Energy strain in three-dimensional protein structures

Vladimir Maiorov and Ruben Abagyan

Background: Steric strain in protein three-dimensional structures is related to unfavorable inter-atomic interactions. The steric strain may be a result of packing or functional requirements, or may indicate an error in the coordinates of a structure. Detailed energy functions are, however, usually considered too noisy for error detection.

Results: After a short energy refinement, a full-atom, detailed energy function becomes a sensitive indicator of errors. The statistics of the energy distribution of amino acid residues in high-resolution crystal structures, represented by models with idealized covalent geometry, were calculated. The interaction energy of each residue with the whole protein structure and with the solvent was considered. Normalized deviations of amino acid residue energies from their average values were used for detecting energy-strained and, therefore, potentially incorrect fragments of a polypeptide chain. Protein three-dimensional structures of different origin (X-ray crystallography, NMR spectroscopy, theoretical models and deliberately misfolded decoys) were compared. Examples of the applications to loop and homology modeling are provided.

Conclusions: Elevated levels of energy strain may point at a problematic fragment in a protein three-dimensional structure of either experimental or theoretical origin. The approach may be useful in model building and refinement, modeling by homology, protein design, folding calculations, and protein structure analysis.

Address: The Skirball Institute of Biomolecular Medicine and The Department of Biochemistry, New York University Medical Center, 540 First Avenue, New York, NY 10016, USA.

Correspondence: Ruben Abagyan
E-mail: abagyan@earth.med.nyu.edu

Key words: conformational analysis, homology modeling, loop modeling, molecular modeling, protein design

Received: **19 January 1998**
Revisions requested: **16 February 1998**
Revisions received: **30 March 1998**
Accepted: **03 April 1998**

Published: **05 June 1998**
http://biomednet.com/elecref/1359027800300259

Folding & Design 05 June 1998, **3**:259–269

Introduction

Identification of strain in protein three-dimensional structures has many important implications for both experimental structure determination and theoretical modeling and design. The term 'steric strain' usually describes unfavorable or disallowed conformations or structural abnormalities of the amino acid residues that are detected by analysis of ϕ/ψ maps, van der Waals clashes, etc. High-resolution X-ray crystal structures from the PDB [1] were analyzed [2,3] and it was shown that some parts of the polypeptide chain manifest higher strain, as a result of packing or functional requirements. An alternative and the most likely source of strain is the error in the coordinates of a structure, ranging from misplaced sidechains and flipped peptide groups to wrong chain tracing [4–6]. Similarly, the dynamic nature of a biopolymer in solution and the ambiguities in peak assignments may result in errors in the structures solved by NMR.

Examination of a three-dimensional structure with respect to experimental (X-ray or NMR) data is a routine procedure [7,8]. Recently, several approaches to structural validations, independent of the direct experimental data, have been developed (see [9] and references therein). Typical 'quality evaluation' involves a comparison of certain characteristics of an analyzed conformation against their distributions observed in a set of high-resolution crystal structures. Studies reported in the literature deal with bond lengths, bond angles, disulfide bond geometry, chirality of Cα atoms, planarity of peptide bonds and aromatic rings, mainchain [3,10] and sidechain [11] torsion angle distributions, inter-residue contacts [12], packing densities and atomic volumes [13,14], inter-atomic distance distribution [15–17], atomic solvation preferences [18], etc. Combined analysis of several features may be particularly advantageous [19–21].

Novotny *et al.* [22] showed that vacuum energy fails to discriminate between misfolded models and the correct folds. Despite a later study showing good discriminating properties of an extended energy function [23], most methods of protein threading, fold recognition and error detection (see [24–28] for reviews) are based on non-energy scoring functions. The quality of three-dimensional structures in protein folding, homology modeling and protein design calculations can be examined by sequence-to-structure compatibility testing, which is intrinsic to all these methods.

We compare the semi-empirical force-field interaction energy of a single amino acid residue with the whole protein structure and the surrounding solvent with an energy distribution observed in high-resolution crystal structures. Internal coordinate mechanics (ICM; [29]), an efficient method for modeling and conformational analysis of peptides and proteins [30–35], was used in the work. We show

Table 1

Protein three-dimensional structures analyzed in this work.

Category	PDB codes
X-ray*	154l, 1aap.A,B, 1aaz.A,B, 1abo.A,B,C,D, 1acf, 1aky, 1amp, 1arb, 1asu, 1bbh.A,B, 1bgh, 1bit, 1cad, 1cbs, 1cew.I, 1cfb, 1chd, 1cmb.A,B, 1crn, 1cse.E,I, 1csn, 1ctf, 1cus, 1cyo, 1dbs, 1dyr, 1ede, 1edt, 1elt, 1emd, 1esl, 1fas, 1fd2, 1fdn, 1fkb, 1flp, 1fnc, 1frd, 1frr.A, fxd, 1gca, 1gdj, 1gia, 1gmp.A1,B1, 1goa, 1gpr, 1hag.E,I, 1hcr.A, 1hml, 1hms, 1hpi,, iab, 1isu.A,B, 1knb, 1knt, 1kpt.A,B, 1lif, 1lki, 1lmq, 1lst,m2, 1lsy, 1lte, 1lz6, 1mct.A,I, 1mcy, 1mjc, 1mml, 1mol.A,B, 1mrg, 1msc, 1nap.A,B,C,D, 1nar, 1ndc, 1nhk.R.L, 1nif, 1noa, 1npc, 1ofv, 1orb, 1osa, 1paz, 1pbn, 1pmy, 1poc, 1ppa, 1ppe.E,I, 1ppf.E,I, 1ppn, 1ppo, 1r69, 1rcf, 1rds, 1reg_x,y, 1ris, 1rpo, 1rro, 1rtp.1,2,3, 1scs, 1sgp.E,I, 1sgt, 1shg, 1sph.A,B, 1st3, 1sta, 1tca, 1tgs_z,I, 1tgx.A, 1thm, 1thv, 1thx, 1tib, 1tif, 1tig, 1tml, 1try, 1ubi, 1udg, 1ukz, 1utg, 1vhh, 1xnb, 1xso.A,B, 1xyn, 1yea, 21bi, 256b.A,B, 2abk, 2alp, 2ayh, 2aza.A,B, 2baa, 2bop.A, 2cab, 2cba, 2cdv, 2ci2.I, 2cy3, 2dri, 2ebn, 2end, 2erl, 2exo, 2fcr, 2hbg, 2lao, 2lig.A,B,C, 2mcm, 2mhr, 2mlt.A,B, 2ovo, 2phy, 2pia, 2prd, 2sec.E,I, 2sn3, 2spc.A,B, 2tgi, 3blm, 3dfr, 3sdh.A,B, 3wrp, 4dfr.A, 4fxn, 5cpa, 5pal
NMR	1amb, 1apc, 1arr.A,B, 1ata, 1atb, 1bba, 1bbl, 1bbn, 1bds, 1bta, 1cbh, 1ccn, 1cfd, 1chc, 1cod, 1coo, 1cre, 1crq, 1cta.A,B, 1cti, 1cxn, 1dem, 1dis, 1dmc, 1dme, 1ehs, 1epg, 1epi, 1era, 1erg, 1exg, 1fbr, 1fks, 1fkt, 1gdc, 1gfc, 1hcc, 1hcd, 1hce, 1hcs_b, 1hfh, 1hfi, 1hid, 1hks, 1hme, 1hnr, 1hre, 1hrq, 1hsm, 1hsq, 1hum.A,B, 1hwa, 1ife, 1ikl1, 1il8.A,B, 1irl, 1itl, 1put, 1sis, 2cti, 2eti
Models	1aag.L,H, 1abl, 1apk, 1bpk,m2, 1bst, 1btd.A,B, 1flx, 1fvb.L,H, 1fvw.L,H, 1gf1, 1gf2, 1hlh.A,B, 1hli, 1hlj, 1ita, 1itn, 1mca.A,B, 1mtm, 1pfa, 1ssr, 2bpk, 2cp1, 2fvb.L,H, 2fvw.L,H, 2hie, 2hif, 3flx, 3itr, 3its
Misfolded†	1bp2(2paz), 1cbh(1ppt), 1fdx(5rxn), 1hip(3b5c), 1lh1(2i1b), 1p2p(3rn3), 1ppt(1cbh), 1rhd(2cyp), 2cdv(2ssi), 2ci2(2cro), 2cro(2ci2), 2cro(2sn3), 2i1b(1lh1), 2paz(1bp2), 2sn3(2ci2), 2sn3(2cro), 2ssi(2cdv), 3b5c(1hip), 3rn3(1p2p)

