Article No. jmbi.1998.2400 available online at http://www.idealibrary.com on IDEAL J. Mol. Biol. (1999) 285, 1711–1733

JMB

Visualizing and Quantifying Molecular Goodness-of-Fit: Small-probe Contact Dots with Explicit Hydrogen Atoms

J. Michael Word, Simon C. Lovell, Thomas H. LaBean, Hope C. Taylor Michael E. Zalis, Brent K. Presley, Jane S. Richardson and David C. Richardson*

Biochemistry Department Duke University, Durham NC 27710-3711, USA

The technique of small-probe contact dot surfaces is described as a method for calculating and displaying the detailed atomic contacts inside or between molecules. It allows one both to measure and to visualize directly the goodness-of-fit of packing interactions. It requires both highly accurate structures and also the explicit inclusion of all hydrogen atoms and their van der Waals interactions.

A reference dataset of 100 protein structures was chosen on the basis of resolution (1.7 Å or better), crystallographic *R*-value, non-homology, and the absence of any unusual problems. Hydrogen atoms were added in standard geometry and, where needed, with rotational optimization of OH, SH, and NH_3^+ positions. Side-chain amide orientations were corrected where required by NH van der Waals clashes, as described in the accompanying paper. It was determined that, in general, methyl groups pack well in the default staggered conformation, except for the terminal methyl groups of methionine residues, which required rotational optimization. The distribution of serious clashes (i.e. non-H-bond overlap of $\geqslant 0.4$ Å) was studied as a function of resolution, alternate conformations, and temperature factor (*B*), leading to the decision that packing and other structural features would not be analyzed for residues in 'b' alternate conformations or with *B*-factors of 40 or above. At the level of the fine details analyzed here, structural accuracy improves quite significantly over the range from 1.7 to 1.0 Å resolution.

These high-resolution structures show impressively well-fitted packing interactions, with some regions thoroughly interdigitated and other regions somewhat sparser. Lower-resolution structures or model structures could undoubtedly be improved in accuracy by the incorporation of this additional information: for example, nucleic acid structures in non-canonical conformations are often very accurate for the bases and much less reliable for the backbone, whose conformation could be specified better by including explicit H atom geometry and contacts. The contact dots are an extremely sensitive method of finding problem areas, and often they can suggest how to make improvements. They can also provide explanations for structural features that have been described only as empirical regularities, which is illustrated by showing that the commonest rotamer of methionine (a left-handed spiral, with all χ values near $-60°$) is preferred because it provides up to five good H atom van der Waals contacts. This methodology is thus applicable in two different ways: (1) for finding and correcting errors in structure models (either experimental or theoretical); and (2) for analyzing interaction patterns in the molecules themselves.

Keywords: protein internal packing; small-probe contact dots; explicit hydrogen atoms; macromolecular structure validation; methionine rotamers

**Corresponding author*

Present addresses: T. H. LaBean, Duke University Computer Science Department, Durham, NC 27706-90129, USA; M. E. Zalis, Massachusetts General Hospital Radiology Department, Boston, MA 02114, USA.

Abbreviations used: NOE, nuclear Overhauser enhancement; NMR, nuclear magnetic resonance; PDB, Protein Data Bank.

E-mail address of the corresponding author: dcr@kinemage.biochem.duke.edu

0022-2836/99/041711–23 $30.00/0

Introduction

Remarkably ordered arrangements in the interior of 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 (Fedorov *et al*., 1992; Richardson *et al*., 1992; Mutter *et al*., 1992; Betz *et al*., 1993; Kamtekar *et al*., 1993; Fezoui *et al*., 1994; Choma *et al*., 1994; Quinn *et al*., 1994; Houbrechts *et al*., 1995; Smith *et al*., 1995; Axe *et al*., 1996) strongly implies that the ordered packing of natural proteins is important, relatively difficult, and does not happen automatically.

On the other hand, structural and functional tolerance of a substantial fraction of mutations in protein interiors (e.g. Lim & Sauer, 1989; Shortle *et al*., 1990; Hurley *et al*., 1992; Richards & Lim, 1993; Munson *et al*., 1994; Dalal *et al*., 1997) implies that side-chain packing is either not important or not difficult. Recent studies of barnase core mutants (Axe *et al*., 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 have also reached conclusions both for (e.g. Shakhnovich & Finkelstein, 1989) and against (e.g. Behe *et al*., 1991; Bromberg & 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 phage 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 (1997) 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, the packing of the final Met side-chains is excellent (Chothia & Gerstein, 1997), and the message may be that extra degrees of freedom are not the dominant issue for unique packing; 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, macromolecular 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, but despite much progress reliable relative assessments are still not possible. One of the factors contributing to this difficulty is that quantifying the steric fit of a ligand to a macromolecule is equivalent to quantifying protein internal packing quality: both versions of this problem have suffered from the same lack of a good methodology.

''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 packing controversy, for visualizing or quantifying the steric component of ligand binding, or for redesigning a local region to better promote ordered structure. The usual measure by buried surface area (Lee & Richards, 1971; Chothia, 1974) 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 H atoms are usually not considered. On the other hand, the measure of packing density using Voronoi polyhedra (Richards, 1974; 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.

Here, we describe small-probe contact dots as a method for calculating and displaying the detailed atomic contacts inside or between molecules. It allows one both to measure and to visualize

directly the goodness-of-fit of packing interactions. We had developed a simpler form of this method nearly 15 years ago and have used it in analyzing various structural details (Richardson & Richardson, 1987; 1988; 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.

In order to demonstrate that small-probe contact dots incorporate new information that has not previously been utilized, the method is applied in the accompanying paper (Word *et al*., 1999) to the problem of assigning the correct 180° flip for planar amide groups at the ends of asparagine and glutamine side-chains in protein structures. Here, the method is used to make various corrections in a reference dataset of 100 high-resolution protein structures and to illuminate the reasons behind some of the rotamer distributions observed for specific side-chain types.

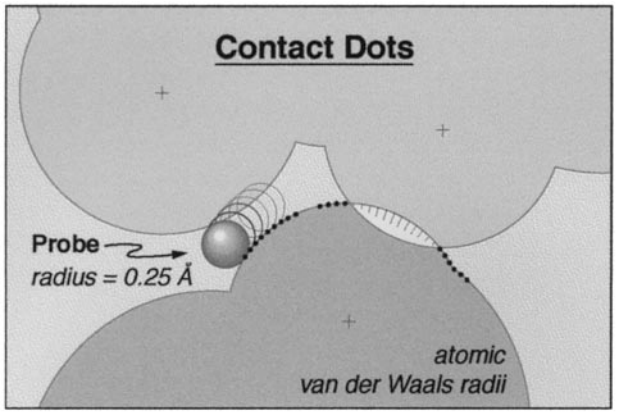

Figure 1. A diagram of the small-probe contact dot algorithm. The small, 0.25 Å radius, probe sphere rolls over the van der Waals surface of each atom, leaving a dot periodically wherever it also touches another atom that is not within three covalent bonds. Where non-H-bonding atoms overlap, the unfavorable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

Procedures

Definition of contact dots

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 Connolly (1983) algorithm 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 (written by J.M.W.), uses a very small probe (typically 0.25 Å in radius rather than 1.4 Å) and leaves a dot when the probe does touch another not-covalently bonded atom (see Figure 1), rather than when it does not touch another atom. Also, small-probe contacts form discontinuous surfaces, the patches of which directly show the location, extent, and shape of close atomic contacts (e.g. Figure 2). Every dot lies on the van der Waals surface of some atom; there are no concave, re-entrant surfaces.

Dot representations

For a visually manageable display of side-chain packing, the default is to show only side-chain-to-side-chain and side-chain-to-main-chain contacts; however main-chain-to-main-chain 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).

One color scheme for these contact dots (e.g. Figure 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 (see Figures 9 and 11) reflects the gap distance between atoms at each dot position: green or yellow for good contact (greens for narrow gaps, yellows for slight overlaps <0.2 Å), pale green dots for H-bonds, blues for wider gaps (>0.25 Å), orange or red spikes for unfavorable interpenetrations, and hot pink spikes for ''clashes'' of $\geqslant$0.4 Å. The default dot density, used for the Figures here, is 16 per Å^2.

