Simple intrasequence difference (SID) analysis: an original method to highlight and rank sub-structural interfaces in protein folds. Application to the folds of bovine pancreatic trypsin inhibitor, phospholipase A₂, chymotrypsin and carboxypeptidase A

Leighton Pritchard¹, Linda Cardle², Susanne Quinn and Mark Dufton³

Department of Pure and Applied Chemistry, University of Strathclyde, 295 Cathedral Street, Glasgow G1 1XL, UK

¹Present address: Cledwyn Building, Institute of Biological Sciences, University of Wales, Aberystwyth SY23 3DD, UK
²Present address: Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, UK
³To whom correspondence should be addressed.
E-mail: mark.dufton@strath.ac.uk

We present Simple Intrasequence Difference (SID) analysis, a novel bioinformatic technique designed to help comprehend the properties of protein fold topologies. The analysis grades numerically every residue position in a given protein 3D structure according to the topological situation of the position in the folded chain. This results in an expression of the potential contribution of each residue position and its vicinity towards the integrity of the molecular conformation. Contiguous highly graded residues delineate the sub-structural interfaces that arise from the presence within the molecular fold of discrete domains and sub-domains. This comprehensive rendering of the internal arrangement of chain interfacing helps predict the potential for site-specific inductions (e.g. via mutations or ligand binding) of conformational change in the fold. Whereas SID analysis of single folds can convey an idea of the basic potential for topological adjustment in the protein family, comparative SID analysis of related folds focuses attention on those areas of the family fold where evolutionary changes, activation events and ligand binding have had the most topological impact. For demonstration, SID analysis is applied to the folds of pancreatic trypsin inhibitor (Kunitz), phospholipase A₂, chymotrypsin and carboxypeptidase A. We find that many of the potentially vulnerable sub-structural interfaces tend to be protected in the fold interior, in many cases stabilised by disulfide bridges spanning the interface. However, the most prominent interfaces tend to be externally accessible, without remedial stabilisation by disulfide bridges. These latter interfaces are associated so closely with the known functional sites that alterations to the interfacial juxtapositions should influence recognition and catalytic behaviour directly. This shows how side chain mutations, chemical modifications and binding events remote from the sites can nevertheless adjust, via interfacial realignment, the conformations and emergent properties of the sites. The close association also provides clear opportunities for interfacial rearrangements to follow intermolecular recognition events in the sites, facilitating translation of the binding into adjustment of the molecular conformation in areas distant from the sites. As a direct consequence of the topological arrangements, a large proportion of the molecular structure has the capacity to shape the character of the functional sites and, conversely, binding at these sites has the potential to radiate influence to the rest of the molecule. For the enzymes considered, the evidence is consistent with the possibility that primary and secondary binding by the substrate enhances catalytic efficiency by imposing conformational change upon the catalytic centre via adjustments to the fold. This influence may be expressed as favourable adjustment of the catalytic geometry, transition state ensemble, energy propagation pathway, or as a physical strain exerted on the substrate bond to be cleaved. The scale of the adjustments, and their importance to the mechanisms, may have been seriously underestimated.

Keywords: conformation/domain/mechanism/prediction/protein

Introduction

Within a single protein molecule, the potential for conformational adjustment differs substantially from locality to locality. Some areas have little scope for residue reorientation within themselves because they are formed from stable substructures. Other areas may have significant potential for relative adjustment because they encompass, or straddle, one or more sub-structural interfaces (Gerstein et al., 1994). An interface in this context can be defined as any conjunction of stable structural elements (e.g. secondary structures, domains, fixed residue ensembles) wherein the relative juxtaposition of the elements is a better prospect for change than the structures of the elements themselves. The relative positions of interactive moieties in recognition sites can be manipulated from remote locations by the reorientation of substructures (e.g. secondary structures, domains), provided that these recognition sites are located astride appropriate interfaces. Adjustment might be induced, for example, by mutations at the interface (as recently highlighted in the case of aspartate aminotransferase) (Oue et al., 1999), or by an intermolecular contact event. Ligand binding at a surface site overlying an interface might also induce a conformational change, either locally or globally, in the protein via indirect impact on the interface (e.g. destabilization of the interface by host residues forming intermolecular interactions with the incoming ligand at the expense of participating in intramolecular interactions).

Intramolecular interfaces in protein structures vary considerably in extent and complexity. A surface-accessible site may have an interface between two, three or more sections of chain beneath it. Each interface might be expected to differ in its sensitivity to perturbing influences. The more complex the interface that underlies a site, the more of the molecule that is required to maintain the disposition of the site, and the more likely it is that reorientation of the recognition moieties could be induced by structural perturbations remote from the site.

Given this background, a systematic search procedure for the detection of vulnerable sub-structural interfaces in protein
structures, which also includes a means of ranking such interfaces according to their likely contribution to molecular stability, becomes a desirable tool for the protein scientist. In particular, such a methodology could assist prediction of those sites on a protein surface where a recognition event could have the most potential for causing a significant change in conformation, and also where in the molecular structure natural or artificial sequence modification is most likely to have a conformational impact and where it may not.

The starting point we chose in developing our methodology was to consider the topological arrangement of the folded polypeptide chain. In natural proteins, these topologies are evidently critical to the constitution and execution of their chemical properties, as tertiary structure is more conserved than primary sequence generally. As is now clear, not only does nature use, and reuse, only a small selection of all possible fold topologies (Richardson, 1981; Zhang and Delisi, 1998), but also in some cases topological similarity exists beyond any obvious amino acid sequence similarity. In the latter instance, either evolutionary preservation of the topology has been a stronger selective pressure compared to sequence preservation, or similar topologies have evolved convergently. As demonstrated by both natural and artificial sequence evolution, the natural topologies also have a great capacity to absorb the effects of multiple side chain mutations without catastrophic effect on fitness (describable as ‘soft failure’) (Prichard and Dufton, 2000). It is this very high level of topological preservation despite primary sequence variation that allows for accurate prediction of similar structures amongst homologous protein sequences.

Most importantly from our point of view, however, the topology provides the basic definition of the number, character, and contribution to stability of the sub-structural interfaces (or intramolecular ‘fault lines’) that can underlie the recognition surfaces. Although some proteins, such as calmodulin and the serpins, undergo gross conformational changes associated with function when the local secondary structural environment changes, or the polypeptide chain is cleaved, the fold and sequence of a protein defines much of the potential for surface adjustment via motion or ligand binding. We investigated how we could search systematically a fold topology in order to highlight its interfaces and grade them according to their likely contribution to molecular stability. Having identified the regions that are theoretically vulnerable, would there be corroborating experimental evidence that the interfaces were associated with sites of intermolecular recognition? If there were, could a prediction be made that the interactive site is modifiable by, or able to generate, some kind of intramolecular conformational change? In this paper, we analyse four well-evolved topologies (Richardson, 1981; Zhang and Delisi, 1998), does nature use, and reuse, only a small selection of all possible fold topologies (Richardson, 1981; Zhang and Delisi, 1998), but also in some cases topological similarity exists beyond any obvious amino acid sequence similarity. In the latter instance, either evolutionary preservation of the topology has been a stronger selective pressure compared to sequence preservation, or similar topologies have evolved convergently.

