Protein Engineering vol.11 no.5 pp.349–359, 1998

Assessment of conformational parameters as predictors of limited proteolytic sites in native protein structures

S.J.Hubbard[1], R.J.Beynon and J.M.Thornton[2]

Department of Biomolecular Sciences, University of Manchester Institute of Science and Technology, PO Box 88, Manchester M60 1QD, [2]Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, UK

[1]To whom correspondence should be addressed

Despite the importance of limited proteolysis in biological systems it is often difficult to rationalize why a proteinase hydrolyses a particular bond, given a simple sequence specificity alone. Understanding of the structural properties limiting the proteolysis represents a first step on the pathway to control and manipulation of this phenomena. An expanded set of nick-sites in proteins of known tertiary structure, cut by both narrow and broad specificity proteinases, has been generated yielding a robust data set of strictly limited sites. A critical evaluation of an expanded set of conformational parameters revealed a strong correlation with limited proteolytic sites, although they are only modest predictors in isolation. The overall predictive power is significantly improved when the conformational parameters are combined in a weighted predictive scheme that permits their relative importance to be compared via a Metropolis search protocol. A subset of the parameters performs equally well demonstrating the key determinants of susceptibility. The derived predictive algorithm has been made available via the internet. Its utility for predicting other surface-correlated features is also discussed.

Keywords: molecular recognition/proteinase/limited proteolysis/prediction/nick-sites

Introduction

Limited proteolysis, the specific fission of only one or a few peptide bonds in a folded protein chain, underpins many important biological functions such as zymogen activation, the blood coagulation cascade and pro-hormone and neuropeptide processing (Ottensen 1967; Neurath and Walsh, 1976; Bond and Beynon, 1987; Price and Johnson, 1989). Generally, the details of the catalytic reaction by which this hydrolysis takes place are well understood, particularly for the serine proteinases (Blow, 1976; Kraut, 1977) where hydrolysis is achieved through nucleophilic attack upon the carbonyl carbon of the peptide bond. However, the global molecular recognition processes are not well understood. Specifically, it is unclear how a proteinase of known sequence specificity recognizes such a limited subset out of the many putative sites of proteolysis in a folded polypeptide chain. For example, trypsin will completely degrade most proteins in denaturing conditions, cleaving at nearly every lysine-X, arginine-X bond (with the partial exception of proline at X). Thus, about 5–10% of the peptide bonds in a typical protein ought to be susceptible to proteolytic attack. However, in native (or near-native) conditions trypsin will cut only a limited number of such bonds (or on occasions none at all) in a native protein fold. The structure and dynamics of the substrate protein must therefore play a role in limiting the proteolysis.

Much of our understanding of the proteinase catalytic mechanism is derived from X-ray crystallographic studies of not only the enzymes, but also complexes with small protein inhibitors, such as BPTI. These protein–protein complexes provide a paradigm for the transition state of the reaction, with the inhibitor reactive site loop bound into the enzyme active site in the manner of a 'perfect' substrate. This canonical conformation is conserved throughout diverse families of small protein inhibitors of serine proteinases although the overall fold and amino acid sequence of these inhibitors are not (Laskowski and Kato, 1980; Bode and Huber, 1992). Using this canonical conformation of the inhibitor reactive site loops as a template it has been shown that limited proteolytic sites are quite different in structure from the idealized inhibitor loops, and they must therefore undergo a conformational change in order to enter the proteinase active site (Hubbard *et al.*, 1991). From modelling experiments it is expected that minimally this must involve a local unfolding step of 10 residues or more prior to recognition and cleavage (Hubbard *et al.*, 1994). Hence, the position of the putative limited proteolytic site (nick-site) with respect to the rest of the substrate tertiary structure, and the inherent flexibility and opportunity for local unfolding must help determine its proteolytic susceptibility. Indeed, the implicit assumption that limited proteolytic sites are at exposed and flexible regions makes limited proteolysis an invaluable structural probe for investigation of protein structure and function (Price and Johnson, 1989; Fontana *et al.*, 1997b).

For predictive purposes, specific proteolytic processing systems have been studied by a number of workers. One such study (Monsalve *et al.*, 1990) observed that proteolytic processing sites in seed proteins are found at sequence sites with a very high probability to form a β-turn. Similarly, a simple scheme to predict Ω-loops from protein amino acid sequences was developed and subsequently applied to the prediction of prohormonal cleavage sites (Bek and Berry, 1990). However, for more general proteolytic systems, it would be a useful first step to be able to predict which sites were most susceptible to limited proteolytic attack for proteins whose tertiary structure is already known. Prior to this it should be established which protein features are responsible for proteolytic susceptibility and their relative importance and weighting. Typically limited proteolytic sites are found at flexible loop regions (as indicated by crystallographic temperature factors or B-values) that are also exposed to the solvent (Fontana *et al.*, 1986; Novotný and Bruccoleri, 1987; Fontana, 1989; Hubbard *et al.*, 1991) and are notably absent in regions of regular secondary structure, especially β-sheets (Fontana, 1989; Hubbard *et al.*, 1994; Fontana *et al.*, 1997a,b). They protrude from the protein surface (Hubbard *et al.*, 1991) and would be expected to be found at regions where the local packing does not inhibit the local unfolding that is deemed necessary.

Previous studies (Hubbard *et al.*, 1991, 1992) considered some of these features and demonstrated their correlation with a small number of tryptic proteolytic sites. A simple prediction scheme was derived from this analysis and was successfully applied to the prediction of limited proteolytic sites of the apo- and holo-forms of the biotin-binding protein avidin (Ellison *et al.*, 1995). Here, we extend these conformational parameter sets to include Ooi numbers, secondary structure parameters and hydrogen bonding. Furthermore, the data set of limited proteolytic sites has been expanded to include sites cut by proteinases other than trypsin and stricter criteria for the definition of 'limited' proteolysis have been applied. A rigorous comparison of the predictive power of these conformational parameters sets has been undertaken. Finally, these conformational parameters have been combined into a predictive algorithm which has been made available to the biological community via the internet.

Materials and methods

Dataset

A list of known limited proteolytic sites was generated by an extensive search of the literature, adding to the previous set of tryptic sites (Hubbard *et al.*, 1991). Sites were further selected according to whether the tertiary structure of the protein (or very close homologue) was known to high resolution, the precise bond cleaved had been determined without ambiguity, and whether the proteolysis itself could truly be deemed 'limited'. For this purpose, second-order rate constants k_2 were estimated from the literature, from a half-life $t_{1/2}$ read from a graph or gel time series, thus:

$$k_2[\mathrm{E}] = \frac{\ln 2}{t_{1/2}}$$

Additionally, digests where the substrate protein could be expected to be largely unfolded or non-native (under high concentrations of denaturing agents or high temperatures) were ignored unless the protein was reported to retain its fold under these circumstances (e.g. by means of retention of activity or structural evidence). Similarly, sites were only included if the structure of the correct apo-/holo-form of the protein was available where this was known to affect the proteolytic susceptibility.

The full list of nick-sites used in this study is given in Table I along with estimates of k_2 and the temperatures at which the reaction took place. One reaction was included which took place above 37°C as the substrate protein was a thermophile which is stable and functional at this temperature.

Calculation of conformational parameters

Calculations were performed on the co-ordinated entries listed in Table I taken from the Brookhaven Databank (Bernstein *et al.*, 1977). The following conformational parameters, and their sub-types, were calculated for each protein.

