

# Conserved features in the active site of nonhomologous serine proteases

Steven C Bagley<sup>1</sup> and Russ B Altman

**Background:** Serine protease activity is critical for many biological processes and has arisen independently in a few different protein families. It is not clear, though, the degree to which these protease families share common biochemical and biophysical properties. We have used a computer program to study the properties that are shared by four serine protease active sites with no overall structural or sequence homology. The program systematically compares the region around the catalytic histidines from the four proteins with a set of noncatalytic histidines, used as controls. It reports the three-dimensional locations and level of statistical significance for those properties that distinguish the catalytic histidines from the noncatalytic ones. The method of analysis is general and can be applied easily to other active sites of interest.

**Results:** As expected, some of the reported properties correspond to previously known features of the serine protease active site, including the catalytic triad and the oxyanion hole. Novel properties are also found, including the spatial distribution of charged, polar, and hydrophobic groups arranged to stabilize the catalytic residues, and a relative abundance of some residues (Val, Tyr, Leu, and Gly) around the active site.

**Conclusions:** Our findings show that in addition to some properties common to all the proteases examined, there are a set of preferred, but not required, properties that can be reliably observed only by aligning the sites and comparing them with carefully selected statistical controls.

## Introduction

The primary means of understanding structure/function relationships in molecular biology has been the study of the structure of individual biomolecules. With the growth of the Protein Data Bank (PDB) [1], the increase in the number of high-resolution protein structures allows analysis not only of the functional properties of single structures, but also of the shared functional properties within classes of proteins. We have previously reported a computer system that takes a set of aligned structures and searches for the biochemical and biophysical properties that distinguish the aligned structures from a set of control structures [2,3]. The method reports its findings as triplets of property/location/abundance, where the property of interest is located in a particular volume with respect to the frame of reference and is found to be either abundant or scarce with respect to the controls. We have reported a preliminary application of this method to the active site of the serine proteases [3]. Although we were able to identify significantly conserved structural features, our sample of proteases contained multiple structures that were closely related (as measured by both sequence homology and structural similarity), and so many of the conserved features were artifacts of a close familial relationship. In order to identify those conserved features that are widely con-

served, and not a consequence of sequential homology, we now report our analysis of the active site of four nonhomologous serine proteases.

The selective forces operating on proteins during their evolution can, in principle, produce similar biological functions through a variety of different detailed structures. Analysis of functional sites that focus solely on amino acids may miss similarities that occur on a different level of detail (e.g. clusters of electronegativity, or the presence of an important chemical group). Therefore, our computational method focuses on neighborhoods of three-dimensional protein structures (sometimes called 3D-motifs), as distinguished from protein sequences. We are concerned with subsets of the entire protein in regions that we call microenvironments or 'sites'. Sites might be active sites, binding sites, or any other region of interest. We describe the protein sites using a redundant set of biochemical and biophysical properties. We include descriptors over a wide range of detail to maximize the chance of finding succinct descriptions. The properties can be roughly classified into five groups: atomic identity (e.g. C, O, and N), functional group (hydroxyl or carbonyl), residue type, secondary structure (helix or strand), and 'other' (charge and hydrophobicity).

Address: Section on Medical Informatics, Stanford University, MSOB X215, Stanford, CA 94305-5479, USA; <sup>1</sup>e-mail: [bagley@smi.stanford.edu](mailto:bagley@smi.stanford.edu).

Correspondence: Russ B Altman  
e-mail: [altman@smi.stanford.edu](mailto:altman@smi.stanford.edu)

**Key words:** active sites, comparative analysis, computer methods, protein structure, serine protease

Received: 28 May 1996  
Revisions requested: 19 Jun 1996  
Revisions received: 08 Jul 1996  
Accepted: 15 Jul 1996

Published: 27 Aug 1996  
Electronic identifier: 1359-0278-001-00371

**Folding & Design** 27 Aug 1996, 1:371–379

© Current Biology Ltd ISSN 1359-0278

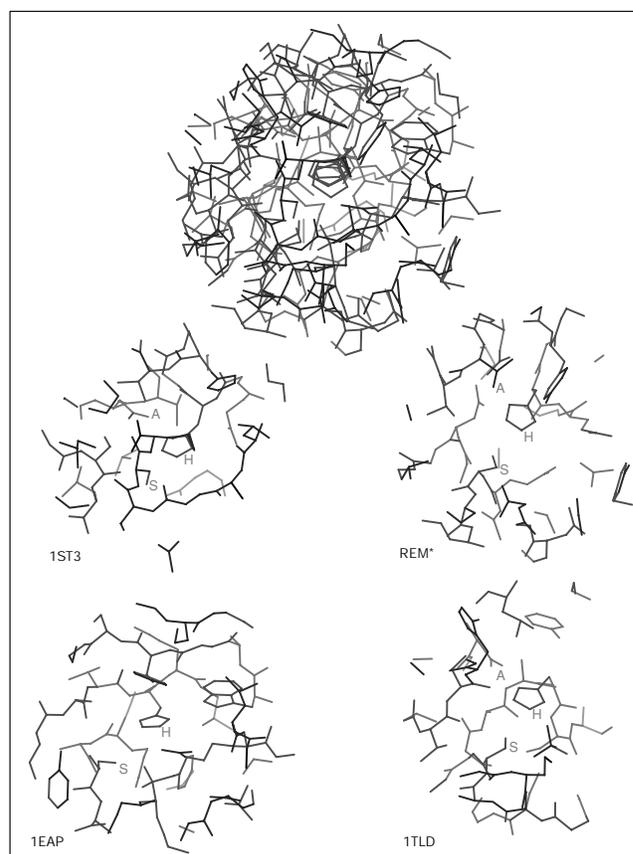
During our analysis of sites, we assign significance to our observations using standard statistical tests in the context of explicitly defined control groups (the ‘nonsites’) against which the sites are compared. Choice of the control group is an important part of the experimental design, serving as an adjustable focus or detail filter. For example, calcium-binding sites might be compared to random protein sites (leading to a description of their anionic environment), or to magnesium-binding sites (leading to a description of the subtle differences between two anionic environments); each is a valid choice of control, but would give rather different results.

