Protein Science (1993), *2*, 1511-1519. Cambridge University Press. Printed in the USA.

Verification of protein structures: Patterns of nonbonded atomic interactions

CHRIS COLOVOS AND TODD O. YEATES

Department of Chemistry & Biochemistry and Molecular Biology Institute, University of California, Los Angeles, California 90024-1569

(RECEIVED December 8, 1992; REVISED MANUSCRIPT RECEIVED June 17, 1993)

Abstract

A novel method for differentiating between correctly and incorrectly determined regions of protein structures based on characteristic atomic interactions is described. Different types of atoms are distributed nonrandomly with respect to each other in proteins. Errors in model building lead to more randomized distributions of the different atom types, which can be distinguished from correct distributions by statistical methods.

Atoms are classified in one of three categories: carbon (C), nitrogen (N), and oxygen (O). This leads to six different combinations of pairwise noncovalently bonded interactions (CC, CN, CO, NN, NO, and OO). A quadratic error function is used to characterize the set of pairwise interactions from nine-residue sliding windows in a database of 96 reliable protein structures. Regions of candidate protein structures that are mistraced or misregistered can then be identified by analysis of the pattern of nonbonded interactions from each window.

Keywords: crystal structure; nonbonded atomic contacts; structure verification

Since the creation of the Brookhaven Protein Data Bank, over 800 structures have been deposited. Although improvements in crystallographic methods and computing power have reduced the time necessary to determine a protein crystal structure, the problem of independent evaluation of model reliability has only begun to be addressed. Limited diffraction resolution and poor phases frequently lead to electron density maps that are difficult to interpret. Preliminary protein models built into ambiguous maps often contain errors of various types that must be corrected during the course of model building and refinement. The different types of errors can be arranged in decreasing order of severity, as follows:

1. Mistracings of residues due to backbone connectivity errors;
2. Misalignments or misregistrations of residues; and
3. Misplacements of side chains.

Attention has become focused on the problem of evaluating the correctness of protein structures since the recent discoveries of serious errors in a number of published structures (Ghosh et al., 1982; McClairen et al., 1986; Chapman et al., 1988; de Vos et al., 1988; Navia et al., 1989).

Several methods of evaluation have been proposed. Of these, the Ramachandran analysis of peptide dihedral angles (Ramachandran & Sasisekharan, 1968) was one of the first to classify allowed and nonallowed conformations; most grossly misfolded structures can be identified in this fashion. Several other methods that physiochemically characterize a protein structure have been described, including the energetics methods of Novotny et al. (1984), the atomic solvation parameter of Eisenberg and McLachlan (1986), and the evaluation of surface polarity (Baumann et al., 1989). Other empirical approaches include analysis of the pairwise likelihood of neighboring residues (Tanaka & Scheraga, 1976; Hendlich et al., 1990), the three-dimensional-one-dimensional profile method of Lüthy et al. (1992), and the fragment matching methods of Jones et al. (1991).

In our study, the six types of noncovalently bonded atom-atom interactions (CC, CN, CO, NN, NO, and OO) in protein crystal structures are considered; we show that the different types occur with nonrandom frequencies in proteins. A classification method based upon this idea will be described and shown to be useful in identifying regions of preliminary models that require adjustment.

Reprint requests to: Todd O. Yeates, Department of Chemistry & Biochemistry, University of California, 405 Hilgard Avenue, Los Angeles, California 90024-1569.

Methods

Database criteria

The database of 96 correct protein structures consists of high-resolution crystal structures selected from the Brookhaven Protein Data Bank (Bernstein et al., 1977; see Table S1 on the Diskette Appendix). The criteria used for selection are: (1) a resolution of 2.5 Å or better, (2) an R-factor of less than 25%, (3) a monomeric or homo-oligomeric structure, (4) the exclusion of prosthetic groups, and (5) good geometry defined by ω, the dihedral angle made by the peptide bond, less than $\pm 15°$ from ideality. Effort was made to include examples from many classes of protein structures (Boberg et al., 1992).

Classification of atomic interactions

If the atoms of a structure are classified as carbon (C), nitrogen (N), or oxygen/sulfur (O), then this gives rise to six distinct interaction types (CC, CN, CO, NN, NO, and OO). Assessment of the nonbonded interactions is subject to the following restrictions: (1) the distance between the two atoms in space is less than some preset limit, typically 3.5 Å, and (2) atoms within the same residue or those that are covalently bonded to each other are not considered. For each protein with n atoms, the fractions of the interactions, f(interaction), are calculated. For example, f(CC) represents the fraction of all pairwise interactions that are of the CC type:

$$f(\mathrm{CC}) = \frac{n_{\mathrm{CC}}}{n_{\mathrm{CC}} + n_{\mathrm{CN}} + n_{\mathrm{CO}} + n_{\mathrm{NN}} + n_{\mathrm{NO}} + n_{\mathrm{OO}}}. \quad (1)$$

Fractions for each interaction type, f, were calculated for each database structure, and the average and the standard deviation of each variable were evaluated (Table 1). To test the correctness of the local structure, a nine-residue sliding window was used. In each window a similar determination of nonbonded interactions was made, where the fraction of each interaction type was calculated by an expression similar to Equation 1 over a nine-residue span. Two additional restrictions are applied: (1) at least one of the two interacting atoms must belong to the window, and (2) an empirical lower limit on the number of interactions is used to screen out any segments that may lie adjacent to structural deletions or in very mobile loops. The averages and standard deviations of these fractions were calculated for nine-residue windows from the entire database of "reliable" structures.

Statistical methods

In order to maximize the discrimination between correct and "initial" structures, information from all six fractional parameters is used. The result from the ith nine-residue window is treated as a six-dimensional vector or observation, $\mathbf{y}_i$, where

$$\mathbf{y}_i = (f(\mathrm{CC}), f(\mathrm{CN}), f(\mathrm{CO}), f(\mathrm{NN}), f(\mathrm{NO}), f(\mathrm{OO}))_{i-4}^{i+4} \quad (2)$$

represents the vector of f(interaction) spanning a nine-residue range centered on the ith residue. Two methods for classifying these observations were evaluated: a Gaussian error function and a convex approximation to the set (not discussed here).