Contact dots are output by Probe as a simple text file of dot lists (vector lists for the bump spikes) in kinemage format (Mime standard: chemical/x-kinemage; format description at ftp://kinemage.biochem.duke.edu), with color, source atom, and contact type specified. Alternative output formats are available for display as graphics objects in the crystallographic model-rebuilding programs O (Jones *et al*., 1991) and XtalView (McRee, 1993). However, the contact dots themselves are most flexible if shown in the Mage display program (written by D.C.R.; Richardson & 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

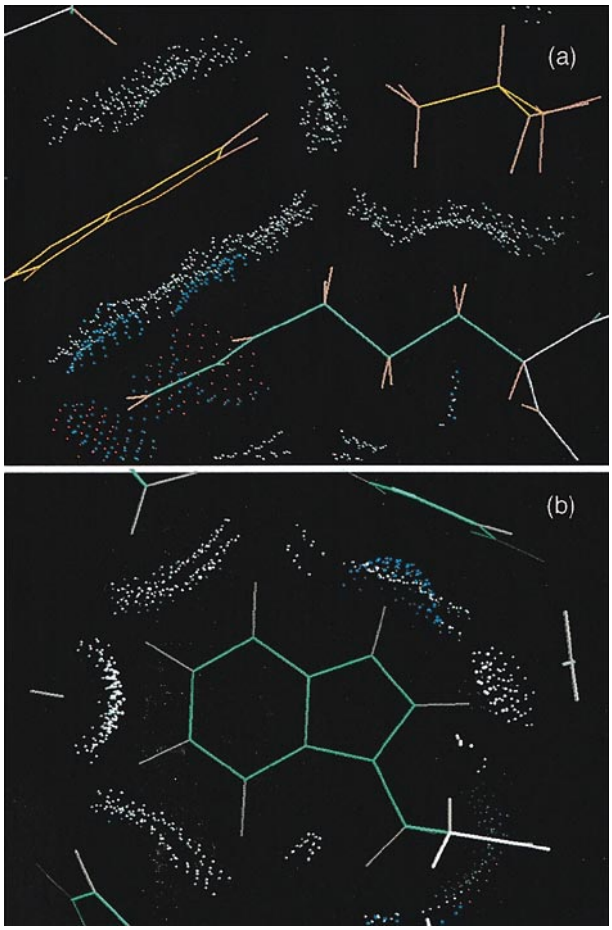

Figure 2. (a) Small-probe contact dots between residues Trp126, Arg95, and Leu91 of wild-type T4 lysozyme 3LZM (Matsumura *et al.*, 1989), colored by atom type: O, red; N, blue; C, white; and H in the color of its bonded atom. This cross-section shows the large flat surface between the Trp ring and the 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.

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 Indigo IIs or O2s used in this work. 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.

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

Adding H atoms

Small-probe contact dots require the use of explicit H atoms, as discussed in Results, including those on small-molecule ligands. The program Reduce (written by J.M.W.) adds them, with standard names, to PDB-format coordinate files (i.e. Protein Data Bank format; Bernstein *et al.*, 1977; http://pdb.pdb.bnl.gov/Format.doc/Format_Home.html) using local geometry and can perform extensive optimizations. The procedure is straightforward for non-rotatable hydrogen atoms, such as methylene or aromatic H. For our initial examination of the 100 structures, we ran Reduce with a simple set of options that added methyl hydrogen atoms in staggered conformation, but left off OH and histidine ring NH hydrogen atoms. The OH hydrogen atoms were also ignored where they were present in the original PDB files, for consistency and also since, in practice, we have found that many of them are incorrectly positioned (see Results). Other H atoms in the original files were left in the same orientation, but their bond lengths were adjusted to our standard values to allow comparison between proteins.

Subsequently, we incorporated rotational optimization of OH, SH, NH_3^+, and methionine methyl H atoms (see Results), based on scores calculated for dot contacts at 1° rotation increments. In the accompanying paper, the individual rotations are reconsidered in the context of optimizing multi-residue H-bond networks and Asn/Gln/His amide or ring flips; the resulting changes are also incorporated into the current reference coordinate sets.

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 *et al.*, 1987). We use the longer values (i.e. nucleus positions) in Reduce, since they are more consistent with the data which was 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 Reduce and Probe is given in Table 1; it includes, for instance, smaller

Table 1. Atomic parameters used in Reduce and Probe

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

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.

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 calculating correct H-bond energies (Hagler *et al*., 1974; Hermans *et al*., 1984), which is of course their dominant mode of interaction. However, van der Waals clashes are indeed essential to analyzing polar-to-non-polar H atom interactions, and also for understanding why groups cannot adopt specific alternative conformations: the counterfactual ''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 the accompanying paper (Word *et al*., 1999), 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 cannot 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.

Reduce handles nucleic acids as well as proteins. It can add hydrogen atoms to those heterogen molecules included in the Protein Data Bank connectivity database (file ftp://pdb.pdb.bnl.gov/pub/resources/hetgroups/het_dictionary.txt) or a similar file if constructed by the user. Here, we modified the standard PDB het dictionary with six additional entries and deprotonated phosphate groups.

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. Here, where water molecules are used they are treated at the level of method 2.

Scoring

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

$$Vol\,(\text{Overlap}) = \sum_{\text{overlap}} l_{sp};$$

$$Vol\,(\text{Hbond}) = \sum_{\text{Hb}} l_{sp}$$

In the extreme, if the gap between dots on the H and 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(b)).

On the other hand, despite indications that certain CH groups can act as H-bond donors (Derewenda *et al*., 1995; Karle *et al*., 1996), the contact dots have not shown unequivocal evidence of such CH$\cdots$O H-bonds, except for H^{δ_2} and H^{ε_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 those typically used in studies of CH$\cdots$O H-bonds (e.g. 1.4 *versus* 1.5 Å for O) The overlaps we see for non-His CH groups are not reproducible or large enough to

Table 2. Very high-resolution, non-redundant protein structures ($n = 100$)

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
1IGD*	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
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
1ISUa	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
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 2—*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 1ctj)	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/8Ca	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
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

Taken from the Protein Data Bank (Bernstein *et al*., 1977) as of January 13, 1997; see Procedures 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; αEF, helical E-F hand; 4hx, four-helix bundle; αHb, globin fold; α hp, multi helix-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 $\geqslant 0.4$ Å per 1000 atoms), after adjustments described in the text; Sc is overall score (contact + Hbond − clash) for the structure.

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, 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, are surfaces rather than 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:

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

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 circle of contact dots surrounding the overlap keeps the overall score favorable, but the circle is not allowed to grow beyond the radius it had at optimum contact. 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 10 and H-bond volume by 4 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:

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

For multi-atom interactions, the contact dots and their scores combine in a more complex way than addition of unmodified pairwise 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.

The Probe program can summarize these scoring data for all or selected parts of an entire structure; alternatively, it can output an 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 at the same time Probe calculates the contacts. For contexts in which elimination of physically impossible atomic overlaps is the main concern (as in Table 3), 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 $\geqslant 0.4$ Å. This list could usefully prioritize the analysis of problem areas, especially during structure refinement.

Choice of reference datasets

The data set of 100 protein structures used for these initial investigations (listed in Table 2) 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 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 ProCheck (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 subunit 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. 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 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 files 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).

Solvent accessibility

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 & Richards, 1971; Shrake & Rupley, 1973). However, we can obtain an analogous measure of solvent accessibility 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. Those dots 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, the standard side-chain areas† were obtained by running the above algor-

† Standard side-chain 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.

ithm for each side-chain rotamer weighted by empirical rotamer occurrence (Dunbrack & Cohen, 1997). These areas are smaller than the traditional accessible-surface areas, 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 these solvent-accessible dots to identify buried atoms, either of the protein or for 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.

Side-chain analyses

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 Nemethy *et al.* (1992), leaving the backbone unchanged and either re-using C^{δ}, or else placing C^{δ} in the plane of 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 resdues that showed significant clashes in their original conformation.

