All-Atom Small-Probe Contact Surface Analysis:

An information-rich description of molecular goodness-of-fit

by J. Michael Word

Department of Biochemistry
Duke University

Date: 1/21/00

Approved:

David C. Richardson, Supervisor

Jane S. Richardson, Supervisor

Michael Been

Jan Hermans

Tao-shih Hsieh

Terrence G. Oas

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biochemistry in the Graduate School of Duke University

2000

Abstract *(Biochemistry)*

All-Atom Small-Probe Contact Surface Analysis:

An information-rich description of molecular goodness-of-fit

by J. Michael Word

Department of Biochemistry
Duke University

Date: 1/21/00

Approved:

David C. Richardson, Supervisor

Jane S. Richardson, Supervisor

Michael Been

Jan Hermans

Tao-shih Hsieh

Terrence G. Oas

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biochemistry in the Graduate School of Duke University

2000

The principal difficulty in protein design is not stability but uniqueness: eliminating disorder from substantially non-native designed proteins has proven refractory. Native proteins, on the other hand, show both order and a remarkable tolerance for mutations. To analyze what native proteins are doing right and what designed proteins are doing wrong, we have developed a detailed representation of close van der Waals packing inside or between molecules: *the all-atom small-probe contact surface*. Its use requires the explicit inclusion of all hydrogen atoms. I wrote the computer programs PROBE (to measure and visualize contacts) and REDUCE (to add and optimize hydrogens) and used them to produce and study a database of 100 high-quality high-resolution protein structures. Flipped side-chain amide orientations (20% of cases) were corrected by REDUCE automatically.

All-atom contact analysis has proven to be a critical tool for structure validation, and contact dot displays have been integrated into popular crystallographic refitting software (O and XTALVIEW) to assist structure determination at all resolutions. An interactive link has been developed between PROBE and the 3D display program MAGE, to facilitate structural analysis. Other laboratories are starting to use these tools.

I designed a λ repressor mutant with a buried tryptophan using contact displays and I created the corresponding plasmid construct. The mutant λ protein is stable, folds cooperatively, and shows a measurable shift in fluorescence upon unfolding. A small SS beta-cross protein I designed for use as a framework for iterative re-

design, when synthesized, proved to be inappropriate for further design because scrambled disulfide forms could not be separated experimentally.

The laboratory produced an improved side-chain rotamer library from 240 high-resolution structures, using modes, with filters for *B*-factor and clashes. Analysis of rotamer data identified systematic misfitting artifacts and produced a rationale for several high or low frequency conformations. An automated procedure was developed to sample contact scores over a range of dihedral angles, producing maps of conformational constraints.

Initial studies have been performed to analyze packing variation in proteins using contours of average local contact density and atom density.

The software, structure databases, and rotamer library are available on the web.

Typographic Conventions Used in the Text

PROGRAMS are set in small-caps Helvetica while commands and the names of command scripts are shown in Helvetica and `file fragments` are in Courier. Our rotamer names are in boldface (i.e. **mttt**).

Acknowledgments

First, I must thank Cate Stewart for putting up with drastic changes in our schedules and our plans so that I could do this thing. A late night for me was usually a late night for you, too, and you endured far too many dinners alone to be counted. While we did manage to go on several big trips, it was the weekend camping trips, or the short getaways to the mountains or the beach, that you and I miss the most—especially this last year. I promise, we will make up for lost time. But, for now, thank you so much for your sincere confidence in me and your recognition that my graduate career could not last forever.

The work presented here was a team effort, involving many hands. Dave and Jane Richardson, being more hands-on than most, taught by example. Absolutely the most creative people I have ever met, they were also completely open to new ideas from others, no matter how embryonic, and encouraged the sort of "fooling-

around" during which so much of my work was conceived. I could not have had better mentors and hope to find ways to continue working closely with them.

While most of the computer programming was done by me, a notable exception is that Dave Richardson wrote and continues to enhance the *essential* graphics programs: MAGE and PREKIN. Certain of these enhancements were a direct result of our discussions; with regularity, I would insist on some complicated new functionality late one evening only to return the next day and find Dave demonstrating that function in a new version of his software.

Also, Dave and Mike Zalis collaborated on SP, the original implementation of small-probe contact surface analysis. Mike visited the lab a few years back and we were able to show him where we had taken this project.

Lab mates Kim Gernert, Hope Taylor, Thom LaBean and Simon "Stan" Lovell all made significant (positive) contributions to my graduate education. Kim introduced me to the original SP program. Her group at Emory went on to build the WEBPROBE interface. Hope provided continuity, much needed in our small group. She also spent a lot of time in an unsuccessful attempt to crystallize the $\lambda^{*}_{6\text{-}85}$(trp) mutant described in chapter 6; in the process, the lab learned the word "coacervate." Thom provided what amounted to "Molecular Biology for Dummies" (there's a book in there Thom!). And Stan—a great colleague, traveling companion, and musician, with an invaluable sense of humor—initiated most of the work on crys-

tallographic applications of small-probe contact surfaces and is the penultimate authority on side-chain rotamers.

The lysozyme mutant work started out in life as Bob Bateman's class project for the Macromolecular Structures course, during his short sabbatical in our laboratory. By necessity and also because he was really fun to be around, Bob and I collaborated from the outset. The result takes up a large part of Chapter 8.

Without Lizbeth Videau the lab might not have been able to keep track of its responsibilities (or references) enough to stay afloat. Her industry helped see to it that Dave and Jane had time to pay attention to my work. She often provided an extra hand or a positive attitude when I needed one.

Interactions with several other laboratories have greatly enhanced my graduate experience. Foremost among these has been my time working with Terry Oas and the great people in his group; thanks for going out of your way to welcome me. I have really enjoyed exchanging ideas about science and scientific programming with Homme Hellinga, Jan Hermans, and their students. It has been especially fun to talk about scientific programming (and books!) with Gene Wickham. Gene was one of the early adopters of my tools and some of his modeling work is discussed in Chapter 4. Thanks also to David King at UC Berkeley for synthesizing and working with me to analyze the SSframe peptide. And, finally, life as we know it would not have been the same without my regular cup of latté and chat with Meta Kuehn.

I have overlooked a lot of other folks who have made my stay here great. I'm sorry, but I there are so many of you that I will just say "I'll miss you."

I have been given opportunities to see the world through the eyes of science by being actively held aloft by many. By my family, certainly. The boost from teachers such as Peggy Hall and Dr. Robert Hargrove is something I still feel. Mercer University generously supported my undergraduate education. But above all, it is my astonishing good fortune to work for GlaxoWellcome, Inc. and to receive their scholarship during my time at Duke. I sincerely appreciate, and will try but can never fully repay, this recognition and assistance.

Contents

Figures

Tables

Abbreviations

ACN . . . . . Acetonitrile.

ANS . . . . . 1-Anilino-naphthalene-8-sulfonate.

ASA . . . . . Accessible Surface Area.

ATP . . . . . Adenosine Triphosphate.

B-factor . . . . . Crystallographic temperature factor.

CD . . . . . Circular Dichroism.

CPK . . . . . Corey–Pauling–Koltun (space filling molecular models).

DEE . . . . . Dead-End Elimination.

DMSO . . . . . Dimethyl Sulfoxide.

DTT . . . . . 1,4-Dithiothreitol.

FMOC . . . . . 9-Fluorenylmethoxycarbonyl.

GdmCl . . . . . Guanidinium Chloride.

GSH . . . . . Reduced Glutathione.

GSSG . . . . . Oxidized Glutathione.

HDV . . . . . Hepatitus Delta Virus.

HPLC . . . . . High Performance Liquid Chromatography.

LB . . . . . Luria Broth.

LCMS . . . . . Liquid Chromatography/Mass Spectroscopy.

MOPS . . . . . 4-Morpholinepropanesulfonic Acid.

NMR . . . . . Nuclear Magnetic Resonance spectroscopy.

PDB . . . . . Protein Data Bank (formerly at Brookhaven—see RCSB).

PDI . . . . . Protein Disulfide Isomerase.

RCSB . . . . . Research Consortium for Structural Biology (current home of the PDB).

ROP . . . . . Repressor of Primer protein.

RT . . . . . Retention Time.

SCSA . . . . . Solvent Contact Surface Area.

SIRAS . . . . . Single Isomorphous Replacement Anomalous Scattering.

UNC . . . . . University of North Carolina.

VDW . . . . . Van der Waals

CHAPTER 1 *Introduction*

The Stubborn Problem of Uniqueness

For nearly two decades, the Richardson laboratory has employed rational protein design, primarily *de novo* design—the development of novel polypeptides with backbone arrangements dissimilar to, and little or no sequence homology to, those observed in nature—as a means of examining our comprehension of the physical principles which relate amino acid sequence to protein secondary and tertiary structure (Erickson *et al.* 1986; Gernert *et al.* 1995; Hecht *et al.* 1990; Quinn *et al.* 1994; Richardson *et al.* 1987a; Richardson *et al.* 1987b; Richardson and Richardson 1987; Richardson and Richardson 1995; Unson *et al.* 1984). The native structure of most globular proteins, resulting from the collective influence of a number of forces and factors, is remarkably compact and reproducible. In order to approximate the work of Mother Nature, the designer of an un-natural protein must strive for a product with an interior packed, on average, as densely as crystals of small organic molecules (Richards 1977), which is precisely organized and yet flexible enough to

perform a specific enzymatic or structural function. The amino acid components the protein designer builds with vary widely in size, shape and physical property; in combination, some interact in stable configurations while others are mutually incompatible. Many useful principles which can guide the design process (e.g. hydrophobic patterning (Dill *et al.* 1995; Kamtekar *et al.* 1993; Sun *et al.* 1995), helix end capping (Presta and Rose 1988; Richardson and Richardson 1988a), discrete side chain conformations (Lovell *et al.* 2000; Ponder and Richards 1987)) have been identified and recently a number of very powerful new computer search algorithms (notably DEE; Desmet *et al.* 1992; Dahiyat and Mayo 1997a, 1997b) have sustained the optimism which has long characterized the protein design field.

Still, we remain a long way from being able to confidently design novel proteins to order. The only well-ordered true *de novo* design to date (Harbury *et al.* 1998) is a coiled-coil with a novel right-handed superhelical twist; a minimalist design which depends on simplification and cooperativity of the repeating structure. With the current incomplete state of knowledge, only a very limited form of rational redesign—adding, removing or altering elements—could be considered reliable for engineering use (Dahiyat and Mayo 1997a; Desjarlais and Handel 1995; Hellinga 1998; Struthers, Cheng, and Imperiali 1996). It's not that researchers have problems designing stable proteins; stability is relatively easy to control (Richardson *et al.* 1992). The hard part is not stability but order. While the concept of *negative design* has helped organize discourse about this issue (Hecht *et al.* 1990; Richardson *et al.* 1992), eliminating disorder from substantially non-native designed proteins has proven refractory.

Remarkably ordered arrangements in the interior of native protein molecules are demonstrated by high-resolution crystalline order in proteins and by the existence of specific through-space NMR couplings between sequentially distant atom pairs. Although we have become quite accustomed to seeing these well-ordered, well-packed arrangements in thousands of X-ray and NMR structures, the quite different, more-or-less molten nature of almost all protein *de novo* designs and randomized cores (Axe, Foster, and Fersht 1996; Betz, Raleigh, and DeGrado 1993; Choma *et al.* 1994; Fedorov *et al.* 1992; Fezoui, Weaver, and Osterhout 1994; Houbrechts *et al.* 1995; Kamtekar *et al.* 1993; Mutter *et al.* 1992; Quinn *et al.* 1994; Richardson *et al.* 1992; Smith *et al.* 1995) strongly implies that the ordered packing of natural proteins is important for function and relatively difficult to attain.

On the other hand, structural and functional tolerance of a substantial fraction of mutations in protein interiors (e.g. Dalal, Balasubramanian, and Regan 1997; Hurley, Baase, and Matthews 1992; Lim and Sauer 1989; Munson *et al.* 1994; Richards and Lim 1993; Shortle, Stites, and Meeker 1990) implies that side-chain packing is either not important or not difficult. Recent studies of barnase core mutants (Axe, Foster, and Fersht 1996) and of the heme-binding properties of randomized helix bundles (Rojas *et al.* 1997) show that low-level activity is compatible with a sizable fraction of conservatively randomized hydrophobic cores, in spite of the well-established sensitivity of detailed functional properties to single core mutations. Theoretical studies also have reached conclusions both for (e.g. Shakhnovich and Finkelstein 1989) and against (e.g. Behe, Lattman, and Rose 1991; Bromberg and Dill 1994) the importance of specifically complementary side-chain packing.

More recently, there have been two direct experimental tests that each seem very convincing but are on opposite sides of this controversy. Gassner *et al.* (1996) solved the crystal structure of a T4 lysozyme mutant with seven methionine substitutions in the hydrophobic core of its larger domain; although less stable than wild-type, it is clearly very well ordered and has 50% activity in spite of the extra side-chain flexibility and different shapes, implying that specific packing is not strongly critical. Dahiyat & Mayo (1997a) used an automated design procedure to redesign the core of the B1 domain of protein G, leaving the backbone fixed and varying the stringency of van der Waals packing (including hydrogen atoms); they produced one sequence at each of four levels of packing stringency and showed that the resulting proteins were well-ordered when designed between 90% and 105% of full van der Waals radii, molten if at < 85%, and unfolded if at > 105%, implying that packing is the dominant factor controlling order. Unfortunately, given the existence of conflicting evidence, neither of these studies can fully settle the question yet: in the T4 lysozyme work, six of the seven mutations were iso-volume Leu → Met in which only one methyl group shifts and the packing of the final Met side-chains is excellent (Chothia and Gerstein 1997); in the protein G work, the calculations did not allow backbone shift, the well-behaved redesign had only three conservative sequence changes from wild-type, and these calculations varying the percentage packing stringency are not the only possible way one could compare the set of sequences. It is important to resolve this basic conflict in how we perceive the nature of protein structure, folding, and evolution. In order to understand the principles involved in forming well-ordered, as opposed to merely stable, macromolecu-

lar structure, one prerequisite will certainly be a clear and detailed representation of local packing quality.

Similarly, ligand docking is vital to the drug design effort and improved methods of quantifying the steric fit of a ligand to a macromolecule may help increase the reliability of such assessments.

Discovery-oriented Science

Often science is advanced—even enabled—by purely technical developments, both great and small. One interesting aspect of this phenomenon is that much of the initial work with a new tool or technique is *discovery-oriented*: often described by researchers as similar to "looking through a new kind of microscope." The work presented here is very much of this type. While it was an outgrowth of my general interest in how protein atoms organize themselves and the Richardson laboratory's long-standing interest in protein design, the precipitating event was when I rewrote an old computer program for generating small-probe contact-dot surfaces (Word *et al.* 1999a), developed many years earlier in our laboratory but little used because it was inconvenient and limited. At the same time, I created a separate program to add hydrogen atoms to coordinate files where they were missing (Word *et al.* 1999b), freeing us from the *implicit hydrogen* approximation commonly employed in molecular modeling. With these more accessible and more capable new tools, we began to examine atomic contacts in a wide variety of structures, surveying this new and unfamiliar visual environment.

It often seemed as if we saw something unexpected everywhere we looked. All-atom contact analysis revealed both the relaxed, beautifully packed nature of proteins as well as scattered mistakes in even atomic-resolution structures. Steric clashes within the sugar-phosphate backbone were more common in non-B-form than in B-form nucleic acid structures. Models for *de novo* designed proteins tended to look unnatural—often dreadful. To better study atomic packing, we built a database of 100 high-resolution X-ray crystal structures from the Protein Data Bank, adding and optimizing all hydrogens and performing certain types of corrections (Word *et al.* 1999a). Along the way, we refined and extended our methods; for example, introducing a numerical measure of goodness-of-fit and automating the correction of flipped sidechain amide groups (Word *et al.* 1999b). We integrated contact analysis into crystallographic refitting programs (Lovell and Word, unpublished; McRee 1999; Richardson and Richardson in press) and used both good contacts and clashes to analyze side-chain rotamers (Lovell *et al.* 1999; Lovell *et al.* 2000). Our research suggests that all-atom steric constraints severely, and perhaps uniquely, limit physically realistic molecular models. In the chapters that follow, I will describe our methodology along with applications of this constraint principle to structure validation, structure determination, protein engineering and theoretical modeling.

CHAPTER 2

The Small-probe Contact Surface Algorithm

Packing and Goodness-of-fit

"Goodness-of-fit" for molecular interfaces, or complementarity of local packing, is surprisingly difficult to define. None of the methods available are suitable either for settling the importance-of-packing controversy, for visualizing or quantifying the steric component of ligand binding, or for redesigning a local region to better promote ordered structure. *Free energies*, although easily defined, are difficult to determine with sufficient accuracy and precision. The usual measure by *buried surface* area (Chothia 1974; Lee and Richards 1971) has been an enormously productive and useful concept; however, it is defined by a water-sized probe sphere, and so considers two atoms effectively touching when they are as much as 2.8 Å apart; it works excellently to measure the size of an interface already known to be well fitted, but cannot discriminate good versus bad packing. Standard energy calculations include both attractive and repulsive *van der Waals* terms and are effective at eliminating bad clashes if all explicit hydrogen atoms are used at full radius; however, it

is a cornerstone of the method that energies are added up across the entire system rather than examined locally, the van der Waals terms are treated purely pairwise, and contacts of polar hydrogen atoms are usually not considered. On the other hand, the measure of packing density using *Voronoi polyhedra* (Richards 1974; Richards 1977) can define the volume of an individual residue or even atom and has been used effectively to study density variation within proteins; however, it has no way of penalizing clashes and would give the same overall value for any rearrangement within a given shell of atoms. Finally, calculated from the portion of each residue's molecular surface that water cannot access, *occluded surface* (OS) packing values (Pattabiraman, Ward, and Fleming 1995) are sensitive to the local arrangement of atoms, are discriminating of bad packing, and are in a number of ways comparable to the results of our approach. However, the occluded surface combines both close and long-range contacts (again, out to 2.8 Å), is somewhat time-consuming to calculate, and is primarily designed to identify typical and deviant packing environments; it is not designed to highlight excluded volume effects.

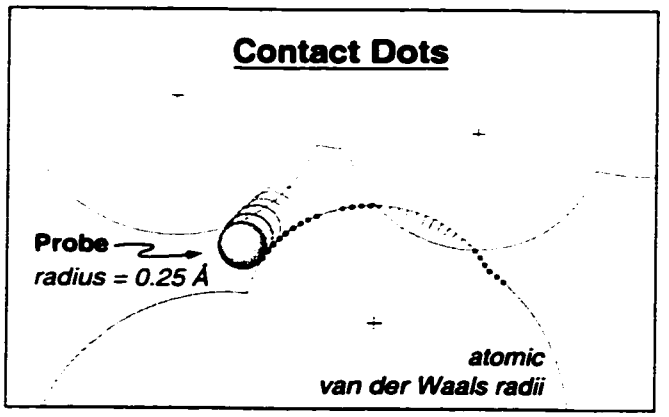
Below, *all-atom small-probe contact analysis* is described as a method for calculating and displaying the detailed atomic contacts inside or between molecules. It allows one to measure and to visualize directly the goodness-of-fit of packing interactions. A simpler form of this method, a program named SP developed in our laboratory nearly 15 years ago, has been used in analyzing various structural details (Richardson and Richardson 1987, 1988b, 1990; Richardson *et al.* 1992). However, the power and convenience of the new PROBE program, the speed of current graphics displays, and most especially the high accuracy of many recently determined

protein structures have now combined to create a tool that can produce genuinely new insights.

Definition of Contact Surfaces

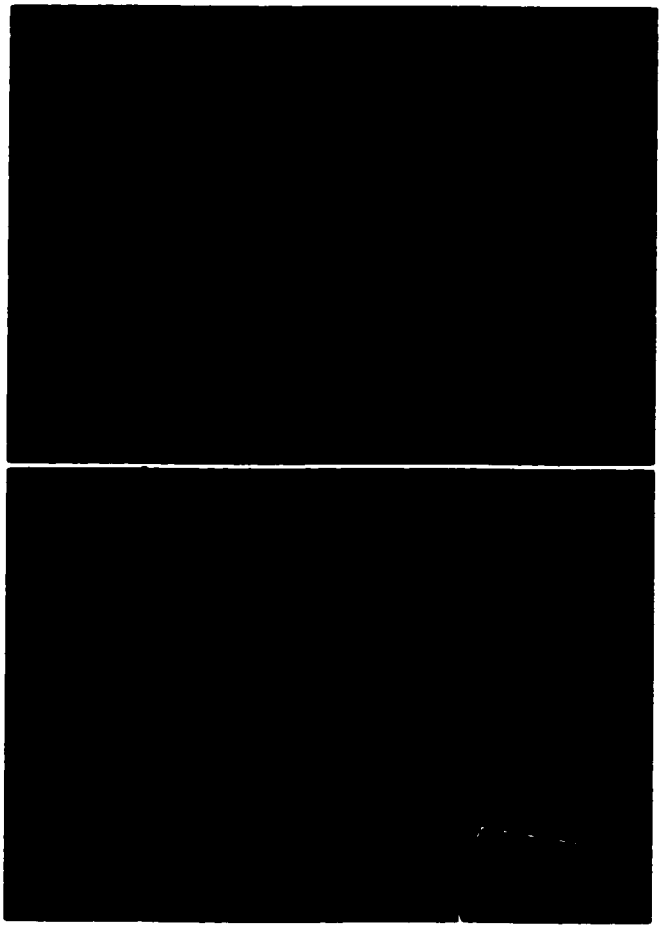
Contact dot surfaces are loosely related to the Lee & Richards (1971) concept of a configuration-dependent exposed surface area. Our implementation is similar to the Langridge-Connolly algorithm (Langridge *et al.* 1981) for showing solvent-accessible molecular surfaces, in that a spherical probe is rolled around the van der Waals surface of each atom, visiting each of a set of predefined points, and a dot is drawn if certain tests are satisfied in that position. The differences are that the contact dot algorithm, as implemented in the program PROBE (Word *et al.* 1999a), uses a very small probe (typically 0.25 Å in radius rather than 1.4 Å), and leaves a dot when the probe touches or partially penetrates another not-covalently-bonded atom (see Figure 2-1), rather than when it *does not* touch another atom. Thus, small-probe contacts form discontinuous surfaces, the patches of which directly show the location, extent, and shape of *close* atomic contacts (e.g. Figures 2-2 and 2-3). Every dot lies on the van der Waals surface of some atom; there are no concave, reentrant surfaces.

FIGURE 2-1. In the small-probe contact dot algorithm, a 0.25 Å radius probe sphere rolls over the van der Waals surface of each atom, leaving a dot periodically wherever it also touches another atom that is not within three covalent bonds. Where non-H-bonding atoms overlap, the unfavorable contact is emphasized by drawing spikes instead of dots.



Both PROBE and the original SP program avoid a computationally expensive $O(n^2)$ search for contacting atoms by dividing space into a lattice of boxes, the sides of each box ($0.2\text{Å} + 2radius_{\text{maxVDW}} + 2radius_{\text{probe}}$) being slightly longer than the largest possible contact distance, and sorting each atom into the appropriate box. Neighbors can then be efficiently located by examining the 27 boxes surrounding an atom's position. Subdividing space in this manner to identify adjacent objects ("hashing") is a standard computational technique (Abagyan, Totrov, and Kuznetsov 1994; Eisenhaber *et al.* 1995; Sedgewick 1990) not limited to Cartesian space.

FIGURE 2-2. (a) Small-probe contact dots between residues Trp126, Arg95, and Leu91 of wild-type T4 lysozyme (Matsumura *et al.* 1989), colored by atom type: O, red; N, blue; C, white; and H in the color of its bonded atom. This cross-section shows the large flat surface between the Trp ring and the Arg guanidinium group, and the interdigitation of methyl and methylene H atoms between Arg and Leu. (b) View into the face of Trp59 in FK506-binding protein 1BKF (Itoh *et al.* 1995). The Trp side-chain NH at top right makes an H-bond to the π electrons of a Phe ring, seen by the overlapping lens shape of blue and white dots.



Atom information for residues, bases and heterogens are read and processed from Protein Data Bank format coordinate files (PDB; Berman *et al.* 2000; Bernstein *et al.* 1977). For more information, see "Using PROBE: Schemes and Patterns" on page 23.

Dot Representations

Dot representations are useful because of the advantages in interactive molecular graphics of both seeing the surface and seeing beyond the surface. Furthermore, dots and spikes are more distinct and easier to pick or label than are continuous surfaces, allowing positive identification of the atoms responsible for each part. Other common illustration techniques which use continuous surfaces—such as CPK spheres (Max 1979; Porter 1978), opaque and semi-transparent molecular surfaces (e.g., GRASP (Nicholls, Bharadwaj, and Honig 1993), SURFNET (Laskowski 1995), RIBBONS (Carson 1997), and RASTER3D (Merritt and Bacon 1997)), and "egg-shell" contact surfaces (MacKenzie, Prestegard, and Engelman 1997)—produce a more familiar appearance in static illustrations, but contact dots are much easier to interpret for serious scientific detail in crowded molecular environments.

Which Contacts are Displayed

For a visually manageable display of side-chain packing, the default is to show only sidechain–sidechain and sidechain–mainchain contacts, but mainchain–mainchain contacts can be included for smaller regions or whenever they are specifically relevant. Small-probe dots can be calculated either for internal contacts within a group of atoms (e.g. an entire protein subunit) or else for the contacts between two specified groups of atoms (e.g. two neighboring alpha-helices, or a ligand and its environment). For details, see descriptions of the various options, beginning on page 24.

The purpose of small-probe contact surfaces is to analyze non-covalent contacts. Thus, in this work, dots are not calculated between atoms connected by two cova-

lent bonds or less. Contacts across dihedral angles (atoms three bonds apart) may profitably be included when analyzing local conformation, but they will show many small bumps because atoms lie closer together in those short-range interactions. For visualizing long-range packing in an entire domain or subunit a good level of clarity is obtained by including contacts of atoms more than four bonds apart if one of them is a H and more than three bonds apart otherwise; this is the default in PROBE. A uniform criterion of > 3 bonds for all atoms is best when evaluating individual residue conformations.

Dot Colors

One color scheme for these contact dots (e.g. Figure 2-2) reflects atom type: C white; N, blue; O, red; S, yellow; and H in the color of its bonded heavy atom. The NH···O hydrogen bonds, then, show as interpenetrating lens shapes in red and blue. Overlapped van der Waals shells of non-polar atoms are emphasized by showing spikes instead of dots: a spike is a line drawn from the dot position to the contact midplane, along the atom radius. An alternative color scheme unique to PROBE (see Table 2-1 and Figure 2-3) reflects the gap distance between atoms at each dot position: green or yellow for good contact (greens for narrow gaps, yellows for slight

TABLE 2-1. Gap Colors in Kinemage Format

Gap (Å)	0.5 to > 0.35	0.35 to > 0.25	0.25 to > 0.15	0.15 to > 0.0	0.0 to > –0.1	–0.1 to > –0.2	–0.2 to > –0.3	–0.3 to > –0.4	≤ –0.4	*H-bond*
Color	blue	sky	sea	green	yellowtint	yellow	orange	red	hotpink	greentint

overlaps), pale green dots for H-bonds, blues for wider gaps, orange or red spikes for unfavorable interpenetrations, and hot pink spikes for "clashes" of ≥ 0.4 Å. The default dot density, used for the figures here, is 16 per $Å^2$.

FIGURE 2-3. A thin slice through a small-probe contact dot kinemage showing the van der Waals interactions of Pro203 and neighboring atoms in Zn elastase at 1.5 Å resolution (PDB file 1EZM). Contact dots are color-coded by the gap between atoms as indicated in Table 2-1. Note the extensive contact and the interdigitation of hydrogens. White markers show the last two points picked (Pro $H^{\gamma 1}$ and one of its contact dots), while the distance between them and the identity of the last one are shown at the lower left. H atoms are from REDUCE, contact dots from PROBE, and the display is in MAGE.

Output Format and Display

Contact dots are output by PROBE as a simple text file of dot lists (vector lists for the clash spikes) in *kinemage* format, with color, source atom, and contact type specified. Alternative output formats are available for display as graphics objects in the crystallographic model-rebuilding programs mentioned in Chapter 7: O (Jones *et al.* 1991) and XTALVIEW (McRee 1993; McRee 1999). However, the contact dots themselves are most flexible if shown in the MAGE display program (Richardson and Richardson 1992, 1994), which supports the alternate color schemes, dot identification by picking, turning on or off groups by atom type or by contacts *versus* clashes *versus* H-bonds, saving many local views within a large structure, and animating between different forms. A *lens* option can restrict display of hydrogen atoms and contact dots to only a region around the last center picked, which allows

real-time viewing of contact-dot kinemages for large proteins on fast Macs and PCs, as well as on the Silicon Graphics Indigo2s or O_2s used in our laboratory. The text and caption windows in MAGE show supporting information chosen by the author of the kinemage; in addition, the text window will include the `USER MOD` records written onto the PDB file header by REDUCE, and the caption documents the PROBE command line that was used to calculate this set of contact dots.

Cavities

Small-probe contact analysis is poor at locating cavities, a limitation many people find non-obvious. A cavity results in a lack of contacts and is not easily distinguished from a filled region which is loosely packed. While PROBE can calculate the solvent contact surface (those points *not* in close atomic contact; see page 26), it can not calculate the reentrant areas to form a continuous molecular van der Waals surface which would clearly delineate a cavity. A technique is presented in Chapter 9 for contouring packing density which does create continuous surfaces but is still relatively insensitive to the distinction between cavities and loose packing. Bohacek and McMartin (1992) creatively employ a (one-sided) solvent accessible surface (Lee and Richards 1971) to visualize ligand binding complementarity, with the ligand as a stick figure resting on that surface. The section below on solvent contact surfaces includes a description of how to use PROBE to generate the solvent accessible surface (both exposed and "buried"), which can display binding site geometry as per Bohacek and McMartin but can also show internal cavities capable of containing solvent. The location and size of cavities, however, are best determined by

other methods (Hubbard, Gross, and Argos 1994; Kleywegt and Jones 1994; Laskowski 1995; Liang, Edelsbrunner, and Woodward 1998; Nicholls, Bharadwaj, and Honig 1993; Vriend 1990).

Hydrogen Atoms

We have found that proper application of small-probe contact analysis requires the use of explicit hydrogen atoms, including those on small-molecule ligands. The program REDUCE (described in Chapter 3 and in Word *et al.* 1999b) adds them to PDB-format coordinate files using local geometry and can perform extensive optimizations of hydrogen position. For comparison, PROBE can also generate contact surfaces using *implicit hydrogens* (also called "united atoms") where hydrogens are ignored and the van der Waals radii of non-hydrogen atoms are increased in compensation.

Parameters

There are small, but for our purposes significant (~ 0.1 Å), differences in the bond lengths used by various refinement or modeling programs, depending mainly on whether the H position is taken as representing the nucleus or the center of the electron cloud (e.g. Iijima, Dunbar, and Marshall 1987). We generate and standardize to the longer values (i.e. nucleus positions) in REDUCE, since they are more consistent with the data used to derive van der Waals radii (Bondi 1964; Gavezzotti 1983). Of course, the effects of bond length and of van der Waals radius for hydrogen atoms

interact strongly for our purposes. The parameter set used in PROBE and REDUCE is given in Table 2-2; it includes, for instance, smaller radii for polar H atoms and those on the edges of aromatic rings. Those radii were decreased both for theoretical reasons of charge polarization (see Gavezzotti 1983) and also because the larger radii produced significant internal clashes for all possible arginine rotamers.

TABLE 2-2. Atomic parameters used in PROBE and REDUCE

A. *Bond lengths (Å)*	
C-H	1.1 Å
N-H, O-H	1.0
S-H	1.3
B. *Van der Waals radii (Å)*	
H	1.17 Å
H (aromatic)	1.0
H (polar)	1.0
C	1.75
C (carbonyl)	1.65
N	1.55
O	1.4
P	1.8
S	1.8

On the other hand, we must justify using any van der Waals terms at all for polar H atoms, since they are set to zero in many energy calculations. Van der Waals terms for polar H atoms have been shown unnecessary for modeling H-bonded systems (Hagler, Huler, and Lifson 1974; SPC, TIP3: Hermans *et al.* 1984), which is of course their dominant mode of interaction. However, the work presented here (especially Chapter 5) shows that van der Waals clashes are indeed essential to analyzing polar–non-polar H atom interactions, and also for understanding why groups

cannot adopt specific alternative conformations: the "negative design" questions that arise in protein design (Hecht *et al.* 1990; Richardson *et al.* 1992) or in considering what would have been the consequences of an alternative side-chain position. The polar H issue is discussed in detail in Chapter 3 (see also, Word *et al.* 1999b), since it is especially crucial for the analysis of side-chain amide conformation.

All these parameters are, of course, compromises. This simple spherical-atom formalism does not allow for the non-uniformities of motion or the real shapes of orbitals, and these radii, which are optimized for long-range interactions, are a little too large for representing the contact interactions around a local dihedral angle. However, they do include a built-in average allowance for the expanding effects of thermal motion, since they were originally derived from accurate small-molecule crystal structures and other experiments in which thermal motion was present.

Water

Water molecules are difficult to include in these procedures, since their hydrogen positions are almost never known. However, their effects can be approximated by one or a combination of the following methods: (1) most roughly, with an asymmetrical PROBE option that uses implicit H for one group (the water molecules) and explicit H for the other group (the protein); (2) by presuming that water molecules can always orient so as to present whatever is needed for each interaction, and therefore using the explicit O radius for van der Waals bumps or to H-bond donors, and an O plus H radius to H-bond acceptors; (3) for well-surrounded water molecules, by orienting appropriately relative to the closest obligate donor or acceptor

and then optimizing rotation around that axis. When PROBE finds water molecules they are usually treated at the level of method 2.

The exception: water molecules with explicit hydrogen atoms (at occupancies above 0.66) are treated like other OH groups by both PROBE and REDUCE. Jan Hermans' DOWSER program (Zhang and Hermans 1996) includes water hydrogens when it positions buried waters and works well in conjunction with our software. The ability to treat some waters in more detail than others can be used to distinguish between structural and solvent water.

When a water molecule (without explicit hydrogens) acts as a hydrogen bond donor and is especially close to the acceptor atom, the covalent bond between the water O and the putative water H is shortened slightly to compensate for the presumed off-axis H-bond geometry in the actual molecule. The OH bond length is reduced by an amount necessary to maintain an approximately constant H-bond overlap volume, equal to that in a strong H-bond with good geometry.

Scoring

Formulation of Scores

Quantitative measures for goodness-of-fit are defined in ways that seek to capture the insights and comparisons gained from the contact-dot visual representation of packing interactions. As in the definition of van der Waals energies, our scoring system is a sum of competing terms, but the contact scores are evaluated per dot, not per atom pair, and are then summed. Hydrogen bonds and other overlaps are quan-

tified by the volume of overlap. Those volumes are easily measured by summing the spike length (l_{sp} : one-half the radial depth the dot penetrates the other surface) at each dot, which is always calculated even though it is not visually displayed for H-bonds. Thus:

$$Vol(Overlap) = \sum_{Overlap} l_{sp} \qquad \textbf{2-1}$$

$$Vol(Hbond) = \sum_{Hb} l_{sp} \qquad \textbf{2-2}$$

In the extreme, if the gap between dots on the H and on the acceptor atom of an H-bond is less than an acceptable lower limit, then that dot is penalized as a clash; the lower limit is set as –0.8 Å for charged salt links and –0.6 Å for other H-bonds. In addition to O and unprotonated His N, potential H-bond acceptors are taken to include S and also the faces of aromatic rings, whose interaction preferences show clearly in the contact dots (e.g. Figure 2-2(b)).

On the other hand, despite indications that certain CH groups can act as H-bond donors (Derewenda, Lee, and Derewenda 1995; Karle, Ranganathan, and Haridas 1996), the contact dots have not shown unequivocal evidence of such CH···O H-bonds, except for $H^{\delta 2}$ and $H^{\varepsilon 1}$ of histidine rings, which would indeed be among the most polar CH groups in proteins. Our van der Waals radii, chosen from data independent of this effect, are all slightly smaller than the ones typically used in studies of CH···O H-bonds (e.g. 1.4 Å *versus* 1.5 Å for O). The overlaps we see for non-His CH groups are not reproducible or large enough to determine correct parameters, and treating them as favorable would not improve our analysis significantly. For

His, this effect raises the number and degree of unavoidable overlaps (some His rings show four potential H-bonds), but since NH···O H-bonds are very much stronger the decisions on possible His ring flips are still made correctly. Therefore, CH groups have not been treated as H-bond donors in the present implementations of our algorithms.

The non-overlapped van der Waals contacts, in contrast with H-bonds or clashes, cannot be defined as volumes, and they need a weighting function similar to that provided visually by the gap-coloring, so that close contacts count more than distant or significantly overlapped ones; slight overlaps should still be favorable in net effect. This can be accomplished with an error-function weighting, so that each non-H-bond, non-clash contact dot is counted with a weight of:

$$w(gap) = e^{-\left(\frac{gap}{err}\right)^2} \qquad \text{2-3}$$

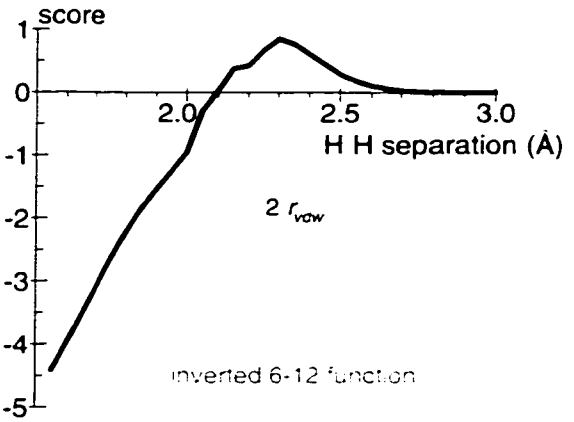
where the gap is the distance from the dot to the other atom's surface, and the error is taken as the probe radius, typically 0.25 Å. The maximum dot weight is thus 1.0 at optimum contact, dropping to $1/e^4 = \sim 0.02$ for the most distant dots allowed by the probe diameter. For slight overlaps, the annulus of contact dots surrounding the overlap keeps the overall score favorable, but the outer edge of this region is restricted to the radius for an optimum contact so there are no contacts around large overlaps. Since the overlap-volume terms and the contact error-function term are not commensurate, an arbitrary but suitable scale factor between them is needed. In practice, multiplying overlap volume by ten and H-bond volume by four before

adding the three terms gives an overall scoring profile similar in shape to the van der Waals function for an isolated pairwise interaction, thus:

$$score = \sum_{dots} w(gap) + 4Vol(Hbond) - 10Vol(Overlap) \qquad 2\text{-}4$$

For multi-atom interactions, the contact dots and their scores combine in a more complex way than addition of unmodified pair-wise terms, but in a way which relates directly to the size and shape of the atomic surfaces that are actually in proximity, including geometrical allowance for how an atom partially shields its neighbor.

FIGURE 2-4. Plot of PROBE scores for interacting CH HC groups at different separations. Van der Waals spheres for the two hydrogens touch at 2.34 Å. For comparison, an inverted Lennard-Jones 6-12 potential is shown in gray.



Processing Scores

The PROBE program can summarize these scoring data for all or selected parts of an entire structure; alternatively, it can output an "unformatted" intermediate file with information at every dot, which is then piped to simple utilities that sort and gather

any desired information per contact, per atom, or per residue. Scores are given both as raw values and also as normalized by possible surface area. That area is calculated by adding up all potential dots on all of the atom surfaces (that is, all dots not inside another covalently-bonded atom), which are accumulated when PROBE calculates the contacts. For contexts in which elimination of physically-impossible atomic overlaps is the main concern (e.g. protein design or structure determination), a *serious clash* is defined as a non-H-bond overlap of 0.4 Å or greater. The *clash score* for a structure is then calculated as the number of serious clashes per thousand atoms (including H). The ordinary contact score is high for a good structure, while the clash score is low for a good structure. A small Unix shell program called clashlistscore analyzes PROBE output to produce a list of atom pairs, scores, and *B*-factors for all clashes with overlap ≥ 0.4 Å. This list could usefully prioritize the analysis of problem areas, especially during structure refinement.

Using PROBE: Schemes and Patterns

Input and Output

A PROBE command is generally structured as follows

probe -optionflags... 'selection patterns'... *coordfile.pdb*... >> *graphics.kin*

PROBE constructs contact surfaces by reading and interpreting the ATOM and HETATM records in one or more PDB coordinate files. Explicit hydrogen atoms have usually been added by REDUCE in a previous step (see Chapter 3). Atoms can be selected for processing by: category (e.g. side chain), element type, atom name,

residue type, residue number, alternate conformation, chain ID, segment ID, file number, occupancy and *B*-factor. Individual models can be selected in files with MODEL records (e.g. NMR structures). Atom properties such as covalent and van der Waals radii and H-bond donor/acceptor status are assigned by interpreting the atom name field—the recently standardized atom type field in columns 73-74 is not consulted. Connectivity is inferred from the distance between pairs of atoms—CONECT records are not consulted (very often, they are not provided). To permit the exploration of contacts for conformations which place atoms in unusually close proximity to their neighbors, a command line option, **-stdbnds**, causes PROBE to consult an internal table of covalent bonds (by atom name) for standard residue types.

Output usually takes the form of graphics in *kinemage* format, although other types of output are available. The graphics is designed to be added to an existing kinemage of the protein structure, so the '>>' symbol is used to append the results.

Self Contacts

There are three general schemes for using PROBE: (1) self contacts, (2) contacts between groups, and (3) exposed van der Waals surface dots. Self contacts use the command line option: **-self**, followed by an atom selection pattern and then one or more input files. (Atom selection patterns will be described in detail later but their use in the following examples should be intuitive.) A list is made of atoms in the input which match the atom selection pattern and this list is scanned for contacts between atoms in the list (except mainchain–mainchain contacts, see below). If the

atom selection pattern is 'all' then all the input atoms are considered. This combination is used frequently enough that the command "probe inputfile >> out.kin" is shorthand for "probe -self 'all' inputfile >> out.kin".

Contacts Between Groups

Contacts between groups are useful for identifying, say, the interface contacts between two subunits or a side chain and its environment. This scheme requires two patterns (one for each side of the contact) and has two modes: -once and -both. With -once, the first pattern is called the *source* and the second the *target*. Dots are placed on the van der Waals surface of source atoms where they are in contact with target atoms. With -both (the commonest usage, with a two-sided appearance like most of the figures here), PROBE goes through the process twice: first the same as with -once and then a second time with source and target atoms swapped, resulting in dots for both sets of atoms. For two subunits, a command like "probe -both 'chainA' 'chainB' sod.pdb >> sod.kin" is often appropriate. For the side-chain packing case, a concise way to generate both sides of the contact is to use -once with overlapping selection patterns: "probe -once 'all' 'sc 57' chymotrypsin.pdb >> chymo.kin". Note that if instead 'sc 57' was the source pattern and 'all' the target, dots would be placed only on the His57 side chain. To only put dots on the neighbors, "probe -once 'not (sc 57)' 'sc 57' chymotrypsin.pdb >> chymo.kin" is required.

Solvent Contact Surface

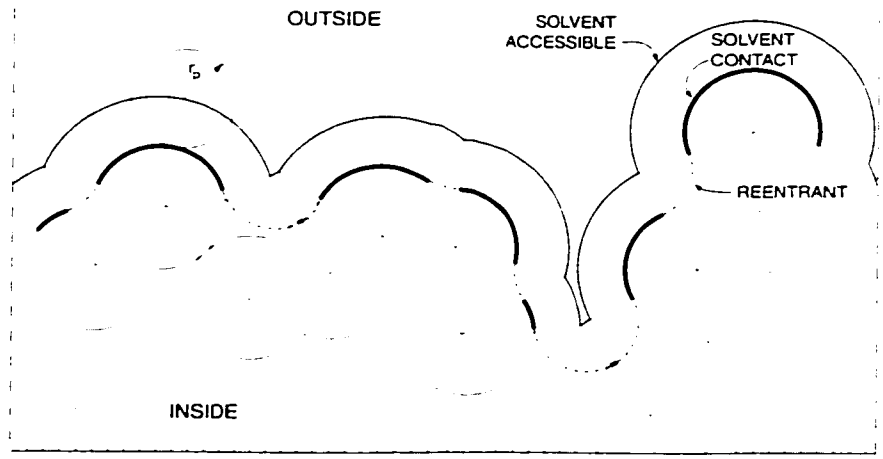
The third scheme is to generate dots for the *solvent contact surface* (SCS) using the option: -out and a single atom selection pattern. This creates the convex part of the molecular van der Waals surface where the probe sphere does *not* touch the surface of any other atom (see Figure 2-5, modified from Richards 1977). The concave *reentrant* part is not generated, so the solvent contact surface consists of multiple patches with a total surface area less than that of the molecular surface.

Importantly, a water size probe (specified using -rad1.4 in the command) produces a solvent contact surface which is directly related to the standard *solvent accessible surface* (the locus of the probe's center as it moves over the van der Waals surface; Lee and Richards 1971) and can therefore be used in estimations of binding affinity and other molecular properties. The solvent accessible surface (AS) is essentially a projection of the solvent contact surface a fixed distance further from the center of each exterior atom. The relationship between the area of the solvent accessible surface (ASA) and the area of the solvent contact surface (SCSA) depends on the average van der Waals radii of exterior atoms, weighted by exposed area. This average radius does not vary much from protein to protein; the ratio ASA/SCSA was estimated empirically to be 4.5 with explicit hydrogens and 3.3 with implicit hydrogens (from 7RSA, 3LZM, 1MCT). For nucleic acids, the estimated ratios are 4.3 and 3.4, respectively (from 1YTB, 1DNS, 1RNA). The general expression relating SCSA and ASA treats each atom type separately.

$$ASA = \sum_{t\,=\,type} SCSA(t)\frac{(r_t + r_p)^2}{r_t^2} \qquad \text{2-5}$$

where $SCSA(t)$ is the solvent contact surface area subtotal and r_t is the van der Waals radius for atoms of type t, and r_p is the probe radius (= 1.4 Å). The summary table for a solvent contact surface generated with the -count option of PROBE (version 2.1 and later) includes an estimate of the ASA computed in this way. At the default dot density of 16 dots per Å^2, surface areas tend to be underestimated by 6-7%. Though increasing the dot density ten fold (-dens160) will make the calculation take almost ten times as long, it will reduce the undercount to under 2%. The average ASA/SCSA ratios listed above are for 160 dots per Å^2; when using 16 dots per Å^2, increase the value by 5%.

FIGURE 2-5. Diagram of the Solvent Accessible Surface, the Solvent Contact Surface, and the Reentrant Molecular Surface (with permission, modified from Richards 1977).



To determine the surface area buried upon dimerization, add the area for each monomer (e.g. chains A and B), calculated separately with

```
probe -dens160 -count -exposed 'chainA not water' dimer.pdb
```

```
probe -dens160 -count -exposed 'chainB not water' dimer.pdb
```

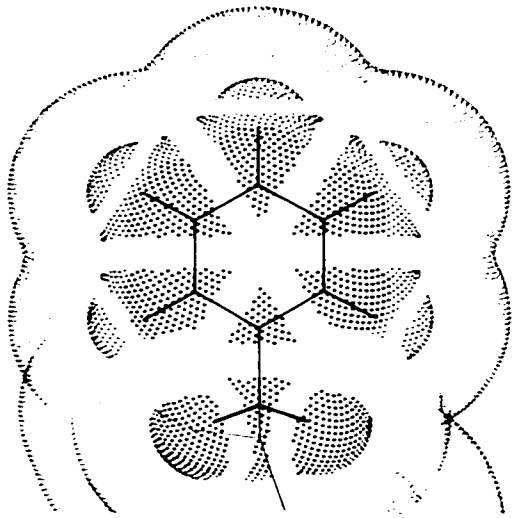
where the option -exposed is a version 2.1 abbreviation for the combination of options: -drop -rad1.4 -out (the -drop option tells PROBE to completely ignore all the atoms which are not selected so that the non-selected chain and the waters will not shield parts of the surface). Then subtract the area for the dimer, calculated by:

```
probe -dens160 -count -exposed 'not water' dimer.pdb
```

The energy of binding is approximately 0.02 kcal/mol/Å^2 of buried solvent accessible surface area (Janin and Chothia 1990). Although we have not yet tried to systematically correlate molecular properties with the solvent contact surface area or, more interestingly, the atomic contact area of a binding interface, it is possible that one of these parameters could surpass the ASA in predicting power.

If the probe size is set to zero (-rad0) then gaps between surface patches are eliminated, converting the solvent contact surface into a CPK-like rendering. At higher dot densities (e.g. -dens100) the bulk of a side chain or ligand is nicely represented, especially when atoms are colored by atom type. Combining a probe size of zero with an offset of 1.4 Å added to all van der Waals radii (-rad0 -addvdw1.4) converts the solvent contact surface into the solvent accessible surface.

FIGURE 2-6. Solvent contact surface (SCS; heavy dots) recessed within the accessible surface (AS; light dots). The concave reentrant parts of the molecular surface are not plotted, leaving gaps between SCS patches. The AS results from an inflation of the SCS by 1.4 Å, and is continuous because the reentrant parts become reduced to infinitely thin seams between convex surface patches. The dot density is 64 per $Å^2$.



Atom Selection Patterns

Mastering atom selection patterns is one of the more challenging tasks when learning how to use PROBE. These patterns are written in their own little language, which borrows from other pattern matching languages but is by necessity idiosyncratic. A pattern consists of one or more key words representing types of atoms (e.g. sc, mc, water, dna, het, all) and a few special characters. The one and three letter codes for standard amino-acids and the three letter codes for nucleic acid bases are all recognized key words. Numbers select residues and a residue range is indicated with a dash ('6-85'). Some frequently used key words have a variable part: 'chainI altA blt40 ogt33' selects atoms from chain "I" in the "A" alternate conformation (where there are alternates, and the single conformation otherwise) with a *B*-factor

less than 40 and an occupancy of greater than 33%. A few key words are modifiers; 'not A' selects everything but 'A'. The most complicated key word is 'within' which selects atoms within a specified range of a given x,y,z position and has the format 'within #.# of #.#, #.#, #.#'. A complete list of PROBE's keywords is generated by the command "probe -help".

The separators between key words control how the key words are combined. Blanks generally mean "boolean and" so the pattern 'nitrogen sc' selects only side-chain nitrogen atoms. Commas mean "boolean or" so the pattern 'sc F,Y,W,H chainE' selects the aromatic side-chain atoms in chain E—note that the commas are applied before the blanks (they have higher precedence). There are two ways to control precedence, with parentheses and with the vertical bar ("|"), a low precedence "boolean or". For example, to select all the atoms in a PDB file except for those in the side chain of Asp102 which we get from a second PDB file with a different conformation, we can use the pattern 'file1 not (sc 102) | file2 sc 102'. (This last pattern should be used with the -drop option to ensure the non-selected atoms do not interfere.) One-word patterns such as 'all' do not require quote marks, but otherwise patterns are packaged within either single or double quotes.

Special Interactions

To minimize visual overload from the mundane, several kinds of interactions are ordinarily not considered when generating contact surfaces. In certain situations it may be important to include these interactions. Mainchain–to–mainchain interactions are ignored by default since, in proteins, these contacts are dominated by the

hydrogen bonds defining the secondary structure. In order to visualize how main-chain atoms can come together in interesting ways, the command line option -mc must be used to ensure these contacts are displayed. This is critical when analyzing nucleic acid structures and complexes since the backbone in nucleic acids is much more exposed than in proteins (and is easier to get wrong). *Warning*: it is easy to forget to use -mc and this can cause much grief!

Similarly, if the structure contains coordinates for water molecules, contacts to the waters are shown but water–to–water contacts are generally not displayed. Generating them requires the option -wat2wat.

Bond Distance

As discussed earlier, PROBE does not show contacts between covalently bonded atoms or atoms separated by only a few bonds. The number of bonds can be controlled by command line options: -1, -2, -3, or -4. The default (-4) is a special case: it eliminates contacts between atoms which are up to four bonds apart only if one is a hydrogen and otherwise acts like -3 ignoring contacts for three or fewer bonds.

PROBE considers metals to be bonded to their ligands, therefore metal–ligand contacts are not generated.

Lensing

At times the total number of dots generated by PROBE can be overwhelming, both for the viewer and for the display program, making rotation and redrawing operations sluggish. The kinemage format includes a *lens* feature which restricts the dis-

play of lensed objects to a fixed size region around the center of rotation. Ideally, the detail (such as contacts and hydrogen atoms) would be lensed, while gross features (such as main chain) would always be shown. PROBE supports this facility with the command line option: -lens, which applies to contact dots (better seen in close-up) but not to H-bonds or clashes (which are less dense and for which overviews are informative).

Unformatted Output

Many of the types of contact analysis presented in this and later chapters (e.g. serious clashes by atom) are not performed by PROBE alone, but by combining PROBE output with other software tools inside Unix shell scripts. Most of the task specific programming is done in the AWK language (Aho, Kernighan, and Weinberger 1988). These tools interpret "unformatted" PROBE output (option -u) an easy to parse table that lists: position, partial score, source and target atom details, and other parameters for each dot or spike. As an example, clashlistcluster uses this information to produce tables of serious clashes, starting with the worst, in which entries for clashes in the same area of the molecule are sorted together.

WEBPROBE

Patrick McConnell and Kim Gernert at BIMCORE, Emory University, have created an HTML interface to PROBE called WEBPROBE which is available to the public at http://www.bimcore.emory.edu/Software/MolMod/probe_intro.html. Once the user has uploaded a structure and made selections from the menu of program options, the server will add hydrogens and calculate the contact surface, generally returning

a kinemage. Although it provides access to only a subset of PROBE's capabilities, the web interface is considerably more transparent for the occasional user than PROBE's raw command line interface.

CHAPTER 3 *Explicit Hydrogen Atoms*

Where are the Hydrogens?

Just as low-resolution protein crystal structures are often represented by the C^{α} backbone, published high-resolution structures usually include only the non-hydrogen atoms; occasionally polar hydrogens are represented. Despite the fact that they account for approximately half of all protein atoms, hydrogens do not diffract X-rays strongly and are generally considered an atomic-resolution feature.

Hydrogens do diffract neutrons, allowing protonation states at catalytic sites, hydrogen bond orientations, and even methyl group conformations to be imaged (Kossiakoff and Shteyn 1984; Kossiakoff *et al.* 1991; McDowell and Kossiakoff 1995). Unfortunately, the neutron source at Brookhaven National Laboratory has been shut down leaving the U.S. without neutron diffraction facilities. Stations are available at Grenoble, France, but the planned neutron protein crystallography station at Los Alamos National Laboratory is expected to be heavily oversubscribed

(Langan and Schoenborn 1999). Considering the lack of sources, relatively few neutron structures are available or anticipated.

Surprisingly, NMR structures are frequently published without all the hydrogens, requiring the active removal of hydrogen atom records from the coordinates. This may be a consequence of the training structural biologists receive; most are not used to looking at structures which include H atoms and find such displays confusing.

An often expressed reason against explicit consideration of all hydrogens during crystallographic refinement is that this would double the number of parameters—recklessly lowering the data to parameter ratio. Instead, refinement is usually carried out with expanded van der Waals terms for non-hydrogen atoms to account for the hydrogens' steric bulk (e.g., Brunger 1992). Being radially symmetric, this *implicit hydrogens* or *united atoms* approximation is oblivious to hydrogen orientation. However, fortunately the locations of most hydrogens are largely determined by the locations of the atoms they are bonded to. A refinement protocol which treats these hydrogens as 'riding' on the heavier atoms sees a helpful increase in the number of constraints without significantly increasing the refinement parameters. (In fact, it is not necessary to include hydrogens in the calculation of structure factors as long as they are used to determine energy terms.) It is possible to do this in SHELX (Sheldrick and Schneider 1997), while dihedral angle dynamics in CNS (Brunger *et al.* 1998) would be suited for use in such a manner. This chapter presents a similar approach to geometrically position hydrogen atoms after refinement but prior to analysis.

Adding Hydrogen Atoms to Structural Coordinates

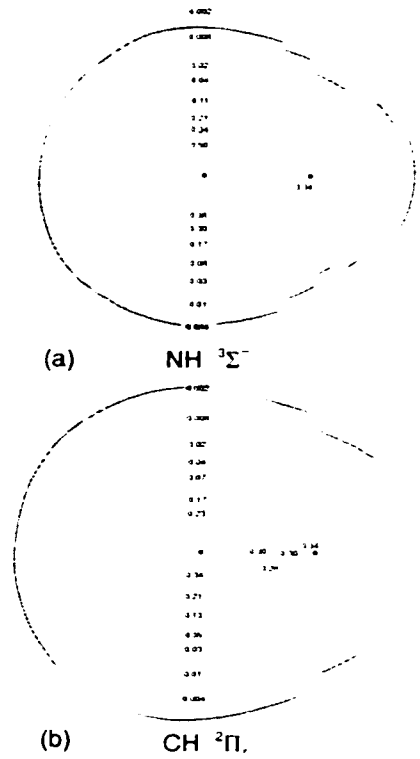
Small-probe contact analysis requires the use of all explicit H atoms. The program REDUCE (Word *et al.* 1999b) adds them to PDB-format coordinate files, using local geometry. Methyl hydrogen atoms are added in staggered positions, and only the Met methyl groups are rotationally optimized. OH, SH, and NH_3^+ hydrogen atoms are rotationally optimized and His protonation is assigned as part of the H-bond network analysis described below. The contact dot algorithm used by PROBE has been re-implemented inside REDUCE and is used to score conformations during optimization. As in PROBE, water molecules are treated by presuming that they can always orient so as to present whatever is needed for each interaction. If water molecules are extremely close, we restrict their H-bonding score to a reasonable value. More information on the treatment of waters, the choice of parameters for bond lengths and van der Waals radii, and further details of the contact-dot methodology are explained in Chapter 2.