For an n-dimensional normal distribution, the probability function, $P(\mathbf{x}_i)$, takes the following form:

$$P(\mathbf{x}_i) \propto e^{-\mathbf{x}_i{}^{\mathrm{T}}\mathbf{B}\mathbf{x}_i}, \quad (3)$$

where $\mathbf{B}$ is a symmetric positive-definite matrix that provides a quadratic description of an elliptical error function (see Appendix). In our case, $\mathbf{y}_i$ is the vector whose coordinates represent the set of linearly independent fractions of the different types of interactions, as defined by Equation 2 and $\mathbf{x}_i = \mathbf{y}_i - \bar{\mathbf{y}}$. When using fractional values, only five of the six parameters are independent. The appropriate matrix $\mathbf{B}$ is calculated from the distribution of vectors, $\mathbf{x}_i$, from the database (see Appendix). The classification problem then reduces to the matrix multiplication of $\mathbf{x}_i{}^{\mathrm{T}}\mathbf{B}\mathbf{x}_i$ for each window. In order to account for the possible non-Gaussian nature of the distribution, the 95% confidence limit for the error term, $\mathbf{x}_i{}^{\mathrm{T}}\mathbf{B}\mathbf{x}_i$, is obtained empirically. This confidence limit must of course be interpreted with caution when applied to structures determined at lower resolution, since low-resolution structures are not represented in our database.

Results and discussion

The fractions of nonbonded pairwise interactions within a specified distance limit were calculated for each of 96 reliable protein structures (Table S1, Diskette Appendix). These fractions were calculated at a variety of different distance cutoffs (3.00–4.75 Å), and the average and standard deviation (data not shown) for each interaction type at every distance limit were calculated (Table 1). When the average fractions are normalized according to the relative abundance of C, N, and O atoms, distinct preferences are indicated by the *nonrandom* frequencies of the different interaction types (Table 1). Preference ratios are defined as the observed fractions divided by the expected fractions calculated from the relative abundance of C, N, and O; a ratio of unity would represent random association of atom types, or the lack of any preference. For distance limits of 3.5 Å and less, CC, NO, and NN interactions are more abundant than predicted for a random association of atoms, while CN and OO fractions are disfavored; CO behaves randomly. The elevated value of f(NO) must re-

Table 1. *Pairwise association of different atom types*[a]

Distance (Å)	Total Number of Interactions		Observed Fractions			Expected Fractions			Preference Ratios		
			C	N	O	C	N	O	C	N	O
3.00	120640	C	0.263	0.019	0.275	0.168	0.228	0.255	1.565	0.083	1.077
		N		0.114	0.310		0.078	0.174		1.470	1.787
		O			0.019			0.097			0.196
3.25	178483	C	0.251	0.097	0.269	0.188	0.233	0.259	1.333	0.416	1.038
		N		0.082	0.276		0.072	0.160		1.137	1.722
		O			0.026			0.089			0.292
3.50	263006	C	0.225	0.142	0.281	0.191	0.227	0.265	1.181	0.624	1.062
		N		0.071	0.237		0.068	0.158		1.046	1.501
		O			0.044			0.092			0.479
3.75	390828	C	0.247	0.142	0.308	0.223	0.218	0.281	1.109	0.651	1.095
		N		0.067	0.186		0.053	0.138		1.256	1.351
		O			0.051			0.089			0.574
4.00	550313	C	0.279	0.147	0.331	0.268	0.203	0.297	1.040	0.724	1.113
		N		0.050	0.145		0.038	0.113		1.302	1.289
		O			0.049			0.082			0.595
4.25	726713	C	0.284	0.164	0.341	0.288	0.199	0.297	0.987	0.824	1.147
		N		0.043	0.121		0.034	0.103		1.250	1.177
		O			0.046			0.077			0.600
4.50	906896	C	0.321	0.168	0.322	0.320	0.200	0.290	1.002	0.841	1.109
		N		0.041	0.103		0.031	0.091		1.316	1.138
		O			0.044			0.066			0.669
4.75	Not Calculated	C	0.359	0.174	0.295	0.352	0.202	0.281	1.019	0.862	1.051
		N		0.038	0.090		0.029	0.080		1.315	1.119
		O			0.044			0.056			0.787

[a] The expected values of the six interaction types (CC, CN, CO, NN, NO, and OO) can be obtained from the calculated fractions of the three atom types (C, N, and O). The ratio of observed over expected yields the preference ratios (a ratio of 1 indicates independent association of atom types). The number of interactions is the total number of unique pairwise interactions in the database of reliable structures.

flect extensive hydrogen bonding, especially in the backbone, and the low f(OO) value may be rationalized by repulsion between fully or partially charged oxygen atoms. Other deviations from unity may be a result of different bonding geometries for C, N, and O. As the distance limit is increased, all of the fractions except for NN tend to a randomized state (i.e., a ratio of unity).

The observation that different atom types are distributed nonrandomly with respect to each other suggested that incorrect structures might be discriminated on this basis. The fractions of interactions for four initial protein models that contained structural errors were calculated under similar conditions. Results for the CC and NO interaction fractions, f(CC) and f(NO), are shown in Figure 1A and B. All four of the initial models that were tested fell two or more standard deviations from the average value for the f(CC) and f(NO) at most distance limits tested, while f(CN), f(CO), f(NN), and f(OO) did not clearly discriminate between structurally correct and incorrect structures (data not shown). It is logical to assume that NO interactions are a measure of hydrogen bonding between atoms, but the exact meaning of the CC term is unclear. In a correct structure, the fraction of hydrogen bonded atoms is optimized (Baker & Hubbard, 1984; Stickle et al., 1992), and in an incorrect structure one would expect the fraction of hydrogen bonded atoms to be lower.

