Protein Engineering vol.2 no.3 pp.193–199, 1988

Coordinated amino acid changes in homologous protein families*

D.Altschuh, T.Vernet[1], P.Berti[1], D.Moras and K.Nagai[2]

Institut de Biologie Moléculaire et Cellulaire du CNRS, 15 rue Descartes, 67084 Strasbourg Cédex, France, [1]Conseil National de Recherches Canada, Institut de Recherche en Biotechnologie, 6100 Avenue Royalmount, Montréal, Québec H4P 2R2, Canada and [2]Medical Research Council, Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, UK

*Issued as NRCC BRI no. 26327

In the tobamovirus coat protein family, amino acid residues at some spatially close positions are found to be substituted in a coordinated manner [Altschuh *et al.* (1987) *J. Mol. Biol.*, 193, 693]. Therefore, these positions show an identical pattern of amino acid substitutions when amino acid sequences of these homologous proteins are aligned. Based on this principle, coordinated substitutions have been searched for in three additional protein families: serine proteases, cysteine proteases and the haemoglobins. Coordinated changes have been found in all three protein families mostly within structurally constrained regions. This method works with a varying degree of success depending on the function of the proteins, the range of sequence similarities and the number of sequences considered. By relaxing the criteria for residue selection, the method was adapted to cover a broader range of protein families and to study regions of the proteins having weaker structural constraints. The information derived by these methods provides a general guide for engineering of a large variety of proteins to analyse structure–function relationships.

Key words: protein structure/sequence similarities/protein engineering/proteases/haemoglobins

Introduction

Within the serine protease, cysteine protease and globin families, the tertiary structure is remarkably similar despite considerable variation in the amino acid sequences. Conserved residues or conservative changes alone cannot account for structural conservation. Thus the structures must be stabilized in different ways in related proteins. Knowing which alternative residues produce a given local structure would improve our understanding of protein folding and residue interactions. A modified function of a member of an enzyme family may arise from a small number of substitutions. The ability to predict which particular residues among a large number of substitutions determine functional differences would be particularly useful for protein engineering.

A recent analysis of tobamovirus coat proteins showed that some amino acid replacements in these related proteins are coordinated. Positions in the protein sequence that are replaced together, and therefore show the same replacement pattern in related sequences, are mostly in spatial proximity in the crystallographic structure of the coat protein disk of tobacco mosaic virus *vulgare*. This strongly suggests that mutations are not stabilized independently of each other in the tobamovirus protein family (Altschuh *et al.*, 1987), and that these coordinated mutations may play important roles in stabilizing protein structure and/or in defining functional diversity.

A simple method of finding coordinated changes in homologous sequences and relating these changes with structure has previously been described (Altschuh *et al.*, 1987). In order to investigate whether this method can be applied successfully to protein families other than tobamoviruses, we analysed three additional families with various degrees of sequence similarities, and for which several sequences and at least one crystallographic structure are available: serine proteases (11 sequences with similarities from 22 to 44%), cysteine proteases (12 sequences with similarities ranging from 25 to 95%) and haemoglobin beta chain (12 sequences with similarities ranging from 40 to 97%).

Coordinated amino acid substitutions were found in all three protein families. Optimal conditions for detecting such changes are highly constrained structures and low sequence similarities. A modification of the original method is described for an extended analysis of the pattern of substitutions in less-constrained structures. Results are illustrated by describing some groups of residues that are substituted together in the two protease and the haemoglobin families.

Materials and methods

Source of sequences

Serine proteases. The sequences used were the same as those aligned by Greer (1981): chymotrypsin, trypsin, elastase, haptoglobin heavy chain, kallikrein, factor IX_a (Christmas factor), factor X_a (Stuart factor), plasmin B chain, group-specific protease, thrombin B chain and bacterial trypsin.

Cysteine proteases. The twelve cysteine protease sequences used in this study are: papain (Cohen *et al.*, 1986), actinidin (Praekelt *et al.*, 1988), *Dictyostelium discoideum* cysteine proteases 1 and 2 (Pears *et al.*, 1985), mouse cathepsin L (Troen *et al.*, 1987), rat cathepsin L (Ishidoh *et al.*, 1987a), chicken cathepsin L (Wada *et al.*, 1987), rat cathepsin H (Ishidoh *et al.*, 1987b), aleurain (Whittier *et al.*, 1987) and human, mouse and rat cathepsins B (Chan *et al.*, 1986).