Polar H van der Waals

Some effort will be made to justify using any van der Waals terms at all for polar H atoms, since they are set to zero in many energy calculations. For instance, Hagler *et al.* (1974) optimized force-field parameters to agree with crystal dimensions, vaporization energies, etc. for a variety of small molecules with H-bonded amide groups, reaching the conclusion that van der Waals terms from the polar H atoms do not make a significant contribution beyond what can be fit with only electrostatic terms and van der Waals interactions for the heavy atoms. Similarly, Berendsen *et al.* (1981) found polar H van der Waals terms unnecessary for obtaining a satisfac-

tory model of water interactions. Those conclusions are undoubtedly justified for the cases analyzed, where each amide or water molecule interacts through H-bonds. However, such studies do not address the issue of what parameters are needed in the less typical but still fairly common cases where non-polar groups contact amide H atoms. In the analysis of side-chain amide conformations described in Chapter 5, such parameters were absolutely essential for evaluating the counterfactual non-polar–to–amide clashes that are characteristic of wrong flip choices for side-chain amide groups. Our approach returns to application of very simple physical principles: as stated for instance in the classic crystallographic text by Stout & Jensen (1968), "Any postulated arrangement of atoms... must fulfil the simple steric requirement that no two atoms should approach closer than the sum of their van der Waals radii unless they are bonded together."

FIGURE 3-1. Total molecular charge density contours in atomic units (Figure modified from Bader, Keaveny, and Cade 1967), overlaid with, in (a) a nitrogen van der Waals radius of 1.55 Å and a hydrogen radius of 1.0 Å at a distance of 1.0Å; (b) a carbon radius of 1.75 Å and a hydrogen radius of 1.17 Å at a distance of 1.1 Å.



In order to decide the best polar H radius for use in calculating contact dots, we have analyzed the spacings actually seen for non-polar-to-amide contacts and then compared the proposed radii to the shape of electron density distributions calculated by quantum mechanics using Hartree-Fock wavefunctions. Figure 3-1(a) shows calculated electron density contours for an NH group (Bader, Keaveny, and Cade 1967), the nearly round shape of which has been used to argue for a negligible effect of the polar H van der Waals terms (Hagler, Huler, and Lifson 1974). However, equivalent contours lie about 0.5 Å farther out in the H direction than else-

where, a difference that can be quite crucial in tightly packed regions. Shown overlaid in Figure 3-1(a) are the simple radii of 1.55 Å for N and 1.0 Å for the polar H, which fit the shape of the electron-density contours almost perfectly. For comparison, Figure 3-1(b) shows the equivalent overlay for the non-polar CH group: standard radii fit the calculated electron density equally well in both cases. The difference in contour level matched is within the uncertainty in our present parameters and has the advantage of keeping the NH radii conservatively on the small side.

Het Dictionary

REDUCE can add hydrogens to those 'heterogen' molecules included in the Protein Data Bank connectivity database (file ftp://ftp.rcsb.org/pub/pdb/data/monomers/het_dictionary.txt) or a similar file if constructed by the user. A modified version of the standard PDB het dictionary, with six additional entries and deprotonated phosphate groups, was used to build the *Top100DB1* (see Chapter 4) and is distributed with our software. The geometry used to position "het" hydrogens is based on the hybridization inferred from the element (determined by parsing the atom name) and the number of covalent bonds listed in the connection table for the attached non-hydrogen atom. The stereochemistry of geminal and methyl hydrogens is determined by their order in the non-hydrogen atom's list of connections.

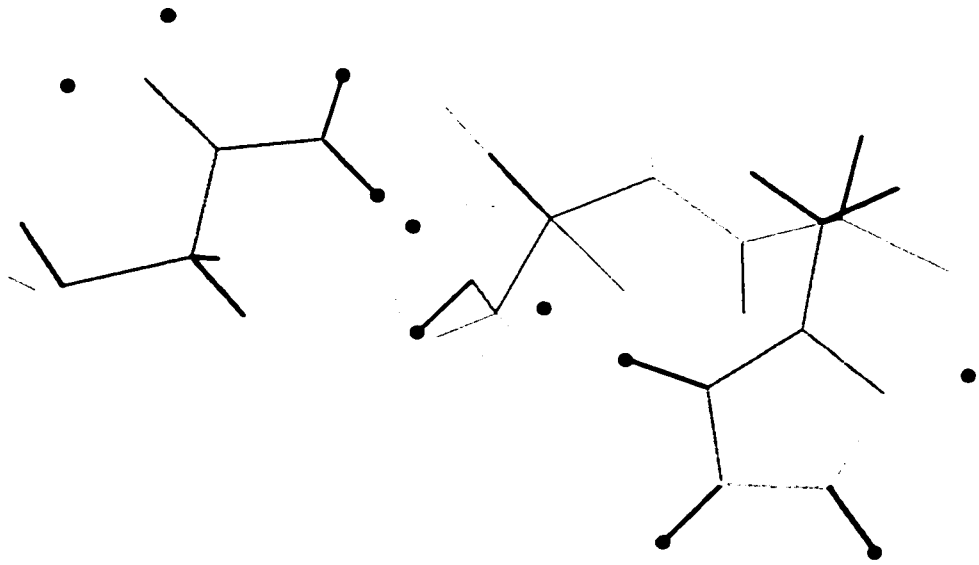
Optimizing Adjustable Groups: Rotations, Flips, and Cliques

Much of the work on optimization described below was motivated by the problem of determining proper orientations for Asn and Gln side-chain amides (see Chapter

5). Since side-chain amide groups often interact either with one another or with other hydrogen atoms (such as OH groups) that require positional optimization, definitive assignments of amide flips must be done within interacting networks rather than individually.

When REDUCE's -build option is used, these interacting closed sets of side chains with flippable groups or rotatable H atoms (most often local H-bond networks) are analyzed in a series of steps. First, REDUCE identifies all metal-liganding or covalently modified groups (listed for Asn/Gln in Table 4-1) and fixes their orientations to the original positions. Then all potentially-interacting pairs and larger closed sets among the remaining side chains or heterogen groups are identified by considering their full range of possible hydrogen positions (in both orientations for flips and each 10° for rotations) along with the positions of flippable heavy atoms. At each such position, a sphere is placed with the van der Waals radius of the corresponding atom. If a sphere from one adjustable group overlaps a sphere of another group, then those two groups can interact. Such pairwise interactions are then gathered into disjoint sets, which we call *cliques*, in the sense that their members all interact internally, but not with any adjustable group outside the clique. The cliques do not propagate through water molecules, because a water molecule is assumed here to be able to act as either an H-bond donor or acceptor independently, even if it makes more than one polar interaction. Only water molecules with $B < 40$ and occupancy ≥ 0.66 are used in our analyses. If a side chain has alternate conformations, the 'a' is used but not the 'b'.

FIGURE 3-2. Diagram of hydrogen placement in an Asn/Ser/His clique, where balls represent potential hydrogen positions. The close interaction between Asn and Ser generates 5 out of the 7 trial OH orientations. This clique would require $2 \times 7 \times 6 = 84$ permutations to be considered.



An Asn or Gln amide has only two possible states, and all of its interactions must switch between donor and acceptor in synchrony. A His has two flip states for the ring as a whole; within each of those we consider three possible protonation states (H only on N^{ε}, H only on N^{δ}, or doubly protonated with a small score penalty of 0.05), so that its two H-bonds usually but not always change in correlation. Double deprotonation is allowed only if the His ligands two metal ions, such as for His61 in the 1XSO superoxide dismutase (Carugo *et al.* 1996).

For fully rotatable H atoms, before undertaking the combinatorial step, we use the following process to reduce the number of states that must be considered. For each OH or SH, orientations for the rotatable H atom are selected in the direction of each one of the potential H-bond acceptors surrounding it. If the acceptor is too close for

an acceptable straight-line H-bond, then potential H orientations are also defined 15° (and, if necessary, also 30°) on either side of it. Finally, an additional orientation is located which avoids these acceptors and has minimal interaction with all surrounding atoms. For an OH surrounded by three H-bond acceptors, for instance, we would thus define four potential positions to be tried in the combinatorial search; but if the acceptors were all close, there would be ten potential H positions (three near each acceptor and one that avoids them all). Each Met methyl and each Lys or N-terminal NH_3^+ is considered in four possible orientations, 30° apart. Each side chain in an interacting clique has a fairly small number of possible H arrangement states that must be considered. When combined, the permutations multiply and, in rare cases often involving multimers or crystal contacts, can reach astronomical levels, preventing calculation. Since the largest clique found in our *Top100DB1* reference structures contains six members (see Chapter 4; in a later set of 240 proteins, one clique had eight members), an exhaustive search was computationally tractable in practice. (For alternative strategies, see "Working With Large Cliques" on page 48.)

The isolated OH, SH, NH_3^+, and Met methyl groups are rotationally optimized in REDUCE by sampling every preassigned orientation and then testing 1° increments around the best one; the final rotation angle and score for the best position are reported, and the optimized H atom is added to the output PDB file. Each clique of two or more is then analyzed by an exhaustive search through all combinations of the preassigned potential H positions for all residues in the clique, plus a final rotational optimization around the best preassigned position. For each residue, the pro-

gram accumulates its best score, the best total score for its clique, and (for Asn, Gln, or His) the best total clique score found with this residue in its opposite flip state. Those scores and the best assignment (e.g. "`FLIP+both NH`" for a His) are written both to the screen and onto the header of the output PDB file, and H atoms for the chosen clique conformation are added to the PDB output file. Each Asn/Gln/His (unless fixed in advance by a covalent modification) is assigned to one of four categories: *Keep* (`K`), *Flip* (`F`), double-*Clash* (`C`), or *unknown* (`X`). An adjustable penalty can be applied to the difference between the best score and the best flipped score, in order to automatically leave the marginal cases in the state originally assigned by the depositors. Alternatively, they can be examined and the flip state decided by the user.

Using REDUCE

Building Hydrogens

In most circumstances, the recommended command when using REDUCE to add hydrogens to a PDB file and standardize the bond lengths of existing hydrogens is

```
reduce -build coordfile.pdb > coordfileH.pdb
```

which includes the optimization of adjustable groups (see page 39). Calculation generally takes several minutes but occasionally take much longer. When greater speed is important, the -build option can be dropped; hydrogens will still be added, but not His side-chain NH hydrogens, and side-chains will not be flipped. For even greater speed, but even less accuracy, adding -nooh and -noadj will skip the OH

and SH hydrogens and eliminate optimization altogether. Input is from the specified PDB format coordinate file and the new, updated PDB coordinates are written to "standard output," here redirected to a file with the '>' symbol. A full list of program options is generated with: reduce -help.

Disulfides, covalent modifications, and connection of the ribose-phosphate nucleic acid backbone, are recognized and any hydrogens eliminated by bonding are skipped. When an amino acid main-chain nitrogen is not connected to the preceding residue or some other group, REDUCE treats it as the N-terminus and constructs an NH_3^+ *only* if the residue number is less than or equal to an adjustable limit (1, by default). Otherwise, it considers the residue the observable beginning of an actually-connected fragment and does not protonate the nitrogen. REDUCE does not protonate carboxylates (including the C-terminus) because it does not specifically consider pH, instead modeling a neutral environment.

Hydrogens are positioned with respect to the covalently bonded neighbors and these are identified *by name*. Non-standard atom names are the primary cause of missing or misplaced hydrogens. If REDUCE tries to process a file which contains hydrogens with non-standard names, the existing hydrogens may not be recognized and may interfere with the generation of new hydrogens. The solution may be to remove existing hydrogens before further processing (see "Trimming hydrogens from a file" below).

Fixing an Orientation

At times it is useful to control the flip state or rotation angle of an adjustable group when adding hydrogens, either because the correct orientation has already been established, allowing the optimization time to be reduced, or because a non-optimal orientation is sought. As an example of the latter, we have developed procedures (flipNQkin and flipHISkin) which generate kinemages that permit animation between the optimal and the alternative flip states for Asn, Gln and His side chains; for REDUCE to calculate that second animation state, the N/Q or H orientations must be fixed non-optimally. These animations allow the viewer to judge the sensitivity of the residue's local environment to a side-chain flip.

The name of a file containing an entry for each fixed orientation residue is passed to REDUCE after the -fix option. The colon delimited format is similar to the orientation data REDUCE prints in the file header (`action:residueID:comment`), and because spacing matters in the residue identifier string, the easiest way to produce this file is to copy and edit `USER MOD` records from REDUCE output. The action can be one of three kinds depending on residue type: `O` to leave in the original orientation, `F` to flip the orientation, and `R#` to rotate a dihedral to an angle of #°. Using either `O` or `F` with His side chains allows the protonation state to vary; to specify a particular orientation and protonation state use `F#` where # is the number of the state (1-3 for the original orientation with H (1) only on $N^{\varepsilon 2}$, (2) only on $N^{\delta 1}$, or (3) doubly protonated; 4-6 for the corresponding three flipped states).

Trimming Hydrogens From a File

In addition to adding hydrogens, REDUCE can remove hydrogens from a file with the -trim option. This can be used, for example, to update the orientation of Asn/Gln/His side chains where H atoms are not wanted: first build the hydrogens and then trim them back out. Trimming can occasionally be fooled if a hydrogen has been given a non-standard name by some other program. The most common example of this comes from left-justified atom names favored by some researchers: gamma hydrogens masquerade as mercury atoms! In this case, manual editing may be required.

Atom Naming Conventions

PDB format provides 4 spaces for the atom name, broken into: two for the right-justified element symbol, one for position, and one for stereochemistry. A C^{α} is thus named '_CA_', while a calcium is named 'CA__' (the underscores represent blanks). The short branch methyl of isoleucine has atoms '_CG2', '1HG2', '2HG2', and '3HG2'; the hydrogens following a rule in which the stereochemistry "wraps" to the first space. Unfortunately, there are a large number of non-standard PDB-like file formats in use—each refinement and modeling program seems to have come up with their own. REDUCE and PROBE use atom names to identify the various components of a side chain and, though they try to be accommodating, these programs work most reliably with standard PDB files. To continue with the left-justified atom names example, the alpha carbons can't be hydrogenated because they have been calcified.

In a move sure to cause even more confusion, in October of 1999 the RCSB announced they will offer coordinate files in both the existing standard and an alternative standard based on NMR nomenclature (Markley *et al.* 1998). Atoms are renumbered in the new convention—beta methylene hydrogens '1HB_' and '2HB_' become '2HB_' and '3HB_', respectively—requiring changes to any software which relies on these names. For several years I have been working to convince developers of macromolecular software tools (e.g. O, XTALVIEW, X-PLOR and CNS) to adopt the PDB standard for coordinate exchange. We use and make available my tool pdbcns for converting between standard and X-PLOR/CNS nonstandard nomenclature. Now there is some question which nomenclature is *the* standard. Currently, we plan to develop versions of REDUCE and PROBE which support the new convention, and perhaps both conventions, as soon as possible.

This new version of REDUCE will also provide a way to process files generated by X-PLOR or CNS which do not use the PDB chain-ID to identify different chains, but instead use the segment-ID field to group residues. Currently, these files cause problems because REDUCE identifies some atoms by a name which includes the chain-ID but not the segment-ID (e.g., see Chapter 7). A similar difficulty is encountered when working with files having alternate conformations which are not marked as such. Although we would like to be able to correctly process every coordinate file, regardless of origin, REDUCE works best with files that conform to PDB standards.

Working With Large Cliques

The current version of REDUCE uses brute-force enumeration to optimize the conformations of adjustable groups. If a clique of adjustable groups is too large (> ~7) this sort of search technique is inadequate—the enumeration will be abandoned and these groups will be left in their initial conformations. The cutoff point is based on the total number of permutations, which the user can control with the -limit# option. Although we are considering more powerful search techniques for these situations, some work-around strategies have been developed.

Examination of the clique may reveal that the orientations of one or more groups are obvious; for instance, they may interact with obligate H-bond donors or acceptors. By fixing the orientation of these groups (as described above), the total number of permutations is reduced. This is especially effective if it breaks the clique into smaller sub-cliques or singletons.

An alternative way to break up cliques is to rotate all the methionine CH_3s and lysine/N-terminus NH_3^+s in an initial pass, then keep them fixed in a second pass.

reduce -nooh *inputfile* | reduce -build -norotmet -norotnh3 - > *outputfileH*

The single dash towards the end of the command line tells REDUCE to read data piped ('|') from the first pass rather than from a file. A few NH_3^+ H-bonds may have inferior geometry with this two-pass approach but the result is otherwise comparable to using -build alone and can be combined with the first approach, if necessary. With this technique, cliques requiring many hours to process have been converted into several smaller problems which were all be solved in a matter of minutes.

CHAPTER 4

Contact Surfaces as a Graphical and Quantitative Validation Tool

> "The explorer Sebastian Cabot (1474?—1557) brought back from America a dancing bear. This animal radically misled him when he came to write his treatise on the bears of America, which do not dance, mostly." (Barthelme 1998, pg. 34)

The Value of Good Examples

The motivating idea behind the creation of PROBE was that protein design and engineering might be advanced through the use of contact surfaces to identify and gain some understanding of the various ways in which atoms pack in protein molecules. Specifically, packing *exemplars* would be compiled and studied from a catalog of protein structures of the highest precision and accuracy. Bias would be minimized by studying a wide variety of different structural types. Minor inconsistencies in the data (such as non-standard nomenclature) would be regularized, to facilitate comparisons between structures.

The development of this resource, called the *Top100DB1*, and structural information gleaned from it will be the primary subject of the next two chapters. Studies of the *Top100DB1* have revealed an abundance of tight, complementary packing throughout proteins, confirmed that proteins are relaxed in the overwhelming majority of circumstances, and contributed to the production of new side-chain rotamer sets. While building this data set, all-atom small-probe contact surfaces proved to be a discriminating structure validation technique—with small-probe contacts you can not get away with much. Packing quality improves with improvements in resolution and *B*-factor, but even atomic-resolution structures often contain mistakes rendered glaring by PROBE. When hydrogens were added to the structures, numerous clashes for Asn, Gln, and His side chains alerted us to the need to find and fix the many which were in 180° "flipped" orientations. The work to reorient these flipped side chains is discussed separately in Chapter 5 and in (Word *et al.* 1999b). This chapter presents the progressive improvement of the *Top100DB1* in general, with lessons learned from these and other structures.†

Choice of Reference Datasets

The data set of 100 protein structures used for these initial investigations (listed in Table 4-1 on page 52) was chosen by resolution, *R*-value, non-homology, and absence of any unusual problems (unusual amino acids, sequence heterogeneity, sequence by X-ray, substantial disordered backbone regions, really large deviations

† Note: Several other people contributed substantially to the work presented here; see (Word *et al.* 1999a) from which this chapter was adapted.

from standard bond geometry, no *B*-factors, etc.). The starting point for the list was the PDB index of January 13, 1997, sorted by resolution; duplicate, homologous, and problem structures were gradually culled, with high resolution as the most important single criterion. All files accepted here were crystallographically determined, with a resolution of 1.7 Å or better, a residual (*R*-value) of 20% or better, and an overall *G*-factor from PRO-CHECK (Laskowski *et al.* 1993) of –0.6 or better. No pair has as high as 30% sequence homology but, more stringently, no more than two examples were included from any known group of related proteins (e.g. only two trypsin-like serine proteases). Mutants were not used if there was a wild-type structure of fairly similar resolution. Packing quality was evaluated only in the results, not in the choice of datasets. Only one copy of identical subunits is included: typically the A subunit, except for 1CPCb, 1EDMb and 1LUCb, where either the authors specified that sub-unit B is preferable or there is a large difference in extent of disordered regions. Also, for ROP protein, Trp repressor, and HIV protease, whose dimer contacts form a large fraction of their cores, a second identical subunit is included as part of the environment to which contacts are calculated, but the atom or residue count is that of the monomer.

If U^2 (atomic displacement) values were reported in place of *B*-factors, they were converted by $B = 8\pi^2 U^2$. To ensure consistent treatment in PROBE and REDUCE, various minor problems with nomenclature or with placement of existing H atoms were corrected in the files. The most common naming problems involved alternate conformations or "het" groups: for example, a residue for which one alternate conformation had an 'a' flag but the other had no flag, or where atom names do not

TABLE 4-1. Top100DB1: very high-resolution, non-redundant protein structures[a]

IDcode	Resol	R%	Refin. program	Tertiary structure	No. of amino acids	Name	Comments	ClSc	Sc
1ETM	0.89	6.5	FMLS/VP	smS	13	Heat-stable enterotoxin	H, aniso	0.0	34
1LKK	1.0	13.3	ShelX	β	105	Tyr kinase SH2	H, aniso	11.4	63
2ERL*	1.0	12.9	ShelX	smS	40	Pheromone ER-1	H, aniso	16.7	41
1BPI*	1.1	14.6	ShelX	smS	58	BPTI	125°K (H refined)	14.8	65
1CNR	1.05	10.5	ProLSQ	smS	46	Crambin, no seq het	H, 150°K	0.0	58
1CTJ*	1.1	13.8	ShelX	smM	89	Cytochrome C6	Aniso (H refined)	8.7	55
1GD*	1.1	19.3	ProL,Xpl	β	61	Protein G		5.7	59
1IRO	1.1	9.0	ShelX	smS	54	Rubredoxin, *Clostr*	H, aniso	16.1	61
1RGEa*	1.15	10.9	ShelX	β	96	RNase SA	H, aniso	2.4	65
1IFC	1.19	16.9	Xpl,TNT	β ud	132	Fatty-acid bdg,intest.	Two conf's	10.4	46
1AMM	1.2	18.5	Restrain	Gkβ	174	Gamma-B-crystallin	150°K	19.8	79
1ARB	1.2	14.9	ProLSQ	Gkβ	268	Achromobacter protease		4.6	73
1CSE	1.2	17.8	EREF	α/β,sm	274,71	Subtilisin/eglin		30.9	56
1JBC*	1.2	11.8	ShelX	Gkβ	237	Concanavalin A	120°K (H refined)	4.8	72
1NOT	1.2	17.8	Xplor	smS	13	G1-alpha conotoxin		0.0	47
1CUS	1.25	15.8	Xplor	α/β	200	Cutinase H (polar)		0.6	65
7RSA*	1.26	15.0	ProLSQ	β	124	RNase A	H,D	0.0	68
1FUS	1.3	18.7	ProLSQ	β	106	RNase F1	(10% to 1rge)	6.1	63
1PTX	1.3	14.8	Xpl,ProL	smS	64	Potent toxin		7.0	66
1RRO	1.3	17.6	ProLSQ	αEF	108	Rat oncomodulin		11.2	55
1AAC*	1.31	15.5	Xplor	Gkβ	105	Amicyanin	(26% to 1plc)	5.1	60
1PLC	1.33	15.0	ProLSQ	Gkβ	99	Plastocyanin H		14.3	56
4PTP	1.34	17.1	ProLSQ	Gkβ	223	B-trypsin/DIFP	(like 1arb)	18.8	58
5P21	1.35	19.6	Xplor	α/β	166	P21 ras		7.8	57

TABLE 4-1. Top100DB1: very high-resolution, non-redundant protein structures[a] (Continued)

IDcode	Resol	R%	Refin. program	Tertiary structure	No. of amino acids	Name	Comments	ClSc	Sc
1BENab	1.4	15.4	ProFFT	smS	21/30	Insulin	H	21.4	44
1RCF*	1.4	13.9	Xpl,ShlX	α/β	169	Flavodoxin, *Anabaena*	H	14.9	65
1SGPi	1.4	17.1	TNT	smS	51	(SGPB)/ovomucoid inhib		30.0	59
1XYZa	1.4	18.3	Xplor	βα8	347	Xylanase		10.9	61
256B*	1.4	16.4	ProLSQ	4hx	106	Cytochrome B562		17.7	54
2CTC*	1.4	16.1	ProLSQ	α/β	307	Carboxypeptidase A		10.4	71
2IHL	1.4	16.5	ProLSQ	α	129	Quail lysozyme		16.9	63
2OLB*	1.4	18.3	ProLSQ	α/β,β	517	Oligo-pept binding prot	123°K	15.9	75
2PHY	1.4	18.6	Xplor	β	125	Photoactive yellow protein		8.7	54
3EBX*	1.4	14.0	ProLSQ	smS	62	Erabutoxin		5.5	51
3SDHa*	1.4	15.9	Xpl,ProL	αHb	146	Clam Hb (homodimer)		11.0	55
bio1RPO	1.4	18.9	Xplor	4hx	(2x)61	ROP protein dimer, mutant		14.3	48
2END	1.45	16.1	ProLSQ	α	138	Endonuclease V		18.7	57
2RN2	1.48	19.5	ProLSQ	β	155	RNase H		17.2	51
1XSOa	1.49	10.4	Xpl,ShlX	Gkβ	150	CuZn SOD, *Xenopus*		2.9	62
8ABP	1.49	17.5	ProLSQ	α/β	306	Arabinose-binding prot		10.7	62
1CKAa*	1.5	17.4	Xplor	β	57	c-crk SH3 domain	113°K, H (polar)	10.6	69
1EDMb	1.5	15.7	Xplor	smS	39	Factor IX EGF		3.6	49
1EZM	1.5	17.6	Xplor	β,α	301	Zn elastase, *Pseudomonas*		5.5	66
1HSUa	1.5	17.3	TNT	smM	62	HiPIP		2.2	56
1LUCb	1.5	18.2	TNT	βα8	324	Luciferase	113°K	11.8	71
1MLA	1.5	18.4	Xplor	α/β,β	309	Malonyl CoA carrier prot		7.0	63
1POA	1.5	14.3	Xpl,PrFF	smS	118	P-lipase A2, cobra		10.8	60
1RIE*	1.5	18.7	Xplor	smM	129	Rieske Fe-S protein	100°K	11.4	64

TABLE 4-1. Top100DB1: very high-resolution, non-redundant protein structures[a] (Continued)

IDcode	Resol	R%	Refin. program	Tertiary structure	No. of amino acids	Name	Comments	ClSc	Sc
1WHI	1.5	18.9	Xplor	β ud	122	L14 ribosomal protein		15.4	48
2MCM*	1.5	16.2	ProLSQ	Gkβ	112	Macromycin		14.7	51
3B5C*	1.5	16.0	ProL,PrFF	4hx	93	Cytochrome B5		10.7	63
2CBA	1.54	15.1	ProFFT	α/β	260	Carbonic anhydrase II		8.3	63
3GRS	1.54	18.6	TNT	α/β	478	Glutathione reductase		16.5	54
1LIT	1.55	18.0	Xplor	β	144	Pancreatic stone inhib.	113°K	17.8	60
1RA9*	1.55	16.9	TNT	α/β	159	DHFR, *E. coli*		18.5	54
1TCA	1.55	15.7	Xplor	α/β	317	Lipase, *Candida*		6.2	64
1HFC	1.56	17.4	Xpl,ProL	β	169	Fibroblast collagenase		10.2	62
1ADS	1.6	20.0	Xplor	βα8	315	Aldose reductase		13.6	64
1ARU	1.6	17.8	Xplor	α	344	Fungal peroxidase		12.7	66
1BKF*	1.6	18.7	Xplor	β	107	FK506-binding protein		13.1	55
1DAD	1.6	18.3	Xplor	α/β	224	Dethiobiotin synth/ADP		9.1	56
1LAM	1.6	17.2	Xplor	α/β	484	Leu aminopeptidase (2 Zn)	123°K	10.2	67
1MCTi*	1.6	16.7	Xplor	smS	28	Squash trypsin inhib	H (polar)	12.6	39
1MRJ*	1.6	17.3	Xplor	~α/β	247	Trichosanthin/adenine	H (polar)	14.3	55
1NFP	1.6	17.5	ProLSQ	~βα8	228	LuxF flavoprotein		14.6	58
1NIF	1.6	17.5	Xplor	Gkβ	340	Nitrite reductase	some H	15.6	52
1PHB	1.6	19.0	ProLSQ	α	414	Cyt P450/camphor		28.8	49
1PTF	1.6	15.6	Xpl,ProL	β	88	His P-carrier		13.2	54
1SMD*	1.6	18.4	ProLSQ	βα8	496	Salivary amylase		15.3	66
1XIC	1.6	15.2	ProLSQ	βα8	388	Xylose isomerase/xylose		5.2	62
2AYH*	1.6	14.3	TNT	Gkβ	214	Beta glucanase		10.1	68
2ER7	1.6	14.2	Restrain	β	330	Endothiapepsin		14.5	63

TABLE 4-1. Top100DB1: very high-resolution, non-redundant protein structures[a] (Continued)

IDcode	Resol	R%	Refin. program	Tertiary structure	No. of amino acids	Name	Comments	ClSc	Sc
2RHE	1.6	14.9	Rst recip-sp	Gkβ	114	Rhe VL dimer		21.3	48
3PTE	1.6	14.8	Xplor	α/β	349	D-Ala transpeptidase		7.8	70
451C*	1.6	18.7	EREF	smM	82	Cyt C551, reduced	(23% to 1etj)	11.8	54
4FGF	1.6	16.1	TNT	tref	149	Fibroblast growth factor		24.0	49
1AKY*	1.63	19.4	Xplor	α/β	221	Adenyl kinase	H (polar)	11.3	53
2CPL	1.63	18.0	Xplor	β	165	Cyclophilin		6.3	64
1KAP	1.64	18.5	Xplor	β hx	479	Alkaline protease/Zn/8 Ca	some H	6.4	65
1CEM	1.65	16.2	Xplor	α hp	363	Cellulase		3.0	72
1CNV	1.65	17.2	Xplor	βα8	299	Concanavalin B	(~ chitinase)	8.0	58
1PHP*	1.65	15.6	Xpl,ProL	α/β	394	P-glycerate kinase		9.8	57
1SNC*	1.65	16.1	ProLSQ	olb	149	Staph nuclease		44.0	35
1SRIa	1.65	17.5	ProL,FFT	β ud	121	Streptavidin/haba		37.1	34
bio2WRP	1.65	18.0	ProL,FFT	α	(2x)107	Trp repressor		18.9	47
1CPCb	1.66	18.1	EREF	αHb	172	Phycocyanin		23.5	49
3CHY*	1.66	15.1	ProFFT	α/β	128	Che Y		19.0	50
2CCYa*	1.67	18.8	ProLSQ	4hx	128	Cytochrome C'	(16% to 256b)	14.0	48
1OSA	1.68	19.4	Xplor	αEF	148	Calmodulin	(27% to 1rro)	9.8	47
2TRXa	1.68	16.5	ProFFT	α/β	108	Thioredoxin		6.1	55
2HFT*	1.69	20.4	ProL,Xpl	Gkβ	218	Tissue factor	(18% to 2rhe)100°K	12.5	54
2MHR*	1.7/1.3	15.8	ProLSQ	4hx	118	Myohemerythrin	H	12.6	56
1DIFab	1.7	19.8	Xplor	β	(2x)99	HIV protease dimer	H	7.6	54
1FNC	1.7	14.9	TNT	α/β,β	314	Ferredoxin reductase		24.8	50
1FXD	1.7	15.7	ProFFT	SmM	58	Ferredoxin II, Fe3S4		8.9	54
1KNB	1.7	15.8	Xplor	β	196	Adenovirus knob domain	H (polar)	8.6	56

TABLE 4-1. Top100DB1: very high-resolution, non-redundant protein structures[a] (Continued)

IDcode	Resol	R%	Refin. program	Tertiary structure	No. of amino acids	Name	Comments	ClSc	Sc
1TTAa	1.7	16.8	ProLsq	Gkβ	127	Transthyretin		19.8	39
2BOPa	1.7	20.1	ProL.,Xpl	β	85	Papil'virus E2 transcrF/DNA		22.9	46
2MSBa*	1.7	17.4	Xplor	β	115	Mannose-binding protein		11.7	63
3LZM	1.7	15.7	TNT	β,α	164	T4 lysozyme		9.6	60

a. Taken from the Protein Data Bank (Bernstein *et al.* 1977) as of January 13, 1997; see text for the selection criteria. The file IDcode is followed by the subunit(s) used; if preceded by bio, the biological dimer of identical subunits was used, generated from crystallographic symmetry. An asterisk (*) means that structure-factor data are available in the PDB.

The resolution is given in Å and the R-value (residual) in %. The refinement programs used in these structure determinations were: X-plor (Brunger), ProLSQ (Konnert, Hendrickson), ShelX (Sheldrick), TNT (Tronrud, Ten Eyck, Matthews), ProFFT (Hendrickson, Konnert, Finzel), EREF (Jack, Levitt), Restrain (Moss), FMLS/VP (Sato).

Abbreviations used for tertiary-structure types are; α, helical; β, sheet; α/β, alpha-beta; αEF, helical E-F hand; 4hx, four-helix bundle; αHb, globin fold; α hp, multihelix-hairpin; βα8, TIM barrel; β ud, up & down β barrel; Gkβ, Greek key β barrel; tref, β trefoil; olb, β oligo-binding fold; SmS, small SS-rich; SmM, small metal-rich.

Comments include whether H atoms were present in the PDB file, the degree of sequence homology when two related proteins are used, and the temperature of data collection if it was noted to be below 200 K.

ClSc is the clashscore (number of atomic overlaps ≥ 0.4 Å per 1000 low-*B* side-chain atoms; computed with **clashlistsc**), after adjustments described in the text; Sc is overall score (contact + 4 Hbond − 10 clash; computed with **scscore**) for low-*B* side chains in the structure.

match those in the PDB het group dictionary. Even in these excellent structures there are rare instances of highly deviant bond angles which were not noticed and fixed by the depositors. Those involving non-hydrogen atoms cannot be addressed without refinement against the experimental data (we rejected files with more than a few of these), but those involving hydrogen atoms we have corrected (e.g. the methylene groups in file 1BEN). Whenever any change was made to a coordinate file, including H addition, it was described in a "USER MOD" record (a standard PDB format type) prepended to the top of the file, and atoms added or changed were flagged beyond column 80.

For the NMR study, we examined three models each, including the minimized average structure if there was one, for a selection of files representing different laboratories and different suites of structure calculation and refinement programs. An excellent set by contact dot criteria is: 1XOB (Jeng *et al.* 1994), 1-2CBH (Kraulis *et al.* 1989), 1CCN and 1CCM (Bonvin *et al.* 1993), 1YUG (Moy *et al.* 1993), 1AGT (Krezel *et al.* 1995), and 1CFE (Fernandez *et al.* 1997).

Methods

Solvent Accessibility

In normal operation, PROBE produces dots that are always at the van der Waals surface of some atom, whereas solvent accessibility is classically measured out on the surface traced by the center of a 1.4 Å radius probe (Lee and Richards 1971; Shrake and Rupley 1973). However, we can obtain an analogous measure of solvent acces-

sibility by asking PROBE to produce dots only where the atom is touched by a 1.4 Å radius probe that intersects no other protein atom. This solvent contact surface (see Chapter 2) may either be displayed, or else counted up (usually per side-chain), divided by dot density, and normalized by a standard side-chain area to give a percentage solvent accessibility. Except for proline, where we used only C^{γ}-*exo* and C^{γ}-*endo* conformations, these standard side-chain areas† were obtained by running the above algorithm for each side-chain rotamer weighted by empirical rotamer occurrence (Dunbrack and Cohen 1997). Solvent contact surface areas are smaller than the traditional accessible surface areas (which PROBE can also determine), since they are measured on the atom surface; however, the relative solvent accessibility percentages obtained by the two methods are quite comparable. We also use solvent contact surfaces to identify buried atoms, both of proteins and of water molecules. The 1.4 Å probe radius finds surface water molecules exposed even when they are in crevices, while it still reports as buried even multiple water molecules in tight cavities, since H-bonding puts a water nearer than 1.4 Å to some of its neighbor protein atoms.

Proline Pucker

For analysis of proline ring pucker, test proline residues were substituted using either the C^{γ}-*endo* or the C^{γ}-*exo* geometry given by Némethy *et al.* (1992), leaving the backbone unchanged and either re-using C^{δ}, or else placing C^{δ} in the plane of

† Standard side-chain solvent contact surface areas in $Å^2$: Ala, 15.6; Arg, 47.1; Asn, 29.5; Asp, 29.2; Cys, 30.0; Gln, 35.6; Glu, 35.7; Gly, 4.4; His, 58.2; Ile, 37.6; Leu, 36.2; Lys, 38.9; Met, 41.0; Phe, 51.8; Pro, 24.6; Ser, 19.6; Thr, 28.0; Trp, 66.6; Tyr, 51.9; Val, 31.4.

the peptide, whichever produced the least distortion between the idealized ring and the pre-existing backbone. The *X*–Pro peptide bond length was not adjusted. Replacements were tried only for proline residues that showed significant clashes in their original conformation.

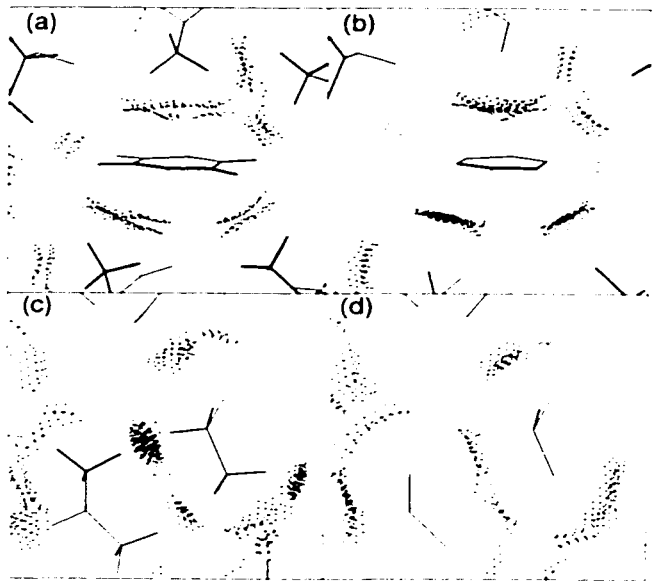
All statistical tests were performed with the STATA-5 package for Macintosh, STATA Corporation, College Station, Texas. Plots were made with Microsoft EXCEL and modified in Adobe ILLUSTRATOR.

Good Packing

The first obvious result from calculating contact surfaces is that these protein structures show impressively well-fitted packing interactions. Side-chain atoms touch their neighbors all around, and the hydrogen atoms interdigitate neatly (e.g. Figure 2-2). Even methyl groups, with a relatively low barrier to rotation, are amazingly relaxed inside proteins, nearly all of them fitting excellently in staggered conformation (e.g. Figure 4-1(a)); they do *not* have room to spin freely. The more accurately determined the structure, the better the packing, and the best of the currently available structures are shown by this independent and highly sensitive analysis to be beautifully accurate. Using standard parameters (see Chapter 2), a very large fraction of atom contacts are found to lie within about ± 0.2 Å of exactly touching (that is, of being separated by the sum of their radii). Significantly disallowed overlaps are nearly absent, most peripheral atoms make contact, and many are optimally positioned between two, three, or more good contacts. In a well-

packed core region, it is rare that a torsion angle can be rotated much in either direction without producing clashes.

FIGURE 4-1. (a) and (b) Comparison of contact dots calculated with explicit *versus* implicit ("united atom") hydrogen atoms, for an Ala/Phe interaction in 3LZM. The explicit dots in (a) show that the Ala methyl is positioned almost perfectly for contact with the Phe ring, while the other contacts shown by implicit dots are much sparser and rounder, in spite of the fact that the refinement had not used explicit H. (c) and (d) Comparison of explicit-H *versus* implicit-H contact dots for a Leu/Val interaction in the theoretical model of a designed protein, 1SSR. The implicit-H contacts in (d) look just as good as the real ones did in T4 lysozyme (b), but with explicit H dots the clashes are shown to be abnormally large.



Most proteins seem to contain some regions with very tight packing and other regions that are sparser, as was also seen earlier using volume criteria (Richards 1977). For example, in T4 lysozyme the region around Trp126 and Arg95 is very tightly packed (Figure 2-2(a)) while a core region near Leu99 is sparse.

Explicit Hydrogen Atoms

These elaborately well-fitted protein cores are observed only when two conditions are met simultaneously: the structure must have been determined with very great accuracy and all H atoms must be represented explicitly in the subsequent analysis. If a "united atom" *implicit* hydrogen representation is used, contacts generally occur at good distances, but they are sparser, broader, smoother, and very much less sensitive to either rotations or displacements. Model structures built with tools using implicit H atoms, for purposes such as completely de novo designs (e.g. PDB file 1SSR) look just as good as experimentally determined protein structures when contacts are calculated with implicit hydrogen atoms (Figure 4-1(d) *versus* (b)). However, if full explicit hydrogen atoms are added geometrically to those designed models, they are seen to have numerous bad clashes (Figure 4-1(c)), whereas if H atoms are added geometrically to high-resolution X-ray structures the contact surfaces fit very well (Figure 4-1(a)), even if hydrogen atoms were not used in the refinement. Since such designed models often look good by most criteria (such as sequence-structure "threading"; e.g. Bowie, Luthy, and Eisenberg 1991) but are much less well ordered than natural proteins when they are actually produced (e.g. Betz, Raleigh, and DeGrado 1993; Richardson *et al.* 1992), protein design requires more stringent model-evaluation criteria such as the explicit H contact dots.

Since rather few sets of design model coordinates have been publicly deposited and methods are not always described in complete detail, it is not possible to make a systematic comparative analysis. However, there are a few PDB files for models built from scratch using all hydrogen contacts: 2SLK, for example (Fossey *et al.*

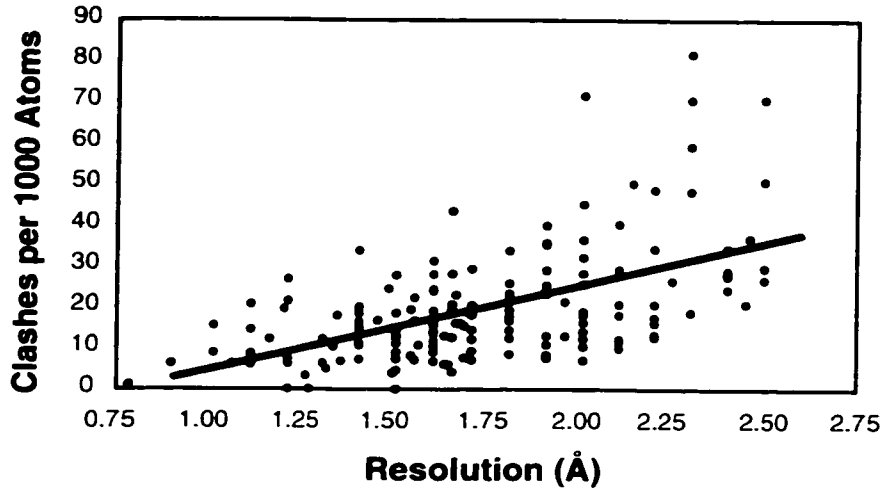
1991), shows excellent contact dots, although unfortunately there is no detailed experimental structure of silk form I from which to judge the correctness of its details. Some of the most successful recent protein redesigns (e.g. Dahiyat and Mayo 1997a; Desjarlais and Handel 1995; Struthers, Cheng, and Imperiali 1996) were modeled using explicit H contacts between (although not within) residues; some cases have achieved well-ordered structures in which only limited regions depart significantly in conformation from the design. Another factor undoubtedly contributing to their success is that although many side-chains were redesigned, the backbone was kept precisely as in a particular known protein rather than built *de novo*; this reduces the likelihood of inadvertently pointing hydrogen atoms at each other in impossible orientations like those in Figure 4-1(c), or of choosing unfavorable backbone geometries to connect secondary structures.

High Resolution

For the *Top100DB1* set of reference proteins, and for other examples as well, the visual appearance of the contact dots, the absence of serious clashes, the density of favorable contacts, and the overall packing evaluation score (see Chapter 2) are all generally related to resolution. For instance, Figure 4-2 plots the number of serious clashes (overlap ≥ 0.4 Å) per 1000 atoms as a function of resolution, for the *Top100DB1* proteins and 80 additional proteins at medium resolution. Although the scatter is high, there is a significantly positive slope for the clash *versus* resolution regression line, as measured by an F test ($p < 0.001$). For the sort of fine detail shown by the contact dots, this relationship indicates that structural accuracy is still

improving noticeably down near 1 Å resolution. However, the number of clashes per 1000 atoms is not significantly related to protein size.

FIGURE 4-2. The number of serious atom clashes (overlap > 0.4 Å) per 1000 atoms, for each of the *Top100DB1* proteins plus 80 more at medium resolution, plotted versus resolution in Å; these values include the high *B*-factor and flipped-amide clashes. Although the scatter is high, the regression line shows a significant relationship.



The very best scoring structures we have found are those at extremely high resolution (around 1.3 Å or higher) which also incorporated in their refinement either calculation of full-radius van der Waals interactions for explicit H atoms, e.g. the 1JBC concanavalin A (Parkin, Rupp, and Hope 1996) or the 1RGE ribonuclease SA (Sevcik *et al.* 1996) using SHELX (Sheldrick and Schneider 1997), or for the 1CNR crambin (Yamano and Teeter 1994) using PROLSQ, or else hydrogen data from neutron diffraction, e.g. for the 7RSA ribonuclease A at 1.26 Å (Wlodawer *et al.* 1988). Figure 4-3 shows just the significantly overlapping (> 0.25 Å) van der Waals contacts for all conformers of the entire ribonuclease A molecule of file 7RSA. If

one uses only conformation 'a' where there are alternate conformations, there are only a few atomic overlaps that make it past this threshold, but not a single severe clash ≥ 0.4 Å even on the outside of the molecule where one might expect less order.

FIGURE 4-3. C^{α} backbone, plus spikes for all overlaps > 0.25 Å, for the 7RSA ribonuclease A (Wlodawer *et al.* 1988). When only non-alternate and 'a' alternate conformations are included, this is the cleanest of all the *Top100DB1* structures, with only a few small overlaps that barely reach this level. However, three severe clashes can be seen, each of which involves a residue in the 'b' alternate conformation.

Small-probe contacts may prove useful during the refitting of lower resolution structures. See Jim Kiefer's comments in Chapter 7 regarding his experiences with a 3 Å cyclooxygenase-2 structure.

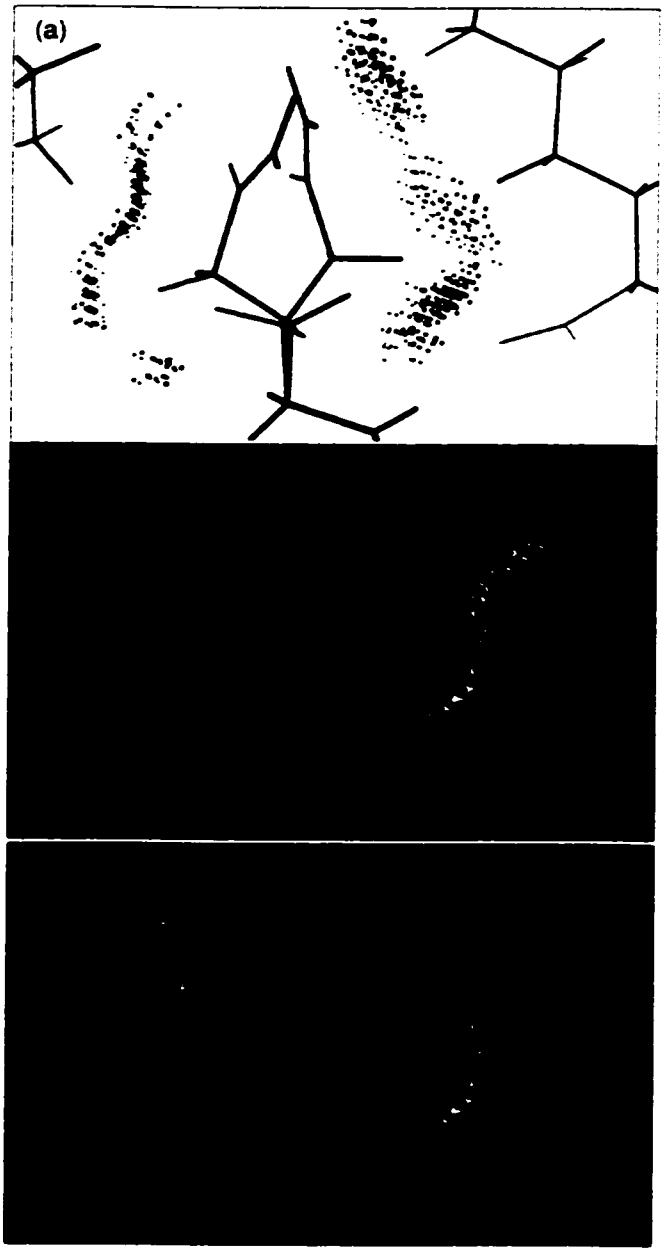
Alternate Conformations and B-factors

In Figure 4-3, if one considers conformation 'b' rather than conformation 'a', then there are three very severe clashes, with overlaps of 0.5 Å to 0.7 Å, making the point that even in such a highly accurate structure 'b' conformations are prone to errors. Of the three clashes, one is very easily correctable: it only requires defining 'a' and 'b' conformations for the OH hydrogen of Thr100, which must necessarily move out of the way, to let O^{γ} be an H-bond acceptor, when Lys98 swings into its 'b' position. It is, therefore, an example of a clash caused by a sub-threshold, unidentified disorder in a neighboring residue; understandably, these occur fairly often and at times require insight and experience to resolve. The second clash, between Lys104b and Ala102, cannot be resolved without examination of the electron density and re-refinement. The third serious clash, of Gln11b, is the most interesting. The two conformations of Gln11 both have good geometry and favorable χ angles. As can be seen in Figure 4-4(a), they are well defined and separated for most of their length, and they form favorable, well-fitted contacts against opposite sides of the over-large cavity left by the surrounding structure. The perpendicular view of Figure 4-4(b) shows that the clashes are between the NH_2 group of Gln11b and the low-*B* side-chain of Leu35, suggesting that the problem is due to an incorrect 180° flip of the amide group. Figure 4-4(c) shows the contact dots after exchanging the N and O atoms of Gln11b, with the clashes cured and new favorable contacts, including a possible weak H-bond to His12 at the ribonuclease active site, which could have implications for its titration behavior. All the other side-chain amide groups are correctly oriented in file 7RSA, because it incorporated earlier neutron-diffrac-

tion data allowing direct visualization of hydrogen and deuterium atoms. However, Gln11b is an especially difficult case since the occupancy is only 0.33 and the potential amide H positions overlap those for Gln11a, intermingling the electron density.

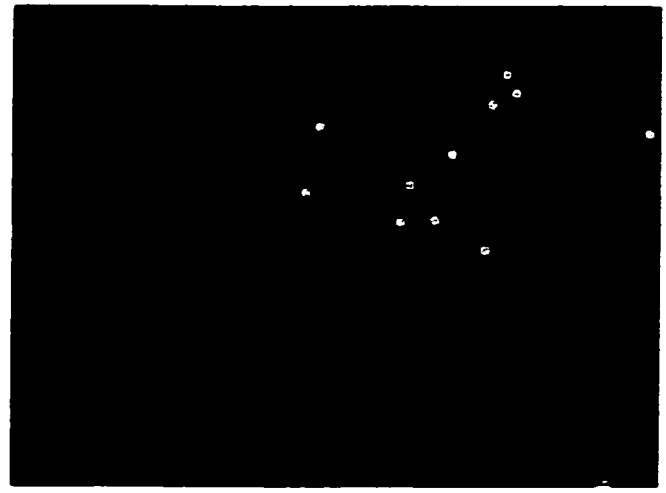
It should not be especially surprising that 'b' conformations are prone to errors, since they should have an occupancy of 0.5 or less and seldom have completely well-separated electron density. The 'a' alternate conformations share those difficulties, but to a less severe level. The problems of 'b' conformations have been exacerbated by the fact that the geometry-checking programs in common use do not, in their default mode, look at 'b' conformations. This is unfortunate because 'b' conformations need the extra information of geometrical constraints even more than the rest of the structure, since they are less well determined by the experimental data. PROBE allows the analysis of 'b' conformations. However, because 'b' conformations are much more likely to be problematic, for our further analyses we look only at 'a' alternate conformations.

FIGURE 4-4. (a) Front view of 7RSA Gln11, with 'a' and 'b' alternate conformations. Each has favorable χ angles, and they hug opposite sides of the available space. (b) Side view of 7RSA Gln11b, showing that it clashes with low *B*-factor atoms of the adjacent Leu side-chain. (c) After flip of the amide, contacts are greatly improved, including a weak H-bond to the His N^{ε} on the left (pale green dots).



Other high-resolution structures, e.g. the 1XSO superoxide dismutase at 1.5 Å resolution (Carugo *et al.* 1996) or the 3LZM wild-type T4 lysozyme at 1.7 Å (Matsumura *et al.* 1989), refined without the use of explicit hydrogen bumps, can show equally excellent packing throughout the interior, especially if they were carefully examined and refit by hand, but they almost always have some bad clashes in regions of high temperature factors (*B*-factors) on the outside. For comparison with the excellent internal packing of Figure 2-2(a) and Figure 4-1(a) for T4 lysozyme, Figure 4-5 shows a surface region with high *B*-factors where the large, red clash overlaps clearly represent physically impossible relative atom positions. In other words, if the positions of the heavier atoms are determined with high enough accuracy, then geometrically added hydrogen atoms will indeed show good packing, but that is obviously not true if the heavy-atom positions are less accurate: for high *B*-factors, 'b' alternate conformations, and lower resolutions. Even the otherwise respectable level of 2 Å resolution is marginal for showing packing details.

FIGURE 4-6. A surface region of 3LZM, colored by *B*-factor. The less mobile interior parts have almost no significant overlaps, while the high *B*-factor loops at the top (in yellow) show a number of physically impossible clashes.

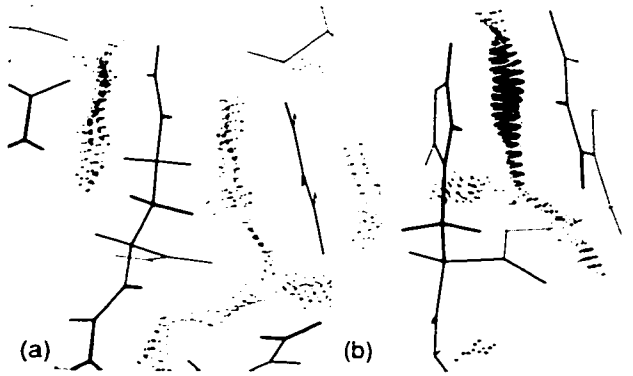


NMR Structures

Similarly, the interior regions show good packing in the very best determined NMR structures: those with many NOE (nuclear Overhauser effect) constraints per residue, stereospecific assignments, and suitable refinement protocols. The identification of close contacts between explicit individual hydrogen atoms is an integral part of NMR structure determination, and so NMR is capable in the best cases of representing packing very accurately (see Figure 4-6(a)). However, other regions of those same structures nearly always show bad clashes (e.g. Figure 4-6(b)), often on the surface where there are fewer NOEs and perhaps disorder as well. Among the ensemble of models calculated for a given NMR structure, including minimized average models, the specific clashes usually differ, but their distributions are similar. It should be possible to eliminate most such clashes by including full-radius van

der Waals terms as lower distance limits for nearby atom pairs in the final stages of refinement. Although that cannot guarantee correct atom positions in the absence of enough experimental constraints, it should help substantially for borderline cases.

FIGURE 4-6. (a) Dot contacts around an interior Gln sidechain in the cellobiohydrolase NMR structure 1CBH (Kraulis *et al.* 1989), showing excellent fit. (b) A Phe-His clash on the outside of the same structure.

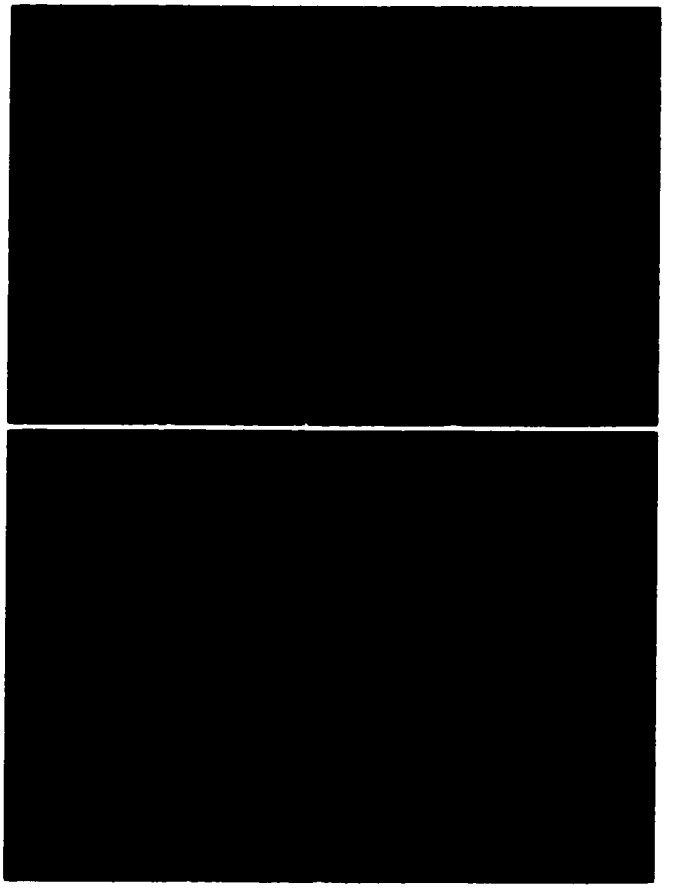


Nucleic Acid Structures

Crystal structures of small oligonucleotides solved at high resolution, e.g. 3DNB (Prive, Yanagi, and Dickerson 1991), 284D (Salisbury *et al.* 1997), or 244D (Laughlan *et al.* 1994) at 1.1-1.5 Å resolution, almost always show excellent packing throughout, as do canonical *B*-form or *A*-form double helices, e.g. 1OSU (Wahl, Rao, and Sundaralingam 1996), 7BNA (Holbrook, Dickerson, and Kim 1985), or 2BOP (Hegde *et al.* 1992) at 1.4-1.9 Å, whose conformations have been very thoroughly characterized. However, the crystal structures of large DNA or RNA molecules have usually been determined only to resolutions in the range of 2 to 3 Å. Small-probe contact dots calculated for such structures show an interesting and

revealing pattern, as shown in Figure 4-7, which compares a *B*-DNA structure with a less regular large RNA. The packing between the bases is beautiful; their flat shape and few degrees of freedom allow very accurate positioning. However, for structures with non-canonical conformations, e.g. 299D (Scott *et al.* 1996), 4TNA (Hingerty, Brown, and Jack 1978), and 1YTF (Tan *et al.* 1996) at 2.5-3.0 Å and even for Z-form DNA in 131D (Bancroft *et al.* 1994) or 1D53 (Kumar *et al.* 1992) at 1.0-1.5 Å, there are usually serious clashes along the backbone, which has very few observable atoms per degree of freedom. For such structures, the determination of backbone conformation would presumably be improved significantly by the incorporation of explicit H atoms and their van der Waals repulsions in the refinement and/or by the diagnostic use of contact dots. In this regard, we are pleased to see that Foloppe and Mackerell (2000) are developing an all-atom force field with specific attention to nucleic acid backbone.

