[18] New Tools and Data for Improving Structures, Using All-atom Contacts

By Jane S. Richardson, W. Bryan Arendall III, And David C. Richardson

The methodology of macromolecular crystallography is mature, powerful, and effective, and it has transformed our understanding of biology at the molecular level. However, anyone who has done such a structure knows there are imperfections well worth correcting: occasional mistakes, and difficult places where no alternative seems right. In recent years, macromolecular structures have been improved significantly (that is, made more accurate for a given resolution and data quality) by the use of validation tools such as the free R factor[1] and Ramachandran-plot criteria.[2] Much of the sensitivity and the power of those tools derives from their independence of the target function being optimized in refinement. We have recently developed a new suite of validation tools based on the fact that the van der Waals contacts of hydrogen atoms are almost never part of the refinement target function, yet they yield a large set of powerful constraints on allowed conformations. These new techniques, reviewed here, promise to make significant further improvements in the accuracy and reliability of protein and nucleic acid crystal structures.

[1] A. T. Brunger, *Nature* **355**, 472-475 (1992).

[2] R. A. Laskowski, M. W. Macarthur, D. S. Moss, and J. M. Thornton, *J. Appl. Crystallogr.* **26**, 283-291 (1993).

The all-atom contact technique depends, of course, on adding the hydrogen atoms, most of which are completely determined to a suitable accuracy by the positions of the heavier atoms. This is done by our program REDUCE as described and discussed in Word *et al.*[3] H atoms are placed at ideal bond lengths and angles, with methyls staggered except for terminal Met methyls. Entire local H-bond networks are optimized, including rotation of OH, SH, NH_3 , etc. and 180° flips of Asn, Gln, and His, but with a simplified model for waters.

With hydrogen atoms present, the PROBE program can calculate all-atom contacts.[4] It uses a very small probe sphere of radius 0.25Å, in an algorithm close to the inverse of the Connolly solvent-accessible-surface calculation.[5] Instead of leaving dots where the probe does not intersect another atom, PROBE leaves dots where the small probe does intersect a not-covalently-bonded atom. The result is paired patches of contact surface wherever atoms are within 0.5Å of touching, as shown in Fig. 1. Either for graphical display or for numerical scoring, there are three terms: favorable van der Waals contacts (shown by green and blue dots); favorable overlaps of H-bond donor and acceptor atoms (shown by pillows of pale green dots); and unfavorable overlaps of other atom pairs, shown as "spikes" of increasingly violent red colors as the atomic clash becomes more physically impossible beyond about 0.4Å overlap. The two favorable contact terms evaluate the local goodness-of-fit inside or between molecules. However, the equilibrium structure of a real molecule can have no large clashes, so a crucial criterion for validation of crystallographically derived structural models is the avoidance of serious atomic clashes - a surprisingly demanding requirement once all H atoms are included. Strategies for making use of that criterion will be the topic of this chapter, along with an update of geometrical criteria that complement the all-atom contact analysis.

Geometrical Validation Criteria From Updated Survey Of Database

Three circumstances have motivated us to update the traditional geometrical criteria for protein structure validation: (1) development of the all-atom contact method, which can discriminate one major category of physically impossible from possible conformations; (2) the need to omit high B-factor examples, which surprisingly had very seldom been done before; and (3) the greatly expanded number of structures now available at very high resolution.

[3] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson, *J. Mol. Biol.* **285**, 1735-1747 (1999b).

[4] J. M. Word, S. C. Lovell, T. H. LaBean, H. C. Taylor, M. E. Zalis, B. K. Presley, J. S. Richardson, and D. C. Richardson, *J. Mol. Biol.* **285**, 1711-1733 (1999a).

[5] M. L. Connolly, *Science* **221**, 709-713 (1983).

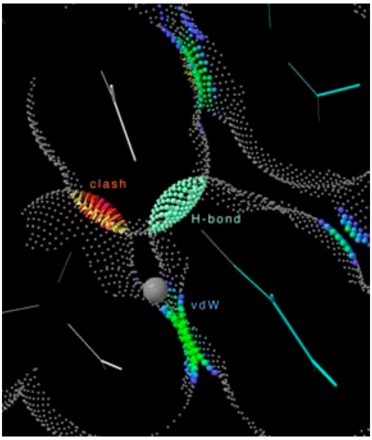


Figure 1. Slice through a small section of protein structure (backbone white, sidechains cyan) showing the relation of all-atom contact surfaces (colored dots) to the atomic van der Waals surfaces (gray dots) and to the 0.25Å radius probe sphere (gray ball) used in the calculation. The probe sphere is rolled over the surface of each atom, leaving a contact dot only when the probe touches another not-covalently-bonded atom. The dots are colored by local gap width between the two atoms: blue near maximum 0.5Å separation, shading to bright green at perfect van der Waals contact (0-Å) gap). When suitable H-bond donor and acceptor atoms overlap, the dots are pale green, forming lens shapes. When incompatible atoms interpenetrate, their overlap is emphasized with "spikes" instead of dots, and with colors from yellow for negligible overlaps to bright reds and pinks for serious clash overlaps ≥ 0.4Å. Kinemage-format contact dots also carry color information about their source atom (e.g., O red, S yellow); in MAGE, one can toggle between the two color schemes. For black-and-white figures, careful attention must be paid to the different appearance of dots (favorable) and spikes (unfavorable). Figure produced in MAGE.[6,7]

Filtering for quality at the local level (e.g., by B-factor) as well as at the whole-structure level (e.g., by resolution) can remove much of the noise in empirical distributions of conformational features, while plotting occurrence as a function of quality indicator can identify and allow removal of some kinds of systematic errors. Therefore, we have revisited the classical rotamer and ϕ,ψ criteria and have proposed the use of Cβ deviation as a single measure encapsulating the most important aspects of bond angle distortions.

[6] D. C. Richardson and J. S. Richardson, *Prot. Sci.* **1**, 3-9 (1992).

[7] J. S. Richardson and D. C. Richardson, *in* "International Tables for Crystallography" (M. G. Rossmann and E. Arnold, eds.), Vol. F: "Crystallography of Biological Macromolecules", pp. 727-730, Kluwer Academic, Dordrecht, The Netherlands, 2001.

Side-Chain rotamers

It has been known since Ponder and Richards[8] that not only do individual sidechain χ angles show distinct preferences (e.g., staggered for tetrahedral geometry), but also there are strong preferences for and against particular combinations of those angles over and above what would be predicted by multiplying the individual distributions. Favorable local energy minima in the multidimensional χ space are known as rotamers, and many authors have compiled libraries of sidechain rotamers.[8-13] Such libraries are often used when fitting models to electron density maps, and either $\chi 1$ alone or $\chi 1$-$\chi 2$ distributions are also used as structure-validation criteria.[2]

Sidechain rotamer libraries are a very powerful and productive tool, but some are too sparse and, now that all-atom contact analysis is available, it can be seen that all previous libraries included at least some physically impossible rotamers. Since H-atom contacts are not refined, even high-resolution structures can have impossible clashes in regions where the electron density was ambiguous; if those bad conformations were systematic errors that occur more often than random (such as flipped-over sidechains) then they made their way into rotamer compilations, where again their H-atom contacts were not checked. Putative rotamers with serious internal clashes also occur in some libraries because of methodological idiosyncracies. Once incorrect rotamers were listed in libraries, then a vicious cycle made them occur even more often in the experimental structures. Figure 2 shows a sample of such cases from earlier rotamer libraries; the spikes show severe unfavorable H-atom contacts internal to each of these defined conformations.

The primary goal of our new "penultimate" rotamer library[14] was nearly complete coverage of the high-quality database, using rotamers located from the empirical distributions but avoiding the problems described above, so that each rotamer represents a physically reasonable local energy minimum. That library was compiled from a non-redundant database of 240 structures at 1.7Å or better satisfying various other quality and relevance criteria[14], and individual sidechains were omitted if they had any atom with B ≥ 40, alternate confor-

[8] J. W. Ponder and F. M. Richards, *J. Mol. Biol.* **193**, 775-791 (1987).

[9] P. Tuffery, C. Etchebest, S. Hazout, and R. Lavery, *J. Biomolec. Struct. & Dyn.* **8**, 1267-1289 (1991).

[10] T. A. Jones, J.-Y. Zou, S. W. Cowan, and M. Kjeldgaard, *Acta Crystallogr. A* **47**, 110-119 (1991).

[11] M. De Maeyer, J. Desmet, and I. Lasters, *Folding & Design* **2**, 53-66 (1997).

[12] H. Schrauber, F. Eisenhaber, and P. Argos, *J. Mol. Biol.* **230**, 592-612 (1993).

[13] R. L. Dunbrack and F. E. Cohen, *Protein Sci.* **6**, 1661-1681 (1997).

[14] S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson, *Proteins: Struct. Funct. Genet.* **40**, 389-408 (2000).