Solvent accessibility

The accessible surface area of each individual residue in each protein was calculated using the method of Lee and Richards (1977) using a 1.4 Å probe and the atomic radii data set of Chothia (1976). Where present and appropriate, heteroatoms (excluding water and similar solvent molecules) were considered for the calculation of atomic accessibilities. Absolute residue accessibilities were calculated simply as the sum of the atomic accessible areas for each residue. Summed residue accessible surfaces were also expressed as relative percentage accessibilities of the exposed state, taking the latter from extended tripeptides of Ala-X-Ala for each amino acid type X. Nick-sites are already known to correlate with solvent exposure (Novotný and Bruccoleri, 1987; Vita *et al.*, 1988; Fontana, 1989; Hubbard *et al.*, 1991).

Protrusion index

The residue protrusion index was calculated by the method of Taylor *et al.* (1983) whereby an equimomental ellipsoid is calculated about the molecular centre of mass approximating the protein shape. Successive similar ellipsoidal shells are assigned containing increasing 10-percentiles of the protein atoms. Atoms are then assigned a score of 0 to 9 signifying the outermost ellipsoidal shell in which each atom lies: 0 for the core through to 9 for the outer, 'protruding' atoms. Calculations were performed using either solely α-carbon atoms or using all atoms and averaging over each residue to obtain a final residue score. Both methods were compared.

Residue-averaged temperature factors

In the absence of more detailed solution data on protein flexibility, atomic temperature factors were used as a measure of mobility, as they have been previously shown to be correlated with limited proteolytic susceptibility (Vita *et al.*, 1988; Fontana, 1989; Hubbard *et al.*, 1991). Four residue-averaged measures were considered using α-carbons only, backbone atoms, side chain atoms and all the atoms in a given residue.

Ooi numbers

Since nick-sites might be expected to be located at regions of weak packing, Ooi numbers were calculated for each residue as a simple and fast measure (Nishikawa and Ooi, 1986). The residue Ooi number is simply the number of other α-carbon atoms within a fixed radius of the residue's α-carbon. Two cut-off radii values were compared: 8 and 14 Å.

Secondary structure parameters

As nick-sites are not prevalent in regions of regular secondary structure (Fontana, 1989; Hubbard *et al.*, 1994; Fontana *et al.*, 1997a,b) particularly β-strands, this was formulated in a simple manner by three scores, one for each of the three secondary structure states helix, strand or coil. An additional penalty score for cysteine residues participating in a disulphide bridge was also applied, subtracted from other secondary structure scores at that residue position. States were assigned to each residue using the method of Kabsch and Sander (1983) where residues were classed as either helix (H), strand (E) and all others coil. Multiple combinations of the four values were compared, ranging from 0.0 to 1.0. The optimal combination was defined using a Metropolis search procedure, discussed later, yielding optimal weights for the four standard states of helix 0.5, strand 0.0, coil 1.0 and disulphide penalty 0.4, based on their ability to discriminate nick-sites from residues in general. This reflects the implausibility of locating nick-sites in β-structure and their rarity in α-helix shown by modelling experiments and prior observation (Hubbard *et al.* 1994; Fontana 1989).

Main chain hydrogen bonding

As noted for loop-closure modelling experiments on putative tryptic sites of elastase, the true nick-site region makes relatively few main chain hydrogen bonds to other regions of the protein (Hubbard *et al.*, 1994) indicating that local unfolding regions are constrained by the fewest intramolecular inter-

Table I. Limited proteolytic sites used in this study

Protein	Source	Proteinase code	Databank	P^1 residue	References	k_2 rate constant (estimate) $\mu M^{-1}.min^{-1}$	Temperature (if known; °C)
Narrow specificity proteinase sites							
Staphylococcal nuclease	*S.aureus*	trypsin	1SNO	Lys5 Lys48 Lys49	Taniuchi *et al.*, 1967 Taniuchi and Anfinsen, 1968	–	25
Elastase	Porcine	trypsin	3EST	Arg125	Ghelis *et al.*, 1978	–	4
Calmodulin	Bovine	trypsin	1LIN	Lys77	Walsh *et al.*, 1977	–	20
Trypsinogen	Bovine	trypsin	1TGN	Lys145	Higaki and Light, 1985, Hermodson *et al.*, 1973	–	25
Thaumatin	*T.daniellii*	trypsin	1THV	Arg119, Lys163	Stephen, 1993	0.02	25
Aspartate aminotransferase	Pig heart	trypsin	2CST‡	Lys19, Arg25	Iriarte *et al.*, 1984	0.005	37
Major Urinary Protein	Mouse	trypsin	1MUP	Arg12	Wu, CY. (UMIT) Personal communication.	0.0007	25
Hemocyanin subunit A	Spiny Lobster	trypsin	1HCY[a]	Lys174, Lys175	Neuteboom *et al.*, 1992	<0.03	37
SH-I neurotoxin	Anemone	trypsin	2SH1	Arg13	Monks *et al.*, 1994	<0.21	37
Cytochrome c	Horse	trypsin	1CRC	Lys55	Hu *et al.*, 1996	<0.09	37
Hirudin	*H.manillensis*	trypsin	5HIR[a]	Lys47	Vindigni *et al.*, 1994	>0.009	37
Cellular retinol binding protein-II	Mouse	arg-C	1OPA[a]	Arg30	Jamison *et al.*, 1994	0.06	37
Hirudin	Leech	V8-proteinase	5HIR[a]	Glu43	Vindigni *et al.*, 1994	0.009	37
Cellular retinol binding protein-II	Mouse	V8-proteinase	1OPA[a]	Glu17	Jamison *et al.*, 1994	–	37
Broad specificity proteinase sites							
Pancreatic lipase	Horse	chymotrypsin	1HPL	Leu410	Aboudalham *et al.*, 1992	0.04	25
Avidin (apo)	Chicken	proteinase k	1AVE	Thr40 Asn42	Ellison *et al.*, 1995	0.002	30
Thermolysin	*B.thermoproteo.*	subtilisin	8TLN	Thr4 Thr224 Gln225	Vita *et al.*, 1985	0.006	25
Ribonuclease A	Bovine	subtilisin	5RSA	Ala20	Richards and Vithayathil, 1959	–	3
Triose Phosphate Isomerase	Yeast	subtilisin	1YPI	Leu174	Sun *et al.*, 1993	–	37
Hemocyanin A	Spiny Lobster	subtilisin	1HCY	Thr173	Vereijken *et al.*, 1982	–	37
Ribonuclease A	Bovine	elastase	5RSA	Ser18 Ala21	Rupley and Scheraga, 1963	–	–
Aspartate aminotransferase	*S.solfataricus*	thermolysin	2CST‡	Gly28	Arnone *et al.*, 1992	0.01	60
Cellular retinol binding protein-II	Mouse	thermolysin	1OPA‡	Ile32	Jamison *et al.*, 1994	–	37

[a]Close homologue.

Table II. Proteinase primary sequence specificity requirements

Proteinase	Amino acid specificity
Trypsin	P_1 = Arg, Lys, $P_{1'} \neq$ Pro
Chymotrypsin	P_1 = Trp, Tyr, Phe, Leu, Met
Subtilisin	$P_1 \neq$ Arg, Lys
Proteinase K	$P_1 \neq$ Arg, Lys, Asp, Glu
Elastase	P_1 = Ala,Val., Leu, Ile, Gly, Ser
Thermolysin	$P_{1'}$ = Leu, Phe, Ile, Val., Met, Ala
Arg-C	P_1 = Arg
V8-proteinase	P_1 = Glu

actions. Non-local hydrogen bonding was quantified by counting the number of backbone–backbone hydrogen bonds in a given loop region made to residues outside of the loop. In addition, where possible and appropriate, hydrogen bonds to heteroatom polar groups were also added to the sum for each loop region. Hydrogen bonds were calculated using a distance cut-off of below 3.5 Å and an angular cut-off of above 90° at the amide hydrogen (except for heteroatoms). This was evaluated for loop lengths of 6, 8, 10 and 12 residues with the putative nick-site situated in the centre of the loop at the 3rd, 4th, 5th and 6th position respectively.