The serine protease class of molecules has been studied from a number of perspectives, including homology modeling [4] and detailed analysis of substrate specificity [5]. It is also the subject of numerous drug design activities [5,6]. Some essential structural features have been identified as necessary for the catalytic activity [7]. First, there is a ‘catalytic triad’ comprising the sidechains from aspartic acid, histidine and serine residues, which provide most of the catalytic activity. Second, the tetrahedral transition state is stabilized by the presence of two NH groups that form hydrogen bonds to the substrate’s oxygen atom; this structure is called the ‘oxyanion hole’ [8]. Third, the types of residues that can bind at the active site vary with the structure of the enzyme’s ‘specificity pocket’. The serine protease families in the PDB include the chymotrypsin family (chymotrypsin, trypsin, elastase, etc.), the subtilisin family (bacterial serine proteases), wheat serine carboxypeptidase, and an artificially produced antibody with proteolytic activity (detailed in the Materials and methods section). The mechanisms of serine protease specificity are an area of great interest, but are not the object of the study reported here. Equally intriguing is the observation that the four very different protein folds appear to function in a similar manner. Our experiment is designed to find those features that are common to the four families of serine proteases. Presumably, these common features support the biochemical mechanism implementing the common observed function of proteolysis of the substrate backbone, abstracting away from the details of each family. Using one structure from each family gives enough data to achieve some sense of significance and avoids the problems of bias due to family underrepresentation or overrepresentation. It also focuses our attention on features common to the serine protease families and not those, such as substrate-binding specificity, that vary across the proteases. A priori, we would expect to find most of the known features, and perhaps to find other features that have not previously been mentioned. In our analysis below we address the issue of reconciling whether these other features are novel or artifactual, and the degree to which statistical significance implies biochemical significance.

## Results and discussion

The structures of the four sites, individually and superimposed, are shown in Figure 1. A kinemage that allows interactive viewing of the full 3D structures, along with the precise regions of significance, can be found in the Supplementary material (published with this paper on the internet). The resolution of our findings is approximately 1.2 Å because of our sampling grid and the collection of statistics over small volumes. Table 1 shows the significant regions, sorted by decreasing statistical significance (taking the best significance for the whole cluster of related results). Each row lists the property name, the range of P-levels for the cluster of that property, whether that property is more abundant in the sites or nonsites, and the residues in the protein 1TLD that contain atoms in the implicated region, using the familiar chymotrypsin family sequence numbering.

**Figure 1**



Four unrelated serine protease molecules. The four serine protease structures used in this study are shown in structural alignment (as described in the text) at the top. Below, each individual structure (1ST3, REM\*, 1EAP, and 1TLD) is shown from the same viewing angle. The catalytic histidine (H), serine (S) and aspartic acid (A) are labeled. 1EAP has no catalytic aspartic acid. Consult the Supplementary material (published with this paper on the internet) for a kinemage containing a 3D presentation of the structures and results.

**Table 1****Rank ordered list of significant results.**

Rank	Property	P-levels	Abundance	Residues in 1TLD
1	HYDROXYL	0.001–0.001	Site	Ser195
2	RESIDUE-NAME-IS-SER	0.001–0.005	Site	Gly196; Ser195; Ser214
3	MOBILITY	0.001–0.005	Site	
4	NEG-CHARGE	0.001–0.005	Site	Ala56; His57; Asp102
5	SECONDARY-STRUCTURE2-IS-COIL	0.001–0.005	Site	Asp194; Ser195; Leu99
6	SECONDARY-STRUCTURE1-IS-COIL	0.001–0.005	Site	Asp194; Ser195; Leu99
7	RESIDUE-CLASS2-IS-POLAR	0.001–0.005	Site	Gly193; Gly196; Ser195
8	RESIDUE-CLASS1-IS-POLAR	0.001–0.005	Site	Gly196; Ser195
9	ATOM-NAME-IS-C	0.001–0.01	Site	Ile212; Asp102
10	RESIDUE-CLASS1-IS-UNKNOWN	0.002–0.005	Site	Gly196; Gly193
11	ATOM-NAME-IS-O	0.002–0.005	Site	Ser195
12	AMIDE	0.002–0.01	Site	Gly196; Ser195
13	ATOM-NAME-IS-N	0.002–0.01	Site	Gly196; Gly193; Ser195
14	RESIDUE-NAME-IS-LEU	0.005–0.005	Site	
15	RESIDUE-NAME-IS-GLY	0.005–0.005	Site	Gly193
16	RING-SYSTEM	0.005–0.005	Site	
17	SECONDARY-STRUCTURE2-IS-HET	0.005–0.005	Site	Gly193
18	SECONDARY-STRUCTURE1-IS-HET	0.005–0.005	Site	Gly193
19	RESIDUE-CLASS2-IS-UNKNOWN	0.005–0.005	Site	Gly193
20	RESIDUE-CLASS2-IS-ACIDIC	0.005–0.005	Site	Asp102
21	RESIDUE-CLASS1-IS-CHARGED	0.005–0.005	Site	Asp102
22	RESIDUE-NAME-IS-VAL	0.005–0.005	Site	
23	RESIDUE-NAME-IS-TYR	0.005–0.005	Site	Asp194
24	RESIDUE-NAME-IS-ASP	0.005–0.005	Site	Asp102
25	CHARGE	0.005–0.005	Site	Ala56; Asp102; His57
26	HYDROPHOBICITY	0.005–0.005	Site	
27	CHARGE-WITH-HIS	0.005–0.01	Site	Ala56; His57
28	CHARGE-WITH-HIS	0.01–0.01	Nonsite	Ser214; Ser195
29	HYDROPHOBICITY	0.01–0.01	Nonsite	
30	VDW-VOLUME	0.01–0.01	Site	Ile212
31	B-FACTOR	0.01–0.01	Site	Ser195; Ile212
32	ATOM-NAME-IS-ANY	0.01–0.01	Site	Asp102; Ile212