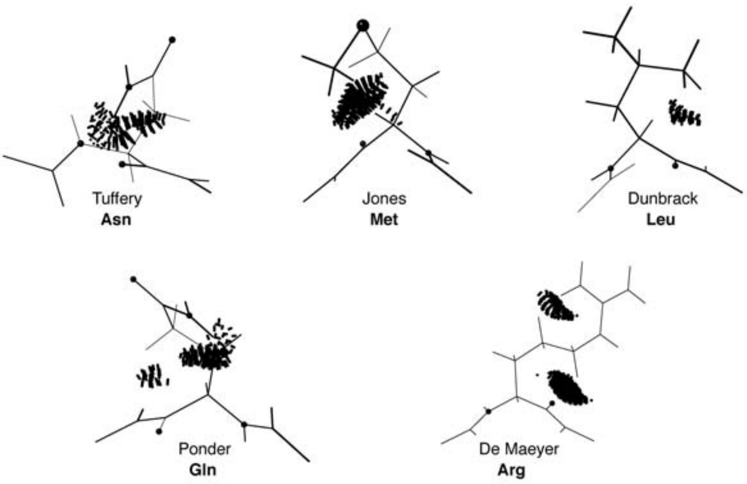


Fig. 2. Examples of defined rotamers that have serious internal clashes, taken from previous rotamer libraries: Ponder and Richards,[8] Tuffery *et al.*,[9], Jones *et al.*,[10], Dunbrack and Cohen,[13] and De Maeyer *et al.*,[12] In addition to the stick-figure for the relevant residue in ideal geometry,[28] only the spikes for clash overlaps are shown; all include one or more serious clashes ≥ 0.4Å. Figure produced in MAGE.[6,7]

mations, serious clashes, covalent modifications, or (for Asn, Gln, His) an uncertain flip state.[3] Rotamers were defined as the modal (peak) values in the smoothed distribution rather than as mean values to avoid dependence on a priori bin or range definitions, to allow for skewed distributions, and to correspond better with energy minima. Wherever compatible with the data, related rotamers were given common χ angles (producing common atom positions), to avoid having users choose between rotamers based on differences that are not statisically significant.

The most general result from this new side-chain survey is that the conformational distributions are even more tightly clustered than observed previously. The weighted average of all $\chi 1$ standard deviations is now only 8.6°, compared with 15.3° for Ponder and Richards[8] and 12.4° for Dunbrack and Cohen.[13] Figure 3A shows the distribution of $\chi 1$-$\chi 2$ values for Met. Note both that the clusters are not always centered exactly on the staggered values, and also that occurrence frequencies often differ greatly from those predicted by the individual χ values (e.g., $\chi 1$ minus is most common overall, but is nearly absent if $\chi 2$ is plus). Figure 3B shows the superimposed sidechains (from 100 proteins) for all Met with $\chi 2$ trans and $\chi 1$ either minus or trans; note the two well-separated

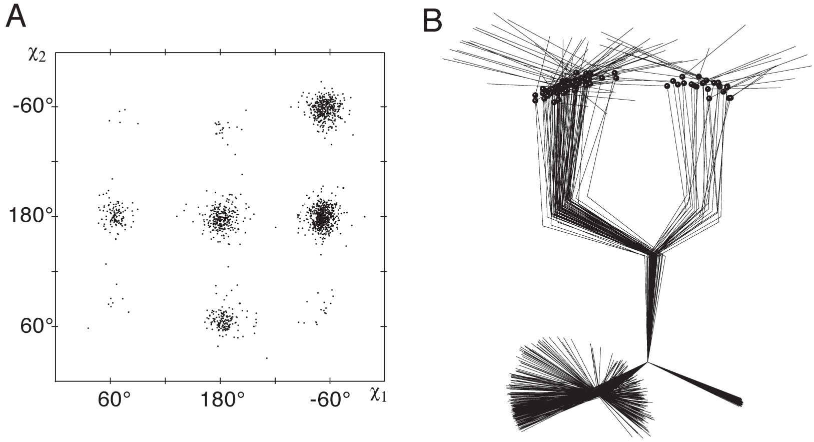


Fig. 3. Rotamers of methionine. (A) χ_1-χ_2 plot for all Met in the Top500 database with B < 40. Although all staggered combinations occur, only 5 of the 9 are common. Note that some clusters are significantly shifted, and also that occurrence frequencies are very different than what would be predicted by multiplying the χ_1 and the χ_2 preferences. (B) All Met examples superimposed that have $\chi_1\chi_2$ trans-trans or plus-trans, from the first 100 structures of the Top 500. Note that the sulfur positions (marked with balls) form two tight and distinct clusters. The Cδ atoms form six looser but still well-defined clusters. Figure produced in MAGE.[6,7]

clusters for the S atom positions. Adjacent rotamers are nearly always quite distinct, as shown even for the extreme case of Lys in Fig. 7 of Lovell *et al.*[14] For all 18 movable sidechain types in the quality-filtered dataset, we found 94.5% to be "rotameric" (within the ranges around defined rotamers, usually ± 30° in each χ angle). The library includes a total of 152 rotamers, or an average of 8 1/2 per movable residue type, from a maximum of 34 for Arg down to 2 for Pro (only Cγ exo and Cγ endo puckers). The rotamers are tabulated in Lovell *et al.*[14] and are available on our website[15] in various forms, including drop-in files for use in O[10] or XtalView.[16,17]

For most residue types, different secondary structures (α-helix, β-sheet, lefthanded, and other) show different occurrence frequencies for each rotamer, but the modal χ values stay the same. Therefore, percentage occurrences for each case are given in the penultimate library tables, but the list of possible rotamers is common. For Asn and Asp, however, the set of modal positions varies with

[15] Richardson Lab, <http://kinemage.biochem.duke.edu>, Duke University, Durham, NC, 2002.

[16] D. E. McRee, "Practical Protein Crystallography", Academic Press, San Diego, 1993.

[17] D. E. McRee, *J. Struc. Biol.* **125**, 156-165 (1999).

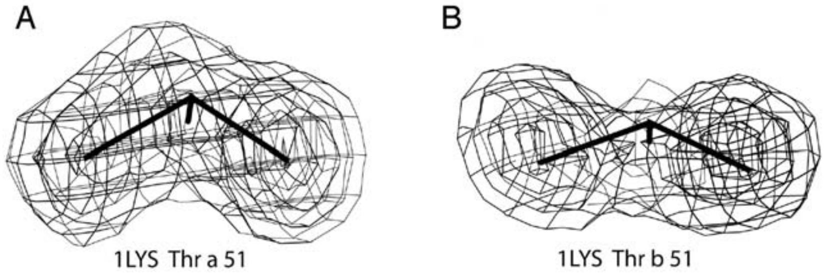


Fig. 4. Electron density contours for two Thr sidechains of the 1LYS[19] hen-egg lysozyme. (A) Thr 51 of molecule A with the expected boomerang-shaped density for the tetrahedral branch around the Cβ. (B) Thr 51 of molecule B, with approximately straight-across density; this occurs relatively often and makes it easy to fit the sidechain eclipsed and backwards, as happened here.

secondary structure[18], so that backbone-dependent rotamers are needed and are defined only for those two residues. For example, α-helical Asn with $\chi 1$ minus shows two close but quite distinct clusters at $\chi 2$ = -20° and at $\chi 2$ = -80°; the latter conformation makes an Nδ H-bond to the i-4 CO, while the former has the flat face of the sidechain amide packed against the i-4 backbone.[18] Lefthanded (or +ϕ) Asn with $\chi 1$ minus, however, shows a single peak at $\chi 2$ = -30° and for β-sheet Asn the peak is at $\chi 2$ = -50°.

Another observation resulting from the rotamer survey was that crystal structures are prone to occasional systematic errors which can be recognized and expunged from rotamer libraries and also from individual structures. One cause of such errors can be electron density for a tetrahedrally branched sidechain that is straight across (as in Fig. 4B) rather than boomerang-shaped, and which is therefore easy to fit incorrectly with the group rotated by 180°. Such examples for Thr are analyzed below, and the more complex case of Leu is presented in detail in Lovell *et al.*[14] The incorrect fitting can be distinguished very definitively, because it has distorted bond angles, eclipsed χ values, atomic clashes, and an increased occurrence rate at high B and low resolution.

Overall, the use of a good rotamer library makes sidechain fitting more accurate as well as faster. Not every side chain is rotameric, since strained contacts or several good H-bonds can dictate an otherwise unfavorable conformation; however, such cases are surprisingly rare and should be accepted only when

[18] S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson, *Proc. Natl.Acad.Sci. USA* **96**, 400-405 (1999).

[19] K. Harata, *Acta Crystallogr. D Biol Crystallogr.* **50**, 250-257 (1994).

there are good physical reasons and when no rotamer can fit acceptably. In particular, partially disordered surface sidechains should always be fit rotameric or as mixtures of rotamers, since there are no interactions to force them away from the local minima.

