*Protein Science* (1995), *4*:622–635. Cambridge University Press. Printed in the USA.


# Characterizing the microenvironment surrounding protein sites

STEVEN C. BAGLEY AND RUSS B. ALTMAN

Section on Medical Informatics, Stanford University School of Medicine, MSOB X-215, Stanford, California 94305-5479

(RECEIVED November 4, 1994; ACCEPTED January 23, 1995)

**Abstract**

Sites are microenvironments within a biomolecular structure, distinguished by their structural or functional role. A site can be defined by a three-dimensional location and a local neighborhood around this location in which the structure or function exists. We have developed a computer system to facilitate structural analysis (both qualitative and quantitative) of biomolecular sites. Our system automatically examines the spatial distributions of biophysical and biochemical properties, and reports those regions within a site where the distribution of these properties differs significantly from control nonsites. The properties range from simple atom-based characteristics such as charge to polypeptide-based characteristics such as type of secondary structure. Our analysis of sites uses nonsites as controls, providing a baseline for the quantitative assessment of the significance of the features that are uncovered. In this paper, we use radial distributions of properties to study three well-known sites (the binding sites for calcium, the milieu of disulfide bridges, and the serine protease active site). We demonstrate that the system automatically finds many of the previously described features of these sites and augments these features with some new details. In some cases, we cannot confirm the statistical significance of previously reported features. Our results demonstrate that analysis of protein structure is sensitive to assumptions about background distributions, and that these distributions should be considered explicitly during structural analyses.

**Keywords:** biophysical properties; calcium binding; computational biology; disulfide bridges; microenvironment; protein structure analysis; serine proteases; software

Central to molecular biology is the determination of macromolecular structure and the analysis of how structural elements produce an observed function. The principles by which structure relates to function have been elucidated in a piecemeal fashion, from work on single structures or small classes of structures. Computational assistance has come primarily in the form of graphical methods for scientific visualization and from special purpose programs for analyzing individual biophysical properties (such as solvent accessibility or electrostatic fields). Unfortunately, studying structures individually entails a risk of missing important relationships that would be revealed by pooling relevant data. The expected surfeit of protein structures provides an opportunity to develop tools for automatically examining biological structures and producing useful representations of the key biophysical and biochemical features. The utility of a general purpose system for producing these representations would extend from medical/pharmaceutical applications (model-based drug design, comparing pharmacological activities) to industrial applications (understanding structural stability, protein engineering).

In this paper we describe a computational tool for analyzing protein sites—microenvironments within a structure distinguished by their structural or functional roles. We define a site as a region within a macromolecule with a central location and a surrounding neighborhood. In principal, a site could include the entire molecule, but we focus on sites that involve proper subsets of the molecule using a neighborhood with a 10-Å radius. Sites can be significant because of their structural role (for example, the site where a disulfide bond forms), their functional role (the active site of a serine protease) or both (the site of calcium binding). The most basic representation of a site is the set of atoms within it, along with their three-dimensional coordinates. We have created a system that augments this representation with the spatial distribution of user-defined properties. These properties include labels designating the types of atoms, chemical groups, amino acids, and secondary structures. They also include simple biophysical characteristics such as charge, polarity, mobility, and solvent accessibility.

The distribution of a property is computed by dividing the total volume of a site into subvolumes and computing the prev-

Reprint requests to: Russ B. Altman, Section on Medical Informatics, Stanford University School of Medicine, MSOB X-215, Stanford, California 94305-5479; e-mail: altman@camis.stanford.edu.

alence of the property within each of these subvolumes. Such distributions can be computed for sites, as well as for other microenvironments that are taken as nonsites. We have built the

interactions (Chakrabarti, 1990a, 1990b, 1993, 1994), patterns of hydration (Roe & Teeter, 1993), protein side-chain interactions (Singh & Thornton [1992] and others). For each of these

each of the three sites studied, we present an analysis of the features that are reported as significant by the program, along with a discussion of similar features reported in the literature. The Electronic Appendix contains the detailed significance measures, and a ranking of the significance of all findings for each of the three sites studied.

