Characterizing oriented protein structural sites using biochemical properties

Steven C. Bagley, Liping Wei, Carol Cheng, and Russ B. Altman

Section on Medical Informatics
Stanford University School of Medicine, MSOB X-215
Stanford, CA, USA 94305-5479, (415) 723-6979
bagley, wei, cheng, altman@camis.stanford.edu
http://www-camis.stanford.edu/projects/helix/features.html

Abstract

A protein site is a region of a three-dimensional protein structure with a distinguishing functional or structural role. Certain sites recur in different protein structures (for example catalytic sites, calcium binding sites, and some types of turns), but maintain critical shared features. To facilitate the analysis of such protein sites, we have developed a computer system for analyzing the spatial distributions of biochemical properties around a site. The system takes a set of similar sites and a set of control nonsites, and finds differences between them. Specifically, it compares distributions of the properties surrounding the sites with those surrounding the nonsites, and reports statistically significant differences. In this paper, we use our method to analyze the features in the active site of the serine protease enzymes. We compare the use of radial distributions (shells) with 3-D grids (blocks) in the analysis of the active site. We demonstrate three different strategies for focusing attention on significant findings, based on properties of interest, spatial volumes of interest, and on the level of statistical significance. Finally, we show that the program automatically identifies conserved sequential, secondary structural and biophysical features of the serine protease active site, using noncatalytic histidine residues as a control environment.

Introduction

One goal of molecular biology—to uncover the relationship between macromolecular structure and biological function—creates important demands for computational assistance. For structure analysis much of the assistance to date has been in the form of tools for scientific visualization, along with algorithms for computing particular biochemical properties (such as solvent accessibility (Kabsch and Sander, 1983) or energy fields (Goodford, 1985)). These tools often focus on single properties and individual protein structures. Although important in preliminary investigations, such a focus potentially misses relationships that would only become apparent in larger data sets. The ongoing enterprise of protein structure elucidation is providing an ever-growing database of atomic coordinates, and it is now possible to perform statistical analyses on the features that come together to form specific structural milieus.

In this paper we report on a computational tool for analyzing local regions (called site microenvironments) in sets of three-dimensional protein structures. These sites are regions that are of structural or functional interest, but which are incompletely understood. The system builds a representation of the atomic positions, and augments it with spatial distributions of biochemical and biophysical properties. These properties (currently numbering about 30) include labels of atom and residue names, common functional groups (e.g., carbonyl and hydroxyl groups), secondary structure types, and physical quantities such as hydrophobicity, mobility, and charge. The abundance of each property in a small volume within the sites is compared to its abundance in the corresponding volume of a set of control nonsites. If the distribution of values in the sites differ from those in the nonsites to a statistically significant degree, the program reports the distinguishing property and the associated spatial volume. These property/volume pairs are preliminary hypotheses about the nature of the site, and can be used to guide further investigation.

This paper is an extension of (Bagley and Altman, 1995) , which reported on the use of radial distributions of key biochemical and biophysical features in the analysis of three different sites: Ca^{2+} binding sites, Cys-Cys bonding (disulfide bridges), and serine protease active sites. As one example, the oxygen rich shells that surround Ca^{2+} ions showed up clearly in the radial distributions. The program also found many critical features of the disulfide environment, although some previously reported features were missing or observed only at low levels of statistical significance. For the serine proteases, the focus of this paper, the radial distribution showed the essential elements of the active site, but without reference to their orientation around the classical "catalytic triad" of histidine, serine and aspartic acid. As noted in the earlier paper, one disadvantage of using a radial distribution is that it ignores the relative orientation of the features within the shells. This is not a problem if the site turns out to be organized either with complete spherical symmetry or with regard only to the distance from a central location (for example, if only inverse-square electrostatic forces were at play).

However, it is expected that most sites are not isotropic. Therefore, we introduce here the use of an oriented method for analyzing sites with richer local structure once they have been rotated into a common alignment.

There are several themes in this work. First, the analysis uses a redundant, intermediate-level vocabulary of atom and residue properties, instead of using only atomic position information. The atoms of the amino acids determine the biochemical characteristics of a site, but many characteristics can be realized in multiple ways. In a sense, it matters less which atoms are in the site and more what biochemical environment they create. By moving from atomic level descriptions to descriptions of the properties that are ultimately selected for by evolution, we may simultaneously reduce the amount of raw data that needs to be considered, and move to a description language that better reflects the factors likely to define the site. The utility of property-based representations has already been shown for inverse protein-folding (Bowie et al., 1991) and characterizing catalytic residues (Zvelebil and Sternberg, 1988).