Ramachandran plots

The Ramachandran plot is especially useful as a geometrical validation criterion because ϕ and ψ are not part of the target function for refinement.[20] The percentage of residues found within the most favored ϕ,ψ regions correlates strongly with resolution and is now standardly reported in protein structure papers, while individual "outlier" residues in accurate structures are taken to indicate either possible errors or potentially-interesting strained conformations. The pioneering, and still most widely used, ϕ,ψ criteria are those in ProCheck,[2] which have the considerable advantage of defining multiple levels of core, allowed, and generously allowed regions; however, they have the serious drawback of being based on old and inaccurate data (the entire PDB from 1990, including structures at 3.5Å resolution and residues with $B > 100$), which made it impossible to locate those outer regions correctly. In reaction, Kleywegt and Jones[21] chose to define only one "strictly allowed" boundary at 98% of a much more accurate dataset, which provides a better validation criterion but does not address the issue of identifying outliers.

We have used a new 500-protein database at 1.8Å resolution or better, B-factor filtering (keeping only residues with all backbone $B < 30$), and density-dependent smoothing to update the assignment of favored, allowed, and outlier regions in ϕ,ψ space.[22] This quality-filtered database of about 100,000 residues defines the allowed-but-disfavored regions quite clearly, because there now are essentially no points at all in the strongly disallowed regions which occupy nearly 60% of the plot area. Figure 5A shows that new ϕ,ψ distribution for the general case: i.e., for all residues not Gly, Pro, or pre-Pro. The omission of pre-Pro has the effect of deleting an area around $\phi = -130°$, $\psi = +80°$ (below left of β) from the favored region. Other than that difference, the inner smoothed contour enclosing 98% of the data (our "favored" region) matches quite exactly with the allowed region of Kleywegt and Jones.[21] In addition, in order to separate the somewhat-disfavored but "allowed" conformations from the strongly disallowed "outlier" regions, an outer contour is shown that includes 99.95% of the high-quality data. Note that there is a "shoal" of disfavored-but-allowed

[20] A. L. Morris, M. W. MacArthur, E. G. Hutchinson, and J. M. Thornton, *Proteins: Struct. Funct. Genet.* **12**, 345-364 (1992).

[21] G. J. Kleywegt and T. A. Jones, *Structure* **4**, 1395-1400 (1996).

[22] S. C. Lovell, I. W. Davis, W. B. Arendall, III, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson, *Proteins: Struct.Funct. Genet.* **50**, 437-(2002).

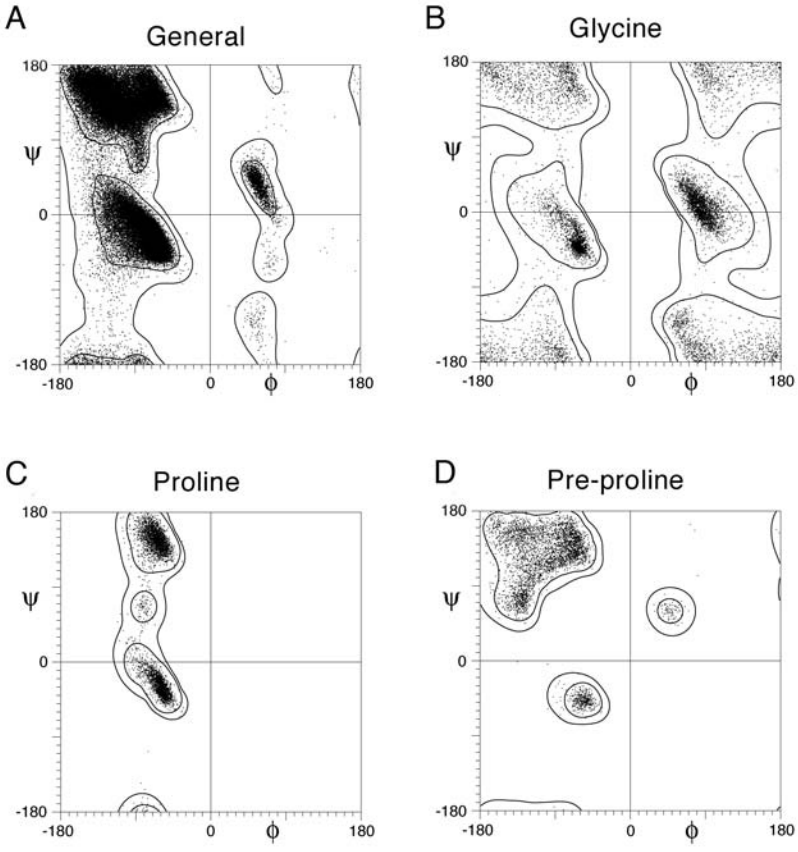


Fig. 5. ϕ,ψ plots for all data from the Top500 structures with backbone B-factors < 30; density-dependent smoothing was used in calculating contours, as explained in Lovell *et al.*[22] (A) The general case, of 81,234 non-Gly, non-Pro, non-prePro residues. The inner contour encloses the "favored" region and 98% of the data; the outer contour encloses the "allowed" 99.95% of the data; the "outlier" region outside encompasses 58.5% of the plot area but almost no high-quality data points. (B) The 7,705 Gly residues, which are more permissive than other residues but still have an extensive forbidden region around ϕ = 0°. Favored and allowed contours for the individual-residue plots are at 98% and 99.8%; those for Gly are symmetrized around the center. c) The 4,415 Pro residues, which are limited to −100° < ϕ < -50°, with major peaks in the polyPro and α regions and a small intermediate peak in the inverse-γ region. (D) The 4,014 prePro (residues that precede Pro but are not Gly or Pro), which are also constrained significantly by clashes with the following Pro Cδ. Reproduced from Lovell *et al.*[22] Figure produced in MAGE.[6,7]

conformations that winds down through the plot near ϕ = +70°. This shoal includes the γ-turn and II' turn regions and many of the examples occur at active sites or binding sites,[23-25] but these conformations are categorized as forbidden by prior validation tools.

Separate Ramachandran plots for Gly residues have sometimes been provided, but their outliers are not penalized in summary statistics; thus, by default they are considered completely permissive. That is unfortunate, since Gly is actually harder to fit correctly than residues with observable Cβ atoms. Figure 5B plots the Gly ϕ,ψ distribution for our quality-filtered database, showing that, although Gly is much more permissive, it still has extensive regions of forbidden conformation, mainly near $\phi = 0°$. The inversion symmetry of Gly occurrence frequencies around the 0°, 0° center point is broken numerically by the special usefulness of Lα glycines, but the outlines of the Gly favored and allowed regions are symmetrical and have been averaged here to improve their accuracy. Note that for all three individual-residue plots, there is only enough data to define a well-behaved outer contour at 99.8% rather than 99.95%. However, both general and individual cases have inner contours at 98%, so those scores can all be combined when evaluating a structure.

Pro is, of course, a special case, because the closed ring constrains ϕ to be near -70°. Interestingly, as well as the classic poly-Pro and α regions, Pro shows a small but well-defined intermediate peak in the inverse-γ region (Fig. 5C). That intermediate peak does not occur for cis Pro or for Pro that are pre-Pro. Both γ and inverse-γ conformations are slightly strained, but are stabilized by a CO (i -1) to NH(i + 1) H-bond that cannot occur for either cis-Pro or pre-Pro.

Other amino acid types besides Gly and Pro show distinctive relative peak heights in their ϕ, ψ distributions (our data, and Hovmoller *et al.*[26]), but their outlines at the 98% level are nearly indistinguishable. However, residues that precede Pro show a very distinctive pattern, as seen in Fig. 5D. They make up most examples in the “pre-Pro” region[27] near $\phi = -130°$, $\psi = +80°$, but they disfavor α and completely forbid the inverse-γ and "bridge" regions between α and β.

These new Ramachandran-plot criteria based on a larger, higher-resolution, quality-filtered dataset can be run on any file, using the MolProbity server accessed through our web site[15] or on the RamPage site at http://www-cryst.bioc.cam.ac.uk/rampage . They discriminate cleanly and robustly between allowed vs outlier backbone conformations, showing which individual residues should be considered either worrisome or interesting. These

[23] B. W. Matthews, *Macromolecules* **5**, 818-819 (1972).

[24] O. Herzberg and J. Moult, *Proteins: Struct. Funct. Genet.* **11**, 223-229 (1991).

[25] K. Gunasekaran, C. Ramakrishnan, and P. Balaram, *J. Mol. Biol.* **264**, 191-198 (1996).

[26] S. Hovmoller, T. Zhou, and T. Ohlson, *Acta Crystallogr. D Biol Crystallogr.* **58**, 768-776 (2002).

[27] P. A. Karplus, *Protein Sci.* **5**, 1406-1420 (1996).

new criteria also provide explicit evaluations for the special cases of Gly, Pro, and pre-Pro residues, allowing all amino-acid types to contribute to an overall Ramachandran-plot validation score at the 98% level.