### *$Ca^{2+}$ binding sites*

Our results for the calcium binding sites are displayed in Figure 1 and Kinemage 1. The key findings can be summarized upon inspection: there is a statistically significant excess of negatively charged, acidic, oxygen-rich, mostly Asp and Glu moieties at radii 2–7 Å. Conversely, there is a relative paucity of hydrophobic (particularly Leu, Val, and Tyr), nonpolar moieties. These findings are consistent with the general understanding of cation binding sites, and particularly, the studies that have noted that metal sites in proteins are commonly surrounded by an inner shell of hydrophilic ligands and outer shell of carbon-containing groups (Yamashita et al., 1990; Nayal & Di Cera, 1994). The program produced other significant findings:

1. The property any-atom shows a sparsely occupied shell at 1–2 Å (indicating empty space) and relatively concentrated shells at 2–5 Å (indicating tightly packed atoms). The shells 0–3 Å are deficient in carbons, 2–3 Å is deficient in nitrogens, and 2–3,

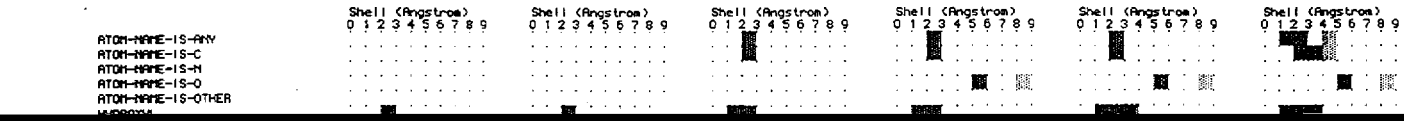
4–5, and 6–7 Å have a greater than expected concentration of oxygen. This distribution is as expected given the spacing of the coordinating oxygen shells and their van der Waals radii. The results of Yamashita et al. (1990) and Nayal and Di Cera (1994) suggest that the oxygen shell around a calcium ion is surrounded by a larger shell of nitrogen. There is a nitrogen shell at 4–5 significant at $P < 0.02$, just below the threshold for the other results reported here.

2. The amide group is underrepresented in shells 2–3 and overrepresented in shells 4–5. The oxygen-supplying carbonyl groups are strongly represented in the shell 2–3, 4–5, and 6–7, similar to the oxygen distribution seen at the atomic level.

*Cys bonding sites*

Our property grid results for the Cys disulfide environment versus Cys nonbonding environments are displayed in Figure 2 and Kinemage 2. The key features of the neighborhood surrounding a cysteine that participates in a disulfide bond is the occurrence of a neighboring cysteine at 2–6 Å, and this is a trivial consequence of our experimental setup. The other key features of the disulfide environment are an abundance of Tyr and a relative paucity of His and Ile. Muskal et al. (1990) noted the abundance of Tyr and attributed it to its hybrid hydrophobic/polar character. There is a relative lack of helical residues in the neighborhood and an increase in polar, especially hydroxyl

## Disulfide Bridge



Serine Protease Active Site

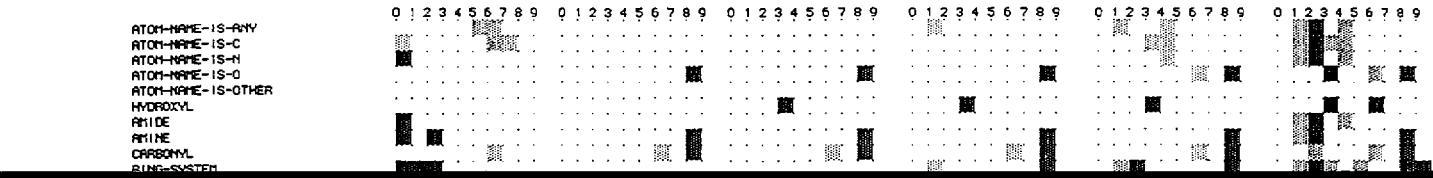


**Table 1.** *Summary of sensitivity analyses*[a]

| | Became insignificant | | Reversed significance | | Became significant | | No change | |
|---|---|---|---|---|---|---|---|---|
| | *N* | % | *N* | % | *N* | % | *N* | % |
| Small grid | 10 | 2.3 | 0 | 0.0 | 14 | 3.2 | 410 | 94.5 |
| Big grid | 11 | 2.5 | 0 | 0.0 | 16 | 3.7 | 407 | 93.8 |
| Big VDW | 11 | 2.5 | 0 | 0.0 | 13 | 3.1 | 410 | 94.5 |
| Small VDW | 9 | 2.1 | 2 | 0.5 | 14 | 3.3 | 409 | 94.2 |
| Smaller sample | 19 | 4.4 | 0 | 0.0 | 7 | 1.7 | 408 | 94.0 |
| Resample | 26 | 6.0 | 0 | 0.0 | 12 | 2.9 | 396 | 91.2 |
| Change control | 6 | 1.0 | 0 | 0.0 | 64 | 10.4 | 550 | 88.7 |