The choice of sphere radius is important as amino acid side chains are of varying effective length and direction, and thus influence, depending on their orientation and environment. A constant sphere radius generalizes such factors and concentrates on relations due to the fold topology. For this analysis, the sphere radius was set at a constant value of 7 Å. This radius has previously been found to be useful in similar clustering methods based on the Cα backbone (Cardle and Dufton, 1994, 1997), and represents the approximate direct influence of any substitution at the site under consideration (Bajaj and Blundell, 1984).

**Step 1: Definition of a volume around each residue**

The spatial volume considered around each residue should be a suitable (in the context of the information one expects to gain) approximation of the range and direction of the physicochemical influence exerted and experienced by that residue. It is nominally possible to construct a volume of any arbitrary geometry and size about any component of the considered residue. For our primary investigation, however, we considered only the simple, computationally inexpensive, approach of constructing a sphere about the α-carbon of each residue. This spherical volume, as well as being easy to assess and automate, is relatively insensitive to structural changes between homologues other than a gross change in the fold. Thus, while it is a suitable basis for assessing properties of the overall fold that do not involve, or are relatively unaffected by, details of the individual side chain character and 3D orientation of each residue, it is unsuitable for investigating the interactions between individual side chains. Such analyses would be better served by, for example, including considerations of side chain orientation in the considered volume.

**Step 2: Clustering**

Each residue position in the folded protein chain is considered in turn, proceeding in sequential order from the N-terminus to the C-terminus. Other residue positions are added to form a cluster about the considered residue if they are found within the defined (here, spherical) volume about the α-carbon. For this investigation, residues were added to the cluster only if their Cα atoms were contained within the defined volume. Other rules for determining whether a residue should be included in the cluster (e.g. hydrophobic residues only, atoms in the volume other than Cα, etc.) are possible, however. Again, our decision reflects the desire to investigate gross...
Step 3: Scoring the cluster

Defining the volume around a residue’s α-carbon atom and determining clusters are general procedures that can be applied to several different modes of analysis (e.g. Cardle and Dufton, 1994, 1997). It is the nature of the scoring mechanism that individualises this method and from which we have derived its name (SID: Simple Intrasquence Difference). At its simplest, SID scoring calculates a value related to the maximum chain separation of residues found within the cluster obtained in step 2 above. Three such methods of scoring are described here, though more are possible.

(a) Simple difference (highest – lowest, HL)

By this method, the SID score for a cluster is the chain separation of the two most chain-extreme residues in that cluster, i.e. the lowest residue number found in the cluster subtracted from the highest.

This score indicates how distant in the primary sequence the most extreme parts of the structure that have packed into the same volume of space. Hence a cluster that contains parts of two or more chain segments that are far apart on the polypeptide chain will have a high score, while a cluster that contains only a single contiguous section of chain will have a low score (Figure 1, Table I). A cluster overlapping an interface will have a high score, whereas a cluster found within an element will have a low score.

There may be an ‘intuitive’ expectation that the highest scores will originate from residue positions close in sequence to the chain termini. In practice, only for small proteins does there appear to be a ‘terminal effect’ relating to the score distribution, and even then it seems to be more a reflection of intrinsic fold properties than methodological artefact. With large folds, high SID scores can occur anywhere in the sequence.

(b) Simple difference (greatest gap, GG)

This scoring method requires the ordering of all the positions found within a cluster into numerical order. The difference between consecutive residue numbers is then calculated. The largest such value is taken to be the SID score for that cluster.

Where only two non-overlapping chain segments are found within a cluster, a high score is obtained. Where only a single segment of chain is found, a low score results. However, the important distinction from the HL score is that when more than two segments of chain are found in the cluster, the GG score gives a score that is lower than the corresponding HL score.

Where the GG score is equal to one (1), the cluster encompasses only chain-local residues within the structure. The core positions of these clusters may represent the nuclei of structurally stable and cohesive elements of local structure. Substitutions of these side chains might be expected to result primarily in local structural changes.

(c) Differential score (DIFF)

The GG score is most useful when subtracted from the HL score to produce the differential score:

$$\text{DIFF} = \text{HL} - \text{GG}$$

Where the GG and HL scores are similar (as in single local sections of structure and junctions between two chain-distant structures) DIFF is small. Where HL > GG, i.e. at the junction of three or more sections of structure, DIFF is large (Figure 1, Table I). A low differential score may thus indicate a simple interface between two sections of chain or an element. A large differential score, however, would likely indicate a multi-way interface between three or more sections of structure. Such junctions between several sections of structure are potentially significant, as they are difficult to stabilize by conventional secondary structural arrangements, and so represent a strong potential for relative motion.

The type of interface enclosed by a particular cluster can
be identified absolutely by consideration of the HL, GG and DIFF scores for that cluster, and the number of residues that contribute to the cluster (Figure 1).

(d) Comparing SID scores across members of a protein family

In addition to calculating SID scores for single exemplars of a protein family, it is also desirable to find out how SID scores for the family fold may be changed as a result of primary sequence evolution, activation mechanisms, ligand binding, chemical modification and so forth. When comparing SID data for homologous folds, or different states of the same fold, we obtained the simple standard deviation in score for every residue position in the chain because the crystal structures available represent only a small sample of the population of available conformations. This measure conveys the inherent variation in the score for each cluster and can be taken as a guide to regions of potential motion or adjustment in the molecule.

Where the standard deviation is close to zero, there is little or no variation in SID score across the structures, which implies that the 3D environment around the position in question is narrowly defined, and not prone to conformational changes. Where the deviation is large, however, the particular members of the cluster vary depending on the specific structure, which may imply some structural or conformational rearrangement between different forms that merits further investigation.

Step 4: Selection of ‘significant’ SID scores

The outcome of the SID calculation for a 3D structure is a value for each residue position in the chain that reflects its topological situation. In view of the fact that fold topologies are so highly conserved in protein families, the results can be extrapolated to other related variants with a confidence in proportion to the extent of the sequence homology. All the SID values obtained for a chain fold are potentially diagnostic because they express site-specific topological ‘viewpoints’ on behalf of the residues concerned. By extension, the magnitude of a SID value can therefore be treated as a reflection of the impact that a substitution, modification or perturbation of the residue side chain might have on the molecular conformation as a whole. There is no threshold definable above or below which a SID score for a residue position in a fold can be regarded as truly ‘significant’ or ‘non-significant’. Similarly, although it is valid to compare scores between different variations or modifications of the same family fold, it is not appropriate to try and compare scores between different fold types (i.e. different protein families). The principal difficulty here is that the range of SID scores for a fold will depend on its overall chain length. The power of SID analysis lies in its ability to assess the characteristics of specific protein topologies internally from the standpoints of each of the residues that comprise its backbone, and in the rendering of that complex 3D information into a simple numerical output for easy comprehension.

To extract ‘significance’ from the SID scores, there are different ways of proceeding depending on whether only one fold is being considered, or a series of similar folds are being compared. For an individual fold, the residue positions can be ordered from the highest scoring to the lowest, and working from the top score downwards, their clustering in the 3D fold can be examined (see below for our approach). The effect of this procedure is to highlight the most prominent interfaces in the molecular fold topology (in terms of linear chain separation) first and then to either extend these interfaces, or nucleate less prominent new ones, by systematically taking into account lower and lower scores. Alternatively, the lowest SID scores can be considered first, such that residue positions with relatively ‘minor’ roles are clustered first to define apparently autonomously folded structural elements, sub-domains and domains. Thereafter, consideration of the higher scores brings into focus the interfaces between these elements and finally the most marked of the internal interfaces. Whichever route is chosen, the complete list of residue positions needs to be traversed to convey to an observer the relative internal contributions of residue positions and their 3D ensembles towards defining and maintaining the fold topology and its capacity for adjustment.