Cβ Deviations

Another traditional validation criterion is the deviation from ideality of covalent bond lengths and bond angles, and we have found the bond angles, in particular, to be very useful. Bond lengths are so tightly restrained that they can signal local fitting problems only in the rare cases where the refinement parameters were set up incorrectly. Bond angles, on the other hand, are very frequently where the distortions end up when a local region is not simultaneously compatible with good geometry and also with the data. Existing tools such as ProCheck do an excellent job of reporting bond angle distortions from ideal geometry (usually taken as Engh & Huber[28]), appropriately scaled as number of standard deviations off. The summary score for such deviations works well as one component in an overall quality evaluation. However, in spite of the wealth of local information provided, those lists are seldom in practice used to find and correct fitting errors, for two reasons. One difficulty is simply the huge number of angles involved and the relatively smooth distribution of deviation sizes. Another difficulty is that at branches, such as Cα, a bad atom placement can be masked if the change is split between two of the angles.

We have developed Cβ deviation as a single-number measure of bond angle distortions at the Cα,[22] that most critical junction where backbone and sidechain must reconcile any disagreements. Cβ deviation is calculated by building a Cβ in ideal geometry out from the backbone (splitting the difference if the τ angle is non-ideal) and then measuring its distance from the reported Cβ position. All the Cβ deviations for a protein can either be listed or be displayed as the radius of balls shown on the 3D structure as in Fig. 6, where the large ones (>0.25 or 0.3Å) stand out clearly (especially in color on the computer screen). If almost all Cβ deviations in a structure are small (< 0.1Å) but a few are large, those cases usually signal misfit sidechains [such as the Leu in Fig. 6 (inset) or the Thr cases described below]; these are the cases most valuable for structure improvement. If many of the deviations are large, then the details of the refinement strategy should probably be questioned. Large Cβ deviations for residues with alternate conformations, however, usually just mean that the backbone also has alternate conformations that were not modeled. The direction of the Cβ deviations can be plotted in MolProbity (expressed as a torsion angle

[28] R. A. Engh and R. Huber, *Acta Crystallogr. A* **47**, 392-400 (1991).

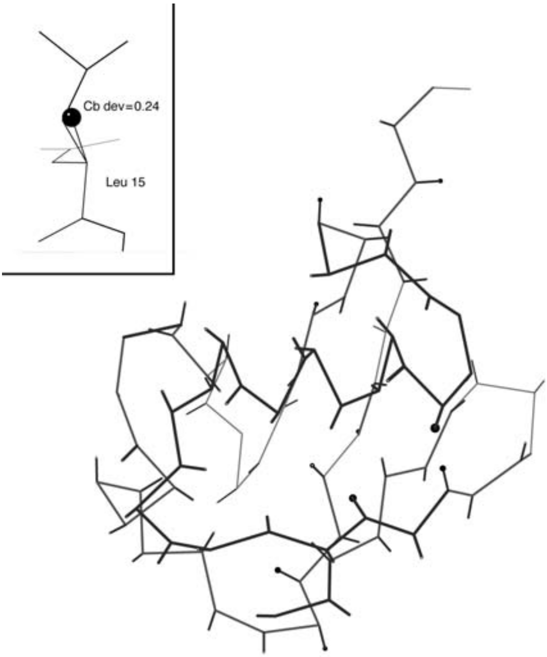


Fig. 6. Deviations of observed Cβ from the ideal position calculated from the backbone atoms; the deviation is shown as the radius of a ball centered at the ideal Cβ. All Cβ deviations for the 1UBQ[29] ubiquitin structure are shown with the Cα backbone; most of the balls are too small to see, but the few large Cβ deviations show up clearly, especially when viewed on-screen in color. *Inset:* Closeup of Leu 15, with a large Cβ deviation of 0.24Å (radius of the ball, centered on the ideal Cβ position). This sidechain was fit reversed into ambiguous electron density, as is fairly common for Leu,[14] but it can be convincingly corrected by idealization and change in both χ_1 and χ_2. Figure produced in MAGE.[6,7]

from N); a strong asymmetry in their distribution can be diagnostic of a missing or incorrectly weighted angle restraint.

All-Atom Contacts for Validation and Improvement

The more revolutionary side of these structure validation tools involves the addition of hydrogens and the use of all-atom contacts. The theory behind these new methods itself had to be validated and the algorithms and parameters optimized. This was done by showing their convergence to agreement with the complete structural details found in proteins as resolution increases to 1Å and beyond (see Fig. 7) and as B-factors decrease. Some atomic-resolution structures are perfect by these criteria, with good rotamers, extensive tight packing,

[29] S. Vijay-Kumar, C. E. Bugg, and W. J. Cook, *J. Mol. Biol.* **194**, 531-544 (1987).

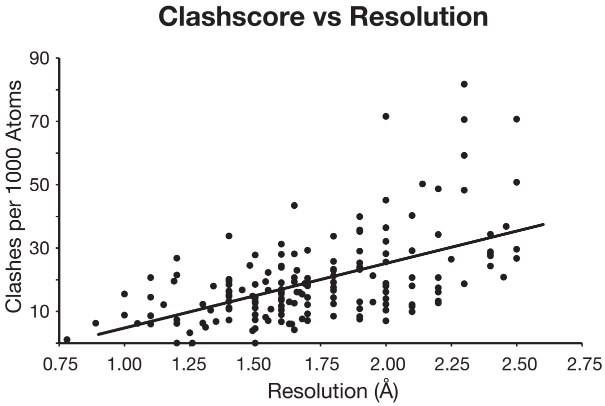


Fig. 7. All-atom clashscore (number of serious overlaps ≥ 0.4Å per 1000 atoms, after correcting amide flips) plotted versus resolution, for 328 non-homologous protein structures between 0.8 and 2.5Å resolution. The relationship is highly significant and is still improving down near 1Å. Reproduced from Richardson.[33]

and no non-H-bond atomic overlaps as much as 0.4Å; examples are conotoxin (1NOT[30]), ribonuclease A (7RSA[31]), and of course the extreme case of crambin (1EJG[32]) at 0.54Å resolution. Other structures at atomic resolution show this same sort of perfection nearly everywhere but have a few isolated clashes, usually on the surface, such as the highly accurate neutral protease at 1Å resolution whose contacts are shown in Fig. 8A. Figure 8B is a closeup illustrating the sort of well-fit packing found in the structure, with hydrogens interdigitated, good contact surfaces all around the Tyr ring, and almost every atom making contacts at ideal distances. At this resolution, the few minor problems found in such a structure may only be with waters, alternate conformations, or high-B regions, but it would still be worth fixing them. The real goal, however, is to remove many of the more numerous clashes seen at lower resolutions and take those structures closer to what would have been found from higher-resolution data.

The first step in this process of diagnosing and improving a structure is to add hydrogens with Reduce, calculate the all-atom contacts with Probe,

[30] L. W. Guddat, J. A. Martin, L. Shan, A. B. Edmundson, and W. R. Gray, *Biochemistry* **35**, 11329-11335 (1996).

[31] A. Wlodawer, L. A. Svensson, L. Sjölin, and G. L. Gilliland, *Biochemistry* **27**, 2705-2717 (1988).

[32] C. Jelsch, M. M. Teeter, V. Lamzin, V. Pichon-Lesme, R. H. Blessing, and C. Lecomte, *Proc. Natl. Acad. Sci. USA* **97**, 3171-3176 (2000).

[33] J. S. Richardson, in "Structural Bioinformatics" (P. E. Bourne and H. Weissig, Eds.). John Wiley & Sons, Inc., New York, 2003.

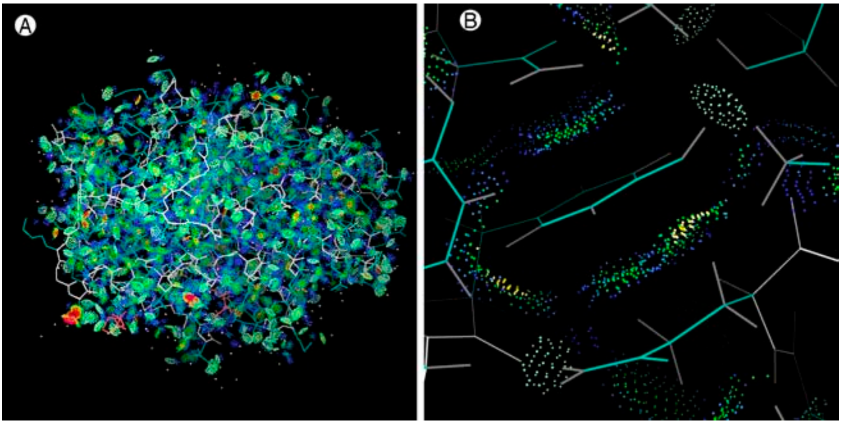


Fig. 8. All-atom contacts for the 1EB6[34] neutral protease at 1Å resolution. (A)Contacts for the entire structure, with a high density of blue and green throughout, showing good H-bonding and well-packed van der Waals interactions. The model has only two serious clashes (red spikes), both on the surface. (B)Closeup on a thin slice of contacts around Tyr 106, to illustrate the detailed appearance of a good model in a well-packed protein interior. Note the three sidechain H-bonds (pale green lenses of dots) and the almost-continuous favorable van der Waals contacts (in blue, green, and yellow) that surround the Tyr ring. Figure produced in Mage.[6,7]

and view the result in Mage. If everything is turned off but the backbone and the serious clashes (overlaps ≥ 0.4Å, red spikes on a display), one can quickly find the trouble spots in even a very large structure. Figure 9 shows the 1LUC luciferase as a typical example; it has about a dozen serious clashes per 1000 atoms, which is better than average for 1.5Å resolution. With such a kinemage display, it is easy to zoom in on a clash, turn the details back on, and analyze its cause; for instance, two of the clashes in the bottom helices are high-B Glx Cγ atoms fit such that a methylene H overlaps the i-4 O atom. For those who prefer to work from a list of clashes, that option is also available. The next few sections will discuss methods for handling problems with different structural features.