[a] For each experiment listed in a row, the changes to the number of property/volume pairs is shown. The reference for all comparisons is the results with a significance threshold of $P < 0.01$. For each new property plot (such as those shown in Figs. 1, 2, 3), we compared the individual boxes within the array with the reference plot. We counted which boxes became insignificant in the new experiment, reversed significance, became significant, or had no change. The first six rows correspond to columns two through seven in Figure 1. The last row (Change control) refers to the experiment in changing control nonsites for serine proteases and corresponds to the first column of Figure 3.

## Discussion

characterized by residues that arise from coils and bends, but

The most significant biological results of our experiments are contained in the plots of significance shown in Figures 1, 2, and 3. During the evolution of a site, the selective pressure is on the ability of the molecule to create an effective milieu (for the desired structure or function). The amino acids provide a basis set of chemical groups that can contribute certain characteristics to the site, but many characteristics can be realized in multiple ways. By temporarily abandoning a view of sites as groups of amino acids, and instead concentrating on the chemical milieu in important locations, we may gain insight into the critical factors that define the site. In essence, the plots shown in Figures 1

that arise from helical elements. Nevertheless, these observations are clear and consistent in our experiments. Although individual exceptions to these aggregate observations can be found, the weight of evidence toward certain types of structural components may be useful for engineering applications as well as for understanding functional roles.

There are several important issues related to the assessment and evaluation of our results and the method in general. Central to our method is the explicit use of a control group as the baseline for the statistical testing. The prominence of the control group in the calculation suggests that care should be taken

small to medium-sized globular proteins, with an overrepresentation of enzymes. Thus, if we used every example of each site, we would have a somewhat biased sample. Instead, we did not use every available site, but an arbitrary sampling of sites from among the alternatives. Strictly speaking, therefore, our results only apply to the sites listed in Table 2. However, as detailed in the Results section, most of the findings in these two systems

**Table 2.** *IDs and sample sizes for each of the proteins analyzed*

| PDB ID[a] | N sites[b] | N nonsites[c] |
|---|---|---|
| **Calcium binding site (site = calcium binding, nonsite = arbitrary)** | | |

As currently implemented, we cannot detect multiple subpopulations within the site or nonsite samples. If these occur, they could be detected by searching for correlations within the feature plots that show dependence of the contents of one property/volume pair on the occurrence of another property/volume pair. The implementation reported here simply reports on marginal distributions of properties and tests for significance of each property/volume pair independently of all other pairs.

Our technique has several features that may make it attractive for exploratory (or confirmatory) analysis of sites within macromolecules. First, it is general purpose. Any environment that is represented as a set of atomic positions can be studied without modifying any of the code. In this paper, we studied three totally different types of sites. Second, the system is modular. The set of properties is easily extended for special purpose analyses. The property calculations are independent of each other, and new properties can be added by writing a small amount of program code. Third, the method analyzes the property distributions within a reasonable statistical framework, with straightforward algorithms. Yet, it relaxes some assumptions that may have limited previous approaches: the control group

the sites and nonsites differ with respect to the distribution of user-defined physical properties.

The algorithmic implementation has, conceptually, four components: (1) a three-dimensional grid for accumulating information about property distributions; (2) a set of property definitions that allow the value of a property at each grid cell to be evaluated; (3) a "collector" that groups grid cells together to form distributions (in the current implementation, the collectors combine grid boxes that are within a shell around the central point of the sites/nonsites[1]; and (4) a method for testing significance. Each will be described in turn.

### *Grid*

The central spatial representation in the computation is a three-dimensional grid that holds the properties of protein's atoms. The grid is cubic, with a unit-cell diagonal chosen to be the length of a carbon–oxygen single bond (giving an edge length of 0.826 Å), so that two nearby atoms rarely occupy the same cell. The axes of the grid are determined by the coordinate system specified in the PDB file. The value stored in each grid cell is determined by the properties. In the current implementation,

cell is taken to fall inside a volume if its center point lies inside the volume. Currently the only collection volume used is a shell of thickness 1 Å. (We have also experimented with spherical volumes and shell thickness of 2 Å as reported in the sensitivity analysis, and these produce similar results.) Each collector sums the values in consecutive shells of 1 Å thickness (out to a user-defined maximum radius), and returns a vector of summed property values, one value for each collection shell. The collection process is shown graphically in Figure 4.[2]