For the analysis of methionine rotamers, side-chain dihedral angles were calculated with locally developed software. Since, in the context of rotamer analyses, the *gauche*$^{+}$, *gauche*$^{-}$ terminology has been used equally often with each of two opposite meanings, we instead here use the abbreviations **p** for the **p**lus bin, **t** for the ***t**rans* bin, and **m** for the **m**inus bin, with divisions at 0° and $\pm 120^{\circ}$. For comparison of theoretical contacts and clashes in rotamer conformations, two sets of ideal-geometry residues were constructed, one using parameters from ECEPP (Momany *et al.*, 1975; Nemethy *et al.*, 1992) and the other by Engh & Huber (1991). They were examined in Mage, including rotation of conformational angles. The differences between the two parameter sets were not crucial for the Met results described here, but they are for some of the other amino acids. 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; structure Figures were made in Mage and modified in Adobe Photoshop or Illustrator.

Program and data availability

The annotated list of 100 high-resolution structures, the coordinate files with H atoms added, and optimized and the various corrections from this and the accompanying paper made, and a set of contact-dot kinemage files, plus the programs Reduce, Probe, Prekin, and Mage, the modified het dictionary, and several supplementary utilities and scripts (such as Dang, which calculates dihedral angles) are available from either the anonymous FTP site (ftp://kinemage.biochem.duke.edu) or the Web site (http://kinemage.biochem.duke.edu). Probe is a generic Unix C program; the current version (v2) of Reduce that includes H-bond optimization is in C^{++}. Mage and Prekin are available in Mac, PC, Linux, and SGI Unix versions, and can be compiled to run on most Unix platforms where Motif is available. A stripped-down version of Mage written in Java is used on our Web site to provide real-time interactive display of small kinemages with contact dots.

Results

The first obvious result from calculating contact dots 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). Even methyl groups, with a rather low barrier to rotation, are amazingly relaxed inside proteins, nearly all of them fitting excellently in staggered conformation (e.g. Figure 3(a)). 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 Procedures), 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 bond angle can be rotated much in either direction without producing clashes.

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(a)) while a core region near Leu99 is sparser.

Explicit hydrogen atoms

In order to see these elaborately well-fitted protein cores, two conditions must both be met: the structure must have been determined with very great accuracy, and all H atoms must be represented explicitly. 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 3(d)

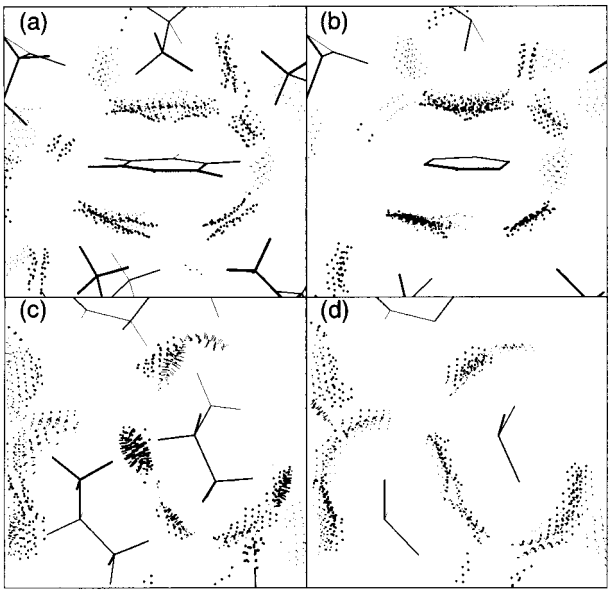

Figure 3. (a) and (b) Comparison of contact dots calculated with explicit *versus* implicit (''united atom'') hydrogen atoms, for an Ala/Phe interaction in 3LZM. The explicit dots in (a) show that the Ala methyl is positioned almost perfectly for contact with the Phe ring, while the contacts shown by implicit dots are much sparser and rounder. Note that refinement did not use 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 (c), but with explicit-H dots the inadvertent clashes are shown to be disastrous.

versus (b)). However, if full explicit hydrogen atoms are added geometrically to those models, they are seen to have numerous bad clashes (Figure 3(c)), whereas if H atoms are added geometrically to high-resolution X-ray structures the contact surfaces fit very well (Figure 3(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 *et al.*, 1991) but are much less well ordered than natural proteins when they are actually produced (e.g. Richardson *et al.*, 1992; Betz *et al.*, 1993), it is important to develop more stringent model-evaluation criteria such as the explicit H contact dots.

Since rather few sets of 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. Desjarlais & Handel, 1995; Struthers *et al.*, 1996; Dahiyat & Mayo, 1997) were modelled using explicit H contacts between (although not within) residues; some cases have achieved well-ordered structures in which only local 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 3(c), or of choosing unfavorable backbone geometries to connect secondary structures.

High resolution

For the current set of 100 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 Procedures) are all generally related to resolution. For instance, Figure 4 plots the number of serious clashes (overlap $\geqslant 0.4$ Å) per 1000 atoms as a function of resolution, for the 100 reference proteins. 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.

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 *et al.*, 1996) or the 1RGE ribonuclease SA (Sevcik *et al.*, 1996) using ShelX (Sheldrick & Schneider, 1997), or for the 1CNR crambin (Yamano & 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 5 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 $\geqslant 0.4$ Å even on the outside of the molecule where one might expect less order.

Alternate conformations: *B*-factors

In Figure 5, 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 cor-

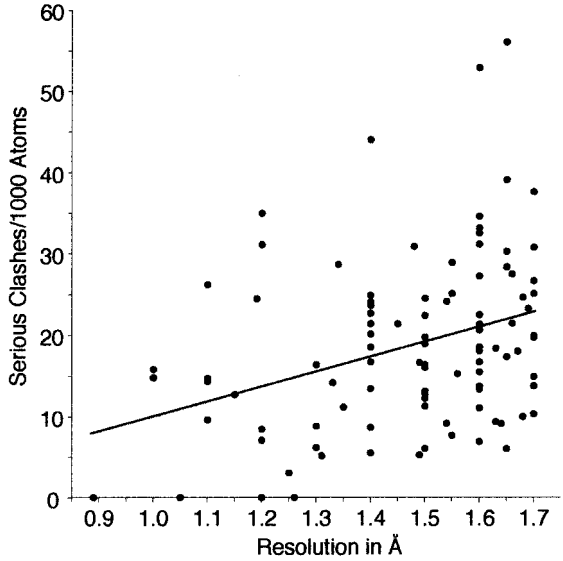

Figure 4. The number of serious atom clashes (overlap >0.4 Å) per 1000 atoms, for each of the 100 reference proteins, plotted *versus* resolution in Å; these values include the high *B*-factor and amide-flip clashes. Although the scatter is high, the regression line shows a significant relationship.

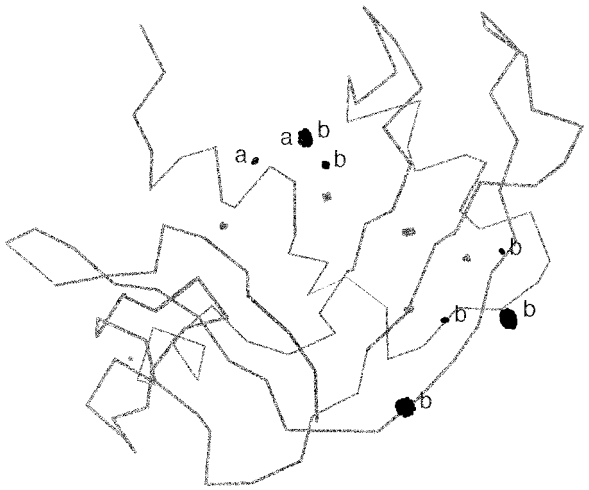

Figure 5. C^{α} backbone, plus spikes for all overlaps >0.25 Å, for the 7RSA ribonuclease A (Wlodawer *et al.*, 1988). When only non-alternate and 'a' alternate conformations are included, this is the cleanest of all the 100 structures, with only a few small overlaps that barely reach this level. However, three severe clashes can be seen, each of which involves a residue in the 'b' alternate conformation.

rectible: 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. 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 6(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 6(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 6(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-diffraction 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.

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.