Assessment of conformational parameters

Residue scores were calculated for each conformational parameter for each protein. The residue parameters were then smoothed using a fixed window length *n*, assigning averaged scores to the putative P_1 residue in a $P_{n/2}$ to $P_{n/2}'$ window (Schechter and Berger, 1967). Values of *n* ranging from 4 to 20 in steps of 2 were evaluated. This encompasses the minimal segment length likely to be necessary for proteolytic recognition (the P_2-P_2' region centred about the scissile peptide) and the minimum number of residues deemed necessary for local unfolding and subsequent cleavage (Hubbard *et al.*, 1994). Residues at the N- and C-termini were smoothed by averaging over the reduced number of residues lying within the smoothing window. The smoothed scores were then normalized to lie within the range 0.0 and 1.0 to give a residue prediction score $N_x(i)$ for each parameter *x* at each residue position *i*. For Ooi numbers and the non-local hydrogen bonding term, normalized scores were also inverted by subtraction from unity to favour more weakly packed and hydrogen bonded residue positions. Mean values were then calculated for three subsets of the data set residues; all residues, putative nick-sites (those satisfying the primary sequence requirements of the attacking proteinase) and the nick-sites themselves. The primary sequence requirements of the proteinases considered in this study are listed in Table II.

Individual conformational parameters were assessed via two simple functions designed to show how well each feature discriminates the true nick-sites from residues in general and other putative (but uncut) nick-sites. These functions are shown below:

$$DF_{all} = \frac{\overline{P_{nicksites}} - \overline{P_{all}}}{\sigma_{all}} \quad \text{and} \quad DF_{nick} = \frac{\overline{P_{nicksites}} - \overline{P_{putative}}}{\sigma_{putative}}$$

where *P* is the normalized score for any given parameter, $\overline{P}$ is the mean parameter score for a given subset of residues and σ the standard deviation of a given mean value. This yields two 'discrimination factor' scores (DF_{all} and DF_{nick}) that vary according to how much outside a given distribution the mean nick-site score lies. The origins of these functions are illustrated in Figure 1 which shows hypothetical distributions of normalized parameter scores.

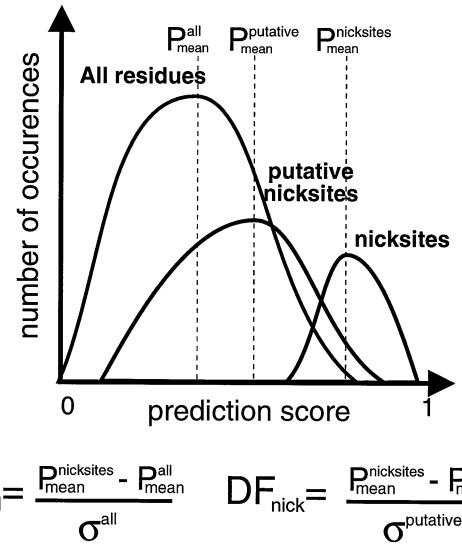

Fig. 1. Origins of optimization parameters for nick-site prediction. Three hypothetical distributions of amino acid scores are illustrated, for all residues, lysine/arginines and for nick-sites. Each distribution may be characterized by a mean values P_{mean} and a standard deviation σ. A discrimination factor *D* may be evaluated for the deviation of the nicksite mean score from either of the other two distributions.

Additionally, the individual parameter scores were sorted for each protein in the data set and a mean rank value R_{mean} calculated for each parameter for every window size under consideration.

Prediction of limited proteolytic sites

Predictions were made by first calculating the normalized residue scores for each of the six selected conformational parameters described above. Normalized scores N_x for each parameter *x* were then combined to give a final prediction score *P(i)* at each residue position *I*, using a weight for each parameter w_x:

$$P(i) = \frac{\sum_{x=1}^{x=n_f} w_x N_x(i)}{n_f} \quad \text{where } n_f = \text{number of valid features}$$

In practice, the number of valid feature scores was either 5 or 6 depending on whether temperature factors were available for that protein. The relative importance of each parameter was adjusted via a weighting scheme. Weights were originally set between 0.0 and 1.0 and then normalized so their mean was set to unity, thus:

$$\frac{\sum_{x=1}^{x=n_f} w_x}{n_f} = 1.0$$

Assessment of prediction scores and prediction optimization

Predictions were assessed via the same discrimination factors DF_{all} and DF_{nicks} and mean rank R_{mean} as for the individual conformational parameters although the weighted prediction scores were used instead of the individual parameter score. However, in order to optimize the prediction, these simple