165 X-ray crystal structures, 61 NMR structures, and 29 model structures are shown. Chain identifiers are indicated after the code if a PDB file contains more than one chain. *Only X-ray crystal structures with a resolution of 2.0 Å or better were considered. †A selection of the misfolded models generated by Holm and Sander [18]. Several models with some atoms missing and/or with the crystallographic resolution of the three-dimensional structure worse than 2.0 Å were excluded. A(B), the sequence of protein A was threaded onto the three-dimensional structure of protein B.

that after careful structural refinement ('regularization') the normalized energy strain per residue is a straightforward and sensitive measure for identifying both local and global conformational strain. Several applications of the method to protein three-dimensional structure analysis and modeling are given. The approach may be useful in model building and refinement, modeling by homology, protein design, folding calculations, and structural analysis.

Results and discussion

A representative set of high-resolution PDB protein structures (Table 1) represented by idealized covalent geometry models was refined by a regularization procedure as described in the Materials and methods section. The ECEPP/3 energy force field [36–38], extended by the solvation and the sidechain entropy terms [39], was used. The total energy of each analyzed protein structure was calculated and partitioned between the residues constituting the protein structure (see Equations 2, 3 and 4 below). Energy distributions for all amino acid residue types were then calculated and described by the average values (E_{av}) and standard deviations (E_{sd}; Table 2, Figure 1). In addition, energy distributions were also derived for different types of the secondary structures (Figure 1). Average residue energies (E_{av}) of the residues of the same type in the helical conformation and the extended conformation did not significantly deviate from the overall distribution. In contrast, residues in the coil conformation consistently manifested 1.9–5.9 kcal/mol higher values than that in the helical conformation. Similarly, the difference between the residues in the coil and the extended conformations varied from 0.5 kcal/mol to 6.1 kcal/mol. These results are in agreement with common knowledge that mainchain flexibility and packing constraints of the polypeptide chain often require loop fragments to adopt less favorable conformations than the core of the structure. If the conformation of the fragment is reliably determined, then the energy strain is to be

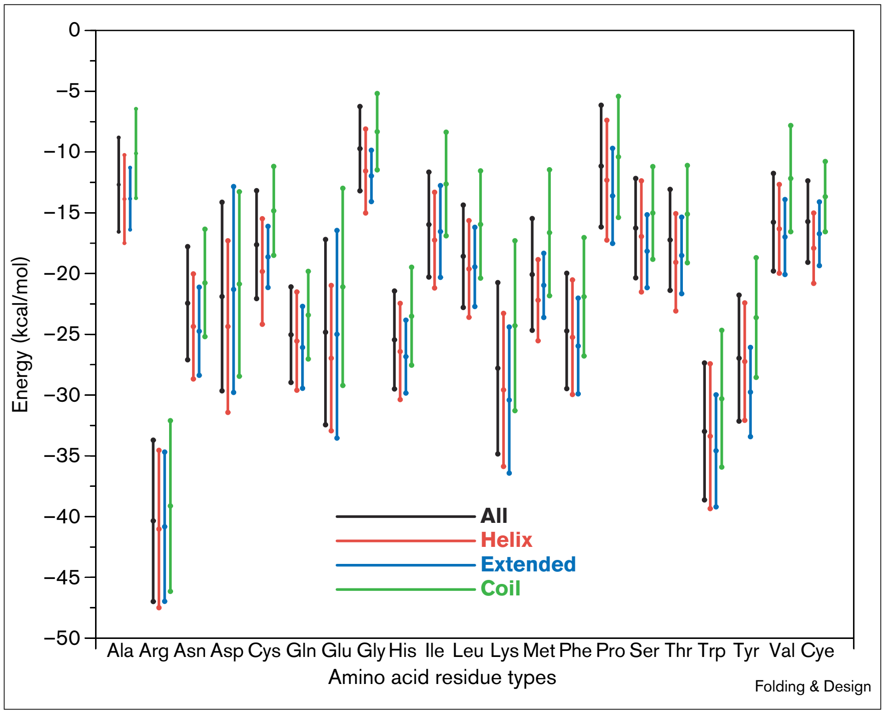
Table 2

Average energies and their standard deviations for the standard amino acid residue types derived from a set of high-resolution X-ray crystal structures.

Residue type	E_{av} (kcal/mol)	E_{sd} (kcal/mol)	N_{occ}*
Ala	−12.68	3.87	1186
Arg	−40.21	6.66	223
Asn	−22.44	4.61	365
Asp	−21.87	7.84	384
Cys†	−17.62/−15.72	4.44/3.40	98/172
Gln	−24.90	4.03	216
Glu	−24.78	7.55	234
Gly	−9.71	3.48	1977
His	−25.43	3.98	207
Ile	−15.91	4.31	625
Leu	−18.52	4.30	858
Lys	−27.88	7.01	190
Met	−20.06	4.59	151
Phe	−24.79	4.69	462
Pro	−11.16	4.97	507
Ser	−16.22	4.32	588
Thr	−17.19	4.16	708
Trp	−33.02	5.59	208
Tyr	−26.69	5.69	364
Val	−15.80	4.02	896

*Number of occurrences of each residue type in the training set. †The two values are for cysteine and cystine amino acids, respectively. E_{av}, average energies; and E_{sd}, the standard deviations.