Second, it is important when studying structural and functional features to provide some grounding for the term "significant". The complete atomic description of a single macromolecule contains a large amount of raw data such that it is difficult to separate the critical features from the merely accidental. Comparison of a set of structures with a background distribution can provide such a separation if the background distribution is chosen carefully. Standard tests of statistical significance then can be used to compute levels of significance, providing objective measures that can be used in cross-study comparisons. Although statistical significance does not guarantee biochemical importance, it does introduce a level of rigor in building preliminary hypotheses.

Third, the use of an explicit control group provides an adjustable focus for determining the kind and degree of important properties that are to be reported. The simplest assumed background is spatial uniformity, which has been successfully used in studies of atomic and residue positions (Warme and Morgan, 1978; Singh and Thornton, 1992). However, it is useful to move beyond assumptions of uniformity by choosing the background distributions from actual protein structures. Explicit choice of the control group can adjust the amount of extraneous findings that are reported. For example, in studying the binding sites of calcium ions, using randomly chosen atoms as the control group highlights the difference between binding and non-binding regions, while using other cations as the control group (for example magnesium or zinc) emphasizes the details specific to the binding of the calcium cations.

Methods

The goal of the algorithm is to produce a succinct characterization of the significant differences in the occurrence of a property in a set of sites with respect to a set of nonsites. For a given property, we fill a three-dimensional grid with the property values computed for the atoms in the sites, and a separate grid with the values for the nonsites. The values of the property within a volume of interest can be collected to form a distribution of values associated with that volume. The distribution for the site instances is compared with the distribution for the nonsite instances; if these distributions differ to a statistically significant degree, then the property name and the region of the microenvironment (the collection volume) are reported. The original algorithm, as described in (Bagley and Altman, 1995) collected property values over concentric shells and so all features were radially averaged. In order to analyze sites in an orientation-sensitive manner, we can divide the site into a collection of cubic volumes (instead of concentric shells), and perform our averaging over these volumes without losing information about orientation.

The site and nonsite files are prepared by extracting regions from the atomic coordinate files stored in the Protein Data Bank (Bernstein et al., 1977). The sites are identified by the user by their three-dimensional position and a radius, typically 10 Å. The nonsites forming the control group are defined similarly. Each site or nonsite instance is stored in a separate file as a list of atoms.

The property values are placed in a three-dimensional grid with cubical cells having an edge length chosen so that only rarely will two atoms occupy a single cell (roughly 0.83 Å). The properties span a wide range of biochemical and biophysical parameters; the list of properties used in this paper is shown in Figure 2. They can be grouped into classes: atom-based (the identity of the atom, one of C, O, N, other, or any), functional-group-based (what functional group is the atom a member of), residue-based (what residue is the atom a member of), secondary structure-based (the secondary structure of the atom's residue), and a handful of others (hydrophobicity, mobility, charge). The precise definitions of all the properties is detailed in (Bagley & Altman, 1995). Each property is computed by a separate subroutine; the list of properties to be used on any invocation is set by the user.

The cubical cells that contain property values are too small to analyze for statistically significant occurrence of properties. They are also much smaller, in general, than the root mean squared deviation between even very similar