FIGURE 4-7. Small-probe dots for (a) the regular *B*-DNA double helix in 7BNA (Holbrook, Dickerson, and Kim 1985), showing excellent contacts throughout. (b) Part of the large hammerhead RNA in 299D (Scott *et al.* 1996), showing excellent contact for the bases but severe clashes for the H atoms along the backbone.



Gene Wickham explored the conformational constraints on the cleavage site phosphate of the genomic hepatitis delta virus (HDV) ribozyme with our analysis tools (Wickham and Word 1999). Working from a 2.3 Å crystal structure of the 3' RNA cleavage product (1CX0; Ferré-D'Amaré, Zhou, and Doudna 1998), Wickham used MAGE's construction tools to build a phosphate group attached to the 5' hydroxyl, added hydrogens with REDUCE, and varied the conformation of the new phosphate

and the bridging O5' both manually in MAGE and in an automated search with autobondrot (described in Chapter 8). A map of contacts between the articulated phosphate and static atoms in the crystal structure, for a full range of P–O5', O5'–C5', C5'–C4' dihedral angles, showed that only a restricted set of conformations (all of which reposition O5') were without large clashes. The goal of this work was to gain insight into the structure of the uncleaved ribozyme. The contact analysis indicates that unless the overall structure deviates somewhat, then the ribose-phosphate backbone must thread through this narrow conformational opening, restricting the modeling possibilities significantly.

Misfolded Structure Database

Our contact surface methods were unable to distinguish the real structures from the decoys in a database of crystal structures and deliberately misfolded structures—where the same sequence was threaded onto a backbone from another structure of the same length—developed by Holm and Sander (1992) to test their solvation free energy model. The failure to discriminate based on packing considerations is explained by observing that a number of the real structures in this database are of poor quality (6 have more than 50 serious clashes/1000 atoms) while the misfolded models have been minimized to improve their packing. The result is that contacts for both sets of structures look equally bad.

Progressive Improvement of the Reference Datasets

Our long-term goal in developing this method is to study the distribution and significance of favorable packing interactions, in order to understand their possible role in structural uniqueness. A pre-condition for such studies, however, is to assemble a set of reference structures with all explicit hydrogen atoms and completely free of any large, physically unrealistic atomic clashes, at least in their interiors. That process has turned out to be surprisingly complex, interesting in its own right, and helped us refine our methods. It includes three components: (1) choice of the reference proteins (explained above) and appropriate exclusion of locally disordered parts; (2) optimization of strategies for the addition and placement of explicit H atoms; and (3) a quite conservative and limited set of corrections to the coordinates or assignments in the original files. All changes are documented in the headers of those modified coordinate files.

The starting basis set of 100 reference proteins incorporates nomenclature corrections, geometrical addition of H atoms not originally present, and rotational optimization of the new OH hydrogen groups one at a time; only the first alternate conformation was used, others were ignored. At this stage there were large interior clashes in seven of the files, which involved pre-existing OH hydrogen atoms. These turned out to be an easily corrected artifact: apparently, during much of the time when these files were deposited, the widely-used X-PLOR refinement program (Brunger 1992) had a bug that systematically placed OH hydrogen groups toward rather than away from neighboring donor H atoms. Although that problem has now been corrected, for consistency we routinely strip out, recalculate, and rotate any

pre-existing OH groups. The first two entries in Table 4-2 give the average clash score (number of clash overlaps ≥ 0.4 Å per 1000 atoms) for the 100 reference proteins before and after all the OH hydrogen atoms have been rotationally optimized one at a time.

TABLE 4-2. Progressive improvement of *Top10DB1* PDB structures

	Average clashscore	Files affected (%)
Original PDB	19.1	
Rotate OH		7
	18.8	
Omit B ≥ 40 atoms		80
	14.5	
Rotate Met-CH_3		36
	14.1	
Flip side-chain amide groups		66
	12.5	

B-factor Cutoff

The next obvious step is to exclude interactions for which one or both atoms have high temperature factors. Although *B*-factors are not entirely equivalent between different refinement protocols, the regions with very high *B*-factors are always prone to problems (see Figure 4-5). These problems are usually due either to choice of a poor geometry within a region of very low and spread-out electron density, or else to correctly following an average density which has impossible geometry because of the way it averages multiple alternative conformations. Occasionally, the high *B*-factor is a result of incorrect local fitting rather than simply reflecting diffuse electron density. Few of these situations are correctable without re-refine-

ment and many would require additional data, so the only reasonable strategy for our present purposes is to ignore clashes with high *B*-factor atoms.

FIGURE 4-8. Bars plot clashes per 1000 atoms as a function of *B*-factor, showing that high *B*-factor atoms are enormously more likely to clash with their neighbors. Also plotted is the cumulative percentage of atoms included below a *B*-factor cutoff at that value. Rejecting all atoms with $B > 40$ loses only 5% of the atoms.

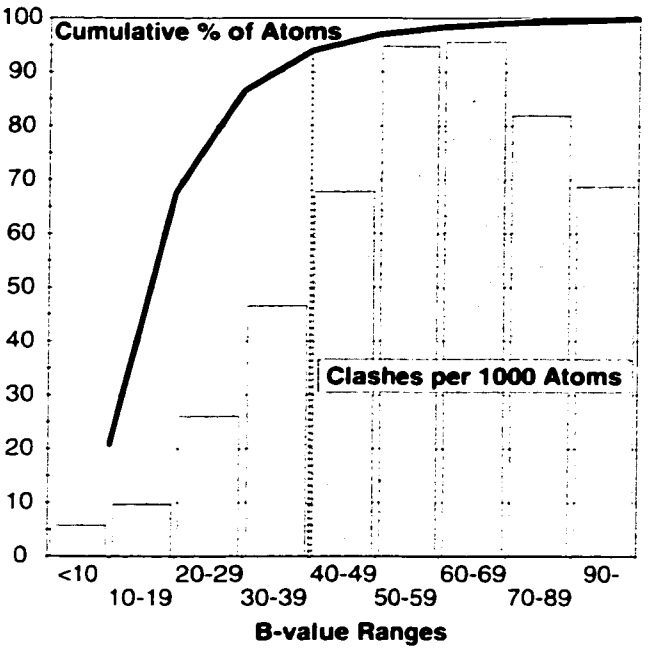


Figure 4-8 plots the number of severe clashes per 1000 atoms *versus* the *B*-factor range; atoms with $B > 50$ are about ten times as likely to have severe clashes as atoms with *B*-factors of 10 to 20. The fraction of clashes falls off again for the very highest *B* ranges, since most of those atoms are out where they have almost no neighbors. Strictly speaking, the reported *B*-factors are not directly comparable between structures, due to variations in data reduction, solvent treatment, estimates of intensity fall-off, and application of either global or local *B*-restraints. It is possi-

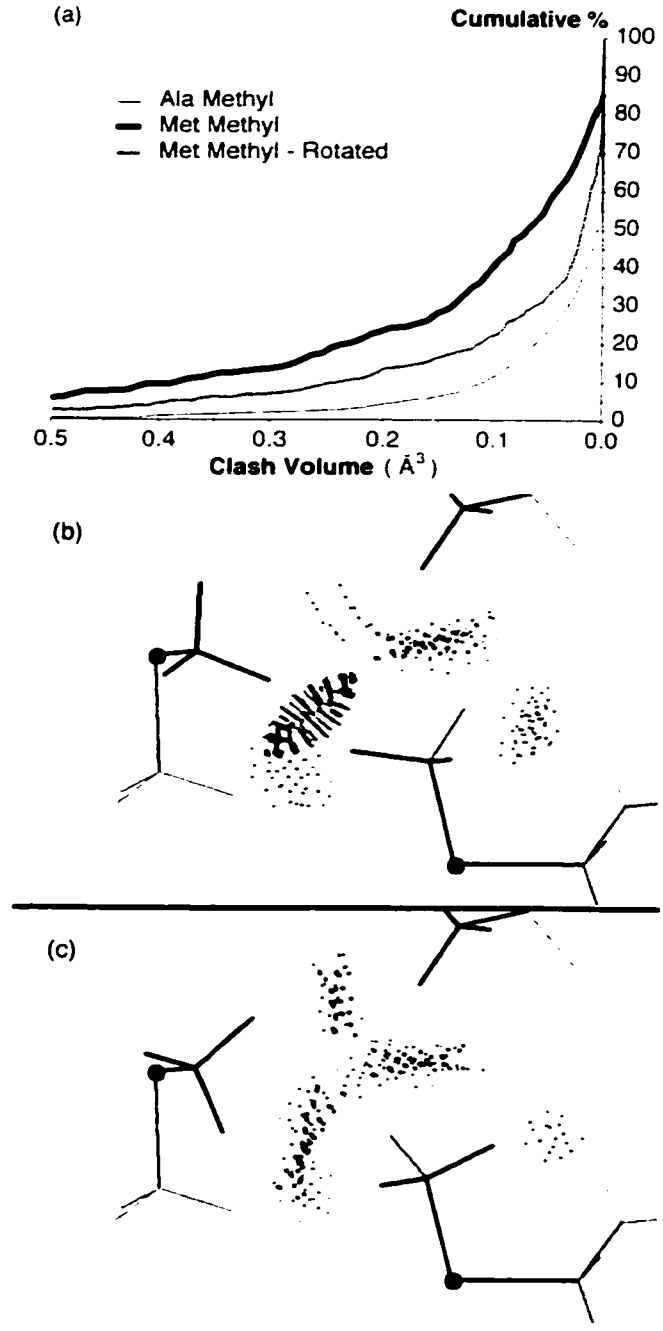
ble to partially compensate for such differences by normalizing *B*-factors by the mean and standard deviation in each structure (Carugo and Argos 1997; MacArthur and Thornton 1999). However, we feel it is preferable to use the simpler absolute values, for two reasons: (1) a high *B*-factor will smear out calculated electron density, unless artificially re-sharpened, no matter what its origin, and (2) differences in the actual level of molecular disorder are often larger than the methodological effects. For example, many atomic-resolution structures have no disordered loops and, therefore, have no high *B*-factors.

B-factors for an individual side chain may be high for various reasons, including thermal motion, static disorder, or phase problems. One of the most important reasons, however, is the possibility of a side-chain misfitting. Refinement of a misfit side chain can either move the atom back into density or increase the *B*-factors, depending on the details of the local environment and the weights of the *B*-factor restraints. Whatever the cause, the conformation of a high *B*-factor side chain is less reliable than for a low *B*-factor equivalent and should not be included in an analysis that depends on accurate details. $B < 40$ was chosen as a conservative cutoff criterion, which keeps approximately 95% of the atoms while rejecting those whose clashes are most likely to be artifacts of misfitting or mobility. The third score entry in Table 4-2 shows the average score improvement obtained by considering only atoms with $B < 40$, for all 100 proteins of the reference data set. The *B*-factor cutoff makes the largest average improvement of any of the steps described here.

Methionine Methyl Groups

After removal of the high *B*-value atoms, the next set of clashes to stand out were a specific subset of the methyl H atoms added in staggered conformation: overwhelmingly, the terminal methyl groups of methionine side-chains. In hindsight, this distinctive behavior of Met methyl rotations seems very reasonable, since they have a much lower barrier to rotation (perhaps 1 kcal/mol *versus* 3 kcal/mol). In file 7RSA, which incorporates neutron diffraction data that can directly locate H atoms, the Met methyl groups are found as much as 36° away from staggered, while 16° is the largest rotation found for the more numerous Ala methyl groups.

FIGURE 4-9. (a) Cumulative distribution of methyl clash volumes, showing that the terminal methyl groups of Met are very much more likely to have large clashes than the β methyl groups of Ala. (b) Two interacting Met methyl groups from squash trypsin inhibitor of file 1MCT chain I (Huang, Liu, and Tang 1993); staggered configuration, with severe clash; (c) both methyl groups rotated, with good contacts.


(a)
Cumulative %
100
90
80
70
60
50
40
30
20
10
0
Ala Methyl
Met Methyl
Met Methyl - Rotated
0.5
0.4
0.3
0.2
0.1
0.0
Clash Volume (Å^3)
(b)
(c)

To further document this difference in packing seen for Met versus other methyl groups, Figure 4-9(a) compares cumulative distributions of the clash volumes found for all of the Met *versus* all of the Ala side-chain methyl groups. Consequently, the Met methyl groups (and only the Met methyl groups) were rotationally optimized to eliminate atomic overlaps, by an initial search at 30° intervals, followed by a 1° search around the best of those positions. Since only one type of contact is involved (without the donor-acceptor ambiguity of H-bonding), the simple algorithm is well behaved even if two such methyl groups can touch one another, which happened for eight cases in the dataset. Figure 4-9(b) and (c) show the dramatic improvement for an interacting pair of Met methyl groups in the inhibitor of PDB file 1MCT, before and after optimization. The differences between the third and fourth score entries in Table 4-2 show the clashscore improvement obtained just by rotational optimization of Met methyl groups, which makes a quite substantial difference for some of the files.

Of course, the real surprise is how seldom any rotation is needed to achieve good packing around other methyl groups. In lists of remaining clashes, there are a fair number of serious intra-residue clashes that might be relieved by methyl rotation: the commonest are between the two branches of an Ile, or between a Leu methyl group and its local backbone. However, these are no more common or more severe than intra-residue clashes involving methylene atoms (usually found in long side-chains, especially Lys), which could only be fixed by moving heavy atoms. For the Leu H^{δ}-H^{α} clashes, often C^{δ} is also too close to H^{α}, so that methyl rotations alone would not be sufficient. There are highly populated non-clashing rotamers only 10-

15° away in χ, and these Leu self-clashes are significantly less common at the highest resolutions and lowest *B*-factors. Similarly, all of the serious Ile H^{δ}-H^{γ} clashes have the two carbon atoms overlapping by 0.5 Å or more, so that methyl rotations alone cannot fix them (in contrast to the situation for Met methyl groups). Only rarely does a non-Met methyl group need rotation against sequentially distant atoms. In summary, therefore, although some perturbations of methyl groups from staggered orientation must undoubtedly occur, the occasional serious clashes seen for non-Met methyl groups seem predominantly due to mispositioning of the methyl carbon atom, rather than due to a need for large rotations of the methyl hydrogen atoms. In the current version of REDUCE, we have chosen to leave all non-Met methyl groups in the staggered position, since the addition of so many degrees of freedom is hard to justify by the small improvement attainable.

The final set of dataset changes documented in Table 4-2 involves full optimization of local H-bond networks, considering the movable-H groups on side-chains, N termini, and heterogen groups, including rotation of OH, SH, NH_3^+, and Met methyl groups, side-chain amide flip for Asn and Gln, and ring flip and protonation state for His. That process is the main subject of (Word *et al.* 1999b) and is described in detail there and in chapter 5. Each of the three main modifications to the dataset summarized in Table 4-2 (exclusion of high *B*-factor atoms, rotation of Met methyl groups, and optimization of H-bond networks) results in a very significant ($p \leq 0.001$ or better) improvement in the clash scores, as measured by a "paired-*t*" test of differences in means. In Table 4-1, the final clashscore is listed for each of the 100

files and also the standard combined score (including contacts as well as clashes and H-bonds), normalized by surface area (see Chapter 2).

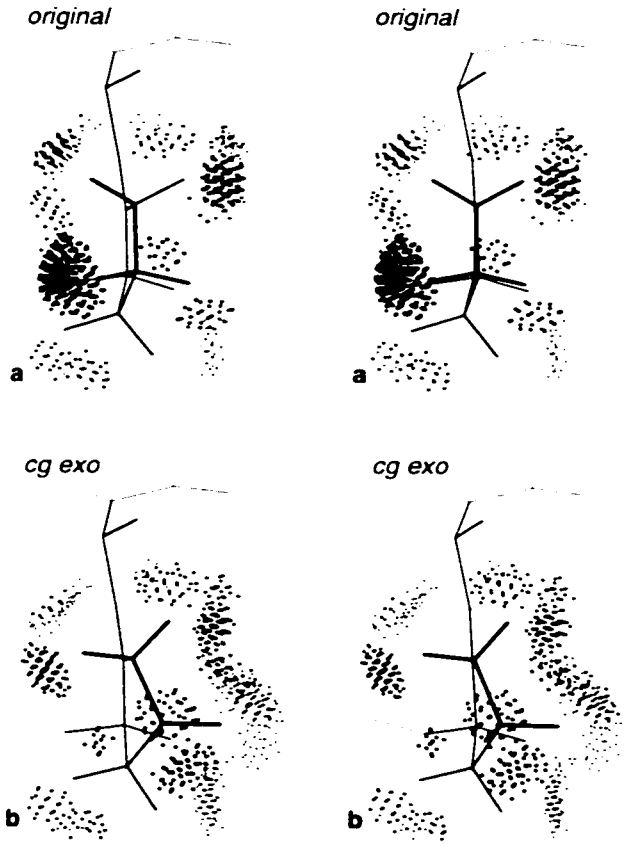
Proline Pucker

Of the remaining clashes, an interesting set involves bumps of Pro side-chains either with the preceding residue or with sequentially distant residues. Since at this resolution identifying *cis versus trans* isomers cannot be a problem except in disordered regions, the most likely difficulty is assignment of Pro ring pucker. Many refinement programs allow three, five, or more states of Pro pucker, sometimes even allowing flat rings. However, Némethy *et al.* (1992) have argued very convincingly, from a survey of highly accurate small-molecule crystal structures, that proline residues can actually adopt only two pucker states: C^{γ}-*endo* or C^{γ}-*exo*. Fortunately those two states have a nearly planar C^{δ}-N-C^{α}-C^{β} dihedral angle, so that it is possible to switch proline pucker as a local change with fairly minimal effect on backbone geometry.

For a sample of 12 proline residues with serious clashes, we tried substituting either a C^{γ}-*endo* or a C^{γ}-*exo* ring in standard geometry (see "Proline Pucker" on page 58). Even with no other adjustable parameters, all but two of them showed significantly improved packing, judged visually and by contact-dot scores. Figure 4-10 shows the most dramatic example, for a completely buried Pro in 1EZM that initially had a modest C^{β}-*endo* pucker and three bad clashes; with standard C^{γ}-*exo* geometry, not only do the clashes disappear, but much new favorable contact is formed. For that

side-chain, the normalized score improved from –4.8 to +106.6; the average score improvement was 41.6. We have not actually altered any of the proline residues in our database files, since those changes would move non-H atoms relative to the electron density. However, the success of such simple replacements argues strongly that restriction to only C^{γ}-*endo* or C^{γ}-*exo* ring pucker would improve refinement of proline residues. When the electron density for a Pro ring appears flat, it might best be fit as a mixture of C^{γ}-*exo* and *endo* conformations.

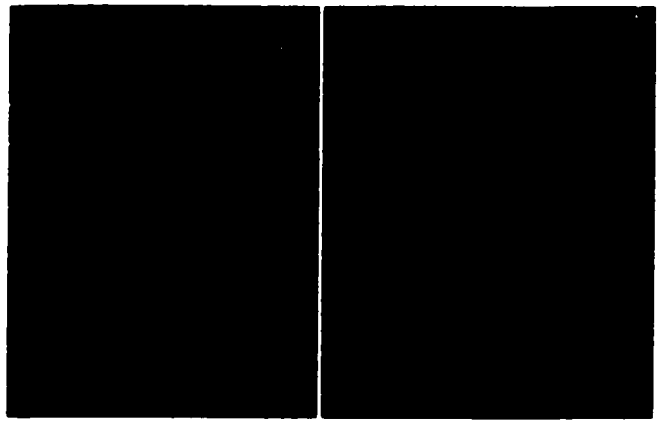
FIGURE 4-10. Pro131 from the Zn elastase 1EZM (Thayer, Flaherty, and McKay 1991), with contact dots (stereo). (a) Original configuration with three serious clashes; (b) ring in C^{γ}-*exo* conformation, with excellent contacts.



Glycine Clashes

Serious clashes involving $1H^{\alpha}$ or $2H^{\alpha}$ of glycine residues are approximately three times as common per H^{α} atom as for any other amino acid. This should not be surprising, since the absence of an observable C^{β} makes ϕ,ψ angles considerably less accurate in glycine residues (see Richardson 1981). One example is the contact of Gly67 with Trp74 in *Escherichia coli* dihydrofolate reductase, for file 4DFR at 1.7 Å resolution and for file 1RA9 at 1.55 Å resolution. In the former structure, the Gly $1H^{\alpha}$ and the Trp ring atoms clash by 0.6 Å and $2H^{\alpha}$ is turned away (Figure 4-11(a)), while in the latter structure both H^{α} atoms contact the ring favorably (Figure 4-11(b)). This improvement results from rotation of the 67-68 peptide and improved planarity of the 65-66 peptide, bringing the ϕ,ψ angles of the glycine from an unfavorable –45°, 73° to a favorable –72°, 143°, in the common polyproline II conformation. In this comparison, the correction presumably came about because higher-resolution data showed the positions of backbone atoms more accurately. The high *B*-factor of Gly67 in file 4DFR (~65) is probably a symptom rather than a cause of the incorrect conformation, since the *B*-factor is only 20 in 1RA9. It seems likely that even at intermediate resolutions many such errors in glycine conformation could be corrected by refinement with hydrogen van der Waals terms.

FIGURE 4-11. (a) Clash of high *B*-factor Gly67 $1H^{\alpha}$ with the ring of Trp74, in *E. coli* DHFR (file 4DFR at 1.7 Å resolution; Bolin *et al.* 1982). (b) Contact of low *B*-factor Gly67 $1H^{\alpha}$ and $2H^{\alpha}$ with the ring of Trp74, in *E. coli* DHFR (file 1RA9 at 1.55 Å resolution; Sawaya and Kraut 1997).



Conclusions

Probably the most important general conclusion from the analysis in this chapter is that explicit hydrogen atoms and their contacts are crucial to detailed and specific interactions between and within molecules. Certainly no analysis of packing inside proteins or of ligand binding can afford to omit them.

The technique of small-probe contact dots demonstrably makes available new information that was not being used. It applies very simple geometrical analysis to explicit hydrogen atoms and van der Waals contacts, and then makes the effects directly visible, either to observation or to quantitative analysis. Because it is exceedingly sensitive and because it concentrates on aspects largely orthogonal to the terms used in most refinement or modeling calculations, it can act as a general-purpose "mine canary" to detect any of a wide range of problems. It is nearly

impossible to do anything wrong inside a protein structure without it showing up in clashes of the contact dots. An especially valuable aspect is that quite often an examination of the local pattern of dot contacts can actually suggest how to fix the problem.

These analyses emphasize the truly revolutionary accuracy and detail attainable in the new wave of protein crystallographic structures at atomic resolution (Dauter, Lamzin, and Wilson 1997). Explicit H atoms were not used in the refinement of more than a handful of the 100 reference structures, yet the other atom positions are so well determined that the implied hydrogen atoms fit in place beautifully. The contact dots and the very high-resolution structures validate one another: the uniformly high contact scores and relatively clash-free interiors, especially in the higher half of our resolution range (obtained without shifting the position of a single non-H atom), demonstrate both that those structures are nearly error free and that our analysis is looking at details that are real. The development and validation of this method depended on the existence of those structures and could not have been done even five years ago. However, even in the low *B*-factor regions of these excellent structures, there remain a very small number of severe clashes (occasionally also signalled by bad bond lengths and angles), caused either by problems in refinement of a small-molecule bound "heterogen" or by a side-chain trapped in the wrong conformational minimum. Those might be fixed by trial refittings based on both contact-dot and electron-density examination, combined with further refinement against the structure-factor data. Also, now that the radius and scoring parameters have been fairly well optimized using these 100 high-resolution protein

structures, such analysis can also be applied to NMR structure refinement, to the improvement of crystallographic structures at more conventional resolutions, to nucleic acid structures, and to theoretical modeling of structures. Including these additional geometrical restraints is quite analogous to inclusion of bond length and angle terms.

For clash-free regions with all hydrogen atoms added and optimized, the favorable terms in the contact dot interactions can then be used in a different way: to understand and interpret structural features seen in an individual protein or empirical regularities found in comparing structures. For example, it was shown that rotation is often needed for Met methyl groups to achieve good packing, while it is almost never clearly justified for the equilibrium position of other side-chain methyl groups. This gives us a further insight into the surprising extent to which nearly all conformational details are cooperatively relaxed in protein interiors.

Although the algorithms and parameters for the small-probe contact surface method have been carefully chosen, tested, and tuned, they will undoubtedly continue to change and improve. We have started with a highly simplified approach and have added complications only when forced to do so. The contact dots and their scores have forced us to deal with Asn/Gln flips, H-bond networks, *B*-factors, Met methyl and NH_3^+ rotations, H-bonds to ring faces, Pro pucker, and contacts with bound heterogen groups. However, it has proven feasible and even advantageous to keep a simple water model, to avoid most methyl rotations, to ignore CH···O H-bonds, and to use simple exhaustive searches for optimizing H-bond networks. In the future,

smaller radii are probably needed for interactions of atoms separated by few covalent bonds, and we plan to include mid-range electrostatically based effects by atom type for distances between contact and a probe-diameter separation. In general, we will be pursuing the question of how best to incorporate these insights, and probably the methods themselves, into established protocols for energy calculations and structure refinement, as well as protein redesign and *de novo* design. There will also be future advances from the rapidly growing data base of very high-resolution structures.

CHAPTER 5

Hydrogen Atom Contacts in the Choice of Side-chain Amide Orientation

Asn/Gln/His Side-chain Orientation

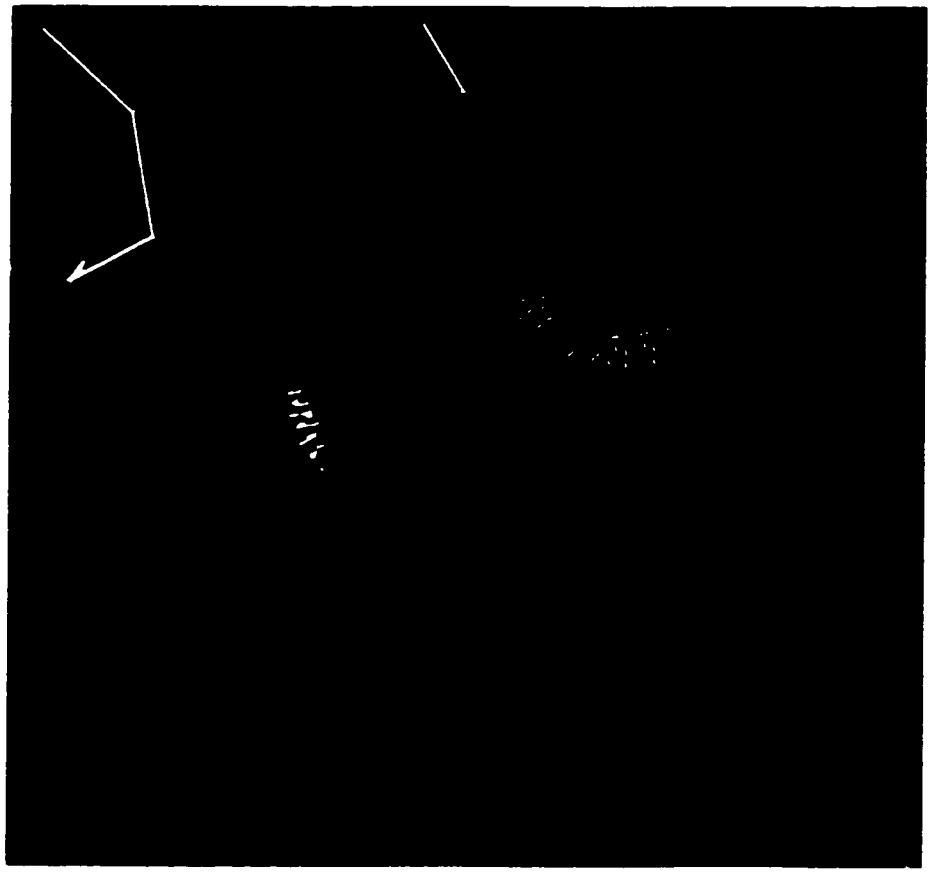
Correct assignments of the NH_2 *versus* the O branches of Asn and Gln side-chain amide groups are a relatively minor part of a protein structure determination, but they can be quite important if the residue is involved in H-bonding at the active site, or if one wants to analyze bound water molecules, H-bond networks, detailed electrostatics, or dynamics simulations. However, such assignments have been considered difficult, and sometimes are not even attempted. The Protein Data Bank (PDB, Berman *et al.* 2000; Bernstein *et al.* 1977) format actually provides a special ambiguous atom designator of A, meaning "either N or O", for use by crystallographers who want to keep the uncertainty explicit. Finding the NH_2 hydrogen atoms or telling apart the N and O atoms by direct observation in the electron-density map is not possible except at extremely high resolution. The distinction, therefore, is almost always made on the indirect evidence of H-bonding possibilities, usually done by inspection as part of the model-fitting process.

The most secure assignments are when the environment includes H-bonding groups that are either obligate donors such as a peptide NH group (which can interact only with the O branch of the amide) or obligate acceptors such as a carbonyl group (which can interact only with the NH_2 branch). In the frequent cases where the surrounding groups are ambiguous donors or acceptors (OH, His, other amide groups, or water molecules), assignment involves analyzing the entire local network of H-bonds. The energy terms in refinement protocols are capable, in principle, of favoring the amide orientation with the best H-bonding, but in practice the 180° flip between the two orientations is a large enough step that minimization will always, and molecular dynamics sometimes, be trapped in one of the local minima. Even after full analysis, many H-bond networks have two equally favorable solutions that involve concerted exchange of all donors and acceptors, many Asn or Gln amide groups are undetermined because they are exposed at the surface, and some make only non-polar interactions. Histidine residues also have a similar assignment problem, since a 180° flip of the imidazole ring exchanges N and C atoms in the δ and ε positions. In place of the amide ambiguities of H-bond donor *versus* acceptor, for His the choice is a more drastic one between a polar, or even charged, NH and a CH with only very weak H-bonding potential; however, imidazole orientation can also be ambiguous.

Several automated methods have been developed to help deal with this problem. HBPLUS (McDonald and Thornton 1994) tries the flip states of each Asn, Gln, and His, and chooses the alternative that minimizes unsatisfied buried H-bonding groups, dividing the prior Asn/Gln/His orientations into highly favored, slightly

favored, indifferent, slightly suspect, and highly suspect; however, it does not deal with pairs or larger interacting groups. NETWORK (Bass *et al.* 1992) analyzes H-bond networks to optimize polar H placement, but does not allow for amide or imidazole flips. WHATIF (Hooft, Sander, and Vriend 1996) deals with both aspects of the problem, including even the assignment of H positions for all crystallographically located water molecules with occupancy > 0.5; it builds in crystal symmetry, and has a penalty bias against flips in marginal cases. Inclusion of the water hydrogen atoms makes the combinatorial problem so huge that it cannot possibly be treated exhaustively, so it is done by a variant of simulated annealing. WHATIF does a thorough and careful job of analyzing the H-bond networks, coming out with a decision for all the ambiguous polar groups; using it would improve assignments for the majority of structures. This feature is just one part of an overall package with many other valuable functionalities. Its disadvantages for amide assignment are that it relies strongly on the positions of water molecules, which are the least reliable feature in macromolecular structures; its output is not convenient, and its answers cannot be critically evaluated because there are no estimates of confidence and the reasons for its choices are well hidden inside a complex, stochastic process.

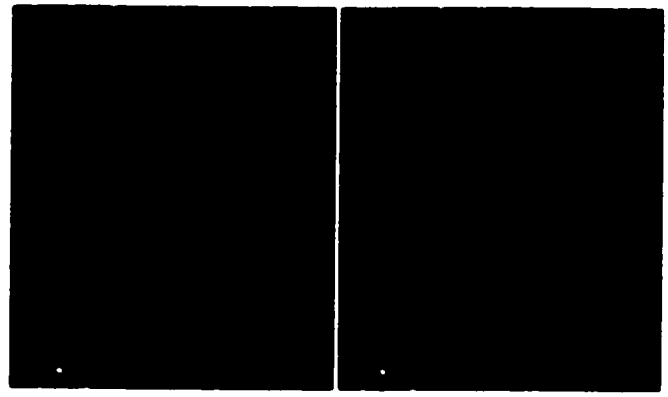
FIGURE 5-1. Small-probe contact dots around Gln71 of cutinase (1CUS), colored by contact gap and including favorable van der Waals contacts (green and blue dots) as well as H-bonds (pale green dots), slight overlaps (short yellow spikes), and clashes (orange or red spikes; none present).



We revisited this problem because our small-probe contact dot methodology uncovered a source of independent new information from the analysis of van der Waals clashes for explicit H atoms, making these decisions less complex and much less subtle. The reasons for a given choice can easily be expressed both as numerical scores and in a visual display that explicitly shows all relevant positive and negative interactions (e.g. Figure 5-1), so that the user can easily evaluate confidence levels for a given choice. The method is applied here to optimizing the H-bond networks

and assigning Asn, Gln, and His flips in a set of very high-resolution crystal structures.

FIGURE 5-2. (a) and (b) Amide flip comparison for Gln90 from the immunoglobulin V_L dimer of 1REI (Epp *et al.* 1975), colored by atom type (O, red; N, blue; C, white) and with clashes emphasized by spikes. There are three H-bonds and no clashes in the correct flip position (a) *versus* no H-bonds and three serious clashes of the NH_2 hydrogen atoms in the incorrect flip position (b).

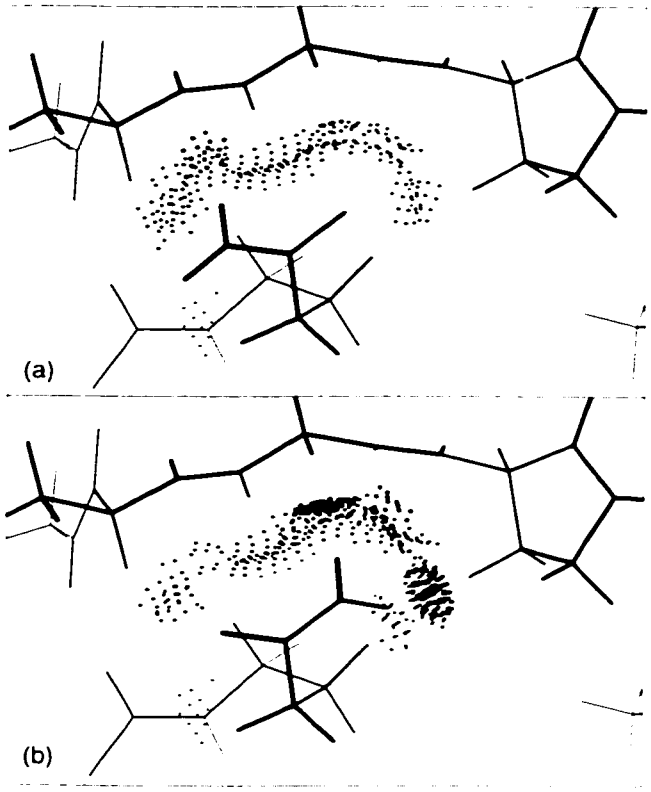


Individual Asn/Gln Examples

The basic point of this new approach is that the van der Waals interactions of polar H atoms are crucial to ruling out incorrect amide orientations, even if they are not necessary for evaluating the energy of correct hydrogen bonding. The NH hydrogen extends 0.6 Å farther than the bare oxygen, which alters the van der Waals interactions so drastically that these amide flip choices generally become blatantly evident rather than subtle, once hydrogen atoms are added by REDUCE (see Chapter 3) and contacts are scored and examined with small-probe dots generated by PROBE (see Chapter 2). Figure 5-2 shows one of the many really obvious cases, which has excellent H-bonding in the correct orientation and extreme van der Waals clashes in

the incorrect orientation; the un-normalized score comparison is +0.3 *versus* −7.0. This Gln could be assigned correctly by any person or computer program explicitly considering it, since the H-bonding is completely unambiguous. However, some cases this obvious were found to be misassigned in our reference dataset, including some in proteins refined using molecular dynamics and even a few in structures at atomic resolution. This particular example, Gln90 from the immunoglobulin V_L dimer 1REI at 2.0 Å resolution, used "A" atom designations to leave the amide assignment explicitly ambiguous.

FIGURE 5-3. (a) and (b) Amide flip comparison for Gln57 in the 1FUS ribonuclease F1 (Vassylyev *et al.* 1993), which has no H-bonding but whose orientation is unambiguously determined by the NH_2 clashes in the position of (b), mainly with the H^{β} of a Pro side chain. In contrast, it fits well against the backbone in (a).



Many other cases cannot be determined by the H-bonds alone, but are unambiguous if the van der Waals interactions are included. Figure 5-3 shows an example (from ribonuclease F1) which has no H-bonds, but where the non-polar interactions accommodate one amide orientation very nicely but not the other. In one flip state the NH_2 group of Gln57 nestles neatly against the backbone, while in the other flip state it collides with a proline side-chain. The score comparison is +1.0 *versus* −1.4.

FIGURE 5-4. (a) and (b) A double amide flip of the Asn128-Gln34 pair in the fungal peroxidase 1ARU (Fukuyama *et al.* 1995) that cannot be resolved just by H-bonding. In the incorrect double flip (b), there is a very bad clash of the Gln NH_2 with H^{α} and a smaller clash of the Asn NH_2, whereas the amide flip state shown in (a) is accommodated well. There is also a shear offset between the amide groups that puts the two NH groups at a further, and more favorable, distance in (a).The contact dots are simplified by showing only the H-bonds and the overlaps, not the attractive van der Waals contacts.

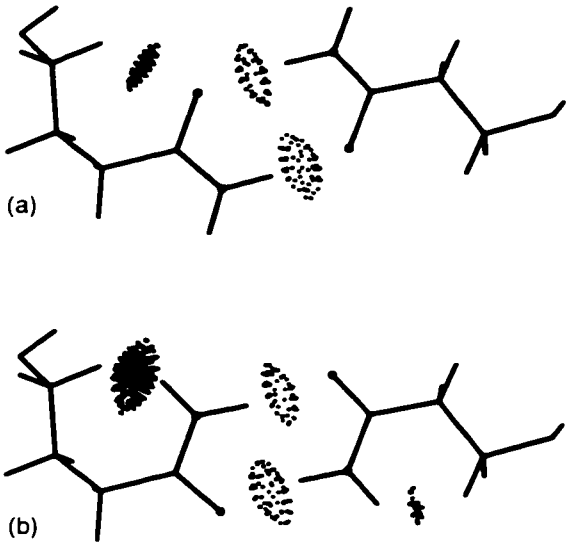


Figure 5-4 illustrates a pair of doubly H-bonded side-chain amide groups from the fungal peroxidase 1ARU, for which two of the four possible orientations are equally good if only H-bonding is considered. Such situations are rather common, either for pairs or for larger H-bond networks, in which switching all donors and acceptors in unison can produce equally good H-bonding. However, as in Figure 5-4, van der Waals clashes usually rule out one of the two best H-bond possibilities: in this case, the original assignment in the coordinate file has good H-bonds but bad clashes of both amide NH_2 groups with their own Ha atoms, while flipping both amide groups gives very much better van der Waals contacts and equally good H-bonding. The score comparison is –0.4 *versus* +3.6 for the original and double-flip

states, respectively (compared with –5.0 and –6.7 for the two single-flip states). A slight shear offset between the two side-chains is visible in Figure 5-4, which puts the two polar H atoms further apart (2.5 Å) in the better orientation. That does not necessarily mean that the amide H radius needs to be larger than 1.0 Å, however, because the two polar H atoms are presumably shifted apart by electrostatic as well as van der Waals repulsion. Unfortunately, our data-set cannot calibrate the consistency of such a shear offset, because it happens that each of the three other doubly H-bonded Asn/Gln pairs has one of the amide groups incorrectly oriented by our criteria, which seems to have caused refinement to slightly distort the interactions.

Systematic Surveys

The *Top100DB1* reference database of 100 proteins contains 1554 unique Asn and Gln residues, 1539 of which have no missing atoms, metal ligation, or covalent modifications. To provide a cross-check on this new methodology (and to identify unusual situations that could be handled by improving the algorithm), the Asn and Gln residues were systematically surveyed three times, each time by a different combination of contact score comparison (see Chapter 2 for more about scores) and inspection of their small-probe contact dots in the MAGE display program. The first time through, I examined each Asn or Gln with $B < 40$ if its individual flipped score was not clearly worse than its original score; I assessed H-bond interactions between multiple residues visually. The above process resulted in flipping 252 Asn/Gln residues (17% of the total) in 71 of the files. For several files the Asn/Gln flip rate was approximately random (near 50%), implying that the amide groups had not

been examined or that the ambiguous A atom designations were used (as in 451C: Matsuura, Takano, and Dickerson 1982).

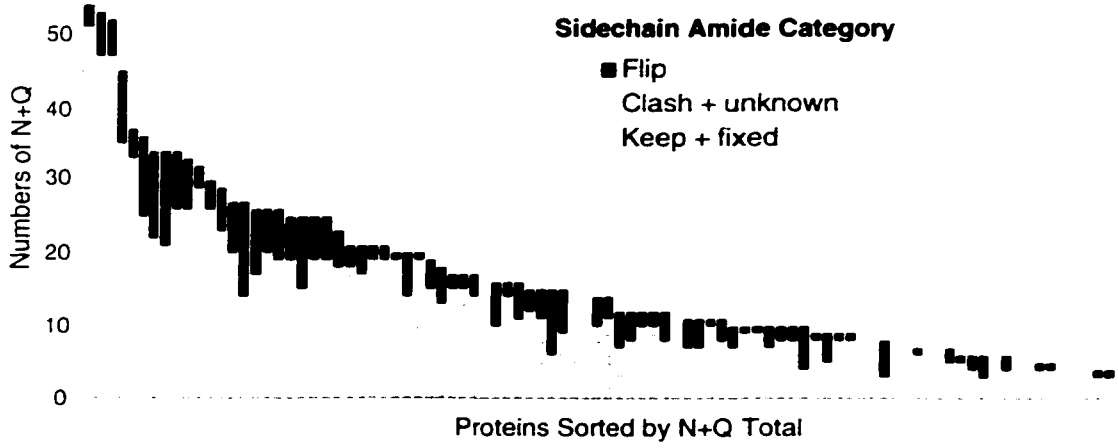
For the second-round survey, an algorithm was developed to analyze full H-bond networks automatically, as part of the REDUCE program. The orientations of 47 His, 17 Cys, ten OH, nine Asn, and three Gln residues are fixed by metal ligation, and three Asn and several Cys, Ser, and Tyr residues by various other covalent modifications; the Asn/Gln modifications are listed in Table 5-1 starting on page 105. This leaves 6548 unique movable-H groups, including Asn, Gln and His residues, OH, SH, NH3+ and Met methyl groups, and similar functionalities in the small-molecule "heterogens" (see Chapter 4 for an explanation of why methyl groups are rotated only in Met side-chains). Those groups were then partitioned into closed sets of local interacting networks or *cliques* (see Chapter 3). Of the movable groups, 5050 were found to be isolated from any other; there are 557 interacting pairs, 94 triples, 14 cliques of four, eight cliques of five, one clique of six, and no larger groupings.

The clique score (using both favorable H-bonds and unfavorable overlaps, including a simple model for interaction with the crystallographic water molecules, but not non-overlapped contacts) is evaluated for all combinations of possible H atom positions, in order to choose the optimal arrangement. Most movable groups have either two, four, or six potential H atom positions, while we restrict an OH or SH to something between two and as many as 18 in the most crowded environments (see chapter 3). Since the largest clique found had only six members, an exhaustive

search is computationally tractable: it takes three hours to do all 100 proteins on an R10000 SGI Indigo2.

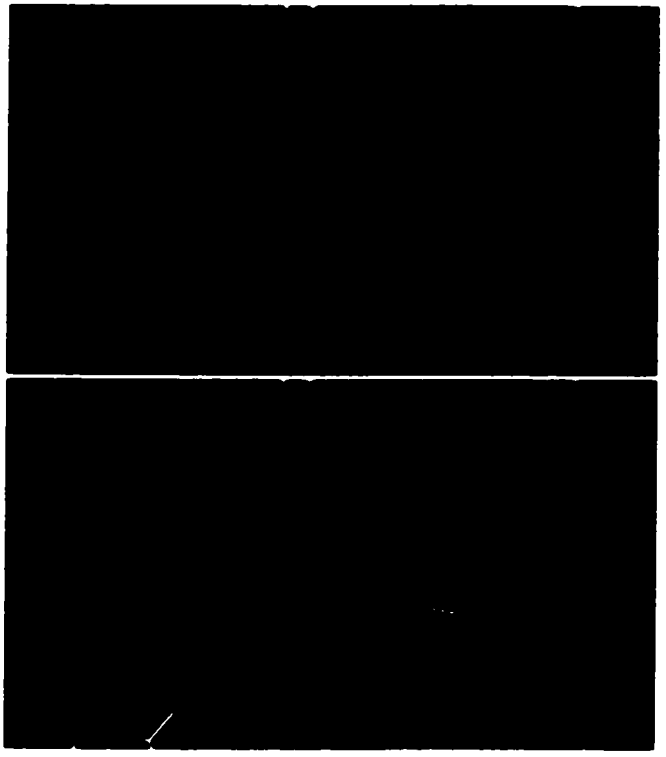
For each residue in a clique, the best total clique score and the best conformation are reported plus, for Asn/Gln/His, the best total clique score found with that residue in the opposite flip state. This score comparison directly shows how sensitive, or well-determined, the flip state is for that specific residue. The H atom positions for the best-scoring arrangement are added to the output PDB format coordinate file, and N *versus* O or N *versus* C identities are switched where indicated for Asn/Gln/His flips. Agreement with decisions made in the first-round survey was used to optimize the value of a penalty against changing the depositor-assigned flip state: the penalty was set at 0.5, which means that the score difference must favor the flipped state by at least 0.5 or REDUCE would not assign a flip. Any flip state for which an atom in the movable group has a serious clash (overlap ≥ 0.4 Å) is flagged with a `!`. If both the best state and the best flip state have clashes, then non-H atoms must be badly placed and *B*-factors are usually high on at least one side of the clash. If it is not practical to evaluate these cases individually, then they should be omitted from any further analysis. For the automated algorithm, therefore, we have adopted the conservative policy of not assigning any flips for these double-clash cases.

FIGURE 5-5. Categories of side-chain amide assignments for each of the *Top100DB1* proteins, sorted in decreasing order of total Asn + Gln residues. Here, the "*Keep + fixed*" category includes the few fixed by covalent modifications, and the "*Clash + unknown*" category includes both the "*C*" (double-clash) and "*X*" (low-score difference) groups. N stands for Asn, Q for Gln.



The result of the automated analysis of round 2 is 100 coordinate files with all H atoms added and optimized, and all changes and assignments documented in the file headers, plus contact dot kinemages which animate between the two flip states, one set for Asn/Gln and another set for His side-chains. Out of the 1554 Asn + Gln side-chains, REDUCE flipped 290, or 19%, of them (see Figure 5-5). The rest were all left in their original orientation, including 49 with bad clashes both ways (3%) and 314 with small score differences between –0.5 and +0.5. Out of 379 His side-chains, 30 were flipped (8%), 27 had bad clashes both ways (7%), and 49 had small score differences. Only 17 Asn, 29 Gln, and 12 His (3% of the total) residues are so completely exposed that they had scores of exactly zero in both orientations.

FIGURE 5-6. (a) Correct *versus* (b) flipped arrangements of the Asn138-His123-His131 H-bonding network from the 1CEM cellulase (Alzari, Souchon, and Dominguez 1996). All four H-bonds are equivalent in the two forms, which are distinguished by clashes of the Asn NH_2 group in (b). van der Waals contact dots are not shown.



As an example of REDUCE's automated analysis of an H-bond network, Figure 5-6 shows the two principal alternatives for the linear Asn138-His123- His131 clique in cellulase (correctly assigned in the PDB file 1CEM); there are two states for the

Because this is a new method, and because we want these modified coordinate files to be a reliable basis for future analyses, we undertook a third-round survey in which both flip alternatives and their contact dots were visually examined in MAGE for all of the Asn/Gln/His residues that the automated algorithm had flipped, for all of the small number of cases where REDUCE (round 2) disagreed with the round 1 assignments, and for all cases in a subset of 20 files. Out of 290 amide flips recommended by REDUCE, only 15 were rejected in round 3 (5% of the flips, or 1% of all Asn/Gln amides groups), of which 12 were declared ambiguous and only three as clear "*Keep*" residues.

During this process, it became obvious that many of the double-clash cases and the marginal cases with low score differences could actually also be assigned an orientation with confidence. Therefore, round 3 was expanded to include visual examination of all Asn/Gln/His in those two categories. Most of the resulting reassignments involve confirming the orientation indicated by the flip scores: for example, Gln61 of 1MLA, malonyl CoA carrier protein (Serre *et al.* 1995), with a score difference of 0.48 was promoted from marginal to flipped, because it can make weak H-bonds both to backbone and to a Glu O^{ε} in the preferable flipped orientation. However, there are a small but significant number of cases where the visual assignment contradicts the direction of the score difference. Sometimes these involve factors that are not yet included in the algorithm but which could be (such as the influence of a charged group slightly too far away to score as an H-bond), while sometimes the analysis involves judging the relative probability of different types of errors in ways it is hard to imagine automating. As an example of the latter sort, Gln260 of 1ARU

(fungal peroxidase) has a score difference of −1.0 *versus* +0.1 and a bad clash of the NH_2 with C^{δ} of Glu325 in the original orientation, yet our judgment agrees unambiguously with the original crystallographic assignment: the low *B*-factor Gln260 in its flipped orientation has only a water H-bond and O^{ε} is very close to two other oxygen atoms, while the original orientation adds a backbone CO H-bond and improves contacts, and a minor rotation of the high *B*-factor Glu325 could convert the clash into an H-bond with the Glu O^{ε}. Such cases should emphasize the point that although the automatic algorithm does a very good job, the most reliable assignments combine the automated analysis with visual inspection.

Making Animated "Flip-kins"

In order to obtain a visual comparison of the alternative Asn/Gln/His flip states, I developed the Unix scripts FlipNQkin and FlipHISkin which use REDUCE output to produce a kinemage that uses animation to permit switching between the two alternative arrangements. These scripts run REDUCE a second time, using the file header information from the previous run to arrange for the optimization of cliques (OH rotations, etc.) with each flippable group fixed in the non-preferred orientation. The kinemage also contains pre-calculated views scaled and centered on each Asn and Gln or His. After examining the animations in MAGE, the user can decide whether to reject any of the automated assignments.

Summary of Final Assignments

The third-round survey results in five categories of Asn/Gln residues summarized in Table 5-1. First is the small subset whose orientation is fixed by metal ligation or covalent modification. The largest category by far (65% of the total) are the 1006 Asn/Gln, with a clear, unambiguous flip assignment that agrees with their identification in the deposited PDB file. Then there are the side-chain amide groups which unambiguously require flipping: 318 Asn/Gln residues (20.5%) in 73 of the 100 files. For 20 of the Asn/Gln side-chains (1%) whose movable group has a severe clash in both orientations, either they or a neighboring group is positioned incorrectly in such an arrangement that it is unclear which would be the correct amide orientation once the problem was fixed. The fifth, final category are 195 still-ambiguous Asn/Gln cases (12.5%), mostly with small score differences ($-0.5 \le D < 0.5$). All examples in the two ambiguous categories are left in their original orientation.

For each of the 100 proteins, sorted by total Asn + Gln residues, Figure 5-5 plots the number of "*Keep + fixed*" Asn/Gln (in medium gray), the number of "*Clash + unknown*" (in light gray), and the number of "*Flip*" Asn/Gln (in dark gray). Twenty-seven files had no flips and four files had 50% or more flips, but overall the distribution is relatively uniform. Somewhat surprisingly, the percentage of amide flips shows no significant relation with resolution. However, the percentage of flips does show a small but significant ($p < 0.017$) positive relation to the size of the protein, perhaps reflecting time limitations for evaluating correct amide orientation as protein size increases.

TABLE 5-1. Round 3 side-chain amide flips of Asn and Gln

PDB	N + Q[a]	Fixed[b]	Keep	Clash[c]	Unk.[d]	Flip
1aac	3		3			
1ads	27		15		5	7
1aky	20		14		5	1
1amm	16		9	1		6
1arb	26		16		1	9
1aru	26	1 CHO	15		4	6
1benAB	6		4		1	1
1bkf	6		2		2	2
1bpi	4		4			
1cem	37		24		9	4
1cka	4		2		1	1
1cnr	3		3			
1cnv	25		15		4	6
1cpcB	12	1 CH3	3	1	2	5
1cse	33	2 Ca	21		3	7
1ctj	10		6		1	3
1cus	15		9		3	3
1dad	18		9		4	5
1dif	9		5		3	1
1edmB	5	1 Ca	4			
1etm	1		1			
1ezm	27		9	1	4	13
1fnc	17		12		3	2
1fus	14		11		3	
1fxd	3		3			
1hfc	14		8	1	1	4
1ifc	12		8			4
1igd	4		3			1
1iro	4		4			
1isuA	7		6		1	
1jbc	19	1 Ca	12	1	1	4
1kap	53	1 Ca	44		2	6
1knb	20		10		4	6
1lam	36		24		1	11

TABLE 5-1. Round 3 side-chain amide flips of Asn and Gln (Continued)

PDB	N + Q[a]	Fixed[b]	Keep	Clash[c]	Unk.[d]	Flip
1lit	12		6		4	2
1lkk	10		5	1	3	1
1lucB	34		19		3	12
1mctf	0					
1mla	25		12		3	10
1mrj	29		19		4	6
1nfp	30		18	1	7	4
1nif	21		17		1	3
1not	1		1			
1osa	11	2 Ca	4		1	4
1phb	34		12	1	8	13
1php	15		8		3	4
1plc	7		7			
1poa	14		9		2	3
1ptf	6		3			3
1ptx	5		4			1
1ra9	9		4		1	4
1rcf	17		13		2	2
1rgeA	7		5	1		1
1rie	6		5		1	
1rro	11	1 Ca	4		2	4
1sgpl	5		3		1	1
1smd	52	1 Ca	42		4	5
1snc	10		6		3	1
1sriA	10		6		1	3
1tca	32	1 CHO	28			3
1ttaA	3		2		1	
1whi	6		4			2
1xic	21		14	1	2	4
1xsoA	12		9	1		2
1xyzA	45		27	1	7	10
256bA	12		8			4
2ayh	23		15	1	2	5
2bopA	11	1 Yb	5	2	2	1

TABLE 5-1. Round 3 side-chain amide flips of Asn and Gln (Continued)

PDB	N + Q[a]	Fixed[b]	Keep	Clash[c]	Unk.[d]	Flip
2cba	21		16		3	2
2ccyA	8		6		2	
2cpl	12		12			
2ctc	25		16	2	1	6
2end	9		8			1
2er7	15		4		2	9
2erl	3		1		2	
2hft	21		16		3	2
2ihl	16		12		2	2
2mcm	7		6		1	
2mhr	6		5		1	
2msbA	10	2 Ca	6			2
2olb	54		46	1	4	3
2phy	11		8			3
2rhe	8		7		1	
2rn2	15		8	1		6
2trxA	7		7			
3b5c	5		5			
3chy	10		8			2
3ebx	7		4		1	2
3grs	25		15	1	3	6
3lzm	17		7		7	3
3pte	34		23		3	8
3sdhA	16		11			5
451c	8		2		1	5
4fgf	9		8			1
4ptp	26		13		6	7
5p21	15		7		8	
7rsa	17		14		3	
8abp	20		17		2	1
bio1rpo	5		3		2	
bio2wrp	10		3		1	6
Totals	1554	15	1006	20	195	318

a. Total number of Asn + Gln.

b. Number of amide orientations fixed by covalent modifications: by metals (Ca or Yb), carbohydrates (CHO), or methylation (CH_3).