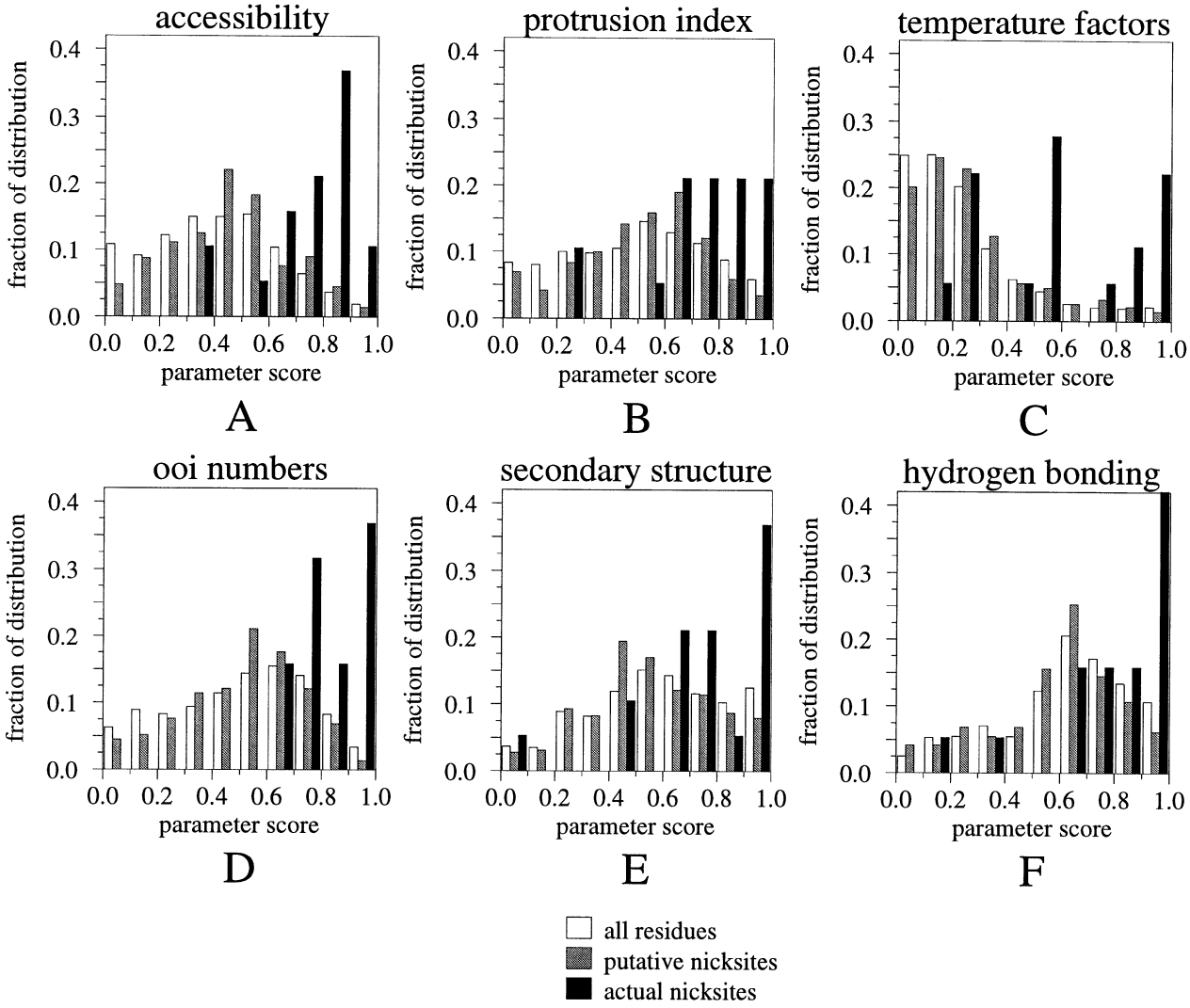

Fig. 2. Distributions of normalized parameter scores for amino acids in the narrow specificity data set. Distributions of scores for amino acids in the narrow specificity data set are shown for the individual conformational parameters after having been normalized so that their sum is 1 over the range of prediction scores 0 to 1. The white bars represent all residues in the dataset, the grey bars putative sites only, and the black bars the actual nick-site residues. Plots are shown for (**A**) relative accessibility, (**B**) protrusion index calculated on Cα-atom positions, (**C**) mean residue temperature factors, (**D**) Ooi numbers with an 14 Å cut-off, (**E**) secondary structure parameters and (**F**) non-local hydrogen bonds outside a loop of 12 residues.

functions were converted to a simple 'energy' function E_{total} via:

$$E_1 = -\ln DF_{all}$$
$$E_2 = -\ln DF_{nick}$$
$$E_3 = -\ln\left(\frac{1}{R_{mean}-1}\right)$$
$$E_{total} = E_1 + E_2 + E_3$$

To assess the relative predictive merits of the parameters, predictions were optimized using a simple Metropolis optimization procedure, where all smoothing windows and feature weights were allowed to change and the resulting prediction 'energy' monitored. Smoothing windows were allowed to vary from 4 to 20 in steps of 2 and weights from 0.0 to 1.0 in steps of 0.1. At each Metropolis step, either one of the window lengths or weights selected at random was increased or decreased, and the resultant prediction 'energy' recalculated. The new windows/weights were accepted according to the Metropolis probability criteria:

$$P_a = e^{-(\Delta E/\alpha)} \qquad \text{where} \qquad \Delta E = E_{new} - E_{old}$$

A Metropolis step was accepted if the probability P_a exceeded a random number cast between 0.0 and 1.0. When the energy decreases P_a always exceeds 1 and the step is accepted. Values for α where found to work well around 0.08. A total of 20 optimizations were run for each of the two nick-site data sets, each for 50 000 steps, taking the lowest energy over the 20 runs as the most optimal.

Results

Choice of conformational parameters

The calculated conformational parameters, and their sub-types scores were assessed for their power to discriminate nick-sites from residues in general and from other putative, but uncut,

Table III. Conformational parameter statistics for nick-sites prediction

Parameter	Energy E_{total}	Optimal window	DF_{all}	DF_{nicks}	Mean rank
Absolute accessibility	–0.54	12	1.37	1.27	2.00
Relative accessibility	–0.54	12	1.37	1.27	2.00
Protrusion Index (α-carbons)	1.53	6	0.91	0.87	4.67
Protrusion Index (all atoms)	1.66	12	0.95	0.91	5.50
B-values (α-carbon only)	–0.23	8	1.71	1.42	2.92
B-values (backbone)	–0.36	4	1.70	1.42	2.67
B-values (sidechain)	–0.26	4	1.63	1.27	2.58
B-values (all atoms)	–0.45	4	1.67	1.34	2.42
Ooi numbers (8 Å)	–0.15	8	1.06	1.11	2.00
Ooi numbers (14 Å)	–0.96	4	1.34	1.31	1.67
Secondary structure set (optimised)	1.02	12	0.87	0.93	3.25
Hydrogen bonding (6 residues)	1.27	12	0.86	0.96	3.92
Hydrogen bonding (8 residues)	0.74	12	0.98	1.10	3.25
Hydrogen bonding (10 residues)	0.46	14	1.04	1.12	2.83
Hydrogen bonding (12 residues)	0.38	14	0.99	1.05	2.50

Table IV. Mean parameter weights and windows from Metropolis optimization analyses

Parameter	Weights		Window size	
	Mean	s.d.	Mean	s.d.
Accessibility	0.84	0.16	14.0	0.0
Protrusion	0.01	0.02	9.6	5.3
B-values	1.00	0.00	4.4	1.0
Ooi numbers	0.48	0.17	6.2	1.5
Secondary structure	0.49	0.18	13.2	2.4
Hydrogen bonding	0.36	0.22	17.4	2.5

nick-sites. The data for the optimal window lengths, as judged by the lowest energy E_{total} , are listed in Table III for the narrow specificity data set. For all the conformational parameters investigated (including all sub-types) the mean nick-sites parameter score is significantly above that of all residues, and indeed above that of putative sites, as judged by the positive discrimination factor scores. This is significant, as all the putative sites cut by narrow specificity proteinase (e.g. lysine/arginine for trypsin, glutamate for V8-proteinase) are likely to be at the surface and at flexible regions anyway. Thus, all the parameters studied here possess some additional predictive power for limited proteolytic sites above that of the inherent physical properties of the amino acids in question. This was also seen to be the case for the smaller number of broad specificity sites listed in Table I (data not shown). The DF_{nicks} scores were typically higher for this second data set as many more residue positions (including some hydrophobic ones) match the broader specificity requirements and consequently the mean parameter scores for the putative residues drop.

Figure 2 illustrates the normalized distributions of parameter scores for the optimal windows for the six selected parameters. The true nick-site residues cluster towards the right-hand-side of the distributions, demonstrating their suitability for prediction parameters. However, as is evident from the mean ranks and the distributions shown in Figure 2, the individual parameters are not perfect predictors. Indeed, the mean rank scores obtained for the parameters tested on the broad specificity data set ranged from 8.0 to 25.0. Some nick-sites clearly possess scores for some parameters that are down in the middle or lower end of the distribution scores.

The optimal smoothing window lengths tended to lie within four and 12 residues for most of the parameters tested against the narrow specificity sites and between six and eight for the broad sites (data not shown). The longer window lengths obtained for the narrow sites were probably due to the nature of the specificity of these proteinases, which cut at residues that would be expected to be at the surface. Hence, better discrimination is obtained by averaging over a large window compared with an isolated exposed/flexible residue. This is consistent with results from modelling experiments where at least 10–12 residues were shown to be involved in the local unfolding required for limited proteolysis (Hubbard *et al.,* 1994).

Some parameters performed better than others. The discrimination factors for temperature factors and accessibility were the highest and those for protrusion index and secondary structure parameters the lowest. In particular, the protrusion index appears to be the weakest predictor of the parameters studied, most probably due to the ellipsoidal approximation to protein shape used in the calculation leading to distorted values for particularly non-ellipsoidal proteins.

Based on the E_{total} data presented in Table III, the following parameter types were selected for the multiple parameter optimization trials: relative accessibility, α-carbon protrusion index, all atom B-values, 14 Å Ooi numbers, secondary structure parameters (helix = 0.5, strand = 0.0, coil = 1.0, disulphide penalty = 0.4), and 12-residue window non-local backbone hydrogen bonding.