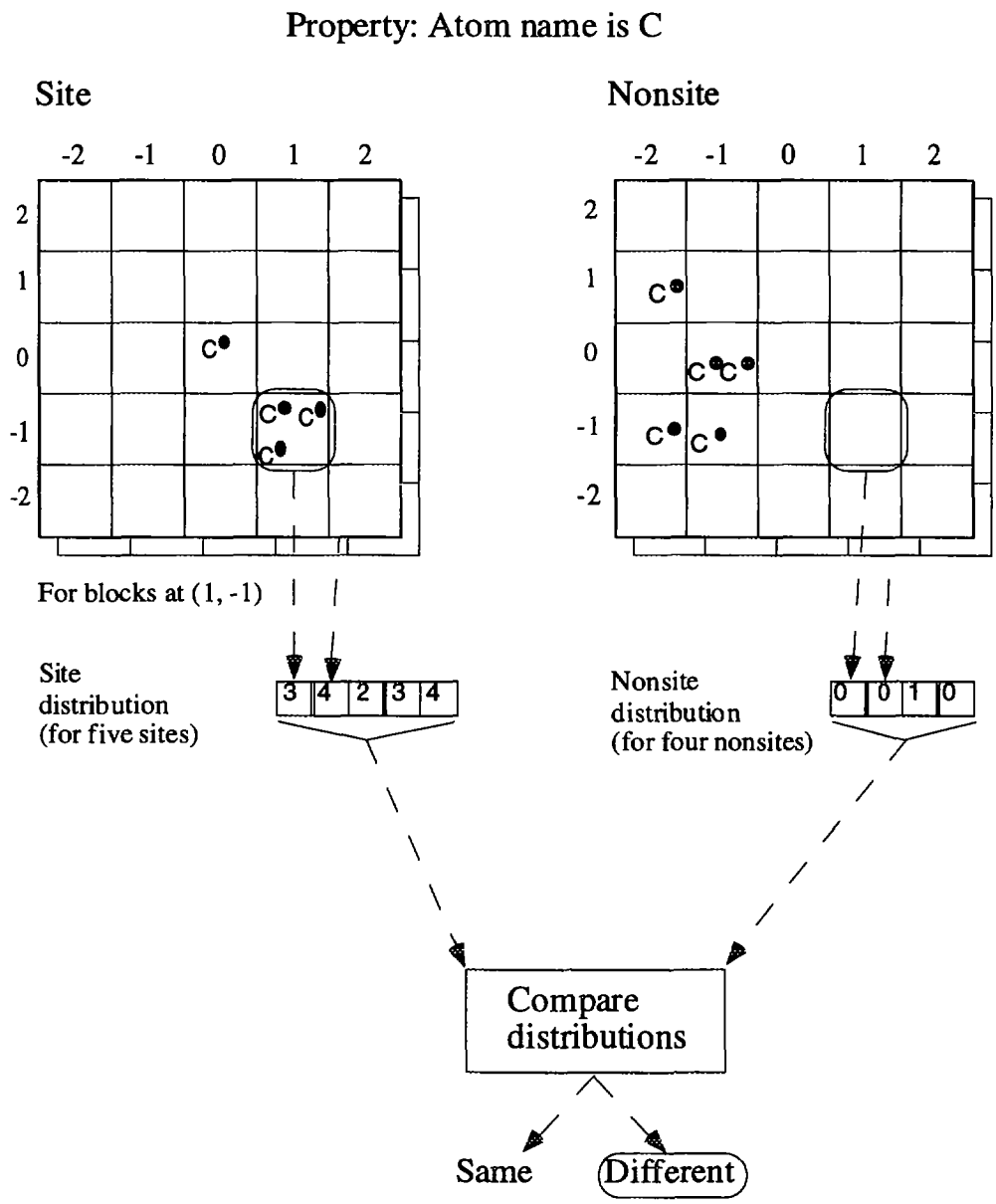

Figure 1. Graphical depiction of the procedure for collecting values of oriented data. A representative site and nonsite are shown in an array of blocks (only two dimensions of which are displayed). Each block is an aggregate of 27 (=3x3x3) underlying grid cells. For each block, the values from all the contained grid cells are summed to produce a single property value. In this example the property is "Atom name is C"; the property value is therefore a count of the number of C atoms falling anywhere inside the block. The block shown at (1,-1) contains three C atoms. The five site instances (of which only the top one is visible) produce five values for this property in this volume. The corresponding nonsite distribution, shown here with four values, can be compared with the site distribution using a non-parametric statistical test (Mann-Whitney rank sum). In this example, the system would conclude that the two distributions are significantly different. The system compares the distribution of every property within every collection volume (block), and reports all property/block pairs that differ significantly.

sites. Thus, we must group the cells into larger volumes that contain sufficient data to support statistical analysis. We call the extraction of the property values out of the grid "collection". We are currently using two collection procedures: radial (cells in spherical shells around the center of the site are aggregated), and oriented (blocks of 27 cells–3x3x3 grid cubes–are aggregated). In both cases, collection reduces all the property values contained in the given volume to the sum of those values (a measure of abundance). For the radial distribution, the data are collected over shells of 1 Å thickness, based on the distance from the center of the site. The oriented collection aggregates grid cells into blocks that are 2.4 Å on each side, and which are addressed using an x-y-z coordinate system relative to the site center. Use of the oriented collection procedure makes sense only if the sites have been aligned to a common coordinate system. The collection stage results in two distributions for each property, one for the site and one for the nonsites, and is illustrated schematically in Figure 1. A pseudocode summary of the procedure is given in the appendix.

Because the site and nonsite distributions are not, in general, normally distributed, they are compared using a non-parametric Mann-Whitney rank sum test (Glantz, 1987) . As with other hypothesis testing procedures, the test compares two distributions to try to reject the null hypothesis (that the two distributions are the same). The threshold of statistical significance is set by the user. Currently, all results with $P > 0.01$ are ignored. The result of the statistical test is a list of pairs of properties and collection volumes that produce P levels at or better than (i.e., below) the threshold. It is important to note that we report significance only of single property-volume pairs, and not the significance of the entire ensemble of pairs. Thus, if we report one hundred pairs with a significance level of $P < 0.01$, we can expect one of these, on average, to be spurious. For the Mann-Whitney test to operate reliably, the larger group (either sites or nonsites) should have at least 9 members which prohibits its use for very small sample sizes.