Haemoglobins. Twelve haemoglobin beta chain sequences were selected among various animal groups: human, bovine, dog, rabbit, chicken, duck, starling, ostrich, carp, goldfish, crocodile and alligator. These sequences were extracted from the NBRF data bank where all references can be found.

Source of structures

The crystallographic structures used for this study are from the Brookhaven data bank. They are papain (Kamphuis *et al.*, 1984), chymotrypsin A (Birktoft and Blow, 1972) and human deoxy haemoglobin beta chain (Fermi *et al.*, 1984).

Alignment procedures

Since the purpose of this paper is methodological, the alignments are not presented here. Alignments and structure–function

1. Alignment of four sequences containing eight positions

Residue position		1	2	3	4	5	6	7	8
Residue type in sequence	1	K	S	L	E	C	T	F	S
	2	K	N	L	D	G	T	F	S
	3	H	N	V	T	G	V	M	A
	4	H	S	F	Q	S	V	M	T

2. The substitution pattern (read vertically) at each position is symbolised by four numbers.

Residue position		1	2	3	4	5	6	7	8
Sequence	1	1	1	1	1	1	1	1	1
	2	1	2	1	2	2	1	1	1
	3	2	2	2	3	2	2	2	2
	4	2	1	3	4	3	2	2	3

3. Analysis methods

Stringent Method: search for positions having an identical pattern of amino acid substitutions

Patterns represented by more than one sequence position are :
Pattern 1 1 2 2 Positions 1, 6, 7
Pattern 1 1 2 3 Positions 3, 8

Relaxed methods: search for positions having one type of amino acid in a set of sequences

- Considering some residue types as identical
Example : If E=D, then position 4 has the pattern '1 1 2 3' instead of '1 2 3 4'.
Groups above are modified as follows :
Pattern 1 1 2 3 Positions 3, 4 and 8

- Searching for almost identical patterns (when a large number of sequences is considered)

- Considering residue conservation in all sequence combinations
Examples :
Pattern 1 1 x x Positions 1, 3, 6, 7, and 8
Pattern x 1 1 x Positions 2 and 5

4. Analysis of spatial proximity between two or more positions from each group using available crystallographic structures.

Fig. 1. Schematic explanation of the analysis methods.

relationships in the protein families will be detailed elsewhere. The residue numbers used are those of the first protein in the alignment.

Serine proteases. The alignment is that of Greer (1981). Chymotrypsin, trypsin, elastase (Greer, 1981) and kallikrein (Bode *et al.*, 1983) are aligned by comparison of the three-dimensional structures.

Cysteine proteases. Cysteine proteases were aligned based on the work of Greer (1981): gaps were introduced only where absolutely necessary and always between segments of regular secondary structure. Papain and actinidin were aligned by comparison of their three-dimensional structures, based on Kamphuis *et al.* (1985). Details of the alignment will be given elsewhere.

Haemoglobins. Because the chain lengths are very similar and the sequence similarities are high, the alignment is obvious except for the exact position of a few gaps.

Analysis methods

Figure 1 gives a schematic explanation of the analysis methods for an alignment of four hypothetical sequences of eight residues. Regions of the sequences containing insertions compared to the parent sequence were excluded from the analysis methods described below since they do not exist in the reference crystallographic structure.

Stringent method: search for residues with coordinated amino acid changes. The stringent method described by Altschuh *et al.* (1987) requires several sequences of homologous proteins and at least one crystallographic structure.

In a first step, groups of positions with identical patterns of amino acid changes are constructed from the sequence alignment. The substitution pattern for each position in an alignment of n sequences is symbolized by a string of n digits: the residue type found in the first sequence of the alignment is assigned number 1. Each time a new residue type is found in one of the following sequences, this number is increased by one.

In a second step, positions with identical patterns are located in the crystallographic structure and the minimum distance between side chain atoms (or main chain atoms for glycine) of such residues is measured to identify spatially close pairs. The structures used are those of the first protein in the alignments:

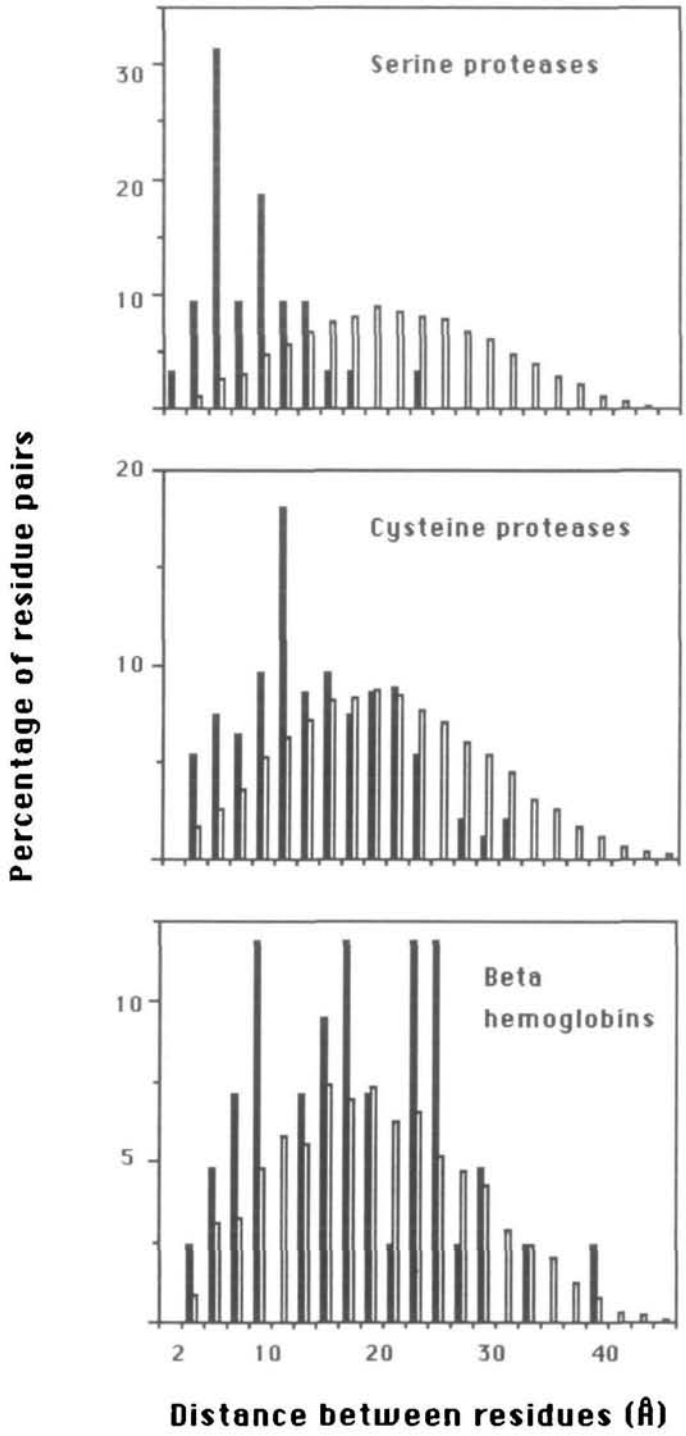

Fig. 2. Distance distribution of all pairs of positions in the protein (white bars) and of pairs selected by the stringent method (black bars) for the three families of proteins.

papain for cysteine proteases, chymotrypsin A for serine proteases and human deoxy haemoglobin for haemoglobins.

Relaxed methods. The first step of the method described above is modified in the following ways:

(i) The analysis can be performed when considering, as identical, two or three amino acid types that are most easily replaced according to the mutation data matrix (Dayhoff *et al.*, 1972, 1978).

(ii) When a very large number of sequences is available it is also useful to select patterns that are not identical but similar. A certain fraction of exceptions is allowed.

(iii) The alignment is searched for positions that contain the same amino acid type in a given set of sequences; that amino acid type must not be found in the other sequences where any combination of substitutions can occur. Thus positions with patterns '1122' and '1123' fall into the same group (Figure 1). Their pattern can be symbolized '11xx' where '1' represents an identical residue type in sequences 1 and 2 and '*x*' represents any residue type except that found in sequences 1 and 2. Similarly, pattern '*xx*11' would contain positions having one type of amino acid in the last two sequences (previously patterns '1233' and '1122').

Groups with >14 positions generated by both the stringent and the relaxed methods were not considered in this study.

Calculation of the significance of results

The distance distribution of pairs selected by the analysis method is compared with that of all pairs in the protein to assess whether or not the positions selected are randomly distributed in the protein structure. The mean distance between positions and the standard deviation are calculated. The histogram of the distance distribution is plotted at 2-Å intervals. Positions conserved in all sequences are excluded.

In order to define an optimal value for spatial proximity, the maximal allowed distance (d) between side chains is varied between 4 and 12 Å. The percentage of pairs with side chain distance of $<d$ Å is calculated for positions selected by the method (R_1) and for all pairs in the protein (R_2). $R_1 = n/N$ where n is the number of spatially close pairs selected by the analysis method, and N is the total number of pairs selected. N is the sum for all groups of $\binom{x}{2} = x!/(x-2)!2!$, where x is the number of positions in each group. R_2 is the ratio of pairs that are spatially close to all possible pairs in the protein. Residues conserved in all sequences are excluded from both R_1 and R_2 calculations.