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 Figures 2(a) and 3(a) for T4 lysozyme, Figure 7 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 not true if the

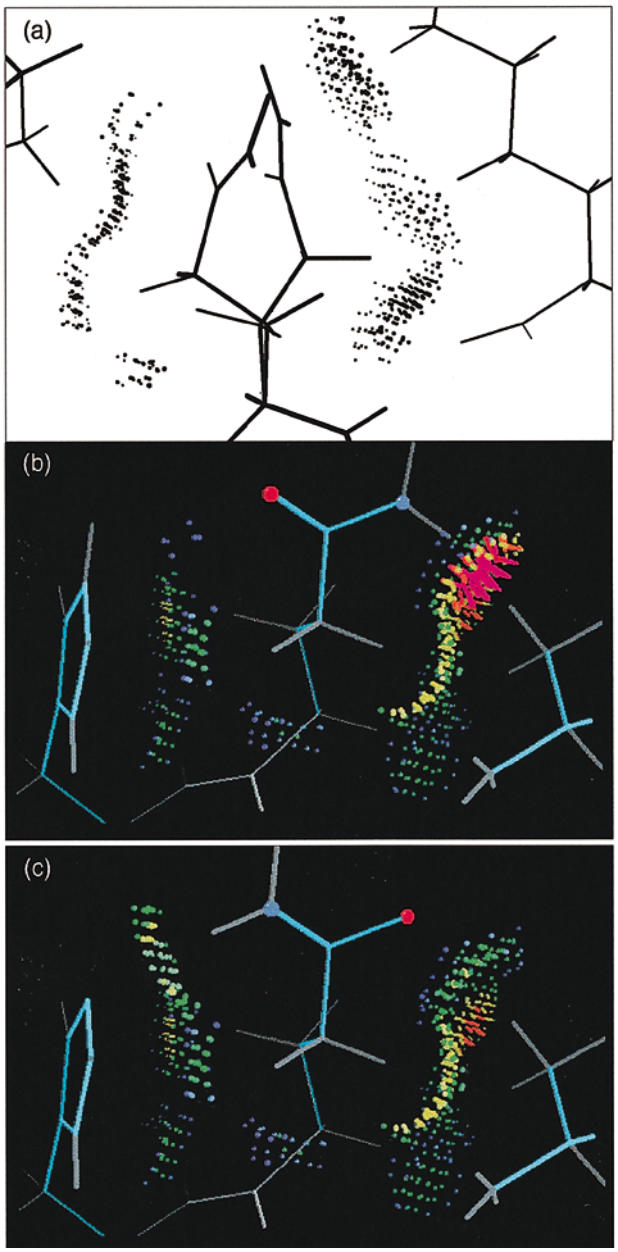

Figure 6. (a) Front view of 7RSA Gln11, with 'a' and 'b' alternate conformations. Each has favorable χ angles, and they hug opposite sides of the available space. (b) Side view of 7RSA Gln11b, showing that it clashes with low *B*-factor atoms of the adjacent Leu side-chain. (c) After flip of the amide, contacts are greatly improved, including a weak H-bond to the His N^{ε} on the left (pale green dots).

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.

NMR structures

Similarly, the interior regions show good packing in the very best determined NMR structures:

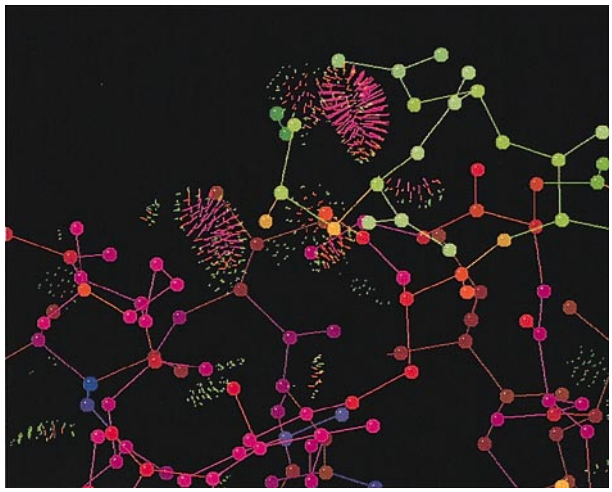

Figure 7. A surface region of 3LZM, colored by *B*-factor. The cooler interior parts have almost no significant overlaps, while the hot, high *B*-value loops at the top (in yellow) show a number of physically impossible clashes.

those with many NOE (nuclear Overhauser enhancement) constraints per residue, stereospecific assignments, and suitable refinement protocols (see Procedures for files used). The distances between explicit individual hydrogen atoms are an integral part of NMR structure determination, and so NMR is capable in the best cases of representing packing very accurately (see Figure 8(a)). However, other regions of those same structures nearly always show bad clashes (e.g. Figure 8(b)), often on the surface where there are fewer NOEs and perhaps motion 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.

Nucleic acid structures

Crystal structures of small oligonucleotides solved at high resolution, e.g. 3DNB (Prive *et al.*, 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 *et al.*, 1996), 7BNA (Holbrook *et al.*, 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

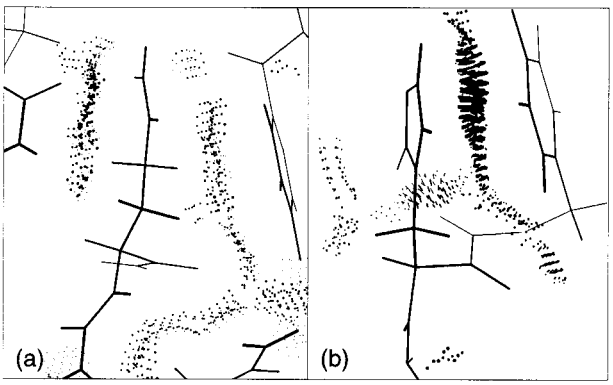

Figure 8. (a) Dot contacts around an interior Gln sidechain in the cellobiohydrolase NMR structure 1CBH (Kraulis *et al.*, 1989), showing excellent fit. (b) A Phe-His clash on the outside of the same structure.

range of 2 to 3 Å. Small-probe contact dots calculated for such structures show an interesting and revealing pattern, as shown in Figure 9, which compares a *B*-DNA structure with a less regular large RNA. The packing is beautiful between the bases, where 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 *et al.*, 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.

Progressive improvement to 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 precondition for such studies, however, is to assemble a set of reference structures with all explicit hydrogen atoms and which are 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 potentially useful. It includes three components: (1) choice of the reference proteins (explained in Procedures) 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.

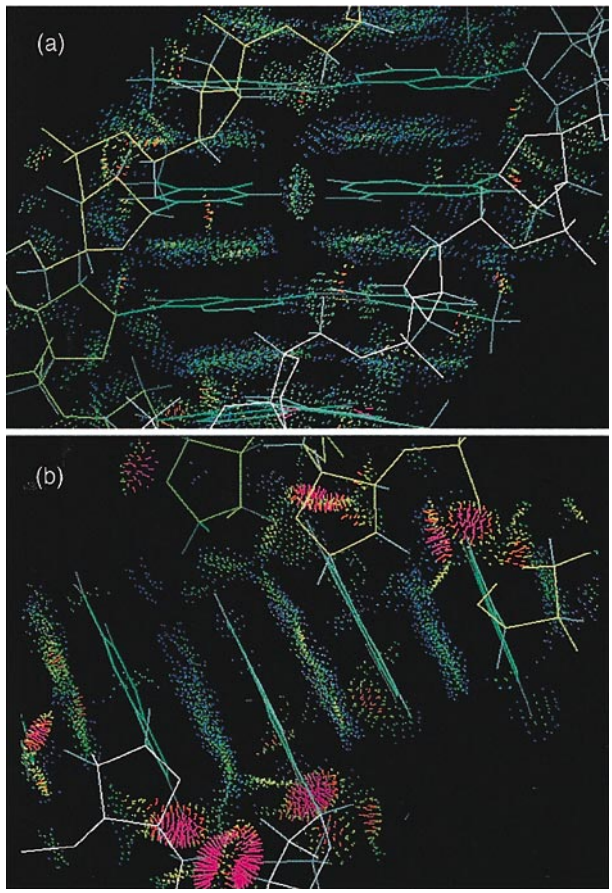

Figure 9. Small-probe dots for (a) the regular *B*-DNA double helix in 7BNA (Holbrook *et al.*, 1985), showing excellent contacts throughout. (b) Part of the large hammerhead RNA in 299D (Scott *et al.*, 1996), showing excellent contact for the bases but severe clashes for the H atoms along the backbone.