A windowing algorithm was used as a test of whether the method would apply to small regions of the protein. The fractions of interactions for each nine-residue

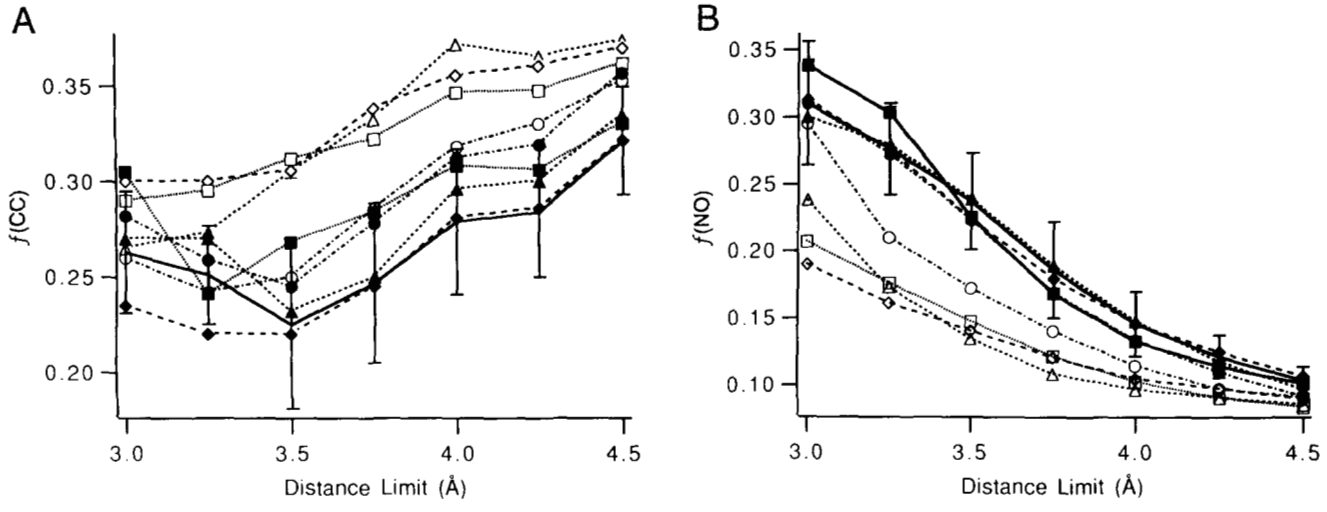

Fig. 1. A: Plot of f(CC) as a function of interatomic distance cutoff limit (see text). **B:** f(NO) as a function of distance cutoff. Error bars represent two standard deviations from the average over the database. Initial and final structures are represented by hollow and solid markers, respectively (circle, rubisco small subunit; triangle, ferredoxin; diamond, *Eco*RI; square, HIV-1).