Determining Asn, Gln, and His orientations

Settling the correct 180° "flip" state of sidechain amides and His rings is the simplest kind of structure improvement, because it is purely local to the sidechain and does not significantly affect agreement with the diffraction data. Since the H atoms are not observed and the N-versus-O or N-versus-C atoms

[34] K. E. McAuley, Y. Jia-Xing, E. J. Dodson, J. Lehmbeck, P. R. Ostergaard, and K. S. Wilson, *Acta Crystallogr. D Biol Crystallogr.* **57**, 1571-1578 (2001).

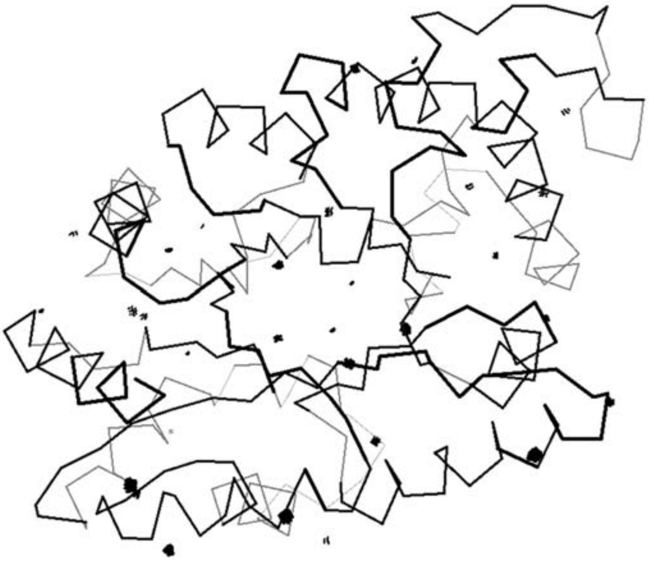


Fig. 9. Cα backbone and serious clashes ≥ 0.4Å (clumps of what should be red spikes) for the B subunit of 1LUC[35] luciferase at 1.5Å resolution. Such a kinemage display quickly locates all the problem areas even in a large structure; one then zooms in and turns on the rest of the model and contact details to study the problem. Figure produced in MAGE.[6,7]

cannot be told apart in the electron density except at extremely high resolution, this distinction traditionally relies just on the comparison of H-bonding energies and is considered subtle and difficult. However, as shown in Word *et al.*,[3] when the amide hydrogens are added and their potential clashes considered, then 80% to 85% of the Asn and Gln flips can be decided quite unambiguously; most of the rest are highly exposed and are probably sampling multiple orientations. Histidine is more complex because different protonation states must be considered, at least in terms of their steric and H-bonding effects (we do not attempt to determine pK_a values). Also the very much weaker but not negligible H-bonding capacity of the ring CH groups means that sometimes all 4 positions have acceptors nearby; however, the CH···O distances are considerably longer than for NH···O or OH···N, and the correct 180° His flip orientation is nearly always clear when both H-bonding networks and all-atom clashes are considered.

For Asn and Gln a number of distinct cases can occur. If there is an obli-

[35] A. J. Fisher, T. B. Thompson, J. B. Thoden, T. O. Baldwin, and I. Rayment, *J. Biol. Chem.* **271**, 21956-21968 (1996).

gate donor or acceptor in good geometry on one side or the other, then simple inspection should always get it right; simulated-annealing refinement will generally get it right but not always: because the two alternatives are not explicitly sampled, the lack of an H-bond can be tolerated, and H-atom clashes are not considered. If the nearby H-bonding groups are ambiguous (e.g., OH, water, or another Asn/Gln/His), then the entire local H-bond network must be considered; a large network is difficult to optimize by inspection, but one best arrangement usually stands out if all alternatives are searched, as is done by REDUCE when run with the "-build" option. If the amide has only van der Waals contacts and no H-bonding, or if the H-bonds are equivalent in the two orientations, then traditional methods cannot tell the alternatives apart. However, the much greater size of the NH_2 group means that all-atom contacts can usually provide a clear answer, as shown for an interesting example in Fig. 10, where all H-bonds are equivalent for the two best conformations but a very large clash of the Gln Hε1 with its Hα in one form makes the choice quite unambiguous. If contact scores for the two possible orientations are nearly equal, REDUCE declares the

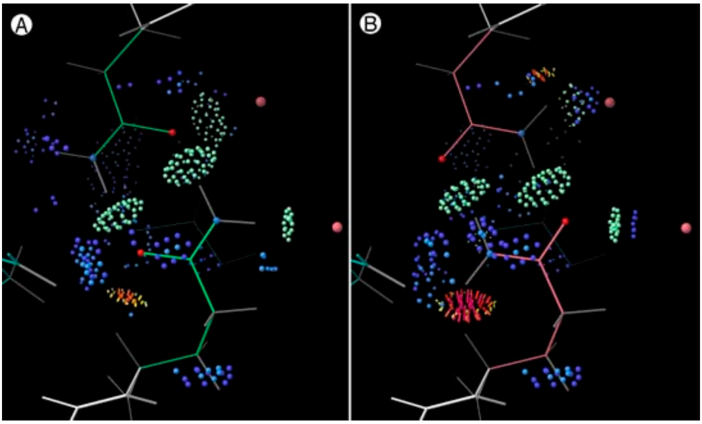


Fig. 10. The use of all-atom contacts to determine the correct 180° "flip" state of an Asn-Gln pair from the 1ARU peroxidase,[36] in context of the local H-bond network. The two best flip states (A and B) are shown, each of which forms the linking pair of H-bonds and is equivalent in water H-bonds and in favorable van der Waals contacts. The choice is clear, however, because conformation (B) makes a serious clash with the Gln Hα. Kinemages that animate between all such paired alternatives are produced by the Flipkin script or on the MOLPROBITY site.[15] Figure produced in MAGE.[6,7]

[36] K. Fukuyama, N. Kunishima, F. Amada, T. Kubota, and H. Matsubara, *J. Biol. Chem.* **270**, 21884-21892 (1995).

case undecided and leaves the orientation as it was.

The above-descriibed procedure is done automatically by the -build option of REDUCE, producing an output PDB-format file with corrected Asn/Gln/His orientations and comments on their scores, as well as including all the new and optimized H atoms. The automatic algorithm is quite reliable, and accepting its results will be an improvement in essentially all cases. However, a script called Flipkin is available for producing a kinemage file that lets the user evaluate critically each of the choices by animating between the two alternatives with all their interactions, as seen for the pair in Fig. 10A and B. Although Asn/Gln/His orientations are a relatively minor aspect of protein structure, they have large effects on H-bonding, electrostatics, and water structure, and they can be really critical to function if they are at an active site or binding site. Running REDUCE is very fast and simple, and should be done on every protein structure before it is deposited.

Other sidechains: Thr/Val/Ile; Leu; Met

At less than atomic resolution the electron density at tetrahedral carbons is fairly often ambiguous, as was illustrated in Fig. 4B. This leads to a significant probability of misfitting, either manually or in refinement, with a $\chi 1$ off by 180° for a β-branched sidechain, or both $\chi 1$ and $\chi 2$ off in a more complex pattern for Leu.[14,37] For Met the S density may be too round with little indication of Cγ or Cε, allowing the sidechain model to reach the S from the wrong direction by changing both $\chi 1$ and $\chi 2$. Such misfittings can be readily located and fixed using the combined criteria of all-atom clashes, Cβ deviations, and bad rotamers.