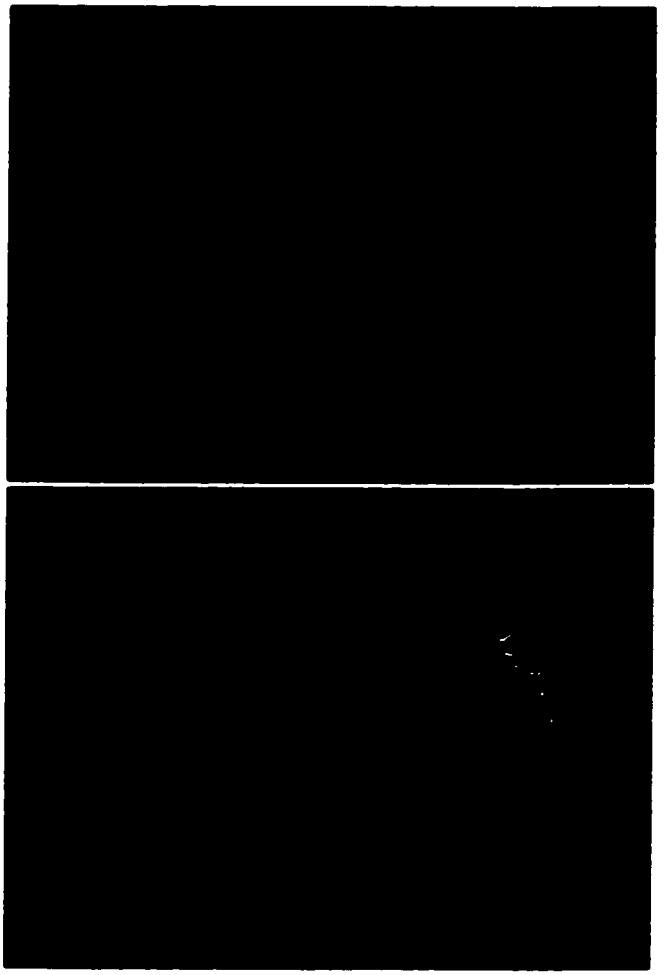
c. Number with severe clashes (overlap ≥ 0.4 Å) in both orientations.

d. Number classified as "Unknown" (score difference > 0.5), even after individual examination.

For histidine side-chains†, there are 47 fixed by metal ligation, 250.5 which unambiguously should be kept in their original orientation (66%), 37.5 which unambiguously require flipping (only 10%), 13 with unresolvable clashes both ways (3.4%), and 31 ambiguous cases with small score differences (8.2%). The non-integral values occur at a dimer interface, as discussed below. Histidine residues show fewer flips than Asn/Gln amide groups, but more often have unresolvable clashes. Those unresolved clashes are almost all with O atoms and, therefore, may be cases of CH···O H-bonding (Derewenda, Lee, and Derewenda 1995) in His rings.

† The assigned histidine protonation states in the *Top100DB1* are: neutral with H on $N^{\delta 1}$ in 35% of cases, neutral with H on $N^{\varepsilon 2}$ in 56% of cases, and a charged doubly protonated His in 9% of cases (118:185:30 respectively).

FIGURE 5-7. (a) and (b) Evidence for dynamic equilibrium in the flip state of His54 in the scorpion toxin 1PTX (Housset *et al.* 1994). Although this His ring has good *B*-factors and makes good contact with the structure behind it, each possible ring flip position makes an N^{δ} H-bond to a well-ordered water molecule that clashes with the other flip state. Although those clashes can be partly mitigated by considering CH···O H-bonds, presumably His54 occupies both conformations.

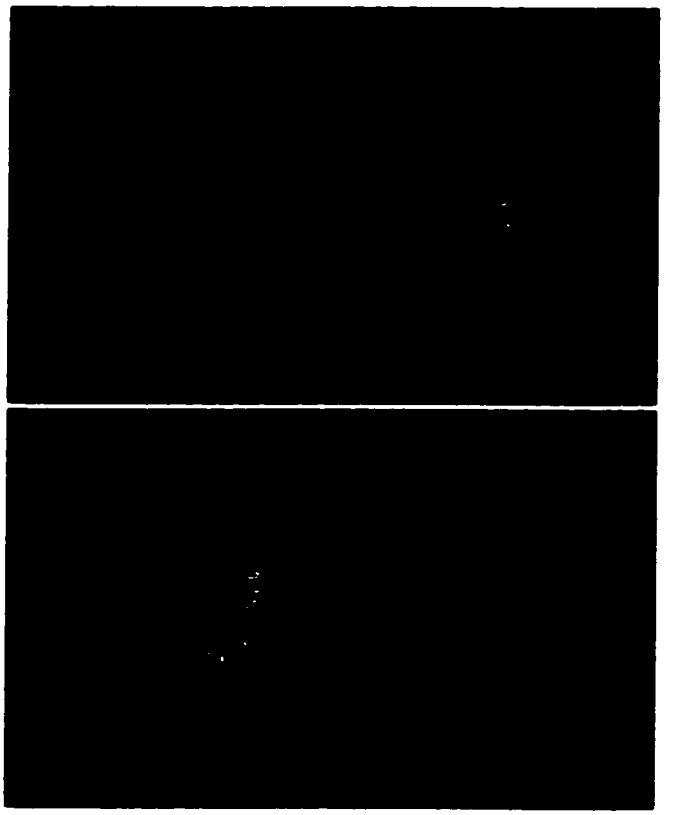


The marginal Asn/Gln/His cases with a small score difference include: (1) completely exposed side-chains with no neighboring atoms or self clashes; (2) H-bond networks with no external clashes to any alternative H positions and nearly equal scores when donors and acceptors are all switched; (3) H-bond networks across

symmetrical dimer interfaces; and (4) conflicts in which each flip state has a favorable interaction incompatible with the other state. Some of these examples make it clear that a flip state can genuinely alternate between two possibilities. For the ROP protein dimer-interface clique of SerA42a-HisA46-HisB46-SerB42b in 1RPO (Vlassi *et al.* 1994), both the Ser alternate conformation and the His flip state must differ across the two subunits; 1RPO His46 shows up as a half-integer value in the His assignments above, because the two interacting copies must be positioned asymmetrically, with one flipped and the other not. Figure 5-7 shows the two flip states of His54 in 1PTX (scorpion toxin), each of which has one good H-bond to a water molecule which clashes in the other flip state. Allowance for the positive effect of CH···O H-bonding would decrease the severity of the clashes and allow a water molecule to stay (a bit farther out) when the ring flipped. However, this histidine residue almost certainly occurs in a mixture of the two orientations.

The cases originally assigned as double-clash by REDUCE predominantly consist of side-chains that are themselves well defined but are bumped by another slightly mispositioned group, often with high *B*-factors; for example, the double-clash Asn366 in 2OLB (oligopeptide-binding protein, Tame *et al.* 1995) overlaps H^{δ} of Arg413, but both from the scores of +0.2! versus –8.8!, and from our visual examination, it is clear that four good H-bonds in the original orientation are clearly preferable to just one in the flipped orientation. Such cases were reassigned in round 3.

FIGURE 5-8. (a) and (b) Asn12 of 1LUCb luciferase (Fisher *et al.* 1996), an example whose interactions are consistent with possible side-chain deamidation. The left-hand side is clearly compatible only with an O^{δ} that can H-bond tightly with two backbone NH groups as in (a), rather than an N^{δ} that would clash impossibly as in (b). The fact that the Arg guanidinium group was observed in this close position also implies an O^{δ} on the right-hand branch of the Asn, both for steric and for electrostatic compatibility.



Three of the Asn double-clash examples are consistent with the possibility of chemical de-amidation of the side-chain: 1CSE Asn158E (subtilisin, Bode, Papamokos, and Musil 1987), 1HFC Asn206 (collagenase, Spurlino *et al.* 1994), and 1LUC Asn12 (luciferase, Fisher *et al.* 1996), which is shown in Figure 5-8. The tight interactions around Asn12, including two H-bonds to backbone NH groups, definitely require an O^{δ} for the left-hand branch of the amide. There is some room around the

arginine guanidinium group, and in the unmodified protein it would need to move farther to the right, away from what would then be the NH_2 group of Asn. However, the fact that the Arg was observed this close to the side-chain of residue 12 in the crystal structure provides circumstantial evidence that Asn12 may have become de-amidated to Asp.

Overall, less than 14% of all the Asn/Gln amide orientations in the 100 proteins remain undetermined here, once H atom van der Waals interactions are considered. Some of those ambiguous cases represent our inadequate level of knowledge, but most of them probably indeed occur in the protein molecule as a mixture of both orientations.

Discussion

The present analysis departs from common practice in two main ways: the use of small-probe contact dots for explicit visualization and quantification of molecular interactions, and the placement of all hydrogen atoms, both polar and non-polar, and inclusion of their van der Waals as well as H-bonding contributions. That additional information made the process of orienting side-chain amide groups much more straightforward and more often definitive than the H-bonding analyses used in previous work. H-optimized coordinate files for the 100 high-resolution proteins of our dataset are now available for further structural analysis. The programs REDUCE, PROBE, and MAGE are available for adding and optimizing H atoms, for analyzing

the contacts in known macromolecular structures, and to help in the determination of new structures.

It appears that most side-chain amide groups do indeed have surroundings in the equilibrium protein structure that enforce a unique, and readily identifiable, amide orientation. Assigning those orientations correctly will help in the details of refinement and the identification of water molecules in crystal structure determinations, and will aid in analyses of hydrogen bonding, water structure, side-chain conformations, and ligand binding. Nielsen *et al.* (1999) have shown that optimization side-chain amide orientation can improve both the results of pK_a calculations and the results of electrostatics calculations. The discovery of serious internal clashes in Asn/Gln side-chains as they are fit even in the highly accurate structures of our dataset (e.g. the severe Gln H^{ε}-H^{α} clash in Figure 5-4(b)) implies that there are problems with existing side-chain rotamer libraries. Chapter 7 describes the development of a new rotamer library that address that issue.

Even using H-bonds and van der Waals interactions, there still remain between 10 and 15% of the side-chain amide groups (and also of the His rings) whose orientation is ambiguous. A few of these cases are due to unresolved problems in the coordinates, and there would be somewhat more such cases in lower-resolution structures. However, at least 10% of these side-chains are probably in dynamic equilibrium in the actual protein molecules, such that both flip states would be significantly populated. For some of the unassigned cases (e.g. Figure 5-7) the amide or ring plane is well defined but the flip is not, while for some of the fully exposed,

high *B*-factor cases (more common for Gln than Asn) even the plane orientation is probably dynamic. Thus we feel that the 14% of side-chain amide groups left unassigned here mainly represent not a failure of the method, but a successful identification of the cases that should not be assigned. It also follows that such unassignable cases should be omitted or treated separately in any statistical analyses of side-chain conformations or interactions. To this end, we flag these cases in the headers of the modified PDB files.

Here, we have tried to start out with a simple and straightforward model and add complications only when we are convinced of their necessity and are also sure that they will contribute in the right direction even for the pathological cases caused by occasional coordinate errors. Our initial aim was simply adding hydrogen atoms in order to use contact dots to quantitatively analyze interior packing in proteins. It was immediately obvious this could not be done without first correcting side-chain amide flips, which led to the study described here. In addition, both the ring flip and protonation state of histidine residues had to be considered, although that treatment was developed only far enough to avoid incorrect His influence on Asn/Gln orientations, since a more complete analysis of His protonation equilibria would require detailed electrostatics and, at times, knowledge of the pH.

The major, completely necessary, complication in the present method is of course the combinatorial analysis of the H-bond network cliques. The tractability of the clique analysis depends, in turn, upon keeping several other aspects of the model simple: Met methyl groups are the only ones rotated; the probe radius is set to zero

so that only H-bond and overlap terms enter the combinatorial search; and, most importantly, interactions with crystallographically located water molecules are included but water-water interactions are not. Inclusion of water molecules is crucial for success of the algorithm, but a simplified model that treats their possible H-bonds as completely independent has worked quite well. We preferred not to attempt explicit placement of hydrogen atoms on the water molecules, because the errors in position or occupancy for a significant fraction of water molecules (for example, those that are impossibly close to side-chain atoms) could often produce problems that would propagate through the network of water molecules. Interactions across crystal contacts were also deliberately omitted, because we are more interested in what can be learned about the molecular structure than in the crystal structure for its own sake. There are, of course, many efficient search algorithms that could be applied to optimizing the clique scores. However, since the simplest and most guaranteed method of complete enumeration is indeed fast enough for the actual cases found here, it was the preferred choice.

All-atom contact analysis, in general, is based on geometry and atom types, rather than on energies. In particular, its treatment of electrostatics is indirect and, so far, only short range: hydrogen-bonding interactions based on degree of atomic overlap, with charged H-bonds stronger only because they can overlap further. Certainly those short-range H-bonds are the most dominant single factor affecting amide orientation. As a future modification, we could incorporate a weighting scheme for the non-overlap contact dots based on atom types, in order to add mid-range electrostatic preferences for distances between grazing contact and a probe diameter fur-

ther out. We do not intend to add long-range electrostatics, however. It would not combine easily with the geometrical terms, and any simple, dielectric-based treatment is likely to overestimate the contribution.

Although adding and optimizing hydrogen atoms and determining side-chain amide orientations is an apparently simple set of tasks, the number and detail of considerations involved and the variety of unexpected special cases that occur in 100 protein structures are very large. Therefore, the automated algorithm in REDUCE is gradually developing into an expert system. Fortunately, for the most part those developments make it more robust and easier to use.

CHAPTER 6

Small-probe Contact Surfaces in Protein Design

The small-probe methodology was revived and expanded, in large part, to improve protein design. The hope was that by understanding the nature of close atomic contacts in the best available examples, we could learn how to design new proteins with both stable and unique structures. So far, the most useful principles to emerge are that hydrogens are crucial for design or redesign and that all-atom steric constraints are very restrictive. The use of small-probe contacts to assist protein design is just getting started, and only a few examples are available.

Monomeric Lambda Repressor Tryptophan Mutant

Folding kinetics in the Oas laboratory for the stabilized G46A/G48A double mutant of the monomeric form of λ repressor ($\lambda^{*}_{6\text{-}85}$) using NMR line shape analysis in concentrated urea showed that, in the absence of denaturant, this protein was folding at a faster rate than had been reported for any other protein (Burton *et al.* 1996).

Following an early test of PROBE where I surveyed contacts for tryptophan side chains, Jane Richardson and I were asked to try to design a variant of $\lambda^{*}_{6\text{-}85}$ with a buried Trp, permitting fluorescence stopped-flow studies that could confirm the NMR results. Stopped-flow would also be used to search for folding intermediates. Demonstration that extremely fast folding proteins are two-state—without populated discrete folding intermediates—would argue against the idea that efficient folding in proteins must progress along a distinct pathway with well defined intermediates and in favor of the new concept of folding funnels (Dill and Chan 1997; Wolynes, Onuchic, and Thirumalai 1995). In addition, if the NMR results were indeed confirmed, that method could be applied more confidently in situations that exploit its unique capabilities (e. g., the study of very fast folding proteins under conditions where they are somewhat unstable).

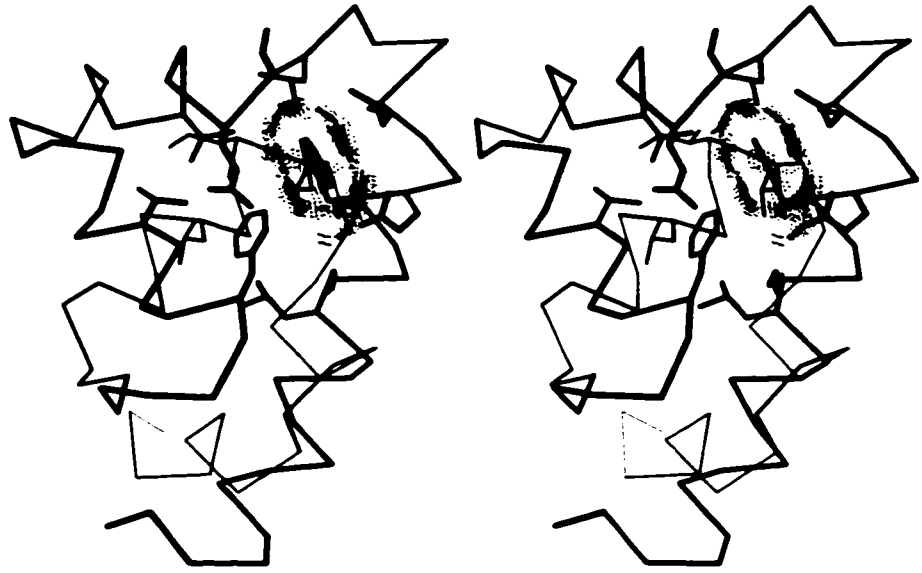
Modeling of the Mutant

A computer model of $\lambda^{*}_{6\text{-}85}$ was created from the coordinates of chain 4 (the better-ordered of the protein N-terminal domains) in the 1.8 Å Protein Data Bank entry 1LMB for the λ repressor/DNA complex (Beamer and Pabo 1992). Previous studies have shown that monomeric $\lambda_{6\text{-}85}$ has the same structure in solution as the corresponding segment in this crystal structure of the full length N-terminal domain dimer bound to DNA (Huang and Oas 1995). The β carbons for the G46A and G48A substitutions were added using standard geometries (Engh and Huber 1991). All hydrogens were also added in standard geometries using REDUCE. A tryptophan substitution was tried at the position of each of the aromatics (Y22, F51, Y60, and F76) and at several other buried residues (L18, V36, M40, L50, L57, L65, and

L69). The Trp in each substituted model was examined using the MAGE display program (Richardson and Richardson 1992), manually rotating over a full range of sidechain χ^1 and χ^2 angles to determine if there is enough room to accommodate the large side chain without significant changes to the conformation of neighboring side chains. Interactive, constrained energy minimization of nearby side chains was done with SCULPT (Surles *et al.* 1994); no changes were made to the positions of backbone atoms. For the most promising conformations, small probe van der Waals contact surfaces (Chapter 2 and Word *et al.* 1999a) were examined to ensure that all atoms fit favorably with their surroundings, including the interactions of all explicit H atoms. This criterion is very stringent and eliminates most candidates.

A F51W mutant clashes slightly with side-chains L18 and L65 in any possible arrangement, and so would require either an additional compensating mutation or some shift of the backbone. Tyr60 is exposed in the native protein; therefore a Trp at this location was not considered useful as a fluorescent label for folding experiments. Modeled replacements of residues 18, 36, 40, 50, 57, 65, 69, and 76 appeared much too crowded to permit the mutation.

FIGURE 6-1. Stereo rendering of small-probe contact dots for Trp22 in the model for $\lambda^{*}_{6\text{-}85}$(trp), with the C^{α} backbone and neighboring side chains, showing excellent packing of the tryptophan. All H atoms (not shown) were present in the contact dot calculation.



Residue 22 was considered the best candidate for substitution (Figure 6-1). Trp22 in the resulting $\lambda^{*}_{6\text{-}85}$(trp) variant packs well at $\chi^1 = 174°$ and $\chi^2 = 81°$—a favorable rotamer (e.g., Lovell *et al.* 2000; Ponder and Richards 1987), with Phe51 and Lys26 contacting opposite faces of the Trp ring. SCULPT minimization made only very minor adjustments to nearby sidechain positions. The conformation is very good, but the sidechain NH has no hydrogen-bonding partner and the hydrophobic end of the indole ring has some solvent exposure. Comparable tryptophan conformations are observed in crystal structures, and the contact score for the modeled new Trp in the model structure is within the top 6% found for the 279 Trp side-chains in the *Top100DB1* database of high-resolution crystal structures (see Chapter 4 and Word *et al.* 1999a).

The $\lambda_{6\text{-}85}$ gene sequence contains a large number of AT bases near residue 22, and there was some concern that the primer for the Y22W mutation might not anneal reliably. Mutations were therefore tried at both Y22W and F51W. Since our first choice was, in fact, successful, only it is described below.

Mutagenesis

I produced the construct used during expression of the mutant protein (during a short rotation in the Oas laboratory) by site-specific mutagenesis (Kunkel, Roberts, and Zakour 1987) on the uracylated single-stranded DNA plasmid pWL46,48 (derived from pAED4, Doering 1992), encoding the 6-85 fragment of Lambda repressor with glycine 46 and glycine 48 changed to alanines (Burton *et al.* 1996). The 38-base mutagenic DNA primer GGACGCACGCAGGTTGAAAG-CAATTTGGGAAAAAAAGA was ordered from Genosys (The Woodlands, Texas). It includes a silent mutation in alanine 15 which eliminates a BsshII restriction site to facilitate the identification of mutant clones, and it was designed with attention to avoiding the formation of duplexes, hairpins, and false priming sites using the programs DNA STRIDER 1.1 and OLIGO 4.0 for the Macintosh. Annealing from 70°C down to 34°C was done in one hour.

In vitro mutagenesis reaction products were transfected (Hanahan 1983) into the ung+ *E. coli* strain JM83 which, in the Kunkel procedure, increases the mutation efficiency by destroying the uracyl-containing template. Cells were grown overnight in LB+Ampicillin at 37°C and plasmid DNA was extracted with a Qiagen miniprep kit. Agarose gel electrophoresis of AseI/NsiI and AseI/BsshII restriction

digests showed 67% of the clones were mutants. The correct mutations were confirmed by automated DNA sequencing (thanks to generous support from the GlaxoWellcome Sequencing Core Facility).

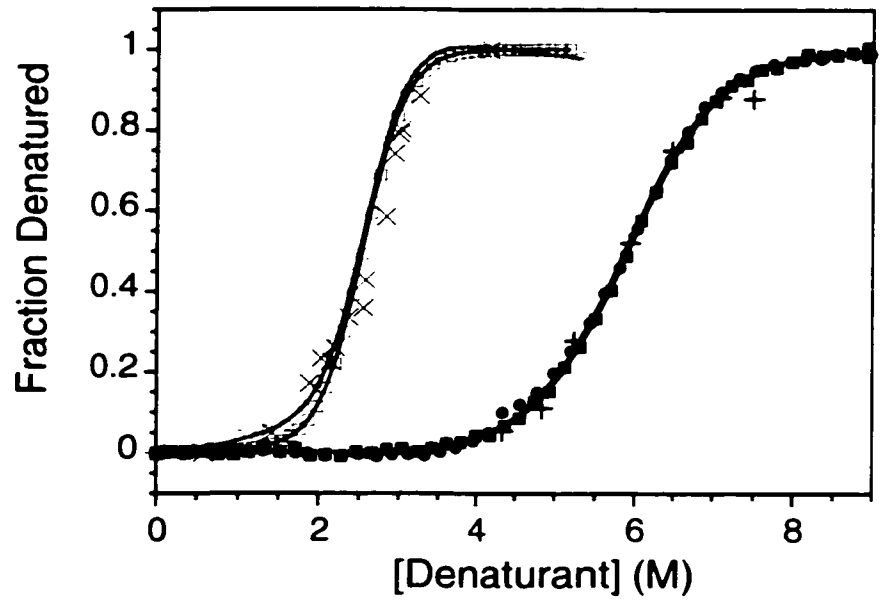
To bypass problems from possible mutations elsewhere in the plasmid, the mutated Lambda gene was extracted with the endonucleases NdeI and SalI, purified by electrophoresis on 0.9% agarose, GeneCleaned, and ligated into the plasmid pAED4 in JM109 cells.

Results

The expression, purification, and characterization of $\lambda^{*}_{6\text{-}85}$(trp) was accomplished by Sina Ghaemmaghami and Randall Burton (Ghaemmaghami *et al.* 1998). Circular dichroism suggested addition of the bulky indole group had not significantly altered the structure. It had also not appreciably destabilized the protein. Denaturation studies in both urea and guanidinium chloride yielded very cooperative titration curves (Figure 6-2) and unfolding free energies (ΔG_u) of about 5.6 kcal/mol at 20°C. The background, $\lambda^{*}_{6\text{-}85}$, has a similar ΔG_u value of about 5.4 kcal/mol at 25°C (Myers and Oas 1999). Remarkable agreement between results from various conditions and methods was used as evidence that $\lambda^{*}_{6\text{-}85}$(trp) folds in a *thermodynamically* two-state manner (Ghaemmaghami *et al.* 1998). A 20 nm frequency shift in UV fluorescence between native and denatured states was measured. With this observation, we had achieved our specific design goal for this mutant: the introduction of a detectable indicator for rapid folding kinetics measurements. Ghaemmaghami and Burton used this to show that stopped flow fluorescence matches the

results from NMR line shape results. This and other studies in the Oas lab have demonstrated that NMR line shape analysis is a reliable kinetics tool. When folding data of either method are extrapolated back to zero denaturant, the folding time for $\lambda'_{6\text{-}85}$(trp) is estimated to be 20 μs at 20°C. No intermediates were detected with stopped flow, confirming the two-state model and supporting the "new view" of protein folding funnels.

FIGURE 6-2. Comparison of fraction denatured calculations for $\lambda'_{6\text{-}85}$(trp) by equilibrium and kinetic techniques. Equilibrium titration as monitored by CD at 220 nm: GdmCl as denaturant (□), urea as denaturant (■). Equilibrium titration as monitored by fluorescence: GdmCl as denaturant (○), urea as denaturant (●). Superimposed are the fraction denatured as measured by NMR line shape analysis (+) and as measured by fractional amplitude of the stopped-flow traces (×). Experiments are performed at 20°C in 100 nM NaCl, 20 nM K_2PO_4 (pH 8.0). The titration curves were fit to a two state equation (solid lines). Reprinted from (Ghaemmaghami *et al.* 1998).



SSFrame

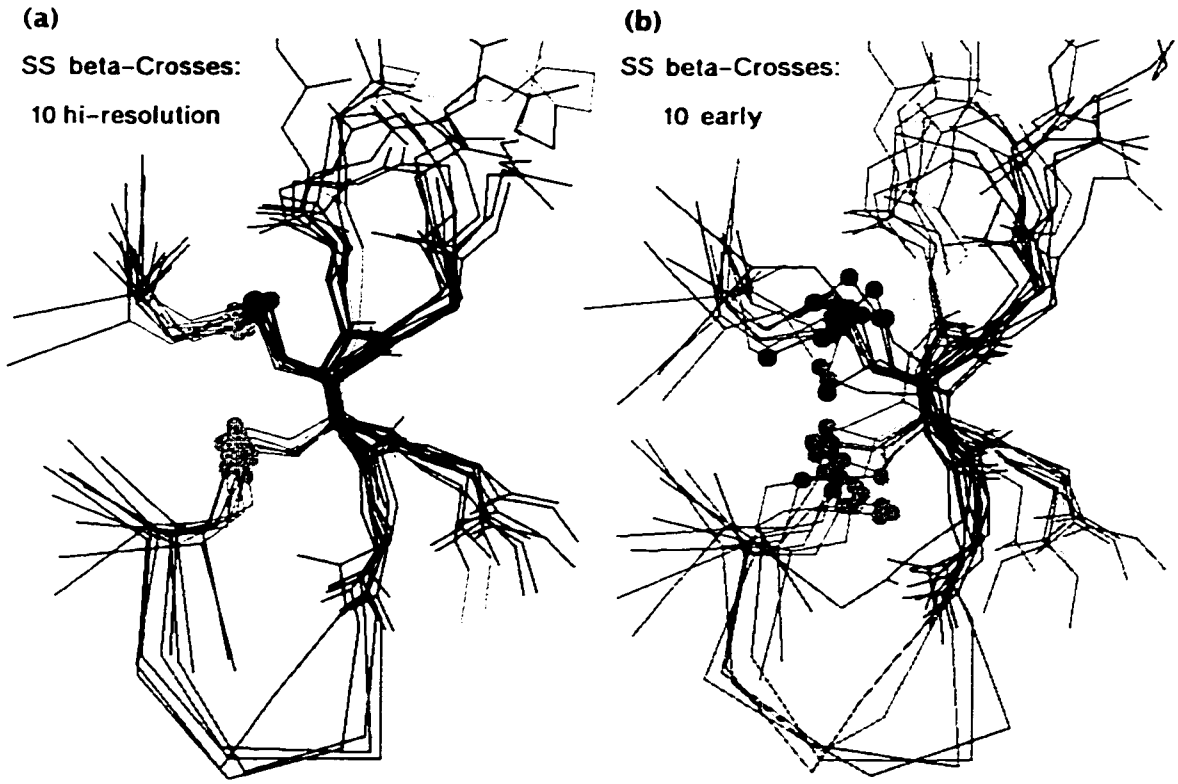
Exploring an SS-beta Cross as a Framework for Iterative Design

In spite of its strengths, full *de novo* design has several liabilities. Results are often difficult to interpret because all the design choices are made at the same time. The designed proteins have not been well ordered. Furthermore, the design–results–redesign cycle time can be years, severely limiting the number of interesting questions which can be explored. To accelerate our rate of progress, I felt it was time for the group to consider altering our design process.

Earlier work, including *Felix*, *Betadoublet* and *Alacoilin* in our lab (Hecht *et al.* 1990; Quinn *et al.* 1994; Gernert *et al.* 1995), showed that protein secondary structure and approximate shape can be engineered. Our more recent work has been focused on getting around the frustrating lack of structural uniqueness in *de novo* designed globular proteins; an ironic problem given the remarkable tolerance of natural proteins for mutations. This remains the key hurdle in protein design today. In *SScorin* (Richardson and Richardson 1995), Jane Richardson attempted to stabilize her design with a disulfide core common to a diverse set of small proteins. Unfortunately, the wrong disulfide connectivity was observed in the one SScorin made to date. Meanwhile, other groups had obtained well-ordered structures from increasingly thorough redesigns on the backbone of single natural proteins; most notably, Barbara Imperiali and co-workers had successfully used incremental redesign to engineer a stable 23-residue peptide based on the zinc finger motif having a unique fold in the absence of metals (Struthers, Cheng, and Imperiali 1996). The SSframe work described here was a re-thinking of the SScorin project, using a dif-

ferent and a more accurately determined paradigm starting structure, new design tools and an incremental redesign rather than all-at-once *de novo* design methodology.

FIGURE 6-3. (a) Alignment of high-resolution β-cross proteins. Disulfides are shown with balls on sulfur. Fragments of the following are represented: bitter gourd trypsin inhibitor (1MCTi), squash trypsin inhibitor (3CTI), potent toxin (1PTX), scorpion toxin v3 (2SN3), wheat germ agglutinin (9WGA-d1), TGF-β (1TGI), gonadotropin (1HCNa), cellobiohydrolase *C-term* NMR structure (1CBH), colipase (both 1LPBa-d1 & 1LPA-d2). (b) Superposition of less-accurate earlier SS β-cross structures: trypsin inhibitor NMR structure (2ETI), charybdotoxin NMR structure (2CRD), scyllatoxin NMR structure (1SCY-m1), platelet derived growth factor (1PDGa), wheat germ agglutinin (3WGA-d1), cellobiohydrolase *C-term* NMR structure (1CBH), Bowman-Birk inhibitor π-II (1PI2-d2), ω-agatoxin NMR structure (1OAV), potato carboxypeptidase A inhibitor (4CPA), β nerve growth factor (1BET).



The choice of molecular structure on which we based the SSframe project grew out of this laboratory's studies of various aspects of protein disulfide linkages. We and others had noted similarities in the topology and disulfide geometry of a diverse set of toxins, protease inhibitors and hormones (Drenth *et al.* 1980; Isaacs 1995; Lin and Nussinov 1995; Narasimhan *et al.* 1994; Richardson 1992). Variously referred to in the literature as toxin-agglutinin folds and 'cystine knots', their potential as frameworks for protein design has been commented on (Vita *et al.* 1995). Three-dimensional alignment on the conserved pair of cystines which project from adjacent strands of an anti-parallel β hairpin in a set of unrelated high resolution X-ray structures shows a remarkably consistent geometry (Figure 6-3(a)) that suggests that this is an especially favorable arrangement worth pursuing as a design target. This compact motif has been named the 'disulfide β-cross' and extensively described by Harrison and Sternberg (1996).

SScorin (1SSR) is a 56-residue *de novo* design built around a β-cross core. The design methodology we use generates a model of the most probable conformer from a (mostly) fixed main-chain geometry and this geometry must be known quite precisely and accurately to facilitate placement of the side chains. NMR based structures and lower-resolution X-ray structures often do not provide an unambiguous set of coordinates with the precision or accuracy required (see Figure 6-3(b) for a superimposition of the dataset available at the time). We suspected that the failure of *SScorin* to oxidize to a single isomer could be due in part to the fact that the main-chain design included a large contribution from the NMR structure, 2ETI, which turns out to have an incorrect SS conformer. Also, SCULPT used a 'united

atom' force field, and we knew by then that H's were crucial (see 1SSR clashes in Figure 4-1(c)).

The criteria for selecting a molecule as a basis for incremental redesign in the present study were as follows.

- The molecule should exemplify the 'β-cross' disulfide motif.
- The molecule should be small, with few extraneous elements.
- A high-resolution (< 2.0 Å) X-ray and/or neutron diffraction structure had to be available.
- This structure must have a relaxed, plausible geometry, whose core matched the superposition in Figure 6-3.

The 28-residue trypsin inhibitor isolated from bitter gourd seeds, MCTI-II, in a 1.6 Å crystal structure of the complex with trypsin (1MCT) by Huang, et. al. (1993; 1992), met all the above requirements. Furthermore, it is a charming molecule which presents, in miniature, helix, sheet, loop, turn and crosslink—all the major elements of protein structure. Successful designs based on it would be more compelling than single-architecture designs such as helix bundles.

Design

Our plan was to pare down MCTI-II to a minimal core, using the design tools SCULPT for interactive refinement and the newly developed PROBE for contact surface analysis. We would then develop a bacterial system for overexpression of this core peptide. This would be time consuming for the first plasmid, but future design

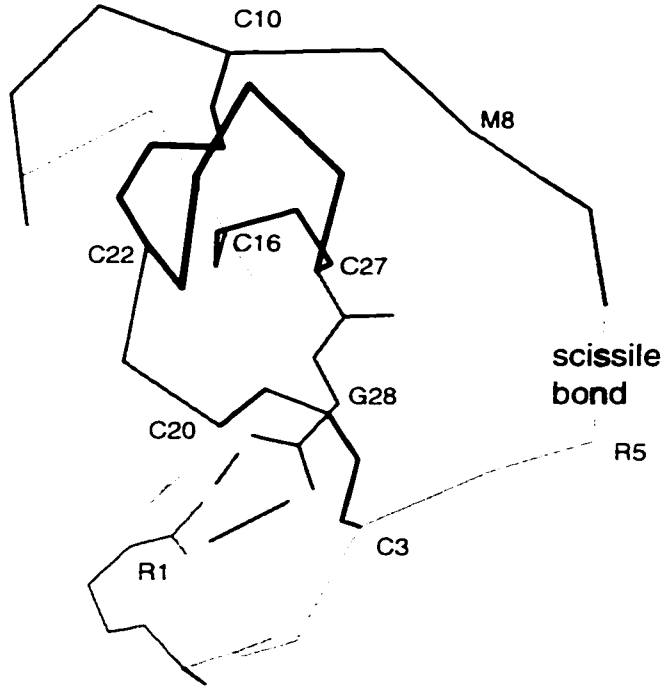
steps could go more quickly with cassette or PCR mutagenesis. Because small proteins like these (< 30 amino acids) are subject to proteolytic degradation *in vivo*, it would be desirable to express them as fusion proteins with leader sequences which target the protein into inclusion bodies for protection (Kojima, Miyoshi, and Miura 1996). One such system is TrpLE (Yansura 1990) in which a methionine is introduced at the junction between leader and peptide, and the leader is cleaved with cyanogen bromide. This cleavage method would bar methionine from the design sequence. Other methods such as thrombin cut sites or acid labile sequences are also possible, each with its own design considerations.

Another way to make these proteins, which takes advantage of their small size, is to use automated peptide synthesis. Because the Richardson laboratory had no experience with peptide synthesis and the required expertise could take a very long time to develop, we collaborated with Dr. David King at UC Berkeley. In addition to his extensive experience with peptide synthesis, Dr. King provides world-class facilities for the analysis of synthesis products. Prior to starting any molecular biology, an initial round of peptide synthesis and analysis in his laboratory would allow us to rapidly evaluate whether the β-cross disulfide core designs are stable and uniquely structured enough to support future design cycles.

Once peptides had been synthesized or expressed, purified, and oxidized, the folded material would be enzymatically cut at digestion sites built into each loop and examined with LC/MS to determine which disulfide crosslinks were formed. If sufficient amounts of "native" disulfide isomers are isolated, CD, NMR, mass spec-

troscopy and perhaps even crystallography would be used to assess their stability and provide structural information for use in the next stage of design.

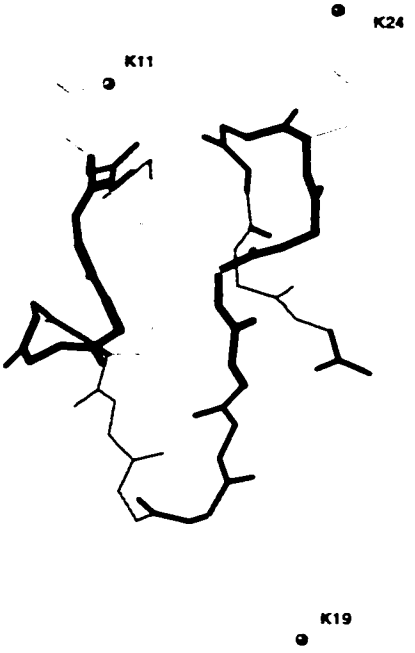
FIGURE 6-4. Trypsin inhibitor MCTI-II C^{α} structure, showing trypsin-active-site-susceptible bond, disulfide bonds, and Arg1-*C-term* salt bridge. Inhibitory loop (in gray) has been removed in SSframe designs.



MCTI-II contains six cysteine residues that could potentially bond in 15 different combinations. The native cross-links are [3-20], [10-22] and [16-27]. To restrict the number of possible disulfide isomers, we wanted to eliminate Cys3 and Cys20 which are not part of the β-cross. The N-terminal 8 residues, containing side chains that function to inhibit trypsin, form a loop that bows away from the rest of the molecule, interacting only at the ends (Figure 6-4). Elimination of this loop will do

away with Cys3 and leave a compact fragment containing just the features essential to a structural core.

FIGURE 6-5. Main chain used for both *SSframe1* and *SSframe2*, with lysines indicating the placement of tryptic digest sites for analysis of disulfide connectivity.



Analysis of cross-links was facilitated by inclusion of lysine or arginine in each of the three loops. This would allow us to cut the oxidized material at these residues by digestion with Lys-C or trypsin. The resulting fragments can be separated using capillary LC and analyzed by mass spectroscopy to assign the disulfide connectivity. Residues 11, 19 and 24, shown on Figure 6-8 as sites of disagreement between the sequence inferred from X-ray data and that from DNA sequencing, were each made lysine in *SSframe1* and *SSframe2* (Figure 6-5). Each is situated in a turn where it should be accessible to the enzyme and each either matches or is of com-

parable shape to one of the residue assignments at the same position in the native protein. In order to limit the cut sites to just these three and to avoid crowding of the enzymatic binding site, which might lower cutting efficiency, I also made the mutation R12T. The resulting CKT sequence that begins the first loop in these designs is also observed in a very similar structure, the carboxypeptidase inhibitor of PDB file 4CPAi.

With the removal of the N-terminal loop, three related issues needed to be addressed: (1) finding an alternative to Cys20, (2) covering the exposed face of loop 16-20, and (3) protecting the C-terminus. Each of these will be described in turn.

Cys Replacement

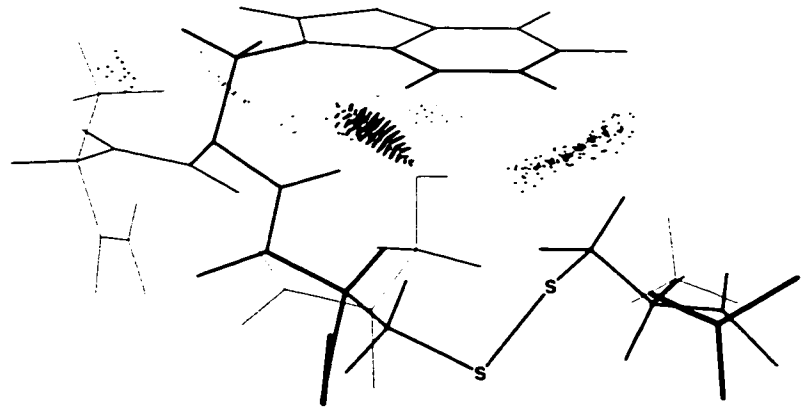
Because of the reactivity of the sulfhydryl group, free cysteines are infrequent in nature (Richardson 1981). When the loop containing Cys3 was removed, it was important to replace its disulfide partner, Cys20, with some other amino acid which would pack well in this spot. The most common replacement for cysteine is serine which has a similar shape. The two prefer different environments, however, and so this is not always the most conservative replacement. In gonadotropin and cellobiohydrolase a threonine occupies a comparable location. In all, I considered serine, threonine, alanine and leucine at residue 20. Packing was evaluated by using SCULPT to position the side chains and examining close contacts between van der Waals shells in small-probe contact surfaces, calculated by PROBE and displayed in MAGE, with the intention of avoiding 'bumps' and maximizing positive van der Waals interactions. In addition, hydrogen bonding opportunities were considered

for serine and threonine. The best contacts were observed for leucine and this is what we used in the designs.

Exposed Face

Removal of the inhibitory loop leaves exposed a pocket on one face of the turn from residues 16 to 20 which had been loosely covered by hydrophobic side chains (Cys3, Pro4, Ile6, Met8). We wanted to cover back over this pocket (especially the [16-27] disulfide) by placing an aromatic at position 17. An aromatic side chain would also be valuable as a spectroscopic label, as there are no other aromatics in the core fragment. The packing of a tryptophan at this location was explored in combination with alanine or serine at position 20. Through the use of both MAGE animation of van der Waals shells and SCULPT modeling, a conformation was found (W17 $\chi^1 = 64°$, $\chi^2 = 98°$) that places the tryptophan in extensive good contact with Ala18, Ser20 or Ala20, and Cys27.

FIGURE 6-6. Model of early stage in SSframe designs with spikes to indicate a 'clash' between $C^{\varepsilon 3}$ and $H^{\varepsilon 3}$ of Trp17 and its own amide hydrogen. Favorable packing interactions with Ala18, Ser20 and Cys27 are shown as contact dots. An aromatic residue was not included in the final design sequences.

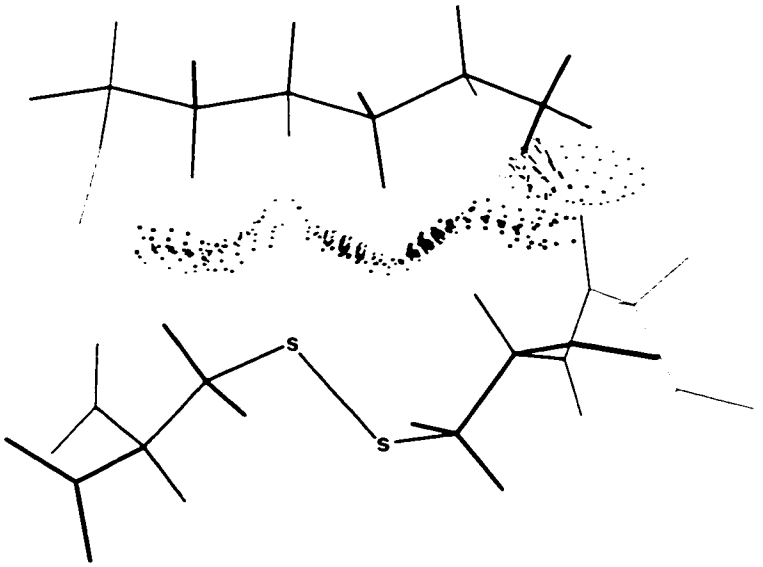


Generation of contact surfaces for this configuration, however, revealed a severe van der Waals overlap between the indole ring of Trp17 and its own amide proton which the other tools did not indicate (Figure 6-6). The MAGE animations available at the time were awkward to set up and intended as a first pass scan of conformations, so only the sidechain–sidechain interactions were modeled. The version of SCULPT available at the time did not include non-polar hydrogens and tended to generate models with over-tight packing. The ϕ and ψ angles for Trp17 (–89° and 154°, respectively) are constrained by the backbone geometry required to form the [16-27] disulfide bond. Tryptophans at similar χ angles and positive ψs in a survey of Trp conformations I compiled previously (data not shown), have ϕs ranging from –135° to –165°. A ϕ of around –90° mandates different χ angles but at these angles the Trp does not pack well. Modeling with Phe and Tyr show that none of the aromatic side chains can pack well at position 17. *SScorin* has a Phe17 with a ϕ of –91°, leading me to propose the resulting bad fit as another factor contributing to the scrambling of its disulfides. In our experience, small-probe contact surfaces, with explicit representation of hydrogens, are much more critical modeling tools than those previously used in our designs; in this case they ruled out use of an aromatic at position 17.

In MCTI-II, there are methionines at residues 8 and 17. Anticipating eventually isolating protein expressed from pTCLE (a modified TrpLE plasmid, Calderone, Stevens, and Oas 1996), it made sense to mutate Met17 and clip the fusion protein with CNBr at Met8. Once it became clear that the mutation M17W was not satisfactory, the design was reconsidered. In light of the fact that the initial product

would be the result of solid state peptide synthesis. SSframe was separated into two designs, differing in their degree of change from MCTI-II. *SSframe1* is changed minimally from the MCTI-II sequence and includes both Met8 and Met17, while *SSframe2* contains additional modifications that might enhance stability and accommodate CNBr cleavage. After reviewing side-chain preferences in related structures and examining packing in models with Lys, Arg, Leu, and Tyr at position 8 and Asp, Glu, Leu and Ala at position 17, I decided *SSframe2* would incorporate Lys8 and Glu17, which interact through a salt bridge. Both designs cover the [16-27] disulfide with an extended aliphatic side chain, making very good contacts (Figure 6-7).

FIGURE 6-7. Packing of Lys8 over disulfide [16-27] in *SSframe2* showing extended regions of good contacts, a key design goal. In this SCULPT generated model, the hydrogen bond between the amine of Lys8 and the main-chain carbonyl of Asp15 is identified as slightly too close by our method (spikes).



C-terminus

The C-terminal carboxylate in MCTI-II forms a salt-link with the guanidinium of arginine 1: an interaction we believe is important to the stability of this part of the structure. An Arg or Lys at residue 19 could potentially provide a similar stabilizing interaction, but Lys is too short and strained χ angles are required for Arg to reach. An alternative way to stabilize the C-terminus is to convert it to an amide. The resins for solid state synthesis of amidated peptides do not yield as high a quality product but this was not a problem because sufficient amounts of peptide were purified to enable this pilot study. With bacterial expression, we would initially try to fold the peptide without any C-terminal modification but if this was unsatisfactory, we could use the C-terminus alpha amidating enzyme that is used to produce active calcitonin (Yabuta 1995). Both *SSframe1* and *SSframe2* include a C-terminal amide.

To further bolster stability at the C-terminus, which is a glycine with its additional backbone degrees of freedom, I modeled the triple modification of side chains on adjacent strands of the beta hairpin: I21T, V23I, G28T. The idea was to try and lash down that terminal residue with an internal hydrogen bond and one from residue 21, along with increased van der Waals interactions with residue 23. The model showed that Thr21 cannot make the desired H-bond but Ile21 could make great contacts with Ile23 as could Thr28. Therefore, these changes were made to positions 23 and 28 in *SSframe2*.

In a last attempt to introduce a well packed aromatic sidechain into the design, both Trp and Tyr were tried as replacements for His26. In neither case could the aromatic ring generate more than a tiny amount of contact surface. I resolved to leave the histidine in the first design and consider replacing it with a tyrosine in future design steps.

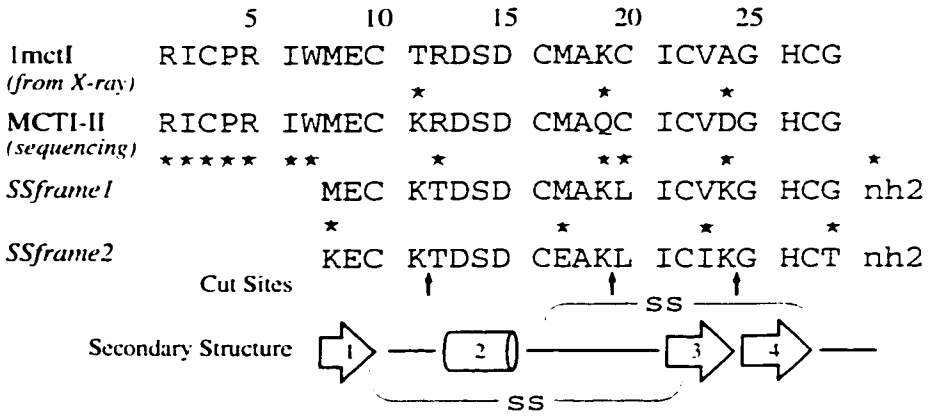
Final Model

The final model consists of a small three strand β sheet with a +2x,−1 topology. The crossover connection contains one turn of a 3_{10} a helix with a web of six hydrogen bonds. There are two β hairpin turns of type II' and type I' respectively. The two disulfides emerge from across the wide pair of hydrogen bonds on adjacent strands of anti-parallel β sheet to connect above and below the helical segment. The cross-links overlap one another in sequence. Disulfide [10-22] is a right-handed spiral and [16-27] is a left-handed spiral. The γ sulfurs of [16-27] pack neatly across the C^{β}-S^{γ} bond of Cys10 at a 90° angle to form the β-cross. Both SSframe designs are 21 residues long not counting the terminal amide.

TABLE 6-1. Calculated SSframe Properties

	Mol Wt (reduced)	PI	Net Charge (pH ~7.5)	Similarity to MCTI-II
SSframe1	2272	7.0	+1	81% (17/21)
SSframe2	2325	6.0	+1	62% (13/21)

FIGURE 6-8. SSframe Sequences. Stars mark changes from the sequence above, arrows mark designed cleavage sites, and diagram below shows secondary structure and SS connectivity.



Design Tools

SCULPT's energy potential is probably not accurate enough to predict backbone conformation, so my modeling in SCULPT was done by allowing side chains to move with respect to a frozen backbone (except at the C-terminus). SCULPT has an explicit representation of only polar hydrogens—the others are taken care of through expanded van der Waals terms for heavy atoms (i.e., "united atom" implicit hydrogens). Our studies of contact surfaces in high resolution structures have convinced us that explicit representation of all hydrogen atoms is necessary to distinguish good packing from bad. We have expanded SCULPT's parameter files to include non-polar hydrogens and this is what I used for later stages of modeling, but at that time we had not yet determined the appropriate van der Waals terms for this parameter set. The contact surfaces in our earlier design model of *Felix* looked terrible—a riot of bumps—and *SScorin*, which was designed with SCULPT but not

PROBE, was clearly too tight. The SSframe designs, therefore, used PROBE extensively for both side-chain selection and conformation.

Synthesis and Oxidation

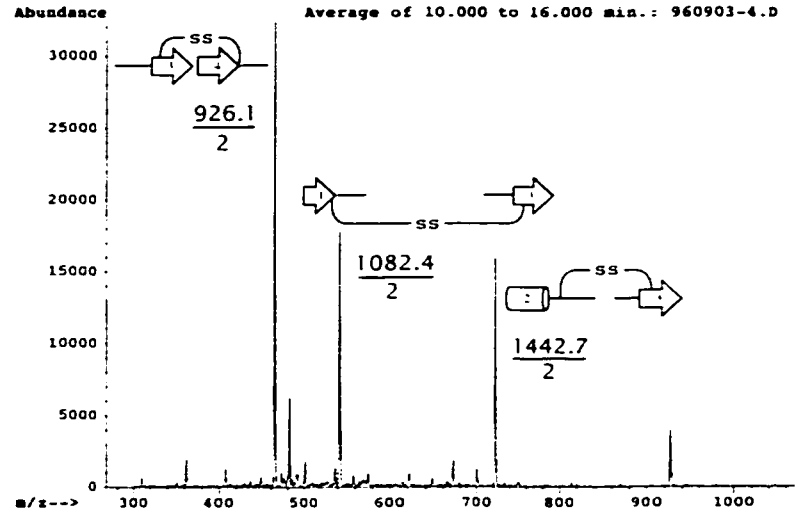
Dr. David King used FMOC chemistry (NovaBiochem) on an Applied Biosystems automated peptide synthesizer to prepare 119 mg crude *SSframe1* in three days. Preparative HPLC on a Vydac C18 reverse phase column of half this crude product yielded 8.5 mg of lyophilized peptide with an estimated purity (as judged by HPLC) of 95%. Similar processing of *SSframe2* yielded 20.4 mg of purified peptide from 110 mg of cleavage product.

At Berkeley, a trial oxidation of *SSframe1* was carried out as follows: 100 µg of purified peptide was dissolved in a mixture of 20 µl n-propanol/30 µl water in a 1ml glass vial. To this was added 50 µl 1 M phosphate buffer (pH 8.5), 3 µl DMSO and 100 µl water. The solution was stirred in air at room temperature for 6 hours. Samples were taken hourly for analysis by reverse phase HPLC (Vydac C18, 2.1x250 mm column). The starting peak with a retention time of 20 min developed into a half dozen peaks then quickly resolved into a single product peak running at 18 min. Mass spectroscopy indicated the loss of 4 mass units, corresponding to the formation of two disulfide bonds.

Encouraged by this result, the above conditions were scaled up and the entire 8.5 mg of pure *SSframe1* was oxidized overnight in a hood then purified by reverse phase HPLC. An overnight tryptic digest of 100 µg in 100 mM tris buffer at pH 8.5 resulted in 5 main peaks on capillary HPLC. Electrospray mass spectroscopy of

these peaks revealed the following connectivity between the four expected cleavage fragments: (1-2),(1-3),(1-4),(2-3),(2-4),(3-4). The native connectivity is (1-3) and (2-4). One of the HPLC peaks (see Figure 6-9) contained multiple species allowing an estimation of relative ratios of disulfide isomers from the intensity of the molecular ion: (1-3):(2-3):(3-4) = 1:1:2.

FIGURE 6-9. Electrospray mass spectrum of one fraction from an overnight digestion of *SSframe1* with trypsin. All possible disulfide connectivities are observed. Masses are for the doubly ionized species. Compare secondary structure fragments to diagram in Figure 6-8.



Scrambled disulfide bonds were also observed when *SSframe2* was oxidized using this protocol. That each of the disulfide isomers ran at the same retention time on analytical reverse phase HPLC (0.5%/min ACN/H2O gradient) suggested these isomers had a very similar shape and would therefore be difficult to separate and purify. In hindsight this is not too surprising, since the design puts the four cysteines almost equally close together in space.

For comparison, Dr. King synthesized a peptide we named *kkwt*, identical to the wild-type MCTI-II sequence except for the mutations Q19K/D24K to aid analysis. During a visit to his laboratory, I purified 4.7 mg of *kkwt*, 4.2mg of *SSframe1* and 4.1mg of *SSframe2* for additional folding studies. I ran a 20 hour trial oxidation on 65 µg of *kkwt* in n-propanol and it appeared to fold a bit slower than the others and had several smaller product peaks in addition to the main one. Oxidized *kkwt* has not yet been analyzed for disulfide connectivity.

It is possible that the oxidation conditions we had been using were too rapid, giving rise to kinetic trapping of non-native disulfide bonds. Back at Duke, I set up a series of folding experiments (described below) in a reduced/oxidized glutathione buffer designed to promote the breaking and re-making of SS bonds (Creighton *et al.* 1993; Freedman 1995; Kojima, Miyoshi, and Miura 1996; Price-carter, Gray, and Goldenberg 1996a; Price-carter, Gray, and Goldenberg 1996b; Ruddon, Sherman, and Bedows 1996). The reactions were followed by C18 reverse phase HPLC on a narrow bore column using the following gradient profile unless otherwise indicated: buffer A=0.1% TFA in water, buffer B=0.085% TFA in ACN; 11-21%B at 2.5%/min, 21-50%B at 0.5%/min and 50-66%B at 2%/min; flow rate: 0.3ml/min.

To define the retention times of the unfolded peptides, I reduced samples of *SSframe1*, *SSframe2* and *kkwt* at a concentration of 20 µM with 50 mM DTT in 0.1 M Tris buffer (pH 8.5) for 3 hours at 25°. Each sample gave a single main reduced peak (40 min, 33 min and 53 min, respectively), and each sample developed two additional, more polar peaks (at 22 min and 30 min) if allowed to stand overnight at 5°C.

Oxidation of *SSframe1* (20 μM) was set up at both pH 7.3 (0.1 M MOPS) and pH 8.5 (0.1 M Tris), each pH at glutathione concentrations of 2 mM GSH/1 mM GSSG and of 1 mM GSH/0.5 mM GSSG. In addition, a sample was prepared at pH 7.3 and 2 mM GSH/1 mM GSSG containing 0.3 μM protein disulfide isomerase (PDI). A control was run at pH 7.3 with no glutathione. Buffers were deoxygenated by bubbling with N_2 for 10 min. Total reaction volumes were 1 ml. These six samples were stirred in capped 1.5 ml vials and 100 μl samples were removed and quenched with 5 μl of concentrated H_3PO_4 and frozen at the following times: 5 min, 1 hr, 17 hr, 50 hr and 72 hr.

Under these conditions, folding appeared to proceed more slowly than in n-propanol. Each of the oxidation reactions generate a number of transient HPLC peaks which go away at later time points. Samples at pH 8.5 oxidized faster than those at pH 7.3 but went to a similar end point. Increasing glutathione concentration also speeds oxidation. The sample with PDI oxidized faster than those without. After 50 hours all reactions containing glutathione had produced a major product peak at RT = 27 min flanked by the smaller peaks at 23 min and 32 min which are seen at all time points. By the same time, the control had only a tiny peak at 27 min but appreciable amounts of the other two peaks.

Since each of the above conditions appeared to produce the same end point, I chose to perform the oxidation of *SSframe2* and *kkwt* under a single set of conditions (pH 7.3, 2 mM GSH/1 mM GSSG) I hoped would promote slow oxidation. After 27 hours, *SSframe2* produces a main product peak at 23 min, again flanked by peaks now at 22 min and 31 min. The two early peaks run very close together making

purification of larger samples more difficult. *kkwt* oxidizes to two main products with retention times of 40 min and 43 min. Interestingly, the 22 min and 31 min peaks have nearly disappeared after 10 min when *kkwt* is oxidized and they are missing from some *SSframe1* oxidation samples which sat in acid in a freezer for more than a week. I do not yet know what these peaks contain; because they do not vary with peptide, they appear to be contaminants.