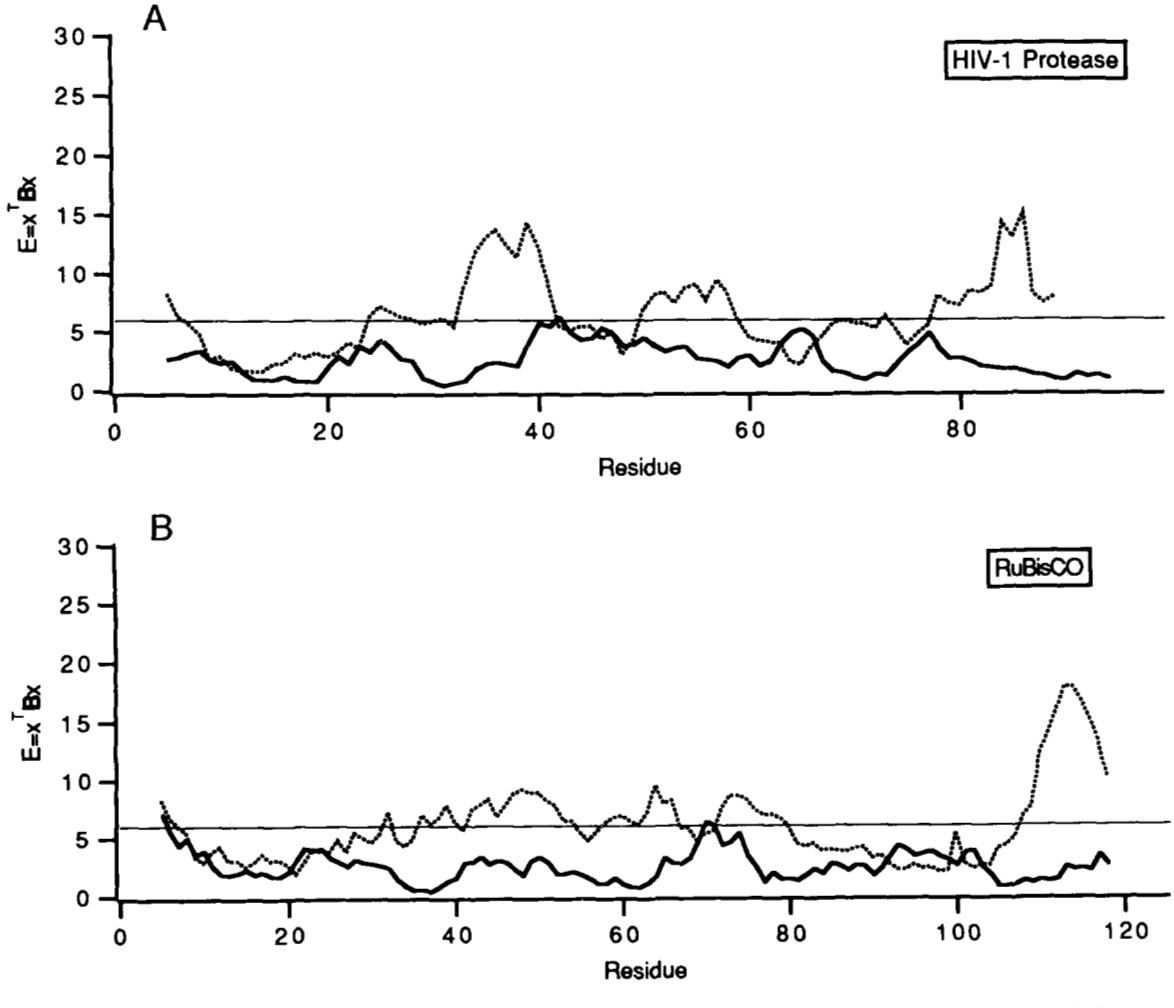

Fig. 2. Plots of the error function, $\mathbf{E} = \mathbf{x}_i^{\mathrm{T}}\mathbf{B}\mathbf{x}_i$, in a nine-residue sliding window. The solid bold line represents the final structure, the broken line the initial structure, and the thin solid line the 95% confidence limit. **A:** HIV-1 protease. **B:** Rubisco small subunit. **C:** *Eco*RI. **D:** PRAI-IGPS. (*Continues on facing page.*)

window for each structure in the 96-structure database (Table S1, Diskette Appendix) were determined, as were those for several initial structures (to be discussed later), at predetermined distance cutoff limits. The average and standard deviation for each interaction type in the database were calculated for each distance limit and were nearly identical to the values calculated over whole structures (Table 1). Each interaction type was plotted as a function of the residue at the window center, for each of the initial structures. As anticipated, the incorrect regions of the initial models showed decreased $f(\mathrm{NO})$ ($f(\mathrm{NO}_{\mathrm{structure}}) < f(\overline{\mathrm{NO}}_{\mathrm{database}}) - 2\sigma_{\mathrm{NO}}$) as well as elevated $f(\mathrm{CC})$ ($f(\mathrm{CC}_{\mathrm{structure}}) > f(\overline{\mathrm{CC}}_{\mathrm{database}}) + 2\sigma_{\mathrm{CC}}$) over the incorrect regions (not shown).

As described in Methods, the best discrimination between correct and incorrect patterns of atomic interactions was obtained by analyzing all six interaction types simultaneously. The entire set of observations formed by the nine-residue windows from the database of reliable structures (Table S1, Diskette Appendix) is characterized as a five-dimensional normal distribution (the sixth fractional value is not independent of the first five). The statistics were evaluated at distance cutoff limits ranging from 3.00 to 4.75 Å, in 0.25-Å intervals.

The 95% confidence limit for the distribution was determined empirically and was found to correspond to an error function ($\mathbf{x}_i^{\mathrm{T}}\mathbf{B}\mathbf{x}_i$) of approximately 6. Several initial and final models were tested to verify the method. The best identification of incorrect regions was obtained with distance cutoffs ranging from 3.00 to 3.75 Å.

An initial model of the human immunodeficiency virus (HIV) protease dimer (Navia et al., 1989) contained structural errors near the C-terminus. Our program shows error function values much greater than the 95% confidence limit for the C-terminus (residues 80–99) as well as residues 30–40. The corrected structures, 4HVP (Wlodawer et al., 1989) (Fig. 2A) and 5HVP (Fitzgerald et al., 1990), display all but 2% of residues under the 95% confidence limit for a distance cutoff of 3.50 Å.

A large fraction of the small subunit of the initial model of rubisco (Chapman et al., 1988) was built in the reverse direction (Schreuder et al., 1990). In our program, the results at distance limits of 3.00–3.75 Å clearly indicate errors in this subunit (Fig. 2B). The final model for the