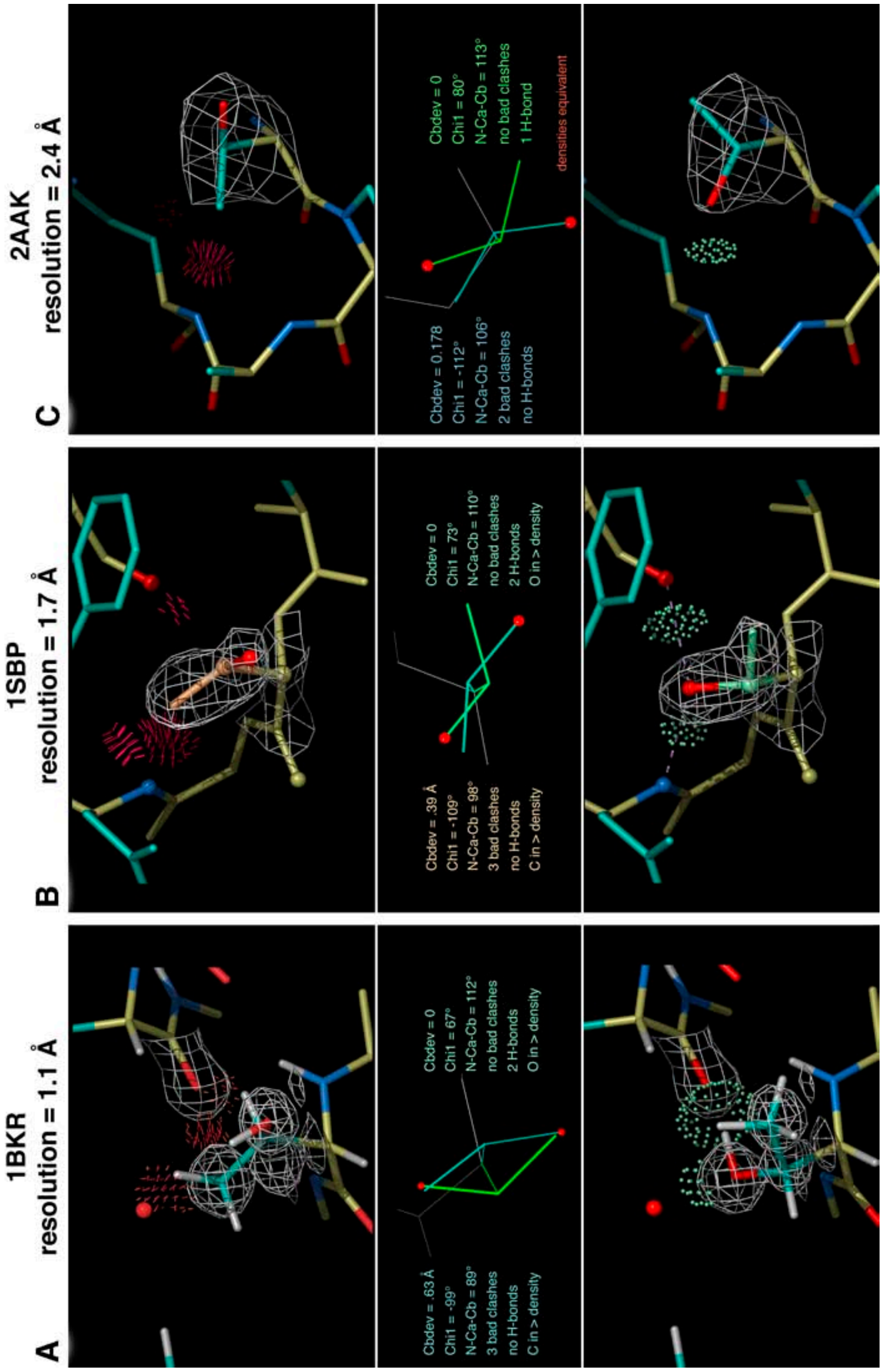
This process will be illustrated here by three examples of backwards Thr residues at different resolutions: 1BKR[38] Thr101 at 1.1Å, 1SBP[39] Thr32 at 1.7Å, and 2AAK[40] Thr3 at 2.Å. (Such problems are neither unusual nor reprehensible; these are well-determined structures, and two of them have lower-than-average clashscores for their resolution. The clashes are really obvious with our tools, but could not have been seen with classic non-H methods.) This process of diagnosis and correction is needed more often at lower resolution, but precise data and tighter coupling of map and model at high resolution make it clearer

[37] C. Lee and S. Subbiah, *J. Mol. Biol.* **217**, 373-388 (1991).

[38] S. Banuelos, M. Saraste, and K. D. Carugo, *Structure* **6**, 1419-1431 (1998).

[39] J. J. He and F. A. Quiocho, *Protein Sci.* **2**, 1643-1647 (1993).

[40] W. J. Cook, L. C. Jeffrey, M. L. Sullivan, and R. D. Vierstra, *J. Biol. Chem.* **267**, 15116-15121 (1992).


A
1BKR
resolution = 1.1 Å
Cbdev = .63 Å
Chi1 = -99°
N-Ca-Cb = 89°
3 bad clashes
no H-bonds
C in > density
Cbdev = 0
Chi1 = 67°
N-Ca-Cb = 112°
no bad clashes
2 H-bonds
O in > density
B
1SBP
resolution = 1.7 Å
Cbdev = .39 Å
Chi1 = -109°
N-Ca-Cb = 98°
3 bad clashes
no H-bonds
C in > density
Cbdev = 0
Chi1 = 73°
N-Ca-Cb = 110°
no bad clashes
2 H-bonds
O in > density
C
2AAK
resolution = 2.4 Å
Cbdev = 0.178
Chi1 = -112°
N-Ca-Cb = 106°
2 bad clashes
no H-bonds
Cbdev = 0
Chi1 = 80°
N-Ca-Cb = 113°
no bad clashes
1 H-bond
densities equivalent

what is happening.

Figure 11 lays out the information for these three Thr cases in columns A, B, and C. The top row shows the 2Fo-Fc electron density contours and the all-atom contacts (just clashes and H-bonds) for each Thr in the deposited structures. The Cβ deviations are significant and the methyls clash badly in each case, often with polar atoms; this suggests that the Oγ might be more suitable in that position. For each case the Cβ was idealized and a new sidechain was added in ideal geometry using the interactive MAGE/PROBE system; all three Thr rotamers were tried and the best one had its χ_1 angle and OH orientation optimized for contacts, for H-bonding, and for overall position near the original coordinates. The bottom row of Fig. 11 show the new H-bonds, excellent all-atom contacts, and good fit to density in the resulting position, obtained with ideal geometry, good rotamers, and no backbone movement. Those statistics are given in the center panels of Fig. 11, and the original and corrected sidechains are shown superimposed. Note that the pairs of sidechain models occupy approximately the same regions of space, so that it is easy to see how both might fit into somewhat ambiguous electron density. In order to fit the density given the initial choice of a backwards rotamer, however, the Cβ positions had to be significantly distorted in all cases; the distortions are actually greater at higher resolution, because the diffraction data are, quite correctly, given a stronger weight relative to geometry. In each case the new version is a better, or at least equivalent, fit to the density even without further refinement. There is simply no question that the original conformations were trapped in the wrong local minimum and that the change is an improvement. Such changes, when refined, lower the residual somewhat and lower the free-R even more (un-

[41] J. M. Word, R. C. Bateman, Jr., B. K. Presley, S. C. Lovell, and D. C. Richardson, *Protein. Sci.* **9**, 2251-2259 (2000).

[42] W. L. DeLano, http://www.pymol.org, Delano Scientific, site hosted by Sourceforge.net, 2002.

Fig. 11. The evidence and process of diagnosing and repairing backward-fit Thr sidechains at 3 different resolutions, in (A) the 1BKR[38] actin-binding domain at 1.1Å, (B) the 1SBP[39] sulfate-binding protein at 1.7Å, and (C) the 2AAK[40] ubiquitin-conjugating enzyme at 2.4Å. *Top row:* Models as deposited [plus the added H atoms in (A)], the 2Fo-Fc sidechain density, and the all-atom clashes (red spikes) and H-bonds (none). In each case, the interactive MAGE/PROBE system[41] was used to idealize the Cβ, try all rotamers, and optimize the one with the best contacts and overlap with the original model. *Middle row:* Original and revised models superimposed, with statistics comparing their quality before and after; all measures improve strongly. *Bottom row:* Revised ideal-geometry models, their new H-bonds (pale green dots), and the 2Fo-Fc densities calculated from the new models (the densities do not change). Top and bottom panels produced in PyMol.[42]

published results, 2002).

As a side note, the high-resolution density in Fig. 11A (calculated with isotropic B's) is quite unambiguous, with a small but clear peak at the position of the correct Cβ. The reason this was not noticed is probably because the map calculated with anisotropic B-factors, as deposited, smears the density equally between the correct and the incorrect Cβ positions, thus obscuring the problem.

Cases equivalent to those in Fig. 11 are also found for Val and Ile. For Leu and Met the analysis is a bit more complex and less dramatic but basically very similar (see Lovell *et al.*,[14] and the discussion of Fig. 14 below). These situations where a sidechain has been fit backwards are only one subset of the possible problem conformations. They are well worth fixing because they have large consequences for atom positions, and they are easy to correct as well as to diagnose with the new all-atom contact tools, in either numerical or displayed forms.

At the surface: multiple-conformation sidechains; waters

On average, structures are considerably less well-defined at the molecular surface, and in many cases the information is simply not there to define correctly the ensemble of positions for a mobile loop or a highly-disordered sidechain. However for intermediate cases with some suggestive density, the combination of rotamer and all-atom contact information can help substantially in assigning a good model. High-resolution structures show that sidechains nearly always occupy discrete and relaxed conformations (i.e., rotamers) even in the core, in spite of the many favorable or unfavorable interactions which could potentially pull or push the equilibrium conformation away from that local optimum. Therefore, in a partially-disordered surface position where the constraints are necessarily weaker, there is no justification for fitting a sidechain conformation as significantly non-rotameric or, of course, as having a serious clash.