The second column contains the name of the property found to be significantly different in sites and nonsites. The third column provides the Mann–Whitney estimate of significance. The fourth column

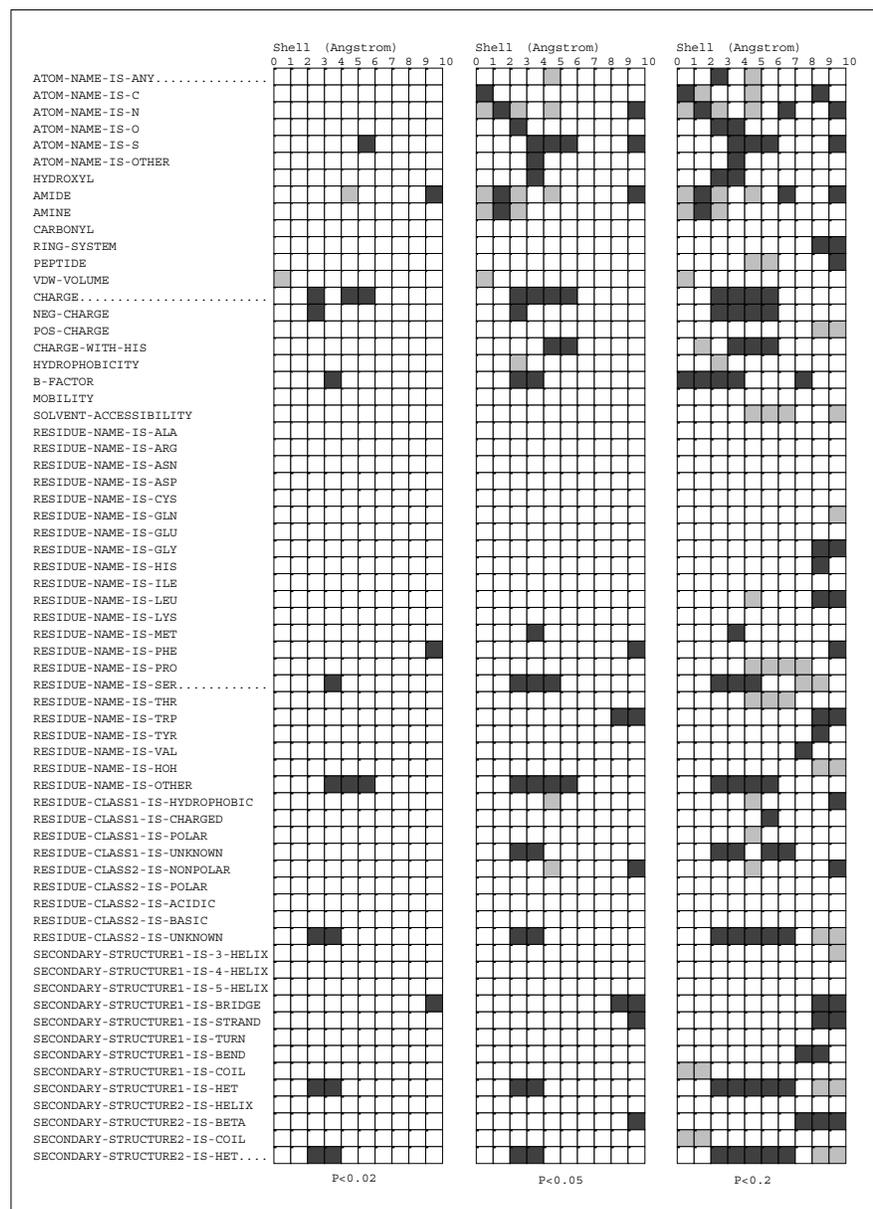
indicates whether the property was more abundant in sites or nonsites. The final column lists residues in the structure of 1TLD that are near the volume in which the significant difference is observed.

The analysis of these superimposed sites using only a radial analysis (shown in Fig. 2), centered on the NE2 of the catalytic histidine, reveals only the features that were used to align the sites: the catalytic histidine, serine, and aspartic acid and oxyanion hole nitrogens. This finding suggests that, in the case of serine protease active sites, it is important to use an oriented grid of features to provide greater spatial resolution. Analyzing the radial distribution of properties is not sufficient to uncover more subtle conserved features. Upon completing an oriented analysis, we find two classes of findings: those that are expected, and those that are novel.