*Testing for significant differences*

The products of the collection stage are site and nonsite distributions. A site distribution for a given property and collection volume contains all the values that were collected for that property/volume pair across all the protein site instances (and thus contains as many values as there are instances). A nonsite distribution is formed analogously. The two distributions are compared for statistical significance. Because these values are not, in general, normally distributed, a nonparametric test (the Mann–Whitney rank-sum test [Glantz, 1987]) is used to compare the distributions to see if the null hypothesis (that the two distributions are the same) can be rejected. All property/volume pairs producing results significant to a user-defined level are displayed in a two-dimensional plot (such as in Figures 1, 2, and 3). The significance level for these experiments was $P < 0.01$. Note that although the rank-sum test is invoked many times, each site and nonsite distribution is tested only once, for the property/volume pair from which it was formed and independently from all other possible pairs. The significance level therefore applies to each of those individual tests, not to a global hypothesis about the site microenvironment (no such hypothesis is formed by the system).

The program is written in generic Common Lisp and currently runs on two platforms, Macintosh Common Lisp on the Apple Macintosh, and Hewlett-Packard/Lucid Common Lisp on the Hewlett-Packard 720 series workstation. Those interested in the program code should contact the authors.

The algorithm can be summarized as follows:

INPUT: Set of sites (positive examples), set of nonsites (negative, control examples), set of properties of interest

For each property,

1. Create a grid for site properties
2. For each site,
   - 2.1. Center site on grid; clear grid
   - 2.2. Add value of property for site into grid
   - 2.3. Collect all values within volumes of interest, to produce a list of volume/value pairs giving the site distribution
3. Create a grid for nonsite properties
4. For each nonsite,
   - 4.1. Center nonsite on grid; clear grid
   - 4.2. Add value of property for nonsite into grid
   - 4.3. Collect all values within volumes of interest, to produce a list of volume/value pairs giving the nonsite distribution
5. Compare site distribution with nonsite distribution, and

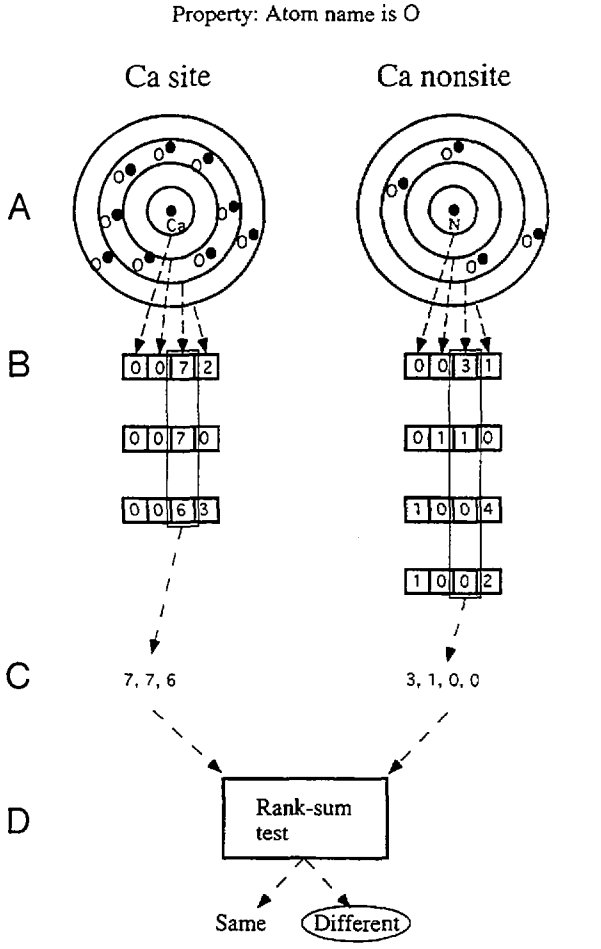



**Fig. 4.** Summary of procedure used to detect significant features. A representative calcium site and nonsite are shown, in the context of the property "atom name is oxygen." This figure illustrates how the system would conclude that the third shell has significantly more oxygens in calcium sites than in nonsites. **A:** Shells are formed around each site or nonsite, and values of the property of interest within the grid cells lying in each shell are summed. **B:** Sums are recorded as a vector, one sum for each shell. **C:** Values for a property/volume pair (in this example, property = oxygen, volume = shell number 3) are collected for all sites to form the site distribution, and analogously over nonsites for the nonsite distribution. **D:** Site and nonsite distributions are compared using the nonparametric Mann–Whitney rank sum test.