The starting basis set for the 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 (see Procedures); all alternate conformations are 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 3 give the average clash score (number of clash overlaps $\geqslant 0.4$ Å per 1000 atoms) for the 100 reference proteins before and after all the OH hydrogen

Table 3. Progressive improvement of PDB datasets

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

atoms have been rotationally optimized one at a time.

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 7). 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 dynamic 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 correctible without re-refinement and many would require additional data, so the most reasonable strategy for our present purposes is to ignore clashes with high *B*-factor atoms.

Figure 10 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. $B < 40$ was chosen as a conservative cutoff criterion, which keeps more than 95% of the atoms while rejecting those whose clashes are most likely to be artifacts of their mobility. The third score entry in Table 3 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.

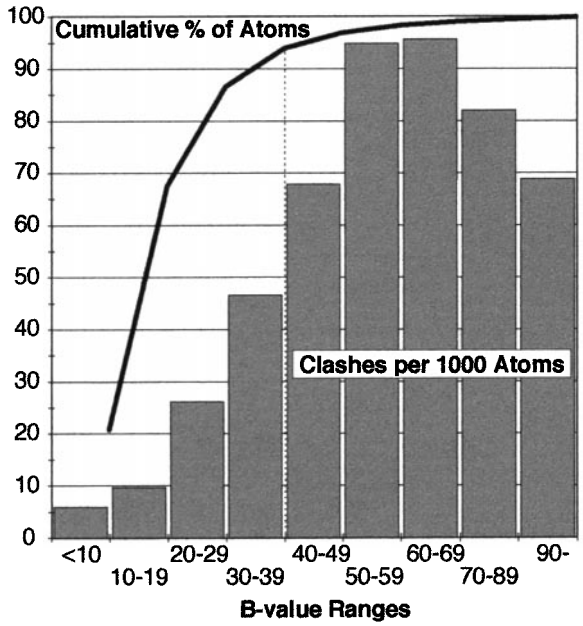

Figure 10. Bars plot clashes per 1000 atoms as a function of *B*-factor, showing that high *B*-factor atoms are enormously more likely to clash with their neighbors. Also plotted is the cumulative percentage of atoms included below a *B*-factor cutoff at that value. Rejecting all atoms with $B > 40$ loses only 5% of the atoms.

Methionine methyl groups

After removal of the high *B*-factor 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 due to the longer C-S bond and the absence of hydrogen atoms on the S. 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.

To further document this difference in packing seen for Met *versus* other methyl groups, Figure 11(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 11(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 3 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

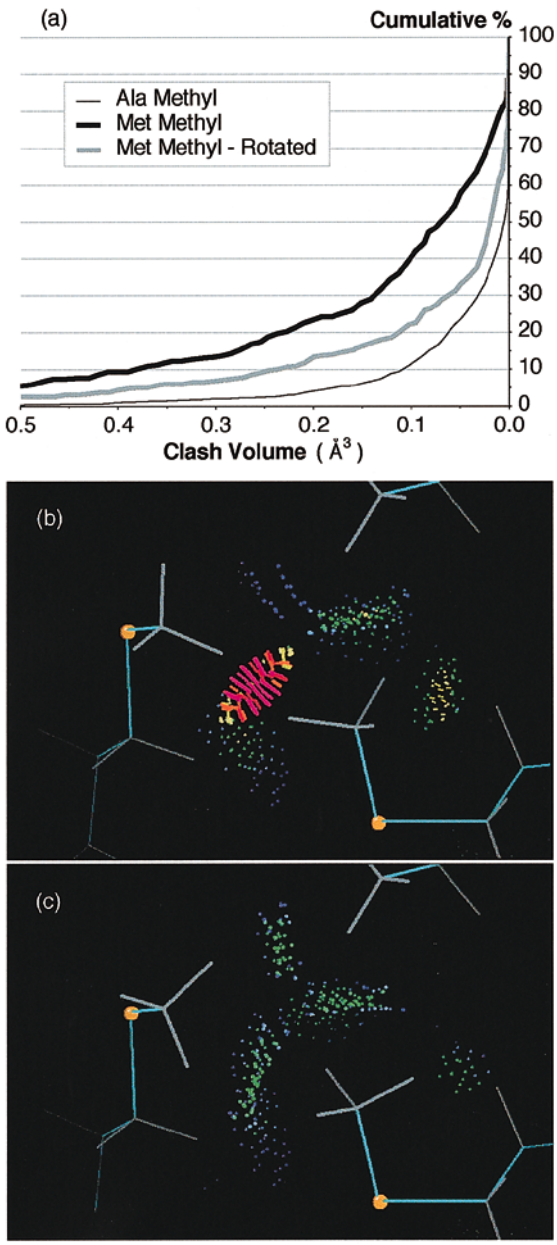

Figure 11. (a) Cumulative distribution of methyl clash volumes, showing that the terminal methyl groups of Met are very much more likely to have large clashes than the β methyl groups of Ala. (b) Two interacting Met methyl groups from squash trypsin inhibitor of file 1MCT chain I (Huang *et al.*, 1993): staggered configuration, with severe clash; (c) both methyl groups rotated, with good contacts.

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 3 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 the accompanying paper (Word *et al.*, 1999) and is described in detail there. Each of the three main modifications to the dataset summarized in Table 3 (exclusion of high *B*-factor atoms, rotation of Met methyl groups, and optimization of H-bond networks) results in a very significant ($p \leqslant 0.001$ or better) improvement in the clash scores, as measured by a "paired-*t*" test of differences in means. In Table 2, 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 Procedures).

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, Nemethy *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 Procedures). Even with no other adjustable parameters, all but two of them showed significantly improved packing, judged visually and by contact-dot scores. Figure 12 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.

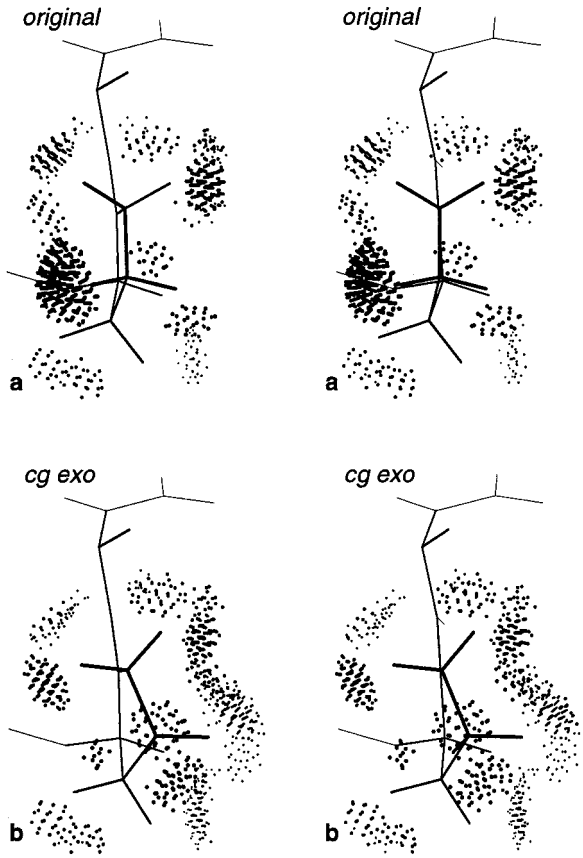

Figure 12. Pro131 from the Zn elastase 1EZM (Thayer *et al.*, 1991), with contact dots (stereo). (a) Original configuration with three serious clashes; (b) ring in C^{γ}-*exo* conformation, with excellent contacts.

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 13(a)), while in the latter structure both H^{α} atoms contact the ring favorably (Figure 13(b)). This improvement results from rotation of the 67-8 peptide and improved planarity of the 65-6 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.

Met rotamers

As well as helping improve the accuracy of structure determination and quantifying packing contacts, the interactions shown by small-probe contact dots 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. They share with the rotamers of aliphatic and unbranched side-chains the advantage that there is no question, for χ_1 and χ_2 at least, that each χ angle clusters into three bins around the three possible staggered values.