Calculation of accessible surface area of residues

The accessible surface area (Lee and Richards, 1971) for each residue was calculated from the crystallographic coordinates data using a program developed by Richmond (1984).

Table 1. Distance between all residue pairs and between selected residue pairs

		Mean distance between residues	Standard deviation
Serine proteases	All pairs	20.8	8.5
	Selected pairs	8.3	4.7
Cysteine proteases	All pairs	20.1	8.6
	Selected pairs	13.8	6.7
Beta haemoglobins	All pairs	19.1	8.4
	Selected pairs	17.4	8.1

Results

Stringent method

Significance of results. Results are significant if the number of spatially close pairs found is larger than that expected from a random selection. Figure 2 shows a comparison of the distance distribution of pairs selected by the analysis method with that of all pairs in the protein. The mean distance between positions and the standard deviation are given in Table I. These results

Number of positions in groups		2	3	4	5	6	7	8	9	10	11	12	13	14
Ser proteases	a	4	0	0	0	0	0	1	0	0	0	0	0	0
	b	1	0	0	0	0	0	1	0	0	0	0	0	0
Cys proteases	a	7	7	0	0	0	0	0	0	0	0	1	0	0
	b	1	2	0	0	0	0	0	0	0	0	1	0	0
β hemoglobin	a	6	2	5	0	0	0	0	0	0	0	0	0	0
	b	0	0	3	0	0	0	0	0	0	0	0	0	0

Fig. 3. Stringent method: size distribution of groups of residues. (**a**) Total number of groups; (**b**) number of groups containing residues whose side chains are <5 Å apart (see Figure 4 for a detailed list of positions).

clearly demonstrate that positions selected tend to be closer together than would be expected from a random selection for proteases.

The range of influence between residues in a protein may vary according to such factors as residue type, structural constraints, etc. The success of the selection process was evaluated by comparing the percentage of spatially close pairs selected by the analysis method (R_1) versus all pairs (R_2) at distances ranging from 4 to 12 Å. If R_1/R_2 is close to one, spatially close positions are selected by chance. If R_1/R_2 is greater than one, then at least some positions are close as a result of the selection method. Results are most significant when considering residues whose side chains are <5 Å apart. In these conditions R_1/R_2 is 12.8 for serine proteases, 4.4 for cysteine proteases and 3.2 for haemoglobins. Above this distance, an increasing number of the residues selected are close by chance. However, many positions of the larger groups are within 6 Å of each other (Figures 4 and 5).

Coordinated amino acid changes can be found in all three protein families, but the significance of the results obtained for the haemoglobins is not as clear as for proteases. Differences in the number of sequences analysed, degree of sequence similarity and protein function can account for the variability of results. These points are discussed below.

Effect of residue location in the crystallographic structure. Figure 3 gives for each family of proteins (i) the size distribution of groups selected and (ii) the number of groups containing residues whose side chains are <5 Å apart. These groups are listed in Figure 4 together with the residues' accessible surface values. Almost all positions from the larger groups are clustered together in the structure, although they are far apart in the primary sequence. The average accessible surface area of the residues from the nine groups selected in the proteases (Figure 4) is half of the average area of all residues in the proteins. Residues with replacement patterns 3, 4 and 9 are mostly buried while those with pattern 1 are also in the immediate vicinity of the catalytic site, demonstrating that coordinated changes appear to be found preferentially in structurally constrained regions of proteins.

Effect of sequence similarity on group selection. When using closely related sequences, as for example several mammalian haemoglobins (results not shown), the groups of residues with identical patterns of amino acid changes contain a large number of residues, whose spatial proximity could be fortuitous. In highly divergent sequences, like serine proteases, only a small proportion of the residues can be aligned confidently, restricting the analysis to a fraction of the structure. Identities ranging from ~30 to 70% are optimal for extracting information from sequence alignments.

Effect of the number of sequences on group selection. When the number of sequences in the alignment is increased, each pattern is represented by only one or very few positions because a larger number of substitution patterns is observed. When using a smaller number of sequences, fewer and larger groups are generated including residues that have an identical pattern of substitution by chance. A minimum of about six related protein sequences is required.

Relaxed methods

The relaxed methods were developed to analyse substitutions in less-constrained protein regions not amenable to study by the stringent method, and in large numbers of sequences (Figure 5). Three kinds of modifications are described below.