Figure 1



Average energy values and their standard deviations for the standard amino acid residue types derived from a set of high-resolution protein crystal structures. The center of each vertical line corresponds to the average energy of the given residue type, and the range between the ends marks two standard deviations. Black lines show data derived without secondary structure differentiation. Colored lines correspond to the secondary structure types: red, helical (including 3_{10} helices and π helices); blue, extended (including β-bridged residues); and green, coiled. Residue types are indicated. Cye indicates the cystine residue to distinguish it from the cysteine residue. Secondary structures were assigned after the regularization of each protein structure in the training set by a modified [48] DSSP algorithm [54].

attributed to the folding/packing or functional requirements. Otherwise, it may be an indication of a faulty local conformation for reasons ranging from the impossibility of using a unique conformation to represent a flexible fragment to gross errors made in experimental determination or theoretical modeling.

Energy strain visualization

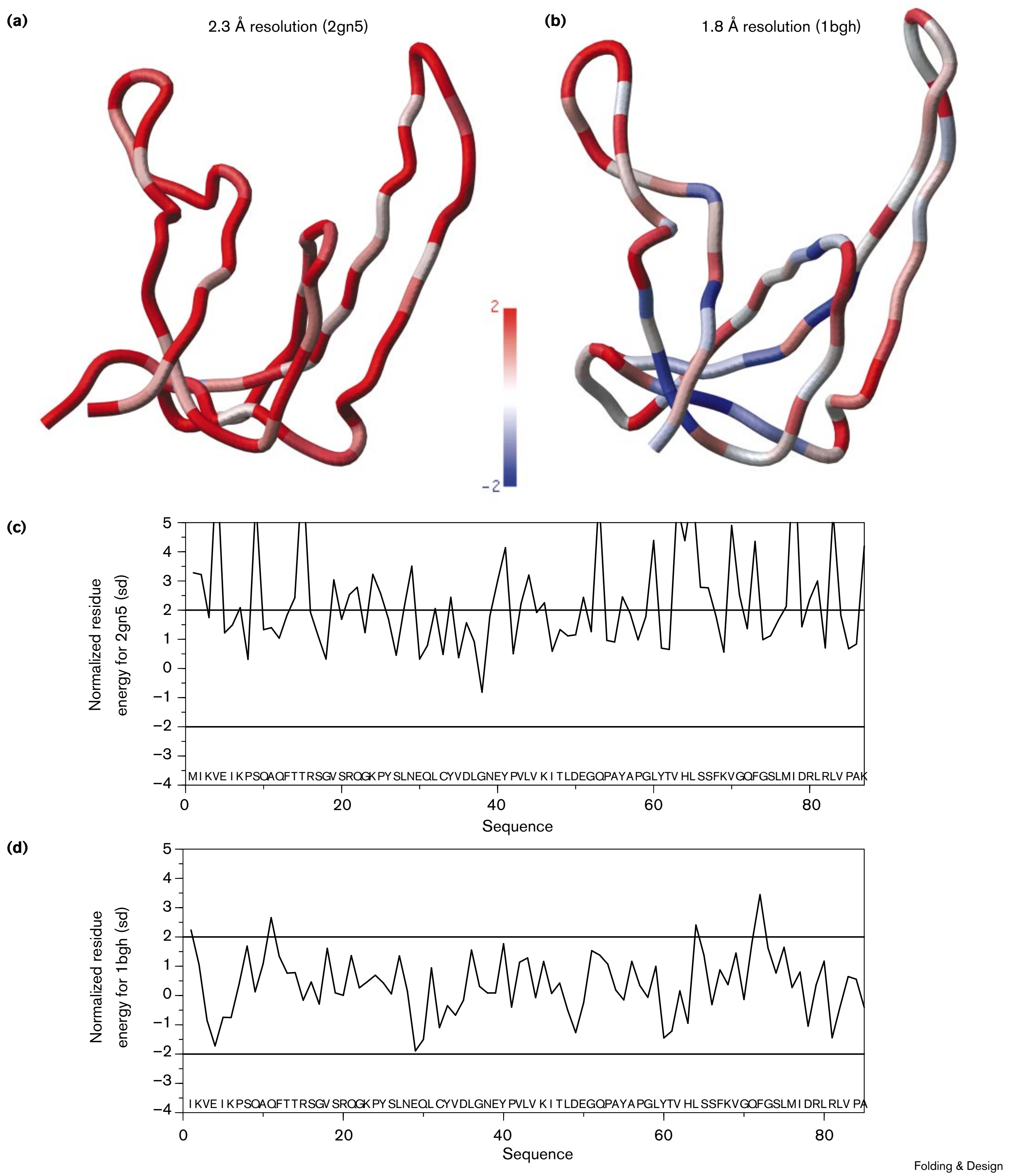
It should be stressed that the calculated energy values can be meaningfully applied to strain analysis if amino acid residues of the same type (i.e. having the same chemical structure) are compared. Only in this case, the difference in energy values reflects different preferences of the compared residues in their *given* conformations to exist in their given protein/solvent environment. To compare the strain condition of residues of different types, we propose a dimensionless 'normalized residue energy' (NRE), relying on the average energy and the standard deviation values (see the Materials and methods section). This measure is independent of the residue type and can be uniformly used for building 'energy strain' profiles and coloring molecular displays, as in the following example.

There are two PDB structures of gene V binding protein solved by X-ray crystallography in the same crystal form. (Figure 2). The first structure (PDB code 2gn5) was solved at 2.3 Å resolution [40]. Later, the second structure (PDB code 1bgh) was independently solved by another group of researchers at a higher resolution, 1.8 Å [41]. The structures differ substantially from each other: the Cα-atom root mean square deviation (rmsd) is 5.2 Å. Backbone displays colored by NRE values clearly highlight the difference in favor of the higher resolution structure. In the 2gn5 conformation, the distribution of the strained residues is approximately uniform (Figure 2a). In contrast, in the 1bgh structure, the strained residues are mainly located in the loop fragments (Figure 2b). Similarly, 39 residues (44% of the total chain length) in the 2gn5 structure have NRE values exceeding the energy strain cutoff level (see the Materials and methods section) compared with only four residues (5%) in the 1bgh structure. Note that the better quality of the 1bgh conformation could be concluded simply by visual inspection of the NRE-colored molecular displays and in a blind test on a comparison of the two conformations, if only their coordinates are given.

A comparison of the three-dimensional structures of different origin

An average NRE value, e_{av}, calculated as the average value of the NRE profile may serve as an energy-based 'quality index' of a three-dimensional structure (see the Materials and methods section). In the above example with gene V binding protein structures, e_{av} = 2.25 for 2gn5 and e_{av} = 0.40 for 1bgh. In a more systematic manner, we compared the distribution of average NRE values for the following categories of protein three-dimensional structures: X-ray crystal structures, NMR structures, theoretical models, and deliberately misfolded models (Table 1).