[2] In order to analyze sites in a manner that is sensitive to orientation, one would define collectors that did not perform radial summation

*Application to $Ca^{2+}$ binding sites*

Calcium ($Ca^{2+}$) is a metal ion commonly bound in proteins. The method was applied to determine which properties correlated with the presence of a calcium binding site. The calcium site was located at the center of the $Ca^{2+}$ ion, to a radius of 7 Å. For a typical binding site, see Figure 5A. The nonsites were chosen randomly from the same proteins from which the sites were selected, with 20 nonsites per protein. The proteins used and the number of sites and nonsites for each protein are shown in Table 2. Proteins were chosen from lists of commonly studied calcium binding proteins.

*Application to disulfide bond sites*

The sulfur atom in a cysteine residue often forms a covalent bond with a sulfur atom in a neighboring cysteine, forming a disulfide bridge. To find the properties correlated with the bonding state of the cysteine, the method was applied, taking the sulfur atom in each cysteine residue as the site's center, out to a radius of 10 Å. For cysteines forming a bridge, this will include the other cysteine residue. A typical site is shown in Figure 5B. The control nonsites were chosen to be cysteines not participating in a disulfide bridge (from proteins containing disulfide bridges, as well as some that do not). The proteins used and the number of sites and nonsites for each protein are listed in Table 2. The proteins were chosen at random from the PDB.

*Application to serine protease active sites*

Central to the proteolytic activity of serine proteases is presence of a catalytic triad, composed of the side chains from Asp, His, and Ser in a particular three-dimensional organization. The active site does not exhibit radial (spherical) symmetry. A typical active site is shown in Figure 5C. The property search was applied to these sites, using the NE2 atom of the His as the center, to a radius of 10 Å. The control nonsites were His residues not in the active site. The proteins used (a selection from the family of serine proteases) and the number of sites and nonsites for each protein are listed in Table 2.

*Sensitivity analysis*

We tested the sensitivity of our results to the choices and assumptions in our method. For each sensitivity test, we changed a parameter (as described below) and then examined the effect on the output representation for changes. We considered four possibilities: a difference between site and control nonsite may have no change, become significant, become insignificant, or reverse significance. (A reversal of significance is the most worrisome situation, because it implies that the parameter is very sensitive to the decision being tested.)

*Grid spacing*

The original grid spacing (0.826 Å) was chosen so that the grid cell diagonal corresponded to the length of a C–O bond, which had the effect of producing very few "collisions," when two atoms both fell into the same cell. As a test, the grid spacing was adjusted upward and downward in turn by 20% (up to 1.00 Å, and down to 0.66 Å), and then all the properties were recomputed for the calcium binding site proteins.

*Shell thickness*

The thickness of the collection shells was originally set at 1 Å. The calcium binding site proteins were rerun using a shell thickness of 2 Å.

*van der Waals radii*

Because in reality each atom is not a sphere of fixed radius, we scaled the van der Waals radii used in the property calculations (taken from the standard Richards sets [Richards, 1974], with augmentation from the literature) upward and downward by 20%, followed by a recalculation of all the properties for the calcium binding site proteins.

*Choice of nonsite controls*

To highlight the significant effect of how the control group (the nonsites) is chosen, a modification of the serine protease experiment was conducted, using randomly chosen atoms as nonsite centers instead of the NE2 atom in His residues not in the active site.

*Size of nonsite sample*

In order to gauge the effects of the sample size, especially on the nonsite group (which we typically have more control over), we compared the results for the calcium binding site run with the number of nonsites reduced by 50% (from 20 to 10). We then ran the analysis again, and compared the output.

*Effect of random sampling*

In order to further gauge the effects of sampling for nonsites, we reran the analysis of calcium binding sites with a different random sample of the same number of nonsites ($n = 20$).

*Statistical significance cutoff*

Finally, in order to test the sensitivity of our method to the definition of significance, we varied the significance level ($P$ value). We chose to conduct this experiment on the disulfide and serine protease environments because the results at standard significance level did not include a number of previously described features (as detailed in the Results).