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 (Chandrasekaran & Ramachandran, 1970; Bhat *et al.*, 1979; James & 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 (see Procedures for nomenclature). Benedetti *et al.* (1983) used additional data from small peptide structures, but

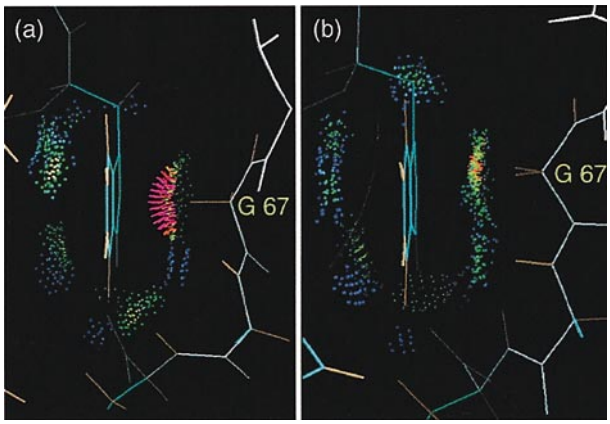

Figure 13. (a) Clash of high *B*-value Gly67 $1H^{\alpha}$ with the ring of Trp74, in *E. coli* DHFR (file 4DFR; Bolin *et al.*, 1982; at 1.7 Å resolution). (b) Contact of low *B*-value Gly67 $1H^{\alpha}$ and $2H^{\alpha}$ with the ring of Trp74, in *E. coli* DHFR (file 1RA9; Sawaya & Kraut, 1997; at 1.55 Å resolution).

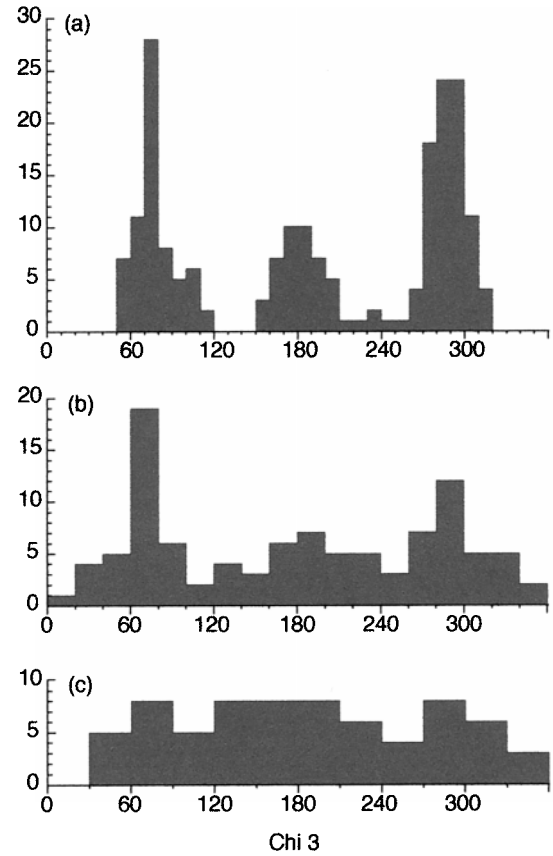

Figure 14. Met χ_3 angle distributions: (a) for the 241 Met side-chains with $B < 30$ in our dataset; (b) for the 90 Met with $B \geqslant 30$ in our dataset; (c) data recorded from Janin *et al.* (1978).

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 14(c)); that observation is quoted by Gellman (1991) and Schrauber *et al.* (1993). Tuffery *et al.* 1991; 1997) 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) have solved the sample size problem by using 518 protein chains now available at 2 Å or better resolution, an order of magnitude more than any of the earlier surveys, and their Met rotamer library is 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), Dunbrack & Cohen (1997) 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 Figure 10, 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 14(a) and (b) compare Met χ_3 values with $B < 30$ *versus* those with *B*-factors for some atom $\geqslant 30$, 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 14(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 ($\mathbf{p} = +75°$, $\mathbf{t} = 180°$, and $\mathbf{m} = -70°$). The χ_3 patterns also differ with χ_2 value; for example, the shoulder seen near $\chi_3 = +100°$ in Figure 14(a) is real, caused by the **mmp** rotamer with a mean of 101° for χ_3.

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 $\pm 75^{\circ}$ (see Figure 15). 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% (their frequencies, means, and standard deviations are listed in Table 4), while seven are completely empty in our 100-protein dataset. As in all previous treatments that included any full rotamers for Met (Ponder & Richards, 1987; McGregor *et al*., 1987; Tuffery *et al*., 1991; Dunbrack & Cohen, 1997), the **mmm** rotamer was found to be the most common. Figure 15 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 (Table 4) 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.

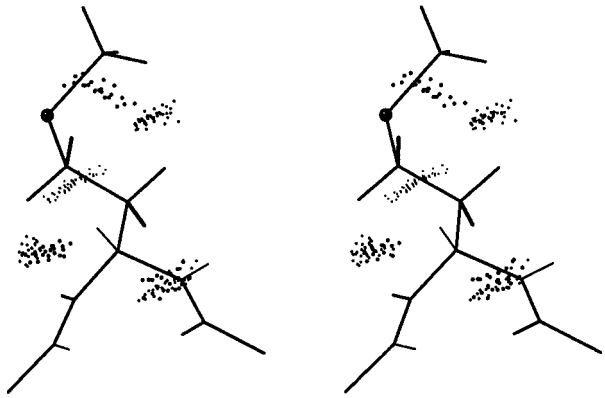

Figure 15. Stereo diagram of the **mmm** rotamer for an ideal-geometry Met residue, with the backbone α-helical, showing contact dots for the five good H atom contacts.

Discussion

Probably the most important general conclusion from this work 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 ''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 *et al*., 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

Table 4. Methionine rotamers from the database of 100 proteins, using side-chains with $B < 30$ and no alternate conformations

Met rotamer	Frequency No.	(%)	χ_1 (deg.) mean	sd	χ_2 (deg.) mean	sd	χ_3 (deg.) mean	sd
mmm	49	20.1	−64	(7)	−62	(9)	−72	(13)
mmp	10	4.1	−65	(5)	−65	(7)	101	(9)
mmt	6	2.5	−67	(11)	−65	(13)	175	(30)
mpm	1	0.4						
mpp	0	-						
mpt	0	-						
mtm	25	10.2	−70	(6)	−177	(15)	−74	(13)
mtp	39	16.0	−69	(6)	178	(11)	73	(13)
mtt	26	10.7	−68	(9)	−178	(8)	−175	(16)
pmm	0	-						
pmp	0	-						
pmt	0	-						
ppm	0	-						
ppp	1	0.4						
ppt	1	0.4						
ptm	9	3.7	64	(7)	−179	(15)	−66	(14)
ptp	7	2.9	65	(6)	−172	(9)	74	(18)
ptt	4	1.6						
tmm	3	1.2						
tmp	0	-						
tmt	1	0.4						
tpm	1	0.4						
tpp	15	6.1	−177	(6)	69	(12)	68	(18)
tpt	6	2.5	−174	(10)	67	(6)	−164	(22)
ttm	16	6.6	−172	(9)	−173	(10)	−67	(12)
ttp	13	5.3	179	(12)	180	(11)	65	(11)
ttt	11	4.5	−169	(9)	174	(12)	177	(18)
	244	99.9						

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 analagous 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. The main examples illustrated here relate to methionine side-chains. 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. 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 $\chi_1 = +60^\circ$ 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. The great value of requiring low *B*-factors as well as high resolution is demonstrated both by the sharpening of Met χ_3 distributions in Figure 14, and by the large clash-score improvements documented for the 100 proteins in Figure 10 and Table 3.

Although the algorithms and parameters for the contact-dot 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.

Acknowledgments

Key conversations and information were provided by Harold Scheraga, Arnie Hagler, Jan Hermans, Peter Kolman, George Sheldrick, Sean Parkin, Jim Kiefer, and Homme Hellinga. We thank Lizbeth Videau for database research. This work was supported by NIH research grant GM-15000, by use of the Duke Comprehensive Cancer Center Shared Resource for Macromolecular Graphics, and by an educational leave for J.M.W. from the Glaxo Wellcome Inc.

References

Axe, D. D., Foster, N. W. & Fersht, A. R. (1996). Active barnase variants with completely random hydrophobic cores. *Proc. Natl Acad. Sci. USA,* **93**, 5590-5594.

Bancroft, D., Williams, L. D., Rich, A. & Egli, M. (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.