Considering residue types equivalent. Residue types with similar side chain volume or chemical character are interchangeable in some cases (Dayhoff *et al.*, 1972, 1978). A less-stringent search is performed by considering some residue types equivalent. In beta globin, position 133 has the same pattern as the spatially close positions 11 and 81 if Leu = Phe (Figure 5, pattern 1). Similarly in serine proteases, the two spatially close positions 30 and 139 have the same pattern if Thr = Ser (Figure 5, pattern 2).

Considering almost identical patterns. When a very large number of protein sequences are analysed, strictly identical patterns among the sequence positions are rare. A larger number of groups can be generated by selecting positions whose pattern is almost identical, that is if exceptions are tolerated (data not shown).

Analysing amino acid conservation in all possible combinations of sequences. Since crystallographic data are not available for all members of a protein family, it is not known whether the same residues still interact after substitution. In regions with low structural constraints, residues that interact in some members of a protein family may contact different residues after substitution. We expect these positions to be conserved in the same sequences, and mutate independently from each other in the remaining sequences. Such residues can be looked for by selecting positions having one type of amino acid in a given set of sequences and any other type of amino acid in the remaining sequences. The groups are too numerous to be listed and a selection of two

Residue number	Amino acid type in sequences 1	2	3	4	5	6	7	8	9	10	11	Spatial proximity to position(s)	Accessibility (Å²)
Serine proteases													
1- Pattern	1	1	1	2	1	1	1	1	1	1	1		
42	C	C	C	T	C	C	C	C	C	C	C	*55*,57,58,195	7
52	V	V	V	L	V	V	V	V	V	V	V		0
55	A	A	A	T	A	A	A	A	A	A	A	*42*,57,58,*195*,*197*	0
57	H	H	H	K	H	H	H	H	H	H	H	42,55,58,195	75
58	C	C	C	N	C	C	C	C	C	C	C	42,*55*,57,195	15
195	S	S	S	A	S	S	S	S	S	S	S	42,*55*,57,58,*197*	21
197	G	G	G	S	G	G	G	G	G	G	G	55,*195*,198	1
198	P	P	P	A	P	P	P	P	P	P	P	197	1
2- Pattern	1	1	1	1	1	2	1	1	1	1	2		
123	L	L	L	L	L	I	L	L	L	L	I	124	46
124	P	P	P	P	P	A	P	P	P	P	A	123	6
Cystein proteases													
3- Pattern	1	1	1	1	1	1	1	1	1	2	2	2	
7	W	W	W	W	W	W	W	W	W	A	A	A	164	30
17	K	K	K	K	K	K	K	K	K	R	R	R	35,50	23
29	S	S	S	S	S	S	S	S	S	G	G	G		2
35	E	E	E	E	E	E	E	E	E	S	S	S	17,50	2
50	E	E	E	E	E	E	E	E	E	A	A	A	17,35,*51*,88	3
51	Q	Q	Q	Q	Q	Q	Q	Q	Q	E	E	E	*50*,88	0
79	G	G	G	G	G	G	G	G	G	L	L	L		3
87	P	P	P	P	P	P	P	P	P	S	S	S	*88*	73
88	Y	Y	Y	Y	Y	Y	Y	Y	Y	H	H	H	50,51,*87*	51
164	V	V	V	V	V	V	V	V	V	L	L	L	7,186	0
166	Y	Y	Y	Y	Y	Y	Y	Y	Y	W	W	W		5
186	Y	Y	Y	Y	Y	Y	Y	Y	Y	F	F	F	164	18
4- Pattern	1	2	3	3	3	3	3	3	3	4	4	4		
32	V	A	G	G	G	G	G	G	G	E	E	E	162	14
85	T	N	S	S	S	S	S	S	S	Y	Y	Y		54
162	A	T	L	L	L	L	L	L	L	R	R	R	32	0
5- Pattern	1	1	2	1	1	1	1	1	1	3	3	3		
131	S	S	A	S	S	S	S	S	S	E	E	E		1
208	Y	Y	T	Y	Y	Y	Y	Y	Y	A	A	A	209	44
209	P	P	S	P	P	P	P	P	P	G	G	G	208	0
6- Pattern	1	2	3	3	3	3	3	3	3	3	3	3		
192	G	N	.	.	.	.	.	.	.	.	.	.	193	50
193	T	V	.	.	.	.	.	.	.	.	.	.	192	79
β hemoglobin													
7- Pattern	1	1	1	1	1	1	1	1	2	2	1	1		
38	T	T	T	T	T	T	T	T	K	K	T	T	39	0
39	Q	Q	Q	Q	Q	Q	Q	Q	R	R	Q	Q	38	46
84	T	T	T	T	T	T	T	T	H	H	T	T		52
127	Q	Q	Q	Q	Q	Q	Q	Q	H	H	Q	Q		72
8- Pattern	1	1	1	1	1	1	1	1	1	2	1	1		
49	S	S	S	S	S	S	S	S	S	C	S	S	53	89
53	A	A	A	A	A	A	A	A	A	D	A	A	49	24
62	A	A	A	A	A	A	A	A	A	E	A	A		56
96	L	L	L	L	L	L	L	L	L	F	L	L		45
9- Pattern	1	1	1	1	2	2	2	2	2	2	2	2		
11	V	V	V	V	I	I	I	I	I	I	I	I		0
74	G	G	G	G	A	A	A	A	A	A	A	A	81	0
81	L	L	L	L	I	I	I	I	I	I	I	I	74	8
108	N	N	N	N	D	D	D	D	D	D	D	D		77