Based on these preliminary refolding results, I performed a 65 hour batch oxidation of approximately 750 µg of *SSframe1* at ~200 µM concentration (again at pH 7.3, 2 mM GSH/1 mM GSSG). The goal was to isolate more than 100 µg of oxidized product—enough to determine disulfide bonding by tryptic digest/LC/MS. I purified the products by HPLC in 4 batches on the same narrow bore reverse phase column I used for analytical work. Afterward, I decided that column overloading had caused co-elution of the main peak with the first contaminant peak. I tried to re-purify this sample on a larger capacity 4.6 mm column (Zorbax C18) but was unable to separate peaks at even very shallow gradients (0.1%/min). It is even possible the contaminating material disappeared from the samples during the re-purification process. Samples from this preparative oxidation of *SSframe1* have not been subjected to thorough analysis.

Discussion

The $\lambda^{*}_{6\text{-}85}$(trp) mutant was a success. Through modeling and a little molecular biology, we produced the desired result: a stable, well behaved protein containing a fluorescent tag with which to monitor rapid protein folding events. All-atom contact

surface analysis identified residue 22 as a potential mutation site and eliminated most of the alternative sites that would otherwise have necessitated trial and error screening. Although the orientation of the Trp22 side chain having the best contacts does not have a hydrogen bonding partner for the ring NH, the good contacts fortunately more than made up for the missing H-bond.

SSframe could not itself be taken to a successful conclusion, but it has had important positive influences on our further projects and methods. It is unknown whether altering the oxidation conditions to include a glutathione redox buffer enriched the amount of native disulfide bonding in the SSframe proteins. Furthermore, some proteins require pro-peptide sequences which guide folding to the active form (Shinde and Inouye 1993); although no pro sequences have been described for MCTI-II, the requirement for a pro sequence or other molecular chaperone has not been ruled out. It is possible that the other isomers of MCTI-II are made in-vivo and degrade more rapidly. If the shapes of the three SSframe disulfide isomers were different enough for chromatographic separation, I could isolate the native isomer for study. But the different ways of connecting the cysteines do not seem to have much effect on the relevant exterior physical properties, probably because all four Cys are very close together in space in the SS β-cross. Directing disulfide formation to the native isomer with protecting groups might not work in the crowded interior of this small protein. Therefore, it may not be possible to obtain samples in high enough purity to study their structural integrity (e.g. by circular dichroism, amide exchange protection, nuclear Overhauser effects).

Natural small SS-rich proteins can also show mixed disulfides when synthesized, or even in vivo, so that these complications do not invalidate the design. However, it does not appear that MCTI-II can provide the desired optimal starting point for quick-cycle incremental design. We would prefer a system where samples could be readily purified and where sample production and analysis facilities were closer at hand. Peptide synthesis is a good way to speed multi-milligram scale production of small proteins. Unfortunately, our use of this technology has been hampered by competing demands on the time of our collaborator. In future studies we may switch to commercial peptide synthesis and locally available mass spectroscopy facilities.

The incremental redesign strategies have indeed become the current standard paradigm in the field (e.g., Dahiyat and Mayo 1997b; Desjarlais and Handel 1995), and have been pursued in our laboratory with the thioredoxin and SymROP projects described below. The productive but awkward modeling for SSframe motivated our development of later tools that are more powerful and much easier to use, especially the interactive MAGE/PROBE system described in Chapter 8.

Designs by Others

Additional protein design efforts, in our laboratory and in a few others, have incorporated small-probe contact information. Several of these are sketched below.

Simon Lovell used the early non-interactive contact displays when he redesigned one of the hydrophobic cores in *E. coli* thioredoxin. Thioredoxin has two distinct

buried cores separated by a β-sheet, permitting iterative cycles of design, expression, and testing, working on a single core at any one time. The protein is small and quite stable (Ladbury *et al.* 1993) so designs can be monitored with NMR (Dyson *et al.* 1990), and a 1.68 Å crystal structure is available for use in model building (2TRX: Katti, LeMaster, and Eklund 1990). The Hellinga laboratory supplied the hyperstable background mutant (D26L plus several Cys → Ser substitutions). In two rounds of design Simon introduced 7 mutations: (1) L24I, V55I, L53A, L103M; (2) D/L26F, L78I, I38D. Collectively, these mutations increase the β–sheet propensity of the dividing layer, fill a cavity, bury an aromatic side chain, and maximize the amount of close contact. PROBE scores for contacts between the 7 mutated residues and to their surroundings are: wildtype = 40.88, round 1 = 27.1, and round 2 = 53.0. They were calculated with PROBE's -count option, which produces quantitative rather than graphical output. CD spectra provide evidence that the secondary structure has been retained: spectra for both mutants are similar to wildtype thioredoxin (but differ slightly from that for the background mutant). In ANS binding experiments, the round 1 design protein binds the fluorescent dye to an extent seen in "molten" structures. However, the round 2 protein does not bind ANS significantly and thus is likely to be well-ordered. Although confirming tests are needed, the iterative redesign of thioredoxin cores appears to have started off on the right foot, showing more unique order for a variant with more changes but a better all-atom contact score.

Following up on David Richardson's realization that the ROP protein architecture contains a latent higher symmetry, the Richardsons worked with Daniel Grell, a

graduate student of Swiss chemist Manfred Mutter, to design a bundle of four identical helices with full 222 symmetry (Grell *et al.* submitted). The four ROP-like short helices were constructed by peptide synthesis and ligated to a small template group using the Template-Assembled Synthetic Protein (TASP) system (Mutter *et al.* 1992). This molecule, a "SymROP," takes full advantage of the potential for high-symmetry inherent in the ROP protein sequence, using two different linker chemistries to place the helices antiparallel. The peptide sequences were varied to increase helical propensity and to optimize side-chain packing, as measured by small-probe contacts. The first SymROPs synthesized have CD spectra indicating high helical content but their low solubility has prevented NMR analysis. Variant SymROPs and other TASPs are planned.

Michael Wisz and Loren Looger, in the Hellinga laboratory at Duke, have been using clashlists, contact dot displays, and packing density contours to monitor their design efforts. Wisz has found our contact analysis tools especially helpful for identifying conflicts in the second shell of residues around the primary mutations that create his metal binding and nascent catalytic sites. Design work in that laboratory now routinely includes all-atom steric analysis, for instance in a final cycle of conformational modification at the end of Looger's Dead-End-Elimination redesign calculations. Those calculations use our new rotamer library (see Chapter 7), which has also been adopted by protein design groups at other institutions, such as Stephen Mayo's laboratory.

CHAPTER 7

Structure Determination and Side-chain Rotamers

Improving Structure Determination

As demonstrated in Chapter 4 and Chapter 5, all-atom small-probe contact surface analysis is a critical tool for macromolecular structure validation, and one which is largely independent of the methods used to produce structures. This independence means contact surface analysis can critique the *methods* as well as the structures. On the other hand, the ideal time for the information in contact surfaces to help determine the correct structure is certainly when the model is being built and refined. Contact analysis can (1) identify mistakes, oversights and systematic problems, (2) automatically fix flipped side chains, (3) suggest reasonable solutions where electron density is poor or there appear to be few constraints, and (4) prioritize refitting tasks to focus attention on the areas with the most conflicts. If contact analysis (or any all-atom steric exclusion technique) becomes routinely used to guide structure determination, our method will lose its independence, but this may be an acceptable cost in return for higher quality structures.

Combining Contacts with Electron Density for Crystallographic Refitting

Improvements made to the crystal structures in the *Top100DB1* involved changes that are largely independent of the electron density: atoms were relabeled but only hydrogens were moved. Corrections which require repositioning of non-hydrogen atoms are best done in the context of the relevant experimental data and constraints—in this case, electron density plus bond lengths and angles. Recognizing this, Simon Lovell constructed a macro for the O crystallographic refitting program (Jones *et al.* 1991) which calls PROBE and then integrates the resulting contact dots and spikes into O's display along with electron density.

O Macros

The macro has since undergone several changes and two different variants are currently in use. One macro generates *self contacts* for a complete, static structure to which hydrogens have been added previously. It permits both a survey of the entire structure and detailed inspection of regions of interest and is easily customized to adjust the level of information produced. To eliminate the need for separate format translation, PROBE was modified to write O graphics format.

The other macro variant can be run on demand as changes are made to the structure. It writes a coordinate file for atoms in groups within 6 Å of the center of rotation. REDUCE then adds hydrogens, feeding the result to PROBE. PROBE creates contacts for clashes and H-bonds only, and these are returned to O for display. The process is a little cumbersome (and requires some fancy manipulation of O's data blocks) but, in practice, takes only a second or two. A minor problem occurs because opti-

mization of adjustable groups is only local in the interest of speed: H-bonds are occasionally underestimated or missed. Hydrogens are re-added each time the macro is run and are used only within PROBE—rather than being constructed once and retained—because O does not robustly support structures that include hydrogens. We spent considerable time developing O parameter files with hydrogens only to discover that a complete, reliable implementation would require the involvement of O's developers. When contacted, they were disinclined to invest the effort to fully support hydrogens as long as the Protein DataBank and the major refinement programs (esp. X-PLOR and CNS) use different hydrogen nomenclatures. Although some progress has recently been made in this area (see "Atom Naming Conventions" on page 46), standardization is an inherently slow process. Consequently, our macros consider all-atom steric effects but do not display the H atoms.

The O package is made available as a tar file at: ftp://kinemage.biochem.duke.edu/UNIXprograms/probeScripts/probewithO.tar. The only O parameter files now provided are our updated rotamer libraries (see below).

XFIT

The popularity of O made it an obvious candidate for adding contact information. Another obvious candidate was XFIT, the refitting component of the crystallographic system XTALVIEW (McRee 1993). A newer program, XFIT is very interactive, and Richardson lab alumnus Duncan McRee was easily persuaded to work with me on a version in which PROBE was tightly integrated (McRee 1999), automatically updating contact surfaces as a structure is adjusted. We used the same

Unix pipe software components created for MAGE/PROBE interaction (described in Chapter 8). This allows XFIT and PROBE to remain separate programs which can evolve independently. PROBE was modified to output graphics in XTALVIEW's format. XFIT was modified to work properly with hydrogens. Each time coordinates are updated, XFIT calls PROBE, forwarding *all* the atoms in a region around the center of rotation. Contact dot density can be adjusted from the XFIT console to favor either responsiveness or display quality. A density of about 10 dots per $Å^2$ results in almost immediate feedback when a clash is introduced or eliminated.

Examples of Crystallographic Use

Trp tRNA Synthetase

Contact analysis revealed 48 clusters containing 79 serious clashes for a 2.3 Å structure of Tryptophan t-RNA synthetase (TrpRS), still in the process of being refined by Charles W. Carter at UNC, Chapel Hill. At this stage of refinement, it is not surprising that inconsistencies exist within the model. However, the clashes can be used to direct further refinement more efficiently. It is our contention that all-atom steric constraints are so restrictive that satisfying them (along with the other usual constraints) will force convergence on the correct structure.

FIGURE 7-1. (a) O display of electron density map with small-probe contact dots, for Met193 in working model of Trp tRNA Synthetase being refined by Carter *et al.* (b) Repositioned conformation for Met193, especially for the methyl, showing improvement in contacts. (c) Electron density map with a lower contour level confirming the new Met conformation.

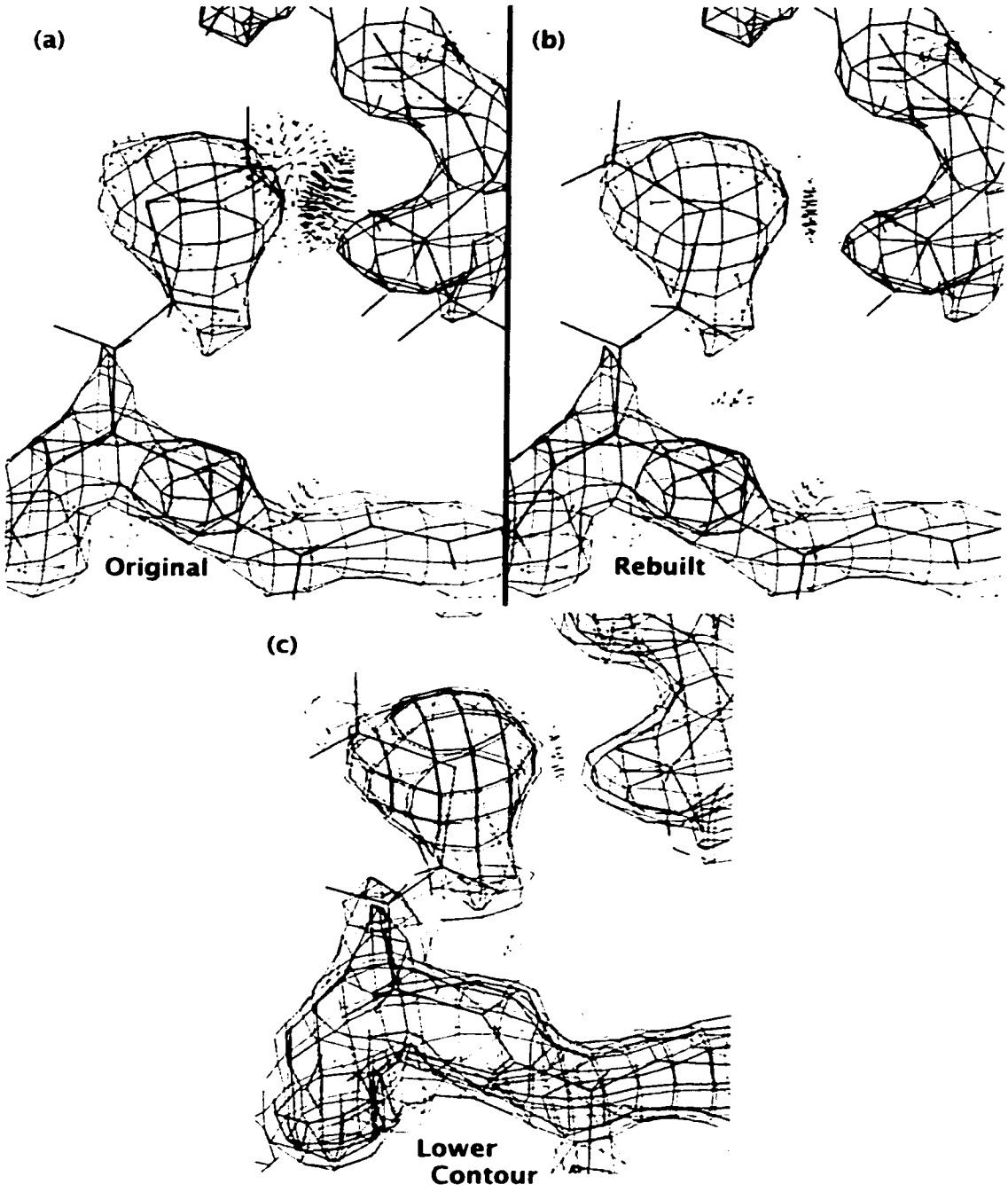

(a)
(b)
Original
Rebuilt
(c)
Lower
Contour

Figure 7-1(a) shows a cluster of the biggest clashes in the molecule. On the left is a *before* view of Methionine 193. The sulfur was originally fit into the strong blob of density but the Met methyl was placed on the wrong side. In the figure it is easy to see this cannot be right, but without the hydrogens and the spikes, it was much easier to overlook. The spikes add new information, similar to bond lengths, which clarifies the structure. A good rotamer was easily found that solves the problem (Figure 7-1(b)). If the electron density is plotted at a lower level, it confirms the new conformation (Figure 7-1(c)).

We anticipate small improvements to the crystallographic residual statistics R and R_{free} as the result of such fixes. From a slightly different perspective, Carter reports that their recent refinement of a 2.25 Å TrpRS/ATP complex against both structure factor amplitudes and phases from a very good SIRAS dataset (but not against real-space terms), carried out with a new implementation of BUSTER and TNT, reduced the number of 'serious clashes' from 270 to 138 (personal communication). This can be taken as an indication that high quality X-ray data should lead to minimal (or no) steric overlap.

COX 2 at 3 Å Resolution

James R. Kiefer, trained in the Beese laboratory and now at Monsanto, has been using the O/PROBE package and found small-probe contact surfaces to be useful guides while refitting structures for poor-quality medium resolution data. Working solely with 3 Å data for one of a series of cyclooxygenase-2 structures, he has identified and fixed clashes and other problems in his model that were unresolved by X-

PLOR, and discovered he had thereby selected the conformations of the corresponding group in a highly refined 2.4 Å COX-2 structure, a structure not consulted until after refitting. In his words, "the working map is unambiguous and X-PLOR doesn't send up any flags about a given area, but it is *wrong*. ...at 3.5-2.8 Å resolution, the definition in the map is so bad that some people don't even try to refine much, because there is too little info to help the builder. Enter dots as a powerful source of supplementary information." (personal communication) In areas where the two structures were not comparable, such supplementary information is highly valued. As well as the usual types of side-chain rotamer problems, examples of the sorts of unusual situations encountered at this resolution that Jim was able to fix with contact information but not by refinement alone include: (1) peptide flips (the main-chain density generally lacked definition, especially of the critical carbonyl Os) and (2) the orientation of aromatic residues (because of initial misfitting, the electron density was spherical). A paper has been submitted describing this structure, including the use of PROBE.

Alternate Conformations at Atomic Resolution

Sean Parkin, at Duke, is using the XFIT/PROBE combination during the final stages of several structures at 1 to 1.3 Å resolution. In most places everything looks perfect, while a few recalcitrant high-*B* regions cannot be satisfactorily fit even with our tools. However, all-atom contact surfaces could correct Asn/Gln flips not shown by H-bond criteria and were especially useful in assigning valid alternate conformations for exposed lysines and to help untangle situations where several interacting residues have multiple conformations. Analysis of the unprecedented 0.54 Å

structure of crambin with Martha Teeter of Boston College showed that PROBE could simplify the task of identifying which water positions are coordinated with each different solvent-exposed side-chain conformation.

Fitting and refinement of the alternate conformations so prevalent at very high resolution can be even more difficult than dealing with single conformations at moderate resolutions. The alternatives, by definition, have only partial occupancy, and their density often overlaps in confusing and ambiguous ways. The tools of all-atom contact analysis can be a big help, since ensuring that each conformer is both internally valid and consistent with its surroundings reduces the number of viable possibilities enormously.

Catalase

A human catalase being refined at 1.4 Å by Christopher Putnam, at Scripps Research Institute, provides an example of when hydrogens cannot be efficiently added with reduce -build. It is large—a tetramer, with 501 residues in each subunit—and contained cliques with 6 members, requiring 7488 permutations each. To streamline this and similar impediments, I developed the method (described at the end of Chapter 3) of positioning the hydrogens in two steps, which broke up each large clique into two sub-cliques which were quickly optimized.

Originally, Chris's structure highlighted an additional problem causing REDUCE to confuse residues in different subunits, stemming from the tendency when using certain refinement programs (e.g., X-PLOR and CNS) to identify separate chains only in the new segment-ID field, ignoring the standard chain-ID field in generated PDB

files. Granted, the PDB format is poorly suited to certain needs and must evolve to support ever more complex structural models, but in this case (10 chains) and most others we have encountered, the chain-ID field would have done the job nicely. Chain-IDs were easily assigned with my format conversion script pdbcns, resolving the problem. Planned future versions of REDUCE will provide a way to group chains using the segment-ID field, while continuing to work with older files which contain unrelated information in those columns. An additional formatting oversight was also discovered by the use of contact dots with catalase: alternate conformations without any flags at all.

The catalase structure is essentially completed, and PROBE is being used as a final cleanup before depositing coordinates. There were no backwards Leu, Met, or Thr conformations, and almost all changes involved Asn, Gln, or His flips. The only example incorrect in all four subunits was the Asn501/Gln475 double salt link similar to the one in Figure 5-4. Review of the USER MOD records written when REDUCE processed the catalase structure, revealed that flipped side chains and overlooked alternate conformations were more common in subunits C and D, while earlier subunits provided a kind of internal standard which confirmed our analysis. This underscores the importance of using automated tools and techniques for prioritizing refitting chores, especially in the very large structures that are increasingly common nowadays: people become fatigued and will overlook things during repetitive tasks.

FIGURE 7-2. Structure and all-atom contacts for His40 of 1MJH (Zarembinski *et al.* 1998): (a) as deposited, with ring CH clashes; (b) with ring flipped, now making good H-bonds and better contacts. View is into the active-site cleft, with the bound ATP at lower left.

Further Usage

Sung Hou Kim's laboratory at Berkeley has solved a structure coded 1MJH at 1.7 Å resolution (Zarembinski *et al.* 1998). A test for the Structural Genomics Initiative, this protein of unknown function from *Methanococcus janeshii* was observed to have a bound ATP molecule in each chain of the dimer. Collaborative examination with REDUCE and PROBE showed a well-fit structure with very few problems; the most significant one was an extremely clear flipped histidine (Figure 7-2 (a) *versus* (b)) in the active site cleft.

Doug Freymann, at Northwestern, is using the O/PROBE package on high-resolution structures of several forms of the 54H signal recognition particle homologue GTPase. He identified an incorrectly rotated Thr, a Pro with mixed ring pucker, and a flipped Gln involved in a conformation-dependent H-bond network at the active site.

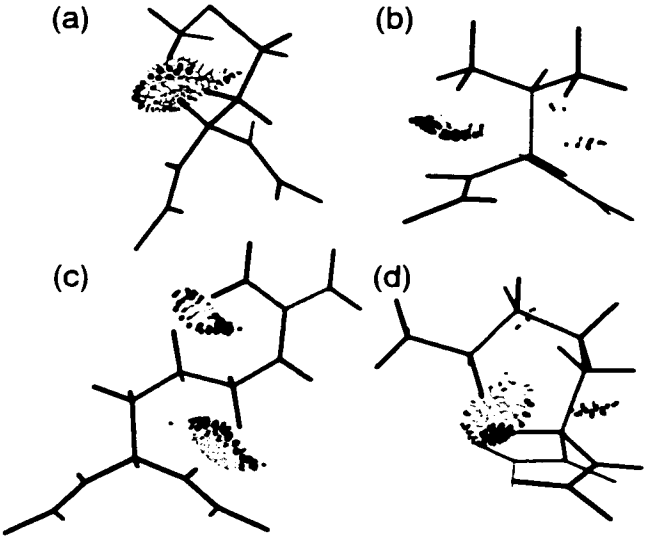
Among uses in Lorena Beese's laboratory, at Duke, a particularly interesting example is a flipped His found by Shawn Johnson in a high-resolution DNA polymerase structure. The phases were accurate enough even in this extremely large protein that contouring the map at a higher level showed distinct peaks for the two ring N positions that confirmed REDUCE's changed assignment.

Improved Side-chain Rotamers

Both the Asn/Gln flip analyses in Chapter 5 and the crystallographic work convinced us that side-chain rotamer libraries needed improvement. This constitutes one further step in a long history of work in this area. The observation that side-chain conformations form discrete multi-dimensional clusters led to the original development of a *rotamer* library (Ponder and Richards 1987), with updates by other groups as more structures became available (DeMaeyer, Desmet, and Lasters 1997; 1997; Dunbrack and Karplus 1993; 1994; McGregor, Islam, and Sternberg 1987; Schrauber, Eisenhaber, and Argos 1993; 1997; Tuffery *et al.* 1991). Side-chain rotamers are useful for model building and refitting (Jones *et al.* 1991; McRee 1993; 1999; Perrakis, Morris, and Lamzin 1999), structure validation

(Hooft, Sander, and Vriend 1996; Laskowski *et al.* 1993), and automated protein design (Dahiyat and Mayo 1997a; Desjarlais and Handel 1995). Unfortunately, all published rotamer libraries are found by our methodology to contain some rotamers with impossible internal atomic overlaps if built in ideal geometry with all hydrogen atoms. Examples are shown in Figure 7-3.

FIGURE 7-3. Examples of rotamers, previously published or currently in use, which show serious internal van der Waals clashes when built in standard geometry (Engh and Huber 1991). (a) Met **tpm** (Jones *et al.* 1991), (b) Val 120° (Schrauber, Eisenhaber, and Argos 1993), (c) Arg **tmtp** (DeMaeyer, Desmet, and Lasters 1997), and (d) Lys **mmpt** (Tuffery, Etchebest, and Hazout 1997). Only the van der Waals overlaps are shown (spikes), as calculated with PROBE.



In published work on contact surface methods, we described improved rotamer sets for Met (Word *et al.* 1999a) and for Asn/Gln (Lovell *et al.* 1999) side chains. Taking advantage of the recent number of new high quality, high resolution structures we have now produced a complete new, "penultimate" rotamer library (Lovell *et al.* 2000). The complete set of 153 rotamers is tabulated in Appendix 2. New features

of this library include a database of 240 structures at 1.7 Å resolution or better (except for the initial Met work which used the *Top100DB1*), the use of filters to omit side chains with high *B*-factors, serious clashes or alternative conformations, correction of Asn/Gln/His flips and omission of indeterminate cases, the use of modes rather than means, the listing of common-atom values where appropriate, and the careful testing for cases that result from systematic misfitting artifacts and their removal from the listed rotamers. Although Simon Lovell was first author on most of this work, I also contributed significantly; the following description will emphasize my contributions.

Angles

For the analysis of rotamers, side-chain dihedral angles were calculated with DANG, a program I developed for tabulating various structural measurements. Since, in the context of rotamer analyses, the *gauche*$^+$, *gauche*$^-$ terminology has been used equally often with each of two opposite meanings (that is, *gauche*$^+$ sometimes means +60° and sometimes means –60°), we instead use the abbreviations **p** for **p**lus χ angles (near +60°), **t** for **t**rans, and **m** for **m**inus, while χ angles centered near other values are quoted to the nearest 10°, or 5° if we felt the data justified a more accurate description. Thus, rotamers have names like Lys **mttt** or Tyr **p**90°. For comparison of theoretical contacts and clashes in potential rotamer conformations, ideal-geometry residues were constructed using parameters from Engh and Huber (1991). They were examined interactively with MAGE/PROBE, including rotation of conformational angles, and were used for the autobondrot conforma-

tional analyses described in "Rationalizing Observed Conformations with a Hard-sphere Model" on page 173.

Distribution Modes

Previous studies of side-chain conformation had used the mean to characterize the central tendency of rotamer populations. However, because means and the bins defining them can produce artifacts, I suggested the use of the *mode*—the most frequent conformation—as a way to avoid this bias. My tools for Gaussian smoothing and contouring multidimensional distributions were modified for use in determining the rotamer modal values.

A modest amount of smoothing of the distribution of measured χ angles permits the mode to be robustly estimated. The degree (width) of smoothing employed is a function of how different two measurements have to be before they are no longer considered equivalent. Smoothing was done by placing a Gaussian mask over each data point and summing the mask values at grid points spaced every 1°. The mask had a half-width at half-height of 1° for 1-dimensional data, 2° for 2-dimensional, 4° for 3-dimensional, and 6° for 4-dimensional data. Regardless of the dimensionality, each mask had an integral of 1. The rotamer was then defined as the position of the local maximum for the summed masks.

Peak half-widths (analogous to standard deviations as used with means) were defined as ± the angular distance from the modal value at which the summed mask function is half of the maximum for that peak. The artifactual broadening ("blur-

ring") of peaks caused by the mask width was corrected according to the following scheme:

$$\text{corrected width} = \sqrt{(\text{distribution width})^2 - (\text{mask width})^2} \qquad \text{7-1}$$

Half-width at half-height can be converted to standard deviation, if the distribution is normal, by dividing by 1.1774 (for a normal distribution the height at 1σ is 0.606, so that the half-width is larger than σ). We have found, however, that almost all of our rotamer distributions are in fact platykurtic (i.e., flatter-topped and steeper-sided than a normal distribution), so that a standard deviation calculated from the set of points would be even smaller than the above estimate.

For all amino acids, scatter-plot kinemages of the raw χ angle distributions, the modes, and the contours for the summed mask functions were displayed in MAGE. For Arg and Lys, a separate 3-D kinemage was made for each χ^1 (**p**, **m**, and **t**). Multiple peaks and asymmetries were evaluated; since each point carries its identity (file and residue) in the kinemage, a sample of outliers was identified and examined in the context of their 3-D structures.

The use of modes *versus* means has four important advantages. First, there is no assumption of a Gaussian shape to the distribution, which is implicit when mean and standard deviation are calculated. Many χ angle distributions are skewed, particularly for terminal χ angles of those side chains that have planar functional groups or where the rotamer is close to a clashing position. Examples include the Arg rotamers with χ^4 ±85°, where the guanidinium clashes with H^{δ} if the angle

changes to ±70°. Modal values locate the most preferred conformation reliably in these cases.

A second advantage of modes is that no *a priori* assumptions need be made about number and location of peaks in the distribution, whereas prior to calculating means and standard deviations, it is necessary to divide the data into bins. Inappropriate boundaries between bins can lead to inclusion of data in the wrong bin, pulling both means away from their true positions. For example, we have shown that amide χ modal positions most commonly occur near ±30° (Lovell *et al.* 1999), while previous treatments have drawn bins around *a priori* assumed means at ±90° (McGregor, Islam, and Sternberg 1987), 0° and ±60° (Dunbrack and Cohen 1997), or ±165° (Carugo and Argos 1997). With such discrepant definitions, means sometimes merely represent the centers of bins, giving little information about preferred conformations.

The third advantage of modes arises when two or more peaks are close together, such as for the leucine χ^1/χ^2 case discussed below. Drawing bins at 0°, 120° and -120° puts two separate peaks in the **tt** and in the **mp** regions. The mean for each of these two regions lies in between the clusters, which has resulted in clashing Leu **tt** and **mp** rotamers for every previous library. In contrast, determining the modes shows two distinct peaks 60-70° apart which can be analyzed separately.

Lastly, if the observed distribution is converted to an energy equivalent, it is the mode rather than the mean that corresponds to the lowest-energy conformation.

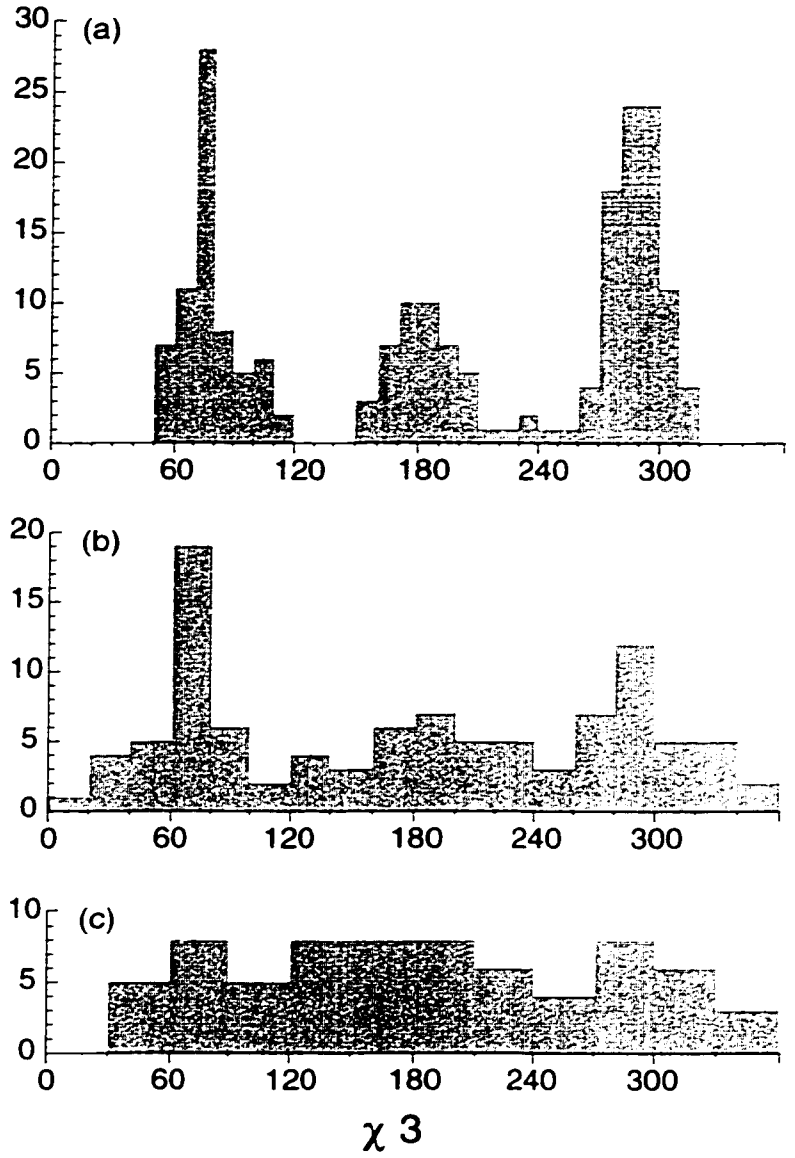
A disadvantage of the modal-value approach is that it requires smoothing to determine the mode reliably. In addition, for clusters with low total population, both mean and mode are susceptible to statistical fluctuations, but the mode is somewhat more so. The common-atom angle definitions, although adopted for other reasons, help avoid most of the small-population problems.

It has recently been found (MacArthur and Thornton 1999) that rotamer mean values change systematically with resolution, at least in part because of averaging between unresolved alternate conformations, which produces skewed distributions at lower resolution. In one case (Leu), part of the shift is caused by a misfitting more common at low resolution (see below), but the general point remains valid and important. Although anomalous behavior of the means uncovered this interesting relationship, the modal values as seen in the data described in that study do not shift from the rotameric positions, again suggesting that modes are preferable for most purposes.

Met Rotamers

As well as helping improve the accuracy of structure determination and quantifying packing contacts, the interactions shown by small-probe contact surfaces can provide more memorable and intuitive illustrations for conformational regularities already understood, or they can provide explanations for conformational features described only as empirical regularities. As one example, we will consider the side-chain rotamers of methionine.

FIGURE 7-4. Met χ^3 angle distributions: (a) for the 241 Met side chains with $B <$ 30 in the *Top100DB1*; (b) for the 90 Met with $B \geq 30$ in our dataset; (c) data recorded from Janin *et al.* (1978).

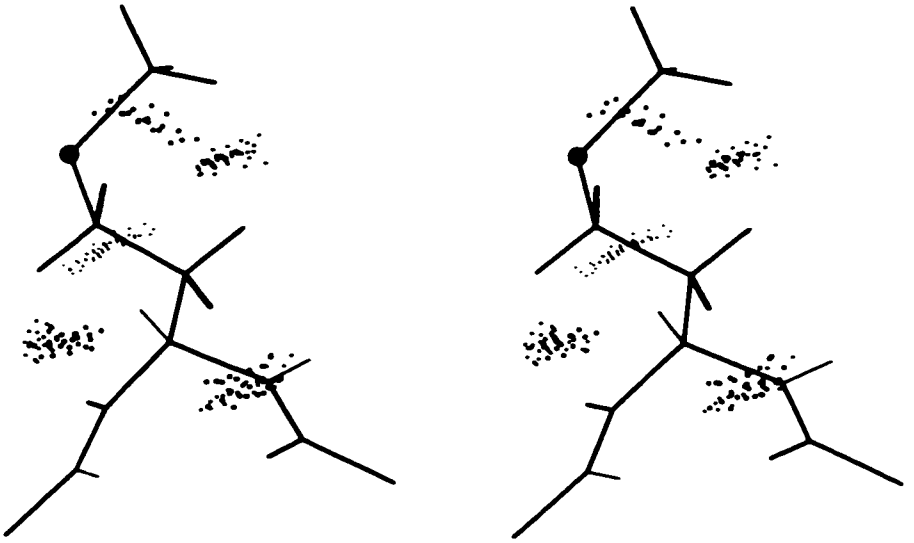


Because methionine is one of the rarest amino acids and has three variable χ angles, its conformational preferences have historically been poorly described. Met χ^1 and χ^2 are not a problem, because their single-angle distributions are very like those of Glu, Gln, Arg, and Lys: +65° rarest in both cases, with –65° preferred for χ^1 and *trans* for χ^2. However, early χ angle surveys (Bhat, Sasisekharan, and Vijayan 1979; Chandrasekaran and Ramachandran 1970; James and Sielecki 1983) were forced to omit Met χ^3 altogether because they had too few examples for analysis, while Ponder & Richards (1987) list only one rotamer with all three angles: **mmm**, indeed now confirmed to be the most common Met rotamer. Benedetti *et al.* (1983) used additional data from small peptide structures, but they combined Met χ^3 with Arg and Lys χ^3, which now turn out to have quite different distributions. Janin *et al.* (1978), using 19 proteins at up to 2.5 Å resolution, found an essentially flat distribution across the whole range of Met χ^3, except for a dip at 0° (see Figure 7-4(c)); that observation is quoted by Gellman (1991) and Schrauber *et al.* (1993). Tuffery *et al.* (1997; 1991) energy minimized the side-chains before clustering; they used all three χ angles in defining discrete Met rotamers, but did not show distributions or standard deviations, and mean χ^3 values for their rotamers span the entire range except near 0°, including two nearly eclipsed at 131° and 140°. The rotamers provided in the O rebuilding program (Jones *et al.* 1991) generally follow Ponder & Richards (1987), but for Met they assume trans χ^3 where it was not specified (which is actually never the most common alternative) and include one impossible **tpm** rotamer with a 0.8 Å clash of C^{ε} to H^{α}.

Dunbrack & Cohen (1997) solved the sample size problem by using 518 protein chains at 2 Å or better resolution, an order of magnitude more than any of the earlier surveys, and their Met rotamer library was clearly the best so far. However, like all of their predecessors, in order to maximize sample size (which they need for study of ϕ,ψ dependence), they use all residues, including those with high *B*-factors, which adds in a component with high random noise. As essentially every one of these authors has pointed out, long external side-chains are often poorly determined, especially toward their ends. Crystallographic *B*-factors are explicitly designed for identifying uncertain regions and, as documented in Chapter 4, a *B*-factor cutoff can eliminate a large fraction of the problems without deleting too large a fraction of the data. In order to see this effect for χ angle distributions, Figure 7-4(a) and (b) compare Met χ^3 values with $B < 30$ *versus* those with *B*-factors for some atom ≥ 30, for the methionine residues in our database. Higher *B*-factors, as well as lower resolution, act to spread out what should be a sharply clustered distribution.

The most obvious conclusion from the high-resolution, low-*B* distribution in Figure 7-4(a) is that, in contrast to most earlier analyses, we find Met χ^3 to be quite "rotameric", with 94% of χ^3 values clustered within 30° of the three means (**p** = +75°, **t** = 180°, and **m** = –70°). The χ^3 patterns also differ with χ^2 value: for example, the shoulder seen near $\chi^3 = +100°$ in Figure 7-4(a) is real, caused by the **mmp** rotamer with a mean of 101° for χ^3.

FIGURE 7-5. Stereo diagram of the **mmm** rotamer for an ideal-geometry Met residue, with the backbone α-helical, showing contact dots for the five good H atom contacts.



The second conclusion is that in marked contrast to the strong *trans* preference seen for aliphatic χ angles (e.g. in Lys), the χ^3 values for Met prefer a *gauche* conformation over *trans*, with −70° the most favored (the **p:t:m** ratios are 35:23:42 % in our data and 36:23:41 % for Dunbrack & Cohen (1997)). Gellman (1991) first discussed this issue, pointing out that the modest clash at *gauche* values for aliphatics is absent for Met χ^3, but stating strong puzzlement that *gauche* is actually preferred rather than just less disfavored. Using contact dots calculated for a Met side-chain with idealized geometry, we can show that there is substantial favorable H^{ε}–H^{β} and H^{ε}–H^{γ} contact when χ^3 is near ± 75° (see Figure 7-5). For *trans* angles, the dots show a slight H^{ε} -H^{γ} contact in aliphatic side-chains but none at all for Met, because of the longer C-S bond.

Of the 27 possible rotamer bins for methionine, we find only about half to be significantly populated: 13 have frequencies > 2%, while seven are completely empty in our 100-protein dataset. As in all previous treatments that included any full rotamers for Met (Dunbrack and Cohen 1997; McGregor, Islam, and Sternberg 1987; Ponder and Richards 1987; Tuffery *et al.* 1991), the **mmm** rotamer was found to be the most common. Figure 7-5 illustrates that the **mmm** rotamer can make five good H atom contacts if the backbone is in the α conformation; it has four good contacts in β conformation. The two next-most-common Met rotamers share a similar pattern of three such contacts, but in **mtp** $2H^{\varepsilon}$ touches $2H^{\beta}$, while in **mtm** $3H^{\varepsilon}$ touches $1H^{\beta}$. An analogous mirroring of **mmm** to produce **mpp** does not occur because the S atom would clash with backbone. Avoidance of clashes is indeed the strongest constraint, but patterns of conformational preference can be better explained if favorable contacts are taken into account.

Our observed occurrence frequencies for all of the Met rotamers agree closely with the backbone-independent distributions for Met given by Dunbrack & Cohen (1997): the percentage population for 23 of the 27 possible rotamers agrees to within ± 1%, and within ± 3% for the other four. Most of those small differences come from higher contrast in our data: we see 20% rather than 17% of the most common rotamer (**mmm**), and a total of only 5% rather than 7% in the 14 least populated ones. We believe this represents an improvement in accuracy, due to the use of a *B*-factor cutoff. Mean χ values for the populated Met rotamers differ from those of Dunbrack & Cohen (1997) by a population-weighted average of only 2.4°. Since the two databases and methodologies are quite independent, this agreement

implies that the mean angle and population values are reliable. The remarkable thing, however, is that our data produces these same answers with only one-eighth as many methionine residues (244 *versus* 2068). In spite of the smaller dataset, our rotamer peaks are more sharply defined: for the 13 populated rotamers, none have standard deviations significantly higher than those from Dunbrack & Cohen (1997), while 44% are significantly lower by *F* test at the 5% level and many are only half as large. These results are a tribute to the merits of both *B*-factor cutoffs and very high-resolution data.

Lysine–the Statistically Simple Side Chain

Lys and Arg, with four χ angles each, have 81 possible staggered rotamers; Met, with three χ angles, has 27. Jane Richardson proposed exploring whether a compact description would suffice, with just a few rules that applied to multiple cases. The attempt failed for Arg and Met, where analogous sets of rotamers show relative frequencies differing by factors of 3 or more, presumably responding to circumstances such as non-uniform patterns of possible H-bond partners. However, Lys rotamers show reproducible patterns of relative frequencies that can be accurately predicted using only a few physically-reasonable parameters, as shown in Table 7-1. Two parameters are the relative preferences for χ^1 **t** (0.65) and **p** (0.13) as a fraction of **m**. Two additional parameters are penalties for the "*syn*-pentane" (Dunbrack and Karplus 1994) conflicts that occur when adjacent *gauche* angles change signs (**mp** or **pm**); one of those penalty factors (0.11) applies for χ^2/χ^3 or χ^3/χ^4 on the unbranched side chain, and a more severe one (estimated as 0.05) applies for χ^1/χ^2 which has backbone atoms on one end. Such *syn*-pentane cases also result in shifted

TABLE 7-1. Lysine rotamer simplified predictions[a]

	pred.[b]	obs.		pred.	obs.		pred.	obs.
pppp	0.0	0	tppp	**1.5**	**2**	mppp	0.1	0
pppt	0.1	0	tppt	**6.4**	**3**	mppt	0.5	0
pppm	0.0	0	tppm	0.2	0	mppm	0.0	0
pptp	0.1	0	tptp	**7.4**	**11**	mptp	0.7	0
pptt	0.3	0	tptt	**32.2**	**32**	mptt	**2.4**	**4**
pptm	0.1	0	tptm	**7.4**	**7**	mptm	0.6	0
ppmp	0.0	0	tpmp	0.0	0	mpmp	0.0	0
ppmt	0.0	0	tpmt	0.7	0	mpmt	0.1	0
ppmm	0.0	0	tpmm	0.2	0	mpmm	0.0	0
ptpp	**1.5**	**1**	ttpp	**7.4**	**12**	mtpp	**11.2**	**12**
ptpt	**6.3**	**7**	ttpt	**32.2**	**25**	mtpt	**48.8**	**38**
ptpm	0.2	0	ttpm	0.8	**2**	mtpm	**1.2**	**1**
pttp	**7.3**	**13**	tttp	**37.0**	**49**	mttp	**56.1**	**42**
pttt	**31.7**	**29**	tttt	**161.0**	**162**	mttt	**244.0**	**244**
pttm	**7.3**	**8**	tttm	**37.0**	**37**	mttm	**56.1**	**56**
ptmp	0.2	0	ttmp	0.8	0	mtmp	**1.2**	**2**
ptmt	**6.3**	**5**	ttmt	**32.2**	**20**	mtmt	**48.8**	**40**
ptmm	**1.5**	**2**	ttmm	**7.4**	**5**	mtmm	**11.2**	**12**
pmpp	0.0	0	tmpp	0.0	0	mmpp	0.2	0
pmpt	0.0	0	tmpt	0.0	0	mmpt	**1.1**	0
pmpm	0.0	0	tmpm	0.0	0	mmpm	0.0	0
pmtp	0.1	0	tmtp	0.4	0	mmtp	**11.2**	**9**
pmtt	0.3	0	tmtt	**1.6**	0	mmtt	**48.8**	**77**
pmtm	0.1	0	tmtm	0.4	**2**	mmtm	**11.2**	**18**
pmmp	0.0	0	tmmp	0.0	0	mmmp	0.2	**2**
pmmt	0.1	0	tmmt	0.3	**1**	mmmt	**9.8**	**10**
pmmm	0.0	0	tmmm	0.1	**1**	mmmm	**2.2**	**1**

a. Rules: value set equal for rotamer with greatest number of observations (boxed); for the rest, multiply by a factor of 0.2 for each *gauche* χ^2 or χ^3; by 0.23 for each *gauche* χ^4; by 0.65 for χ^1 **t**; 0.13 for χ^2 **p**; by an additional 0.11 for χ^2/χ^3 or χ^3/χ^4 **mp** or **pm**; and by an additional 0.05 for χ^1/χ^2 **pp**, **tm**, **mp**, or **pm**.

b. Predicted and observed frequencies of 1.0 or greater are shown in bold.

χ values to avoid the clash, such as $\chi^1 = -90°$ for Lys **mptt** or $\chi^3 = 103°$ for Met **mmp**.

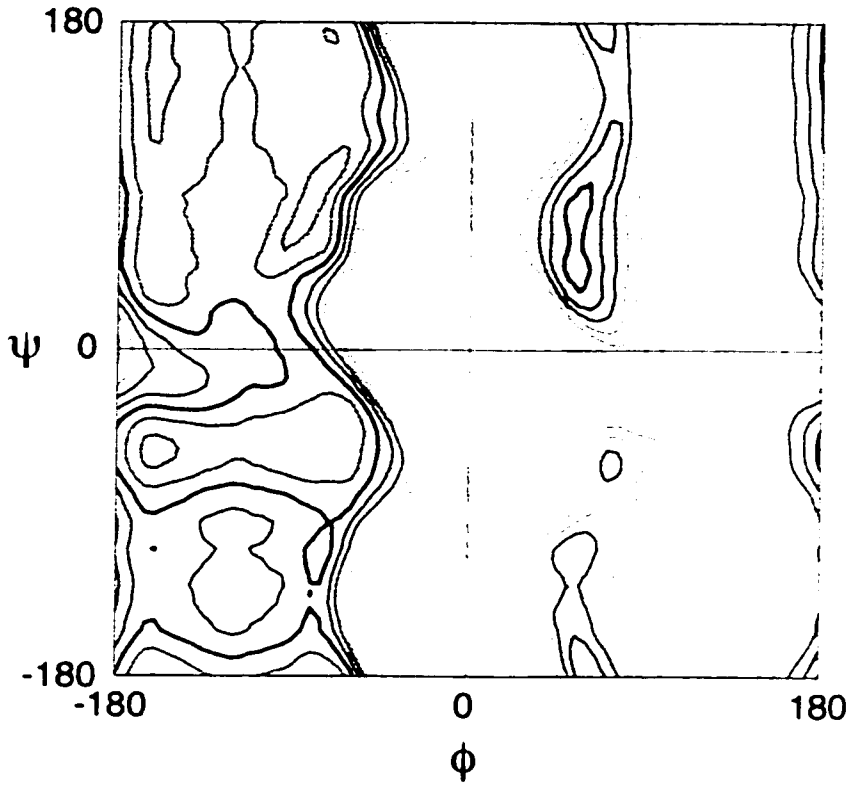
The most interesting parameter is the penalty for having a *gauche* angle in Lys χ^2, χ^3 or χ^4. It can be estimated separately for one-*gauche*, two-*gauche* and three-*gauche* rotamers relative to the cases where χ^2, χ^3, and χ^4 are *trans*, avoiding any comparisons that contain **mp** or **pm** combinations: those factors are found to be 0.21, $(0.22)^2$, and $(0.20)^3$, suggesting that the parameters are independent and simply multiplicative. Also, the experimentally-measured energy difference of 0.89 kcal/mol between *gauche* and *trans* butane (Wiberg and Murcko 1988) can be converted using the Boltzman relationship $E = RT\ln P = 0.592\ln P = 1.364\log P$ to give a *gauche* factor of 0.22.

An overall least-squares fit of the parameter values to all 81 Lys rotamer frequencies was done by minimizing the sum of squares using Mathematica (Wolfram Research 1996). Since the Lys NH_3^+ terminal group is smaller than a methyl, six parameters were used, and the *gauche* penalty was fit with one value for χ^2 and χ^3 and a separate value for χ^4, which came out as 0.20 and 0.23 respectively. The predictions in Table 7-1 are obtained by multiplying all applicable factors for each rotamer; the correlation coefficient between predicted and observed (Pearson's r) is 0.993. An estimate of the relative rotamer pseudo-energies can be made by adding up the energies for each applicable penalty factor, so that, for instance, **mpmp** acts as though it is 6 Kcal/mol less favored than **mttt** and does not occur in our data set.

Note that the strong preference of Lys for *trans* χ angles (about 1:5:1 **m:t:p**) is real and is not a result of fitting disordered side chains as trans. Not only are the ratios consistent across all rotamers, but also the contrast is lower, not higher, at high *B* and is significantly lower for χ^4, which is not consistent with an effect from increasing uncertainty. In contrast, Met χ^3 prefers *gauche* by more than 2:1, since the *gauche* form not only does not clash, but actually has favorable H-atom van der Waals contacts (see page 167).

Presumably the reason Lys behaves in such a statistically simple fashion is that although the end makes charged H-bonds, the geometry of those interactions is relatively unconstrained with the side chain having so many degrees of freedom that it can usually get to its appropriate position without strain. Given that the high-resolution, low-*B* lysines very seldom have any χ angles as much as 30° from staggered and populate the less-favored rotamers only as often as dictated by their pseudo-energy differences, it seems completely unjustified ever to fit partially disordered lysines with eclipsed angles or poor rotamers.

FIGURE 7-6. Ramachandran-like map of allowable main-chain ϕ and ψ angles based on small-probe contact scores calculated using autobondrot.



Rationalizing Observed Conformations with a Hard-sphere Model

Ramachandran Map

An almost universal step in protein structure validation categorizes observed main-chain conformations as either 'allowed', 'marginal', or 'disallowed' based on the steric constraints on polypeptide chain conformation, worked out by Ramachandran et. al. (1963) and famously summarized in their ϕ, ψ map. A similar diagram can be constructed from all-atom small-probe contacts (Figure 7-6). Using a prototype of

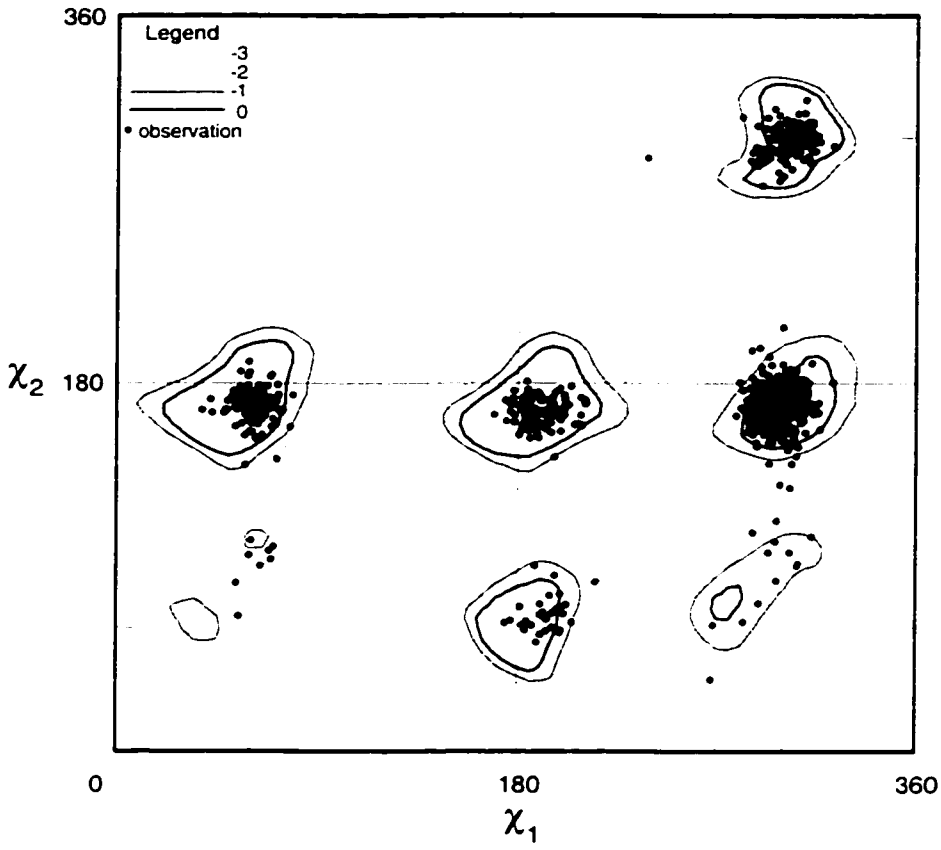
the *autobondrot* function described in the next chapter, the -3 -self PROBE score (see Chapter 2) for an articulatable computer model of a single ideal-geometry alanine dipeptide was sampled every 5° at all discrete combinations over the full range of ϕ and ψ main-chain dihedral angles. In the figure, contact scores at each conformation have been contoured; the shaded regions have favorable van der Waals interactions and H-bonds, while open regions have unfavorable clashes. Slightly unfavorable regions (such as "the shoal": $\phi = 60°$, $\psi = 170$ to $-60°$) are occasionally observed, whereas the region in the center near $\phi = 0°$, $\psi = 0°$ has large clashes which make it an impossible conformation. The correspondence between our map and the classic Ramachandran map confirms the set of parameters used to generate our map. Since our method did not include any electrostatic repulsion there are small regions with acceptable scores near the alpha and beta conformations that actual peptides do not populate (e.g., near -160, -55 where successive NH's touch), just as in the classic map.

Side-chain Maps

Here we show the extent to which our all-atom hard-sphere model, combined with a simple torsional potential, describes and rationalizes observed distributions of side-chain conformations. High-quality, high-resolution data continues to confirm even further the "rotameric" concentration of protein side chains into discrete sets of allowed χ angles. As the quality of the data on sidechain conformations has improved, the distributions of the "rotamers" have sharpened. Furthermore, signifi-

cantly non-rotameric outliers almost always turn out to be doing something interesting.

FIGURE 7-7. Contour map of small-probe contact scores (plus a torsional term) for all side-chain conformations, including methyl rotations, of an isolated isoleucine calculated using **autobondrot**, along with conformations measured for Ile's with $B < 20$ from the database of 240 structures described in the text.



The χ^1-χ^2 map for isoleucine in Figure 7-7, showing the match between contours for all-atom probe scores and points for the most reliable experimental observations, serves to illustrate that non-polar side chains are almost completely constrained by van der Waals and torsional constraints. The contact scores were

calculated from conformations in standard Engh and Huber (1991) geometry, where the side-chain dihedrals were sampled every 5°. Both methyls were also allowed to rotate, which they did up to 20°, and the maximum score was used. These scores are independent of backbone conformation because the main chain was chopped down to just N, C^{α}, H^{α}, and C. The *total score* is the sum of the PROBE score plus a cosine function representing a torsional barrier for both χ^1 and χ^2; total scores above −1 are said to be "acceptable." See Chapter 8 for a more complete description of the *total score* and the acceptability criterion. The individual Ile conformations (along with those for other side chains discussed below) are from the database of 240 crystal structures at a resolution of 1.7 Å or better, developed to create the improved rotamer library described above. Side chains were rejected if they had multiple conformations, a van der Waals overlap of ≥ 0.4 Å, or a *B*-factor ≥ 20.

The level of correspondence between observation and model is similarly good for leucine (Figure 7-8), with one interesting exception. Observations in the circled regions (labeled **tt*** and **mp***) with poor PROBE scores appear to have been fit with χ^2 "backwards"—a type of error which has long been known to occur with Val and Thr side chains (in which case it produces a doubly-eclipsed conformation) when the electron density for the branched terminal group is ambiguous in shape; standard refinement programs will usually distort the model rather than rotate the χ angle by the 140°- 180° needed.

FIGURE 7-8. Contour map of small-probe contact scores (plus a torsional term) for all side-chain conformations, including methyl rotations, of an isolated leucine calculated using autobondrot, along with conformations measured for Leus with $B < 20$ from the database of 240 structures described in the text. "Misfit" conformations **tt*** and **mp*** are circled, with arrows pointing to a preferable rotamer that occupies approximately the same region in space.