The suggested strategy, therefore, is to cycle through the best available rotamer library (see above for our current recommendation) doing modest local torsion adjustments for each; to reject those that have irreconcilable clashes with reliable parts of the surrounding structure; and to accept a small number of conformations (two, or perhaps three) for which there is evidence in the density, giving some preference to any that can make favorable H-bond or van der Waals interactions. Their relative occupancies can then be refined, discarding any that behave badly, and then allowing the positions to shift. At high resolution the backbone can be refined separately for alternate conformations (or, as an approximate surrogate, the Cβ positions), but otherwise the Cβ should be kept common. Such a strategy is easier than unconstrained hand-fitting through the confusing density produced by multiple conformations, and it produces more physically reasonable models. It is also, we hope, potentially automatable.

An even more pervasive difficulty at the molecular surface is distinguishing solvent peaks from partially ordered sidechains from noise peaks. All-atom contacts help by showing explicitly the H-bonds and contacts made by proposed waters. Ones that make no interactions at all are suspect (if symmetry-related molecules have been considered), while ones with one or more good H-bonds are almost certainly correct. The most interesting cases are waters that sit too close to non-polar sidechain atoms, such as the high-resolution example in Fig. 12A, where two waters clash badly with an Asp methylene hydrogen. Sometimes such "water" peaks can be caused by the next atoms in an unmodeled alternate sidechain conformation, whose covalently-bonded distance from the methylene C is, of course, much shorter than the possible distance for a water. The density peaks for the clashing waters in Fig. 12A are in almost exactly the right positions to fit the two Oδ atoms in the most common Asp rotamer, as shown in Fig. 12B, and there is even density for the Cγ. Surprisingly, in other cases it often happens that the water peak is round and well-defined, with relatively high occupancy and low B, while the details of the sidechain conformation are less reliable, with higher B values and/or more ambiguous density. Fig. 12C shows such a case, where the two χ angles flanking the clashing Lys Cε can each be changed slightly to swing the methylene sideways enough to relieve the clash and also fit the low but clear electron density better (Fig. 12D).

In both the above cases, and with solvent problems in general, the electron density is the final arbiter, with the help of H-bond and rotamer considerations. However, without the all-atom clashes one would not have known which waters needed careful attention.

Ligand binding and other molecular contacts

The contacts between a protein and a small-molecule ligand, or between two macromolecules, are completely analogous to the atomic contacts inside a molecule and are treatable in the same way. REDUCE can add and optimize hydrogens on nucleic acids as well as on proteins, and on small-molecule "heterogens" using the PDB het atom dictionary (to which the specifications for a novel ligand can be added if necessary).

There are two quite different uses for all-atom contact information at such interfaces. One is to identify and fix any conflicts that indicate fitting errors, which is important since small molecules are on average less well-deter-

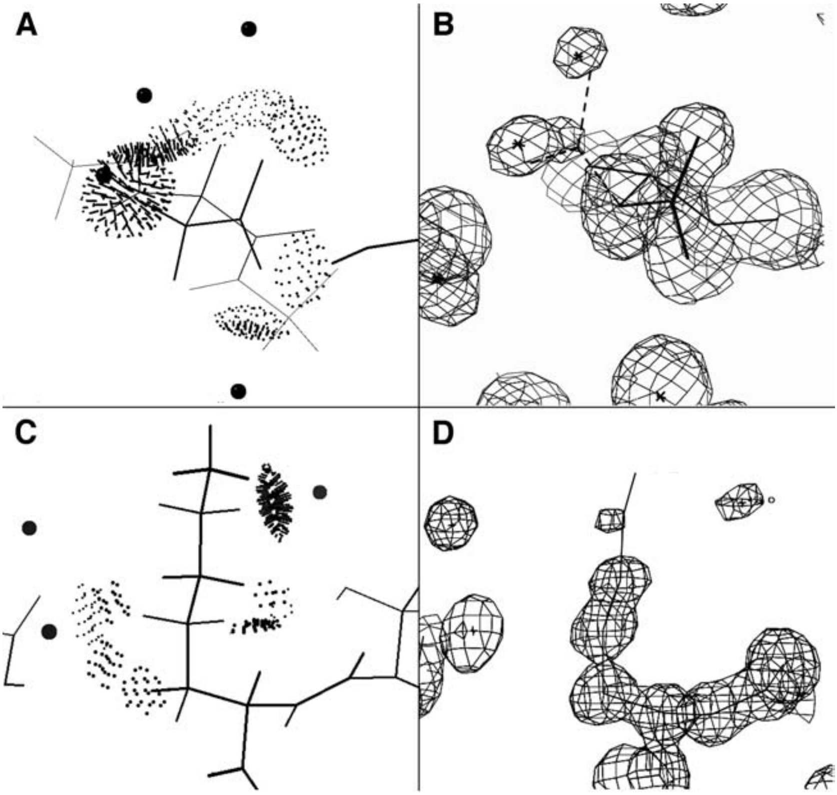


Fig. 12. Two different kinds of clashes between surface sidechains and waters [black balls in (A) and (C); asterisks in (B) and (D). (A) Asp-9 of 1EB6[34] makes four H-bonds (dots at right), two of them to the upper and lower waters, but its Cβ and Hβ clash seriously with the two waters at left. (B) The dotted-line model shows that those two waters are very close to the positions expected for the 2 Oδ atoms of the commonest Asp rotamer, and there is even density for the Cγ. The density peaks of those two "waters", therefore, almost certainly represent an alternate conformation of the Asp. The backbone is probably slightly different for this second conformation, allowing the Cβ and atoms beyond to shift up and left to match the density even better. (C) Lys 184 of the 1CXQ[43] ASV integrase core has a water seriously clashing with an ε methylene H. (D) The electron density shows that this time the water is probably correct and the sidechain somewhat displaced, since the Cε density is weak and shifted left, away from the water. Figure produced in MAGE.[6,7]

mined than parts of the protein with similar B-values, presumably because their properties and parameters are less familiar to either the crystallographers or the refinement protocols. The second use is to characterize, either visually or numerically, the atom-by-atom details of the interface and to analyze the conse-

[43] J. Lubkowski, Z. Dauter, F. Yang, J. Alexandratos, G. Merkel, A. M. Skalka, and A. Wlodawer, *Biochemistry*. **38**, 13512-13522 (1999).

quences of mutations or of ligand modifications. Interface mutations, especially, can be easily and effectively characterized with the interactive MAGE/PROBE system described below.

Backbone: protein and nucleic acid

Polypeptide backbone outside of regular secondary structure is inherently harder to model or modify than sidechains, since it is connected in three directions at each residue. On the other hand, those constraints make large-scale mistakes less likely, and there are usually fewer mainchain-mainchain clashes than sidechain-sidechain clashes in a given protein structure. Those all-atom clashes, large backbone bond-angle distortions, or ϕ,ψ values in forbidden regions of the applicable Ramachandran plot (see above) are indicators of backbone problems. High-B regions are, as always, difficult to fit correctly. Glycine, with no information contribution from a Cβ, is the amino acid with the highest risk of backbone error; an example is shown in Fig. 13 of Word *et al.*[4] Abrupt changes of backbone direction (aside from the well-understood H-bonded "tight turns") fairly often show clashes, especially at relatively low resolution where the electron density across a close mainchain contact may be continuous and make the two sides look even closer than they actually are. We have not yet developed user-friendly or automatic tools for correcting backbone; for now, hand-rebuilding should seek to idealize bad angles or torsions and pull apart close approaches that have bad clashes. Often a concerted motion of two consecutive peptides can correct a local problem with minimal displacement of the surrounding structure.

Diagnosis and correction of backbone clashes is even more important for nucleic acids, especially for the complex, non-repeating conformations common in large, biologically significant RNA molecules. The bases, and especially base pairs, produce clear, easily interpretable electron density even at 2.5-3Å resolution, and their stacking and H-bond interactions are typically determined very accurately. Base-base all-atom contacts nearly always look really excellent. Nucleic acid backbone, on the other hand, is very difficult to fit correctly because it has 6 degrees of freedom per residue, with only the phosphate as a clear marker whose atom(s) can be positioned accurately at essentially any resolution from the local density height and shape. Adding on the H atoms and considering their contacts adds valuable additional constraints. Figure 13 illustrates the typical situation for a short section of RNA in the 1JJ2[44] ribosome structure, where all-atom contacts are excellent for the bases and for the backbone of one nucleotide, but the following nucleotide in a similar overall conformation shows large, physically impossible overlaps (indicated by spikes rather than

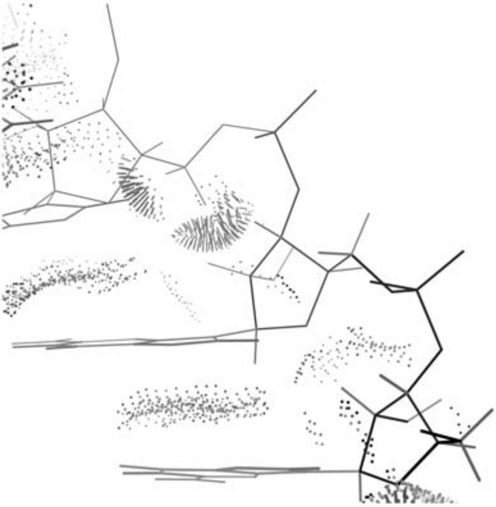


Fig. 13. Residues 544-545 of the 23S RNA from the 1JJ2[44] 50S ribosome, illustrating the use of all-atom contact analysis for nucleic acids. The base-base contacts are excellent, as is typical in RNA or DNA structures because the bases can readily be positioned accurately. Backbone, however, with 6 variable torsions per residue, causes problems more frequently. Here the contacts are very good for the lower nucleotide, but the upper nucleotide has severe clashes around C5´. The overall conformations are quite similar (both near A-form), but probably two adjacent backbone torsions are somewhat wrong in the latter case, swinging the C5´ out of place.

dots) that mean the detailed local conformation must be incorrect. Such analysis can quickly find the problem areas in even a very large RNA molecule, to help concentrate rebuilding efforts where they are most needed. We hope soon to tabulate a library of backbone rotamers for nucleic acids, giving a broad set of trial replacement structures to substitute for clashing conformations of similar overall shapes, such as the pair in Fig. 13.