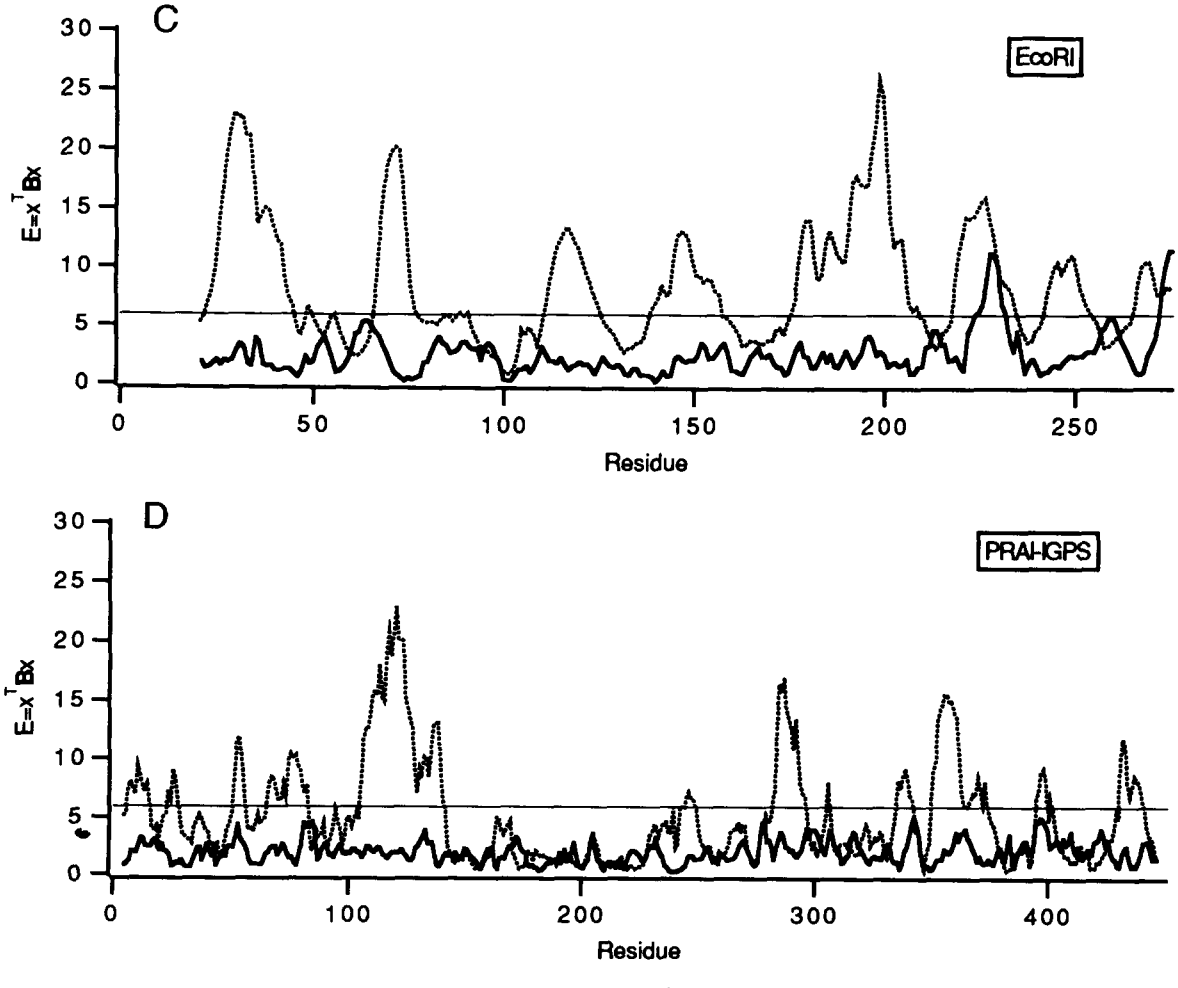

Fig. 2. *Continued.*

rubisco small subunit (Curmi et al., 1992) shows only 2% of the windows outside the 95% confidence limit at a distance of 3.50 Å.

The preliminary model of the endonuclease *Eco*RI (McClairen et al., 1986) contained several structural errors and misconnections. At distance limits of 3.00–3.75 Å, all of the incorrect regions are identified as suspect (residues 20–50, 60–170, 208–233, 240–C-terminus), while all but 5% of the corrected structure (Kim et al., 1990) lies within the 95% confidence limit (Fig. 2C) at a distance cutoff of 3.50 Å.

Perhaps the most striking difference between initial and final models was obtained with the PRAI-IGPS bifunctional enzyme from the tryptophan biosynthetic pathway. The initial MIR model of this structure (Priestle et al., 1987) contained several misregistrations (residues 45–150) and other regions of poorly defined structure (residues 250–300, 350–400). All of these regions were detected using distance limits of 3.00–3.75 Å, while the final model of the bifunctional enzyme (Wilmanns et al., 1992) behaved ideally (Fig. 2D). A comparison between Cα positions (i.e., the Euclidean distance between Cα positions) for the initial and final PRAI-IGPS structures (Wilmanns, 1990) (Fig. 3) and the plot of the error function versus residue show a strong correspondence between Cα shifts and an elevated error function ($E > 95\%$ confidence limit). Note that the region from residues 45 to 150 contained misregistrations, whereas residues centered about residue 100 were not misregistered (Wilmanns, 1990). The error function shows the majority of the region from residues 45 to 150 as being suspect except for the region about residue 100 (Fig. 2D). A comparison of Figures 2D and 3 suggests that the method described here is sensitive to errors on the order of 1.5 Å in Cα positions.

The program was also able to properly identify the incorrect regions in other initial protein models not described here. Furthermore, the purposely misfolded structures of Novotny et al. (1984) fell completely outside of the acceptable region of the database because they did not satisfy the criteria for the minimum number of interactions (see Methods). Both structures, the vl-like-hemerythrin and the hemerythrin-like-vl, had a much greater surface area and volume than their respective native structures. Because the atomic density of these models was unusually low, the number of interactions was below the allowed limit. However, the program was not successful in identifying incorrect model structures that had undergone extensive energy minimization in the absence of experimental constraints. More recent models of the vl-like-hemerythrin structures (Novotny et al., 1988), which have surface areas and volumes comparable to the native form of hemerythrin, were not identified as being misfolded. Novotny et al. (1988) have shown that surface polarity is an important criterion in distinguishing between misfolded structures, and we have not yet explicitly taken this parameter into account.

Survey of the PDB

Using this method, 298 monomeric structures from the PDB (July 1991 release; Bernstein et al., 1977) were analyzed. Of these structures, those for which fewer than 80% of the windows were within the 95% confidence limit were flagged. At distance limits of 3.00, 3.25, 3.50, and 3.75 Å, respectively, 92%, 94%, 97%, and 96% of the models (not including unrefined, old refinement methods, nonexperimental models, or NMR structures) met the 80% criterion. All 298 structures used in calculating the above percentages are listed in Table S2 on the Diskette Appendix. A graphical representation of the tested PDB entries is shown in Figure 4, where it is evident that only a few structures have more than 20% of their structure below the 95% confidence limit and therefore are somewhat questionable.