Fig. 4. List of groups selected with the stringent method. For each group, the amino acid substitution pattern is given, followed by a list of residue numbers with corresponding amino acid types in the sequences analysed, spatial proximity to other positions of the group and surface accessibility (Å²). Spatial proximity is indicated in normal font if side chains are <5 Å apart, and in italics if they are separated by 5–6 Å.

examples are given in Figure 5. This type of substitution pattern can be observed for instance in pairs of residues known to form salt or disulphide bridges. A clear example is depicted in Figure 5, pattern 3, where the two cysteines 136 and 201, known to form a disulphide bridge in chymotrypsin A (Birktoft and Blow, 1972), are always conserved together. The residue types at positions 136 and 201 are not coordinated in proteins which do not contain cysteines at these positions. Similarly, when amino acids which form salt bridges in human haemoglobin are replaced by uncharged residues, their partners are no longer selectively

Residue number	Amino acid type in sequences 1	2	3	4	5	6	7	8	9	10	11	12	Spatial proximity to position(s)	
(1) Considering equivalent residue types														
β hemoglobin														
1- Pattern	1	1	1	1	2	2	2	2	2	2	2	2		
11	V	V	V	V	I	I	I	I	I	I	I	I	133	
74	G	G	G	G	A	A	A	A	A	A	A	A	81	
81	L	L	L	L	I	I	I	I	I	I	I	I	74,*133*	
108	N	N	N	N	D	D	D	D	D	D	D	D		
133	V	V	V	V	L	L	L	L	L	L	F	F	11,*81*	**if L = F**
Serine proteases														
2- Pattern	1	1	1	1	1	1	1	1	2	1	2			
30	Q	Q	Q	Q	Q	Q	Q	Q	M	Q	M		139	
139	T	S	T	S	S	S	S	T	A	T	A		30	**if T = S**

(2) Searching for almost identical patterns

This variation of the method is useful when a large number of sequences is considered (no example given here).

(3) Considering amino acid conservation in all possible combinations of sequences

Serine proteases													
3- Pattern	1	1	1	x	1	x	x	1	1	x	x		
136	C	C	C	g	C	g	g	C	C	g	f		201
201	C	C	C	v	C	t	t	C	C	m	r		136
β hemoglobin													
4- Pattern	1	1	1	1	x	x	x	x	x	x	x	x	
3	L	L	L	L	w	w	w	w	f	f	w	w	11,133
11	V	V	V	V	i	i	i	i	i	i	i	i	3,133
23	V	V	V	V	c	c	c	c	c	c	l	l	
29	G	G	G	G	a	a	a	a	s	s	a	a	
61	K	K	K	K	r	r	q	r	q	q	a	a	
72	S	S	S	S	g	g	g	g	g	g	m	e	*74*
74	G	G	G	G	a	a	a	a	a	a	a	a	*72*,81
81	L	L	L	L	i	i	i	i	i	i	i	i	74,*133*,136
108	N	N	N	N	d	d	d	d	d	d	d	d	*109*
109	V	V	V	V	i	i	i	i	i	i	c	c	108
133	V	V	V	V	l	l	l	l	l	l	f	f	3,11,*81*,*136*
135	A	A	A	A	r	r	r	r	r	r	s	s	*136*
136	G	G	G	G	v	v	v	v	q	q	v	v	81,*133*,135
143	H	H	H	H	r	r	r	r	a	a	r	r	

Fig. 5. Examples of positions selected with the relaxed methods in β haemoglobins and serine proteases. In patterns 3 and 4, the sequences considered are represented by the number '1'; all other sequences are represented by the letter 'x'. Residue types in the set of sequences considered are indicated in upper case letters. See also legend to Figure 4.

maintained and are also replaced by uncharged residues (unpublished results), illustrating the probable variability in the geometry of contacts formed by polar residues.