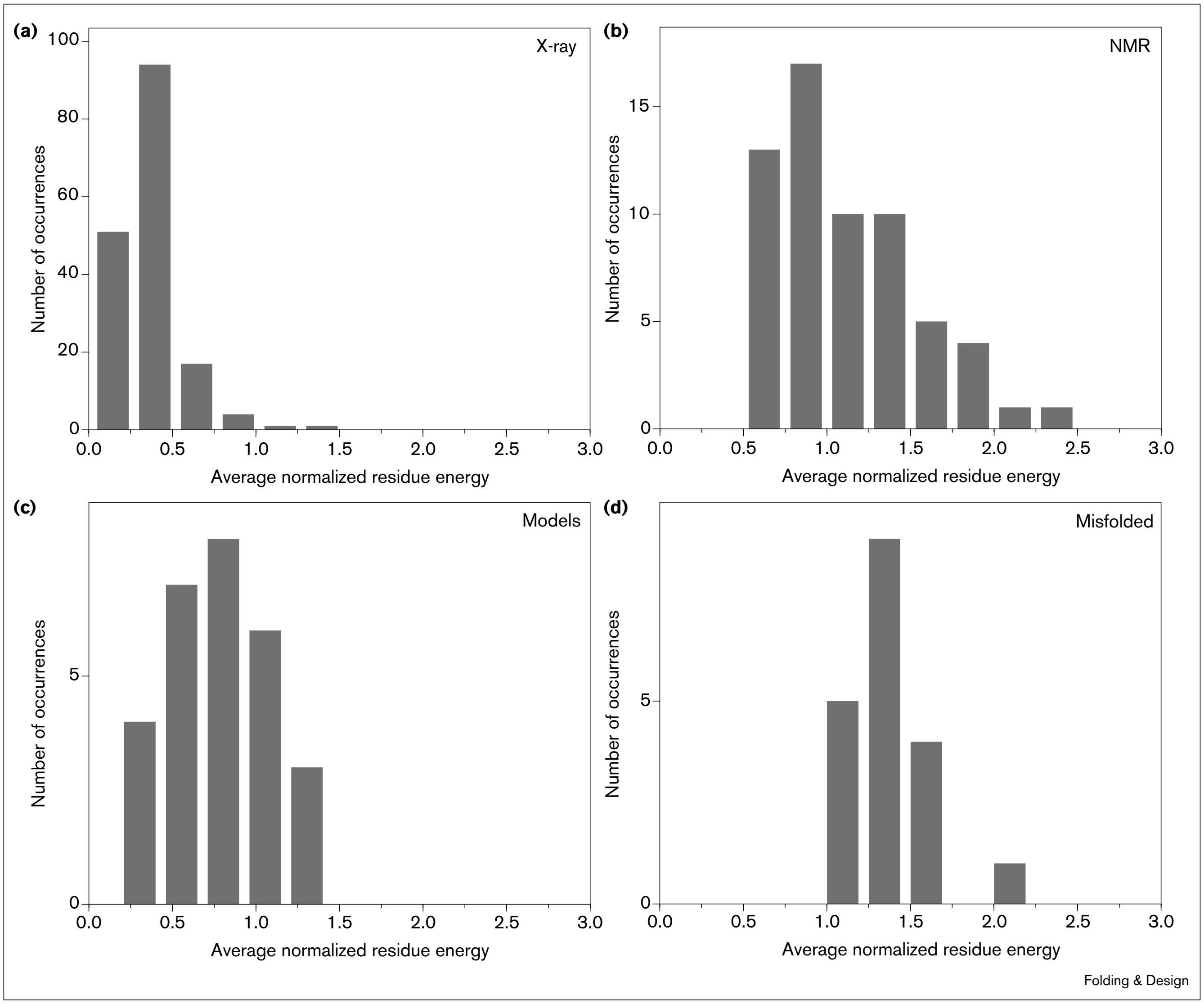
Figure 2



Folding & Design

Two structures of gene V binding protein solved in the same crystal form. **(a)** PDB code 2gn5, resolution 2.3 Å [40]; **(b)** PDB code 1bgh, resolution 1.8 Å [41]. Residues are colored according to the normalized residue energy (NRE) values (see the Materials and methods section). The polypeptide chain of the 1bgh structure is two residues shorter than that of 2gn5. The Cα atom rmsd of the two structures over 85 overlapping residues is 5.2 Å. NRE profiles are shown for **(c)** 2gn5 and **(d)** 1bgh.

Figure 3



Distributions of average NRE values for four categories of three-dimensional protein structures: **(a)** X-ray crystal structures; **(b)** NMR structures; **(c)** theoretical models; and **(d)** misfolded models generated by Holm and Sander [18].

The deliberately misfolded models included a set of decoys built by threading an amino acid sequence of one protein onto the three-dimensional structure of another having the same chain length [18]. Four structural categories were characterized using average NRE histograms (Figure 3) and overlap numbers ω (Table 3).

X-ray crystal structures were ranked first in the comparison of average-NRE distribution (Figure 3a); this is expected because only high-resolution (2 Å or better) and well-refined structures were considered. For the whole set of representative crystal structures used in the work (Table 1), the correlation coefficient between the average NRE e_{av} and the X-ray resolution was 0.231. This poor correlation is likely to

Table 3

A pairwise comparison of the distributions of average NRE values* of different categories of protein three-dimensional structures.

First set	Second set	Overlap number ω†
X-ray	NMR	−0.95
X-ray	Models	−0.72
X-ray	Misfolded	−0.99
NMR	Models	0.45
NMR	Misfolded	−0.48
Models	Misfolded	−0.91

*Calculated according to Equation 8. †Calculated according to Equation 9.

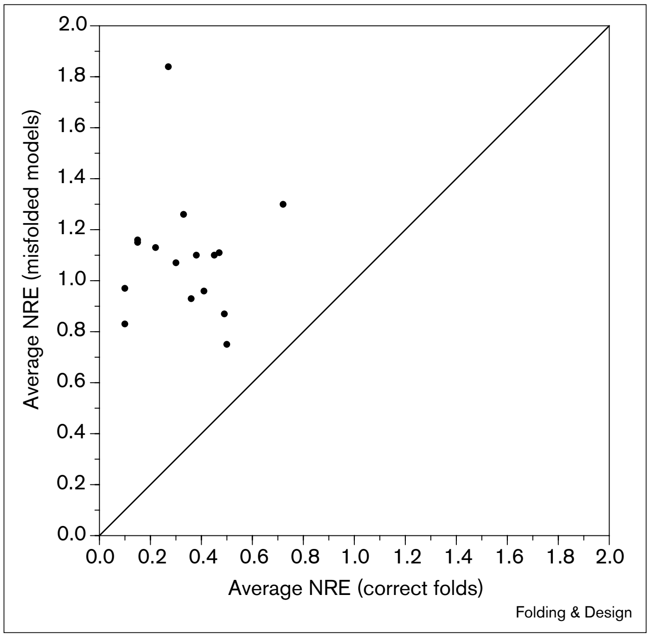
be a result of the narrow range of X-ray resolution of the analyzed structures: we observed a good approximation of true distribution of energy strain not influenced by the error in the coordinates. With regard to NMR structures, their average NRE e_{av} distribution was relatively broad, with 15% of the total number of structures manifesting $e_{av} > 2.0$ (Figure 3b). Interestingly, the theoretical models were found less strained than the NMR structures (Figure 3c).