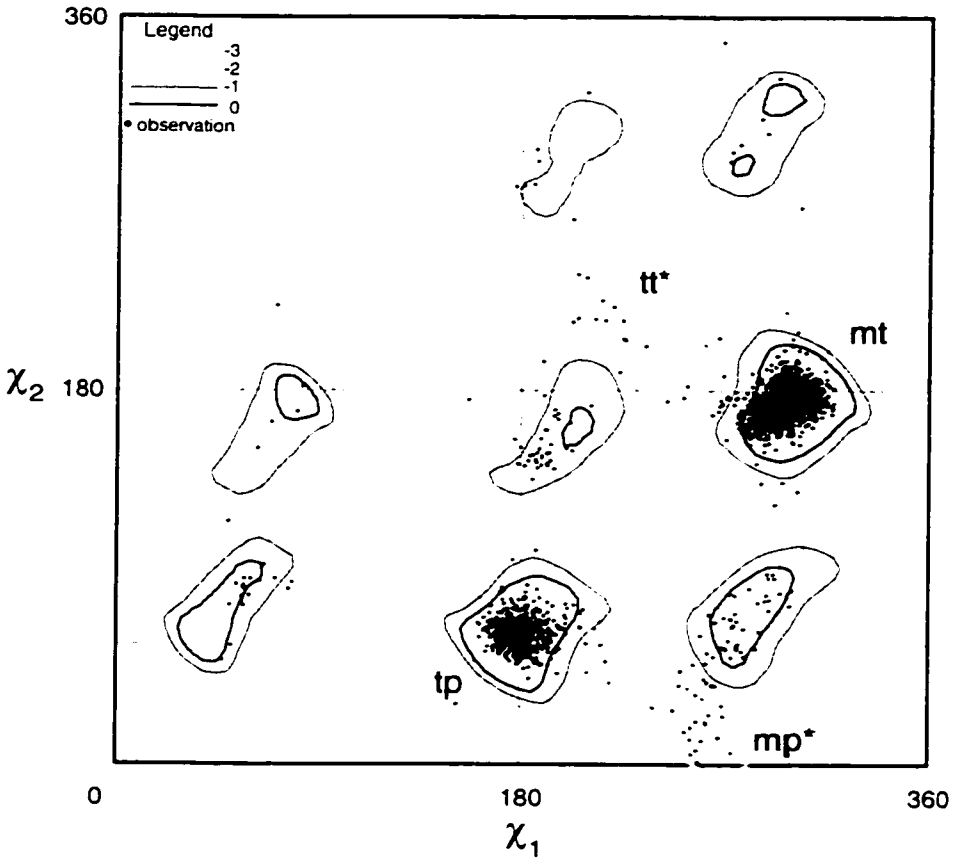
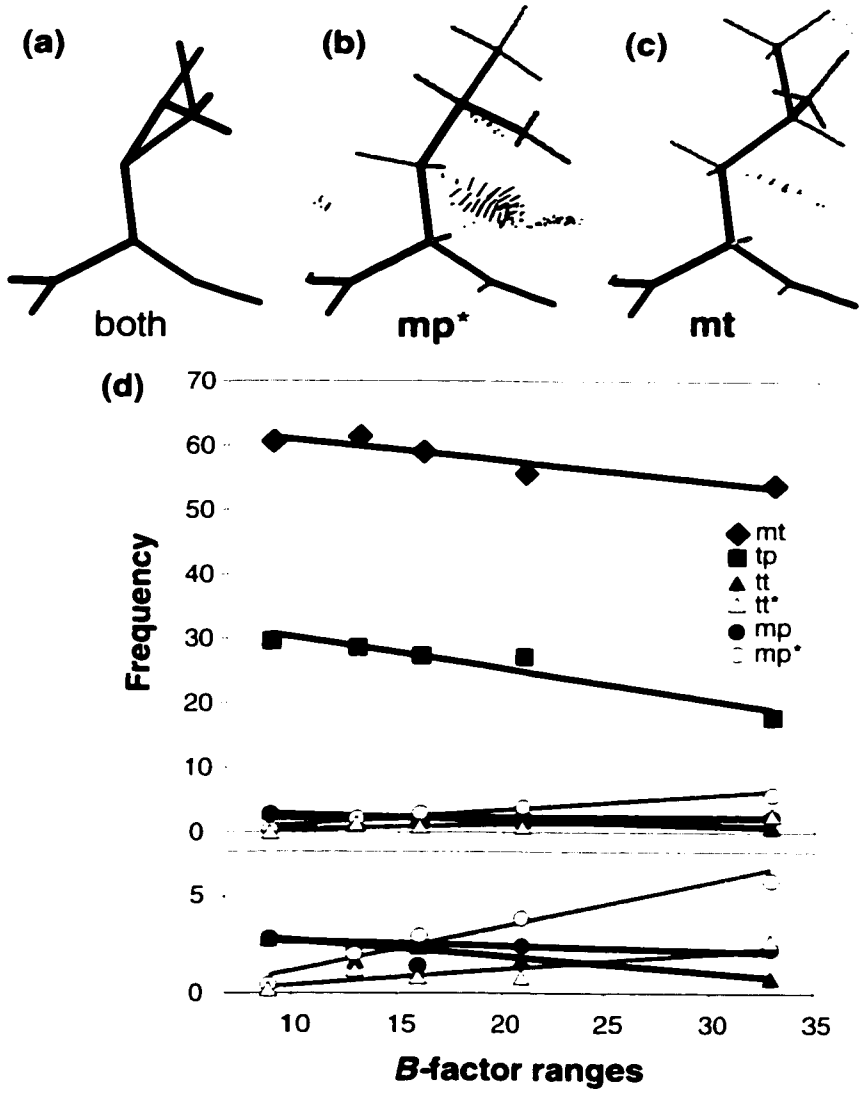


FIGURE 7-9. (a) Superposition of a favorable Leu conformation (**mt**) and its misfit partner (**mp***), without H atoms. (b,c) A comparison of the structures and their contacts for a genuine rotamer (**mt**, at right) *versus* its misfit partner (**mp***, center); only overlapping contacts are shown, as calculated with PROBE and displayed in MAGE. (d) Correlation of rotamer frequency with *B*-factor for both genuine and misfit Leu rotamers. *B*-factor bins were constructed to contain the same number of points in each bin for the whole distribution. The lower panel is an enlargement of the bottom section of the main plot to show more clearly the slope of the line for the rarer rotamers. Systematically misfit rotamers (**tt*** and **mp***) are indicated by open symbols.



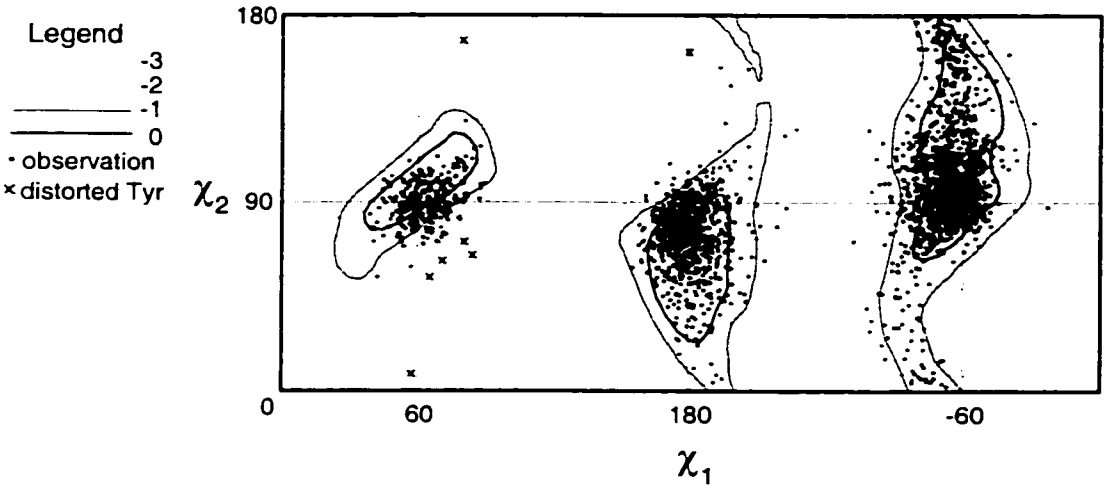
Another way of describing this situation is that Leu has two pairs of conformations that occupy approximately the same physical space: **mt** *versus* **mp*** and **tp** *versus* **tt***, one pair of which is shown in Figure 7-9(a). The ability to superimpose the C^{δ} atoms of Leu by rotating χ^1 by 30° to 40° and χ^2 by 140° to 150° from some starting positions has been noted before, mainly for **mt** *versus* **mp*** (Dunbrack and Karplus 1993; Lee and Subbiah 1991; MacArthur and Thornton 1999; Petrella, Lazaridis, and Karplus 1998). However, none of those authors reached a conclusion as to which of the apparently equivalent conformations is preferable. On the other hand, Kuszewski et al. (1996) discussed the probability of Leu misfittings and changed the less common **mp*** and **tt*** forms by 40° and 140° in their data, but they gave no additional evidence besides the inherent plausibility of that decision. Here and in Lovell *et al.* (2000) we analyze other sources of information that can resolve this ambiguity.

For each of the above pairs, one conformation is one of the two highly-favorable major Leu rotamers (Figure 7-9(c)), while the other alternative has a severe clash when built in standard geometry with explicit hydrogens (Figure 7-9(b)). **mp*** has an atomic overlap of 0.6 Å between the $C^{\delta 1}$ and the H^{α}, which of course is why it lies in an unfavorable region of the Leu PROBE-score map in Figure 7-8. The strongest piece of evidence for rotamer correctness is a positive correlation with map quality (i.e., either resolution or *B*-factor), whereas a misfit conformation should correlate negatively. Figure 7-9(d) shows the variation of Leu rotamer occurrence with *B*-factor. Those rotamers we define as genuine become more common as *B*-

factor decreases, whereas the flipped conformations **tt*** and **mp*** become less common.

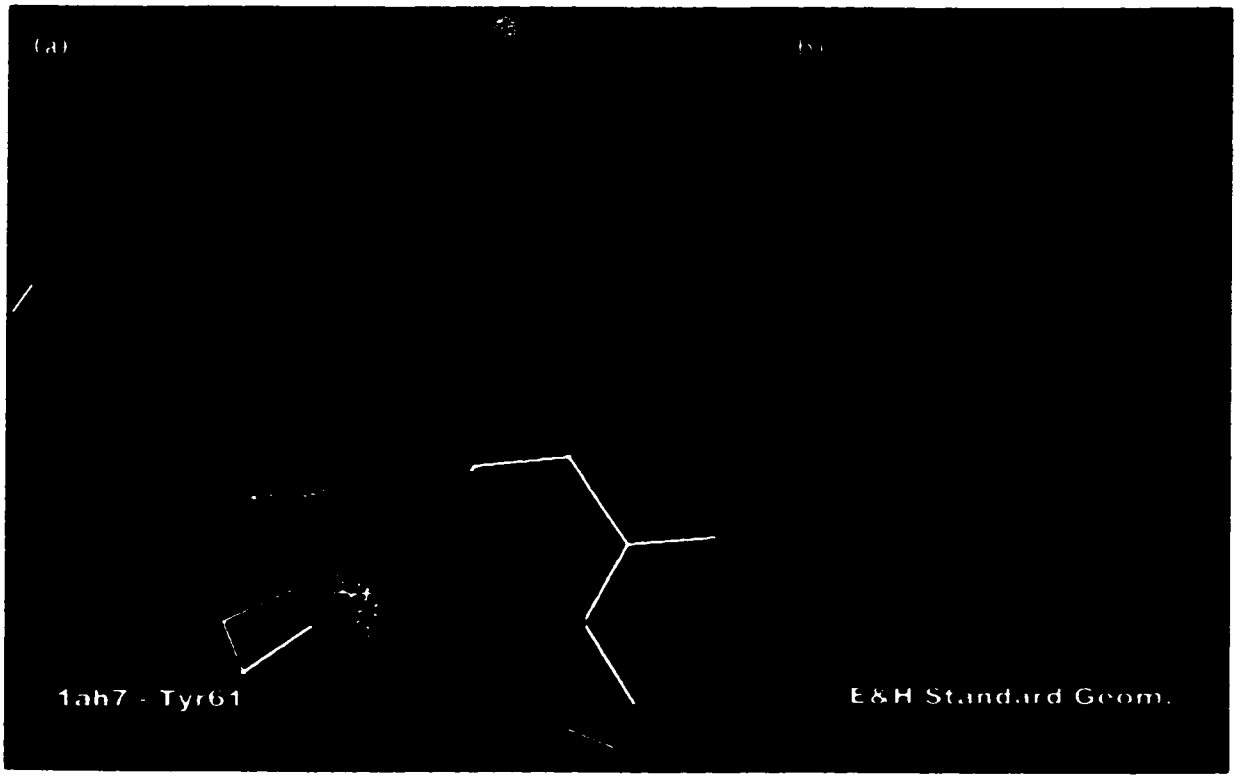
For all of the above reasons, we conclude that the **mp*** and **tt*** conformations are very unlikely to be correct. We simply omit them from our data rather than transforming them to the two major peaks, because these misfittings usually cause movement of backbone atoms, and their transformed coordinates would be unreliable. After the backward leucines are omitted, there remains a valid rotamer cluster in each of the **tt** and **mp** areas (Figure 7-8) which is clash-free and shows the correct *B*-factor dependence (Figure 7-9(d)). Because **tt*** and **mp*** are more numerous at lower resolution and higher *B*, every previous rotamer defined for Leu **tt** or **mp** has either been between the two clusters or in the incorrect one, setting up an unfortunate feedback cycle in which bad rotamers and individual misfit side chains reinforce one another's plausibility. By omitting these misfit examples, we both improve our rotamer library and also improve the match between observed χ distributions and calculated PROBE-score maps.

FIGURE 7-10. Contour map of small-probe contact scores (plus a torsional term along χ^1) for all side-chain conformations of an isolated phenylalanine residue calculated using **autobondrot**, along with side-chain conformations measured for Phe and Tyr residues with $B \leq 20$ from the database of 240 high-resolution structures described in the text. Conformational outliers for tyrosines with distorted bond angles are marked x.



The three allowed χ^1 regions for phenylalanine and tyrosine shown in Figure 7-10 each have markedly different shapes, reflecting the different sizes of the residue's N, H^{α}, and C atoms, whose collisions cause the forbidden regions. The contoured scores include a torsional term only for χ^1. Conformational outliers, marked with an X, are all tyrosine residues with distorted bond angles. Their angles around C^{α} and C^{β} increase by about 4° to avoid large clashes with the main chain. Inspection of these cases shows that they are unlikely to be misfit; each is well packed with surrounding atoms and several form H-bonds to main chain.

FIGURE 7-11. Structure and all-atom contacts for a well-determined tyrosine with an outlier conformation, Ty61 of 1AH7 (Hough *et al.* 1989): (a) in the deposited conformation, with distorted bond angles but excellent contacts. (b) as built with the same χ angles in ideal geometry, showing a serious clash between ring and backbone.



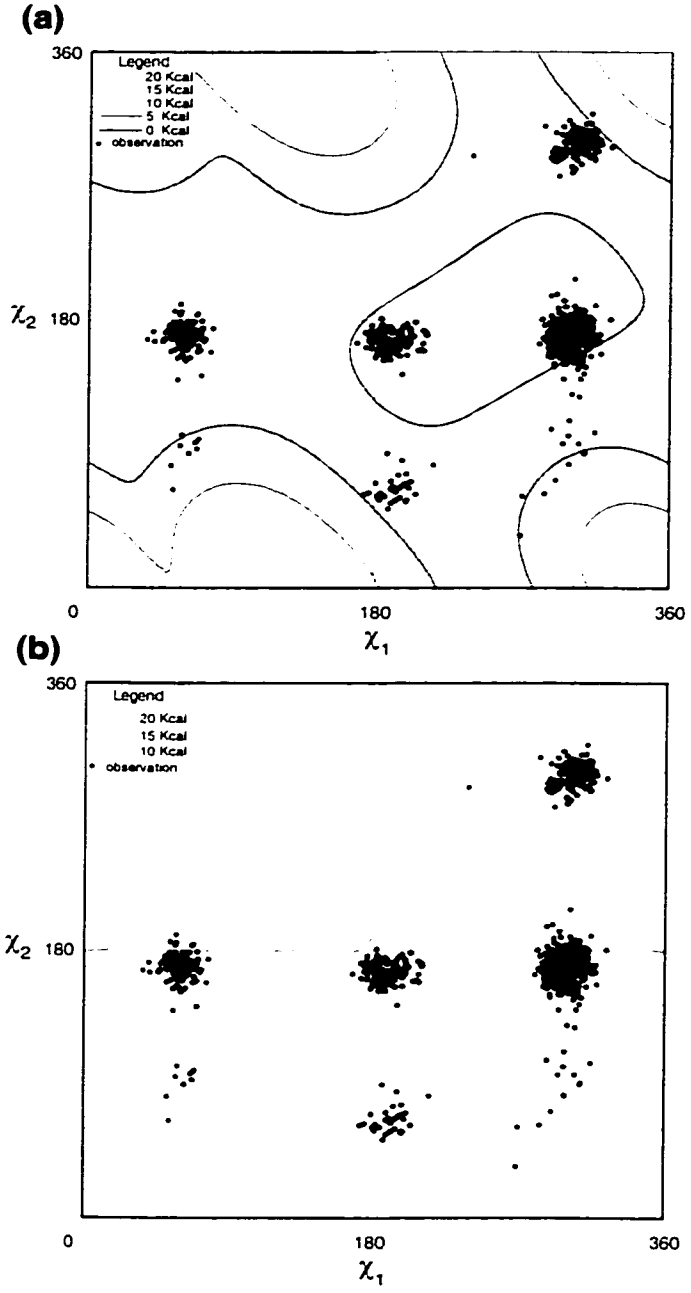
As a specific example, outlier tyrosine 61 from phospholipase C (1AH7, $\chi^1 = 79°$, $\chi^2 = 168°$) in Figure 7-11(a) makes excellent contact with nearby side chains and does not clash with the backbone. Such a low-*B* aromatic with this level of contact surface complementarity is virtually certain to be fit correctly. In that same conformation, however, the standard-geometry model used to generate PROBE scores shows an overlap of 0.6 Å with the backbone carbonyl carbon (red spikes in Figure 7-11(b)). The difference is due to wider bond angles in Tyr61, especially 119° for C^{α}–C^{β}–C^{γ} which is nearly 3σ above the Engh and Huber mean (1991).

In general, side chains adopt a relaxed conformation and information about rotameric conformations is very useful when building a structure, but rotamers are not the whole story. An outlier conformation is not necessarily a mistake: a molecule may require it for function, for example, and have a way to compensate. All-atom contact surface analysis distinguishes these rare, strained side chains from their relaxed neighbors.

Comparison to Energy

Recent successful protein (re-)designs (e.g., Dahiyat and Mayo 1997b; Desjarlais and Handel 1995) have essentially all used explicit hydrogens—as we do—and have shown that successful modeling of hydrophobic packing in protein interiors requires van der Waals and torsional potentials. These are usually described in energy terms, but energy maps are often broader and less discriminating than the geometrical models described here. PROBE scores are not energies *per se*; instead they characterize surface complementarity.

FIGURE 7-12. Energy contours for isoleucine side-chain conformations calculated using the CHARMM force field as implemented in X-PLOR, combined with conformations measured for isoleucines with B < 20 from the database of 240 high-resolution structures described in the text. (a) Contours calculated using the united atom approximation and the default set of energy terms. (b) Improved match to observations using explicit hydrogens and only van der Waals and torsional energy terms.


(a)
360
Legend
20 Kcal
15 Kcal
10 Kcal
5 Kcal
0 Kcal
observation
χ2 180
0
180
360
χ1
(b)
360
Legend
20 Kcal
15 Kcal
10 Kcal
observation
χ2 180
0
180
360
χ1

Comparison of the energy contours for Ile side-chain conformations in Figure 7-12(a) with the score contours in Figure 7-7, is striking. The energies use the CHARMM force field with "united atom" *implicit* hydrogens, as implemented in X-PLOR (Lovell and Word, manuscript in preparation). Although the observations fall within low energy regions, there are large areas with similarly low energies but containing no observations. Using an approach more similar to contact analysis—including explicit hydrogens and calculating energies based on only the dihedral and van der Waals terms (the position of the methyls is also optimized)—the match to the span of observed conformations is significantly improved (Figure 7-12(b)), although contact scores (Figure 7-7) are still the better match.

Contact scores for isolated residues are good at identifying the boundary between acceptable and unacceptable conformations; the match to relative frequency is less accurate. For instance, the scores do not properly rank the two most frequent leucine rotamers. Other factors, such as interactions with neighboring residues and backbone, may be required before either scores or energies can be used to predict likelihood.

The contact analysis shown here for non-polar side chains appears, somewhat surprisingly, also effective at predicting relaxed conformations for more polar residues. Work is in progress to compare score maps with observed conformations for all the amino acids as part of a project to develop a new "rotamericity" scale for use in structure validation. Rule based models, such as the system shown above for lysine, do not seem to be widely applicable.

Our deliberately simplistic hard-sphere model of atomic interaction emphasizes the dominance of steric and torsional effects in determining side-chain conformation in actual proteins. The inclusion of van der Waals terms for explicit hydrogens, done with increasing frequency in the computational chemistry community, is probably the most important extension being advanced in current methods for accurate energy calculations.

Discussion

Significance of Contact Analysis Tools for Crystallography

Protein crystallography, molecular graphics, and refinement (or energy calculations in general) are mature fields. It is rather surprising, then, that the simple, geometric approach of the all-atom contact dot method has something new to contribute that bridges those areas. It leads to easy and intuitive visualization of interactions that were previously hidden or implicit, and in quantitative form can actually account for observed conformational distributions better than the traditional energy calculations do. The revolutionary accuracy of atomic-resolution crystal structures can not only be directly visualized but, surprisingly, can be improved in some details. Most importantly, the contact analysis technique can enhance the accuracy and reliability of structures at any given level of resolution.

In addition to use in individual structure determinations, the all-atom steric analysis methods lend themselves very well to further automation of the structure determination process with minimal sacrifice of quality control. Both those aspects are

increasingly valuable in this time of database growth and the structural genomics initiative (Montelione and Anderson 1999; NIGMS 1999; Sali 1998). A widespread effort is beginning: to solve a structure in each protein family—on the order of 10,000 structures—requiring substantial increases in speed, automation, and reliability at every stage of the process. Although the production and crystallization steps are probably the most difficult, the fitting, refinement, and validation stages will also need to be done with much less time-consuming human intervention than at present. We are very eager for the results of such work, but are concerned about the level of accuracy that current black-box methods would achieve. Since all-atom contact analysis seems to provide a novel, independent, and very sensitive method of detecting and fixing errors, its application to structural genomics will be both timely and important. The addition and optimization of hydrogens and the correction of Asn/Gln/His flips are already automated. The fitting tools and rotamer library are already well integrated into current structure-solution methods, but an important future direction will be for further automation and integration into the systems now being developed for high-throughput crystallography.

Comparison of the New Rotamer Library with Earlier Ones

For the simpler amino acids and the most common rotamers, all libraries, of course, agree quite well, at least in existence and position if not always in probability. For the rarer rotamers and the more difficult residue types (including Lys, Arg, Met, Leu, Gln, Glu, Asn, Asp, and Pro), there are at least three factors governing disagreements between this and previous work. Growth in the database is crucial to

such efforts, but here it is not the most decisive issue: our raw data are essentially indistinguishable from those of Dunbrack and Cohen (Dunbrack and Cohen 1997).

The second factor is the development of our new methods for optimizing explicit H positions (Word *et al.* 1999b) and representing all-atom contacts clearly and dramatically (Word *et al.* 1999a). If graphics such as Figure 7-3 and Figure 7-9(b) had been available to earlier authors, their rotamer lists would almost certainly have been affected. The all-atom contact analysis, in both visual and quantitative forms, was essential when discarding from the present library a significant number of previous rotamers now shown to represent flipped amides or systematic fitting errors. On the other hand, this process helped in validation of a relatively large set of well-behaved rotamers down to the level of 1-2% occurrence probability.

A third, more complex, factor covers differences in choice of definitions and methodologies. Some disagreements arise from blurring the distinction between a true rotamer (i.e., a locally favored conformation with clustered examples) and an arbitrary sample point in conformation space. Many computational uses of rotamers require additional sampling within the allowed regions, but such sample points are not real rotamers because their spacing and position depends on their intended use, not on the properties of the side chain conformations.

Most earlier work used the mean (average) value as the rotamer position, whereas we use the mode (peak occurrence), which has the important advantages of corresponding to the local energy minima and of being sensitive to closely-spaced peaks while independent of skewed peak shape or of arbitrarily-defined bins. As was done

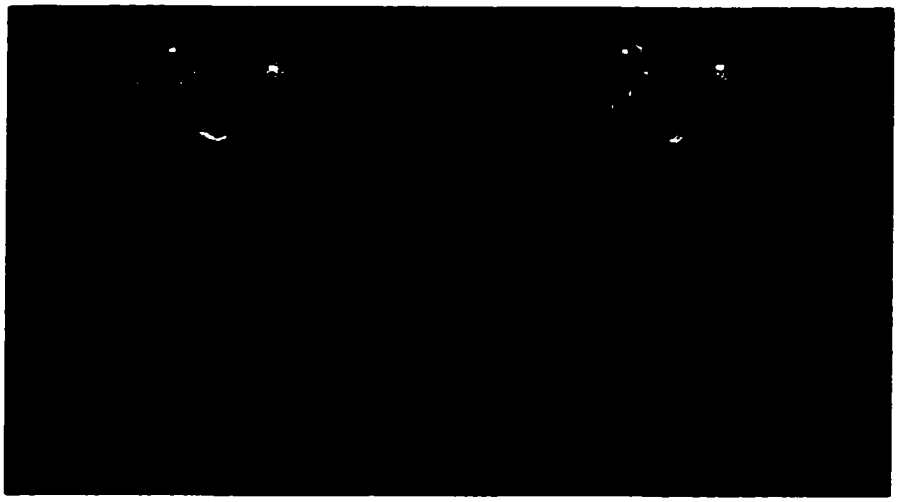
by DeMaeyer et al. (1997), we also list "common-atom" rotamer positions with common χ angles for cases that have similar data and equivalent subsets of geometry and contacts. This streamlines some applications, and it avoids the danger of choosing between rotamers based on a difference that is not statistically significant.

Differing treatment, as well as size, of the database used is an important methodological issue. In compiling the new library, the number of side chains analyzed is reduced by eliminating those with uncertain conformations. In general, when a side chain has been shown to be either wrong or uncertain we simply omit it from the compiled data, since any correction process not using the experimental data would be highly suspect. The only exceptions are the 180° flips of side-chain amides or imidazoles which we do correct in unambiguous cases, and the orientation of movable hydrogen positions, neither of which affects agreement with the X-ray data significantly. A larger database is clearly desirable when trying to distinguish signal (correct rotamers) from random statistical noise, since the signal-to-noise ratio often increases as the square root of the number of observations. However, that relationship holds only if the data is of uniform quality and if the errors are random, neither of which is the case for side-chain conformations. In fact, since low-resolution, high *B*-factor data is most susceptible to systematic errors, adding such observations will degrade rather than improve the results. In effect, we filter out the noise rather than attempting to amplify the signal.

The great value of requiring low *B*-factors as well as high resolution is demonstrated by the sharp and distinct distributions now obtained even for difficult cases

such as Met χ^3 (Figure 7-4(a)), Asn (Lovell *et al.* 1999), and Lys (Figure 7-13). The analysis of Met χ angles shows that although the existence of the major rotamers is determined mainly by atomic clashes (e.g. the three staggered values for χ^1 and χ^2, and the absence of $\chi^1 = +60°$ on helices), the exact position and relative populations of those rotamers are often determined by favorable atomic contacts, such as those that make *gauche* rather than *trans* χ^3 preferred for Met.

FIGURE 7-13. Stereo pair showing all examples superimposed, for three neighboring Lys rotamers: **mtpt** (blue), **ttmt** (yellow), and **ttpt** (green). Balls indicate the mean N^{ζ} position for each rotamer. **mtpt** and **ttmt** diverge for C^{γ} and C^{δ}, but their distributions for C^{ε} and N^{ζ} rejoin and coincide, resulting in almost identical mean N^{ζ} positions. **ttpt** is one of the closest possible nearby rotamers, but its terminal distribution is completely distinct. Individual side chain examples are superimposed onto ideal-geometry N, C, C^{α}, C^{β} atoms using PROFIT and displayed in MAGE.



Significance of Tight Rotamer Clusters

Overall, these results show even more strongly than before that protein side-chain conformations do indeed occur as well-defined rotamers. To illustrate the overall level of rotamer clustering in Cartesian space for real side chains, Figure 7-13

shows the superposition of all examples in our database of three neighboring Lys rotamers: **mtpt**, **ttmt** and **ttpt**. Even at the terminal atom the clusters are tight, despite the distribution at each angle having a significant spread. The distributions of NH_3^+ positions for **mtpt** (blue) and **ttmt** (yellow) are completely overlapping with means only 0.27 Å apart, whereas the NH_3^+ distribution for the near-neighbor rotamer **ttpt** (green) is well separated from the others in its own distinct location 2.1 Å away. The standard deviation of Lys NH_3^+ atom positions in a given rotamer is about 0.8 Å which certainly seems narrow enough to confirm the practical utility of rotamers: even with four χ angles, the rotamer clusters are crisply distinct.

Dissemination of these Methods

See Appendix 1 for availability of the software and files described in this chapter. As described above, our contact-dot tools are already being used in model refitting by a number of laboratories across the country. An article on the crystallographic uses of MAGE, PROBE, and related programs was invited for the new Volume F of the International Tables for Crystallography, now in press (Richardson and Richardson in press). After a demonstration David Richardson and I gave at the 1999 Mid-Atlantic Crystallography meeting, Gary Gilliland invited the Richardsons to present a workshop on all-atom contact analysis at the Macromolecular Crystallography course at Cold Spring Harbor, October 13-26, 1999. It was well received and will be repeated next year. Knowledge of these new methods will spread from our Web site (see Appendix 1) and links at the O and XTALVIEW sites, and as the uses by collaborators reach publication.

The new rotamer library will probably spread very quickly, since the methodology is already widely used and changing libraries is extremely simple. Several protein design labs have already requested the library in advance of publication (Mayo, Handel, Hellinga). Most crystallographers who are shown the new library simply replace their old rotamer files with the new ones. Morten Kjeldgaard, one of the O authors, has done a comparative fitting test with the new library and intends to make it the new standard distributed with O.

The new rotamer results show that even for the low-*B*, interior side chains, where both favorable and unfavorable interactions are most likely to distort preferred conformations, only very rarely are examples seen > 30° away from rotameric values. It follows that in high-*B*, exposed positions there is no justification for fitting side chains in non-rotameric conformations.

CHAPTER 8

Evaluation of Contacts for Conformational Alternatives

Modeling Substitution Mutations

Substitutions—replacing one amino acid side chain with another—are an important class of protein mutations. In nature, a substitution can arise due to an error in replication, transcription, or translation or even through post-translational reactions. In molecular biology, numerous techniques are routinely employed to deliberately alter the amino acid type at a specific sequence location. A single substitution may give rise to either subtle or profound changes in molecular properties (e.g., Glykos, Cesareni, and Kokkinidis 1999).

When planning a substitution mutation or when analyzing experimental results obtained from such a mutant, several questions related to the tertiary molecular structure arise: Is sufficient space available for the new side chain? Would other side chains or the backbone have to move to accommodate the new side chain? If it is compatible, is more than one conformation available to the new side chain? Is

there unoccupied space adjacent to the side chain? Are there hydrogen bonding or electrostatic partners nearby in a position where they can interact?

Even in the absence of high-resolution three-dimensional structures for both the background and mutant proteins, preventing direct examination of these questions, many conclusions can be drawn by extrapolating from a single accurate structure. The techniques described in this chapter, interactive MAGE/PROBE and autobondrot, can take a static set of atomic coordinates (with all H atoms added) and permit some parts to move by rotations around one or more axes. Typically these are side-chain atoms and the axes define the side-chain torsion angles: χ^1, χ^2, etc. where χ^1 is defined as the dihedral angle around the central bond for the N–C^{α}–C^{β}–C^{γ} atoms, and the other χ angles are measured progressively further out the side chain. At each conformation sampled, PROBE calculates the small-probe contact surface for interactions between the side-chain atoms and other nearby atoms. This contact surface can be represented visually as a set of contact dots and clash spikes on the van der Waals surface of atoms at points less than 0.5 Å from the surface of a non-bonded nearby atom. Alternatively, the contact surface can be summarized by a numerical contact score with terms for H-bonds, unfavorable overlaps, and favorable van der Waals contacts. Contact dots show detailed surface complementarity; the contact score is not an energy and as currently formulated the better the packing, the more positive the score. For a detailed description of small-probe contact dot surfaces, contact scores, and the PROBE program, see Chapter 2.

Simple mutation analysis with PROBE contact dots is done either interactively in the MAGE display (Richardson and Richardson 1992; 1999) or by generating plots of contact score *versus* conformation. From such exploration of how the contact surface varies with changes in side-chain conformation, one can readily determine whether or not nearby groups, in their given positions, can accommodate the substituted side chain and over what range of angles. This simple process provides most of the payoff obtainable by a predictive analysis. However, if desired, an additional step can be done to test the benefit of moving another side chain. Movement of main chain, however, is presumed to have unpredictable and perhaps widespread consequences and is not attempted here.

For this sort of analysis to be reliable, the reference structure needs to be accurate enough that the positions of groups near the site of mutation can be trusted to better than 1/2 Å; in our experience, crystal structures better than 2.0 Å resolution are desirable. NMR structures with more than about 20 restraints per residue in the region of interest should support this method, but we have not tested such use. Even given a high-resolution structure, reliable analysis cannot be performed in regions of disorder as indicated by high temperature factors (*B*-factors) or multiple conformations.

Another requirement is that hydrogen atoms must be explicitly included in the molecular models, because van der Waals surface complementarity is very sensitive to the local geometry. Although commonly employed, implicit representations of hydrogen atoms (in which the van der Waals radius of the non-hydrogen atom is

increased) ignore the individual hydrogens' significant influence on local packing (e.g., leucine conformation in Petrella, Lazaridis, and Karplus 1998). See Chapter 3 for a description of the program REDUCE which adds hydrogens to structures and will optimize the positions of movable ones.

Many other approaches to modeling sidechain substitutions have been described. One popular family of techniques involves building a model of the mutant and performing energy minimization or molecular dynamics runs. Some or all of the side chains and prosthetic groups are free to move. Depending on the force field and strategy, the backbone is either fixed or free to move. Comparison of both energy and structure with the non-mutated background is used to distinguish mutational alternatives. This family of techniques is powerful, but energy-based methods almost always give a plausible result—they rarely say *no*. Ironically, this means that *yes* may not always mean yes, and it is left up to the user to decide how much change is unacceptable. Explicit estimation of the change in stability ($\Delta\Delta G_u$) for a substitution involves an *alchemical* thermodynamic cycle (Zeng *et al.* 1999) in which atoms and even charges are added or removed; this can be very effective for relatively small conformational changes but is not easy for the non-expert to perform and is computationally demanding. The method described here provides an easy way of answering a more limited question: is or is not the proposed mutation compatible with an essentially unchanged surrounding structure? This question is probably the most important one that can be asked in advance, because if the surroundings must move, then even if the mutant is stable (which it may well be) there is no straightforward way to analyze the meaning of whatever changed properties

are observed. Similarly, if a mutation is found to produce unexpected consequences, this simple analysis can show in hindsight whether conformational changes are likely to be the confounding factor.

MAGE/PROBE and Autobondrot

The computer program MAGE displays interactive three-dimensional graphics from a *kinemage*, a structured text file containing object descriptors, coordinates, identifiers and display parameters (Richardson and Richardson 1992; 1994; 1999). MAGE versions 5.4 and above for Unix, Linux and Microsoft Windows (95 or later) have been modified to interact with separate external programs which generate new geometric objects that are then incorporated into the current graphics display.

This program–to–program communication is used in the work described here in two ways. First, a rotatable side chain (either mutated or not) can be constructed with the assistance of PREKIN. PREKIN is the primary feeder program for MAGE, designed to read Protein Data Bank (PDB) format atomic coordinates and create a kinemage to illustrate them with a choice of styles or subset selections. The version of PREKIN used for this work is 5.4. Although PREKIN can independently construct a kinemage with a mutation or a rotatable side chain, the MAGE/PREKIN link can conveniently be used to modify a structure while it is being viewed and analyzed in MAGE. Clicking on a residue, and then choosing the "Remote Update..." tool sets up a command instructing PREKIN to either make the selected residue rotatable or to replace it with a rotatable version of a different amino acid. PREKIN uses Engh and

Huber (1991) geometry to build mutant side chains. Because the bond angles for the C^{β} are fairly often distorted, in our predictive models the C^{β} position for the mutant was standardized (setting bond length and bond angles to Engh and Huber values) to better represent an un-strained side chain; models using the actual mutant crystal structures were built with the reported C^{β} coordinates.

The second way MAGE is used to interact with external programs is by calling PROBE, to generate the same kind of small-probe contact dots described in Chapter 2. Once the link is set up, the generated contact information is coupled with the rotation of torsion angles in MAGE: each time an angle is adjusted, MAGE calls PROBE, passes PROBE the modified coordinates, reads the results and displays the updated contact dots and clash spikes. Despite the fact that PROBE is being restarted each time the conformational angles are adjusted, the process is very quick, allowing interactive exploration. Usually, PROBE is instructed to combine the rotated coordinates from MAGE with the static atomic coordinates from the original PDB file. To do this, MAGE generates a command line which uses PROBE's flexible method of selecting sets of atoms to make the rotated coordinates supersede the originals. XTALVIEW has recently been modified to call PROBE using the same mechanism (McRee 1999) providing an interactive display of steric constraints during crystallographic map fitting. Also, this interaction with PROBE can be simulated one step at a time by macros in the 'O' fitting program (Jones *et al.* 1991). (The use of PROBE with XTALVIEW and O are discussed in chapter 7.)

In both interactive cases, MAGE communicates with external programs using pipes, a computing facility widely supported under Unix, Linux and other operating systems. The external programs do not have to be modified to work with this technique as long as they can read ATOM records from the "standard input" stream and write geometry to "standard output" in kinemage format.

The MAGE/PREKIN and MAGE/PROBE links operate locally. Although pipes are a simple, robust technique for communication between programs running on the same machine, the way we are using them is not well suited for communication between programs running on separate machines since the overhead of repeatedly re-starting a program remotely can be much greater than doing the same locally. In order for MAGE to efficiently call programs remotely, we would need to modify the remote program to act more like a server by maintaining a persistent connection. Currently, on the Macintosh a relatively high process creation overhead also seems to recommend a persistent connection. Versions of PROBE which persistently communicate with MAGE may be developed in the future but it is clear that one of the advantages of our current implementation is flexibility in the choice of external programs.

The MAGE/PROBE link described above permits interactive exploration of conformational space. If this space is large because several torsion angles must be explored together, sampling by hand is inadequate and a more automated method of examining conformations is needed. PROBE version 2.0 has a new feature called 'autobondrot' to iterate over a range of dihedral angles, determining the contact

surface at each step and calculating a numerical score. This score, along with each of the angular coordinates, is written out in tabular format. The autobondrot procedure is controlled by a rotation script (a .rotscr file) which specifies the dihedral axes, angle ranges and sampling frequency, and the static and rotated atoms. A conversion tool (mkrotscr) will translate any rotations in a kinemage file (e.g., side-chain rotations setup by PREKIN) into the dihedral scan pattern in a .rotscr file, which the user can customize further if needed. This same system was used to generate the main-chain and side-chain conformation maps in Chapter 7.

FIGURE 8-1. Example of a .rot script executable command file

```
probe -q -stdbonds -3 -drop -once "file1" \
  "file1 | file2 alta not water not(sc 61)" -auto - 1ah7H <<END_OF_INPUT
atom      1  cb  tyr    61      34.219  17.937   4.659  1.00  0.00
bondrot:chi1:78.7:  0:359:5:33.138:18.517: 5.531:34.219:17.937: 4.659
cos:-3:60:3:
atom      1 1hb  tyr    61      34.766  18.777   4.206  1.00  0.00
atom      1 2hb  tyr    61      34.927  17.409   5.315  1.00  0.00
atom      1  cg  tyr    61      33.836  16.989   3.546  1.00  0.00
bondrot:chi2:-11.8:  0:179:10:34.219:17.937: 4.659:33.836:16.989: 3.546
atom      1  cd1 tyr    61      32.578  16.433   3.418  1.00  0.00
atom      1  cd2 tyr    61      34.803  16.657   2.603  1.00  0.00
atom      1  ce1 tyr    61      32.294  15.554   2.393  1.00  0.00
atom      1  ce2 tyr    61      34.520  15.798   1.551  1.00  0.00
atom      1  cz  tyr    61      33.249  15.259   1.456  1.00  0.00
atom      1  hd1 tyr    61      31.793  16.694   4.142  1.00  0.00
atom      1  hd2 tyr    61      35.813  17.084   2.693  1.00  0.00
atom      1  he1 tyr    61      31.299  15.089   2.328  1.00  0.00
atom      1  he2 tyr    61      35.291  15.550   0.807  1.00  0.00
atom      1  oh  tyr    61      32.991  14.372   0.421  1.00  0.00
atom      1  hh  tyr    61      33.803  14.287  -0.156  1.00  0.00
END_OF_INPUT
```

Relying on exhaustive enumeration at a sampling rate of 10 to 20 conformations per second on a 250 MHz R10000 Silicon Graphics workstation, autobondrot is most appropriate for scans involving 1 to 4 torsion angles. For most uses, side

chains with up to 3 torsion angles are sufficiently well sampled every 5° in χ^1, 5° or 10° in χ^2, and 10° or 15° in χ^3. The longer side chains of Lys and Arg with 4 torsion angles can be rapidly surveyed by centering χ^1 at a staggered angle (e.g. –60°) and sampling each χ every 120°, 5°, 10° and 15°, respectively. Each of the three blocks with the same χ^1 can then be plotted separately in 3-D. Sampling only "rotameric" conformations where side chains have previously been observed in a number of proteins can speed the scanning somewhat, but the benefits of a comprehensive survey map usually outweigh the marginal increase in speed. The autobondrot scores for an isolated side chain recapitulate the boundaries of allowed regions within which the rotamers occur, so that the information is included implicitly.

Scoring

PROBE scores are determined by dividing the contact dots into categories, scaling each appropriately and summing (see Chapter 2 for a description of how scores are computed). In the procedure described here, scores were calculated only for movable atoms in the rotated side chain. An important step for many types of conformational scans is the addition of a penalty function representing a torsional barrier to bond rotation, as is done in most force fields. This is necessary even with explicit hydrogen van der Waals contacts included, because the primary torsional effect is an intrinsic property of the bond, related to the hybridization of the bonded atoms (Momany *et al.* 1975; Streitwieser and Heathcock 1976). The torsional penalty varies with the cosine of the dihedral angle (e.g. from 0 at staggered angles to s at eclipsed angles) and is added to the PROBE score as follows

$$\text{TotalScore}(\chi^*) = \text{ProbeScore}(\chi^*) + \frac{1}{2}\sum_i s_i(1 - \cos[n_i(\chi^i - \delta_i)]) \qquad \text{8-1}$$

where χ^* refers to the conformation of the group as a whole, s_i is the scale factor between PROBE score and torsional strain for the ith χ dihedral angle, taken to be −3.0 for most Csp^3–Csp^3 bond dihedrals, 0.0 for dihedrals adjacent to flat groups (such as χ^2 of Phe or Asp), and 0.0 for dihedrals that rotate an OH or SH group, n_i is the number of barriers in a full rotation and δ_i is the phase angle where the penalty vanishes. For χ^1 of all residues except Gly, Ala or Pro, $n = 3$ and $\delta = 60°$.

The torsional scale factor of −3.0 for Csp^3–Csp^3 bonds was selected to approximately match PROBE scores to observed side chain distributions from a database of 240 high-resolution structures, as was the conformation acceptance criterion: *TotalScore* > −1.0 (Word and Lovell, unpublished results, see Chapter 7). Structures with scores above the acceptance cutoff, even if they clash slightly, are assumed capable of adjusting nearby side chains, shifting main chain, or deviating modestly from standard geometry in response to any residual strain without greatly destabilizing the protein. Precise torsional scale factors for S-C, S-S, and C-N bonds and several other special cases have not been determined, they should almost certainly be smaller than −3.0; by default, we set them to 0.0.

The output PROBE scores for each combination of dihedral angles is most readily analyzed graphically. Software developed in-house (KIN2DCONT and KIN3DCONT) is used to generate 2D or 3D contour maps in kinemage format from tabular data.

The contour maps summarize the trends in the data, while individual values can be identified by clicking on the contours in MAGE.

Although the autobondrot procedure is described here in terms of sampling torsional angles, it can also rotate around or translate along arbitrary directions in space. This could be used to perform simple docking calculations (a capability we have not yet exploited).

See Appendix 1 for information on obtaining the programs used in this work (MAGE and PREKIN by DCR and PROBE, REDUCE, KIN2DCONT, KIN3DCONT and mkrotscr by JMW).

Applications in Other Laboratories

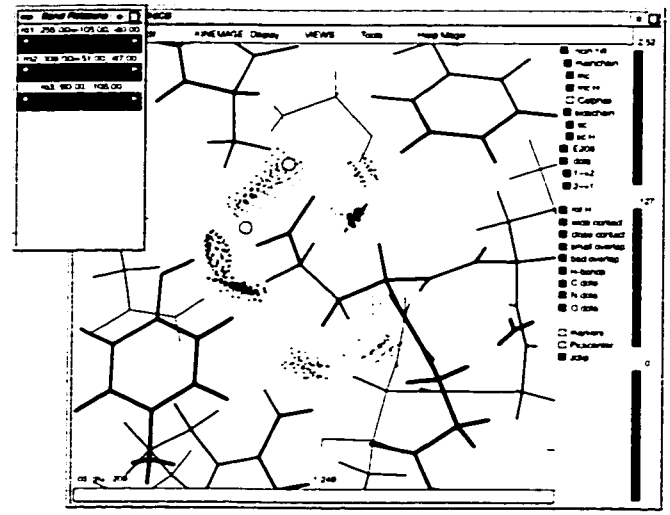
In the Oas laboratory, Greg Kapp used an early version of autobondrot to analyze the potential range of motion for a mutated tyrosine in the presence of neighboring groups in an attempt to explain the lack of expected NOE peaks in certain λ repressor NMR spectra.

Gene Wickham, in the Been laboratory, made full use of the MAGE's construction tools, the MAGE/PROBE interface, and the conformational search capabilities provided by autobondrot, to model a phosphate attached to O5' of an HDV ribozyme cleavage product. This work is described in more detail in "Nucleic Acid Structures" on page 70.

Ricin

Castor bean seeds contain the toxic protein ricin whose A chain (RTA) is an enzyme which inactivates eukaryotic ribosomes by removing a specific conserved adenine base from 28S rRNA (Endo *et al.* 1987). Yeast genetic studies (Frankel *et al.* 1989) and analogy with the homologous Shiga-like toxin I (Hovde *et al.* 1988) suggested that glutamic acid 177 may be required for catalysis. In an attempt to determine the precise role of Glu177, Schlossman and co-workers (1989) cloned and tested several active site mutants. The mutation E177D lowered activity by a factor of 80 but, disturbingly, E177A lowered activity only by a factor of 20. At the time, Jane Richardson observed that in the wild type ricin crystal structure Glu208 appeared to be close enough to stand in and rescue activity (Frankel *et al.* 1990). The double mutant testing this hypothesis, E177A-E208D, was found to be completely inactive. A crystal structure of the E177A RTA mutant (Kim *et al.* 1992) confirmed that Glu208 does indeed adopt the proposed alternative conformation.

FIGURE 8-2. MAGE/PROBE display of ricin E177A active site mutant showing the alternative "rescue" conformation of Glu208, with interactive contact dots. The structure is from the wild-type ricin A chain structure 1IFT (Weston *et al.* 1994). REDUCE was used to add hydrogens and a kinemage was created with PREKIN with an Ala substituted for Glu at position 177 and a rotatable Glu208 side chain. Adjusting the sliders in the upper left results in rotation of side-chain torsion angles χ^1, χ^2 and χ^3 of Glu208 and the recalculation of the contact dots. In the interactive display, contact surfaces are color coded to indicate van der Waals surface complementarity. Gray balls mark the locations of Glu177's oxygens in the crystal structure.



The modeling used to predict the role of Glu208 consisted of noting that, in the 2.3 Å ricin structure 1RTC (Mlsna *et al.* 1993), the side chain could be oriented to reach the active site and speculating that it must work, based on the odd experimental result. Without that result, the best one can say when modeling various conformational alternatives is whether the model is consistent with accepted physical principles. Chief among these is that there must be enough room to fit a side chain into a given position without deforming chemical bonds or putting atoms on top of one another. Such steric constraints are particularly limiting when all the hydrogen atoms are explicitly represented in the model. Figure 8-2 is a kinemage of an E177A model (based on the more recent 1.8 Å RTA structure 1IFT, Weston *et al.*

1994) showing dots for the atomic contacts to Glu208. In this picture, the Glu is positioned to hydrogen bond with Tyr123's phenolic oxygen (as Glu177 did in the wildtype), as represented by the lens-shaped set of dots to the left. Contact dots are on the van der Waals surface of an atom; clicking on a dot displays the atom name. The kinemage is interactive; as the χ angles are varied, the contact dots are automatically updated. Exploring different conformations, the great majority of them are found to clash with the atoms of neighboring groups. The conformation shown in Figure 8-2, however, places the carboxylic acid in approximately the same spot as Glu177 previously occupied, without introducing steric conflicts.

Simple interactive exploration suggested that there are at least two conformations which accomplish the goal of avoiding clashes and putting the carboxylate in the right spot. Are there any others? And how large are the acceptable regions? An automated conformational survey was performed over all three side-chain torsion angles for Glu208 using the coordinates from 1IFT. The autobondrot procedure was used to sample each angle at 5° increments (10° for χ^3) and calculate a PROBE score for the side chain assuming fixed conformations for the main chain and all the other side-chain atoms (waters were not considered). The PROBE score was combined with a torsional factor for χ^1 and χ^2, but not χ^3, as described above. The scan required 81 minutes on our Silicon Graphics Indigo2 workstation.

FIGURE 8-3. Face (a) and side (b) views of a three-dimensional contour map summarizing an **autobondrot** scan of Glu208 side-chain conformations in a model of the ricin E177A mutant. The model is constructed from the high resolution wild-type structure 1IFT, in which Glu177 was trimmed back to Ala, waters were omitted, and hydrogens were added with **REDUCE**. The gray contour mesh encloses favorable conformations where the contact score is greater than −1. The heavy black lines enclose the two regions (near the top and bottom) where the carboxylate oxygens for Glu208 have a summed distance of less than 2.5 Å from those in the original Glu177. The full range is shown for each torsion angle, so opposite edges wrap and a line which extends beyond the right edge continues on the left. Solid points mark: (WT) the conformation Glu208 had in the wild type structure, (Pred) Jane Richardson's original modeling prediction, and (Obs) the conformation observed in the crystal structure of the E177A mutant (Kim *et al.* 1992).

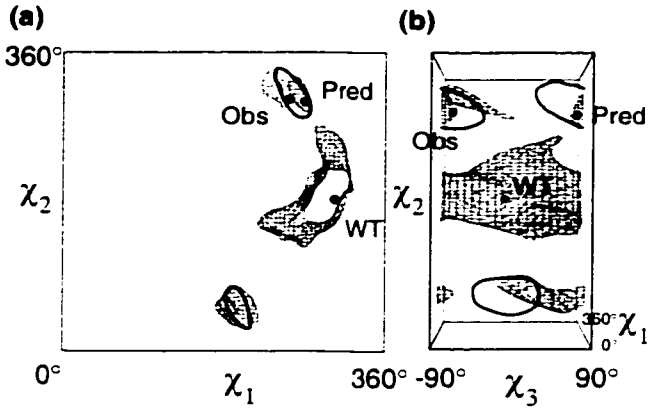


Figure 8-3 shows the results of this scan, a map of the three-dimensional space which describes Glu208 side-chain conformational freedom. The grey mesh encloses conformations in which Glu208 has acceptable contact scores (> −1). Conformations outside of this mesh are clashing with neighboring atoms. Also, there are indeed only two conformational regions, outlined in black, which put the carboxylate oxygens in approximately the same position as those in Glu177 (the two are centered at $\chi^1 = -105°$, $\chi^2 = -51°$, $\chi^3 = 90°$, and $\chi^1 = -167°$, $\chi^2 = 53°$, $\chi^3 = -5°$). Each of these regions encloses a narrow range of conformations which also have acceptable contacts—candidates for how Glu208 might be rescuing enzymatic

activity in the mutant. The PROBE scores do not distinguish between these two alternatives and each is handicapped by a nearly eclipsed side-chain dihedral. In the end, the first region (at top in Figure 8-3) is favored by a slight shift of Arg180 to ion pair with Glu208; this conformation was the one observed in the E177A mutant crystal structure (Kim *et al.* 1992).

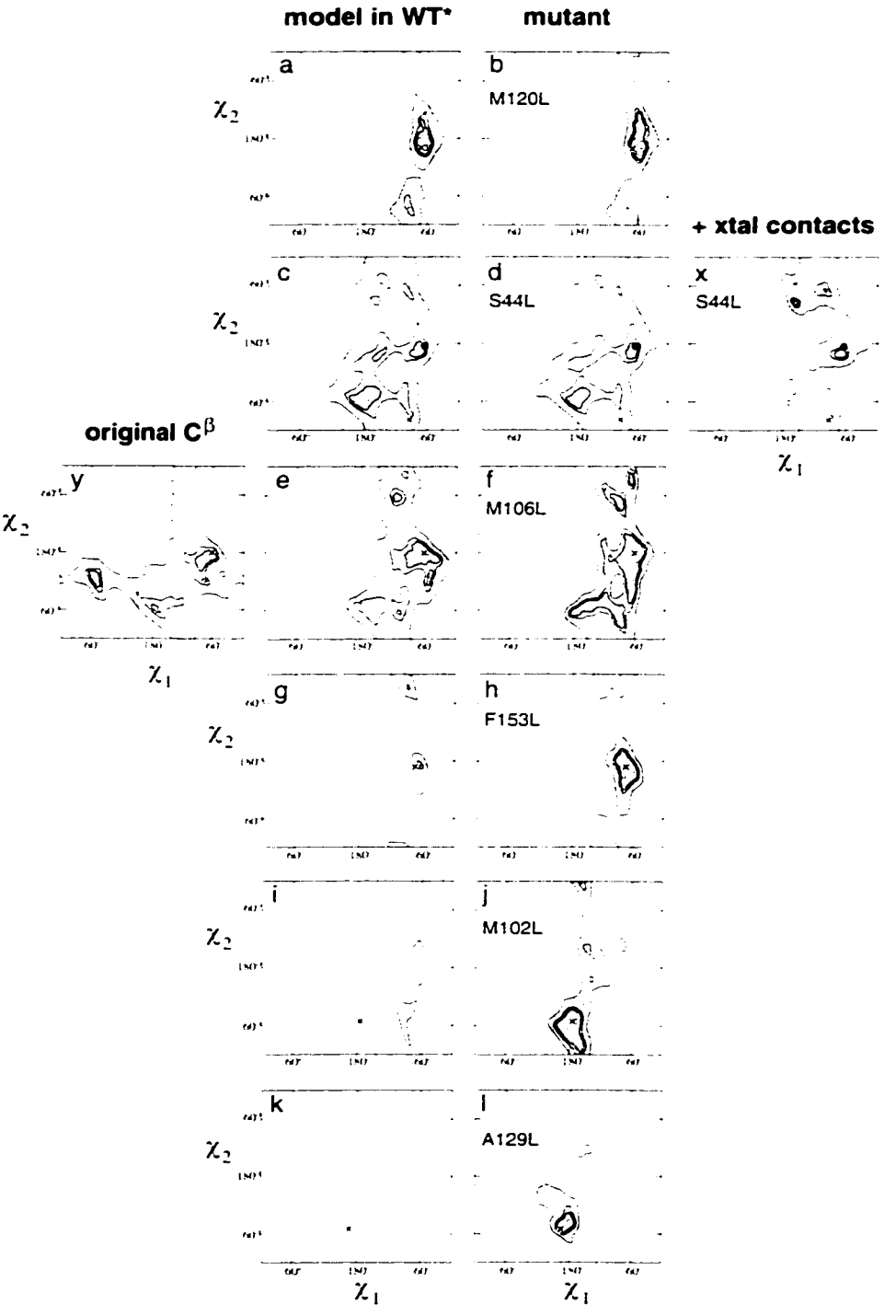
Lysozyme

Bacteriophage T4 lysozyme is in many ways a molecular 'lab rat'—a system which is well behaved and well characterized and which can be readily manipulated for study. Over the past three decades, Matthews and co-workers have extensively studied T4 lysozyme, developing a large and unique database of mutants for which there are activity and stability measurements and in most cases high-resolution X-ray crystal structures (Matthews 1995). This database has been gleaned, by that laboratory and by others, to yield many significant insights about the sensitivity of protein stability and structures to mutations and the determinants of those effects (e.g., Baldwin *et al.* 1993, 1996; Blaber *et al.* 1993; Faber and Matthews 1990; Gassner, Baase, and Matthews 1996; Karpusas *et al.* 1989; Vetter *et al.* 1996). In the process, the early simple description of "temperature sensitive" mutants has evolved into a sophisticated understanding of the thermodynamics of protein stability.

Here we make use of this resource, comparing models of several mutants based on the "pseudo-wildtype" structure WT* (C54T/C95A with PDB code 1L63; Matsumura and Matthews 1989) with the observed structure for each mutant protein.

We have chosen six mutants, of resolution to 1.7-2.05 Å, which substitute a single leucine for either alanine, serine, phenylalanine or methionine. Analysis of just $x \rightarrow$ L mutations facilitates case-to-case comparisons, and the set covers a range of local environments as well as a range of responses to the mutation. In each case, structural changes (or lack thereof) in response to the mutation were described in detail in the primary reference for each structure, which we will quote from; the intention here is to focus on the limited but very useful inferences that can be made when using conformational modeling of steric interactions with fixed backbones and neighboring side-chains. It should be noted that each of the selected mutations was successful in yielding a stable, well-ordered protein.