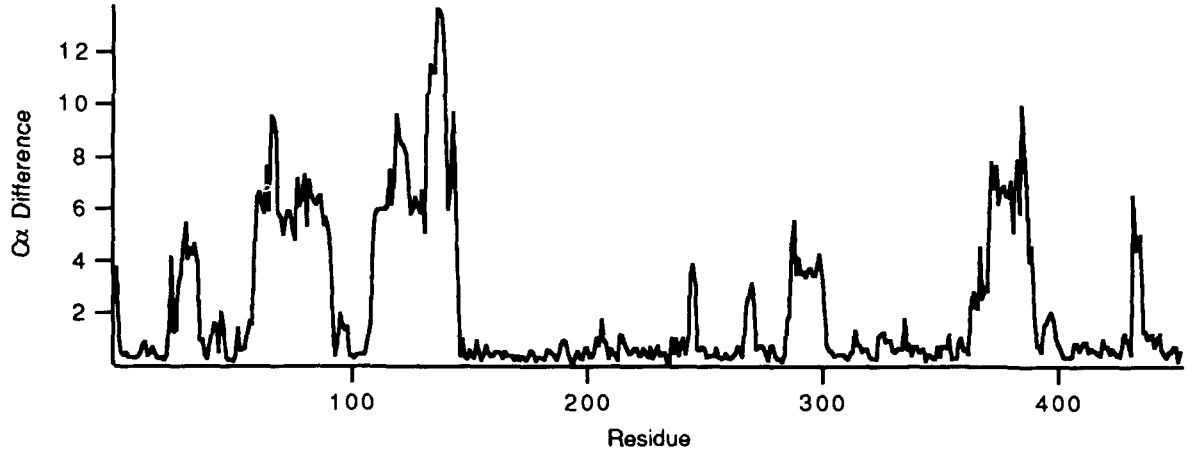

Fig. 3. Plot of distances between corresponding Cα positions in initial and final models of PRAI-IGPS as a function of residue.

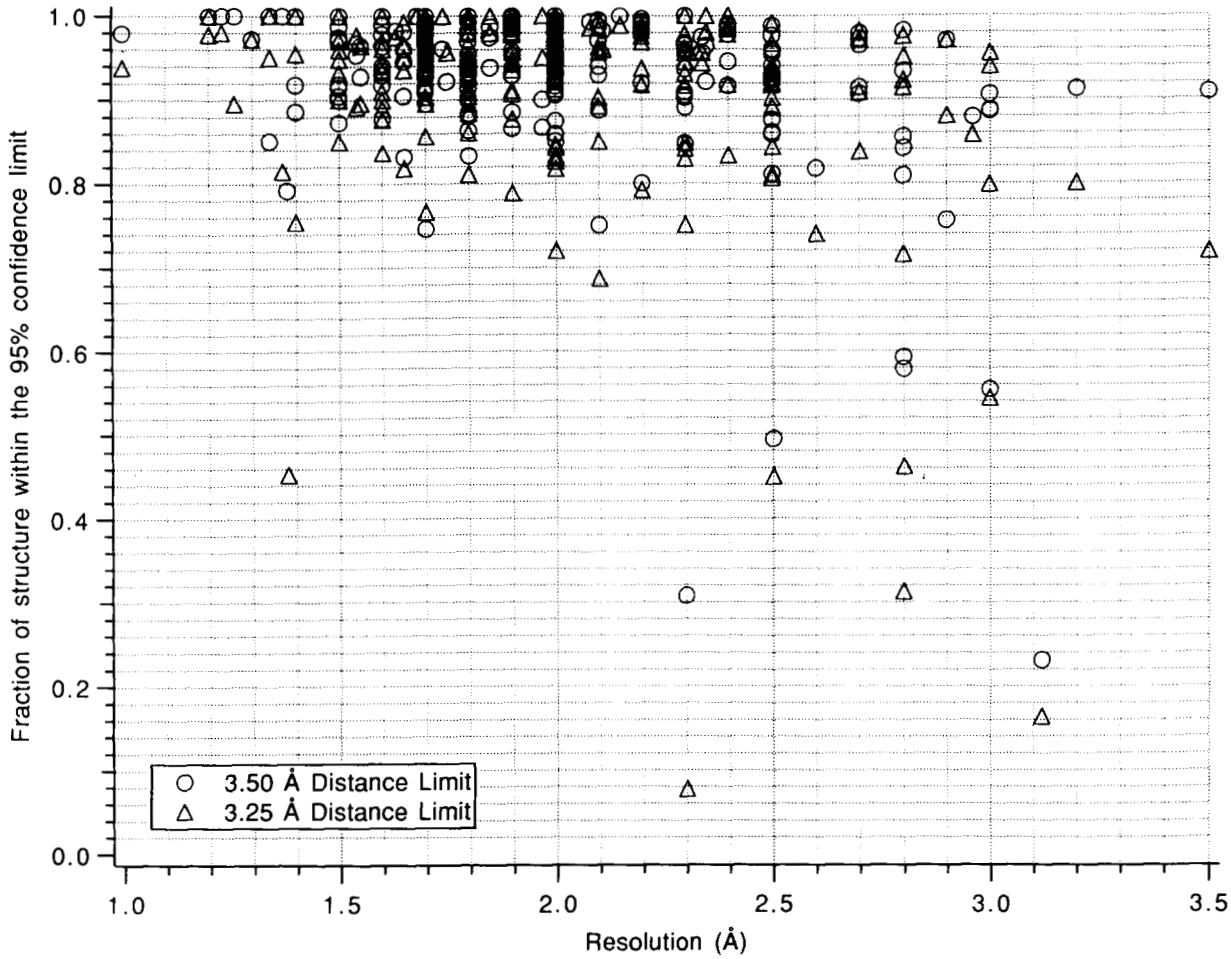

Fig. 4. Scatter plot of the PDB survey data in Table S2, Diskette Appendix.

As expected, there is a general correlation between resolution and the measure of structure quality. In addition, there appear to be a few high-resolution structures that our method suggests may contain errors, while the majority of medium- and low-resolution structures behave well. There are cases where a structure was originally determined at low resolution and then later reported at a higher resolution with both structures performing well by our criteria (see 2FNR and 1FNR in Table S2, Diskette Appendix).

For several of the structures in the PDB that appear to be outliers according to our analysis, there is other evidence to support the possibility of errors. For example, a recent redetermination of the crystal structure of the gene V protein (M. Skinner, unpubl.) has demonstrated multiple misregistrations in 2GN5. In addition, 1FXB has been replaced by 2FXB in more recent versions of the PDB. Cα differences between these two models are as high as 4.7 Å in regions where the error function for 1FXB was high (data not shown). The structure of the snake neurotoxin, 1NXB and 5EBX, has been determined by two independent groups at high resolution. The two models are very similar overall, but there are Cα differences in excess of 4 Å at residues 9 and 46. These regions of 1NXB were identified by our method as suspect. For xylose isomerase, Henrick, Collyer, and Blow (Protein Data Bank entry 4XIA, 1989) have pointed out significant differences between two homologous structures (3XIA and 4XIA), and Bryant (pers. comm.) has shown that the sequence alignment of the two proteins differs from the structural alignment in most of the secondary structural elements; 3XIA was an outlier in our survey. Whether other outliers contain serious or minor errors or represent false positives of our method remains to be tested.