It is likely that there are many reasons for an elevated strain level in the NMR conformations, when compared with the theoretical models. First, the elevated strain level may be a result of difficulties in representing an NMR solution structure with a set of conformations or a 'mean' structure. Second, a typical conformation derived from the NMR data is a compromise between the conformational (steric clashes, energy, etc.) and experimental (inter-atomic distance constraints and torsion angle restraints) requirements. Thus, further relaxation of the NMR-derived structures seems achievable, but only at the expense of increasing the number of violations of the experimental constraints and restraints [42]. Third, the analyzed theoretical models were mostly those modeled by homology, so they simply inherited the high quality of their three-dimensional prototypes. Alternatively, the elevated strain level may be attributed to more loosely interpretable (compared to NMR) restraints imposed on a model; the relaxation of the generated conformations may, therefore, be performed more extensively.

The misfolded structures were ranked the lowest and clearly were separated from all other categories (Figure 3d). In terms of the average NRE e_{av}, the misfolded models were always worse than the corresponding correct folds (Figure 4). This is also true if non-normalized energies are compared (data not shown).

It has long been recognized that vacuum empirical force-field energy can give comparable values for both a correct fold and a deliberately misfolded structure with accurately positioned sidechains [22]. It was also shown, however, that incorporation of a solvation term improves the recognition [23]. The use of easily refinable ICM models, in which torsion angles represent essential degrees of freedom, and the full-atom extended energy seem to provide a computationally tractable approach to realistically describe the distribution of strain along the chain. Recognition of the misfolded structures by energy is possible provided that the structures are regularized and the solvation and the sidechain entropy terms are taken into consideration. As a result of compensatory effects, variations of the energy components or related characteristics (such as hydrogen bonding, the number of buried polar and charged groups, the exposed surface of hydrophobic residues, torsion angle distribution, atomic contacts, etc.) may be misleadingly large. Thus, an accurately calculated energy should be the ultimate measure of structural quality.

Figure 4

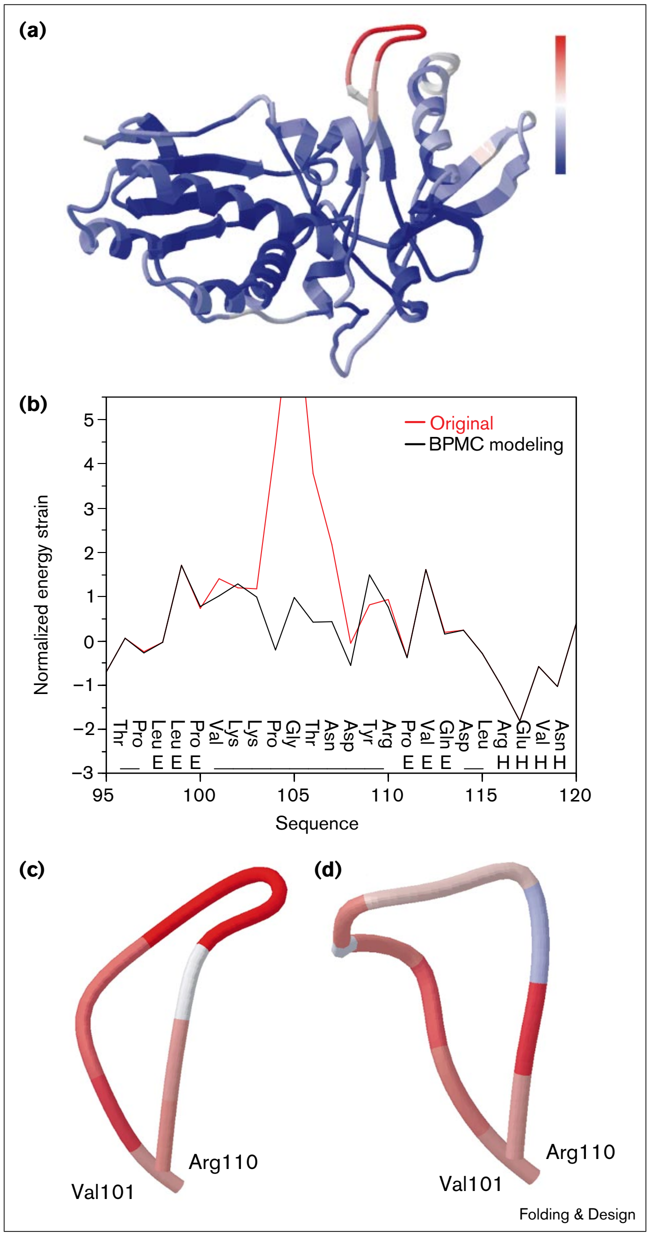


A comparison of misfolded models [18] with the corresponding correct folds: a scatter plot of the average NRE values (e_{av}). The misfolded models showed a substantially higher strain energy level than the correctly folded structures. Average NRE values were calculated according to Equation 8.

Optimization and correction of loop conformations

Energy strain identification applied to the ICM loop modeling methodology [43–45] allows one to obtain a visual feedback during loop calculations. As a test case, we chose the X-ray crystal structure of the *Moloney murine* leukemia virus reverse transcriptase solved at 1.8 Å resolution. Loop fragment Lys102–Tyr109 was found to have rather high B-factor values: an average residue B-factor equal to 66.1 $Å^2$ compared with 21.8 $Å^2$ for the whole protein (Figure 5a). The NRE profile of the regularized crystal structure identified strong energy strain in the central part of the loop (Figure 5b). We performed a biased probability Monte-Carlo (BPMC) search for a lower energy conformation of an extended fragment, Val101–Arg110, which includes the loop Lys102–Tyr109 (see the Materials and methods section). A set of 50 fragment conformations with 1.2–2.3 Å of mainchain atom rmsd from the regularized crystal structure was generated. The best conformation found had an energy of –1241.5 kcal/mol compared with –1173.0 kcal/mol for the original crystal structure. Correspondingly, the NRE values were substantially lower and at the level of the remaining part of the structure. (Figure 5b). The two conformations differed from each other by 2.0 Å of mainchain atom rmsd. Of course, it is difficult to guarantee that the conformation found in the BPMC search is realized in the crystal. On the other hand, it seems unlikely that the original, strongly strained conformation may exist in the

Figure 5



A search for a low-energy alternative to an energy-strained loop of the *Moloney murine* leukemia virus reverse transcriptase structure. **(a)** A ribbon representation of the backbone conformation colored by B-factor values. The whole range of average residue B-factor values (8.2–81.9 Å^2) corresponds to the color palette from blue to white to red. The loop Lys102–Tyr109 manifesting high B-factor values appears in red. The NRE color code is the same as in Figure 2. **(b)** NRE profiles of the fragment Val101–Arg110. The regularized crystal structure is shown in red; the energy is −1173.0 kcal/mol. The lowest-energy conformation found during the BPMC search is shown in black; the energy is −1241.5 kcal/mol. **(c)** The original regularized X-ray crystal conformation of the fragment. **(d)** The lowest energy conformation found during the BPMC search. The orientation in figures (a), (c) and (d) is the same. The conformations in (c) and (d) are colored by the NRE values; positions of the boundary residues Val101 and Arg110 are marked. The mainchain atom rmsd between the conformations is 2.0 Å.

crystal structure. Perhaps certain flexibility is intrinsic to the fragment, resulting in the poorer quality of the diffraction data in this region. This example illustrates how the described methodology can be implemented in a real structure determination or refinement protocol. If there are several variants possible under circumstances of insufficient or ambiguous experimental data, then the selection of the conformation for a fragment may be governed by the energy strain analysis (J. Williams and R. Wierenga, personal communication).