By introducing these modifications to the method, a large number of groups are formed, among which only few are significant. When considering only those cases where most residues from a group are spatially related, interesting clusters of residues that were missed by the original method are observed. Relaxed methods are therefore valuable tools in the detailed analysis of amino acid substitutions in protein families.

Discussion

Optimal conditions for finding coordinated amino acid changes using the stringent method

Highly constrained structures. The method is based on the assumption that residues linked by a specific role mutate in a coordinated manner. This is only true if the geometry of contacts is the same in all related proteins which explains the requirement for highly constrained structures. In a less-constrained structure, residues that interact in one member of a protein family could contact other residues after mutation. Such residues will not have an identical substitution pattern.

Coordinated mutations were easily characterized in tobamovirus coat proteins. However, the study of other types of protein families led to more ambiguous results. Viral coat proteins must be able to aggregate reversibly to produce a capsid containing the nucleic acid. Constraints exist not only on molecular size and shape, but also on surface regions that interact to form quaternary structures. These stringent structural requirements could favour the occurrence of coordinated changes.

In proteins, such as immunoglobins and insulins, mutations can be accommodated by minor shifts of elements of secondary structure (Lesk and Chothia, 1982; Chothia *et al.*, 1983). Structurally common regions of protein families include major elements of secondary structure and the active site (Chothia and Lesk, 1986). Coordinated changes are expected in such regions only if constraints on local shape are high.

Low sequence similarity. The method is not useful for analysing proteins with high sequence similarity because each pattern is represented by a large number of sequence positions. The

proximity in space of these positions could be fortuitous. The existence of coordinated changes in such closely related proteins cannot be excluded but substitutions are likely to be conservative or occur in regions that have no structural constraints. In highly divergent proteins, a large number of changes have accumulated, reducing the number of positions that only fortuitously have the same pattern of substitution. The passage from one set of amino acids having a given role to another set having a similar role must have occurred in many steps, involving sequential substitutions of this set of residues and possibly neighbouring ones as well. Only the analysis of stable states for a given interaction would allow the detection of coordinated changes.

Tobamovirus coat proteins represent an optimal case for this type of analysis: three out of seven available sequences have $<50\%$ similarity with the tobacco mosaic virus *vulgare* protein as well as with each other, while retaining similar chain lengths, facilitating the alignment. The whole structure of this small protein is likely to be highly constrained. In proteases, only a portion of the structure is constrained but coordinated changes also occur within this portion.

Extension of the stringent method: the relaxed methods. Applying the relaxed methods generates a much larger number of groups which require extensive screening to select for meaningful information. E.g. amino acids with similar side chain volume or chemical character can be considered equivalent in some conditions only. According to the Dayhoff mutation matrices, Ile is easily replaced by Leu. However, Nagai *et al.* (1987) showed that Ile and Leu have different effects when they replace a Val residue in the oxygen binding pocket of beta globin. Therefore, it is necessary to consider all possibilities when analysing substitutions in sequence alignments. Assessing the functional significance of amino acid conservation in a set of sequences requires careful consideration of the amino acid type and local environment of the selected positions. Groups of residues may be incomplete or contain unrelated residues. However, interesting correlations were found in this way, particularly for charged residues with long side chains.

Effect of amino acid changes. In some cases, coordinated changes may be complementary. In cysteine proteases for example, the pairs of substitutions Val32→Gly and Ala162→Leu (Figure 4, pattern 4) or Val164→Leu and Tyr186→Phe (Figure 4, pattern 3) should not drastically alter local structure. On the contrary, the Val32→Glu and Ala162→Arg substitutions (Figure 4, pattern 4) could modify the structure and/or chemical environment, thus influencing enzyme function. It is possible that other modifications, not detected by the method described, compensate for the substitutions at positions 32 and 162. Interactions that involve polar residues with long side chains often undergo extensive changes. It is difficult to predict to what extent these changes could modify structure or function. The haemoglobin sequences analysed include fish haemoglobins which have different functional properties. The groups of residues selected could be responsible for these differences. Among serine proteases, haptoglobin is able to bind substrates but has no catalytic activity. Kurosky *et al.* (1980) have noted that two positions corresponding to the active site in other proteases are occupied by a different residue type in haptoglobin. The analysis performed here shows that six additional positions that are conserved in all other serine proteases are also substituted in haptoglobin.