**Supplementary material in Electronic Appendix**

The Electronic Appendix (SUPLEMNT directory, Bagley.SUP subdirectory) contains quantitative presentations of the property/volume plots at the standard conditions ($P < 0.01$ significance threshold), for the Ca binding site (first experiment in Fig. 1, file Bagley.ca), disulfide bonding environment (second experiment in Fig. 2, file Bagley.cys), and serine protease active site (second experiment in Fig. 3, file Bagley.his). Each entry contains the significance threshold (the $P$ value), whose sign indicates if the mean value in sites is greater (+) or less than (−) the control group. The rank of the cell is given in parentheses, with (1) being the most significant. The ranks are calculated with full precision of significance values, to break ties.

The Electronic Appendix also contains kinemages of a calcium binding site, a disulfide bridge, and a serine protease active site.

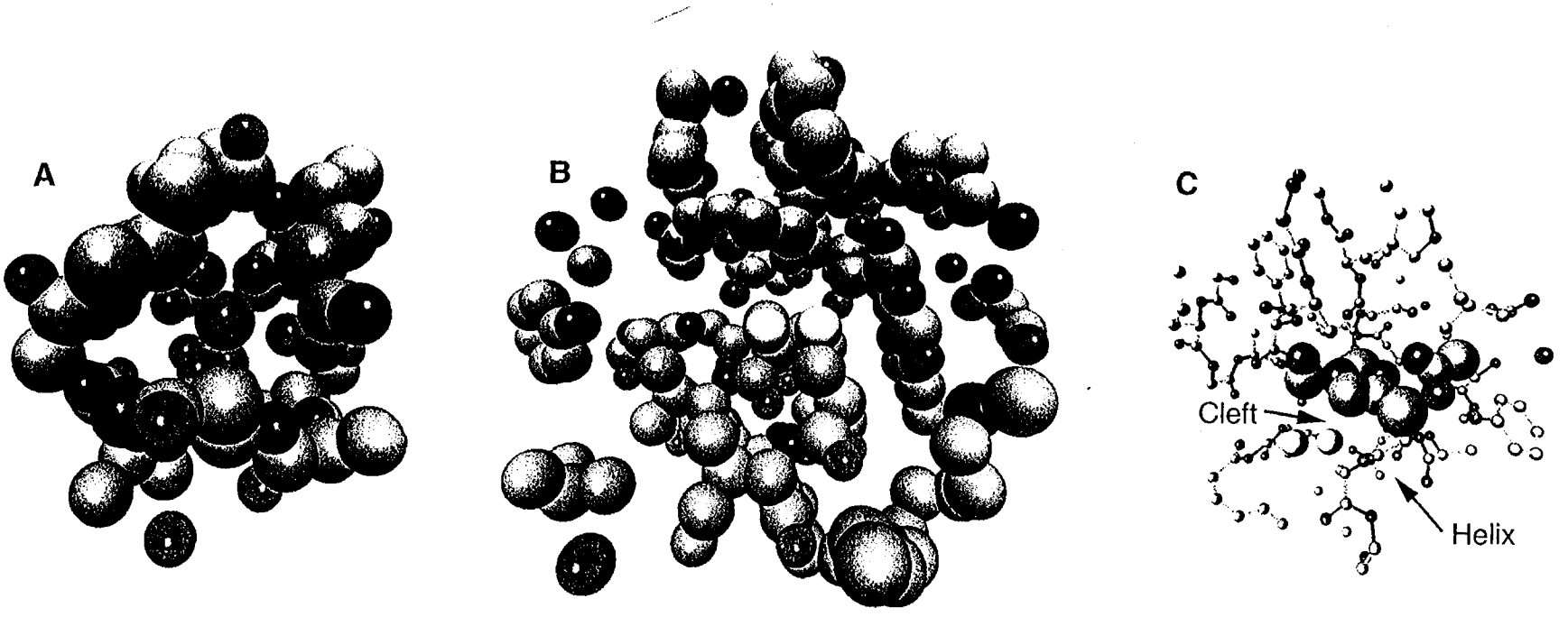



**Fig. 5.** **A:** $Ca^{2+}$ binding site of $\beta$-trypsin (4PTP) is shown. The van der Waals radii have been scaled by 0.5 to make the neighborhood more visible. Carbon atoms are light blue, oxygen atoms are red, nitrogen atoms are blue, calcium is green. This site is typical of those used in the calculation of significant properties for calcium sites and demonstrates the difficulty in systematically determining which structural features are consistently present and significant over many such examples. A kinemage view of this site appears as Kinemage 1. **B:** Disulfide bridge from glutathione reductase (3GRS) is shown. Coloring scheme is as in Figure 5, with sulfur atoms drawn yellow, and phosphates (as well as ambiguous nitrogen/carbon atoms) drawn pink. The van der Waals radii have been scaled by 0.5 to make the neighborhood more visible. This site is one of the sites used in the calculation of significant properties for disulfide bridge sites. A portion of the planar flavin ring system that occurs close to the disulfide bridge is shown to the right of the sulfur. All atoms reported in the PDB file and within the radius of interest are used in these calculations. A kinemage view of this site appears as Kinemage 2. **C:** Active site from $\gamma$-chymotrypsin (1GCT), a serine protease, is shown. Only the atoms in the catalytic triad are shown full scale. The nearby cysteines are shown half size; all other atoms are greatly reduced in size. The binding cleft and $3_{10}$-helix that are seen in the proteases are labeled. A kinemage view of this site appears as Kinemage 3.

## Acknowledgments

R.B.A. is a Culpeper Foundation Medical scholar and is supported by NIH grant LM-05652. This work was supported by computing resources provided by the Stanford University CAMIS project, which is funded under grant number LM05305.