For the radial distribution, the results are plotted in a two-dimensional display indexed by property and shell radius (Figure 2). The oriented distributions require the use of further data reduction followed by a 3D graphics display. The collection volumes for the oriented distributions are cubic blocks in a three-dimensional space. To cluster the blocks, we compute the connected sets (connected components), as determined by the next-door neighbor relation (sharing a face, edge, or corner). Each connected set is displayed using a cloud of dots dispersed over the volume, which provides a visual indication of its location and extent (Figure 3).

The system is written in Common Lisp, and has been tested in MCL 2.0 for the Apple Macintosh. The running time for a set of site and nonsites of the size reported in this paper is several hours. A translation of the procedure into C is being tested.

The serine protease data set

To facilitate the comparison of features found by the program with those already reported in the literature, we chose the active site of the serine protease family, which is known to have a rich and interesting three-dimensional organization (Warshel et al., 1989; Greer, 1990; Zhou et al., 1994; Perona and Craik, 1995) . The activity of the site is due to the catalytic triad: a His with Ser and Asp residues that are nearby in space, but not close in the protein sequence. To prepare the data set, the sites were defined to be the active sites of six serine proteases, with the NE2 atom of the His ring used as the site center, and including all atoms within a radius of 10Å. For the control group, the nonsites were centered on those His residues (also at the NE2 with a 10Å radius) found in the same proteins that are not in the catalytic triad. The purpose of this control is to remove the His and its local effects from the analysis in an attempt to better define the properties of the surrounding environment relevant to proteolysis, and to avoid rediscovering a list of features typically associated with His residues. The control residues were drawn from the same proteins, but could be drawn from unrelated proteins, depending on the goals of the user. The number of nonsites varies for each protein. The Brookhaven IDs, names, and number of sites (and nonsites) for the six proteins used in this study are 1ARB, Achromobacter protease I, 1 (5); 2GCT, γ-chymotrypsin 1 (1); 1SGT, trypsin, 1 (0); 1TON, tonin, 1(6); 3EST, native elastase 1 (5); and 4PTP, β-trypsin 1 (2). All of these proteins are in the trypsin family of serine proteases, and are structurally very similar; therefore, our study will emphasizes properties common to members of this family, and de-emphasizes those features distinguishing the individual family members.

For the oriented analysis, sites and nonsite files must be in a common coordinate system with the proper orientation. The PDB coordinates of the protein 4PTP were arbitrarily chosen as the common coordinate system. The site files were transformed by translation (to bring the site centers into coincidence) and rotation (to produce the smallest RMS distance between the site atoms and the 4PTP atoms, measured at the Cα locations of the catalytic His, Asp, and Ser). A similar transformation was applied to the nonsites using the backbone atoms of the defining His residue (at Cα, C, and O of the residue and the N of the next residue).

Results and Discussion

The results for the radial distributions of the serine protease active sites are shown in Figure 2. We focus here on a few of properties that relate directly to the (well documented)