If desired, one can focus on ‘exceptional’ interfaces in a fold rather than developing a complete view of the topological implications, by determining ‘outliers’ in the SID data. To do this, the median and the upper (UQ) and lower (LQ) quartiles of the SID score distribution (HL, GG or differential) should be determined. The inter-quartile distance (IQD) is then defined as (UQ – LQ). Clusters that lie outside the range defined by (UQ + 1.5 × IQD) > SID score > (LQ – 1.5 × IQD) are classed as outliers. We used the boxplot function of Kaleida- graph v3.07 (Synergy Software).

For each of the protein folds analysed in this paper, we chose to work down from the highest SID scores so that the most significant sub-structural interfaces were highlighted first. From there, we incorporated successively lower scores to detect lesser and ‘tributary’ interfaces such that the molecule became divided successively into quasi-autonomous domains and sub-domains. To simplify data presentation, limit the length of this paper and demonstrate the potential of SID analysis in its fundamental form, our reporting is based on the ‘HL’ calculation option and the illustrations are of the ‘HL’ data. The ‘GG’ and subsequent ‘DIFF’ calculation options (results not shown) were employed in the comparison of uncomplexed and complexed conformations to define more closely the nature of the differences between them.

When comparing the SID data for two or more related folds (i.e. either folds of related variants, or different states of one fold), the interest lies in discovering which residue positions have ‘sensed’ the largest changes in their topological situation. For example, do they sense a lowering in SID value (such as might be caused by a local interface widening or opening) or an increase in value (interface narrowing or closing)? Again, what constitutes a quantitatively ‘significant’ or ‘insignificant’ change is internally defined and difference values obtained for one fold type have no inherent comparability with such values obtained for a contrasting fold type. Having been made aware of the localities in the fold that have given rise to the most pronounced differences in SID score, one must return to a comparison of the original 3D structures to establish the actual scale of the changes.

In our examination of the data, we were particularly interested in prominent intramolecular interfaces that appeared to be vulnerable to intermolecular recognition events. Such interfaces, if disturbed, would have the potential to instigate an adjustment of the chain topology, and therefore a conformational change in the protein, in response to the binding. If such interfaces could be associated with known binding properties, then involvement of conformational changes in the translation of the binding into an action (e.g. catalysis or activation) needed to be considered seriously. Thus, we needed
We selected four exemplars of the best known and experimentally well-studied protein folds to examine initially, namely those of bovine pancreatic trypsin inhibitor (BPTI) (PDB code 4pti), secretory phospholipase A2 (2bp2), chymotrypsin (2gmt) and carboxypeptidase A (2ctb), from the PDB database (http://pdb-browsers.ebi.ac.uk). This selection of folds covers sequence lengths from 58 to 307 residues and the folds contain various mixtures of classical secondary structure elements, and in the two larger examples, readily identifiable domains. Brief descriptions of the four families are given below.

The pancreatic trypsin inhibitor fold is composed of around 58 amino acid residues, which are formed into a twin-stranded β-sheet with two helical sections packed against it. The fold, which is stabilised by three disulfide bridges, is representative of a large family found throughout living organisms. Most examples (BPTI included) act as serine proteinase inhibitors (e.g. of the chymotrypsin considered herein) by binding strongly within the enzyme’s substrate binding site. Some examples found in snake venom can block potassium (dendrotoxin) or calcium (calicidin) channels and are thus neurotoxic. These actions (protease inhibition and neurotoxicity) are usually mutually exclusive.

The phospholipase A2 fold is comprised of about 123 residues and contains five sections of helix, a twin-stranded β-sheet and several disulfide bridges whose location varies between enzyme subclasses. The fold is representative of a large family of enzymes variously capable of hydrolysing 3-sn-phosphoglycerides at the C2 position in either micellar or bilayer form. This example acts primarily on micellar substrates after activation and usually requires a calcium atom for activity. Some examples from snake venom are additionally myotoxic and/or neurotoxic.

The chymotrypsin fold comprises 245 residues formed into two predominantly β-sheeted domains. Chymotrypsin is the founding member of the large serine proteinase family of enzymes (e.g. trypsin, elastase, kallikrein), all of which are capable of cleaving protein sequences internally, but with different specificities.

The carboxypeptidase A fold contains some 307 residues formed into a central multistrand β-sheet with helices packed against it. Its main enzymic activity is to cleave C-terminal residues from protein substrates according to specificity, and the enzyme requires a zinc atom for catalytic activity.

A supporting reason for the choice of these folds was that 3D data also exists for precursor folds, and folds of sequence variants and the fold as complexed to other ligands. This allows for the controlled comparison of structurally, functionally and sequentially induced changes to the fold. The CATH structural (http://www.biochem.ucl.ac.uk/bsm/cath) and 3Dee domain definition (http://circinus.ebi.ac.uk:8080/3Dee/help/help_intro.html) databases were also consulted to obtain the current view of the sub-domain structure of the folds.

Results
Pancreatic trypsin inhibitor fold (Figures 3 and 4)

Interfaces highlighted by SID. The highest SID score is 57 and is shared by residue positions 1, 57 and 58. These positions and their localities emphasize the interfacing of the N- and C-termini (which mark the start and end of helices respectively). Consideration of successively lower SID scores highlights more of this interface until residue position 23 and its neighbours 29 and 43 become included. Positions 23/29 lie opposite each other in the central, twisted, β-sheet, and form the centre of a packing interface for the N- and C-terminal helices against this β-sheet. Position 43 is located within one of two notable chain segments in the molecule that do not adopt classical secondary structures. This segment crosses over the edge of the β-sheet in the vicinity of position 23. Hence, the highest SID scores (i.e. 57–50) relate to a multi-way interface between

Step 5: Cluster merging
Cluster merging is the process by which the clusters obtained in the scoring stage of SID analysis are combined to visualise the structural elements and interfaces. It operates identically to the procedure of cluster merging used by regiovariation analysis (Cardle and Dufton, 1994, 1997) as an iterative process.

First, clusters are ranked according to their SID score (either highest to lowest or vice versa). For the purposes of explanation we ranked the clusters from highest to lowest SID score. Beginning the first ‘round’ at the clusters with the highest SID score, we examine the members of all the clusters. If two or more clusters share a common residue position, then the members of those clusters are combined to form a single larger cluster. Once this procedure has been completed for all clusters with that SID score, the process moves on to the next highest SID score. In this ‘round’, all the clusters at that score are compared both with each other and with the existing merged and unmerged clusters from the last round. Again, if any clusters share a common member, they are merged. This procedure is repeated for progressively lower SID scores until all clusters have been examined, or all positions have been allocated to a single cluster.

When merging takes place from lowest to highest SID score clusters, the stable structural elements are identified. When merging clusters goes in the opposite direction, from highest to lowest SID score, the interfaces between those elements are seen.