Available Tools and Modes of Use

The most basic and flexible way to do all-atom contact analysis is to download the programs and run them from the command line in Linux or Unix. They are all available from our web site,[15] including documentation and examples; open source is also available for modifications or for other compilations. Reduce is written in C++ and the other programs are in C. Utility scripts

[44] D. J. Klein, T. M. Schmeing, P. B. Moore, and T. A. Steitz, *EMBO J.* **20**, 4214-4221 (2001).

such as Flipkin or Clashlist are in Perl or Awk. The display program MAGE and its input utility PREKIN are available also for Mac OS9 or OSX and for Windows operating systems, and either Java or KiNG MAGE that provides 3D display directly on the Web.

Interactive MAGE/PROBE for testing changes

The remote update function in MAGE provides a convenient and powerful system for evaluating the consequences of an isolated change (a mutation, or a conformational change for one or a small cluster of sidechains) within the context of the surrounding structure.[41] While looking at a kinemage display of contacts for a whole structure, the user can ask PREKIN to set up rotations for an idealized and/or mutated sidechain, including a hypertext list of rotamers added to the text window. Then PROBE is asked to update the display of all-atom contacts as the rotamer is changed or the individual χ angles adjusted. This system was used to obtain the corrected Thr conformations in Fig. 11A-C. It can be run in Linux, Unix, Windows, or Mac OSX operating systems.

Use while rebuilding in O or XtalView

Simplified versions of the all-atom contact display for clashes and H-bonds can be invoked while doing model-to-map fitting in O[10] or XtalView,[16,17] giving the ability to evaluate contact and electron density information at the same time. PROBE, and for O REDUCE, must be installed and are called on demand to produce the updated contact display. Macros and instructions for doing this are available at http://kinemage.biochem.duke.edu. Figure 14 shows an example of such use, at an intermediate stage in refinement of a Trp tRNA synthetase complex, then at 2.9Å resolution. The original conformation and a suggested revision for the Met 193 sidechain fit the density about equally well, but the atomic clashes (and the rotamer quality) differentiate between them quite unambiguously. This conformational change was validated by three later datasets at 1.7Å resolution[45] and used in the 1I6K, 1I6L, and 1I6M coordinates.

The MolProbity web server for structure validation

Prominently featured on the kinemage web site is a service called MOLPROBITY that lets the user run our programs and scripts on an uploaded PDB-format file. The results are presented in tabular form, as rotatable kinemages in Java MAGE, or as files for downloading. Geometric validation tools evaluate ϕ,ψ

[45] P. Retailleau, Y. H. Yin, M. Hu, J. Roach, G. Bricogne, C. Vonrhein, P. Roversi, E. Blanc, R. M. Sweet, and C. W. Carter, *Acta Crystallogr. D Biol Crystallogr.* **57**, 1595-1608 (2001).

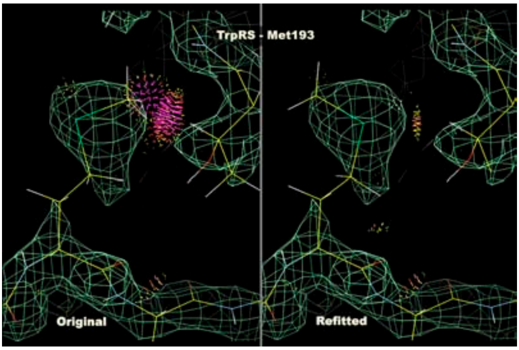


Fig. 14. Using all-atom clashes and H-bonds during model rebuilding in O,[10] with an example from Trp tRNA synthetase[45] when still at 2.9Å resolution. The Met-193 electron density has a gap between Cβ and Cγ and is round for the sulfur, not showing Cε. The original conformation fits reasonably well, but is not a good rotamer and has several bad clashes (red spikes). Trying good rotamers[14] gave a revised conformation that fits the density equally well and fits very much better into the surrounding model. Higher-resolution data confirmed that the change was correct.[45]

values, rotamers, and Cβ deviations, as described above. Reduce is run to add and optimize H atoms, and Probe to calculate all-atom contacts; sidechain contacts and backbone-backbone contacts are each viewable in separate Java kinemages. The flipkin script is run to produce kinemages for evaluating the proposed flips of Asn, Gln, and His sidechains; if the user disagrees with any of Reduce's decisions, the process can easily be re-run with the desired changes. MolProbity is the easiest way to try out these new methods, and it provides a detailed and thorough analysis of macromolecular crystal structures either during or at the end of refinement.

Discussion

Detailed and critical analysis of the contacts and geometry inside macromolecules shows that the structures are remarkably relaxed and at the same time remarkably well packed. Most interior atoms make contacts at near-ideal distances, sidechains are closely rotameric, methyls are well-staggered, and bond angles distort only slightly. The great majority of observed large deviations from ideality (eclipsed χ angles, bond angles 10° off, etc.) show the clear hallmarks of errors. Local regions with significantly strained conformations do indeed occur, but they are quite rare and usually involve either unusual ligands or else places where the unusual conformation is needed for biological function. Atomic clashes greater than 0.5Å are not physically possible, while large geometrical distortions should only be accepted when the data truly rule out a more regular conformation and when there are either compensating interactions or reasons for the strain.

Depending on the nature of the problem and the way refinement is handled, fitting errors may result either in all-atom clashes, or in bond and torsion angle distortions, or in both. However, it is very difficult to get any part of the molecular interior in the wrong local minimum without it showing up clearly by one of those criteria. Such problems often will show up in difference maps but, surprisingly, often do not. Therefore, a powerful addition to crystallographic fitting and refinement is the combined use of all-atom contact and geometrical validation tools, especially since that combination can often show how to correct the problems. Many of the benefits of all-atom contact analysis could be obtained by using all hydrogens and their contacts directly in refinement, which is occasionally now done at very high resolution but might actually prove quite useful at lower resolution where the extra constraints are more needed. However, it is also a considerable advantage to keep these criteria independent of the refinement target function, so that their evaluation stays unbiased and sensitive. Therefore, we currently recommend that refinement usually be done without explicit H atom contacts but that all-atom contact analysis be consulted periodically either in conjunction with, or as a partial replacement for, manual rebuilding. Then at the end, a structure can be checked one more time before deposition. In either case, the emphasis would be on cases where a conformation is in the wrong local minimum and where an improved fitting can be suggested.

In obtaining such improved fittings for sidechains in the wrong local minimum at high resolution, it turns out to be surprisingly effective to leave the backbone unchanged, idealize the sidechain geometry, and choose the best-fitting conformation near a good rotamer. Faced with the task of optimizing an incorrect rotamer to the diffraction data, refinement algorithms seldom produce the needed shift of Cβ by changing ϕ and ψ (presumably because ignoring off-diagonal terms in the matrix makes it difficult to achieve such concerted motions), but instead simply distort bond angles. This behavior is actually fortunate, however, because it makes the correction of such misfittings quite easy. At lower resolution, it is more often necessary to shift backbone when correcting sidechains. While coordinates obtained from Asn/Gln/His flips are quite accurate, the results of any of the other correction procedures described here are expected to be close but not optimal and should always be submitted to further cycles of refinement.

In the near future we plan to further combine multiple validation criteria for unified presentation, as lists tied to the 1-D sequence, as 2-D interaction plots, and as graphics shown jointly on the 3-D structure. We are also developing updated validation criteria for nucleic acids and better methods for adjusting backbone. Future directions for the development of all-atom contact analysis center first of all on further automation, both for convenience of use and also for suitability in application to the high-throughput structure determination needed in structural genomics. At high resolution, these methods should be able to validate where they already exist, or to produce where they are close, "superstructures" with no atomic clashes, with near-ideal geometry, and with free R as good or better than before. At lower resolution, it should be possible to bring the structures significantly closer to what would have been found at high resolution.

Acknowledgements

This work was supported by NIH grants GM-15000 and GM-61302.