Use in homology modeling and a comparison with directional atomic contact index

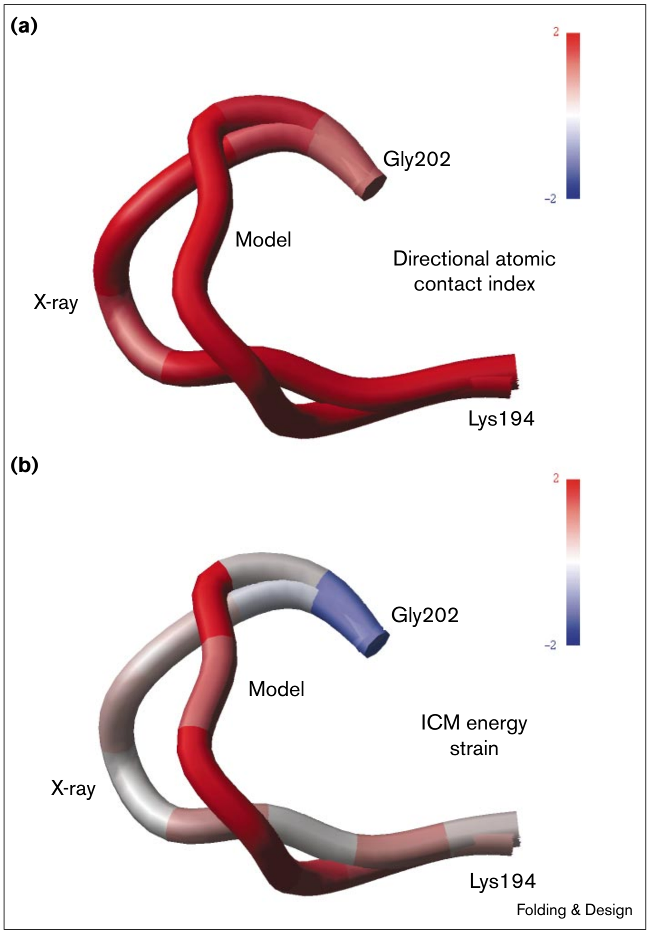
A model of the 216-residue *Carica papaya* caricain [46] (PDB code 1meg) was built on the basis of 212-residue papain three-dimensional structure (PDB code 1ppn; [47]). The two proteins show 70% sequence identity and have a Cα-atom rmsd = 1.1 Å for 208 aligned residues. Modeling was performed using the ICM homology modeling method [43,48]. The total mainchain rmsd between the crystal structure of 1meg and the generated caricain conformations was = 1.8 Å. The most significant differences between the conformations were found in three loops where the rmsd > 4.0 Å: Gln99–Ala105, Gly167–Tyr174 and Lys194–Gly202. The nine-residue loop Lys194–Gly202 was chosen for an exemplary comparison of the NRE and the directional atomic contact index (quality index; [15]). Values of both indices per residue were calculated for the caricain crystal structure and the model conformations, and molecular displays of two loop conformations were colored according to the calculated values of the quality index (Figure 6a) and NRE (Figure 6b). (To use the same coloring scheme, we used applied a linear transformation to the quality index as explained in the Materials and methods section).

For a qualitative comparison of the energy strain with the quality index, smoothed (over the nine residues of the loop Lys194–Gly202) values of each index, expressed as a standard deviation of the values observed in the X-ray structure, were used. In this example of the incorrect conformation recognition, the energy strain method manifested higher sensitivity than the quality index analysis. The calculated values for the energy strain for the crystal and the model were 0.43 and 2.38, respectively, whereas for the quality index values they were 1.95 and 2.55, respectively (Figure 6). Although in this example the energy strain analysis shows better results, a systematic comparison, on a statistically significant sample of structures, with the quality index as well as other error-detection methods is necessary for an objective evaluation of the performance of the method.

Energy minimization: when to stop

The proposed method of energy strain localization may give a practical guideline for the duration of the energy optimization. Generally, a lower energy minimum can be

Figure 6



Correct (X-ray crystal) and incorrect conformations of the loop Lys194–Gly202 of the caricain Asp158→Gln mutant [46] analyzed with the use of **(a)** Vriend and Sander's [15] directional atomic contact index (quality index) and **(b)** the energy strain. The incorrect loop conformation is from a homology model; coloring is according to (a) the linearly transformed quality index (see text) and (b) NRE values (see text). N-terminal and C-terminal residues are marked.

achieved in a longer computational run, but the time of the calculations is always limited. Calculations are usually stopped when no significant energy changes occur for a specified number of energy function calls, or the number of calls exceeds a specified limit. Such interruption of the optimization process may not mean that an energy minimum has been reached. Rather, it may result from the features of the optimization algorithm applied to a complicated energy landscape. In a better case, the optimization procedure may definitely indicate that an energy minimum is found. In either case, an independent evaluation of how low the 'low-energy' minimum really is seems to be helpful when planning a computational experiment or analyzing its results.

There are many useful theoretical approaches for structural error detection, including inverse protein folding and protein threading methods (see [24–28] and references therein). As a rule, they use simplified pseudo-energies and/or ad hoc scoring schemes. They are, therefore, focused on a low-resolution structural analysis and are not particularly sensitive to higher resolution details. Discrimination between detailed conformational models, in which the addition of one atomic group makes a difference, is quite important. For example, the discrimination may be the selection among several generated loop conformations, resolving ambiguities in charged sidechain positions, finding a problem fragment in a designed protein, etc. In this regard, the method presented in this work may be helpful for identification and visualization of problem conformations in protein design, folding and homology modeling and refinement, as well as a general tool of structural quality evaluation.