Choice of proteins
We selected four exemplars of the best known and experimentally well-studied protein folds to examine initially, namely those of bovine pancreatic trypsin inhibitor (BPTI) (PDB code 4pti), secretory phospholipase A2 (2bp2), chymotrypsin (2gmt) and carboxypeptidase A (2ctb), from the PDB database (http://pdb-browsers.ebi.ac.uk). This selection of folds covers sequence lengths from 58 to 307 residues and the folds contain various mixtures of classical secondary structure elements, and in the two larger examples, readily identifiable domains. Brief descriptions of the four families are given below.
the terminal helices, the β-sheet and one of the unstructured segments. Within this locale is found one disulfide bridge, that links the two helices together (5–55), and a second bridge that links the C-terminal helix to the β-sheet (30–51). Continuation down the scale of SID scores serves to expand the interface already defined to include more positions within the N-terminal helix and positions associated with the interface of the anti-parallel strands that run between the helices and the 14–38 disulfide bridge.

Surface-accessible interfaces. The highest scoring SID area is for the most part densely packed, showing that the major structural interfaces in the molecule are mostly inaccessible from the solvent phase. However, the residues that comprise the N- and C-terminal segments have both high SID scores and relatively high external accessibility. The fold’s vulnerability here is very effectively remedied by the 5–55 disulfide bridge. A separate region of the fold with high SID score/high external accessibility involves the unstructured anti-parallel strands that connect the two terminal helices with the β-sheet. A portion of one of these anti-parallel strands (namely that leading from the N-terminal helix) provides the major part of the interaction with serine proteinases. Again, as with the
chain termini, the introduction of a disulfide bridge (this time at 14–38) has limited much of the potential conformational vulnerability by linking the two strands together. The β-sheet formation also has a similar stabilizing effect. What seems to remain in this fold, therefore, is the potential for small-scale relative adjustments of the 8–13 and 39–44 strands.

Interfacial adjustments related to evolution and enzyme binding. Several 3D structures for different BPTI variants have been obtained, both as free forms and as complexed to serine proteinases (Bode and Huber, 1992). This enables alterations to conformation (expressed as variations in SID scores for homologous positions) caused by sequence evolution and enzyme binding to be discerned.

The uncomplexed versions show significant variations around positions 28, 29, 47, 54 and 56–58. Lesser alterations are found around 9, 10, 13, 30, 33, 44 and 48. These suggest that evolutionary changes have mainly altered the interfacing of the C-terminal helix with the C-terminal side of the β-sheet, while the smaller changes have mostly related to the interfacing of the two anti-parallel 'non-classically structured' strands with each other and the opposite side of the β-sheet. The C-terminal helix is connected directly to one of the unstructured

Fig. 3. Graphs of SID score (HL) for each residue position in the sequences of the four folds analysed. Pancreatic trypsin inhibitor, 4pti; phospholipase A₂, 2bp2; chymotrypsin, 2gmt; and carboxypeptidase A, 2ctb.
strands by the 44–47 segment that stretches round the edge of the sheet.

Upon complexation with serine proteinase enzymes, interfacial adjustment is seen around positions 6, 25, 28, 47, 54, 56–58. Smaller events occur around 1–2, 5, 9–10, 18–19, 21, 29, 30, 44, 46, 48 and 52. These changes are similar to those observed above, but are more extensive, this time also involving the N-terminal helix interfacing with the β-sheet.

Apparently, whereas natural sequence variation has consequences primarily for the interfacing of the C-terminal helix to the remainder of the molecule (and thereby one of the ‘unstructured strands’ to which it is connected), binding to an enzyme also changes the interfacing of the N-terminal helix such that both ‘unstructured’ strands may be affected.

**Phospholipase A2 fold (Figures 3 and 5)**

**Interfaces highlighted by SID.** The highest SID score is 99 and is achieved by the vicinities of residue positions 8, 12 and 103. In the fold, the localities of these three positions overlap such that three separate secondary structure elements are encompassed. For the most part, the locale highlights the interfacing between the end segment of the N-terminal helix (helix A) with one of the two major helices that form the core of the molecule (helix E). Also included is a smaller interface between helix A and the hairpin turn of the β-sheet (i.e. positions 12, 80, 81). In other words, the region is a ‘sandwiching’ of helix A between helix E and the β-sheet. Part of this locale constitutes the ‘hydrophobic channel’ through which the phospholipid substrate is thought to be drawn (Scott et al., 1991; Arni and Ward, 1996). Taking into account lower SID scores down to 91 expands this region and introduces two satellite areas that are initially unconnected to the original. The second site is formed by an interface of the non-classically structured segments 24–36 and 115–123. The third is another three-element interface between the N-terminal residue and two ‘transitional’ segments, namely where helix D joins to the β-sheet, and where the β-sheet joins helix E (residues 1, 66–73, 92, 95).

The helix A/helix E interface has a moderate packing density, but there is no disulfide bridge to add covalent stability, so some interfacial adjustment may be possible. There is such a bridge stabilizing the smaller interface nearby between the N-terminal helix and the β-sheet (11–77), but it is not present in other members of this enzyme family (i.e. it is present in Type I phospholipases, such as the bovine pancreatic enzyme, but not in Type II phospholipases) (Yang, 1994). The main core of the molecule is formed by two long helices (C and E) that are linked by two disulfide bridges, and the interface between helices A and E can be considered peripheral to this.
core. The satellite region of high SID score involving the non-classically structured C-terminal vicinity includes two disulfide bridges (27–123 and 29–45) that stabilise the interface and position it relative to helix C, respectively. The locale permits the entry and co-ordination of a calcium ion, which is, in some variants, essential for catalytic activity (Scott and Sigler, 1994). The metal ion adds substantially to local stability because its liganding maintains the loop structure in the 24–36 strand and effectively cross-links residue 49 in helix C (aspartate) with the residues at 28, 30 and 32 (in much the same fashion as the 29–45 disulfide bridge). In some Type II phospholipases, the calcium is replaced by the side chain of a lysine residue at position 49 (Harris, 1991). Throughout the Type II phospholipases, the C-terminal is extended, the extension being anchored to helix C by an additional disulfide bridge between position 51 and the C-terminus.

The remaining high scoring site involving the N-terminal residue is lent immediate stability by the 61–91 disulfide bridge and to a lesser extent by the 84–96 and 11–77 bridges (as noted above, the 11–77 bridge is absent in Type II phospholipases). Part of this locale (helix D) is missing in the majority of the known variants because five residues (62–66) have been deleted in them relative to the pancreatic enzyme.

Surface-accessible interfaces. When the SID scores are judged against local residue packing density, the two most externally accessible and high scoring locales are the calcium binding vicinity and the area that includes the N-terminus. Clearly, the calcium binding area has to permit entry of the metal ion and the formation around it of the correct ligand geometry (the metal can be removed or added without major denaturation). The vicinity of positions 69 and 72 that encompasses position 1 has a lower accessibility in the proenzyme because of the presence of an N-terminal extension that is removed in the activation process. In other members of the family, accessibility is further modified by the absence of helix D.