## References

Barlow DJ, Thornton JM. 1986. The distribution of charged groups in proteins. *Biopolymers 25*:1717–1733.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

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

Branden C, Tooze J. 1991. *Introduction to protein structure*. New York: Garland Publishing, Inc.

Chakrabarti P. 1990a. Interaction of metal ions with carboxylic and carboxamide groups in protein structures. *Protein Eng 4*:49–56.

Chakrabarti P. 1990b. Geometry of interaction of metal ions with histidine residues in protein structures. *Protein Eng 4*:57–63.

Chakrabarti P. 1993. Anion binding sites in protein structures. *J Mol Biol 234*:463–482.

Chakrabarti P. 1994. Conformational analysis of carboxylate and carboxamide side-chains bound to cations. *J Mol Biol 239*:306–314.

Fiser A, Cserzo M, Tudos E, Simon I. 1992. Different sequence environments of cysteines and half cystines in proteins: Application to predict disulfide forming residues. *FEBS Lett 302*:117–120.

Glantz SA. 1987. *Primer of biostatistics*. New York: McGraw-Hill Book Company.

Goodford PJ. 1985. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem 28*:849–857.

Greer J. 1990. Comparative modeling methods: Application to the family of the mammalian serine proteases. *Proteins Struct Funct Genet 7*:317–334.

Jernigan R, Raghunathan G, Bahar I. 1994. Characterization of interactions and metal ion binding sites in proteins. *Curr Opin Struct Biol 4*:256–263.

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

Karlin S, Zuker M, Brocchieri L. 1994. Measuring residue associations in protein structures: Possible implications for protein folding. *J Mol Biol 239*:227–248.

Klingler TM, Brutlag DL. 1993. Detection of correlations in tRNA sequences with structural implications. *First International Conference on Intelligent Systems for Molecular Biology*. Menlo Park, California: AAAI Press.

Klingler TM, Brutlag DL. 1994. Discovering structural correlations in $\alpha$-helices. *Protein Sci 3*:1847–1857.

Korn AP, Burnett RM. 1991. Distribution and complementarity of hydropathy in multisubunit proteins. *Proteins Struct Funct Genet 9*:37–55.

Mitchell JBO, Nandi CL, McDonald IK, Thornton JM, Price SL. 1994. Amino/aromatic interactions in proteins: Is the evidence stacked against hydrogen bonding? *J Mol Biol 239*:315–331.

Muskal SM, Holbrook SR, Kim SH. 1990. Prediction of the disulfide-bonding state of cysteine in proteins. *Protein Eng 3*:667–672.

Nayal M, Di Cera E. 1994. Predicting $Ca^{2+}$-binding sites in proteins. *Proc Natl Acad Sci USA 91*:817–821.

Perutz M. 1992. *Protein structure: New approaches to disease and therapy*. New York: W.H. Freeman and Company.

Reid KSC, Lindley PF, Thornton JM. 1985. Sulphur–aromatic interactions in proteins. *FEBS Lett 190*:209–213.

Richards FM. 1974. The interpretation of protein structures: Total volume, group volume distributions and packing density. *J Mol Biol 82*:1–14.