Materials and methods

Internal coordinate mechanics and energy calculations

The ICM method is described in detail elsewhere [29,39]; here, we describe the setup used in this work. Standard torsion angles of the polypeptide chain (namely mainchain φ, ψ, ω and sidechain χ) were allowed to vary ('free variables'). All bond lengths, bond angles, phase angles and the remaining torsion angles (e.g. φ torsion of proline residues, sidechain torsions of the aromatic rings, etc.) were kept unchanged ('fixed variables'). The total energy E_{total} was calculated with ECEPP/3 force field [36–38] extended by the solvation energy and the sidechain entropy contributions:

$$E_{total} = E_{vw} + E_{14} + E_{hb} + E_{to} + E_{el} + E_{ss} + E_{tz} + E_{so} + E_{en} \qquad (1)$$

where the terms correspond to van der Waals interactions (vw), 1–4 nonbonded interactions (14), hydrogen bonding (hb), torsion energy (to), electrostatic interactions (el), disulfide bond constraints (ss), solvation energy (so) and sidechain entropy (en) [39]. The solvation energy was calculated using new atomic solvation parameters [49]. An additional term, included in the energy function during regularization and loop-modeling calculations (see below), was the 'tether' contribution (tz). (Tethers are defined as harmonic restraints confining atoms of the model to the corresponding atom locations of a given three-dimensional template structure.) The cutoff distance for truncation of the van der Waals and electrostatic interactions was set to 7.5 Å, and for hydrogen bonding interactions it was set to 3.0 Å. The electrostatic energy was calculated with a distance-dependent dielectric constant $D_{diel} = 4r$.

Regularization

To perform energy calculations by ICM, a protein three-dimensional structure should be 'regularized' (i.e. be represented by an all-atom energy-refined model (ICM model) of the polypeptide chain in internal coordinates with idealized covalent geometry [29,31,35]). The maximal number of iterations during the annealing and the energy relaxation steps of the protocol was set to 40. Iteration cycles were stopped if the energy difference between the two successive steps was < 0.5 kcal/mol. Tether restraints to the crystallographic atom positions were switched on at the beginning of the annealing step, with the initial weight equal to 1 kcal/mol A^2. At each of the following iterations, the tether weight was recalculated according to Equation 2 from reference [35]. A criterion of the regularization convergence was arrival at a negative value of the total energy E_{total}. In fact, for all structures from the training set (see below) the convergence was reached, the total energy was substantially lower obeying a linear relationship $E_{total} = 13.7–10.1N_{res}$ kcal/mol with the correlation coefficient equal to –0.980. (Here, N_{res} is the total number of residues in the protein structure.) If the regularization failed to converge, it was considered an indication that the accuracy of an input three-dimensional structure was not sufficient for energy calculations, and such structures were disregarded. After the regularization, the total energy of a structure was calculated without the tether term contribution.

Protein dataset

A representative collection of 277 PDB proteins (Table 1) included both experimental X-ray and NMR structures and theoretical models with chain lengths of 28–338 residues. Solvent molecules, non-peptide ligands, prosthetic groups and complexed ions were omitted. In addition, a series of deliberately misfolded decoys [18], generated by threading the amino acid sequence of one protein onto the three-dimensional structure of another, was considered (Table 1). ICM models were built by the regularization procedure described above. The rmsd difference between the heavy atoms of the regularized structures and the corresponding ICM models varied from 0.1 Å to 0.5 Å.

Our experience shows that an X-ray resolution equal to 2 Å or better correlates well with an accuracy of structure determination necessary for successful regularization and full-atom, detailed energy calculations (data not shown). Thus, a training set used for the derivation of the energy parameters included 114 monomeric X-ray crystal structures solved at a resolution ≤ 2.0 Å (see the Accession numbers section). Residues with the average B-factor > 20.0 Å^2 were omitted. The derived parameters may be somewhat biased as a result of the selection for the calibration of only well-ordered fragments of the structures. It was considered more important, however, to exclude from consideration all questionable and, therefore, possibly strained fragments. N-terminal and C-terminal residues were also excluded to avoid effects of a different chemical structure neighborhood at the termini of the polypeptide chain.

Partition of the total energy between the residues

The total energy of a protein three-dimensional structure E_{total} in Equation 1 may be rewritten as a sum of the contributions associated with N_{res} residues constituting the polypeptide chain:

$$E_{total} = \sum_{i=1}^{N_{res}} E(i) \tag{2}$$

where each *i*-th residue's contribution E(*i*):

$$E(i) = E_{vw}(i) + E_{14}(i) + E_{hb}(i) + E_{to}(i) + E_{el}(i) + E_{ss}(i) + E_{so}(i) + E_{en}(i) \tag{3}$$

In turn, each energy term $E_{eterm}(i)$ representing pairwise inter-atomic interactions (vw, 14, hb, el, ss) may be presented as a sum of two components, namely $E_{eterm}(i)^{(intra)}$, describing intra-residue interactions, and $E_{term}(i)^{(inter)}$, describing interactions of the atoms of given residue *i* with all other atoms not belonging to this residue:

$$E_{eterm}(i) = E_{eterm}(i)^{(intra)} + 0.5E_{eterm}(i)^{(inter)} \tag{4}$$

Here, *eterm* stands for any of the pairwise energy terms (vw, 14, hb, el, ss). The coefficient equal to 0.5 on the right-hand side of Equation 4 is necessary to avoid counting $E_{eterm}(i)^{(inter)}$ twice, once for each of two interacting residues. The remaining energy components, namely torsion (to), solvation (so) and sidechain entropy (en), are not pairwise, and they did not require such consideration. The resulting energy per residue, E(*i*), described by Equations 2, 3 and 4 represents the sum of the internal interactions of a residue and its interactions with the rest of the protein structure and with the solvent.

Derivation of energy statistics

For all proteins in the training set, the energy contributions E(*k*,*X*), $\{k = 1, .., N_{occ}(X)\}$ of each residue *k* of type *X* were calculated according to Equations 2, 3 and 4. $N_{occ}(X)$ stands for total number of occurrences of the residue type *X* in the set. In total, there were 21 residue types, namely 19 standard amino acid residue types, and cysteine and cystine residues, which were considered separately as having different chemical structures: *X* = Ala, Arg, .., Cys (cysteine), Cye (cystine), .. , Val. Resulting distributions were described by the average values:

$$E_{av}(X) = \frac{\sum_{k=1}^{N_{occ}} E(k,X)}{N_{occ}} \tag{5}$$

and the standard deviations (Table 2):

$$E_{sd}(X) = \left(\frac{\left[\sum_{k=1}^{N_{occ}} E(k,X) - E_{av}(X) \right]^2}{N_{occ}} \right)^{1/2} \tag{6}$$

In a given protein structure with N_{res} residues, residue *i* of type *X* having energy E(*i*,*X*) was characterized by a dimensionless normalized residue energy, NRE:

$$e_i = \frac{E(i,X) - E_{av}(X)}{E_{sd}(X)} \tag{7}$$

A series of NRE values for the whole protein structure, $\{e_i\} = E(i,X)$, where $i = 1, 2, ..., N_{res}$, defines an NRE profile, a representation of energy strain distribution along the polypeptide chain. Molecular displays of the analyzed structures were colored by their NRE values. Low and high NRE values were marked by blue and red colors, respectively. Different protein structures may have different ranges of the NRE distributions. Thus, to preserve the same coloring scheme, we introduced an energy strain cutoff, e_{cut}, which was set to 2. For coloring purposes, the NRE values beyond the range $[-e_{cut}; e_{cut}]$ were truncated to the boundary values, $-e_{cut}$ and e_{cut}, respectively.