The first class of results are those expected as a direct consequence of our site alignment procedure. In general, the

central histidine (His57 in 1TLD) has been controlled for in the nonsites and so should have only a minimal presence in the results. The serine and aspartic acid residues of the catalytic triad should appear in abundance in the sites, as seen in Table 1. In addition, the carboxylic carbon of the catalytic aspartic acid was used in the alignment, and so this result (abundance of property ATOM-NAME-IS-C) follows trivially, as does the associated negative charge around Asp102 although the spatial extent of the negative charge is a bit more disperse than might be expected. In addition, the nitrogen atoms that participate in the oxyanion hole, believed to stabilize the transition state (and used for the alignment in the case of REM\*), are observed consistently in the four structures. Also, as expected, the hydroxyl group (and its oxygen) of the catalytic serine is

Figure 2



Radial analysis of protease active sites. The results of the radial analysis are shown in a two-dimensional graphic, with the property names along the left edge. Each strip presents the data for one value of the significance cutoff, shown along the bottom. The columns of each strip are the results for each shell radius (measured in Å), as labeled across the top of each strip. Dark grey boxes indicate property/shell pairs where the site values exceed the nonsite values; light grey boxes mark the converse. Unlabeled boxes show no significant differences between sites and nonsites. Most of the property names are self-explanatory. Mobility is computed as the minimum number of bonds to the nearest  $C\alpha$  or backbone atom. CHARGE-WITH-HIS is a variant of the charge property that places +0.5 on the histidine nitrogens, to account for the possibility of charged histidine sidechains.

seen as a conserved feature. The increased significance of the RESIDUE-NAME-IS-SER relative to RESIDUE-NAME-IS-ASP confirms the observation that “the precise geometric orientation of the Asp is not conserved relative to the Ser–His catalytic diad” [5]. It is also believed that features of secondary structure are not conserved across the three natural serine protease classes [5]. We observe this to be generally true, but do observe that the property reflecting a coil geometry is conserved around the backbone of the catalytic aspartic acid and serine.

The second class of results are those that are unexpected or novel. First, the property reflecting mobility of

sidechains (a measure of the number of freely rotatable bonds between an atom and the associated  $C\alpha$ ) is abundant in the region near the entrance to the active site in trypsin. Although there are no atoms occupying this space in trypsin, the other molecules have highly mobile sidechains, presumably to allow accommodation of the ligand as it enters the active site. Second, we note that there is increased polarity between the catalytic serine and the oxyanion hole, as would be expected given the binding role of this pocket. The oxyanion nitrogens seem to be the most obvious of a larger set of polarizable atoms that can accommodate the oxyanion intermediate (or its absence). Third, and in addition to the oxyanion hole

amide moieties, we observe an additional, relatively conserved amide in the region on the other side of the catalytic aspartic acid, opposite the oxyanion hole. This may reflect an additional stabilizing electronic interaction in the presence of the oxyanion intermediate. Fourth, there is an abundance of negative charge around the catalytic histidine, above the plane of the histidine, on the side opposite the catalytic aspartic acid. This may reflect an additional stabilization of the active site geometry. Finally, we see a significant abundance of leucine, glycine, valine and tyrosine around the active site. These may facilitate nonspecific polypeptide ligand binding and stabilization.

As noted in the Materials and methods, we used a mirror image of the carboxypeptidase structure for the main experiment. In order to assess the robustness of our findings, we ran the experiment after aligning the wheat serine carboxypeptidase in the native conformation, with no mirror image. As expected from the analysis reported in [10], the number of conserved properties observed by our system decreases, including the inability to recognize the oxyanion hole nitrogens, which do not superimpose well without the mirror image. This result highlights the importance of aligning the sites in a manner appropriate for comparison, and we suggest that the mirror image is the right structure for our experiment. We can therefore consider our result to be an independent verification that the active site of wheat carboxypeptidase is an approximate enantiomer of the trypsin-like active sites. Because of our sampling resolution (we average over square volumes with an edge length of 2.4 Å), we are evaluating only the relative physical locations of chemical groups and atoms and not their detailed stereochemistry, which must of course obey normal rules of protein chemistry. None of our findings is dependent on assumptions about stereochemistry.

To assess the false-positive rate, a sensitivity analysis compared two random subsets of nonsite histidines to calibrate the method and ensure that the rate of spurious findings was low. Since there should be no systematic variation between the randomly partitioned nonsites, this experiment provides an estimate of the false-positive rate obtained in the other experiments (due principally to the small number of sites). The experiment was performed for three different, random partitions of the nonsites. The results showed 2, 4, and 13 significant results at the level of significance (0.01) used as an upper cutoff for the data presented. With a uniform sampling of the nonsites, the significance level of  $P < 0.005$  appears to mark the boundary between interesting results that either confirm or extend our knowledge of the active site, and questionable results containing a large fraction of false positives. The two major structural constellations in the active site, the catalytic triad and the oxyanion hole, are robustly detected. The catalytic triad was explicitly used in the site

alignment procedure; the oxyanion hole (used only for the alignment of REM\*) was expected and found.

The choice of sites that are nonhomologous is an important one. In previous work we have demonstrated that the choice of control group is an important experimental consideration and suggested that selecting proteins with a strong evolutionary relationship might result in detecting features that were common but of little functional importance [2]. Of course, statistical significance does not guarantee biochemical significance. First, sampling and problems of uniform representation can cause irrelevant features to be reported (false positives). Second, even without sampling error, the magnitude of the difference between two groups that have been shown to be statistically significant may be small from a biochemical perspective. In spite of these possible problems, the program reduces a large amount of data to a cognitively manageable set; the results are hypotheses worthy of further inspection to either confirm them or to rule them out.