Optimization of prediction parameters

The relative importance of the six parameters was further assessed via Metropolis optimization of the parameters using the prediction algorithm. The quality of nick-site prediction was significantly improved by combining the parameters. Even without optimizing the parameter weights and smoothing windows, in 9 of the 12 narrow data set proteins one of the nick-site residues was the top scoring position satisfying primary sequence requirements. For the narrow specificity data set, the Metropolis runs typically converged quickly within 2000 steps and found a set of weights and smoothing windows that further improved the prediction quality as judged by the prediction energy E_{total}. The simulations always converged to a solution where a true nick-site was ranked as the top scoring residue for all 12 proteins in the narrow specificity data set.

Table V. Linear correlation coefficients between prediction parameters

	Accessibility	Protrusion	B-values	Ooi numbers	Secondary structure	Hydrogen bonding
Accessibility	1.0	0.75	0.48	0.87	0.23	0.35
Protrusion	–	1.0	0.46	0.85	0.16	0.24
B-values	–	–	1.0	0.45	0.08	0.15
Ooi numbers	–	–	–	1.0	0.29	0.39
Secondary structure	–	–	–	–	1.0	0.82
Hydrogen bonding	–	–	–	–	–	1.0

Table VI. Optimized predictions of narrow and broad specificity nick-sites

Optimized energy	Parameter smoothing top scores/windows						Parameter weights						DF_{all}	DF_{nick}	Mean	
	a	b	c	d	e	f	a	b	c	d	e	f			rank	max
narrow E_{total}	14	6	4	6	14	12	0.7	0.0	0.9	0.3	0.6	0.2	1.87	1.71	1.00	12/12
narrow E_{total} (4 parameters)	8	-	4	12	-	18	0.7	-	0.8	0.6	-	0.8	1.83	1.65	1.00	12/12
broad E_{total}	14	6	8	4	6	12	0.8	0.2	0.5	0.2	0.7	0.8	1.73	1.71	2.62	3/8
broad E_{total} (4 parameters)	6	-	12	8	-	12	0.4	-	0.2	0.1	-	0.6	1.71	1.71	2.75	3/8

Parameters: a, accessibility; b, protrusion index; c, temperature factors (B-values); d, Ooi numbers; e, secondary structure parameters; f, non-local hydrogen bonding.

However, the same optimal solution was achieved from slightly different combinations of weights and windows for the six parameters. The mean values of these weights and windows for the lowest energy state over each of the 20 runs are listed in Table IV for the narrow specificity set. These data represent the relative discriminatory potential of the six parameters. As noted before, the protrusion index is a particularly weak predictor with a mean weight of only 0.01. Correspondingly, temperature factors were always the top weighted parameter and are therefore the most significant predictors, closely followed by accessibility. Consistent optimal window lengths were also apparent for some parameters.

The data in Table IV strongly suggested that not all the parameters were strong predictors of proteolytic susceptibility. There is clearly some overlap between the various parameters and indeed, the features are generally highly correlated as shown in Table V. For example, the hydrogen bonding function embodies some features of secondary structure by definition. Hence, the prediction optimizations were re-run with a reduced set of four parameters, removing the protrusion index and the secondary structure term. The results of these and the full six parameter optimizations are shown in Table VI and the data is represented graphically in Figure 3. Using both six and four parameters, a set of windows and weights were found that predict a nick-site residue to be top scoring for every protein in the narrow (high) specificity data set yielding a 100% successful prediction. Although the same success was not achieved with broader sites, the overall prediction was good, and no real loss of predictive quality was achieved using only four parameters. For the broad specificity nick-sites, accessibility and hydrogen bonding parameters dominate more than the temperature factors and secondary structure parameters, although they are both still important.

The prediction problem is considerably more challenging for the broad specificity proteinases due to the increased number of putative residues (for example, almost all residues are deemed to match the subtilisin primary specificity). Despite this, the top-scoring nick-site for each protein is predicted to be in the top 6% of all putative sites and the majority of the very top scoring residues lie within a few residues of one of the true nick-sites. This is well illustrated in Figure 4 by the prediction profiles for the broad data set. In all cases, the nick-sites lie close to the top of profile peaks. Indeed, for 1AVE, 1OPA, 1YPI and 8TLN the nick-sites are located in the tallest peak and the 2CST nick-site is the highest scoring putative site. Nevertheless, the ability to predict the precise site of limited proteolysis in every example seems beyond the scope of this approach and is likely due to the subtleties in primary and secondary subsite recognition as well as steric (local) unfolding factors.

For both data sets, different combinations of window and weight sets produced the same energy minimum. However, common trends were evident. The final weights from the Metropolis runs showed that for the narrow specificity sites, the most important parameters were accessibility, temperature factors, Ooi numbers, secondary structure and hydrogen bonding. Similarly, accessibility, temperature factors and hydrogen bonding were consistently highly weighted for the broad sites. The protrusion index was almost always the most lowly weighted term for both data sets. The optimal window lengths also showed consistencies. The accessibility window converged to 14 for the six parameter optimizations for both narrow and broad specificity sites and the window sizes for hydrogen bonding ranged between 12–18 for the lowest energy solutions. However, the window lengths changed when the optimization was reduced to only four parameters. This might reflect the fact that the system was over-determined with six parameters and the optimal weights and windows found at this level were affected by noise from superfluous parameters.

The prediction protocol optimized on the narrow specificity data set was tested on the broad specificity data set. Unsurprisingly, as shown in Table VII, the prediction scores obtained