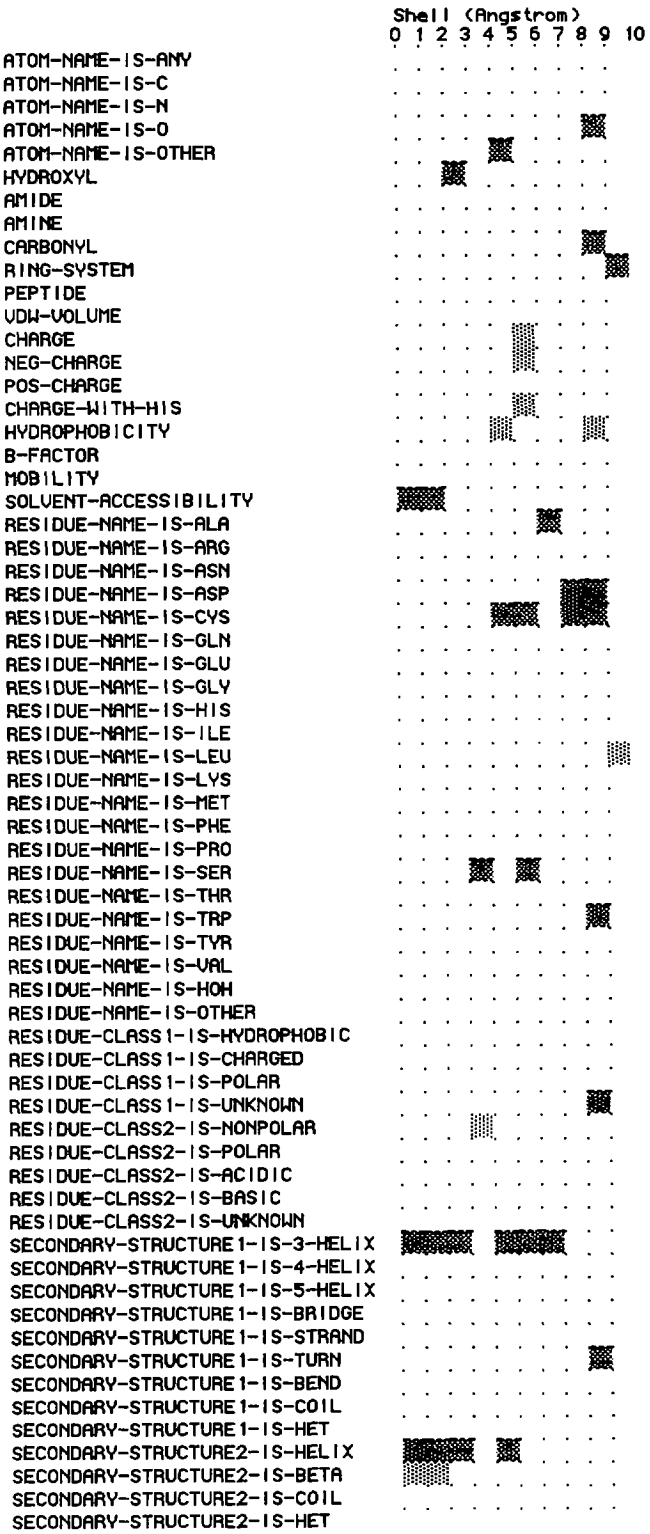

Figure 2. The results for radial distributions around the serine protease active site showing the significant properties and shells. The properties are arrayed on the vertical axis; the shell volumes (by increasing radius) along the horizontal axis. Each non-white cell marks a statistically significant result. Dark gray marks property/shell pairs for which the site values exceeds the controls; light gray marks the converse. For example, atoms from Asp residues in shell 7-9Å (property RESIDUE-NAME-IS-ASP) are more abundant in the sites than the nonsites, and there is a relative lack of beta secondary structure in shell 0-2 Å (property SECONDARY-STRUCTURE2-IS-BETA).

The first group of properties (ATOM-NAME-IS-*) refers to the types of atoms seen in the volumes of interest. The next group refers to chemical groups, and some biophysical parameters. The following group (RESIDUE-NAME-IS-*) refers to the types of amino acids seen. The final sets of features represent two different classifications of amino acids, and two different secondary structural classifications. The features are defined formally in (Bagley & Altman, 1995).

---

characteristics of the active site: the solvent accessibility is high, there are Asp and Ser residues nearby, and one or more 3-10 helices run through the site. In addition, there are Cys residues found in abundance nearby. These residues have not been reported as important to the catalytic activity of the site; instead, they are part of a nearby disulfide bridge joining two polypeptide chains. In this case, we can easily interpret these results in light of the existing literature on serine proteases. However, we would generally like a more detailed view of the spatial arrangement of these properties. Operating with only a radial distribution, it is unclear whether the significant property/volume pairs indicate the true spreading of the property over a shell volume or the existence of a spatially compact feature a given distance from the site center. The distributions for the oriented site data show how this problem can be rectified. The clusters of neighboring "blocks" highlight the spatial extent of significant properties, with only a small blurring of location due to errors in superposition and the discrete nature of the grid. Table 2A lists the clusters for selected properties; some of these are overlaid on a display of the active site in Figure 3.

In analyzing the significant features determined by the block analysis, we can use the results of the radial analysis as a guide. For example, radial collection finds, among other things, shells of solvent accessibility, Asp, Ser, and Cys residues, and 3-10 helices, all above the level found in the control group. Block collection reveals that the abundance of Asp in shells at 7-9Å actually represent two separate concentrations of Asp atoms, one in the catalytic triad, and another nearby, outside of the catalytic triad. The noncatalytic Asp has been shown to be involved in