Our program is designed for exploratory analysis of aligned microenvironments. As such, it is able to suggest a number of hypotheses about serine protease active sites. First, it appears that there is a region in all four molecules which must be either vacant or filled with highly mobile molecules that can serve as a partial gate to the entering substrate (in the region of His65 and Met215 in subtilisin). This observation is similar to published reports that there are relatively flexible loops around the reactive center to allow for an induced fit [9]. Second, there are significant amounts of increased polarity in the region around the oxyanion hole and catalytic serine in the absence of substrate, suggesting a stabilization of the region when not confronted with the oxyanion intermediate. Third, there is a network of partial charges around both the catalytic aspartic acid and histidine, with partial negative charges around the histidine, and some amide-group stabilization of the aspartic acid. Finally, there is an increased abundance of leucine, glycine, valine and tyrosine in locations that suggest they may be involved in nonspecific stabilization of peptide ligand. Each of these hypotheses is based on reliable observations from our data, but the final interpretations require careful energetic evaluation of the active site, and perhaps site-directed mutagenesis to see if the hypothesized effects can be perturbed.

Perhaps the most striking finding is that there are few features that are absolutely conserved over the four protease molecules we studied. After accounting for the catalytic triad and oxyanion hole, there are few observed features present in all four sites. Instead, we have a second class of findings that are found significantly more frequently, but not universally, in the active sites than in the control nonsites. These findings are characterized by being never (or rarely) seen around noncatalytic histidines, but being seen

in two or three of the four catalytic sites. They include the relative abundance of valine, tyrosine, leucine, and glycine, and the configuration of polarity and hydrophobicity around the catalytic histidine. In addition to the absolute requirements for proteolytic activity, there are some relative requirements that may be met in a variety of fashions.

It is generally assumed that the serine protease molecules have the same general mechanism of action. Our results show that there are additional environmental features around the catalytic triad that should be included in detailed hypotheses about enzyme mechanism. Conversely, features that did not appear as significantly conserved may be associated with the process of enzyme specificity, or may play other catalytic roles that affect rate constants or affinities. As we accumulate more information about serine protease molecules with no ancestral relationship, we can dissect the details of these second-order requirements.

We have built a general-purpose computer tool for studying structure/function relationships. The tool is designed to be applicable to any set of protein sites. The basic utility of the tool is established by demonstrating that it uncovered most of the published common features of serine proteases. Even though the serine proteases are very well studied, we expected (and found) some relatively conserved properties that had not been previously recognized. In general, the new findings are subtle and do not occur universally in the four sites. Our results suggest two take-home messages. First, a statistical comparison of superimposed protein environments is able to automatically detect conserved spatial features. In this case, about half the findings are well documented and the other half are new. Second, in the case of the serine proteases there are subtle features, not previously recognized as critical for catalysis, that deserve closer examination in light of their relative abundance in four nonhomologous active sites.

## Materials and methods

The primary goal of our method is to find the properties and locations for which the oriented site data differ from the nonsite (control) group. In this section, we describe how the protein sites were chosen, how

**Table 2**

**Proteins used in this study.**

Family	Name of protein	PDB identifier
Chymotrypsin	$\beta$ -trypsin	1TLD
Subtilisin	Subtilisin bl	1ST3
Carboxypeptidase	Serine carboxypeptidase II	REM*
Antibody	Catalytic antibody	1EAP

The PDB identifiers for the chosen representation of each class of serine protease are listed. Note that the REM\* data have not yet been issued a PDB identifier.

they were aligned to a common coordinate system, how the significant properties and locations were found, and how the resulting data were prepared for presentation. Most of the assumptions underlying various choices in these procedures have been subjected to sensitivity analysis.

### Choice of proteins

We identified structures of serine proteases from four different nonhomologous families: chymotrypsin, subtilisin, carboxypeptidase, and an artificially engineered antibody fragment with proteolytic activity [5,10,11]. One representative protein was chosen from each family in those cases where alternatives were available. The PDB identifiers of the chosen proteins are given in Table 2; the structures at the region of the active site are displayed in Figure 1.

The structure of the antibody 1EAP requires special processing. As noted in [11], a rotation of 180° is needed to put the active oxygen of Ser-H99 into the proper geometry relative to the rest of the active site. This modified structure was used for all of our experiments.

In addition, the structure of wheat serine carboxypeptidase poses a special problem. A recent report [10] has demonstrated, based on the crystal structure of a protein-inhibitor complex, that the geometry of its active site is approximately the mirror image of the active site of trypsin family members. Thus, in order to compare the active sites on equivalent terms, we inverted the wheat serine carboxypeptidase structure, producing a site that could be aligned with the others, as detailed in the next section. Because the structure has not yet appeared in the PDB, we refer to this structure by the identifier REM\*.

### Alignment of protein sites and choice of nonsites

Because our method operates on the protein data considered as three-dimensional structures, the protein active sites must be brought into a common alignment in order to be compared. This was performed by choosing the rotation and translation producing the best RMS fit between a selected subset of the atoms in each structure. One of the