Behe, M. J., Lattman, E. E. & Rose, G. D. (1991). The protein-folding problem: the native fold determines packing, but does packing determine the native fold?. *Proc. Natl Acad. Sci. USA,* **88**, 4195-4199.

Benedetti, E., Morelli, G., Nemethy, G. & Scheraga, H. A. (1983). Statistical and energetic analysis of side-chain conformations in oligopeptides. *Int. J. Pept. Protein Res.* **22**, 1-15.

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

Betz, S. F., Raleigh, D. P. & DeGrado, W. F. (1993). *De novo* protein design: from molten globules to native-like states. *Curr. Opin. Struct. Biol.* **3**, 601-610.

Bhat, T. N., Sasisekharan, V. & Vijayan, M. (1979). An analysis of side-chain conformation in proteins. *Int. J. Pept. Protein. Res.* **13**, 170-184.

Bolin, J. T., Filman, D. J., Matthews, D. A., Hamlin, R. C. & Kraut, J. (1982). Crystal structures of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase refined at 1.7 Å resolution. I. General features and binding of methotrexate. *J. Biol. Chem.* **257**, 13650-13662.

Bondi, A. (1964). van der Waals volumes and radii. *J. Phys. Chem.* **68**, 441-451.

Bonvin, A. M., Rullman, J. A., Lamerichs, R. M., Boelens, R. & Kaptein, R. (1993). ''Ensemble'' iterative relaxation matrix approach: a new NMR refinement protocol applied to the solution structure of crambin. *Proteins: Struct. Funct. Genet.* **15**, 385-400.

Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science,* **253**, 164-170.

Bromberg, S. & Dill, K. A. (1994). Side-chain entropy and packing in proteins. *Protein Sci.* **3**, 997-1009.

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

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

Chandrasekaran, R. & Ramachandran, G. N. (1970). Studies on the conformation of amino acids XI. Analysis of the observed side group conformations in proteins. *Int. J. Pept. Protein Res.* **2**, 223-233.

Choma, C. T., Lear, J. D., Nelson, M. J., Dutton, P. L., Robertson, D. E. & DeGrado, W. F. (1994). Design of a heme-binding four-helix bundle. *J. Am. Chem. Soc.* **116**, 856-865.

Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature,* **248**, 338-339.

Chothia, C. & Gerstein, M. (1997). How far can sequences diverge? *Nature,* **385**, 579-581.

Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science,* **221**, 709-713.

Dahiyat, B. I. & Mayo, S. L. (1997). *De novo* protein design: fully automated sequence selection. *Science,* **278**, 82-87.

Dalal, S., Balasubramanian, S. & Regan, L. (1997). Transmuting α helices and β sheets. *Fold. Des.* **2**, R71-R79.

Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1997). The benefits of atomic resolution. *Curr. Opin. Struct. Biol.* **7**, 681-688.

Derewenda, Z. S., Lee, L. & Derewenda, U. (1995). The occurrence of C-H···O hydrogen bonds in proteins. *J. Mol. Biol.* **252**, 248-262.

Desjarlais, J. R. & Handel, T. M. (1995). *De novo* design of the hydrophobic cores of proteins. *Protein Sci.* **4**, 2006-2018.

Dunbrack, R. L., Jr & Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**, 1661-1681.

Engh, R. A. & Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallog. sect. A,* **47**, 392-400.

Fedorov, A. N., Dolgikh, D. A., Chemeris, V. V., Chernov, B. K., Finkelstein, A. V., Schulga, A. A., Alakhov, Y. B., Kirpichnikov, M. P. & Ptitsyn, O. B. (1992). *De novo* design, synthesis and study of albebetin, a polypeptide with a predetermined three-dimensional structure. *J. Mol. Biol.* **225**, 927-931.

Fernandez, C., Szyperski, T., Bruyere, T., Ramage, P., Mosinger, E. & Wuthrich, K. (1997). NMR solution structure of the pathogenesis-related protein P14A. *J. Mol. Biol.* **266**, 576-593.

Fezoui, Y., Weaver, D. L. & Osterhout, J. J. (1994). *De novo* design and structural characterization of an α-helical hairpin peptide: a model system for the study of protein folding intermediates. *Proc. Natl Acad. Sci. USA,* **91**, 3675-3679.

Fossey, S. A., Nemethy, G., Gibson, K. D. & Scheraga, H. A. (1991). Conformational energy studies of β-sheets of model silk fibroin peptides. i. sheets of poly(Ala-Gly) chains. *Biopolymers,* **31**, 1529-1541.

Gassner, N. C., Baase, W. A. & Matthews, B. W. (1996). A test of the ''jigsaw puzzle'' model for protein folding by multiple methionine substitutions within the core of T4 lysozyme. *Proc. Natl Acad. Sci. USA,* **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. *J. Am. Chem. Soc.* **105**, 5220-5225.

Gellman, S. H. (1991). On the role of methionine residues in the sequence-independent recognition of nonpolar protein surfaces. *Biochemistry,* **30**, 6633-6636.

Hagler, A. T., Huler, E. & Lifson, S. (1974). Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *J. Am. Chem. Soc.* **96**, 5319-5327.

Hecht, M. H., Richardson, J. S., Richardson, D. C. & Ogden, R. C. (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., Grossman, S. R., Laimins, L. A. & Sigler, P. B. (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.

Hermans, J., Berendsen, H. J. C., van Gunsteren, W. F. & Postma, J. P. M. (1984). A consistent empirical potential for water-protein interactions. *Biopolymers,* **23**, 1513-1518.

Hingerty, B., Brown, R. S. & Jack, A. (1978). Further refinement of the structure of yeast $tRNA^{Phe}$. *J. Mol. Biol.* **124**, 523-534.

Holbrook, S. R., Dickerson, R. E. & Kim, S.-H. (1985). Anisotropic thermal-parameter refinement of the DNA dodecamer CGCGAATTCGCG by the segmented rigid-body method. *Acta Crystallog. sect. B,* **41**, 255-262.

Houbrechts, A., Moreau, B., Abagyan, R., Mainfroid, V., Preaux, G., Lamproye, A., Poncin, A., Goormaghtigh, E., Ruysschaert, J.-M., Martial, J. A. & Goraj, K. (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 Eng.* **8**, 249-259.

Huang, Q., Liu, S. & Tang, Y. (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. *J. Mol. Biol.* **229**, 1022-1036.

Hurley, J. H., Baase, W. A. & Matthews, B. W. (1992). Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *J. Mol. Biol.* **224**, 1143-1159.

Iijima, H., Dunbar, J. B. J. & Marshall, G. R. (1987). Calibration of effective van der Waals atomic contact radii for proteins and peptides. *Proteins: Struct. Funct. Genet.* **2**, 330-339.

Itoh, S., DeCenzo, M. T., Livingston, D. J., Pearlman, D. A. & Navia, M. A. (1995). Conformation of FK506 in X-ray structures of its complexes with human recombinant FKBP12 mutants. *Bioorg. Med. Chem. Letters,* **5**, 1983-1988.

James, M. N. G. & Sielecki, A. R. (1983). Structure and refinement of penicillopepsin at 1.8 Å resolution. *J. Mol. Biol.* **163**, 299-361.

Janin, J., Wodak, S., Levitt, M. & Maigret, B. (1978). Conformation of amino acid side-chains in proteins. *J. Mol. Biol.* **125**, 357-386.

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

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallog. sect. A,* **47**, 110-119.

Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science,* **262**, 1680-1685.

Karle, I. L., Ranganathan, D. & Haridas, V. (1996). A persistent preference for layer motifs in self-assemblies of squarates and hydrogen squarates by hydrogen bonding [X-H$\cdots$O; X = N, O, or C]: a crystallographic study of five organic salts. *J. Am. Chem. Soc.* **118**, 7128-7133.

Kraulis, P. J., Clore, G. M., Nilges, M., Jones, T. A., Pettersson, G., Knowles, J. & Gronenborn, A. M. (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., Kasibhatla, C., Hidalgo, P., MacKinnon, R. & Wagner, G. (1995). Solution structure of the potassium channel inhibitor agitoxin 2: caliper for probing channel geometry. *Protein Sci.* **4**, 1478-1489.

Kumar, V. D., Harrison, R. W., Andrews, L. C. & Weber, I. T. (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.