FIGURE 8-4. Autobondrot side-chain conformational maps for six T4 lysozyme mutants, calculated in the context of the static protein structure. For each mutant, total score (PROBE score + torsional potential) contours are shown for a leucine with idealized geometry, both in the "pseudo-wildtype" structure (WT*) and in the observed crystal structure of the mutant. Shaded regions have scores > -1, indicating acceptable side-chain packing. Solid black contours are at scores of -10, -1 and $+1$. Dashed lines are contoured by 10s from -60 to -20 and solid gray contours are by 2s between -10 and -1 and every 1 above that. Conformations observed in the mutant crystal structure are marked with an X. In (x), (c) and (d) the circled-plus sign marks the conformation of the highest-populated leucine rotamer ($\chi^1 = -65°$ and $\chi^2 = 175°$) which occupies a position in space close to the conformation reported for the mutant crystal structure. S44 is on the exterior of the protein and (x) shows the map for the S44L mutant when crystal contacts are considered; the map changes but the conclusion remains. The background structure used in the models (a, c, e, g, i, k and y) has PDB identifier 1L63. The mutant structures used were: (b) 233L, (d and x) 110L, (f) 234L, (h) 1L87, (j) 1L77, and (l) 195L. Hydrogens were added to each structure with REDUCE. The position of the C^β was idealized to standard bond lengths and angles when modeling each of the substitutions in 1L63 (WT*) except in (y) which is the only site where there was a significant difference.


model in WT*
mutant
a
b
M120L
+ xtal contacts
c
d
S44L
x
S44L
original Cβ
y
e
f
M106L
g
h
F153L
i
j
M102L
k
l
A129L
χ2
χ1

The six sites vary in the extent to which the wild-type structure can accommodate the substitution, as revealed in contour maps of the total score (PROBE score + torsional penalty; see "Scoring" on page 201) for all combinations of side-chain dihedral angles, χ^1 and χ^2 (Figure 8-4). The isosteric substitution M120L (Lipscomb *et al.* 1998) is at a site with "partial solvent exposure" near the center of a short helix. The maximum total score for the model (Figure 8-4(a)) is +8 in a conformation with considerable positive van der Waals contacts. In the mutant structure (Figure 8-4(b)), the conformations of the main chain or the neighboring side chains do not substantially change from that in WT*, so the maps are very similar. The regions predicted from the model to accept the substitution (score > −1; shaded areas in Figure 8-4) do indeed contain the observed conformation.

Site S44L (1994; Blaber, Zhang, and Matthews 1993) is in many ways even simpler (Figure 8-4(c,d)). Being "fully solvent exposed," the acceptance regions for both the model and the observed mutant outline the major leucine rotamers and primarily reflect contacts with local main-chain atoms. A complication, however, is that the leucine conformation in the mutant structure (marked with an X) lies outside the acceptance regions, implying considerable strain (score −3). There seems no reason for Leu44 to avoid the more favorable available conformation, since it does not appear to be constrained by its surroundings. When crystal contacts are considered, the score for the reported conformation is not improved, although χ^1 *trans* is revealed to be unfavorable (Figure 8-4(x)). We note that this side chain has been built into a conformation (which we will call **mp*** using nomenclature described in Chapter 7 and in Lovell *et al.* 1999) which is pseudo-symmetric with the conforma-

tion of the un-strained principal leucine rotamer **mt** (marked in Figure 8-4(c,d) with a circle-plus) (Lee and Subbiah 1991; Lovell *et al.* 2000; Petrella, Lazaridis, and Karplus 1998). Both conformations position the terminal methyls in approximately the same location and can often appear to fit less-than-optimal electron density equally well. However, C^{γ} is offset in opposite directions in the two conformations—the difference is about an Ångstrom—and in this case $C^{\delta 1}$ differs by more than $C^{\delta 2}$. A comparison of the crystallographic *B*-factors of the side-chain carbons can be diagnostic; they describe not only atomic motion but also the local quality of the electron density and how well the atom's position matches that density. For Leu44 (*B*-factors: C^{α} 24 Å^2, C^{β} 24 Å^2, C^{γ} 35 Å^2, $C^{\delta 1}$ 35 Å^2, $C^{\delta 2}$ 25 Å^2), the higher value for C^{γ} than for $C^{\delta 2}$ suggests that C^{γ} may not be centered in the electron density and that the rotameric conformation **mt** is probably a better choice for this side chain (score +4). We have previously suggested (Lovell *et al.* 2000) that the leucine **mp*** conformation and an analogous **tt*** conformation are mistakes resulting from ambiguous electron density, the confusing similarity to major rotamers and, more recently, from the recycling of these mistaken conformations from earlier structures into entries in standard rotamer tables. If we accept the suggestion that the actual conformation is **mt**, we see from the similar contour maps that the unstrained local environment for the model of S44L is a good representation of the actual mutant.

Well packed, high-scoring conformations in the model do not ensure that the protein will not change upon mutation. In M106L (Lipscomb *et al.* 1998), a mutation of "the most solvent exposed methionine of T4 lysozyme," the model has a maximum score of +4.6 but the mutant shows a dramatically different conformational

map (Figure 8-4(e,f)). The primary cause is a change in main-chain angles which tilts the C^{α}–C^{β} bond vector by about 27° and displaces C^{β} by almost 1 Å (Figure 8-4(y)). According to Lipscomb *et al.* (1998), "the introduction of the leucine side chain does not result in steric interference with neighboring protein atoms or in the formation of cavities." If the mutant is not forcing the backbone movement, the change may come from the release of strain in WT*. Alternatively, the WT* may be misleading in this region: Met106 in the background structure has significantly non-ideal bond angles and is more disordered than Leu106 in the mutant, and perhaps it adopts the mutant backbone structure (or even the mutant side-chain χ^1) part of the time. In any case, the model correctly predicts that the mutation will work.

In mutant F153L (Figure 8-4(g,h); Eriksson, Baase, and Matthews 1993), the model's acceptance region is very small, with a maximum score of +1. This represents a borderline case, where there are "moderately large shifts of several side-chains toward the mutation site" and shifts in the main chain in part due to slight twisting of the alpha helix containing the mutation by −2° as well as a slight shift of an adjacent helix. The model's score map (Figure 8-4(g)) should be read as indicating that some modest reorganization (which, in any given case, may or may not be available to the protein) would probably be required to accommodate the mutation comfortably.

Finally, mutants M102L (Hurley, Baase, and Matthews 1992) and A129L (Baldwin *et al.* 1996) require significant structural reorganization (Figure 8-4(i,j,k,l)). The models have maximum scores of −2 and −37, respectively. Unlike the other cases

considered above, the change in free energy of unfolding found for the mutant protein relative to WT* is negative (–0.7 kcal/mol for M102L and –1.3 kcal/mol for A129L). That is, these two mutations are destabilizing, and the induced strain from A129L is calculated to be –2.6 kcal/mol (Baldwin *et al.* 1996). The C^{β}s shift by 0.5-0.7 Å and several nearby side-chain atoms move by 1 Å or more; in M102L the C^{ε} of nearby Met106 moves 2.8 Å. For an alanine to leucine mutation in a protein interior this is hardly surprising, but even the isosteric methionine to leucine mutation changes the side-chain shape and reduces the degrees of freedom available to the side chain. In each of these two cases, the model score map clearly signals that the mutation may not be stable but that even if it is, the background structure is not a good stand-in for the mutant structure.

For each mutation, an essential part of the interpretation of the contour map involved going back to the interactive kinemage and studying the contact dots and clashes for conformations in each of the different regions of the map. Even when the map shows good scores, looking at the structure can reveal complicating factors (e.g. buried charges or cavity formation) which are not considered in this method but may destabilize the mutant protein or force the mutant to adopt a different structure. On the other hand, neighboring groups which constrain the mutant side chain may be relatively unconstrained themselves. Autobondrot provides an overview of conformation space, revealing the trends, while critical inspection of the atomic structure and its contacts provides a complementary detailed view of the elements—their identity and the weight they should be given—of which these trends are composed.

Discussion

The methodology described here has a number of limitations. Electrostatics are only used to recognize hydrogen bonds; charge-charge attraction or repulsion is ignored. Our current plans are to extend PROBE to consider electrostatic effects out as far as the probe diameter. Until this is available, the possible effects of any strongly polar groups on the results should be kept in mind. Another limitation is the reliance on fixed bond angles, fixed nearby side chains and fixed main chain position (any alternatives must be set up as separate runs). While it is possible to model some changes to nearby side chains, Matthews and co-workers have convincingly shown that in many cases the main chain changes conformation, which can not be reliably modeled.

In (Baldwin *et al.* 1993) they argue that "protein backbones are more flexible than generally assumed" and that the modeling of side chain steric conformations using fixed main chain and neighboring groups is "overly restrictive." The remarkable tolerance for small → large mutations at some sites within T4 lysozyme and other proteins makes this a very reasonable position on modeling done with the purpose of reliably predicting a protein's response to mutation or perhaps in homology modeling. Given the current lack of know-how in main chain conformational searching, our technique pursues a more modest goal: the conformational modeling described here attempts to predict *whether* the structure *must* change and not to say *how* it changes. In searching molecular conformations, these procedures act as conservative filters and (as seen in the A129L mutation) may very well 'reject' a mutation which later proves to be stable.

Finally, small-probe contact dots are not good at cavity identification and do not calculate cavity size. When replacing a large side chain with a small one, any potential decrease in stability due to cavity creation must be calculated independently and considered along with the contact surface information.

More sophisticated computational methods such as molecular dynamics or interactive minimization, which may have fewer of these limitations, are not without their own; as mentioned above they can be overly compliant. Also, they can have a significant learning curve and often require the expense of a commercial license.

The procedure described here is designed to encourage the bench scientist to consider the molecular geometry when planning or analyzing a mutation. It is a natural extension to simple visual inspection of a molecular model from X-ray crystallography or NMR and promotes exploration. Interactive graphics encourages thinking about spatial arrangements and consideration of the relative impact of nearby groups, complementing the analysis of summary statistics such as contact scores or energies. MAGE, PREKIN and PROBE cooperate to make exploration of side-chain contacts straightforward. Autobondrot surveys conformational space, showing the number and size of favorable regions, usually leading back to the interactive display to model some point on the map.

We agree with Matthews that no modeling method with fixed backbone will be able to predict the conformation or stability of mutants. However, that is not what we are trying to do here. Instead, these MAGE/PROBE/autobondrot methods are intended as screens for whether or not a proposed mutation can be accommodated without

other shifts (so that its results can be interpreted straightforwardly). For the cases examined here, the method indeed works as intended.

CHAPTER 9

Distribution of Packing Density

Packing Trends

A single molecular summary statistic, such as density, energy, or contact score total, while able to provide some insight as to overall packing quality, will ignore variation within a structure. To examine how the level of packing is distributed requires a method which assigns separate values to different parts of a molecule. A value for each residue can be plotted *versus* sequence position to create a molecular profile, often correlated with secondary structure or domain boundaries. For example, Pattabiraman *et al.* (1995) analyze hen egg lysozyme using profiles of per-residue-normalized occluded surface area. The PROVE system uses profiles of volume *Z* score (standard deviations from the group mean) to effectively validate structures (Pontius, Richelle, and Wodak 1996). Interacting clusters emerge when contacts between different residue pairs are presented as a 2D matrix (Ooi and Nishikawa 1973). These contact matrices have been used to identify molecular domains and to define the local elements in a diffusion-collision protein folding model (Myers and

Oas 1999). The semi-2D hierarchy plots of compact clusters by Zehfus and Rose (1986) provide an alternative description of molecular domains. Direct examination of the extent of these structures and the spatial arrangements within them, however, requires assignment of local packing values to some point in each fragment of a 3D structure or to points distributed throughout the space containing the molecule.

Richards (1974) used a 5.6 Å lattice to construct packing density (van der Waals volume divided by Voronoi volume) contours in slices through ribonuclease S and also described packing density deviations in egg lysozyme. He noted that the cores in these proteins were often less dense than the surrounding areas and speculated that proteins might use the dense regions (presumed to be stiffer) to transmit forces across the structure, while under-packed areas may permit "breathing motions."

Statistics for residue contact preferences, compiled from a tessellation of space with Delaunay tetrahedra—the mathematical dual to the Voronoi polyhedra, delineating the space *between* clusters of four neighboring residues as represented by their C^{α}s—has been used to plot the likelihood of each four closest residues within a protein structure (Singh, Tropsha, and Vaisman 1996). Although this useful technique for protein analysis is sometimes described as revealing variations in packing, it does not take the volume of the tetrahedra into account and its results do not correlate with the distribution of mass within proteins.

Pellequer *et al.* (1999) summed our small-probe contact areas for residues in structures of antibody variable fragments, both with and without bound antigen.The area differences between *holo* and *apo*, measuring the effect of binding on compactness,

were tabulated and used to color code each residue. With this representation, they identified residues located far from the binding site which are sensitive to various antigens, and rationalized their effect on antigen recognition.

Contact Contours

We are interested in how packing restricts internal degrees of freedom within proteins, and how such confinement is related to molecular function or stability. To explore objective measures of trends perceived in clumpy aggregates of small-probe contact dots, we calculate and display the amount of self contact area in each cubic Ångstrom throughout any given molecule. Figure 9-1 shows self contact dots for T4 lysozyme (3LZM; Weaver and Matthews 1987) along with contours of the dot density trends. At a packing level of 1 dot/Å^3 (using our default surface dot density of 16 dots/Å^2), the outer contour contains almost all the contacts and approximates a molecular envelope. The inner contour in the figure, at 2.5 dots/Å^3, reveals the type of non-uniform packing described by Richards (1974): loose regions are surrounded by regions with more contacts. The pocket in the lower domain has been shown to be tolerant of mutations (Gassner, Baase, and Matthews 1996; Karpusas *et al.* 1989). Substrate binds in the cleft between the upper and lower domains.

FIGURE 9-1. Small-probe self contacts in T4 lysozyme with contours of average dot density (outer at 1 dot/Å^3 and inner at 2.5 dots/Å^3).

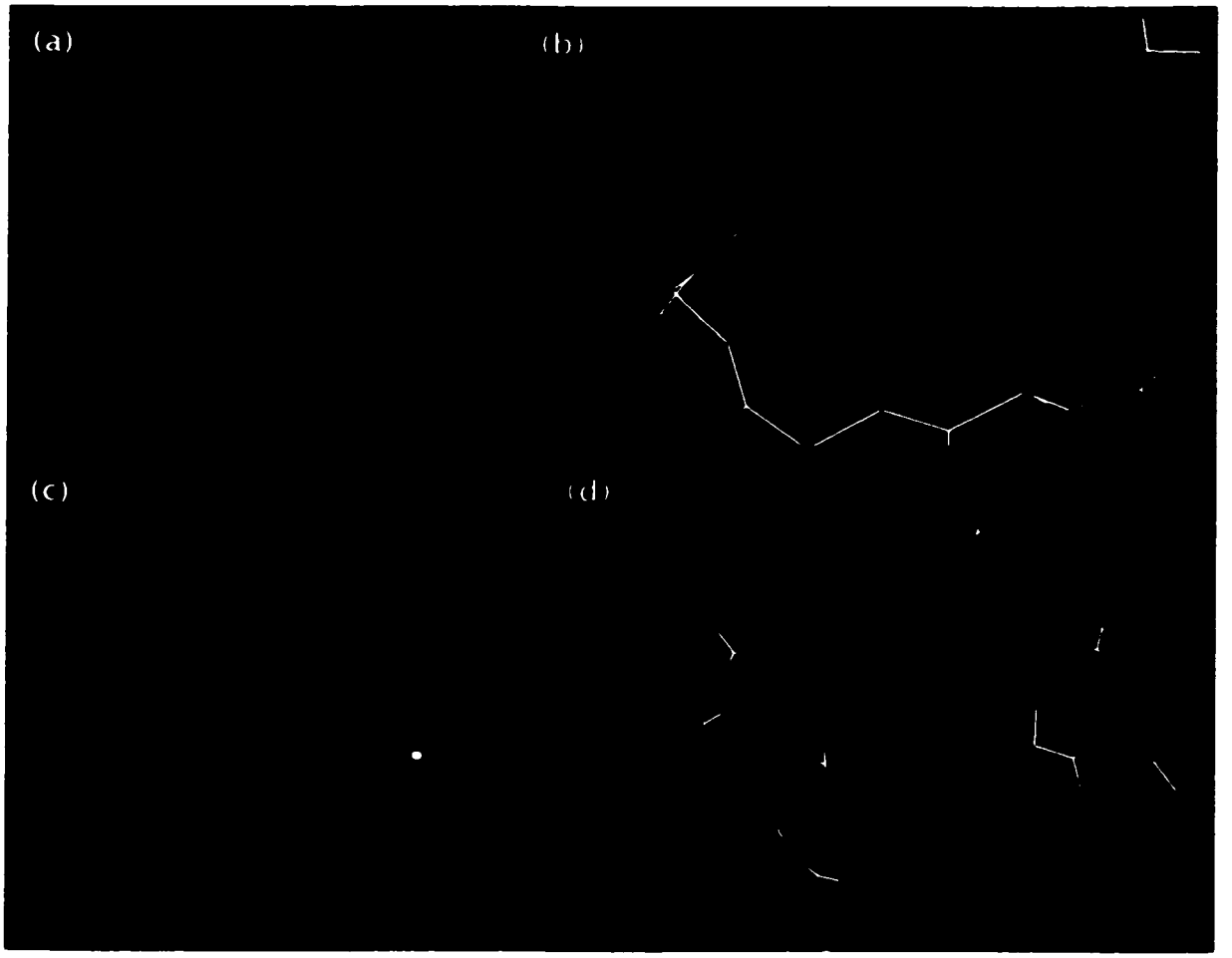
Contours are computed rapidly with the custom program KIN3DCONT. Input consists of a table listing the number of contact dots for each atom in a structure (or the contact area—the count divided by dot density) along with the atom's Cartesian coordinates. A 3D histogram is built by dividing space into a lattice, here 1 Å in each dimension, and binning each atom's dot count. Counts are divided between neighboring lattice points in proportion to how close they are to each. Dots—rather than contact scores—are summed, because the negative contribution of clashes to the score would render clashing regions underpacked; using dots, clashes show high packing. The sums are noisy, varying considerably between adjacent lattice points, so trends are emphasized by light smoothing with a gaussian filter. For

atomic data on a 1 Å grid, a filter with a standard deviation of between 1.5 and 2 Å seems to work well. Contours at arbitrary levels through the data are generated using a *contour tree* algorithm (Zyda 1988). Two dimensional contours for a series of planes cutting through the molecule along each of the x, y and z axes are combined to make a 3D contour mesh. Output from KIN3DCONT (as well as KIN2DCONT and KIN1DLINE) is in either kinemage or PostScript format. A table of smoothed bin totals can also be output for further processing. For example, this function has been used in the production of χ angle frequency distributions, as described in Chapter 7.

Survey Results

Contour map kinemages were generated for each *Top100DB1* structure. In a survey of maps for 13 structures selected at random, contacts were distributed in various ways but none was homogenous. Cellulase (1CEM; Alzari, Souchon, and Dominguez 1996) contains a loosely packed hydrophobic core surrounded by tightly packed helices and a 22 Å long chain of H-bonds; the *c-crk* SH3 domain (1CKA; Wu *et al.* 1995) has a highly packed hydrophobic core; the fungal peroxidase (1ARU; Fukuyama *et al.* 1995) has a central region with many contacts around the heme; flavodoxin/FMN (1RCF; Burkhart *et al.* 1995) has a concentration of contacts in the nucleotide binding pocket; and amylase (1SMD; Ramasubbu *et al.* 1996) shows a number of high-contact zones throughout the molecule.

FIGURE 9-2. Contact density contours for Trp t-RNA synthetase. (a) Overview, showing three high contact density contacts at > 5 dots/Å^3; (b) hydrogen bonds around Lys235; (c) clash from misplaced CH_3 on Met193; (d) extended region of packing near His43.



Detailed study of contours for several proteins, looking at the factors which typically result in high or low average contact area, suggested that by describing packing in terms of contact surface area, the method picked up more than just van der Waals interactions. In the working model of Trp t-RNA synthetase provided by

Charles W. Carter of UNC, Chapel Hill, the three highest concentrations of contact dots (at > 5 dots/Å^3) have three different causes (see Figure 9-2(a)).

Figure 9-2(b)—detail of the rightmost concentration in (a)—shows a region near the anti-codon binding site dominated by hydrogen bonds: four to Lys235, with contributions by Tyr265, Glu225, and both the side chain and main chain of Asp223. This model does not contain any waters, but in other structures, waters are very often at the centers of the highest levels of packing, since they have up to four H-bonds around one oxygen.While these residues are probably very tightly packed by most measures, it would be desirable to have some way to distinguish H-bonds from van der Waals packing, as we do in contact dot displays.

Near the center of the molecule, another concentration of dots surrounds the conserved Met193 (Figure 9-2(c)), where the methyl group was misplaced as described in Chapter 7. Clashes essentially scoop out contacts from one location and pile them in another, and have a much larger impact on the average contact area than on, for example, the average mass density.

Finally, Figure 9-2(d) contains a more extended dog-bone-shaped high-contact region around His43, with some hydrogen bonds but also a great many van der Waals contacts. Carter reports that His43 has been suspected as "a possible transducer to the other subunit" and "it also plays a role in binding a [high affinity] drug candidate" (personal communication). This region really does represent the type of close packing, in this case between adjacent helices, that we were trying to identify in developing our approach.

Alternative Formulations

While H-bonds and clashes were often over-emphasized using the contact-density contours, it appeared that packing near some bulky groups might be underrepresented because large numbers of covalent attachments can leave relatively little potential contact area. Several alternatives were considered in search of a better overall way to characterize how constrained residues are by their surroundings.

Adding Bond Information

Since contacts do not include interactions between bonded atoms, an extra factor can be included for each bond. Lacking any clear theoretical basis, we somewhat arbitrarily made the bond contribution comparable to the contact contribution by placing a value equivalent to 10 dots (or 0.6 $Å^2$ of contact) at the center of covalent bonds to hydrogen and 20 dots (1.2 $Å^2$) at the center of other covalent bonds. While the contours for this system have a similar distribution to those for dot counts alone, the levels tend to be higher along the main chain. This emphasizes secondary structure, which is not really what we were after. The balance between bonds and dot counts has not been optimized for our purposes, primarily because the necessary target function is not immediately obvious. This method has not yet been fully analyzed, but the complexity of juggling information from disparate sources persuaded us to consider other one-component systems.

Percent Coverage

Another one-component method is to scale the contact information to calculate the percentage of each amino-acid residue side chain or main chain covered with dots.

and place this percentage at each atom in the fragment. If a tryptophan sidechain has 40% of its exposed van der Waals surface in close contact, each side-chain atom in that Trp is assigned a value of 40. At the cost of an unwelcome additional level of smoothing (because the percentage applies to the whole group), this system tends to deemphasize water H-bonds. It would require an extra scale factor to deemphasize the main chain.

Atom Volume

The distribution of bonds is essentially the same as the distribution of atoms, and it is a but a short step from thinking about bonds to looking more closely at the atom volume approaches to packing. To take all possible advantage of our use of explicit hydrogens, I developed a table of atom volumes which included consideration of the hydrogens. This table is used by mapvol, an AWK script which processes PDB coordinates to generate input for the contouring software: the volume of each atom and its position.

Volumes were estimated by Monte Carlo integration (Press *et al.* 1988, p237) with custom software (ATVOL). Uniform random sample points inside a minimal bounding box surrounding the target atom are tested for whether they are also inside the atom's clipped van der Waals sphere. Clipping planes bisect the overlap region for covalently bonded neighbors (or all overlapping neighbors, if requested). The fraction of sample points which are within the clipped sphere, multiplied by the volume of the bounding box, yields a volume estimate for the atom in question. Integration by random sampling avoids systematic errors that result from the use of uniform

sampling grids and is simple to implement correctly. Accuracy can by determined by subtracting the estimated volume of the unclipped sphere (calculated at the same time) from the volume based on the formula for the sphere. With 10^7 sample points per atom, the volume is estimated in $Å^3$ to two decimal places. The integration requires too much time to be repeated for each protein. Instead the volume of each atom in each standard amino-acid residue (in Engh and Huber geometry: 1991) is precalculated and used repeatedly. A similar calculation was performed for each atom in the standard nucleotides (using geometry in PDB files 1DNS and 1RNA), and the various heterogens found in the *Top100DB1*.

Contours of atom volume are in dimensionless units of packing density (the fraction of space within the van der Waals shell). With atom volume, we essentially recapitulate the work of Richards (1974; 1977), but with smoothing rather than Voronoi volumes, and with all explicit H atoms. The method is effective and useful. In the Hellinga laboratory, Mike Wisz and Loren Looger are using our atom volume software to monitor designed proteins for loose packing. Its shortcoming, from our point of view, is that it cannot discriminate whether a loosely-fitted group sits in the middle of its enclosing cavity or hugs one of the sides.

Discussion

We want to understand variation in packing. Does loose packing in some protein cores drive tight peripheral packing? For instance, do proteins need to put some residues very close together in order that a few residues can have more freedom, much

like some people at a party might crowd more closely together to let others dance? There are several competing theories of packing and we will probably need surveys that compare detailed contacts, for individual residues in regions of higher and lower packing, to sort this out and choose among them or propose an alternative. The tools I developed will allow the calculation, smoothing, contouring, and display of whichever function or combination of functions the laboratory decides is most suitable for these analyses.

CHAPTER 10 *Concluding Remarks*

All-atom small-probe contact surfaces are information rich, describing molecular goodness-of-fit in terms of van der Waals surface closeness and complementarity, spatial location, extent, type (contact, H-bond, clash), and component identity. Rendered as multi-colored dot surfaces, they can be quite beautiful. They reveal the often astonishing order within large biomolecules. Contacts have allowed us to go beyond the mere tabulation of structural regularities, to develop some understanding of the influences that conspire to produce these regularities. Hydrogens are at the front lines of molecular recognition and even though they are tiny, packing is usually so tight that they cannot be ignored. Contacts are highly critical and just about every problem raises a red flag—important feedback when determining a structure or engineering a new one. It is our belief that close scrutiny of the ever larger set of high quality macromolecular structures with these tools will lead us to a better understand how they come to have their unique forms.

Scientific Programming

One of the most prominent aspects of this work is the influential role played by software tools. The development and creative application of computer programs to solve scientific problems has, in my lifetime, become an accepted—even respected—job description for working scientists. Over the course of a few years, I have built up a collection of programs, used repeatedly in our work and increasingly in other laboratories: a convenient and greatly expanded re-implementation of the small-probe contact surface algorithm, with links to other programs; a novel method for adding the missing hydrogens to structures and automatically fixing certain common mistakes; new programs for measuring, reformatting, summarizing, and visualizing molecular information, along with numerous other handy utilities. These programs represent a form of scientific achievement in themselves, because they help address a long-standing criticism of the kind of "discovery-oriented" research practiced in the Richardson lab—that it requires special talents, and was not sufficiently objective to be duplicated by others—by incorporating many of our insights and heuristics into software products which can be widely disseminated.

Technical Improvements

It is a fact of life that software, like houses, must be maintained or eventually abandoned. My list of technical matters which need attention includes the following:

- Implement the new standard hydrogen naming scheme (Markley *et al.* 1998) in PROBE and REDUCE. Read element type information from the new atom type

field, when available. Add support for the segment-ID field to REDUCE, for files that do not use the chain-ID.

- Extend the size of cliques that can be searched by REDUCE. Genetic algorithms or Dead End Elimination (DEE) may prove useful here.
- Port PROBE and REDUCE to the Macintosh, along with the interactive MAGE/PROBE link. Add contact information to SCULPT.

A more ambitious task, requiring further research, is to include a short-range electrostatics term to the PROBE score. This requires determining how the term is to be calculated and how to scale its contribution to the total score, perhaps using Ramachandran and χ angle maps. It will also have technical ramifications, since it will increase the radii for potential interactions and thus often increase clique sizes.

Collaborations

We will continue to collaborate with other groups (e.g., the RCSB and the groups that maintain O and CNS) to expand the use of explicit hydrogens in macromolecular structure software. This is an idea whose time has come, especially now that computer technology is so much more up to the challenge and since our work has shown how influential hydrogens are in closely packed molecular environments. Likewise, we will seek to have our methodology applied more widely in structure validation software. An all-atom contact surface validation function could prove a helpful addition to the software suite used to support high-throughput structure determination ("structural genomics"). Finally, we will continue to promote the introduction of a step during structural refinement where explicit hydrogens are

added and their van der Waals interactions are taken to full strength. Although numerical instabilities make this perhaps unwise early in refinement, we believe that just as with bond lengths and angles, with good data it should be possible to eventually reach a model which is completely consistent with what we know about excluded volume effects, while remaining faithful to the experimental data.

Other Research Directions

We would like to develop a grading system for side-chain conformations based on contact scores for idealized residues and/or distributions of observed side-chain conformation data. The measured conformation of each residue in protein structure could be graded for rotameric likelihood ("rotamericity") and perhaps color-coded in a display of the structure.

In the maps of PROBE scores *versus* side-chain conformation described in Chapter 8, the side chains from actual crystal structures (not the models) seem to have similarly sized zones of good packing. Is this generally the case? Can we correlate the size of "acceptable" conformation regions with essential side-chain motion? A survey of side-chain packing in our *Top100DB1* could be informative.

We would like to use all-atom contact surfaces to guide the construction of chimeric NMR structures, which combine elements of the various models. Such a combined structure would be a great improvement over the average or the "single best" structure and would also be useful for homology modeling and for determination of 'error-free' consensus structures.

Finally, is there an efficient way to implement contact surface complementarity as a target function in minimization or molecular dynamics/simulated annealing? Can contact surface information be made to work with these systems, perhaps turned into some kind of pseudo-energy? This could allow us to better understand the unique aspects of our surface-based method.

Biological Relevance

The following three questions related to protein design, and macro-molecular structure in general, motivated the work described in previous chapters: (1) What is the nature of how atoms interact in proteins and other biologically important molecules? (2) How can proteins be so tolerant to mutations and still have unique, ordered structures? (3) How restrictive are excluded volume constraints and in what ways can we productively use these constraints to enhance our ability to propose realistic models of macro-molecules?

To address the first question, I have developed and refined software tools which describe atomic interactions at close range, both visually and quantitatively. Our laboratory has produced several databases of high-quality high-resolution protein structures. We have been studying these structures and have begun to publish our findings, including a complete new side-chain rotamer library.

The second question restates the central problem of protein design. While we do not yet know the full answer, our work has shown that natural proteins are internally relaxed for the most part, they have an "inner-repose" which, along with internal

variations in the level of packing and a compliant backbone, may leave room to accommodate a less than perfect fit.

Third, our experience with all-atom contacts makes it clear that steric constraints are very confining. It may well be that the only way that models can satisfy these constraints completely, as well as the experimental data, is to be correct.

APPENDIX 1 *Program and Data Availability*

CD-ROM

Source code for the software listed below and the datasets mentioned on page 239 will be made available on the CD-ROM provided with some copies of this thesis.

Programs

The following is a very brief listing of those programs I developed while in the Richardson laboratory which are most likely to be of continuing interest to others. Further down is a laundry list of Unix scripts used in our work. Source and executables for PROBE, REDUCE and many other programs are available at our FTP/web site: kinemage.biochem.duke.edu.

Key Applications

PROBE is a generic Unix C program for generating small-probe contact surfaces. It has proven to be easy to port to other Unix-like operating systems such as Linux and Windows®. I intend to eventually port PROBE to the Macintosh; perhaps *OS 10* will even allow us to port the MAGE/PROBE link. The current version of PROBE, 2.1.6, includes autobondrot and surface area calculations.

REDUCE is a Unix C++ program for adding and optimizing hydrogens to molecular structure files. It uses C++ templates pretty heavily and was initially hard to port beyond the Silicon Graphics. As C++ compilers have become more capable, versions for Linux and Windows® became available. A Macintosh version is planned. The current version of REDUCE is 2.13.1.

DANG is also a Unix C++ program which generates tables of useful geometric measurements, such as dihedral angles, from PDB format coordinate files. It may have porting difficulties similar to REDUCE.

Also available, for other potential uses, is C source code for the procedure p2sys() which MAGE calls to establish pipes and run external programs.

Contouring

KIN1DLINE, KIN2DCONT, and KIN3DCONT are all C programs which read `value, coordinate` lists to produce line and contour graphics displays, in either kinemage or PostScript format. They construct smoothed histogram data which can be dumped for non-graphical uses.

Utilities

CLUSTER (a C++ program) makes the algorithm REDUCE uses to identify interacting disjoint set (i.e. cliques) available for other uses. For example, the Unix script clashlistcluster uses it to build up groups of clashing residues.

ATVOL is a C program that calculates atomic volume by Monte Carlo integration, as described in "Atom Volume" on page 226.

DERIV is a C program for determining the derivative of noisy data. It was used to analyze distributions of packing data.

BNDLST is a C program. It is really a stripped down version of PROBE which prints a list of covalent and H-bonded neighboring aotms, along with several atomic properties.

Scripts

A large number of command scripts were generated during our work. Most of these are either Unix shell programs, AWK scripts, or some combination. Typically, they run programs such as REDUCE, PROBE, and PREKIN in particular specific ways, piping the results from one program to another. This list is intended to serve as an index to these procedures.

Related to Contacts

The PROBE/O scripts were initially developed by Simon Lovell. We collaborated on the web release.

clashlistscore, clashlistcluster, and clashlistsc produce sorted lists of clashes in various formats. contactscore, atomscore and gapbin format PROBE score information in various ways. bcutsc produces a breakdown of contacts by *B*-factor.

inself uses a standard protocol (probe -3 -quiet -once "sc,het alta blt40 ogt0" "alta blt40" *filename*) to generate dots for internal contacts.

Kinemages

qdotkin (quickly) makes contact dot kinemages from a PDB.

flipNQkin and flipHISkin make animated kinemages for evaluating flips of Asn/Gln or His side chains, using the utility functions mkNQflip, qkinNQa, mkHISflip, qkinHISa.

ca2view creates a *view*, centered on the C^{α}, for each residue in a PDB file.

Related to autobondrot

mkrotscr makes an autobondrot script from a kinemage containing a rotatable group, while pdb2rotscr uses PREKIN to make a .rot script from a PDB file.

For PDB File Manipulation

cutoutpdb marks exterior atoms (accessible to a 1.4 Å radius probe) by making their occupancy negative. h2ocutoutpdb does the same to bulk or surface waters, distinguishing them from buried waters in cavities.

within and withinA identify residues and atoms within a given radius of a specified point .

updateChainID and updateOCC are used to edit the chain-ID and occupancy fields in PDB records (for instance to set H atom occupancies to 1.0 if they had been assigned as 0.0).

rotatePDB, translatePDB, and xformPDB all modify PDB coordinates, for instance to generate the symmetry-related molecules in a multimer.

pdbcns translates between standard PDB format and X-PLOR/CNS atom names, including moving segment-ID information to the standard chain-ID field. Recent changes to the PDB atom name format have not been incorporated. shiftatomname is a handy utility for correcting left-justified atom names in PDB-like files.

For Packing Analysis

scoreDotsAtAtom converts contact dot information into counts of dots per atom.

mapvol assigns a volume to atoms in a PDB file, using a table of values previously compiled with ATVOL.

Coordinate Files, Rotamers, etc.

The annotated *Top100DB1* list of high-resolution structures, the coordinate files with H atoms added, and optimized and the various corrections made, a set of con-

tact-dot kinemage files, and the modified het dictionary are available free at the FTP/web address: kinemage.biochem.duke.edu.

APPENDIX 2 *Rotamer Library*

The complete new side-chain rotamer library described in Chapter 7 is listed below.

Figure A2-1. Rotamers for all standard amino acid side chains

Name	#	%	alpha	beta	other	χ1	χ1	χ2	χ2	χ3	χ3	χ4	χ4	χ1	χ2	χ3	χ4
Arginine						mode	comm.[a]	mode	comm.	mode	comm.	mode	comm.	1/2 Width at 1/2 Height			
ptp85[b]	3	<1%[c]	0%	1%	<1%		**62**		**180**		**65**		**85**				
ptp180	11	1%	0%	2%	2%	71	**62**	171	**180**	65	**65**	-161	**-175**	14	17	10	13
ptt85	16	2%	1%	2%	2%	65	**62**	-178	**180**	-179	**180**	88	**85**	13	14	13	17
ptt180	16	2%	1%	2%	2%	59	**62**	176	**180**	-178	**180**	-177	**180**	15	13	15	19
ptt-85	15	2%	1%	2%	2%	66	**62**	-176	**180**	-178	**180**	-83	**-85**	15	14	12	14
ptm180	6	1%	0%	1%	1%		**62**		**180**		**-65**		**175**				
ptm-85	5	1%	0%	0%	1%		**62**		**180**		**-65**		**-85**				
tpp85	11	1%	3%	1%	<1%	-178	**-177**	57	**65**	57	**65**	85	**85**	13	13	12	15
tpp180	8	1%	1%	0%	1%		**-177**		**65**		**65**		**-175**				
tpt85	20	2%	3%	2%	2%	177	**-177**	64	**65**	180	**180**	86	**85**	14	13	15	14
tpt180	15	2%	3%	1%	1%	179	**-177**	60	**65**	178	**180**	163	**180**	13	17	14	17
ttp85	33	4%	5%	3%	3%	-179	**-177**	177	**180**	65	**65**	83	**85**	14	17	13	15
ttp180	25	3%	5%	3%	1%	-178	**-177**	-178	**180**	65	**65**	-162	**-175**	14	16	14	26
ttp-105	9	1%	1%	1%	1%		**-177**		**180**		**65**		**-105**				
ttt85	19	2%	2%	2%	2%	-175	**-177**	176	**180**	179	**180**	83	**85**	14	14	13	14
ttt180	33	4%	3%	7%	3%	-179	**-177**	177	**180**	-179	**180**	170	**180**	15	13	12	27
ttt-85	26	3%	3%	3%	2%	-179	**-177**	179	**180**	180	**180**	-86	**-85**	15	14	14	15
ttm105	10	1%	2%	1%	<1%	-178	**-177**	170	**180**	-66	**-65**	107	**105**	15	16	15	15
ttm180	13	1%	<1%	4%	1%	180	**-177**	-178	**180**	-67	**-65**	176	**175**	15	12	11	15
ttm-85	28	3%	3%	3%	3%	-175	**-177**	-178	**180**	-65	**-65**	-84	**-85**	14	16	15	14
mtp85	22	2%	2%	3%	2%	-69	**-67**	177	**180**	64	**65**	84	**85**	13	17	13	13
mtp180	45	5%	4%	3%	6%	-65	**-67**	176	**180**	64	**65**	-174	**-175**	12	19	13	19
mtp-105	7	1%	0%	2%	1%	-62	**-67**	179	**180**	67	**65**	-113	**-105**	11	15	13	15
mtt85	34	4%	4%	4%	3%	-67	**-67**	178	**180**	179	**180**	83	**85**	12	19	13	19
mtt180	89	9%	9%	5%	12%	-67	**-67**	-178	**180**	-177	**180**	174	**180**	14	13	13	21
mtt-85	53	6%	4%	7%	6%	-66	**-67**	-177	**180**	-179	**180**	-83	**-85**	13	13	13	13
mtm105	15	2%	1%	1%	2%	-68	**-67**	-179	**180**	-65	**-65**	103	**105**	12	13	13	15
mtm180	48	5%	1%	4%	8%	-68	**-67**	173	**180**	-64	**-65**	180	**175**	14	17	13	30
mtm-85	54	6%	13%	2%	3%	-69	**-67**	-167	**-167**	-63	**-65**	-86	**-85**	14	13	13	13
mmt85	7	1%	1%	1%	1%		**-62**		**-68**		**180**		**85**				
mmt180	18	2%	1%	3%	2%	-63	**-62**	-66	**-68**	-179	**180**	-168	**180**	13	13	10	29
mmt-85	22	2%	<1%	4%	3%	-60	**-62**	-72	**-68**	-178	**180**	-92	**-85**	14	13	15	13
mmm180	11	1%	<1%	2%	2%	-64	**-62**	-74	**-68**	-67	**-65**	172	**175**	14	15	10	13
mmm-85	22	2%	2%	3%	3%	-62	**-62**	-64	**-68**	-61	**-65**	-82	**-85**	14	13	15	13
		82%	79%	81%	84%												
	769/938[d]		234	146	389												

Lysine																	
ptpt	7	1%	0%	2%	<1%		62		180		68		180				
pttp	13	1%	0%	1%	2%	63	62	-170	180	-177	180	72	65	13	14	14	11
pttt	29	2%	0%	4%	3%	63	62	-178	180	178	180	-179	180	13	13	13	10
pttm	8	1%	0%	1%	1%		62		180		180		-65				
ptmt	5	<1%	0%	1%	<1%		62		180		-68		180				
tptp	11	1%	1%	1%	1%	179	-177	59	68	163	180	60	65	13	12	10	11
tptt	32	3%	5%	1%	2%	179	-177	62	68	173	180	171	180	10	10	13	14
tptm	7	1%	1%	1%	<1%		-177		68		180		-65	14	9	12	10
ttpp	12	1%	1%	<1%	1%		-177		180		68		65				
ttpt	25	2%	2%	5%	1%	180	-177	179	180	78	68	179	180	14	12	14	14
tttp	49	4%	5%	5%	3%	-177	-177	180	180	171	180	63	65	14	13	12	12
tttt	162	13%	17%	19%	10%	-177	-177	178	180	179	180	180	180	13	13	15	13
tttm	37	3%	4%	2%	3%	-177	-177	172	180	178	180	-72	-65	12	13	15	13
ttmt	20	2%	2%	4%	1%	-175	-177	-174	180	-69	-68	179	180	14	14	10	15
ttmm	5	<1%	1%	0%	<1%		-177		180		-68		-65				
mptt	4	<1%	0%	0%	1%		-90		68		180		180	10	9	10	13
mtpp	12	1%	1%	1%	1%	-69	-67	-179	180	70	68	67	65	12	13	11	9
mtpt	38	3%	4%	2%	3%	-69	-67	164	180	62	68	-179	180	13	13	14	14
mttp	42	3%	2%	4%	4%	-67	-67	-176	180	174	180	76	65				
mttt	244	20%	23%	14%	21%	-67	-67	176	180	179	180	177	180	14	13	12	14
mttm	56	5%	3%	5%	6%	-67	-67	-179	180	-179	180	-63	-65	13	12	13	14
mtmt	40	3%	6%	2%	3%	-70	-67	-170	180	-66	-68	-175	180	12	13	14	13
mtmm	12	1%	0%	1%	1%	-70	-67	-179	180	-66	-68	-64	-65	12	12	12	11
mmtp	9	1%	<1%	0%	1%		-62		-68		180		65				
mmtt	77	6%	3%	5%	8%	-58	-62	-61	-68	-177	180	-179	180	12	13	13	13
mmtm	18	1%	1%	1%	2%	-59	-62	-69	-68	-176	180	-70	-65	14	12	10	15
mmmt	10	1%	<1%	1%	1%	-59	-62	-58	-68	-75	-68	-174	180	12	13	10	15
		81%	82%	80%	82%												
	984/1209		261	194	529												
Name	#	%	alpha	beta	other	χ^1	χ^1	χ^2	χ^2	χ^3	χ^3		χ^3	χ^1	χ^2	χ^3	
Methionine						mode	**comm.**	mode	**comm.**	mode	**comm.**		range[e]	1/2 Width at 1/2 Height			
ptp	12	2%	1%	3%	3%	68	**62**	-167	**180**	88	**75**			11	17	12	
ptm	17	3%	1%	6%	4%	67	**62**	174	**180**	-78	**-75**			9	10	9	
tpp	30	5%	8%	2%	5%	-177	**-177**	66	**65**	75	**75**			10	15	15	
tpt	9	2%	1%	4%	1%	179	**-177**	67	**65**	-179	**180**			9	8	9	
ttp	28	5%	7%	7%	2%	176	**-177**	178	**180**	73	**75**			10	11	11	
ttt	17	3%	5%	2%	2%	180	**-177**	171	**180**	174	**180**			9	9	19	
ttm	36	7%	3%	10%	8%	-177	**-177**	176	**180**	-78	**-75**			10	10	13	
mtp	92	17%	22%	10%	17%	-68	**-67**	177	**180**	72	**75**			10	12	14	
mtt	43	8%	9%	8%	7%	-67	**-67**	177	**180**	-178	**180**			10	13	15	
mtm	58	11%	12%	11%	9%	-67	**-67**	-177	**180**	-76	**-75**			12	11	16	
mmp	15	3%	3%	1%	4%	-64	**-65**	-63	**-65**	103	**103**			9	10	10	
mmt	10	2%	0%	2%	3%	-63	**-65**	-64	**-65**	180	**180**			12	14	19	
mmm	105	19%	21%	16%	19%	-66	**-65**	-60	**-65**	-67	**-70**			11	13	16	
		86%	91%	84%	83%												
	472/550		175	112	185												

Glutamate														
pt-20°	80	5%	1%	9%	7%	63	**62**	-175	**180**	-18	**-20**	-90 to 90	14	13	23
pm0°	32	2%	0%	0%	4%	71	**70**	-79	**-80**	5	**0**	-50 to 50	14	13	17
tp10°	91	6%	10%	2%	6%	-177	**-177**	65	**65**	13	**10**	-10 to 90	14	13	17
tt 0°	350	24%	25%	42%	18%	-177	**-177**	178	**180**	2	**0**	-90 to 90	14	14	30
tm-20°	17	1%	1%	1%	1%		**-177**		**-80**		**-25**	-50 to 10	13	13	15
mp0°	88	6%	<1%	2%	10%	-65	**-65**	85	**85**	-3	**0**	-60 to 60	14	13	25
mt-10°	484	33%	36%	29%	32%	-67	**-67**	177	**180**	-10	**-10**	-90 to 90	13	16	25
mm-40°	197	13%	19%	7%	12%	-65	**-65**	-58	**-65**	-40	**-40**	-90 to 30	14	14	25
		91%	92%	92%	90%										
	1339/1470		394	225	720										
Glutamine															
pt 20°	37	4%	1%	5%	6%	64	**62**	180	**180**	20	**20**	-90 to 90	13	14	16
pm0°	15	2%	0%	1%	3%		**70**		**-75**		**0**	-60 to 60			
tp-100°	14	2%	4%	2%	<1%		**-177**		**65**		**-100**	-150 to 0			
tp60°	78	9%	13%	9%	7%	-175	**-177**	64	**65**	60	**60**	0 to 90	14	15	24
tt 0°	140	16%	16%	29%	12%	-174	**-177**	173	**180**	-5	**0**	-90 to 90	14	13	40
mp0°	24	3%	<1%	1%	5%		**-65**		**85**		**0**	-60 to 60			
mt -30°	304	35%	40%	26%	36%	-67	**-67**	177	**180**	-25	**-25**	-90 to 90	16	15	37
mm-40°	127	15%	12%	13%	17%	-66	**-65**	-60	**-65**	-40	**-40**	-95 to 0	16	18	26
mm100°	22	3%	4%	1%	2%		**-65**		**-65**		**100**	0 to 150			
		88%	89%	86%	88%										
	761/863		229	137	395										
												χ^2 range			
Aspartate															
p-10°	203	10%	1%	2%	13%	61	**62**	-4	**-10**			-90 to 0	9	19	
p30°	194	9%	1%	5%	12%	65	**62**	9	**30**			0 to 90	8	14	
t0°	438	21%	8%	44%	20%	-176	**-177**	1	**0**			-50 to 50	12	30	
t70°	118	6%	11%	7%	4%	-179	**-177**	65	**65**			50 to 90	12	18	
m -20°	1088	51%	77%	38%	47%	-71	**-70**	-15	**-15**			-90 to 20	10	16	
		96%	97%	95%	96%										
	2041/2124		365	232	1444										
Asparagine															
p-10°	103	7%	0%	1%	10%	63	**62**	-13	**-10**			-90 to 0	8	9	
p30°	132	9%	<1%	7%	12%	64	**62**	34	**30**			0 to 90	6	7	
t-20°	177	12%	5%	21%	12%	-174	**-174**	-20	**-20**			-120 to 0	5	21	
t30°	228	15%	13%	18%	15%	-168	**-177**	31	**30**			0 to 80	14	22	
m-20°	580	39%	65%	28%	33%	-71	**-65**	-23	**-20**			-60 to 10	10	20	
m-80°	118	8%	8%	9%	8%	-71	**-65**	-76	**-75**			-100 to -60	9	9	
m120°	58	4%	3%	3%	4%	-64	**-65**	132	**120**			60 to 160	9	18	
		94%	95%	88%	95%										
	1396/1490		293	179	924										

Name	#	%	alpha	beta	other	χ1 mode	χ1 comm.	χ2 mode	χ2 comm.	χ2 range	χ1 1/2 Width	χ2 1/2 Width at 1/2 Height
Isoleucine												
pp	10	1%	<1%	1%	<1%		**62**		**100**			
pt	216	13%	4%	13%	22%	61	**62**	171	**170**		10	10
tp	36	2%	2%	1%	4%	-169	**-177**	66	**66**		13	11
tt	127	8%	1%	8%	14%	-174	**-177**	167	**165**		13	11
mp	19	1%	0%	2%	1%		**-65**		**100**			
mt	993	60%	81%	58%	41%	-66	**-65**	169	**170**		10	10
mm	242	15%	10%	16%	17%	-57	**-57**	-59	**-60**		10	10
		99%	99%	98%	99%							
	1643/1667		496	629	518							
Leucine												
pp	21	1%	<1%	2%	1%		**62**		**80**			
tp	750	29%	30%	36%	23%	177	**-177**	63	**65**		10	10
tt	49	2%	1%	3%	1%	-172	**-172**	147	**145**	120 to 180	9	9
mp	63	2%	1%	5%	2%	-85	**-85**	66	**65**	45 to 105	11	14
mt	1548	59%	62%	46%	66%	-65	**-65**	174	**175**		11	11
		93%	95%	93%	93%							
	2431/2602		836	644	951							
Histidine												
p-80°	51	9%	0%	6%	13%	60	**62**	-75	**-75**	-120 to -50	10	12
p80°	26	4%	0%	4%	6%	61	**62**	78	**80**	50 to 120	13	10
t-160°	31	5%	5%	14%	1%	-178	**-177**	-163	**-165**	150 to -120	12	20
t-80°	64	11%	17%	9%	9%	-173	**-177**	-81	**-80**	-120 to -50	10	22
t60°	94	16%	24%	17%	12%	-178	**-177**	62	**60**	50 to 120	13	19
m-70°	174	29%	26%	30%	30%	-60	**-65**	-69	**-70**	-120 to -30	11	23
m170°	44	7%	9%	3%	9%	-63	**-65**	165	**165**	120 to -160	10	16
m80°	78	13%	14%	10%	14%	-66	**-65**	83	**80**	50 to 120	11	18
		94%	94%	92%	95%							
	562/598		124	143	295							
Tryptophan												
p-90°	67	11%	2%	13%	14%	58	**62**	-87	**-90**	-130 to -60	12	10
p90°	34	6%	1%	9%	6%	60	**62**	92	**90**	60 to 130	12	8
t-105°	100	16%	27%	10%	14%	178	**-177**	-105	**-105**	-130 to -60	16	14
t90°	109	18%	28%	14%	15%	-178	**-177**	88	**90**	0 to 100	10	11
m-90°	31	5%	0%	7%	7%	-70	**-65**	-87	**-90**	-130 to -60	9	12
m0°	48	8%	15%	2%	8%	-66	**-65**	-4	**-5**	-40 to 20	9	20
m95°	195	32%	22%	43%	29%	-69	**-65**	95	**95**	60 to 130	11	19
		94%	95%	98%	92%							
	584/618		140	175	269							

Tyrosine												
p90°	182	13%	1%	21%	12%	63	**62**	89	**90**	60 to 90, -90 to-60	13	13
t80°	486	34%	55%	25%	30%	176	**-177**	77	**80**	20 to 90, -90 to -75	11	14
m-85°	618	43%	26%	50%	45%	-65	**-65**	-87	**-85**	50 to 90, -90 to -50	11	21
m-30°	124	9%	15%	4%	9%	-64	**-65**	-42	**-30**	-50 to 0, 0 to 50	11	18
		98%	97%	99%	97%							
	1410/1443		290	468	652							
Phenylalanine												
p90°	202	13%	1%	24%	11%	59	**62**	88	**90**	60 to 90, -90 to-60	11	11
t80°	522	33%	57%	18%	29%	177	**-177**	80	**80**	20 to 90, -90 to -75	13	17
m-85°	697	44%	29%	51%	47%	-64	**-65**	-83	**-85**	50 to 90, -90 to -50	12	17
m-30°	149	9%	12%	5%	11%	-64	**-65**	-19	**-30**	-50 to 0, 0 to 50	9	20
		98%	97%	99%	98%							
	1570/1599		389	514	667							
Proline												
Cγ endo	379	44%	23%	54%	43%	30	**30**			15 to 60	7	
Cγ exo	372	43%	68%	28%	44%	-29	**-30**			-60 to -15	6	
cis, Cγ endo	56	6%	0%	1%	7%	31	**30**			15 to 60	5	
		93%	91%	84%	94%							
	807/928		20	57	730							

Name	#	%	alpha	beta	other	χ1 act.	χ1 **com.**	χ1 1/2 Width at 1/2 Height
Threonine								
p	1200	49%	25%	31%	65%	59	**62**	10
t	169	7%	0%	13%	6%	-171	**-175**	6
m	1062	43%	74%	55%	29%	-61	**-65**	7
		99%	100%	99%	99%			
	2431/2447		395	672	1364			

Valine									
p	169	6%	2%	8%	8%	63	**63**	="177"[f]	8
t	1931	73%	90%	72%	63%	175	**175**	="-65"	8
m	526	20%	7%	20%	28%	-64	**-60**	="60"	7
		99%	100%	99%	99%				
	2626/2649		622	1080	924				
Serine									
p	1201	48%	33%	36%	55%	64	**62**		10
t	541	22%	22%	34%	18%	178	**-177**		11
m	714	29%	44%	29%	25%	-65	**-65**		9
		98%	98%	100%	98%				
	2456/2498		350	485	1621				
Cysteine									
p	64	23%	5%	23%	34%	55	**62**		14
t	74	26%	20%	45%	21%	-177	**-177**		10
m	142	50%	75%	32%	43%	-65	**-65**		11
		99%	100%	100%	98%				
	280/285		85	65	130				

[a] "mode" indicates the peak of the smoothed distribution, "comm." indicates the common-atom value (given in bold face).

[b] Mode and 1/2 width at 1/2 height values are not given for minor rotamers.

[c] <1% indicates a value between 0.5% and 0%. 0% indicates no observations.

[d] Total number of rotameric side chains / Total number that pass all data filters.

[e] Ranges used in determining frequencies are normally common-atom values ±30°. Exceptions (always in the terminal χ value) are listed here.

[f] Standard conventions [29,30] result in χ angles being named differently for Val than for Thr and Ile. These figures indicate the equivalent angles.

Bibliography

Abagyan, Ruben A., Maxim M. Totrov, and Dmitry Kuznetsov. 1994. ICM-A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation. *Journal of Computational Chemistry* **15**: 488-506.

Aho, Alfred V., Brian W. Kernighan, and Peter J. Weinberger. 1988. *The AWK Programming Language*. Edited by Michael A. Harrison. Addison-Wesley Series in Computer Science. New York: Addison-Wesley Publishing Company.

Alzari, P.M., H. Souchon, and R. Dominguez. 1996. The crystal structure of endoglucanase CelA, a family 8 glycosyl hydrolase from *Clostridium thermocellum*. *Structure* **4**: 265-275.

Axe, D.D., N.W. Foster, and A.R. Fersht. 1996. Active barnase variants with completely random hydrophobic cores. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 5590-5594.

Bader, R.F.W., I. Keaveny, and P.E. Cade. 1967. Molecular Charge Distributions and Chemical Bonding. II. First-Row Diatomic Hydrides, AH. *Journal of Chemical Physics* **47**: 3381-3402.

Baldwin, Enoch P., Omid Hajiseyedjavadi, Walter A. Baase, and Brian W. Matthews. 1993. The Role of Backbone Flexibility in the Accommodation of Variants That Repack the Core of T4 Lysozyme. *Science* **262**: 1715-1718.

Baldwin, Enoch P., Jian Xu, Omid Hajiseyedjavadi, Walter A. Baase, and Brian W. Matthews. 1996. Thermodynamic and Structural Compensation in "Size-switch" Core Repacking Variants of Bacteriophage T4 Lysozyme. *Journal of Molecular Biology* **259**: 542-559.

Bancroft, D., L.D. Williams, A. Rich, and M. Egli. 1994. The Low-Temperature Crystal Structure of the Pure-Spermine Form of Z-DNA Reveals Binding of a Spermine Molecule in the Minor Groove. *Biochemistry* **33**: 1073-1086.

Barthelme, Donald. 1998. *The Teachings of Don B.: Satires, Parodies, Fables, Illustrated Stories, and Plays of Donald Barthelme*. Edited by Kim A. Herzinger. New York: Vintage International.

Bass, M.B., D.F. Hopkins, W.A.N. Jaquysh, and R.L. Ornstein. 1992. A method for determining the positions of polar hydrogens added to a protein structure that maximizes protein hydrogen bonding. *Proteins* **12**: 266-277.

Beamer, L.J. and C.O. Pabo. 1992. Refined 1.8 Å crystal structure of the λ repressor-operator complex. *Journal of Molecular Biology* **227**: 177-196.

Behe, M.J., E.E. Lattman, and G.D. Rose. 1991. The protein-folding problem: The native fold determines packing, but does packing determine the native fold? *Proceedings of the National Academy of Sciences of the United States of America* **88**: 4195-4199.