Interfacial adjustments related to evolution. Eight experimentally determined 3D structures for variants of this enzyme were submitted to SID analysis and significant variations noted. The most prominent adjustments have been in the vicinities of positions 21, 50 and 121–123, with lesser changes around 11, 22, 33, 47, 68, 73, 77, 80–82, 94, 100, 110, 117 and 132. When the clustering of these positions in the fold is considered, areas similar to those highlighted above are found to be involved, namely around the N-terminal, the C-terminal end of helix A, and the calcium binding domain. There is also some variation in the proximity of the core helices C and E.
Chymotrypsin fold (Figures 3 and 6)

Interfaces highlighted by SID

The highest SID scores (i.e. above 200) relate to two areas in the chain fold with approximately equal prominence. One is the vicinity of positions 189/190 where the N-terminal strand intrudes into domain 2, while the other is around positions 120/121, the ‘central point’ of the sequence about which the molecular fold is approximately palindromic. Positions 120/121 are also the mid-point of the main-chain crossover from domain 1 to domain 2.

The 189/190 locale is at the forefront of the enzyme’s activation and recognition mechanisms. First, it forms a major part of the $S_1$ specificity pocket that recognizes the side chain of the substrate P$_1$ residue (alterations within this pocket are responsible for substrate discrimination amongst different serine proteinases) (Czapinska and Otlewski, 1999). Secondly, the area is altered in the transition between the inactive precursorzymogen (chymotrypsinogen) and the active enzyme (Matthews et al., 1967). A cleavage of the N-terminal strand by another proteinase between positions 15 and 16 allows residue 16 (now bearing a charged $\epsilon$-amino function) to penetrate into the domain structure so as to change the orientation of residue 194 and thereby bring about full catalytic potency. The high SID score is created by residues 16/17 intruding into the area.

The 120/121 environment includes the domain crossover strand, part of the domain 1/domain 2 interface and the original N-terminus. It can be regarded truly as the ‘central point’ of the fold and lies ‘below’ the adjacent catalytic residues His57 (contributed from domain 1) and Ser195 (contributed from domain 2). The 189/190 locale is close by within domain 2.

As progressively lower SID scores are considered, more and more of the domain interface is highlighted so that a stage is reached when all the interface becomes merged with the $S_1$ pocket/’Second N-terminal’ vicinity.

Surface-accessible interfaces. Most outstanding is the original N-terminal strand in the vicinity of residue position 3 where positions 1–5 encounter positions 120–122 of the domain crossover and positions 204–208 of domain 2. The implied vulnerability to external modification is substantially reduced by disulfide bridge 1–122. The bridge links the N-terminal strand to the domain crossover link. However, the potential contribution to molecular stability of this N-terminal chain segment is changed drastically by the activation cleavage between positions 15 and 16, the consequence being that the segment can no longer be regarded as an integral part of the fold (the equivalent N-terminal strand is lost from trypsin because this bridge has no equivalent in trypsin).

The locality of 190 is again prominent and the area of interest broadly comprises strands 16–20, 143–146, 184–194 and 220–226. The need for external accessibility in this region is clear: it has to permit entry of the P$_1$ side chain of a substrate (e.g. a tyrosyl side chain) for recognition purposes. Some stability is given to the area by the 191–220 disulfide bridge, which is wholly contained within it, and the nearby 168–182 bridge. In the transition between delta and alpha/gamma chymotrypsin, positions 147–148 are cleaved out, which would ‘loosen up’ the area.

A third area with high SID score and relatively high accessibility is where position 193 on domain 2 interfaces with strand segments 33–34 and 39–42 on domain 1. Although the domain interface produces high SID scores along its length, it is for the most part densely packed and inaccessible except for this particular region.

Interfacial adjustments related to evolution, activation and inhibitor binding. The folds of seven uncomplexed trypsins were analysed, the results averaged and the positions with the most significant standard deviations in SID score noted. The most prominent conformational variations concerned the environments of positions 30, 103, 139, 141, 203, 221 and 235. More minor variations are seen at 66, 70, 79, 91, 94, 96, 101, 113, 125, 177 and 219. These positions highlight the outer surface of the molecule, particularly the loop extremities. There are also notable clusters at either end of the interface between the two major domains.

To ascertain the fold adjustments involved in the activation of trypsinogen to trypsin, 11 trypsinogen folds were calculated and averaged for comparison with the free enzyme. It is clear from the SID results for these folds that there is substantial conformational variation amongst the trypsinogens, but on activation to trypsin, the fold becomes defined within much more narrow limits. In order of decreasing magnitude, the following positions experience significant alterations in their SID scores as a result of activation and thereafter show much less score variation: 191, 146, 16, 142, 190, 143, 41, 145, 17, 189, 144, 150, 159, 140, 188, 114, 23, 152, 193. These positions are all associated with the setting up of the fully functional $S_1$ specificity pocket architecture (as caused by the intrusion into the second domain of the new chain N-terminal created for the enzyme by the activation process). While 114 lies some distance away in the first domain, it is directly affected by the reorientation of the nearby N-terminal chain segment. In contrast, although positions 79, 103, 125 and 139 also experience change as a result of the activation process, the variation in SID score suggests they move to a less restricted situation. Positions 79 and 139 lie adjacent to the N-terminal strand while 103 and 125 ‘sandwich’ the start of the C-terminal helix, so it can be deduced that the activation process affects the relationship of both chain termini with respect to the major domains. Inspection of the relevant SID scores shows a narrowing of the major domain interface is involved. Only one position experiences a significant alteration in SID score, yet is conformationally ‘fixed’ in both trypsinogen and trypsin. This is position 194, a critical part of the $S_1$ specificity pocket which effectively ‘switches’ from one fixed position to another.

The final comparison that can be made is between the free trypsins and trypsins complexed with inhibitors. Over 60 examples of the complexed trypsin fold were averaged and compared. In order of decreasing magnitude, complex formation changes specifically the SID-measured environments of positions 116, 114, 58, 146, 24, 71 and 187. The changes are generally much smaller than those seen in the activation process, but they also involve a loss of juxtapositional consistency when the different complexes are compared. The positions relate to the surroundings of position 24 (i.e. the locale of interaction between the N-terminus and the $S_1$ specificity pocket), so it is implied that $S_1$ pocket occupancy has either ‘loosened up’ the area, or that different inhibitors change the locale in different ways. A notable change in isolation is seen at position 58, possibly related to the proximity of the catalytically vital histidine 57.

Carboxypeptidase fold (Figures 3 and 7)

Interfaces highlighted by SID. The highest SID scores relate primarily to the 3D confluence of three chain segments within
SID analysis of proteins

Fig. 7. Carboxypeptidase A. Cartoons of chain fold 2cbt (generated by RASWIN, Sayle, 1996) with notable sites circled. (A) Sites of high SID scores (HL calculation). (B) Sites of most significant SID score variation amongst homologues. (C) Sites of most significant changes in SID score as caused by activation. (D) Sites of most significant changes in SID score as caused by inhibitor complexation.

the structure. These are 8–13, 67–85 and 277–294. At lower scores, segments 196–198 and 116–127 become included in the confluence. A satellite site (comprising segment 95–102 and position 305, the C-terminus) arises nearby. Inspection of the fold topology according to the hierarchy of the SID results divides the molecule initially into two domains (1–189 and 190–307) wherein the latter domain is itself an assemblage of sub-domains 200–266 and 272–307. This enzyme is not widely recognized as having a domain structure because of the β-sheet that forms the core of the molecule, but the domains can be readily observed by 'undoing' the hydrogen bonding between the parallel β-strands 60–67 and 189–196.