Roe SM, Teeter MM. 1993. Patterns for prediction of hydration around polar residues in proteins. *J Mol Biol 229*:419–427.

Rooman MJ, Kocher JP, Wodak SJ. 1992. Extracting information on folding from the amino acid sequence: Accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. *Biochemistry 31*:10226–10238.

Sekharuda YC, Sundaralingam M. 1988. A structure–function relationship for the calcium affinities of regulatory proteins containing "EF-hand" pairs. *Protein Eng 2*:139–146.

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

Sippl MJ. 1990. Calculation of conformational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol 213*:859–883.

Walshaw J, Goodfellow JM. 1993. Distribution of solvent molecules around apolar side-chains in protein crystals. *J Mol Biol 231*:392–414.

Warme PK, Morgan RS. 1978. A survey of atomic interactions in 21 proteins. *J Mol Biol 118*:273–287.

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

Yamashita MM, Wesson L, Eisenman G, Eisenberg D. 1990. Where metal ions bind in proteins. *Proc Natl Acad Sci USA 87*:5648–5652.

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

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

## Appendix 1: List of microenvironment properties

This appendix contains the set of biophysical/biochemical properties currently used by the system. Properties marked NC are stored only in the cell containing the nucleus of the atom; properties marked EV are spread out over the electron (van der Waals) volume of the atom.

### *Atom-based properties*

*Atom types.* One of (ANY, CARBON, NITROGEN, OXYGEN, or OTHER). The atom name is entered in the grid at the location of the atom's nucleus. NC.

*Hydrophobicity.* All O and N are −1. Any C directly bonded to an O or an N is 0. All other C are 0. All metal ions (Ca, Cu, Fe, Zn, Mn, Mg) are −2. The S in Cys is −1. All other atoms are 0. EV.

*Charge.* The value is −1/3 for each of CG, OD1, and OD2 in Asp, −1/3 for each of CD, OE1, OE2 in Glu, +1 for NZ in Lys, +1/3 for each of CZ, NH1, NH2 in Arg, +2 for Ca, Cu, Fe, Mg, Mn, Zn, and −1 for Cl, and 0 for all other atoms. EV.

*Charge-with-His.* Similar to charge property, with the addition that His ND1 and His NE2 each are 0.5, and His AD1, His AD2, His AE1, and His AE2 are each 0.25. EV.

### *Chemical group-based properties*

*Hydroxyl.* The value is 1.0 for Ser OG, Thr OG1, or Tyr OH, and 0.5 for Cys SG. 0.0 otherwise. NC.

*Amide.* The value is 1.0 for Asn ND2, Gln NE2, and Pro N, and 0.5 for Arg NH1 and NH2, Asn AD1 and AD2, Gln AE1 and AE2, His ND1 and NE2, and 0.25 for His AD1, AD2, AE1, AE2. 0.0 otherwise. NC.

*Amine.* The value is 1.0 for Arg NE, Lys NZ, and Trp NE1, 0.5 for Arg NH1 and NH2, and His ND1 and NE2, and 0.25 for His AD1, AD2, AE1, AE2. 0.0 otherwise. NC.

*Carbonyl.* The value is 1.0 for backbone O, Asn OD1, and Gln OE1 and 0.5 for Asp OD1 and OD2, Asn AD1 and AD2, Gln AE1 and AE2 and Glu OE1 and OE2. 0.0 otherwise. NC.

*Ring-system.* The value is 1 if the atom is part of a ring system (in His, Phe, Trp, or Tyr). 0 otherwise. NC.

*Peptide.* The value is 1 if the atom is part of the polypeptide backbone. 0 otherwise. EV.

### *Residue-based properties*

*Residue types.* The standard 20 amino acids, or HOH or Other. NC.

*Hydrophobicity classification 1.* One of HYDROPHOBIC (Ala, Ile, Leu, Met, Phe, Pro, Val), CHARGED (Arg, Asp, Glu, Lys), POLAR (Asn, Cys, Gln, His, Ser, Thr, Tyr, Trp), or UNKNOWN (nonstandard residues). NC.

*Hydrophobicity classification 2.* One of NONPOLAR (Ala, Ile, Leu, Met, Phe, Pro, Trp, Val), POLAR (Asn, Cys, Gln, Gly, Ser, Thr, Tyr), ACIDIC (Asp, Glu), or BASIC (Arg, Lys, His), or UNKNOWN (nonstandard residue). NC.