**Table 3**

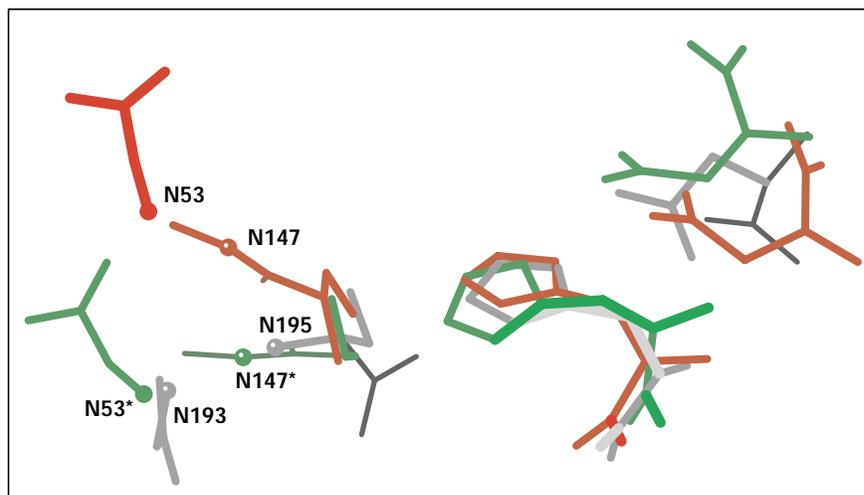
**Atoms used for aligning active sites.**

PDB ID	Pairing of atoms used in alignment					
1ST3	His62 CB	His62 ND1	His62 NE2	Ser215 OG	Asp32 CG	—
1TLD	His57 CB	His57 ND1	His57 NE2	Ser195 OG	Asp102 CG	—
REM*	His397 ND1	His397 CE1	His397 NE2	Ser146 OG	Gly53 N	Tyr147 N
1TLD	His57 ND1	His57 CE1	His57 NE2	Ser195 OG	Gly193 N	Ser195 N
1EAP	His35 CG	His35 ND1	His35 CE1	His35 NE2	His35 CD2	Ser99 OG
1TLD	His57 CG	His57 ND1	His57 CE1	His57 NE2	His57 CD2	Ser195 OG

Alignments used to bring each active site into alignment with the active site of 1TLD.

Figure 3

Alignment of mirror image of 1WHS. The catalytic triad and oxyanion nitrogens of the wheat serine carboxypeptidase active site are shown (black), along with their mirror image (grey), and the trypsin (1TLD) active site to which they were aligned (light grey). The oxyanion nitrogens of 1TLD are labeled (N195 and N193). The oxyanion nitrogens from wheat serine carboxypeptidase are labeled (N53 and N147) as are the corresponding nitrogens after generating a mirror image (N53\* and N147\*). The mirror image data give a better fit to the trypsin site. In particular, the oxyanion nitrogens are displaced by more than 2 Å in the native conformation of 1WHS. As described in the text, the mirror image data (REM\*) were used in the experiments.



proteins (1TLD) was arbitrarily chosen as the basis of the coordinate system; the others structures were transformed into those coordinates. For the engineered antibody and the subtilisin, we took the catalytic triad (Asp, His, Ser) as constituting the essential atoms of the active sites, with some adjustments necessary for missing residues (in particular, the antibody 1EAP lacks Asp in the active site). In general, atoms contributing directly to the site activity (such as trypsin's Ser195 OG) were used. The details of which atoms were used is listed in Table 3. In the case of the wheat serine carboxypeptidase, we first generated the mirror image structure and then fit the structure to 1TLD by aligning the two oxyanion hole nitrogens (Gly53 and Tyr147, with Gly193 and Ser195), the hydroxyl oxygens of the catalytic serines, and the outer ring atoms (ND1, CE1, NE2) of the catalytic histidine. The alignment of REM\* (mirror image) and 1TLD is shown in Figure 3. The alignment shows a good visual match and had the oxyanion hole nitrogens in similar positions [10]. Failure to generate a mirror image results in a good fit of the catalytic aspartic acid, serine and histidine, but incompatible locations for the oxyanion hole nitrogens. For all calculations, each site or nonsite contained all the atoms within a sphere of radius 10 Å, using atom NE2 of the histidine as the center.

The nonsites are the control group. They are chosen to be similar to the sites, except that they lack catalytic activity. We chose histidine residues in the same proteins as were used for the sites. By using non-catalytic histidine residues as controls, we guaranteed that our system would not detect features of the catalytic environment that were directly attributable as 'normal' histidine environmental elements. We transformed the nonsites so that the ring of the nonsite histidine aligned exactly with the histidine of the active site. To avoid bias, the nonsite histidines were aligned to one of the site histidines, chosen at random. In a second set of experiments, we used a larger group of histidine residues from other serine proteases. The PDB identifier and the number of sites and nonsites from each protein are given in Table 4.

#### *Finding statistically significant properties and locations*

The algorithm searches over a 3D grid holding the atoms near the active site for statistically significant properties that distinguish the sites from the nonsites. The details of the grid search are given in [3] and will be repeated only briefly here. The 3D grid contains small, uniformly spaced cubic cells of a size (edge length 0.83 Å) such that each cell is unlikely to contain more than one atom. Properties are associated with atoms and represent atomic identity, residue membership, and a variety of biochemical and biophysical properties. They are defined precisely in the appendix of [2]. For each property, its value is

computed for each atom and inserted into the proper cell. In some cases, the properties intentionally 'leak' into neighboring cells to simulate properties, such as charge, that exhibit field effects.