The area highlighted by SID encompasses the majority of the active site of the enzyme. Most notably, the catalytically essential zinc ion (liganded by residues 69, 72 and 196) (Christianson and Lipscomb, 1989) is contained within it. This metal ion acts as a bridge between the two major domains, complementing their interfacing via the parallel β-sheet formation between strands 60–72 and 187–196. Other residues with suspected catalytic roles (E270), primary binding roles (R127, R145) or secondary binding roles (R71, Y198, F279) are either close by the region or within it (Auld and Vallee, 1987).

Surface-accessible interfaces. The most accessible part of the high scoring area is the conjunction of chain segments 9–13, 71–85 and 277–294, while a secondary focus involves the C-terminal (segments 99–103, 301–307). The third focus is the confluence of segments 56–66, 187–197 and 263–272. The latter area contains E270 and is linked to the most accessible focus via the zinc binding site (i.e. residues 72 and 197).

There is only one disulfide bond in the molecule, and it does not act to stabilize any of the prominent structural interfaces detected. Instead, it lies wholly within a non-classically structured part of the 1–189 domain that is probably too small to have sufficient inherent stability from non-covalent forces.

Interfacial adjustments related to evolution and inhibitor binding. The available 3D data for this enzyme not only includes several variants, but also the proenzyme and the enzyme complexed to small inhibitors. For the uncomplexed activated forms, the most pronounced SID-detected variations occur around positions 58, 73, 182 and 302, with lesser changes around 14, 47, 59, 116, 120, 144, 231, 275 and 297. These correspond to adjustments in the vicinity of the zinc ion and the C-terminus.
The activation process involves substantial alterations, initiated by the newly formed N-terminus, with positions 12, 40, 73 and 144 experiencing the most environmental changes. The positions are in line with one another on one face of the β-sheet such that the N-terminus is able to exert influence upon the zinc locality and Asn/Arg 144/145 which bind the substrate carboxyl group. A second site of adjustment again involves the locality of the C-terminus of the enzyme.

The complexation process also causes adjustment near the N-terminus, centred about position 77, but this time extending onwards over the edge of the β-sheet to its other side (182, 237) via 153–155 and the Zn site. The perturbations around 77 are caused by inhibitor binding in the S1 site while those on the opposite side of the β-sheet are caused by binding in the S1’ site. Yet again there is some adjustment near the C-terminus.

Discussion

Major interfaces tend to be surface accessible and frequently stabilised by disulfide bridges

For the three larger folds, we observed that crowded residue positions tend to generate high SID scores (i.e. the more residues that are clustered within 7 Å of a given residue, the greater is the probability that residues from far away in the sequence will be encompassed). This reflects a tendency for chain interfaces to be buried in the interior of the molecular superstructure (presumably to minimise vulnerability to externally induced fold disruption). However, the highest SID scores seen for each protein fold show a different relationship to local residue packing density. There is a strikingly even distribution across the different packing densities for these prominent scores, with densities of 8–11 residues per 7 Å sphere being the most common. In other words, the most long-range interfaces (in terms of distance apart in the chain sequence of the contributing residues) tend not to be completely buried—they are for the most part only moderately packed localities and therefore potentially of restricted accessibility from the external environment (see Figure 2 for the carboxypeptidase case). In some instances, the individual folds possess more than one prominent substructural interface. These major interfaces appear to represent significant ‘mid-ground’ between ‘surface’ and ‘core’ where only certain ligands can gain access to some of the more important determinants of conformational stability.

Vulnerable interfaces (i.e. high scoring and with apparent external accessibility) are frequently compensated wholly or partially by the presence of disulfide bridges. These provide strong covalent linkages between the distant sequence segments that form the interface (e.g. Kunitz inhibitors, phospholipases, serine proteinases). This is not to imply that high scoring SID areas without disulfide bridges are topologically unstable, merely that a potential for a significant sub-structural realignment might remain if the typical non-covalent stabilizing forces (i.e. the hydrophobic/hydrophobic, charge/charge and charge/aromatic interactions, identified by Brocchieri and Karlin, as being chiefly responsible for the internal stabilization of protein fold arrangements) were not maximized, or were reduced by an external agency (Brocchieri and Karlin, 1995). Likewise, the presence of disulfide bridges across an interface does not remove the possibility of all adjustment, it implies only that any adjustments should be small scale.

Prominent and externally vulnerable interfaces may also accommodate metal atoms (e.g. phospholipases, carboxypeptidases), suggesting that these ‘guests’ are required for the conformational impact of their liganding as well as their particular chemical properties.

SID scores facilitate the division of the molecules into sub-domains

The identification of the most prominent sub-structural interfaces by the SID calculation also indicates which parts of the superstructure may deserve to be identified as semi-autonomous domains. There have been many attempts to systematically detect sub-structural domains in folds based on a range of intuitive and subjective criteria. According to the widely cited CATH and 3Dee protein database assignments (see Materials and methods), the pancreatic trypsin inhibitor, phospholipase and carboxypeptidase folds are all classifiable as single domains. The serine proteinase fold, however, is classified as having two domains, one comprising the N-terminal segment with residues 121–233 and the other the C-terminal segment with residues 24–121. The SID results present a quite different picture, especially in respect of the phospholipase and carboxypeptidase folds.

(i) Pancreatic trypsin inhibitor fold. The primary distinction made by SID in the molecular superstructure is between ‘domains’ 1–38 and 45–56. This implies that the N-terminal strand/β-sheet and the C-terminal helix have a degree of structural independence and that the 38–45 strand is effectively a ‘linker’ segment that traverses the edge of the β-sheet.

(ii) Phospholipase fold. SID produces a primary distinction between residues 1–109 and 110–123 as being quasi-autonomous regions of the molecule. The latter section does not adopt a regular secondary structure and is not as closely packed as the remainder of the chain, relying on the calcium and disulfide bridges for its conformation. The 1–109 domain is divided by SID into further ‘sub-domains’ of 1–71 and 72–109. The interface between these is effectively the anti-parallel alignment of helices C and E that is cross-linked by disulfide bridges, and helix D is part of the loop ‘hinge’ between them (given extra support by the 61–91 bridge). Given this divide, the N-terminal and C-terminal strands are seen as ‘wrap around’ arms that emerge from one sub-domain to embrace the other (i.e. the N-terminal strand emerges from the helix D sub-domain to nestle between helix E and the β-sheet while the C-terminal strand emerges from the helix E domain to embrace the calcium binding area). The disulfide bridges 11–77 (when present), 27–123 and 50–131 (when present) can be viewed as retaining the terminal strands in place against the major ‘domains’.

(iii) Chymotrypsin fold. SID data suggests there are two major domains (i.e. 2–122 and 127–245) similar to those identified by CATH and 3Dee. However, the chain terminal strands are assigned to the domains to which they are linked sequentially, not to those against which they are packed. Hence they can be appreciated as two domains with ‘wrap around arms’ for each other.