Laskowski, R. A., Macarthur, M. W., Moss, D. S. & Thornton, J. M. (1993). ProCheck: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallog.* **26**, 283-291.

Laughlan, G., Murchie, A. I. H., Norman, D. G., Moore, M. H., Moody, P. C. E., Lilley, D. M. J. & Luisi, B. (1994). The high-resolution crystal structure of a parallel-stranded guanine tetraplex. *Science,* **265**, 520-524.

Lee, B. K. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379-400.

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

Matsumura, M., Wozniak, J. A., Sun, D.-P. & Matthews, B. W. (1989). Structural studies of mutants of T4 lysozyme that alter hydrophobic stabilization. *J. Biol. Chem.* **264**, 16059-16066.

McGregor, M. J., Islam, S. A. & Sternberg, M. J. E. (1987). Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J. Mol. Biol.* **198**, 295-310.

McRee, D. E. (1993). *Practical Protein Crystallography*, 1st edit., Academic Press, San Diego.

Momany, F. A., McGuire, R. F., Burgess, A. W. & Scheraga, H. A. (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. *J. Phys. Chem.* **79**, 2361-2381.

Moy, F. J., Li, Y.-C., Rauenbuehler, P., Winkler, M. E., Scheraga, H. A. & Montelione, G. T. (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., O'Brien, R., Sturtevant, J. M. & Regan, L. (1994). Redesigning the hydrophobic core of a four-helix-bundle protein. *Protein Sci.* **3**, 2015-2022.

Mutter, M., Tuchscherer, G. G., Miller, C., Altmann, K.-H., Carey, R. I., Wyss, D. F., Labhardt, A. M. & Rivier, J. E. (1992). Template-assembled synthetic proteins with four-helix-bundle topology. Total chemical synthesis and conformational studies. *J. Am. Chem. Soc.* **114**, 1463-1470.

Nemethy, G., Gibson, K. D., Palmer, K. A., Yoon, C. N., Paterlini, G., Zagari, A., Rumsey, S. & Scheraga, H. A. (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. *J. Phys. Chem.* **96**, 6472-6484.

Parkin, S., Rupp, B. & Hope, H. (1996). Atomic resolution structure of concanavalin A at 120 K. *Acta Crystallog. sect. D,* **52**, 1161-1168.

Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775-791.

Prive, G. G., Yanagi, K. & Dickerson, R. E. (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. *J. Mol. Biol.* **217**, 177-199.

Quinn, T. P., Tweedy, N. B., Williams, R. W., Richardson, J. S. & Richardson, D. C. (1994). Beta doublet: *de novo* design, synthesis and characterization of a novel β sandwich protein. *Proc. Natl Acad. Sci. USA,* **91**, 8747-8751.

Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.* **82**, 1-14.

Richards, F. M. (1977). Area, volumes, packing, and protein structure. *Annu. Rev. Biophys. Bioeng.* **6**, 151-176.

Richards, F. M. & Lim, W. A. (1993). An analysis of packing in the protein folding problem. *Quart. Rev. Biophys.* **26**, 423-498.

Richardson, D. C. & Richardson, J. S. (1990). Protein origami. In *Protein Folding: Deciphering the Second Half of the Genetic Code* (Geirasch, L. & King, J., eds), 1st edit., pp. 5-17, 327-333, American Association Advancement of Science, Washington, DC.

Richardson, D. C. & Richardson, J. S. (1992). The kinemage: a tool for scientific illustration. *Protein Sci.* **1**, 3-9.

Richardson, D. C. & Richardson, J. S. (1994). Kinemages: simple macromolecular graphics for interactive teaching and publication. *Trends Biochem. Sci.* **19**, 135-138.

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. In *Advances in Protein Chemistry* (Anfinsen, C. B., Edsall, J. T. & Richards, F. M., eds), 34th edit., pp. 167-339, Academic Press, New York.

Richardson, J. S. & Richardson, D. C. (1987). Some design principles: betabellin. In *Protein Engineering* (Oxender, D. L. & Fox, C. F., eds), pp. 149-163, 340-341, Alan R. Liss, Inc., New York.

Richardson, J. S. & Richardson, D. C. (1988). 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: Struct. Funct. Genet.* **4**, 229-239.

Richardson, J. S., Richardson, D. C., Tweedy, N. B., Gernert, K. M., Quinn, T. P., Hecht, M. H., Erickson, B. W., Yan, Y., McClain, R. D., Donlan, M. E. & Surles, M. C. (1992). Looking at proteins: representations, folding, packing, and design. *Biophys. J.* **63**, 1186-1209.

Rojas, N. R. L., Kamtekar, S., Simons, C. T., McLean, J. E., Vogel, K. M., Spiro, T. G., Farid, R. S. & Hecht, M. H. (1997). *De novo* heme proteins from designed combinatorial libraries. *Protein Sci.* **6**, 2512-2524.

Salisbury, S. A., Wilson, S. E., Powell, H. R., Kennard, O., Lubini, P., Sheldrick, G. M., Escaja, N., Alazzouzi, E., Grandas, A. & Pedroso, E. (1997). The bi-loop, a new general four-stranded DNA motif. *Proc. Natl Acad. Sci. USA,* **94**, 5515-5518.

Sawaya, M. R. & Kraut, J. (1997). Loop and subdomain movements in the mechanism of *Escherichia coli* dihydrofolate reductase: crystallographic evidence. *Biochemistry,* **36**, 11205-11215.

Schrauber, H., Eisenhaber, F. & Argos, P. (1993). Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J. Mol. Biol.* **230**, 592-612.

Scott, W. G., Murray, J. B., Arnold, J. R. P., Stoddard, B. L. & Klug, A. (1996). Capturing the structure of a catalytic RNA intermediate: the hammerhead ribozyme. *Science,* **274**, 2065-2069.

Sevcik, J., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1996). Ribonuclease from *Streptomyces aureofaciens* at atomic resolution. *Acta Crystallog. sect. D,* **52**, 327-344.

Shakhnovich, E. I. & Finkelstein, A. V. (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. & Schneider, T. R. (1997). SHELX: high resolution refinement. *Methods Enzymol.* **277**, 319-343.

Shortle, D., Stites, W. E. & Meeker, A. K. (1990). Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry,* **29**, 8033-8041.

Shrake, A. & Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **79**, 351-371.

Smith, D. D. S., Pratt, K. A., Sumner, I. G. & Henneke, C. M. (1995). Greek key jellyroll protein motif design: expression and characterization of a first-generation molecule. *Protein Eng.* **8**, 13-20.

Struthers, M. D., Cheng, R. P. & Imperiali, B. (1996). Design of a monomeric 23-residue polypeptide with defined tertiary structure. *Science,* **271**, 342-345.

Tan, S., Hunziker, Y., Sargent, D. F. & Richmond, T. J. (1996). Crystal structure of a yeast TFIIA/TBP/DNA complex. *Nature,* **381**, 127-134.

Thayer, M. M., Flaherty, K. M. & McKay, D. B. (1991). Three-dimensional structure of the elastase of *Pseudomonas aeruginosa* at 1.5- Å resolution. *J. Biol. Chem.* **266**, 2864-2871.

Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1991). A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dynam.* **8**, 1267-1289.

Tuffery, P., Etchebest, C. & Hazout, S. (1997). Prediction of protein side-chain conformations: a study on the influence of backbone accuracy on conformation stability in the rotamer space. *Protein Eng.* **10**, 361-372.

Wahl, M. C., Rao, S. T. & Sundaralingam, M. (1996). The structure of r(UUCGCG) has a 5′-UU-overhang exhibiting Hoogsteen-like *trans* U·U base pairs. *Nature Struct. Biol.* **3**, 24-31.

Wlodawer, A., Svensson, L. A., Sjolin, L. & Gilliland, G. L. (1988). Structure of phosphate-free ribonuclease A refined at 1.26 Å. *Biochemistry,* **27**, 2705-2717.

Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of sidechain amide orientation. *J. Mol. Biol.* **285**, 1735-1747.

Yamano, A. & Teeter, M. M. (1994). Correlated disorder of the pure Pro^{22}/Leu^{25} form of crambin at 150 K refined to 1.05-Å resolution. *J. Biol. Chem.* **269**, 13956-13965.

Edited by J. Thornton

(Received 28 May 1998; received in revised form 2 November 1998; accepted 3 November 1998)