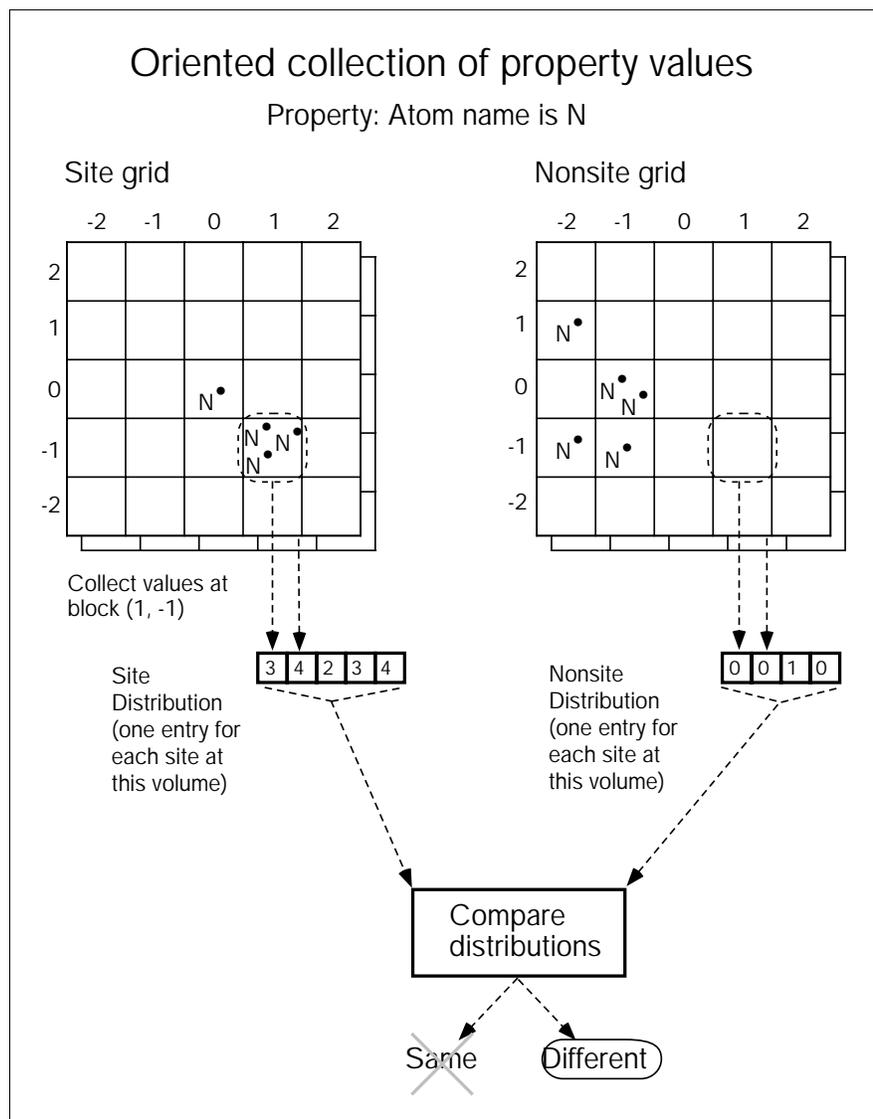
Each grid cell is quite small, and the range of values each will contain does not produce a distribution amenable to statistical analysis. Therefore, the cells are aggregated into larger volumes, a process we refer to as 'collection'. The collection volumes are a way of indexing into the site microenvironment. The current implementation uses two collection schemes: radial (nonoverlapping shells around site center), and oriented (nonoverlapping blocks, each comprising  $3 \times 3 \times 3 = 27$  grid cells). The property value for a collection volume (shell or block) is the sum of the values of the contained cells. Taking the values for corresponding collection volumes (for each of the active sites) produces a distribution for the property values at the volume in question. A similar distribution for the nonsites (same property, same volume) can also be produced and compared to the site distribution. If those two distributions differ to a statistically significant degree, then that property/location pair is reported as being of potential interest. This comparison is repeated for each possible choice of property and location and filtered

Table 4

#### Distribution of sites and nonsites.

PDB ID	No. sites	No. nonsites
1TLD	1	2
1ST3	1	2
1EAP	1	2
REM*	1	2
1ARB	0	2
1GCT	0	1
1TON	0	2
3EST	0	2
1SCA	0	2
2ST1	0	2
<b>Totals</b>	<b>4</b>	<b>19</b>

The number of sites (catalytic histidine residues) and nonsites (noncatalytic histidine residues) taken from each of the proteins used.

**Figure 4**

Summary of the computational method. The analysis of the data in an oriented grid is presented schematically here. The values for a given property are placed in the grid, grouped into larger cells and assembled across a number of sites (or nonsites) to form a distribution. The distributions are compared at each volume and for each property. Differences are assessed statistically using the Mann–Whitney rank sum test.

to retain for further consideration only those results significant to  $P < 0.01$ . The comparison for significance uses a nonparametric test (Mann–Whitney rank sum) because the distributions are often not normally distributed. A schematic diagram of the oriented collection and comparison procedure is shown in Figure 4.

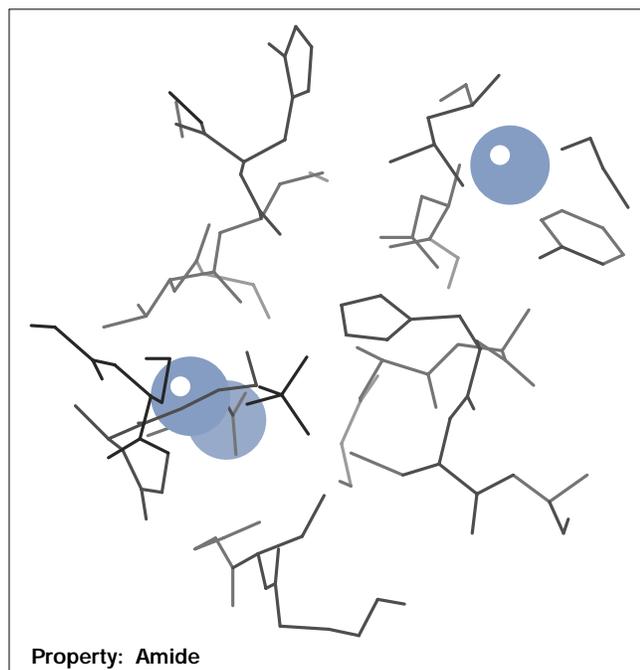
#### Preparing the data for presentation