Property	Local Residues in 4PTP	P-Levels <
A		
Solvent-Accessibility	Gln192, Gly193, Ser214, Trp215, Phe41, Ala56, His57, Tyr94	0.01-0.001
Residue-name-is-Asp	Asp102, Asp194	0.001
Residue-name-is-Cys	Cys42, Cys58	0.001-0.002
Residue-name-is-Ser	Ser195, Ser214	0.001-0.002
2ndry-structure-is-Helix	Ala55, Ala56, His57, Cys58, Tyr59	0.001-0.002
B		
Residue-name-is-Ala	Ala55	0.001
Negative-Charge	*more abundant in nonsites, no residues in volume in 4PTP	0.001
Atom-name-is-N	Ser214, Trp215	0.001
Charge	*more abundant in nonsites, no residues in volume in 4PTP	0.001
Atom-name-is-O	Ser54	0.001
Atom-name-is-S	Cys42, Cys58	0.001
Hydroxyl	Ser54, Cys42	0.001
2ndry-structure1-is-3-Helix	Ala56, Ala55, His57,	0.001
C		
Atom-name-is-N	Gln192, Gly193, Asp194, Ser195	0.01
Peptide	Asp194, Ser195	0.01
Solvent-Accessibility	Gln192, Gly193,	0.01
Residue-class-is-polar	Gln192, Gly193	0.005

Table 2. A selection of results for the oriented distributions at the serine protease active site. Each line of the table contains the data for one property cluster; each cluster is formed from neighboring blocks (cubic volumes). The first column, labeled **Property**, lists the name of the property found to be significant, **Local Residues in 4PTP** lists residues in the protein 4PTP (a typical member of the set of sites) that are contained in the volume in which the listed property is found to be significantly present (or absent). The **P-levels** are the range of significance levels for property/volume pairs. Part **(A)** shows the properties recognized in the radial analysis of the serine protease active site, and found to be significant as well in the oriented analysis of blocks. The key features of the site (conserved Ser, Asp, Cys, helical structure and solvent accessibility) are localized to specific regions, as indicated by the local residues. Part **(B)** lists the most significant properties found by the program, and an indication of the local residues in the associated volumes. In two cases, marked with an *, the property is found to be more abundant in the nonsites in a volume which contains no residues. Finally, part **(C)** shows the key features found in the region of the oxyanion hole, which provides two nitrogen atoms to bind the oxyanion intermediate of the substrate. Not only are the nitrogens recognized as conserved, but the polar, solvent accessible nature of the hole and the fact that it is formed by peptide nitrogens (and not sidechain nitrogens).

substrate binding and stabilization (Perona & Craik, 1995). A similar splitting is also found for the Ser residues. The Cys residues, found in shells 4-6Å and 7-9Å are resolved using block collection into a single region off the site center (shown in Figure 3). The 3-10 helices extend from the center of catalytic triad across the active site (also shown in Figure 3).

We turn now to the analysis of the oriented data in greater detail. Because the serine protease molecules from which the sites are drawn have a high degree of structural homology, we find many features that are significant—even with the noncatalytic histidine environments as a control. In order to focus attention on critical features, we employ three strategies. The first looks at properties of interest based on the results of the radial collectors, as described in the preceding paragraph. The second strategy uses level of significance. Table 2B shows the properties with the highest significance level in our calculation. In addition to a conserved set of alanines (the highest ranking finding), there is a relative lack of negative charge in one region around the catalytic site. There is also an excess of hydroxyl moieties scattered around the site (shown in Figure 3). The chief problem with looking at features that are highly ranked is that the differences in rank may be small, and the biological significance of the features may be low.

The final strategy for focusing attention is based on looking at volumes of interest. One spatially localized feature of the serine proteases is the oxyanion hole: two nitrogen atoms in the enzyme that hydrogen bond to an oxygen atom in the substrate intermediate. In chymotrypsin, the nitrogen atoms are contributed by the amide groups in the backbones of Gly 193 and Ser 195 (Perona and Craik, 1995). The system reports a significant abundance of nitrogen atoms (property ATOM-NAME-IS-N) in the volume near the backbone of Gly 193 and Ser 195. It also finds an abundance of polar residues, solvent-accessibility, charge, and peptide units in these volumes—all consistent with the structure and function of the oxyanion hole (Table 2C).