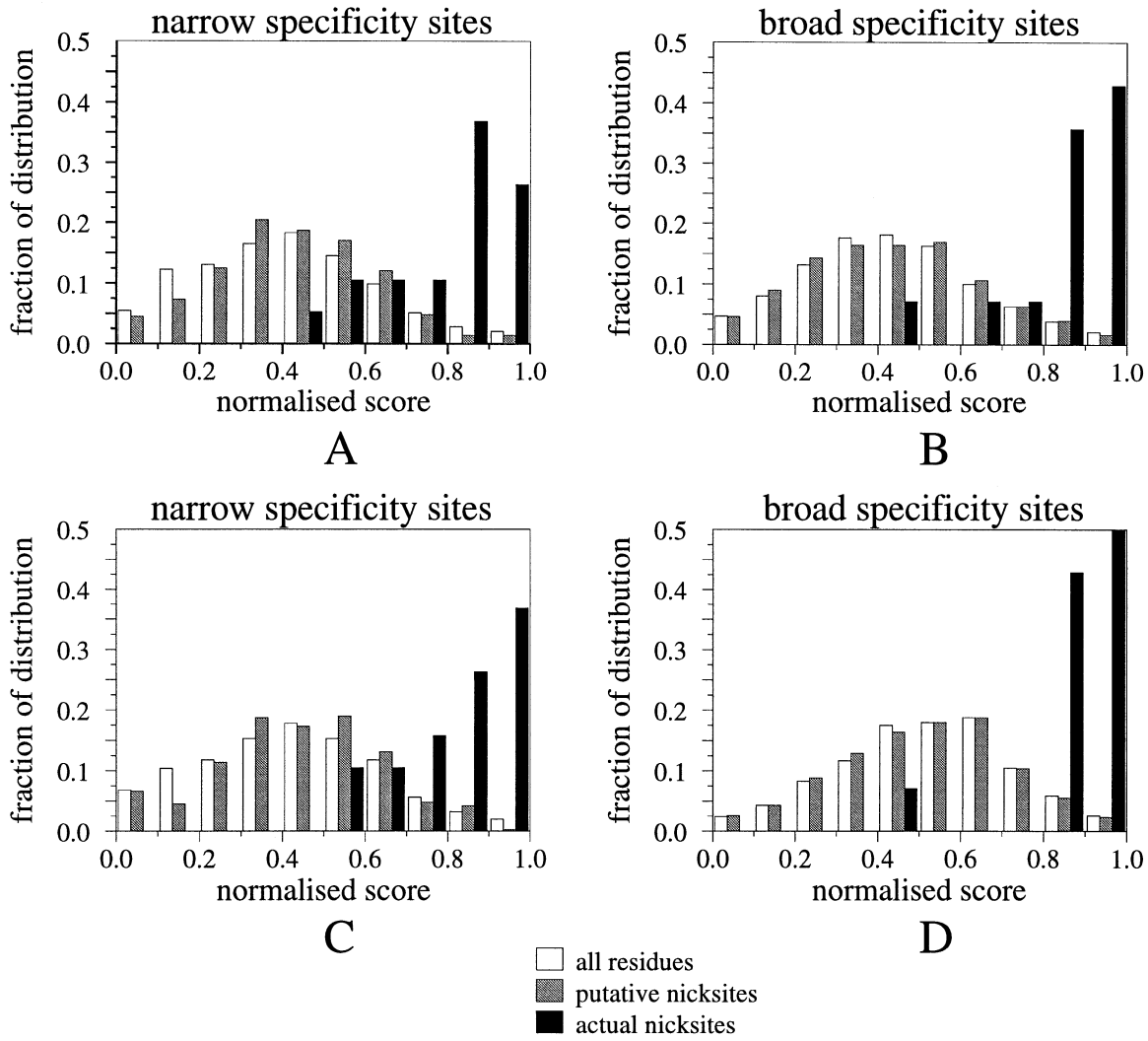

Fig. 3. Distributions of prediction scores for amino acids in the nicksite data sets after optimization. Distributions of amino acid prediction scores after Monte Carlo optimization for (**A**) the narrow specificity proteinase data set and (**B**) the broad specificity proteinase data set, both using the full six parameters. The same profiles are shown in (**C**) for the narrow specificity data set and (**D**) the broad specificity data set, using the reduced set of 4 parameters. The white bars represent all residues in the data set, the grey bars putative sites, and the black bars the nicksite residues.

were slightly inferior to the most optimal weights and windows for broad specificity proteinases. Nevertheless, the prediction is still almost as good, particularly when considering the increased difficulty in predicting broad specificity and the narrow specificity weights and windows may be applied universally to nick-site prediction. This applies when using four or six parameters.

The ability of the algorithm was also tested by a jack-knifing procedure where the protein under consideration is removed from the data set, the weights and parameters were re-optimized without it, and then the prediction is reapplied to that protein. When applied to the narrow specificity data set the overall quality of prediction was only slightly inferior. With six parameters nine out of 12 narrow specificity nick-sites were ranked as top scoring with a mean rank of 1.50. This improved slightly with the reduction to only four parameters yielding a mean rank of 1.3.

These results suggest that the algorithm is a useful predictor of proteolytic susceptibility and that it is worthwhile finding the most optimal parameters to accomplish this. To confirm the validity of the approach, the probability of achieving a perfect 12 out of 12 prediction was calculated by conducting a systematic search of parameter space for the four parameter narrow specificity data set. This was estimated to be only a 0.2% chance. Similarly, randomly chosen weights and windows would only correctly predict a true nick-site for seven out of 12 proteins.

Discussion

A critical analysis of protein conformational parameters has demonstrated their ability to distinguish limited proteolytic sites from other putative cleavage sites for proteins of known structure for proteinases of both narrow (e.g. trypsin) and broad (e.g. subtilisin, thermolysin) specificity. The results confirm earlier conclusions that nick-sites are found at exposed, protruding and flexible regions of protein structure (Fontana *et al.*, 1986; Novotný and Bruccoleri, 1987; Vita *et al.*, 1988; Fontana, 1989; Hubbard *et al.*, 1991, 1994) which are typically loops or turns, rarely in or near helices, but apparently never in extended β-structure. Similarly, substantial segments of contiguous residues local to the scissile peptide must possess

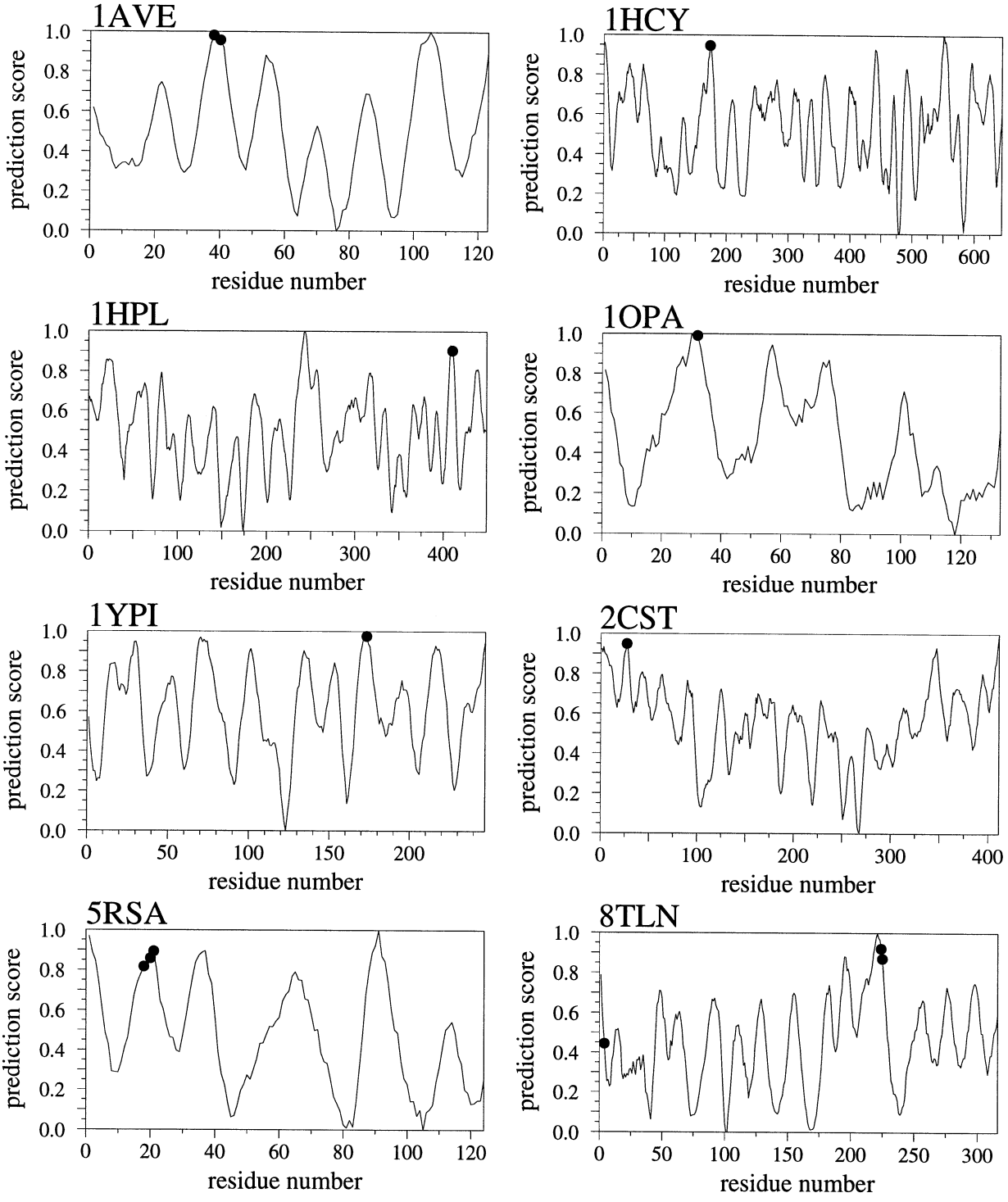

Fig. 4. Prediction profiles for proteins cut by broad specificity proteinases. The prediction profiles for the eight proteins from the broad specificity data set are shown, using the weights and window from the Monte Carlo optimization runs. Limited proteolytic sites are indicated by filled circles on the profiles. The proteins shown are 1AVE, avidin; 1HCY, haemocyanin; 1HPL, lipase; 1OPA, cellular retinol-binding protein II; 1YPI, triose phosphate isomerase; 2CST, aspartate aminotransferase; 5RSA, ribonucelase; 8TLN, thermolysin.

the requisite properties needed to allow local unfolding for subsequent limited proteolytic cleavage.