Although results clearly depend on available sequences, the methods described here emphasize possible relationships between sets of residues among the complexity of interactions that compose protein structures. This information represents a helpful starting point for further theoretical and experimental analysis on structure–function relationships in proteins.

Acknowledgements

We thank Jeremy Tame, Philippe Walter and Fred Jacobs for careful reading of this paper, M.H.V.Van Regenmortel, David Y.Thomas and Andrew Storer for their support throughout this work. We are also grateful to Cyrus Chothia, Giulio Fermi and Anne Bloomer for helpful discussions. T.V. is recipient of a grant from the NRCC/CNRS exchange programme. This paper is dedicated to the memory of our colleague Petr Klein.

References

Altschuh,D., Lesk,A.M., Bloomer,A.C. and Klug,A. (1987) *J. Mol. Biol.*, **193**, 693–707.
Birktoft,J.J. and Blow,D.M. (1972) *J. Mol. Biol.*, **68**, 187–240.
Bode,W., Chen,Z., Bartels,K., Kutzbach,C., Schmidt-Kastner,G. and Bartunik,H. (1983) *J. Mol. Biol.*, **164**, 237–282.
Chan,S.J., San Segundo,B., McCormick,M.B. and Steiner,D.F. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 7721–7725.
Chothia,C. and Lesk,A.M. (1986) *EMBO J.*, **5**, 823–826.
Chothia,C., Lesk,A.M., Dodson,G.G. and Hodgkin,D.C. (1983) *Nature*, **302**, 500–505.
Cohen,L.W., Coghlan,V.M. and Dihel,L.C. (1986) *Gene*, **48**, 219–227.
Dayhoff,M.O., Eck,R.V. and Park,C.M. (1972) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, Vol. 5, pp. 89–99.
Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, Vol. 5, suppl. 3, pp. 345–352.
Fermi,G., Perutz,M.F., Shaanan,B. and Fourme,R. (1984) *J. Mol. Biol.*, **175**, 159–174.
Greer,J. (1981) *J. Mol. Biol.*, **153**, 1027–1042.
Ishidoh,K., Towatari,T., Imajoh,S., Kawasaki,H., Kominami,E., Katunuma,N. and Suzuki,K. (1987a) *FEBS Lett.*, **223**, 69–73.
Ishidoh,K., Imajoh,S., Emori,Y., Ohno,S., Kawasaki,H., Minami,Y., Kominami,E., Katunuma,N. and Suzuki,K. (1987b) *FEBS Lett.*, **226**, 33–37.
Kamphuis,I.G., Kalk,K.H., Swarte,M.B.A. and Drenth,J. (1894) *J. Mol. Biol.*, **179**, 233–256.
Kamphuis,I.G., Drenth,J. and Baker,E.N. (1985) *J. Mol. Biol.*, **182**, 317–329.
Kurosky,A., Barnett,D.R., Lee,T.H., Touchstone,B., Hay,R.E., Arnott,M.S., Bowman,B.H. and Fitch,W.M. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 3388–3392.
Lee,B. and Richards,F.M. (1971) *J. Mol.Biol.*, **55**, 379–400.
Lesk,A.M. and Chothia,C. (1982) *J. Mol. Biol.*, **160**, 325–342.
Nagai,K., Luisi,B., Shih,D., Miyazaki,G., Imai,K., Poyart,C., DeYoung,A., Kwiatkowsky,L., Noble,R.W., Lin,S.H. and Yu,N.T. (1987) *Nature*, **329**, 858–860.
Pears,C.J., Mahbubani,H.M. and Williams,J.G. (1985) *Nucleic Acids Res.*, **13**, 8853–8866.
Praekelt,U.M., Mckee,R.A. and Smith,H. (1988) *Plant Mol. Biol.*, **10**, 193–202.
Richmond,T.J. (1984) *J. Mol. Biol.*, **178**, 63–89.
Troen,B.R., Gal,S. and Gottesman,M.M. (1987) *Biochem. J.*, **246**, 731–735.
Wada,K., Takai,T. and Tanabe,T. (1987) *Eur. J. Biochem.*, **167**, 13–18.
Whittier,R.F., Dean,D.A. and Rogers,J.C. (1987) *Nucleic Acids Res.*, **15**, 2515–2535.

Received on May 2, 1988; revised on July 15, 1988