Conclusions

Different atom types are distributed nonrandomly with respect to each other in proteins because of energetic and geometric effects. More random distributions are expected in incorrect regions of protein structural models. The analysis presented here reliably identifies regions of error in protein crystal structures by examining the statistics of pairwise atomic interactions. This method

should provide a useful tool for model-building and structure verification. It appears to be sensitive to errors in backbone positions on the order of 1.5 Å. One limitation is that the method does not distinguish well between errors on this scale and more severe errors such as occur with mistracing. The approach is also sensitive to the method used for atomic refinement, and unrefined structures generally do not score well. A final limitation is an inability to identify incorrect model structures that have been extensively refined without experimental constraints. A more reliable discrimination of incorrect regions would likely be obtained by combining the present analysis with others, especially those that explicitly consider surface polarity.

The FORTRAN program ERRAT is available from the authors. Analysis of a 300-residue protein takes less than 2 s of CPU time on a DEC 5000/200.

Acknowledgments

We thank the following investigators for kindly allowing use of coordinates of their preliminary structures: Drs. Paula Fitzgerald (HIV-1), John Rosenberg (*Eco*RI), Seunghyon Choe (diphtheria toxin), Matthias Wilmanns (bifunctional enzyme PRAI-IGPS), and David Eisenberg (rubisco small subunit). We also thank Drs. D. Eisenberg, M. Wilmanns, G. Privé, I. Wilson, and S. Choe for useful discussions and Drs. M. Chapman, J. Priestle, and M. Wilmanns for critical reading of the manuscript. This work was supported by USPHS grant GM 31299 and NSF PYI award DMB-9158602 to T.O.Y.

References

Baker, E.N. & Hubbard, R.E. (1984). Hydrogen bonding in globular protein. *Prog. Biophys. Mol. Biol. 44*, 97–179.

Baumann, G., Frommel, C., & Sander, C. (1989). Polarity as a criterion in protein design. *Protein Eng. 2*, 329–334.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., 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.

Boberg, J., Salakoski, T., & Vihinen, M. (1992). Selection of a representative set of structures from Brookhaven Protein Data Bank. *Proteins Struct. Funct. Genet. 14*, 265–276.

Chapman, M.S., Suh, S.W., Curmi, P.M.G., Cascio, D., Smith, W.W., & Eisenberg, D.S. (1988). Tertiary structure of plant RuBisCO: domains and their contacts. *Science 241*, 71–74.

Curmi, P.M.G., Cascio, D., Sweet, R.M., Eisenberg, D., & Schreuder, H. (1992). Crystal structure of the unactivated form of ribulose-1,5-bisphosphate carboxylase/oxygenase from tobacco refined at 2.0-Å resolution. *J. Biol. Chem. 267*, 16980–16989.

de Vos, A.M., Tong, L., Milburn, M.V., Matias, P.M., Jancarik, J., Noguhi, S., Nishimura, S., Miura, K., Ohtsuka, E., & Kim, S.H. (1988). Three-dimensional structure of an oncogene protein: Catalytic domain of human c-H-ras p21. *Science 239*, 888–893.

Eisenberg, D. & McLachlan, A.D. (1986). Solvation energy in protein folding and binding. *Nature 319*, 199–203.

Fitzgerald, P.M.D., McKeever, B.M., van Middlesworth, J.F., Springer, J.P., Heimbach, J.C., Leu, C., Herber, W.K., Dixon, R.A.T., & Darke, P.L. (1990). Crystallographic analysis of a complex between human immunodeficiency virus type 1 protease and acetyl-pepstain at 2.0 Å resolution. *J. Biol. Chem. 265*, 14209–14219.

Ghosh, D., O'Donnel, S., Furry, W., Jr., Robbins, A.H., & Stout, C.D. (1982). Iron-sulfur clusters and protein structure of *Azotobacter* ferredoxin at 2.0 Å resolution. *J. Mol. Biol. 158*, 73–109.

Hendlich, M., Lackenr, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., & Sippl, M. (1990). Identification of native protein folds amongst a large number of incorrect models. *J. Mol. Biol. 216*, 167–180.

Jones, T.A., Zou, J.Y., Cowan, S.W., & Kjelgaard, M. (1991). Improved methods for binding protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A47*, 110–119.

Kim, Y., Grable, G., Love, R., Greene, P.J., & Rosenberg, J.M. (1990). Refinement of *Eco* RI endonuclease crystal structure: A revised chain tracing. *Science 247*, 1307–1309.

Lüthy, R., Bowie, J.U., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature 356*, 83–85.

McClairen, J.A., Frederick, C.A., Wang, B., Greene, P., Boyer, H.W., Grable, J., & Rosenberg, J.M. (1986). Structure of the DNA-*Eco* RI endonuclease recognition complex at 3 Å resolution. *Science 234*, 1526–1541.

Navia, M.A., Fitzgerald, P.M.D., McKeever, B.M., Leu, C., Heimbach, J.C., Herber, W.K., Sigal, I.S., Darke, P.L., & Springer, J.P. (1989). Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. *Nature 337*, 615–620.

Novotny, J., Bruccoleri, R., & Karplus, M. (1984). An analysis of incorrectly folded protein models: Implications for structure predictions. *J. Mol. Biol. 177*, 787–818.

Novotny, J., Rashin, A.A., & Bruccoleri, R.E. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins Struct. Funct. Genet. 4*, 19–30.