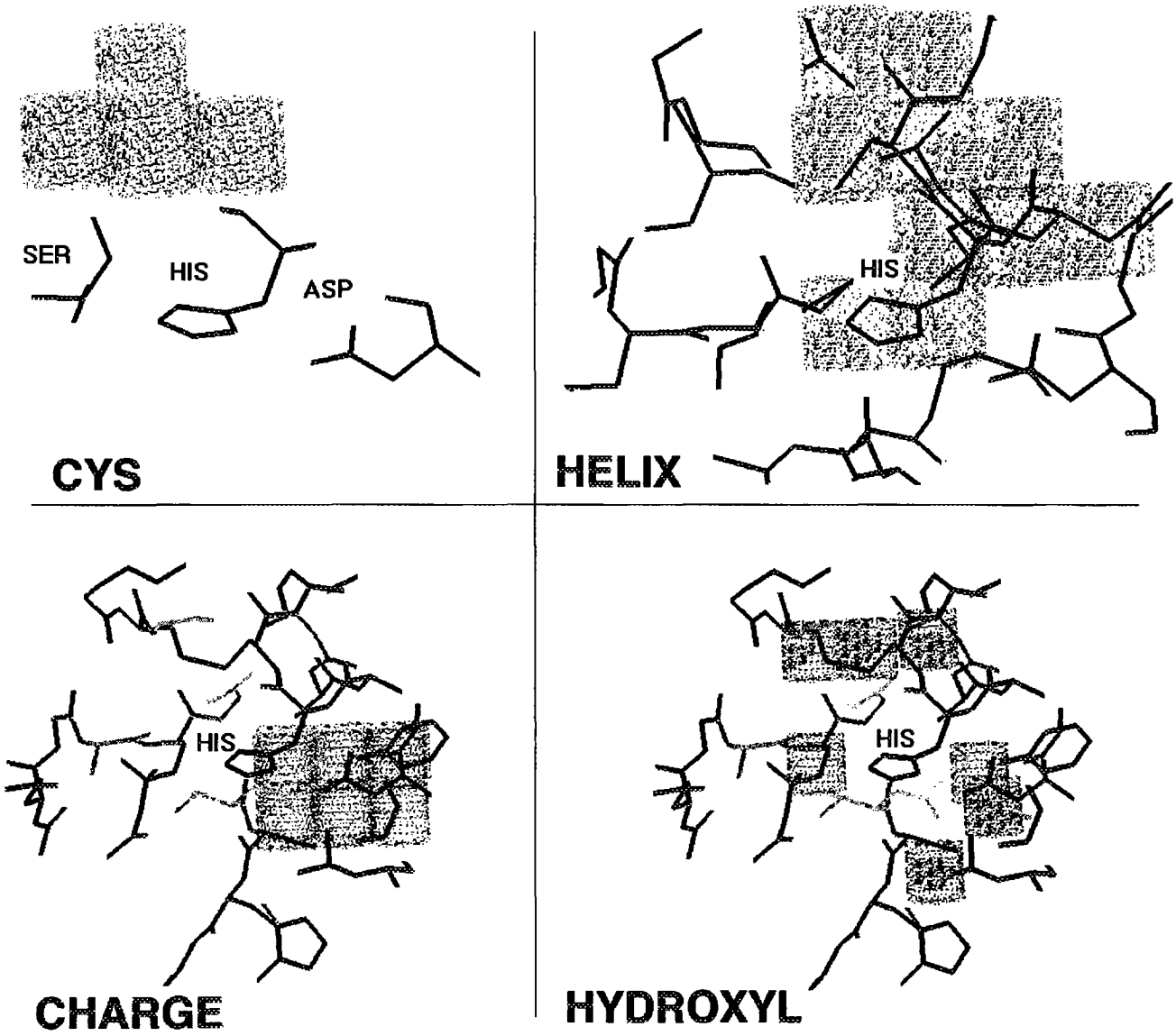

Figure 3. The spatial distributions of four properties are shown using dot clouds to identify the location of the significant volumes. (CYS) The three residues of the catalytic triad are isolated and a cloud of points is drawn in the volumes where an excess of cysteine residues are seen in sites. These correspond to a conserved disulfide bond in this set of structures. (HELIX) The atomic structure of the site region is shown with a cloud of points drawn in volumes for which there is an excess of helical secondary structure. The catalytic HIS is part of a helix. (CHARGE) The active site region is shown with volumes for which there is an excess of negative charge (around the ASP of the catalytic triad). (HYDROXYL) The active site is shown with volumes for which there is an excess of hydroxyl (-OH) moieties. These are dispersed throughout the site, and are critical for substrate binding.

Because all the significant features are reported with respect to the noncatalytic His controls, none of the properties reported can be interpreted as being part of the normal environment surrounding a histidine, since we have explicitly controlled for this environmental effect. Other nonsites could be chosen to accentuate the charge characteristics of the site, or its cavity. These would yield further information about the distinguishing features of the catalytic sites. The choice of sites and nonsites (and how they are superimposed) is critical in determining the kinds of properties that are reported. We anticipate that our program will be used in an iterative fashion, while the control nonsites are refined and modified to provide different perspectives on the sites.

The radial distributions are concise summaries of the key features, and the oriented distributions offer greater spatial resolution. The greater volumes gathered by radial collection of property values provides stronger support for the calculation of statistical significance. For example, a shell between radius 1 Å and 2 Å has a total volume of 29 $Å^3$, while the cubic volumes have a side dimension of 2.4 Å yielding a total volume of 13.8 $Å^3$. On the other hand, the oriented distribution provides better localization, and a greater number of significant findings. The balance between these two representations can be manipulated to provide a manageable set of reported features.

