. Liebman. 1984. Analysis of the domain structure of phosphoglycerate mutase. Biochem. Biophys. Res. Commun. 121: 826–833.

28 Sielecki, A.R., W.A. Hendrickson, C.G. Broughton, L.T.J.

Delbaere, G.D. Brayer and M.N.G. James. 1979. Protein structure refinement: *Streptomyces griseus* serine protease A at 1.8A resolution. J. Mol. Biol. 134: 781–793.

29 Swift, G.H., C.S. Craik, S.J. Stary, C. Quinto, R.G. Lahaie, W.J. Rutter and R.J. MacDonald. 1984. Structure of two related elastase genes expressed in the rat pancreas. J. Biol. Chem. 259: 14271–14278.

30 Weinstein, H., M.N. Liebman and C.A. Venanzi. 1984. Theoretical principles of drug action: the use of enzymes to model receptor recognition and activity. In: New Methods in Drug Research (Makriyannis, I., ed.), pp. 233–246, J.R. Prous Publishing Co., Barcelona.