The protrusion index is a relatively weak predictor as it is an approximation to true protein shape. Large proteins might more usefully be broken down into constituent subunits and/or domains which are more likely to be globular or, alternatively, the protrusion index can be eliminated entirely without compromising the algorithm unduly. As there is some overlap in the parameters as shown by the correlation coefficients in Table V, the chief determinants of limited proteolytic susceptibility may be divided into three groups:

1. Exposure—as characterized by accessibility and Ooi numbers. Although not an absolute requirement, a nick-site is

Table VII. Comparison of predictions for broad specificity nick-sites

Protein code	Number of putative nick-sites	Top scoring nick-site ranking			
		narrow specificity Full 6 parameter set	optimized parameters Reduced 4 parameter set	broad specificity Full 6 parameter set	optimized parameters Reduced 4 parameter set
1AVE	41	2	2	1	1
1HPL	165	3	3	4	4
1HCY	579	14	11	6	7
1OPA	36	2	2	1	2
1YPI	216	9	10	2	2
2CST	149	2	1	1	1
5RSA	109	11	15	2	1
8TLN	294	3	2	4	4

more likely to be situated in a region close to the protein surface so that local unfolding is more easily accomplished.
2. Flexibility—as characterised by X-ray crystallographic temperature factors, which give a measure of the dynamic properties along the protein chain, obviously critical for local unfolding and adaptation to the enzyme's active site.
3. Local interactions—as characterized by secondary structure and hydrogen bonding. A good candidate for local unfolding and adaptation must not be tied down by interactions such as disulphide bridges or hydrogen bonding (such as within regular secondary structure).

Undoubtedly, the key determinant is the ability to unfold locally and adapt to the enzyme's active site. The question remains as to which parameters are the best indicators of this ability. Despite the importance of these three features, an over-reliance on any one may lead to a false prediction. For example, as pointed out by Fontana and co-workers (1997a,b) many putative nick-sites on a protein surface are not cleaved. Similarly, although important, temperature factors are an imperfect measure of the true segmental mobility. This is because they report on static disorder, as well as thermal motions. They are also distorted by intermolecular crystal packing interactions, which reduce the flexibility of those residues involved, manifesting in reduced temperature factors. This is the case for the region containing several of the ribonuclease (5RSA) nick-sites from residues 18–24 which make intermolecular contacts in the crystal. This serves to damp down the apparent mobility of this segment and affect the associated thermal factors. However, it may be possible to partially address this problem of crystal-masked flexibility by accounting for packing affects and modifying atomic B-values (Scheriff *et al.*, 1985).

Similarly, protein–ligand interactions can profoundly effect the proteolytic susceptibility of a protein, stabilizing or destabilizing it (Fontana, 1989; Jamison *et al.*, 1994; Ellison *et al.*, 1995). If the incorrect apo- or holo-protein structure is unavailable the prediction will be affected. In the case of the retinol-binding proteins (1OPA) this is likely to affect the prediction profoundly as the retinol (or analogous ligand) stabilizes the protein and reduces its susceptibility to proteolysis (Jamison *et al.*, 1994). Although there exists a large amount of X-ray and limited proteolysis data, the only structure available for these proteins in the apo-form is the cellular retinol-binding protein II (1OPA). Several other limited proteolytic systems were excluded from this study as only the holo-form crystal structure was available which was resistant to limited proteolysis. Indeed, the prediction algorithm applied to these systems yields rather poor predictions. Use of the incorrect apo- or holo- crystal form had a more deleterious effect on the prediction than use of an alternative structure from another species, as these proteins were generally well predicted. For example, Gly28 in aspartate aminotransferase (2CST) is the highest scoring putative thermolysin site despite the fact that the *S.Solfataricus* structure is not yet available and the chicken heart homologue structure is used instead.

Another factor not considered is the 'steric fit'. Some sites may be geometrically more disposed to local unfolding and subsequent docking. Although this can be assessed (Hubbard *et al.*, 1994) it is compute-intensive and difficult to integrate into the prediction algorithm. Furthermore, the steric accessibility to the active site cleft can vary between proteinases of the same primary specificity such as kallikrein and trypsin (Chen and Bode, 1983).

Put into a biochemical context, limited proteolysis is not a discrete process where every bond is either susceptible or resistant. Dynamics, and hence kinetics, must play a role in whether a particular cleavage will be observed, determined not only by the enzyme and substrate ratio, but also by the diverse range of conditions under which the experiments were conducted. In some cases, several bonds might be accessible to proteolysis in native state conditions, whilst under slightly different (retarding) experimental conditions, only a single site may be observed to be cleaved at a slow rate. This site might score lowly, although it is the highest of the putative sites for a given proteinase. Hence, the ranking of sites is also of great importance. However it is rare that all rate constants for all susceptible sites in a given protein and the true 'rank' order of limited proteolysis are experimentally determined. Thus, not unreasonably, we select the top-scoring nick-site in each protein for the calculation of mean ranking since the true 'first cut' site is rarely unequivocally determined. A further complication arises due to the nature of the data. Because the proteolysis has in some way been limited, there may well be more susceptible sites in the substrate proteins than have been measured experimentally. As there is no way for this to be ratified without re-performing all the experiments under a vast array of differing conditions we have simply divided sites into either nick-sites or 'not nicksites' which may not be universally true. These points are highlighted by the range of reaction temperatures and the second-order rate constants estimated from the literature; the latter ranging over several orders of

magnitude from 0.002 to 0.2 $\mu M^{-1}.min^{-1}$ (Table I). Clearly, these rates are strongly dependent on experimental conditions and it would be folly to try and correlate prediction scores with rate constants given such non-standard experimental conditions. It therefore remains a challenge for the future to distinguish further the true 'susceptibility' of each peptide bond and account for kinetic factors. This would require the ability to predict the energy barrier for each protein segment to unfold locally. Studies in our laboratory are currently underway to achieve this goal.

Despite the limitations discussed here, this approach remains successful for proteinases of different specificity with the prime determinants of susceptibility remaining the same: flexibility, exposure and the ability to unfold locally. It should be stressed, however, that the algorithm is not a definitive prediction tool and the results should be interpreted carefully and with caution given the limitations described here.

This approach also has potential for application to other surface-correlated features of proteins. Properties such as antigenicity are surface-correlated (Thornton *et al.*, 1986) as are post-translational modifications such as glycosylation and phosphorylation. The latter represent additional examples of the constraints placed by tertiary structure on the modification of sequence patterns, for example by *N*-glycosylation in the case of Asn-X-Ser/Thr motif or by cleavage in the case of proteolysis. The potential of this generic approach to these kind of prediction problems is being evaluated in our laboratory.