We note that there are several ways in which the program can be controlled. First, the explicit choice of the control group determines the background against which the protein sites are viewed. Second, the radial and oriented collections provide a convenient balance between detail and spatial focus. Third, the significant results can be pruned along several dimensions to reduce the volume of data while emphasizing salience: significance level, spatial location, spatial extent and property type. The three-dimensional maps of features that are produced by the method, as shown in Figure 3, may be useful as a basis for overlapping three-dimensional sites which have low identity at the atom or residue level, but which form similar three-dimensional biochemical environments.

Conclusion

This paper has presented a method for characterizing protein sites using biochemical properties tested for statistical significance against an explicit control group of nonsites. We have introduced a procedure for forming the distributions using oriented data, and compared it to the use of radial (unoriented) distributions. Our method has several advantages. The method analyzes the property distributions within a reasonable statistical framework, while relaxing some assumptions that may have limited previous approaches: the control group distributions are not spatially uniform, the choice of controls determines which properties are reported as significant, and can be used to remove spurious detail, and the property distributions need not be Gaussian. We have shown that an oriented analysis of sites in a common coordinate system, along with heuristics for data presentation provide a view of the sites which is spatially precise. Finally, we have analyzed the active site of six serine protease molecules, and shown how key conserved features of the microenvironment around the catalytic triad can be uncovered after aligning key atoms from the catalytic triad. In addition to finding the well known geometry of the catalytic residues, the system recognizes the key features of the oxyanion hole (an important feature for binding and stabilization), a conserved (non-catalytic) Asp, and several conserved structural features.

Acknowledgments

RBA is a Culpeper Medical Scholar, and this work is supported by the Culpeper Foundation and NIH LM-05652. Computing environment provided by the CAMIS resource under NIH LM-05305.

Appendix

Pseudocode summary of algorithm:

Input: *a set of sites (positive examples), a set of nonsites (negative examples), set of properties of interest.*

```
For each property,
 1. Create a grid to hold the property.
 2. For each site or nonsite instance:
   2a.  Clear out the grid.
   2b.  For each atom in the site/nonsite
        instance, enter the value of the
        property computed for that atom at its
        location in the grid.
   2c.  Over each collection volume (shell or
        block), sum all the values within the
        volume.  Save these values indexed by
        property and collection volume.
   3.  Reorder the data to produce one site
       distribution and one nonsite distribution
       for each property/volume.
   4.  Repeat steps 2 and 3 for the nonsites.
For each property/volume, compare the site and
nonsite distributions, and report those with
statistical significance exceeding threshold.
```

Output: *a list of property/volume pairs that show significant differences between sites and nonsites, the level of this significance and direction (whether the site values are greater than or less than the nonsite values).*

References

Bagley, SC and Altman, RB. 1995. Characterizing the microenvironment surrounding protein sites. *Protein Science* , 4, in press.

Bernstein, FC, Koetzle, TF, Williams, GJB, Meyer, EFJ, Brice, MD, Rodgers, JR, Kennard, O, Shimanouchi, T and Tasumi, M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology* 112: 535-542.

Bowie, JU, Luthy, R and Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253: 164-170.

Glantz, SA. 1987. *Primer of Biostatistics*. McGraw-Hill Book Company.

Goodford, PJ. 1985. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry* 28: 849-857.

Greer, J. 1990. Comparative modeling methods: application to the family of the mammalian serine proteases. *Proteins* 7: 317-334.

Kabsch, W and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22: 2577-2637.

Perona, JJ and Craik, CS. 1995. Structural basis of substrate specificity in the serine proteases. *Protein Science* 4: 337-360.

Singh, J and Thornton, JM. 1992. *Atlas of protein side-chain interactions.* IRL Press.

Warme, PK and Morgan, RS. 1978. A survey of atomic interactions in 21 proteins. *Journal of Molecular Biology* 118: 273-287.

Warshel, A, Naray-Szabo, G, Sussman, F and Hwang, J-K. 1989. How do serine proteases really work? *Biochemistry* 28: 3629-3637.

Zhou, GW, Guo, J, Huang, W, Fletterick, RJ and Scanlan, TS. 1994. Crystal structure of a catalytic antibody with a serine protease active site. *Science* 265: 1059-1064.

Zvelebil, MJJM and Sternberg, MJE. 1988. Analysis and prediction of the location of catalytic residues in enzymes. *Protein Engineering* 2: 127-138.