Priestle, J.P., Grütter, M.G., White, J.L., Vincent, M.G., Kania, M., Wilson, E., Jardetzky, T.S., Kirschner, K., & Jansonius, J.N. (1987). Three-dimensional structure of the bifunctional enzyme *N*-(5′-phosphoribosyl)anthranilate isomerase-indole-3-glycerol-phosphate synthase from *Escherichia coli*. *Proc. Natl. Acad. Sci. 84*, 5690–5694.

Ramachandran, G.N. & Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Adv. Protein Chem. 23*, 283–437.

Schreuder, H.A., Curmi, P.M.G., Cascio, D., & Eisenberg, D. (1990). The RuBisCO saga—Accuracy and reliability of macromolecular crystal structures. *Proc. CCP4 Study Weekend, 26–27 January 1990*, 73–82.

Stickle, D.F., Presta, L.G., Dill, K.A., & Rose, G.D. (1992). Hydrogen bonding in globular proteins. *J. Mol. Biol. 226*, 1143–1159.

Tanaka, S. & Scheraga, H.A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules 9*, 945–950.

Wilmanns, M. (1990). Ph.D. Thesis, University of Basel, Basel, Switzerland.

Wilmanns, M., Priestle, J.P., Niermann, T., & Jansonius, J.N. (1992). Three-dimensional structure of the bifunctional enzyme phosphoribosylanthranilate isomerase: Indoleglycerolphosphate synthase from *Escherichia coli* refined at 2.0 Å resolution. *J. Mol. Biol. 233*, 477–507.

Wlodawer, A., Miller, M., Jaskolski, M., Sathyanarayana, B.K., Baldwin, E., Weber, I.T., Selk, L., Clawson, L., Schneider, J., & Kent, S.B. (1989). Conserved folding in retroviral proteases: Crystal structure of a synthetic HIV-1 protease. *Science 245*, 616–621.

Appendix: *n*-dimensional normal distribution

Suppose a set of observations in two dimensions follows a normal distribution (Fig. 5). The corresponding probability function, $P(x, y)$, is a general Gaussian:

$$P(x, y) \propto e^{-(ax^2+bxy+cy^2)}, \tag{A1}$$

where the quadratic exponent represents a family of ellipses in the x, y plane, each corresponding to a constant probability $P(x, y)$.

In n dimensions, this generalizes to

$$P(\vec{\mathbf{x}}) \propto e^{-\vec{\mathbf{x}}^T \mathbf{B} \vec{\mathbf{x}}}, \tag{A2}$$

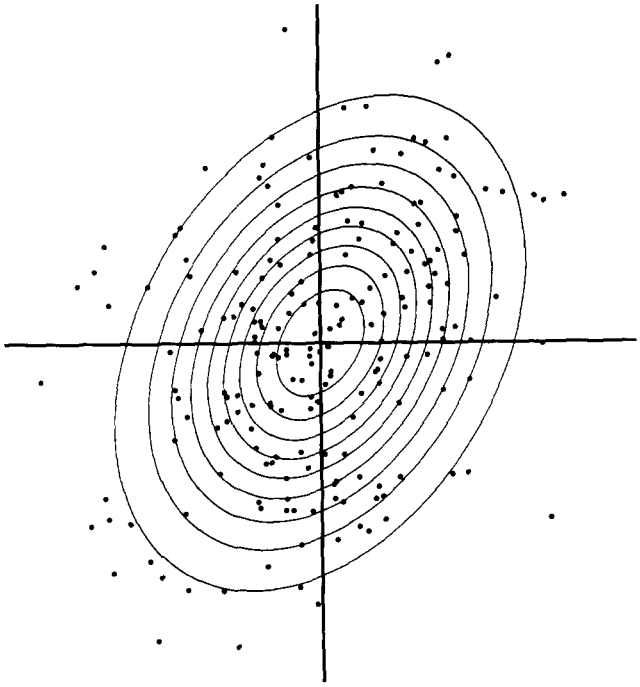

Fig. 5. A normal distribution in two dimensions.

where **B** is a symmetric positive-definite matrix. When the observations, $\vec{x}$, are referred to a new coordinate system by the following linear transformation:

$$\vec{x}' = \mathbf{B}^{1/2}\vec{x}, \tag{A3}$$

then

$$P(\vec{x}') \propto e^{-\vec{x}'^{\mathrm{T}}\vec{x}'} = e^{-|\vec{x}'|^2} \tag{A4}$$

and the Gaussian distribution is spherically symmetric in the new coordinate system. The following covariance values may be obtained in this coordinate system:

$$\langle \vec{x}_i' \vec{x}_j' \rangle = \sigma^2_{\vec{x}_i'\vec{x}_j'} = \int_{\vec{x}'} \vec{x}_i' \vec{x}_j' \, P(\vec{x}') \, d\vec{x}' = \frac{1}{2}\delta_{ij}, \tag{A5}$$

where δ_{ij} is the Kroenecker delta ($\delta_{ij} = 0$, $i \neq j$ and $\delta_{ij} = 1$, $i = j$), so that the covariance matrix

$$\langle \vec{x}'\vec{x}'^{\mathrm{T}} \rangle = \tfrac{1}{2}\mathbf{I}, \tag{A6}$$

where **I** is the identity matrix.

According to Equation A3,

$$\vec{x} = \mathbf{B}^{-1/2}\vec{x}' \tag{A7}$$

and

$$\langle \vec{x}\vec{x}^{\mathrm{T}} \rangle = \langle \mathbf{B}^{-1/2}\vec{x}'\vec{x}'^{\mathrm{T}}\mathbf{B}^{\mathrm{T}-1/2} \rangle. \tag{A8}$$

Equations A8 and A6 give

$$\mathbf{B}^{-1} = 2\langle \vec{x}\vec{x}^{\mathrm{T}} \rangle, \tag{A9}$$

which establishes the relationship between **B** and the set of observations, $\vec{x}$.