(iv) Carboxypeptidase fold. SID suggests that the molecule is divisible primarily into two domains comprising 1–189 and 190–307, with the latter domain being divisible into further sub-domains of 200–266 and 272–307. The smallest ‘autonomous’ domain is the C-terminal helix which lies along the length of the interface between the other two domains. The zinc is found at a confluence of all three sub-domains.
Functional sites are associated with major interfaces

Our evidence suggests that for the four folds examined, the functional and guest binding sites have developed in molecular loci where the various ‘tributary’ interfaces between the domains and sub-domains combine and converge in 3D to create particularly prominent and multi-component interfaces (witness the P* residues in the inhibitors, the N-terminal locale/calcium site/hydrophobic channel in the phospholipases, the S1 pocket/catalytic site in serine proteinases, and the zinc binding/protein cleaving site of carboxypeptidases). We expect this has provided the opportunity for residue substitutions remote from the functional sites, and substitutions scattered over a wide area of the molecule, to have had a bearing on the evolutionary development and ‘tuning’ of the functional sites. These substitutions could have caused adjustments of ‘tributary’ interfaces in their vicinity, with the consequence that effects were relayed ‘downstream’ via the interfacial connectivity to the functional site itself. Presumably, most of the effects would have been received at the functional site as slight changes of juxtaposition amongst the functional site residues. Depending on the character and connectivity of the interfacing paths along which they might travel, mutation-induced adjustments might be amplified, damped or combined to create more subtle effects than would be achieved by any side chain substitution in the site itself. In the lifetime of the protein, a modification or binding remote from the functional site could also bring about an adjustment of the functional site via the interfacial connectivity. Thus, the significance of a specific topology is that it will dictate the ‘directionality’ of the adjustments possible, and also determine from which parts of the molecule mutation or binding events (e.g. allosteric factors, metal ions) will be able to exert an influence on particular components of the functional site.

Interfacial adjustment may also play a mechanistic role

Given the network of tributary interfaces that can be identified as converging upon a functional site, it should also be considered that a binding event in the functional site could radiate influence out to a large proportion of the molecule via the same interfacing network. However, whereas the evolutionary process adjusts the site in small steps from different locations, a ligand binding event (especially in the cases of the serine proteinases and carboxypeptidases) could produce a multipoint impact on a large proportion of the highest order interfaces simultaneously. The combination of the incoming ligand imposing some of its geometry upon the host protein, plus the possibility of some interface-stabilizing residues switching from intra- to intermolecular binding, could precipitate large-scale interfacing adjustments that might impact across the whole of the host protein. There is the clear opportunity, therefore, for global molecular adjustment to be harnessed to any ensuing translation of the binding into a response (e.g. channel opening or catalysis).

With this possibility in mind, we re-examined the four folds above to see if the execution of their prime function had the potential to involve significant internal domain reorientation.

Pancræatic trypsin inhibitor fold

In such a simple molecule, wherein the fold is defined by a β-sheet with two helices packed against it, the adjustment potential resides in altering the juxtapositions of the three elements of secondary structure (Pritchard and Duf ton, 1999). The scope for relative motion is reduced substantially by the 5–55 and 30–51 disulfide bridges, but what remains has consequences primarily for the two strands that link the helices to the β-sheet. This is because the two strands do not adopt classical secondary conformations or have any special anchorage to the sheet. In effect, the two strands are ‘strung out’ from one end of the sheet to the other with alterations in ‘tension’ having the most likely impact upon them. Given the comprehensive experimental data, there is little need for speculation other than to suggest that while adjustment of both the helix/sheet interfaces appears to be an integral part of the enzyme binding process, adjustment of only the C-terminal helix/ sheet interface (and thereby the ‘tension’ of the 36–45 strand) seems to be a profitable evolutionary process. Interestingly, SID defines the C-terminal helix as a more autonomous part of the structure than the N-terminal helix.

Phospholipase A2 fold

The most ‘vulnerable’ aspects of this molecular fold are for the most part compensated for by the strategic locations of disulfide bridges and the liganding of the calcium ion. However, the A helix, which is involved in high scoring regions at each end and is linked to the calcium binding region via helix B, has no covalent anchorage save the 11–77 disulfide bridge, and even this is absent in the Type II variants. Moreover, the topological environment of the start of helix A is variable between family members because of deletions/insertions in and around helix D. SID analysis, coupled with the experimentally observed 3D variation seen between homologues (above), suggests that the helix AB segment (i.e. chain segment 1–30) is that part of the fold most likely to be prone to externally/ internally induced adjustment with respect to the remaining superstructure, perhaps effecting an ‘opening/closing hinge’ movement about its midpoint (the helix AB/helix E crossover). In view of structural variation, it may be disposed differently in different family representatives, and have greater freedom in Type II phospholipases.

In relation to what has been shown experimentally about the enzyme activation and substrate binding, the areas highlighted by SID are associated closely with some of the most critical features. The highest scoring site is located close to where the hydrocarbon chains of phospholipid substrates are expected to bind (i.e. in inhibitor complexes, the hydrocarbon chains lie in the ‘elbow’ created by helices A and B) (Scott et al., 1991; Arni and Ward, 1996). The calcium binding area (the second region highlighted by SID) is expected to bind the phosphate moiety of the substrate, while the remaining high SID region is around the N-terminal (Scott and Sigler, 1994). The latter area plays a central role in activating the enzyme to attack aggregated substrate (e.g. in micellar form). In the proenzyme, an N-terminal extension causes the locality to be disordered and only substrate monomers can be hydrolysed. Removal of the extension causes the newly formed α-amino function to link to D99 via a water molecule and hydrogen bond to residues 4 and 71, actions which set up both the catalytic mechanism and the Interface Recognition Site (IRS) for optimal activity against substrate aggregates (Dijkstra et al., 1983; Scott et al., 1991; Dua et al., 1995). It has been postulated that aggregated substrate is able to stimulate the enzymic action via a conformational change in the protein, mediated by the encounter with the IRS (aggregated substrate can also enhance inhibitor and calcium binding) (Slotboom et al., 1982). More recent findings confirm that binding to lipid/water interfaces causes the N-terminal helix and β-sheet to change position (VandenBerg et al., 1995). Moreover, in
the examples of this enzyme encountered frequently as dimers, the IRS and catalytic site would be inaccessible to substrate unless a conformational change occurred to expose them (Brunie et al., 1985; daSilvaGiotto et al., 1998). Interestingly, in the dimers, the interfacing involves the two most exposed areas of high SID—the calcium binding area of one molecule interacts with the N-terminal area of the second in a symmetrical fashion.

There appear, therefore, to be strong grounds for anticipating some element of conformational adjustment in the mechanism of these enzymes, and the SID results give focus to the idea. Secondary binding of the substrate monomers (i.e. via the hydrocarbon chains) occurs close to the site of highest SID between helices A and B, and the N-terminal locality and the calcium binding area are both secondary areas of high SID that are associated with the mechanistic events. These regions are potentially in communication via change in the conformational disposition of the helix A/helix B segment. The impetus for adjustment could arise from any of the three sites individually or from an encounter between the IRS and an aggregated substrate (in which case the substrate could influence all three high SID areas simultaneously). An important experimental finding of relevance is that catalytic efficiency of the enzyme increases as the hydrocarbon chain length is increased (Slotboom et al., 1982); this would befit a mechanism in which secondary binding at the site of highest SID was translated into an allosteric optimization of catalytic efficiency. As to the possible extent of the adjustment in the molecule, it will be recalled that SID identifies two sub-domains with ‘wrap around’ arms, namely 1–71 and 72–109/101–123. According to the evidence detailed above, the intramolecular docking of one of these arms (the N-terminal one) is influenced directly by binding to substrate hydrocarbon chains and lipid/water interfaces, while the other (the C-terminal one) could be disturbed by any change to the calcium ligand field (e.g. by interaction with the substrate phosphate). Finally, the N-terminal resides near the ‘hinge’ of the two domains (as retained by the 61–91 disulfide bridge and adjusted by deletions/insertions and the β-sheet disposition). All these circumstances are consistent with activation and substrate binding mechanisms that have, either over evolutionary time or in real time, consequences for the juxtaposition of helices C and E. Since the catalytic residues are found at this interface (Janssens et al., 1999), contributed from either side, the ultimate issue would seem to be the adjustability of their precise 3D orientation.