In addition to a textual display of the findings sorted by significance, it is convenient to have a visual presentation. The radially collected distributions can be shown in a two-dimensional format, indexed by shell radius and property (as in Fig. 2). For the oriented collection, blocks of significant results for the same property are clustered together and shown as spheres of the same color, overlaid on a line drawing of the three-dimensional structure of the aligned sites. Figure 5 shows the active site of 1TLD with three spheres indicating where there is a significant abundance of amide groups across the four protease molecules. The amide groups in the lower left of Figure 5 correspond to the location of the oxyanion hole. The amide in the upper right is close to the catalytic aspartic acid. A kinemage computer file (for use with the

MAGE program [12]) with the four aligned serine protease molecules and the full set of sphere labels for the properties listed in Table 1 is included as Supplementary material.

#### Sensitivity analysis

We measured the sensitivity of our experiment to the choice of nonsites. We evaluated the significance level necessary to avoid spurious findings. Even though the test for statistical significance reports a  $P$ -level that measures the false-positive rate, we decided also to measure it experimentally. The nonsites were randomly partitioned into two groups; one group was used as the study group, the other as the control group. Although these groups contain inherent variation, when compared to each other there should be no large cross-group differences, and the number of reported significant property/volume pairs can provide an estimate of the false-positive rate. This estimate is only approximate, because some of the properties are not statistically independent – ATOM-NAME-IS-O and HYDROXYL strongly correlate, for example. We repeated the partition of the nonsites and cross-comparison in three different experiments.

**Figure 5**

Location of abundant amide groups in four serine protease molecules. The structure of 1TLD is shown, with three spheres drawn in the volumes that show an abundance of amide groups across the four protease molecules examined. The closely associated pair of spheres is in the region of the oxyanion nitrogens. The lone sphere is a newly observed, relatively conserved amide functionality near the catalytic aspartic acid. Consult the Supplementary material for a kinemage containing a 3D presentation of the structures and results.

#### Supplementary material available

A kinemage file, suitable for viewing with the MAGE program, is available as Supplementary material (published with this paper on the internet). This file allows the viewing of the four superimposed serine protease structures in 3D and allows the user to examine the location of significant properties using colored spheres overlaid on the structures, as in Figure 5. It also contains the raw data counts for sites and nonsites for all of the significant properties contained in Table 1. (This file is also available over the internet at <http://www-smi.stanford.edu/projects/helix/pubs/folddes96/>)

#### Acknowledgements

RB Altman is a Culpeper Medical Scholar and is supported by NIH grant LM-05652. Some computing facilities were provided by the CAMIS resource, funded by NIH #LM-05305. SJ Remington provided the coordinates for the crystal structure of wheat serine carboxypeptidase bound to chymostatin.

#### References

- Bernstein, F.C., *et al.*, & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.
- Bagley, S.C. & Altman, R.B. (1995). Characterizing the microenvironment surrounding protein sites. *Protein Sci.* **4**, 622–635.
- Bagley, S.C., Wei, L., Cheng, C. & Altman, R.B. (1995). Characterizing oriented protein structural sites using biochemical properties. In *Intelligent Systems for Molecular Biology*. pp. 12–20, Cambridge, England, AAAI Press.
- Greer, J. (1990). Comparative modeling methods: application to the family of the mammalian serine proteases. *Proteins* **7**, 317–334.
- Perona, J.J. & Craik, C.S. (1995). Structural basis of substrate specificity in the serine proteases. *Protein Sci.* **4**, 337–360.
- Warner, P., Green, R.C., Gomes, B. & Strimpler, A.M. (1994). Non-peptidic inhibitors of human leukocyte elastase. 1. The design and synthesis of pyridone-containing inhibitors. *J. Med. Chem.* **37**, 3090–3099.
- Branden, C. & Tooze, J. (1991). *Introduction to Protein Structure*. New York, Garland Publishing, Inc.
- Whiting, A.K. & Peticolas, W.L. (1994). Details of the acyl-enzyme intermediate and the oxyanion hole in serine protease catalysis. *Biochemistry* **33**, 552–561.
- Aulak, K.S., Davis, A.D., Donaldson, V.H. & Harrison, R.A. (1993). Chymotrypsin inhibitory activity of normal C1-inhibitor and a P1 Arg to His mutant: evidence for the presence of overlapping reactive centers. *Protein Sci.* **2**, 727–732.
- Bullock, T.L., Breddam, K. & Remington, S.J. (1996). Peptide aldehyde complexes with wheat serine carboxypeptidase II: implications for the catalytic mechanism and substrate specificity. *J. Mol. Biol.* **255**, 714–725.
- Zhou, G.W., Guo, J., Huang, W., Fletterick, R.J. & Scanlan, T.S. (1994). Crystal structure of a catalytic antibody with a serine protease active site. *Science* **265**, 1059–1064.
- Richardson, D. & Richardson, J. (1992). The kinemage: a tool for scientific communication. *Protein Sci.* **1**, 3–9.

Because **Folding & Design** operates a 'Continuous Publication System' for Research Papers, this paper has been published via the internet before being printed. The paper can be found in the BioMedNet library at <http://BioMedNet.com/> – for further information, see the explanation on the contents page.