Software availability

Nickpred, the program used in this work is available to the biological community via the World Wide Web and via email. Information on the WWW version is located at: http://sjh.bi.umist.ac.uk/nickpred.html. For more information on the Email version, send mail to nickpred@sjh.bi.umist.ac.uk. with the word HELP in the body of the mail, on a line on its own.

Acknowledgements

SJH acknowledges the support of the Wellcome Trust via an Advanced Retraining Fellowship 044959/Z/95/PMG/MJD.

References

Abousalham,A., Chaillan,C., Kerfelec,B. and Foglizzo,E. (1992) *Protein Engng*, **5**, 105–111.
Arnone,M.I., Birolo,L., Giambernini,M., Cubellis,M.V., Nitti,G., Sannia,G. and Marino,G. (1992) *Eur. J. Biochem.*, **204**, 1183–1189.
Bek,E and Berry,R. (1990) *Biochemistry*, **29**, 178–183.
Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 532–542.
Blow,D.M. (1976) *Chem. Res.*, **9**, 145–152.
Bode,W. and Huber,R. (1992) *Eur. J. Biochem.*, **204**, 433–451.
Bond,J.S. and Beynon,R.J. (1987) *Mol. Aspects Med.*, **9**, 175–287.
Chen,Z. and Bode,W. (1983) *J. Mol. Biol.*, **164**, 283–311.
Chothia,C. (1976) *J. Mol. Biol.*, **105**, 1–14.
Ellison,D., Hinton,J., Hubbard,S.J. and Beynon,R.J. (1995) *Protein Sci.*, **4**, 1337–1345.
Fontana,A., Fassina,G., Vita,C., Dalzoppo,D., Zamai,M. and Zambonin,M. (1986) *Biochemistry*, **25**, 1847–1851.
Fontana,A. (1989) In Kotyk,A., Skoda,J., Paces,V. and Kostka,V. (eds) *Highlights of Modern Biochemistry*. VSP International Science Publishers, Zeist, The Netherlands, pp. 1711–1726.
Fontana,A., Zambonin,M., Polverino de Laureto,P., De Fillipis,V., Clementi,A. and Scaramella,E. (1997a) *J. Mol. Biol.*, **226**, 223–230.
Fontana,A., Polverino de Laureto,P., De Fillipis,V., Clementi,A. and Scaramella,E. and Zambonin,M. (1997b) *Fold. Des.*, **1**, 17–26.
Ghelis,C, Tempete-Gaillourdet,M. and Yon,J.M. (1978) *Biochem. Biophys. Res. Commun.*, **84**, 31–36.
Hermodson,M.A., Ericsson,L.H., Neurath,H. and Walsh,K.A. (1973) *Biochemistry*, **12**, 3146–3153.
Higaki,J.N. and Light,A. (1985) *Anal. Biochem.*, **148**, 111–120.
Hu,Y., Fenwick,C. and English,A.M. (1996) *Inorg. Chim. Acta*, **242**, 261–269.
Hubbard,S.J., Campbell,S.F. and Thornton,J.M. (1991) *J. Mol. Biol.*, **220**, 507–530.
Hubbard,S.J., Thornton,J.M. and Campbell,S.F. (1992) *Faraday Discuss.*, **93**, 13–23.
Hubbard,S.J., Eisenmenger,F. and Thornton,J.M. (1994) *Protein Sci.*, **3**, 757–768.
Iriarte,A., Hubert,E., Kraft,K. and Martinez-Carrion,M. (1984) *J. Biol. Chem.*, **259**, 723–728.
Jamison,R.S., Newcomer,M.E. and Ong,D.E. (1994) *Annu. Rev. Biochem.*, **46**, 331–358.
Kabsch,W. and Sander,C. (1993) *Biopolymers*, **22**, 2577–2637.
Kraut,J. (1977) *Ann. Rev. Biochem.*, **46**, 331–358.
Laskowski,M.,Jr. and Kato,I. (1980) *Annu. Rev. Biochem.*, **49**, 593–626.
Lee,B. and Richards,F.M. (1977) *J. Mol. Biol.*, **55**, 379–400.
Monks,S.A., Gould,A.R., Lumley,P.E., Alewood,P.F., Kem,W.R., Goss,N.H. and Norton,R.S. (1994) *Biochim. Biophys. Acta*, **1207**, 93–101.
Monsalve,R.I., Menéndez-Arias,L., López-Otin,C. and Rodriguez,R. (1990) *FEBS Lett.*, **263**, 209–212.
Neurath,H. and Walsh,K.A. (1976) *Proc. Natl Acad. Sci. USA*, **73**, 3825–3832.
Neuteboom,B., Jekel,P.A. and Beintema,J.J. (1992) *Eur. J. Biochem.*, **206**, 243–249.
Nishikawa,K. and Ooi,T. (1986) *J. Biochem.*, **100**, 1043–1047.
Novotný,J. and Bruccoleri,R.E. (1987) *FEBS Lett.*, **211**, 185–189.
Ottensen,M. (1967) *Annu. Rev. Biochem.*, **36**, 55–76.
Price,N.C. and Johnson,C.M. (1989) In Beynon,R.J., Bond,J.S. (eds) *Proteolytic Enzymes: A Practical Approach*. Oxford, UK, IRL Press, pp. 163–180.
Richards,F.M. and Vithayathil,P.J. (1959) *J. Biol. Chem.*, **234**, 1459–1464.
Rupley,J.A. and Scheraga,H.A. (1963) *Biochemistry*, **2**, 421–431.
Schechter,I. and Berger,A. (1967) *Biochem. Biophys. Res. Commun.*, **27**, 157–162.
Sheriff,S., Hendrikson,W.A., Stenkamp,R.E., Sieker,L.C. and Jensen,J.H. (1985) *Proc. Natl Acad. Sci. USA*, **82**, 1104–1107.
Stephen,A. (1993) PhD Thesis, University of Manchester Institute of Science and Technology, Manchester, UK.
Sun,A-Q., Ümit Yüksel,K. and Gracy,R.W. (1993) *J. Biol. Chem.*, **268**, 26872–26878.
Taniuchi,H. and Anfinsen,C.B. (1968) *J. Biol. Chem.*, **243**, 4778–4786.
Taniuchi,H., Anfinsen,C.B. and SodjaA. (1967) *Proc. Natl Acad. Sci. USA*, **58**, 1235–1242.
Taylor,W.R., Thornton,J.M. and Turnell,W.G. (1983) *J. Mol. Graph.*, **1**, 30–38.
Thornton,J.M., Edwards,M.S., Taylor,W.R. and Barlow,D.J. (1986) *EMBO J.*, **5**, 409–413.
Vereijken,J.M., Schwander,E.H., Soeter,N.M., Beintema,J.J. (1982) *Eur. J. Biochem.*, **123**, 283–289.
Vidigni,A., De Filippis,V., Zanotti,G., Visco,C., Orsini,G. and Fontana,A. (1994) *Eur. J. Biochem.*, **226**, 323–333.
Vita,C., Dalzoppo,D. and Fontana,A. (1985) *Biochemistry*, **24**, 1798–1806.
Vita,C., Dalzoppo,D. and Fontana,A. (1988) In Chaiken,I.M., Chiancone,E., Fontana,A., Neri,P., (eds). *Macromolecular Recognition: Principles and Biotechnological Applications*. Clifton, New Jersey: Humana Press.
Walsh,M., Stevens,F.C., Kuznicki,J. and Drabikowski,W. (1977) *J. Biol. Chem.*, **252**, 7440–7443.

Received June 26, 1997; revised January 8, 1998; accepted January 15, 1998