Chymotrypsin fold

With two major domains in the molecule of similar size and fold, and no cross interface disulfide bridges, the largest scale conformational adjustments that can be entertained involve some kind of movement of these domains with respect to each other. The highlighting by SID of the area of the original N-terminus, and of the ‘new’ N-terminus/S1 specificity pocket would be consistent with this notion because it is clear that the N-terminal strand is the main obstacle to domain adjustment. The N-terminal strand is essentially a ‘wrap around arm’ that emerges from domain 1 to embrace domain 2. In chymotrypsinogen, this arm is a substantial double strand loop with the N-terminal anchored at the domain crossover. This would appear to prevent the possibility of any significant movement of the two domains relative to each other. However, in the transition from proenzyme to active enzyme, the N-terminal loop is cut at its point of reversal, removing the restraint that the 1–122 disulfide and interaction of the 1–15 segment with domain 2 could have against domain movement. Thereafter, the ‘new’ N-terminal becomes associated with the disposition of the S1 binding pocket, and in doing so is placed in a situation where it may be affected by substrate binding in the S1 pocket. Thus, the SID results suggest that the greatest potential for conformational adjustment is centred round the S1 pocket and the nearby domain interface, which includes the catalytic residues 57 and 195 responsible for the actual peptide bond cleavage. Moreover, the restraint against domain movement appears to be reduced by activation, while what restraint remains comes under the influence of substrate binding events. It has been inferred from inhibitor binding studies that substrates bind across the domain interface with the scissile peptide bond actually bridging the interface. The majority of the S1 secondary interactions take place with the 120–240 domain while the majority of the S1 secondary interactions take place with the 1–120 domain. This situation is ideal for a mechanism in which a movement of the two large domains is used to strain/distort the scissile peptide bond preparatory to hydrolysis (Dufton, 1990).

Carboxypeptidase A fold

Based on the SID results above, there would appear to be some considerable scope for conformational change in response to substrate binding and the subsequent hydrolytic events. The circumstantial evidence includes:

(i) The active site is disposed at an interface of three sub-domains.
(ii) The zinc atom that participates in catalysis bridges the two major domains.
(iii) The S1’ pocket is in the second domain on one face of the β-sheet while the S1 binding takes place in the first domain on the opposite side of the β-sheet.
(iv) The scissile peptide bond of the substrate is expected to be positioned at the point where the substrate chain crosses over the edge of the β-sheet from one side to the other, immediately above the zinc ion and the two parallel β strands which unite the two domains.
(v) The major interface between the two domains is highly accessible at either end, and moderately accessible along its length on the C-terminal side (the C-terminal helix affords a basic protection for the interface, and since the C-terminal helix can itself be regarded as a sub-domain of the second major domain, there is in effect a three-way sub-domain interface running across the molecule).
(vi) Activation and inhibitor complexation causes fold deviations around the zinc binding/active site area which appear transmitted to the C-terminal area via the C-terminal helix.

The barrier to any major conformational response to binding (i.e. the domains moving with respect to each other) is constituted primarily by the parallel β-sheet formation between strands 60–67 and 189–196, and the zinc liganded between residues 72 and 196. In order to propose a fold adjustment in the mechanism, it would be necessary to suppose that the changes in the ligand field of the zinc atom during catalysis (it is hypothesized that the zinc co-ordinates the tetrahedral intermediates) (Christianson and Lipscomb, 1989) are in some way connected with a temporary destabilisation/adjustment of the domain interface. Whether the behaviour of the zinc could initiate a response by precipitating the loss of the parallel...
inter-strand hydrogen bonding, or whether secondary binding by the substrate (i.e. substrate residues P₁, P₂, P₃, P₄, etc.) near these strands could impose allosterically a catalytically efficient geometry on the zine by adjusting its ligands via the interface, remains open to question. The benefit of causing a domain adjustment is that the scissile bond would become the focus of opposing conformational adjustments in the enzyme by virtue of its location, and the fact that the substrate structure either side of it is anchored in different domains and on opposite sides of the β-sheet. This could distort/strain the peptide bond in such a way as to ease the catalytic task (e.g. by reducing the resonance stabilization and making the peptide bond more eter-like). It is interesting to note that a similar contrivance (i.e. substrate scissile bond ‘stretched’ over a β-sheet edge) is observable for subtilisin. In the case of carboxypeptidase, some role for the C-terminal helix seems likely since not only does it lie along the main interface and have a degree of structural independence, but also there is an unexplained centre of residue conservation around the C-terminus in the enzyme family as a whole (Cardle and Dufton, 1994).

Conclusion
Although SID analysis is very simple in concept and execution, the value of its output is that it presents the human observer with a systematic and quantitative appraisal of the interfacial characteristics of a fold topology. In doing so, it directs the observer’s attention to localities in the chain fold where the conformational status of the molecule is rendered particularly vulnerable to mutational or external influence. This revelation of the consequences that different topologies might have for site-specific inductions of conformational change, and the ways such changes might be propagated or directed in the protein molecule, cannot be obtained as completely from conventional 2D and 3D display techniques. This is because the human brain cannot assimilate the 3D information in sufficient detail from these displays to be able to extrapolate potential properties based on the totality of that information.

Our application of the SID technique suggests that the proteins we have examined have been underestimated in their ability to function as ‘true’ mechanical devices (i.e. to deploy structural movement and induction of strain as a prime means of translating binding events into inhibitory or catalytic consequences) (Williams, 1993). This is not to imply that the ‘stereo-electronic’ catalytic mechanisms that have been proposed for the enzymes are fundamentally incorrect, only that a significant contribution to the process may have remained undetected (e.g. that the proteinases could make use of primary and secondary binding to create a physical distortion of their target substrate bond and link it to a catalytic group realignment adjacent to that bond). In the case of proteinases, this reappraisal goes some way to explaining why, for example, different fold topologies and catalytic groupings are required for aminopeptidase, endopeptidase and carboxypeptidase actions, and for proteinases and oligopeptidases (i.e. that different fold-dependent ‘mechanical’ solutions are required according to the location of the target peptide bond within the substrate chain). Applied to further protein families, the SID technique may help confirm that natural proteins have been based on a limited subset of the possible chain topologies because only a few chain fold patterns have proved to be generally robust and incrementally adjustable when subject to random mutational events and functional selection. Many of these enduring topologies may have the additional virtue of presenting unique, but reliable, conformational responses when other molecules become bound to their exposed sub-structural interfaces.

References

Received September 27, 2001; revised September 1, 2002; accepted December 3, 2002.