Benedetti, E., G. Morelli, G. Nemethy, and H.A. Scheraga. 1983. Statistical and energetic analysis of side-chain conformations in oligopeptides. *International Journal of Peptide and Protein Research* **22**: 1-15.

Berendsen, H.J.C., J.P.M. Postma, W.F. van Gunsteren, and J. Hermans. 1981. Interaction Models for Water in Relation to Protein Hydration. In *Intermolecular Forces*, ed. B. Pullman:331-342: D. Reidel Publishing Company.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Research* **28**, no. 1: 235-242.

Bernstein, F.C., T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology* **112**: 535-542.

Betz, S.F., D.P. Raleigh, and W.F. DeGrado. 1993. *De novo* protein design: from molten globules to native-like states. *Current Opinion in Structural Biology* **3**: 601-610.

Bhat, T.N., V. Sasisekharan, and M. Vijayan. 1979. An Analysis of Side-Chain Conformation in Proteins . *International Journal of Peptide and Protein Research* **13**: 170-184.

Blaber, Michael, Joel D. Lindstrom, Nadine Gassner, Jian Xu, Dirk W. Heinz, and Brian W. Matthews. 1993. Energic Cost and Structural Consequences of Burying a Hydroxyl Group within the Core of a Protein Determined from Ala → Ser and Val → Thr Substitutions in T4 Lysozyme. *Biochemistry* **32**: 11363-11373.

Blaber, Michael, Xue-jun Zhang, Joel D. Lindstrom, Sheila D. Pepiot, Walter A. Baase, and Brian W. Matthews. 1994. Determination of alpha-Helix Propensity within the Context of a Folded Protein. *Journal of Molecular Biology* **235**: 600-624.

Blaber, Michael, Xue-jun Zhang, and Brian W. Matthews. 1993. Structural Basis of Amino Acid alpha Helix Propensity. *Science* **260**: 1637-1640.

Bode, W., E. Papamokos, and D. Musil. 1987. The high-resolution X-ray crystal structure of the complex formed between subtilisin Carlsberg and eglin c, an elastase inhibitor from the leech, *Hirudo medicinalis*. *European Journal of Biochemistry* **166**: 673-692.

Bohacek, Regine S. and Colin McMartin. 1992. Definition and display of steric, hydrophobic, and hydrogen-bonding properties of ligand binding sites in proteins using Lee and Richards accessible surface: validation of a high-resolution graphical tool for drug design. *Journal of Medicinal Chemistry* **35**, no. 10: 1671-1684.

Bolin, J.T., D.J. Filman, D.A. Matthews, R.C. Hamlin, and J. Kraut. 1982. Crystal structures of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase refined at 1.7 Å resolution. I. General features and binding of methotrexate. *The Journal of Biological Chemistry* **257**: 13650-13662.

Bondi, A. 1964. van der Waals Volumes and Radii. *Journal of Physical Chemistry* **68**:3: 441-451.

Bonvin, A.M., J.A. Rullman, R.M. Lamerichs, R. Boelens, and R. Kaptein. 1993. "Ensemble" iterative relaxation matrix approach: a new NMR refinement protocol applied to the solution structure of crambin. *Proteins: Structure, Function, and Genetics* **15**: 385-400.

Bowie, J.U., R. Luthy, and D. Eisenberg. 1991. A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. *Science* **253**: 164-170.

Bromberg, S. and K.A. Dill. 1994. Side-chain entropy and packing in proteins. *Protein Science* **3**: 997-1009.

Brunger, A.T. 1992. *X-PLOR version 3.1: A System for X-ray Crystallography and NMR*. New Haven, CT: Yale University Press.

Brunger, A. T., P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J.-S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson, and G. L. Warren. 1998. Crystallography and NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta Crystallographica, Section D* **54**: 905-921.

Burkhart, B.M., B. Ramakrishnan, H. Yan, R.J. Reedstrom, J.L. Markley, N.A. Straus, and M. Sundaralingam. 1995. Structure of the trigonal form of recombinant oxidized flavodoxin from *Anabaena* 7120 at 1.40 Å resolution. *Acta Crystallographica, Section D* **51**: 318-330.

Burton, Randall E., Guewha S. Huang, Margaret A. Daugherty, Paul W. Fullbright, and Terrence G. Oas. 1996. Microsecond protein folding through a compact transition state. *Journal of Molecular Biology* **263**: 311-322.

Calderone, Tiffany L., Robert D. Stevens, and Terrence G. Oas. 1996. High-level Misincorporation of Lysine for Arginine at AGA Codons in a Fusion Protein Expressed in *Escherichia coli*. *Journal of Molecular Biology* **262**: 407-412.

Carson, Mike. 1997. Ribbons. *Methods In Enzymology* **277**: 493-505.

Carugo, K.D., A. Battistoni, M.T. Carri, F. Polticelli, A. Desideri, G. Rotilio, A. Coda, K.S. Wilson, and M. Bolognesi. 1996. Three-dimensional structure of *Xenopus laevis* Cu,Zn superoxide dismutase *b* determined by X-ray crystallography at 1.5 Å resolution. *Acta Crystallographica, Section D* **52**: 176-188.

Carugo, O. and P. Argos. 1997. Correlation between side-chain mobility and conformation in protein structures. *Protein Engineering* **10**: 777-787.

Chandrasekaran, R. and G.N. Ramachandran. 1970. Studies on the Conformation of Amino Acids XI. Analysis of the Observed Side Group Conformations in Proteins. *International Journal of Protein Research II*: 223-233.

Choma, C.T., J.D. Lear, M.J. Nelson, P.L. Dutton, D.E. Robertson, and W.F. DeGrado. 1994. Design of a Heme-Binding Four-Helix Bundle. *Journal of the American Chemical Society* **116**: 856-865.

Chothia, Cyrus. 1974. Hydrophobic bonding and accessible surface area in proteins. *Nature* **248**: 338-339.

Chothia, Cyrus and Mark Gerstein. 1997. How far can sequences diverge? *Nature* **385**: 579-581.

Creighton, Thomas E., Christopher J. Bagley, Leanne Cooper, Nigel J. Darby, Robert B. Freedman, Johan Kemmink, and Amina Sheikh. 1993. On the Biosynthesis of Bovine Pancreatic Trypsin Inhibitor (BPTI): Structure, Processing, and Disulphide Bond Formation of the Precursor in Vitro and in Microsomes. *Journal of Molecular Biology* **232**: 1176-1196.

Dahiyat, B. I. and S. L. Mayo. 1997a. *De Novo* Protein Design: Fully Automated Sequence Selection. *Science* **278**: 82-87.

Dahiyat, B. I. and S. L. Mayo. 1997b. Probing the role of packing specificity in protein design. *Proceedings of the National Academy of Sciences of the United States of America* **94**: 10172-10177.

Dalal, S., S. Balasubramanian, and L. Regan. 1997. Transmuting α helices and β sheets. *Folding & Design* **2**: R71-R79.

Dauter, Z., V.S. Lamzin, and K.S. Wilson. 1997. The benefits of atomic resolution. *Current Opinion in Structural Biology* **7**: 681-688.

DeMaeyer, Marc, Johan Desmet, and Ignace Lasters. 1997. All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Folding & Design* **2**: 53-66.

Derewenda, Z.S., L. Lee, and U. Derewenda. 1995. The occurrence of C-H · · · O hydrogen bonds in proteins. *Journal of Molecular Biology* **252**: 248-262.

Desjarlais, J. R. and T. M. Handel. 1995. *De novo* design of the hydrophobic cores of proteins. *Protein Science* **4**: 2006-2018.

Desmet, Johan, Marc De Maeyer, Bart Hazes, and Ignace Lasters. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**: 539-542.

Dill, Ken A., Sarina Bromberg, Kaizhi Yue, Klaus M. Fiebig, David P. Yee, Paul D. Thomas, and Hue Sun Chan. 1995. Principles of protein folding—a perspective from simple exact models. *Protein Science* **4**: 561-602.

Dill, Ken A. and Hue Sun Chan. 1997. From Levinthal to pathways to funnels. *Nature Structural Biology* **4**: 10-19.

Doering, Don Shimon. 1992. Functional and structural studies of a small F-actin binding domain. Ph. D., Massachusetts Institute of Technology.

Drenth, Jan, Barbara W. Low, Jane S. Richardson, and Christine S. Writht. 1980. The toxin-agglutinin fold. *Journal of Biological Chemistry* **255**, no. 7: 2652-2655.

Dunbrack, Roland L. and Fred E. Cohen. 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science* **6**: 1661-1681.

Dunbrack, Roland L. and Martin Karplus. 1993. Backbone-dependent Rotamer Library for Proteins: Application to Side-chain Prediction. *Journal of Molecular Biology* **230**: 543-574.

Dunbrack, Roland L. and Martin Karplus. 1994. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nature Structural Biology* **1**: 334-340.

Dyson, H.J., G.P. Pippert, D.A. Case, A. Holmgren, and P.E. Wright. 1990. Three dimensional structure of the reduced form of *E. coli* thioredoxin determined by nuclear magnetic resonance spectroscopy. *Biochemistry* **29**: 4129-4139.

Eisenhaber, Frank, Philip Lijnzaad, Patrick Argos, Chris Sander, and Michael Scharf. 1995. The double cubic lattice method: efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *Journal of Computational Chemistry* **16**, no. 3: 273-284.

Endo, Yaeta, Kazuhiro Mitsui, Mitsuyoshi Motizuki, and Kunio Tsurugi. 1987. The Mechanism of Action of Ricin and Related Toxic Lectins on Eukaryotic Ribosomes. *The Journal of Biological Chemistry* **262**, no. 12: 5908-5912.

Engh, R.A. and R. Huber. 1991. Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement. *Acta Crystallographica, Section A* **47**: 392-400.

Epp, O., E.E. Lattman, M. Schiffer, R. Huber, and W. Palm. 1975. The Molecular Structure of a Dimer Composed of the Variable Portions of the Bence-Jones Protein REI Refined at 2.0 Å Resolution. *Biochemistry* **14**: 4943-4952.

Erickson, Bruce W., S. B. Daniels, P. A. Reddy, Cecilia G. Unson, David C. Richardson, and Jane S. Richardson. 1986. Betabellin: an engineered protein. In *Computer Graphics and Molecular Modeling*, ed. R. Fletterick and M. Zoller:53-57: Cold Spring Harbor Lab.

Eriksson, A. E., Walter A. Baase, and Brian W. Matthews. 1993. Similar Hydrophobic Replacements of Leu99 and Phe153 within the Core of T4 Lysozyme Have Different Structural and Thermodynamic Consequences. *Journal of Molecular Biology* **229**: 747-769.

Faber, H. R. and Brian W. Matthews. 1990. A mutant T4 lysozyme displays 5 different crystal conformations. *Nature* **348**: 263-266.

Fedorov, A.N., D.A. Dolgikh, V.V. Chemeris, B.K. Chernov, A.V. Finkelstein, A.A. Schulga, Y.B. Alakhov, M.P. Kirpichnikov, and O.B. Ptitsyn. 1992. *De novo* Design,

Synthesis and Study of Albebetin, a Polypeptide with a Predetermined Three-dimensional Structure. *Journal of Molecular Biology* **225**: 927-931.

Fernandez, C., T. Szyperski, T. Bruyere, P. Ramage, E. Mosinger, and K. Wuthrich. 1997. NMR Solution Structure of the Pathogenesis-related Protein P14A. *Journal of Molecular Biology* **266**: 576-593.

Ferré-D'Amaré, A. R., K. Zhou, and J. A. Doudna. 1998. Crystal Structure of a Hepatitis Delta Virus Ribozyme. *Nature* **395**, no. 6702: 567-574.

Fezoui, Y., D.L. Weaver, and J.J. Osterhout. 1994. *De novo* design and structural characterization of an α-helical hairpin peptide: A model system for the study of protein folding intermediates. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 3675-3679.

Fisher, A.J., T.B. Thompson, J.B. Thoden, T.O. Baldwin, and I. Rayment. 1996. The 1.5 Å Resolution Crystal Structure of Bacterial Luciferase in Low Salt Conditions. *Journal of Biological Chemistry* **271**: 21956-21968.

Foloppe, Nicolas and Alexander D. Mackerell, Jr. 2000. All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data. *Journal of Computational Chemistry* **21**, no. 2: 86-104.

Fossey, S.A., G. Nemethy, K.D. Gibson, and H.A. Scheraga. 1991. Conformational Energy Studies of β-Sheets of Model Silk Fibroin Peptides. I. Sheets of Poly(Ala-Gly) Chains. *Biopolymers* **31**: 1529-1541.

Frankel, Arthur, David Schlossman, Phil Welsh, Andrew Hertler, David Withers, and Stephen Johnston. 1989. Selection and Characterization of Ricin Toxin A-Chain Mutations in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology* **9**, no. 2: 415-420.

Frankel, A., P. Welsh, J.S. Richardson, and J. D. Robertus. 1990. Role of Arginine 180 and Glutamic Acid 177 of Ricin Toxin A Chain in Enzymatic Inactivation of Ribosomes. *Molecular and Cellular Biology* **10**: 6257-6263.

Freedman, Robert B. 1995. The formation of protein disulfide bonds. *Current Opinion in Structural Biology* **5**: 85-91.

Fukuyama, K., N. Kunishima, F. Amada, T. Kubota, and H. Matsubara. 1995. Crystal structures of cyanide- and triiodide-bound forms of *Arthromyces ramosus* peroxidase at different pH values. *Journal of Biological Chemistry* **270**: 21884-21892.

Gassner, N.C., W.A. Baase, and B.W. Matthews. 1996. A test of the "jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of T4

lysozyme. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 12155-12158.

Gavezzotti, A. 1983. The Calculation of Molecular Volumes and the Use of Volume Analysis in the Investigation of Structured Media and of Solid-State Organic Reactivity. *Journal of the American Chemical Society* **105**: 5220-5225.

Geilman, S.H. 1991. On the Role of Methionine Residues in the Sequence-Independent Recognition of Nonpolar Protein Surfaces. *Biochemistry* **30**: 6633-6636.

Gernert, Kim M., Mark C. Surles, Thomas H. LaBean, Jane S. Richardson, and David C. Richardson. 1995. The Alacoil: A Very Tight Antiparallel Coiled-coil of α-helices. *Protein Science* **4**: 2252-2260.

Ghaemmaghami, Sina, J. Michael Word, Randall E. Burton, Jane S. Richardson, and Terrence G. Oas. 1998. Folding Kinetics of a Fluorescent Variant of Monomeric λ Repressor. *Biochemistry* **37**: 9179-9185.

Glykos, Nicholas M., Gianni Cesareni, and Michael Kokkinidis. 1999. Protein plasticity to the extreme: changing the topology of a 4-α-helical bundle with a single amino acid substitution. *Structure* **7**: 597-603.

Grell, Daniel, Jane S. Richardson, David C. Richardson, and Manfred Mutter. submitted. SymROP: ROP Protein with Identical Helices, Redesigned by All-atom Contact Analysis and Molecular Dynamics. *Journal of Molecular Graphics and Modeling* (in press).

Hagler, A.T., E. Huler, and S. Lifson. 1974. Energy Functions for Peptides and Proteins. I. Derivation of a Consistent Force Field Including the Hydrogen Bond from Amide Crystals. *Journal of the American Chemical Society* **96**:17: 5319-5327.

Hanahan, Douglas. 1983. Studies on transformation of *Escherichia coli* with plasmids. *Journal of Molecular Biology* **166**: 557-580.

Harbury, Pehr B., Joseph J. Plecs, Bruce Tidor, Tom Alber, and Peter S. Kim. 1998. High-resolution protein design with backbone freedom. *Science* **282**: 1462-1467.

Harrison, Paul M. and Michael J. E. Sternberg. 1996. The disulphide β-cross: from cystine geometry and clustering to classification of small disulphide-rich protein folds. *Journal of Molecular Biology* **264**: 603-623.

Hecht, M.H., J.S. Richardson, D.C. Richardson, and R.C. Ogden. 1990. *De Novo* Design, Expression, and Characterization of Felix: A Four-Helix Bundle Protein of Native-Like Sequence. *Science* **249**: 884-891.

Hegde, R.S., S.R. Grossman, L.A. Laimins, and P.B. Sigler. 1992. Crystal structure at 1.7 Å of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target. *Nature* **359**: 505-512.

Hellinga, Homme W. 1998. The construction of metal centers in proteins by rational design. *Folding & DesignFolding & Design* **3**: R1-R8.

Hermans, J., H.J.C. Berendsen, W.F. van Gunsteren, and J.P.M. Postma. 1984. A Consistent Empirical Potential for Water-Protein Interactions. *Biopolymers* **23**: 1513-1518.

Hingerty, B., R.S. Brown, and A. Jack. 1978. Further Refinement of the Structure of Yeast $tRNA^{Phe}$. *Journal of Molecular Biology* **124**: 523-534.

Holbrook, S.R., R.E. Dickerson, and S.-H. Kim. 1985. Anisotropic Thermal-Parameter Refinement of the DNA Dodecamer CGCGAATTCGCG by the Segmented Rigid-Body Method. *Acta Crystallographica, Section B* **41**: 255-262.

Holm, L. and C. Sander. 1992. Evaluation of protein models by atomic solvation preference. *Journal of Molecular Biology* **225**, no. 1: 93-105.

Hooft, R.W.W., C. Sander, and G. Vriend. 1996. Positioning Hydrogen Atoms by Optimizing Hydrogen-Bond Networks in Protein Structures. *Proteins: Structure, Function, and Genetics* **26**: 363-376.

Houbrechts, A., B. Moreau, R. Abagyan, V. Mainfroid, G. Preaux, A. Lamproye, A. Poncin, E. Goormaghtigh, J.-M. Ruysschaert, J.A. Martial, and K. Goraj. 1995. Second-generation octarellins: two new *de novo* $(\beta/\alpha)_8$ polypeptides designed for investigating the influence of β-residue packing on the α/β-barrel structure stability. *Protein Engineering* **8**: 249-259.

Hough, E., L. K. Hansen, B. Birknes, K. Jynge, S. Hansen, A. Hordvik, C. Little, E. Dodson, and Z. Derewenda. 1989. High-Resolution (1.5 Å) Crystal Structure of Phospholipase C from *Bacillus Cereus*. *Nature* **338**: 357-360.

Housset, D., C. Habersetzer-Rochat, J.-P. Astier, and J.C. Fontecilla-Camps. 1994. Crystal Structure of Toxin II from the Scorpion *Androctonus australis* Hector Refined at 1.3 Å Resolution. *Journal of Molecular Biology* **239**: 88-103.

Hovde, Carolyn J., Stephen B. Calderwood, John J. Mekalanos, and R. John Collier. 1988. Evidence that glutamic acid 167 is an active-site residue of Shiga-like toxin I. *Proceedings of the National Academy of Sciences of the United States of America* **85**: 2568-2572.

Huang, Guewha Ssteven and Terrence G. Oas. 1995. Structure and stability of monomeric λ repressor: NMR evidence for two-state folding. *Biochemistry* **34**: 3884-3892.

Huang, Q., S. Liu, and Y. Tang. 1993. Refined 1.6 Å Resolution Crystal Structure of the Complex Formed between Porcine β-Trypsin and MCTI-A, a Trypsin Inhibitor of the Squash Family. *Journal of Molecular Biology* **229**: 1022-1036.

Huang, Qichen, Shengping Liu, Youqi Tang, Fuyue Zeng, and Ruiqing Qian. 1992. Amino acid sequencing of a trypsin inhibitor by refined 1.6 Å X-ray crystal structure of its complex with porcine beta-trypsin. *FEBS Letters* **297**: 143-146.

Hubbard, Simon J., Karl-Heinz Gross, and Patrick Argos. 1994. Intramolecular cavities in globular proteins. *Protein Engineering* **7**, no. 5: 613-626.

Hurley, James H., Walter A. Baase, and Brian W. Matthews. 1992. Design and Structural Analysis of Alternative Hydrophobic Core Packing Arrangements in Bacteriophage T4 Lysozyme. *Journal of Molecular Biology* **224**: 1143-1159.

Iijima, H., J.B.Jr. Dunbar, and G.R. Marshall. 1987. Calibration of Effective van der Waals Atomic Contact Radii for Proteins and Peptides. *Proteins* **2**: 330-339.

Isaacs, Neil W. 1995. Cystine knots. *Current Opinion in Structural Biology* **5**: 391-395.

Itoh, S., M.T. DeCenzo, D.J. Livingston, D.A. Pearlman, and M.A. Navia. 1995. Conformation of FK506 in X-ray structures of its complexes with human recombinant FKBP12 mutants. *Bioorganic & Medicinal Chemistry Letters* **5**: 1983-1988.

James, M.N.G. and A.R. Sielecki. 1983. Structure and Refinement of Penicillopepsin at 1.8 Å Resolution. *Journal of Molecular Biology* **163**: 299-361.

Janin, Joël and Cyrus Chothia. 1990. The Structure of Protein-Protein Recognition Sites. *Journal of Biological Chemistry* **265**, no. 27: 16027-16030.

Janin, J., S. Wodak, M. Levitt, and B. Maigret. 1978. Conformation of Amino Acid Side-chains in Proteins. *Journal of Molecular Biology* **125**: 357-386.

Jeng, M.-F., A.P. Campbell, T. Begley, A. Holmgren, D.A. Case, P.E. Wright, and H.J. Dyson. 1994. High-resolution solution structures of oxidized and reduced *Escherichia coli* thioredoxin. *Structure* **2**: 853-868.

Jones, T.A., J.-Y. Zou, S.W. Cowan, and M. Kjeldgaard. 1991. Improved Methods for Building Protein Models in Electron Density Maps and the Location of Errors in these Models. *Acta Crystallographica, Section A* **47**: 110-119.

Kamtekar, S., J. M. Schiffer, H. Xiong, J. M. Babik, and M. H. Hecht. 1993. Protein Design by Binary Patterning of Polar and Nonpolar Amino Acids. *Science* **262**: 1680-1685.

Karle, I.L., D. Ranganathan, and V. Haridas. 1996. A persistent preference for layer motifs in self-assemblies of squarates and hydrogen squarates by hydrogen bonding [X-H···O; X = N, O, or C]: a crystallographic study of five organic salts. *Journal of the American Chemical Society* **118**: 7128-7133.

Karpusas, Mihail, Walter A. Baase, Masazumi Matsumura, and Brian W. Matthews. 1989. Hydrophobic packing in T4 lysozyme probed by cavity-filling mutants. *Proceedings of the National Academy of Sciences of the United States of America* **86**: 8237-8241.

Katti, S.K., D.M. LeMaster, and H. Eklund. 1990. Crystal structure of thioredoxin from *Escherichia coli* at 1.68 Å resolution. *Journal of Molecular Biology* **212**: 167-184.

Kim, Youngsoo, Debra Mlsna, Arthur F. Monzingo, Michael P. Ready, Art Frankel, and Jon D. Robertus. 1992. Structure of a Ricin Mutant Showing Rescue of Activity by a Noncatalytic Residue. *Biochemistry* **31**, no. 12: 3294-3296.

Kleywegt, Gerard J. and T. Alwyn Jones. 1994. Detection, Delineation, Measurement and Display of Cavities in Macromolecular Structures. *Acta Crystallographica, Section D* **50**: 178-185.

Kojima, Shuichi, Katsunori Miyoshi, and Kin-ichiro Miura. 1996. Synthesis of a squash-type protease inhibitor by gene engineering and effects of replacements of conserved hydrophobic amino acid residues on its inhibitory activity. *Protein Engineering* **9**: 1241-1246.

Kossiakoff, Anthony A. and S. Shteyn. 1984. Effect of protein packing structure on side-chain methyl rotor conformations. *Nature* **311**, no. 5986: 582-583.

Kossiakoff, Anthony A., Mark Ultsch, Stephen White, and Charles Eigenbrot. 1991. Neutron Structure of Subtilisin BPN': Effects of chemical environment on hydrogen-bonding geometries and the pattern of hydrogen-deuterium exchange in secondary structure elements. *Biochemistry* **30**: 1211-1221.

Kraulis, P.J., G.M. Clore, M. Nilges, T.A. Jones, G. Pettersson, J. Knowles, and A.M. Gronenborn. 1989. Determination of the Three-dimensional Structure of the C-Terminal Domain of Cellobiohydrolase I from *Trichoderma reesei*. A study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing. *Biochemistry* **28**: 7241-7257.

Krezel, A.M., C. Kasibhatla, P. Hidalgo, R. MacKinnon, and G. Wagner. 1995. Solution structure of the potassium channel inhibitor agitoxin 2: Caliper for probing channel geometry. *Protein Science* **4**: 1478-1489.

Kumar, V.D., R.W. Harrison, L.C. Andrews, and I.T. Weber. 1992. Crystal Structure at 1.5 Å Resolution of d(CGCICICG), an Octanucleotide Containing Inosine, and Its Comparison with d(CGCG) and d(CGCGCG) Structures. *Biochemistry* **31**: 1541-1550.

Kunkel, Thomas A., John D. Roberts, and Richard A. Zakour. 1987. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Methods In Enzymology* **154**: 367-382.

Kuszewski, J., A.M. Gronenborn, and G.M. Clore. 1996. Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Science* **5**: 1067-1080.

Ladbury, J.E., R. Wynn, H.W. Hellinga, and J.M. Sturtevant. 1993. Stability of oxidized *E. coli* thioredoxin and its dependence on protonation of aspartic acid residue in the 26 position. *Biochemistry* **32**: 7526-7530.

Langan, Paul and Benno P. Schoenborn. 1999. Letter: Need for neutron diffraction instruments. *Science*, 5 November, 1089.

Langridge, Robert, Thomas E. Ferrin, Irwin D. Kuntz, and Michael L. Connolly. 1981. Real-time color graphics in studies of molecular interactions. *Science* **211**, no. 4483: 661-666.

Laskowski, Roman A. 1995. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics* **13**, no. 5: 323-330.

Laskowski, Roman A., M.W. Macarthur, D.S. Moss, and Janet M. Thornton. 1993. ProCheck –A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* **26**: 283-291.

Laughlan, G., A.I.H. Murchie, D.G. Norman, M.H. Moore, P.C.E. Moody, D.M.J. Lilley, and B. Luisi. 1994. The High-Resolution Crystal Structure of a Parallel-Stranded Guanine Tetraplex. *Science* **265**: 520-524.

Lee, B.K. and F.M. Richards. 1971. The Interpretation of Protein Structures: Estimation of Static Accessibility. *Journal of Molecular Biology* **55**: 379-400.

Lee, Christopher and S. Subbiah. 1991. Prediction of Protein Side-chain Conformation by Packing Optimization. *Journal of Molecular Biology* **217**: 373-388.

Liang, J., H. Edelsbrunner, and C. Woodward. 1998. Anatomy of protein pockets and cavities: measurement of binding site geometry and implicaitons for ligand design. *Protein Science* **7**, no. 9: 1884-1897.

Lim, W.A. and R.T. Sauer. 1989. Alternative packing arrangements in the hydrophobic core of λ repressor. *Nature* **339**: 31-36.

Lin, Shuo Liang and Ruth Nussinov. 1995. A disulphide-reinforced structural scaffold shared by small proteins with diverse functions. *Nature Structural Biology* **2**: 835-337.

Lipscomb, Leigh Ann, Nadine C. Gassner, Sheila D. Snow, Aimee M. Eldridge, Walter A. Baase, Devin L. Drew, and Brian W. Matthews. 1998. Context-dependent protein stabilization by methionine-to-leucine substitution shown in T4 lysozyme. *Protein Science* **7**: 765-773.

Lovell, Simon C., J. Michael Word, Jane S. Richardson, and David C. Richardson. 1999. Asparagine and Glutamine Rotamers: *B*-Factor Cutoff and Correction of Amide Flips Yield Distinct Clustering. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 400-405.

Lovell, Simon C., J. Michael Word, Jane S. Richardson, and David C. Richardson. 2000. The penultimate rotamer library. *Proteins: Structure, Function, and Genetics* (submitted).

MacArthur, Malcolm W. and Janet M. Thornton. 1999. Protein side-chain conformation: a systematic variation of χ_1 mean values with resolution - a consequence of multiple rotameric states? *Acta Crystallographica* D **55**: 994-1004.

MacKenzie, Kevin R., James H. Prestegard, and Donald M. Engelman. 1997. A Transmembrane Helix Dimer: Structure and Implications. *Science* **276**: 131-133.

Markley, John L., Ad Bax, Yoji Arata, C. W. Hilbers, Robert Kaptein, Brian D. Sykes, Peter E. Wright, and Kurt Wüthrich. 1998. Recommendations for the Presentation of NMR Structures of Proteins and Nucleic Acids. *Journal of Molecular Biology* **280**: 933-952.

Matsumura, Masazumi and Brian W. Matthews. 1989. Control of Enzyme Activity by an Engineered Disulfide Bond. *Science* **243**: 792-794.

Matsumura, M., J.A. Wozniak, D.-p. Sun, and B.W. Matthews. 1989. Structural studies of mutants of T4 lysozyme that alter hydrophobic stabilization. *Journal of Biological Chemistry* **264**: 16059-16066.

Matsuura, Y., T. Takano, and R.E. Dickerson. 1982. Structure of cytochrome c_{551} from *Pseudomonas aeruginosa* refined at 1.6 Å resolution and comparison of the two redox forms . *Journal of Molecular Biology* **156**: 389-409.

Matthews, B. W. 1995. Studies on protein stability with T4 lysozyme. *Advances in Protein Chemistry* **46**: 249-278.

Max, Nelson L. 1979. ATOMLLL: atoms with shading and highlights. *Computer Graphics* **13**: 165-173.

McDonald, I.K. and J.M. Thornton. 1994. The application of hydrogen bonding analysis in X-ray crystallography to help orientate asparagine, glutamine and histidine side chains. *Protein Engineering* **8**: 217-224.

McDowell, Robert S. and Anthony A. Kossiakoff. 1995. A comparison of neutron diffraction and molecular dynamics structures: Hydroxyl group and water molecule orientations in trypsin. *Journal of Molecular Biology* **250**: 533-570.

McGregor, M.J., S.A. Islam, and M.J.E. Sternberg. 1987. Analysis of the Relationship Between Side-chain Conformation and Secondary Structure in Globular Proteins. *Journal of Molecular Biology* **198**: 295-310.

McRee, D.E. 1993. *Practical Protein Crystallography*. San Diego: Academic Press.

McRee, Duncan E. 1999. XtalView/Xfit - A Versatile Program for Manipulating Atomic Coordinates and Electron Density. *Journal of Structural Biology* **125**, no. 2-3: 156-165.

Merritt, Ethan A. and David J. Bacon. 1997. Raster3D: Photorealistic Molecular Graphics. *Methods In Enzymology* **277**: 505-524.

Mlsna, Debra, Arthur F. Monzingo, Betsy J. Katzin, Stephen Ernst, and Jon D. Robertus. 1993. Structure of recombinant ricin A chain at 2.3 Å. *Protein Science* **2**: 429-435.

Momany, F.A., R.F. McGuire, A.W. Burgess, and H.A. Scheraga. 1975. Energy Parameters in Polypeptides. VII. Geometric Parameters, Partial Atomic Changes, Nonbonded Interactions, Hydrogen Bond Interactions, and Intrinsic Torsional Potentials for the Naturally Occurring Amino Acids. *Journal of Physical Chemistry* **79**: 2361.

Montelione, Gaetano T. and Stephen Anderson. 1999. Structural genomics: keystone for a Human Proteome Project. *Nature Structural Biology* **6**, no. 1: 11-12.

Moy, F.J., Y.-C. Li, P. Rauenbuehler, M.E. Winkler, H.A. Scheraga, and G.T. Montelione. 1993. Solution Structure of Human Type-α Transforming Growth Factor

Determined by Heteronuclear NMR Spectroscopy and Refined by Energy Minimization with Restraints. *Biochemistry* **32**: 7334-7353.

Munson, M., R. O'Brien, J.M. Sturtevant, and L. Regan. 1994. Redesigning the hydrophobic core of a four-helix-bundle protein. *Protein Science* **3**: 2015-2022.

Mutter, M., G.G. Tuchscherer, C. Miller, K.-H. Altmann, R.I. Carey, D.F. Wyss, A.M. Labhardt, and J.E. Rivier. 1992. Template-Assembled Synthetic Proteins with Four-Helix-Bundle Topology. Total Chemical Synthesis and Conformational Studies. *Journal of the American Chemical Society* **114**: 1463-1470.

Myers, Jeffrey K. and Terrence G. Oas. 1999. Contribution of a Buried Hydrogen Bond to λ Repressor Folding Kinetics. *Biochemistry* **38**, no. 21: 6761-6768.

Narasimhan, Lakshmi, Juswinder Singh, Christine Humblet, Kunchur Guruprasad, and Tom Blundell. 1994. Snail and spider toxins share a similar tertiary structure and 'cystine motif'. *Nature Structural Biology* **1**: 850-852.

Némethy, G., K.D. Gibson, K.A. Palmer, C.N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H.A. Scheraga. 1992. Energy Parameters in Polypeptides. 10. Improved Geometrical Parameters and Nonbonded Interactions for Use in the ECEPP/3 Algorithm, with Application to Proline-Containing Peptides. *Journal of Physical Chemistry* **96**: 6472-6484.

Nicholls, Anthony, Ranganathan Bharadwaj, and Barry Honig. 1993. GRASP–Graphical Representation and Analysis of Surface-Properties. *Biophysical Society Journal, 37th Annual Meeting Abstracts* **64**: A166.

Nielsen, J.E., K.V. Andersen, B. Honig, R.W.W. Hooft, G. Klebe, G. Vriend, and R.C. Wade. 1999. Improving macromolecular electrostatics calculations. *Protein Engineering* **12**, no. 8: 657-662.

NIGMS. 1999. Structural Genomics Targets Workshop:14. Bethesda, MD: NIH.

Ooi, T. and K. Nishikawa. 1973. Section describing contact matrices. In *Conformation of Biological Molecules and Polymers*, ed. E. Bergmann and B. Pullmann:173-187. New York: Academic Press.

Parkin, S., B. Rupp, and H. Hope. 1996. Atomic Resolution Structure of Concanavalin A at 120K. *Acta Crystallographica, Section D* **52**: 1161-1168.

Pattabiraman, N., K. B. Ward, and P. J. Fleming. 1995. Occluded Molecular Surface: Analysis of Protein Packing. *Journal of Molecular Recognition* **8**: 334-344.

Pellequer, Jean-Luc, Shu-wen W. Chen, Victoria A. Roberts, John A. Tainer, and Elizabeth D. Getzoff. 1999. Unraveling the effect of changes in conformation and

compactness at the antibody V_L-V_H interface upon antigen binding. *Journal of Molecular Recognition* **12**: 267-275.

Perrakis, Anastassis, Richard Morris, and Victor S. Lamzin. 1999. Automated protein model building combined with iterative structure refinement. *Nature Structural Biology* **6**, no. 5: 458-463.

Petrella, Robert J., Themis Lazaridis, and Martin Karplus. 1998. Protein sidechain conformer prediction: a test of the energy function. *Folding & Design* **3**, no. 5: 353-377.

Ponder, J.W. and F.M. Richards. 1987. Tertiary Templates for Proteins: Use of Packing Criteria in the Enumeration of Allowed Sequences for Different Structural Classes. *Journal of Molecular Biology* **193**: 775-791.

Pontius, Joan, Jean Richelle, and Shoshana J. Wodak. 1996. Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures. *Journal of Molecular Biology* **264**, no. 1: 121-136.

Porter, Thomas K. 1978. Spherical shading. *Computer Graphics* **12**, no. 3: 282-285.

Press, William H., Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. 1988. *Numerical Recipes in C*. Cambridge: Cambridge University Press.

Presta, Leonard G. and George D. Rose. 1988. Helix Signals in Proteins. *Science* **240**: 1632-1641.

Price-carter, M., W. R. Gray, and D. P. Goldenberg. 1996a. Folding of ω-conotoxins. I. Efficient disulfide coupled folding of mature sequences *in vitro*. *Biochemistry* **35**, no. 48: 15537-15546.

Price-carter, M., W. R. Gray, and D. P. Goldenberg. 1996b. Folding of ω-conotoxins. II. Influence of precursor sequences and protein disulfide isomerase. *Biochemistry* **35**, no. 48: 15547-15557.

Prive, G.G., K. Yanagi, and R.E. Dickerson. 1991. Structure of the *B* -DNA Decamer C-C-A-A-C-G-T-T-G-G and Comparison with Isomorphous Decamers C-C-A-A-G-A-T-T-G-G and C-C-A-G-G-C-C-T-G-G. *Journal of Molecular Biology* **217**: 177-199.

Quinn, T.P., N.B. Tweedy, R.W. Williams, J.S. Richardson, and D.C. Richardson. 1994. BetaDoublet: *De Novo* Design, Synthesis and Characterization of a Novel β Sandwich Protein. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 8747-8751.

Ramachandran, G. N., C. Ramakrishnan, and V. Sasisekharan. 1963. Stereochemistry of Polypeptide Chain Configurations. *Journal of Molecular Biology* **7**: 95-99.

Ramasubbu, N., V. Paloth, Y. Luo, G.D. Brayer, and M.J. Levine. 1996. Structure of human salivary α-amylase at 1.6 Å resolution: Implications for its role in the oral cavity. *Acta Crystallographica, Section D* **52**: 435-446.

Richards, Frederic M. 1974. The interpretation of protein structures: total volume, group volume distributions and packing Density. *Journal of Molecular Biology* **82**: 1-14.

Richards, Frederic M. 1977. Area, volumes, packing, and protein structure. *Annual Review of Biophysics and Bioengineering* **6**: 151-176.

Richards, Frederic M. and Windel A. Lim. 1993. An analysis of packing in the protein folding problem. *Quarterly Reviews of Biophysics* **26**: 423-498.

Richardson, David C., F. Colonna, E. van Cutsem, and K. Mortensen. 1987a. Protein design exercises: Betalphacin. *EMBL BIOcomputing Technical Document* **1**: 51-70,110, 117, 121.

Richardson, D.C. and J.S. Richardson. 1990. Protein Origami. In *Protein Folding: Deciphering the Second Half of the Genetic Code*, ed. L. Geirasch and J. King:5-17 and-327-333. Washington, D.C.: American Association for the Advancement of Science.

Richardson, D.C. and J.S. Richardson. 1992. The Kinemage: A Tool for Scientific Illustration. *Protein Science* **1**: 3-9.

Richardson, D.C. and J.S. Richardson. 1994. Kinemages - Simple Macromolecular Graphics for Interactive Teaching and Publication. *Trends in Biochemical Science* **19**: 135-138.

Richardson, Jane S. 1981. The Anatomy and Taxonomy of Protein Structure. In *Advances in Protein Chemistry*, ed. C.B. Anfinsen, J.T. Edsall, and F.M. Richards:167-339. New York: Academic Press.

Richardson, J.S. 1992. The Protein Tourist #4: A survey of proteins in the "small irregular" category of tertiary structure. *Protein Science*. 1. (Kinemages and text on computer disk and website).

Richardson, Jane S., Pat Argos, D. Kneller, D. Osguthorpe, and M. Scharf. 1987b. Protein design exercises: Babarellin. *EMBL BIOcomputing Technical Document* **1**: 21-39, 107, 115, 120.

Richardson, Jane S. and David C. Richardson. 1987. Some Design Principles: Betabellin. In *Protein Engineering*, ed. Dale L. Oxender and C Fred Fox:149-163 and 340-1. New York: Alan R. Liss, Inc.

Richardson, J.S. and D.C. Richardson. 1988a. Amino Acid Preferences for Specific Locations at the Ends of α Helices. *Science* **240**: 1648-1652.

Richardson, J.S. and D.C. Richardson. 1988b. Helix Lap-Joints as Ion-Binding Sites: DNA-Binding Helix Pairs and Ca-Binding "E-F Hands" are Related by Charge and Sequence Reversal. *Proteins* **4**: 229-239.

Richardson, J.S. and D.C. Richardson. 1995. Model Coordinates for SScorin (designed protein). *Brookhaven Protein Data Bank* 1SSR.

Richardson, J.S. and D.C. Richardson. 1999. Kinemage Supplement (of Over 100 Kinemages) to 2nd Edition of Introduction to Protein Structure, by C.-I. Branden and J. Tooze. In *Introduction to Protein Structure*. London: Garland Press.

Richardson, Jane S. and David C. Richardson. in press. MAGE, PROBE, and Kinemages. In *International Tables for Crystallography*, ed. Michael Rossmann and Eddy Arnold, F:25.2.8. Boston: Kluwer Academic Publishers.

Richardson, Jane S., David C. Richardson, Neil B. Tweedy, Kim M. Gernert, Thomas P. Quinn, Michael H. Hecht, Bruce W. Erickson, Yibing Yan, Robert D. McClain, Mary E. Donlan, and Mark C. Surles. 1992. Looking at proteins: representations, folding, packing, and design. *Biophysical Journal* **63**: 1186-1209.

Rojas, N.R.L., S. Kamtekar, C.T. Simons, J.E. McLean, K.M. Vogel, T.G. Spiro, R.S. Farid, and M.H. Hecht. 1997. *De novo* heme proteins from designed combinatorial libraries. *Protein Science* **6**: 2512-2524.

Ruddon, Raymond W., Simon A. Sherman, and Elliott Bedows. 1996. Protein folding in the endoplasmic reticulum: Lessons form the human chorionic gonadotropin β subunit. *Protein Science* **5**: 1443-1452.

Sali, Andrej. 1998. 100,000 protein structures for the biologist. *Nature Structural Biology* **5**, no. 12: 1029-1032.

Salisbury, S.A., S.E. Wilson, H.R. Powell, O. Kennard, P. Lubini, G.M. Sheldrick, N. Escaja, E. Alazzouzi, A. Grandas, and E. Pedroso. 1997. The bi-loop, a new general four-stranded DNA motif. *Proceedings of the National Academy of Sciences of the United States of America* **94**: 5515-5518.

Sawaya, M.R. and J. Kraut. 1997. Loop and subdomain movements in the mechanism of *Escherichia coli* dihydrofolate reductase: crystallographic evidence. *Biochemistry* **36**: 11205-11215.

Schlossman, David, David Withers, Philip Welsh, Audrey Alexander, Jon Robertus, and Arthur Frankel. 1989. Role of Glutamic Acid 177 of the Ricin Toxin A Chain in Enzymatic Inactivation of Ribosomes. *Molecular and Cellular Biology* **9**, no. 11: 5012-5021.

Schrauber, H., F. Eisenhaber, and P. Argos. 1993. Rotamers: To be or not to be? An Analysis of Amino Acid Side-chain Conformations in Globular Proteins. *Journal of Molecular Biology* **230**: 592-612.

Scott, W.G., J.B. Murray, J.R.P. Arnold, B.L. Stoddard, and A. Klug. 1996. Capturing the Structure of a Catalytic RNA Intermediate: The Hammerhead Ribozyme. *Science* **274**: 2065-2069.

Sedgewick, Robert. 1990. *Algorithms in C*. Edited by Michael A Harrison. The Addison-Wesley Series in Computer Science. New York: Addison-Wesley Publishing Company, Inc.

Serre, L., E.C. Verbree, Z. Dauter, A.R. Stuitje, and Z.S. Derewenda. 1995. The *Escherichia coli* malonyl-CoA:acyl carrier protein transacylase at 1.5 Å resolution. *Journal of Biological Chemistry* **270**: 12961-12964.

Sevcik, J., Z. Dauter, V.S. Lamzin, and K.S. Wilson. 1996. Ribonuclease from *Streptomyces aureofaciens* at atomic resolution. *Acta Crystallographica, Section D* **52**: 327-344.

Shakhnovich, E.I. and A.V. Finkelstein. 1989. Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a 1st-order phase transition. *Biopolymers* **28**: 1667-1680.

Sheldrick, G.M. and T.R. Schneider. 1997. SHELX: high resolution refinement. *Methods in Enzymology* **277**: 319-343.

Shinde, Ujwal and Masayori Inouye. 1993. Intramolecular chaperones and protein folding. *Trends in Biochemical Science* **18**: 442-446.

Shortle, D., W.E. Stites, and A.K. Meeker. 1990. Contributions of the Large Hydrophobic Amino Acids to the Stability of Staphylococcal Nuclease. *Biochemistry* **29**: 8033-8041.

Shrake, A. and J.A. Rupley. 1973. Environment and Exposure to Solvent of Protein Atoms. Lysozyme and Insulin. *Journal of Molecular Biology* **79**: 351-371.

Singh, R. K., Alexander Tropsha, and Iosif I. Vaisman. 1996. Delaunay Tessellation of Proteins: Four Body Nearest Neighbor Propensities of Amino Acid Residues. *Journal of Computational Biology* **3**: 213-222.

Smith, D.D.S., K.A. Pratt, I.G. Sumner, and C.M. Henneke. 1995. Greek key jellyroll protein motif design: expression and characterization of a first-generation molecule. *Protein Engineering* **8**: 13-20.

Spurlino, J.C., A.M. Smallwood, D.D. Carlton, T.M. Banks, K.J. Vavra, J.S. Johnson, E.R. Cook, J. Falvo, R.C. Wahl, T.A. Pulvino, J.J. Wendoloski, and D.L. Smith. 1994.

1.56 Å structure of mature truncated human fibroblast collagenase. *Proteins: Structure, Function, and Genetics* **19**: 98-109.

Stout, G.H. and L.H. Jensen. 1968. *X-ray Structure Determination: a Practical Guide*. London: The MacMillan Company, Collier-MacMillan Ltd.

Streitwieser, Jr., Andrew and Clayton C. Heathcock. 1976. *Introduction to Organic Cemistry*. New York: Macmillan Publishing Co., Inc.

Struthers, M. D., R. P. Cheng, and Barbara Imperiali. 1996. Design of a Monomeric 23-Residue Polypeptide with Defined Tertiary Structure. *Science* **271**: 342-345.

Sun, Shaojian, Rachel Brem, Hue Sun Chan, and Ken A. Dill. 1995. Designing amino acid sequences to fold with good hydrophobic cores. *Protein Engineering* **8**, no. 12: 1205-1213.

Surles, M.C., J.S. Richardson, D.C. Richardson, and F.P. Brooks, Jr. 1994. Sculpting proteins interactively: Continual energy minimization embedded in a graphical modeling system. *Protein Science* **3**: 198-210.

Tame, J.R.H., E.J. Dodson, G. Murshudov, C.F. Higgins, and A.J. Wilkinson. 1995. The crystal structures of the oligopeptide-binding protein OppA complexed with tripeptide and tetrapeptide ligands. *Structure* **3**: 1395-1406.

Tan, S., Y. Hunziker, D.F. Sargent, and T.J. Richmond. 1996. Crystal structure of a yeast TFIIA/TBP/DNA complex. *Nature* **381**: 127-134.

Thayer, M.M., K.M. Flaherty, and D.B. McKay. 1991. Three-Dimensional Structure of the Elastase of *Pseudomonas aeruginosa* at 1.5 Å Resolution. *The Journal of Biological Chemistry* **266**: 2864-2871.

Tuffery, P., C. Etchebest, and S. Hazout. 1997. Prediction of protein side chain conformations: a study on the influence of backbone accuracy on conformation stability in the rotamer space. *Protein Engineering* **10**: 361-372.

Tuffery, P., C. Etchebest, S. Hazout, and R. Lavery. 1991. A New Approach to the Rapid Determination of Protein Side Chain Conformations. *Journal of Biomolecular Structure & Dynamics* **8**: 1267-1289.

Unson, Cecilia G., Bruce W. Erickson, David C. Richardson, and Jane S. Richardson. 1984. Protein engineering: design and synthesis of a protein. In *Federation Proceedings*, ed. Karl F. Heumann, 43:1837. St. Louis Missouri: Federation of American Societies for Experimental Bioloby.

Vassylyev, D.G., K. Katayanagi, K. Ishikawa, M. Tsujimoto-Hirano, M. Danno, A. Pahler, O. Matsumoto, M. Matsushima, H. Yoshida, and K. Morikawa. 1993. Crystal

structures of ribonuclease F1 of *Fusarium moniliforme* in its free form and in complex with 2' GMP. *Journal of Molecular Biology* **230**: 979-996.

Vetter, Ingrid R., Walter A. Baase, Dirk W. Heinz, Jian-Ping Xiong, Sheila Snow, and Brian W. Matthews. 1996. Protein structural plasticity exemplified by insertion and deletion mutants in T4 lysozyme. *Protein Science* **5**: 2399-2415.

Vita, Claudio, Christian Roumestand, Flavio Toma, and André Ménez. 1995. Scorpion toxins as natural scaffolds for protein engineering. *Proceedings of the National Academy of Sciences of the United States of America* **92**: 6404-6408.

Vlassi, M., C. Steif, P. Weber, D. Tsernoglou, K.S. Wilson, H.-J. Hinz, and M. Kokkinidis. 1994. Restored heptad pattern continuity does not alter the folding of a four-α-helix bundle. *Nature Structural Biology* **1**: 706-716.

Vriend, Gert. 1990. WHAT IF: A molecular modeling and drug design program. *Journal of Molecular Graphics* **8**, no. 1: 52-56.

Wahl, M.C., S.T. Rao, and M. Sundaralingam. 1996. The structure of r(UUCGCG) has a 5'-UU-overhang exhibiting Hoogsteen-like *trans* U·U base pairs. *Nature Structural Biology* **3**: 24-31.

Weaver, L. H. and B. W. Matthews. 1987. Structure of Bacteriophage T4 Lysozyme Refined at 1.7 Ångstroms Resolution. *Journal of Molecular Biology* **193**: 189-199.

Weston, Simon A., Alec D. Tucker, David R. Thatcher, Dean J. Derbyshire, and Richard A. Pauptit. 1994. X-ray Structure of Recombinant Ricin A-Chain at 1.8 Å Resolution. *Journal of Molecular Biology* **244**: 410-422.

Wiberg, Kenneth B. and Mark A. Murcko. 1988. Rotational Barriers: 2. Energies of Alkane Rotamers. An Examination of Gauche Interactions. *Journal of the American Chemical Society* **110**: 8029-8038.

Wickham, Gene S. and J. Michael Word. 1999. Evaluation of the resultant van der Waals interactions after modeling the cleavage site phosphate into the crystal structure of the HDV genomic ribozyme. *Nucleic Acids Symposium Series*: in press.

Wlodawer, Alexander, L. Anders Svensson, Lennart Sjölin, and Gary L. Gilliland. 1988. Structure of Phosphate-Free Ribonuclease A Refined at 1.26 Å. *Biochemistry* **27**: 2705-2717.

Wolfram Research, Inc. 1996. *Mathematica*. Champaign, IL: Wolfram Research, Inc.

Wolynes, Peter G., Jose N. Onuchic, and D. Thirumalai. 1995. Navigating the folding routes. *Science* **267**, no. 5204: 1619-1620.

Word, J. Michael, Simon C. Lovell, Thomas H. LaBean, Hope C. Taylor, Michael E. Zalis, Brent K. Presley, Jane S. Richardson, and David C. Richardson. 1999a. Visualizing and Quantifying Molecular Goodness-of-Fit: Small-Probe Contact Dots with Explicit Hydrogens. *Journal of Molecular Biology* **285**: 1709-1731.

Word, J. Michael, Simon C. Lovell, Jane S. Richardson, and David C. Richardson. 1999b. Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-chain Amide Orientation. *Journal of Molecular Biology* **285**: 1735-1747.

Wu, X., B. Knudsen, S.M. Feller, J. Zheng, A. Sali, D. Cowburn, H. Hanafusa, and J. Kuriyan. 1995. Structural basis for the specific interaction of lysine-containing proline-rich peptides with the N-terminal SH3 domain of c-Crk. *Structure* **3**: 215-226.

Yamano, A. and M.M. Teeter. 1994. Correlated Disorder of the Pure Pro^{22}/Leu^{25} Form of Crambin at 150K Refined to 1.05 Å Resolution. *Journal of Biological Chemistry* **269**: 13956-13965.

Yansura, D. G. 1990. Expression as *trpE* fusion. *Methods In Enzymology* **185**: 161-166.

Zarembinski, T. I., L. W. Hung, H. J. Mueller-Dieckmann, K. K. Kim, H. Yokota, R. Kim, and S. H. Kim. 1998. Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 15189-15193.

Zehfus, Micheal H. and George D. Rose. 1986. Compact Units in Proteins. *Biochemistry* **25**: 5759-5765.

Zeng, Jun, Masha Fridman, Hiroshi Maruta, Herbert R. Treutlein, and Thomas Simonson. 1999. Protein-protein recognition: An experimental and computational study of the R89K mutation in Raf and its effect on Ras binding. *Protein Science* **8**, no. 1: 50-64.

Zhang, Li and Jan Hermans. 1996. Hydrophilicity of cavities in proteins. *Proteins: Structure, Function, and Genetics* **24**: 433-438.

Zyda, Michael J. 1988. A Decomposable Algorithm for Contour Surface Display Generation. *ACM Transactions on Graphics* **7**, no. 2: 129-148.

Biography

John Michael Word was born in Scottsboro, Alabama on September 6, 1957 to Elizabeth Ann and Billy Fred Word. Being naturally curious, as a toddler Mike became something of an escape artist, often venturing far from home and occasionally requiring several stitches. At the age of 7, he set fire to the house while experimenting with his first, and only, home chemistry set. Surprisingly, his parents continued to support his interest in science. His mother was encouraging whenever experiments were performed *outside* and made sure he could attend any summer programs he was accepted to, including an NSF Summer Science Training Program at the University of Georgia and a University Research Program at Auburn.

In high school Mike did poorly in a dreadful touch typing class—his single most valuable course work to date. Later during high school, he hand ground a 6" mirror

to build a reflecting telescope, and stargazing remains a lifelong hobby. In track he set a North Hall High record in pole vaulting.

When a family crisis intervened, Mike quit high school early and got a job with a company which manufactured poultry vaccines, as a driver and egg candler. In the fall, he attended Mercer University in Macon Georgia on a full missionary scholarship (he is not making this up!) since the school did not provide an adequate academic scholarship. With the help of the College Level Entrance Placement program, he CLEPed out of a full year of credits and made the *huge mistake* of graduating *summa cum laude* in 1978 with a BS in chemistry in three years (thanks also to Drs. Hargrove, James, Furth and Marquardt). He should have stayed and taken lots of advanced math and at least one biology course. Instead he would wait almost two decades before taking biology.

Accepted into the chemistry graduate program at Berkeley, Mike quickly became acclimated to west coast living and started working with Bill Washburn, who did not get tenure and left quickly. With Andrew Streitwieser, Mike studied carbon acidity in phenalenes. At home, he developed a knack for diapering as a busy new father to his son, John. Eventually, it became clear that being a parent required income and Mike got his Masters in 1980 and got a job in Philadelphia.

To do so required agreeing that he knew how to program in FORTRAN so he lied and hurriedly taught himself in the first week at Rohm and Haas. There he was exposed to statistical computing by a great boss, John Ritts. There too, he met his

one great love, Cate Stewart, to whom he remains very attached. Eventually he went on to a company which typesets US patents and from there to SmithKline Beecham where he helped found the Science Fiction computing group with Bill Wood and Frank Tobin.

After deciding it was just too damn cold in Philadelphia and the people aren't very friendly, Mike got a job as a scientific programmer with Glaxo and he and Cate moved to Raleigh, North Carolina, which they recommend highly. After several productive years, in which he worked with the crystallography and structural NMR groups and pitched in on the solution structure of the human *c-src* SH2 domain, Glaxo generously asked Mike to be the first to enter a program permitting select employees to return to graduate school. He was accepted by Duke's biochemistry department, and soon joined the Richardson laboratory, where his strong interest in the visual arts and computer graphics made him a good match.

Mike grew up near the Tennessee river in Alabama, and the Flint river and Lake Lanier in Georgia and consequently developed a love of the water. Right now, he would love to be snorkeling in the Caribbean. He is also a skier and doesn't regret switching to cross-country several years back because it's so much warmer than alpine. He and Cate have a passion for travel which they have only partially indulged while Mike was in school. When he leaves Duke, Mike plans to return to work at GlaxoWellcome but the company has merged and completely reorganized since he went on leave, so he has no idea just what he will be up to.

List of Publications

Xu, Robert X., J. Michael Word, Donald G. Davis, M. L. Rink., D. H. Willard, Jr., and Robert T. Gampe, Jr. 1995. Solution Structure of the Human pp60$^{c\text{-}src}$ SH2 Domain Complexed with a Phosphorylated Tyrosine Pentapeptide. *Biochemistry* **34**: 2107-2121.

Ghaemmaghami, Sina, J. Michael Word, Randall E. Burton, Jane S. Richardson, and Terrence G. Oas. 1998. Folding Kinetics of a Fluorescent Variant of Monomeric λ Repressor. *Biochemistry* **37**: 9179-9185.

Word, J. Michael, Simon C. Lovell, Thomas H. LaBean, Hope C. Taylor, Michael E. Zalis, Brent K. Presley, Jane S. Richardson, and David C. Richardson. 1999. Visualizing and Quantifying Molecular Goodness-of-Fit: Small-Probe Contact Dots with Explicit Hydrogens. *Journal of Molecular Biology* **285**: 1709-1731.

Word, J. Michael, Simon C. Lovell, Jane S. Richardson, and David C. Richardson. 1999. Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-chain Amide Orientation. *Journal of Molecular Biology* **285**: 1735-1747.

Lovell, Simon C., J. Michael Word, Jane S. Richardson, and David C. Richardson. 1999. Asparagine and Glutamine Rotamers: *B*-Factor Cutoff and Correction of Amide Flips Yield Distinct Clustering. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 400-405.

Wickham, Gene S. and J. Michael Word. 1999. Evaluation of the resultant van der Waals interactions after modeling the cleavage site phosphate into the crystal structure of the HDV genomic ribozyme. *Nucleic Acids Symposium Series*: in press.

Lovell, Simon C., J. Michael Word, Jane S. Richardson, and David C. Richardson. 2000. The penultimate rotamer library. *Proteins: Structure, Function, and Genetics* submitted.