A protein three-dimensional structure can be characterized globally by an NRE value averaged over all residues of the chain:

$$e_{av} = \frac{\sum_{i=1}^{N_{res}} e_i}{N_{res}} \tag{8}$$

The lower the value of e_{av}, the better is assumed to be the quality of the analyzed 3D conformation.

It was shown earlier [42,50] that a gross comparison of the distributions, for which the nature is unknown (e.g. non-normal distributions) and a more accurate analysis is complicated or impossible, may be accomplished by an 'overlap number' ω [50]. In this work, we used this parameter to compare the distributions of average NRE e_{av} values calculated for different categories of three-dimensional structures. If, for example, the crystal structures *X*, $X = \{X_i$, where $i = 1, ..., N_X\}$, and the theoretical models *Y*, $Y = \{Y_j$, where $j = 1, ..., N_Y\}$, are compared, then the overlap number is defined as:

$$\omega(X,Y) = \left(1 - \left| \frac{1}{N_X N_Y} \left(\sum_{i=1,N_X} \sum_{j=1,N_Y} \delta(X_i,Y_j) \right) \right| \right) \tag{9}$$

where $\delta(X_i,Y_j) = 1$ if $X_i > Y_j$, 0 if $X_i = Y_j$, and −1 in all other cases. In essence, the parameter relates the overlap between two sets to the expected probability that an element of the second set is greater than an element of the first one, and has the following properties: ω(*X*,*Y*) varies between −1 (all X_i are smaller than all Y_j) and +1 (all X_i are greater than all values of Y_j); $\omega(X,Y) = -\omega(X,Y)$; $|\omega(X,Y)| < |\omega(X,Z)|$ means that set *Y* is 'closer' to set *X* than set *Z*.

Search for optimal loop conformations

The seven-residue loop Lys102–Tyr109 of the X-ray crystal structure of the *Moloney murine* leukemia virus reverse transcriptase; [51]) manifesting elevated values of B-factor was chosen for a test modeling of an alternative conformation. Fragment Val101–Arg110 (defined as loop Lys102–Tyr109 extended by one residue from both endpoints) was considered in the BPMC search [39]. The following two subsets of the standard torsions were set free. First, it was mainchain and sidechain torsions of the residues belonging to the fragment Val101–Arg110. The second subset was sidechain torsions of remaining residues located in the 5 Å vicinity of the fragment. In total, 69 torsions were set free and all the remaining standard torsions were fixed at their values realized in the regularized structure.

In the ICM calculations, the polypeptide chain is built residue by residue, in succession, starting from the N terminus. In connection to loop calculations, special consideration of the following two points is required. First, the relative positions of the non-loop parts of the structure may change after a BPMC random move occurs in the mainchain torsion of a loop residue. Thus, a local deformation of a loop fragment should be taken into account [52]. Second, local minimization is necessary after each random move to resolve inevitable steric clashes. The minimization may also displace the portion of the structure after the loop. Furthermore, the distorted conformation may still be accepted and the distortion may build up. To prevent both complications, the part of the structure after the loop has to be tethered to the corresponding portion of the original crystal structure.

BPMC calculations performed by the ICM program were combined with the generation and run-time update of a 'conformational stack', a list of low-energy conformations [53]. Each time a new conformation was accepted during a BPMC search it was compared to all those from the stack collected so far. If a new conformation was not similar to any of those already collected, a new slot for the conformation was created in the stack. If the new conformation did not exceed the similarity cutoff in comparison to any conformation already in the stack, it substituted for the last one if its energy was lower; otherwise, it was disregarded. If the stack was full, but the BPMC search was continuing and the accepted conformation differed from all those already in the stack, the new conformation replaced the conformation with the highest energy in the stack. The similarity cutoff was set to 2.5 Å mainchain atom rmsd of the generated fragment conformations. The maximum number of conformations simultaneously residing in the stack was set to 50.

Directional atomic contact analysis

The directional atomic contact index (quality index; [15]) values per residue were obtained as a part of the PROCHECK program [19] output (EMBL World Wide Web server at http://biotech.embl.heidelberg.de:8400/cgi-bin/sendquery). Special measures were taken to preserve the same coloring scheme as described above for NRE. The quality index values were multiplied by –1 ($X \Rightarrow -X$) and subjected to a linear transformation $X' = A + BX$. Coefficients A and B were derived from the condition that the quality index values $X' \in [-6.55;\ 6.22]$ obtained for the whole X-ray structure should match the corresponding NRE range, $Y \in [-2.19;\ 4.68]$): $Y = A + BX'$ and were found to be $A = 1.15$ and $B = 0.54$. Thus, transformed X' values was used instead of the original quality index values for coloring (Figure 6).

Accession numbers

The PDB codes for the 114 X-ray crystals structures with a resolution ≤ 2.0 Å: 154l, 1acf, 1aky, 1amp, 1arb, 1asu, 1bgh, 1cad, 1cew.I, 1cfb, 1chd, 1csn, 1ctf, 1cus, 1cyo, 1dbs, 1dyr, 1ede, 1edt, 1elt, 1emd, 1esl, 1fas, 1fkb, 1flp, 1fnc, 1frd, 1fxd, 1gca, 1gdj, 1gia, 1goa, 1gpr, 1hcr.A, 1hms, 1hpi, 1iab, 1knb, 1knt, 1lki, 1lmq, 1lte, 1mcy, 1mjc, 1mml, 1mrg, 1msc, 1nar, 1ndc, 1nif, 1noa, 1npc, 1orb, 1osa, 1pbn, 1pmy, 1poc, 1ppa, 1ppo, 1r69, 1rcf, 1rds, 1ris, 1rpo, 1rro, 1scs, 1shg, 1sta, 1tca, 1tgx.A, 1thm, 1thv, 1thx, 1tib, 1tif, 1tig, 1tml, 1try, 1ubi, 1udg, 1ukz, 1utg, 1vhh, 1xnb, 1yea, 21bi, 2abk, 2alp, 2ayh, 2baa, 2bop.A, 2cba, 2cdv, 2ci2.I, 2cy3, 2dri, 2ebn, 2end, 2exo, 2hbg, 2lao, 2mhr, 2ovo, 2phy, 2pia, 2prd, 2sn3, 2tgi, 3blm, 3dfr, 3wrp, 4dfr.A and 4fxn. (If a PDB structure includes more than one polypeptide chain, the chain identifier is indicated after the PDB code.) The *Moloney murine* leukemia virus reverse transcriptase structure was taken from the PDB (PDB code 1mml).

Acknowledgements

This work was supported by grants DE-FG0296ER62268 (U.S. Department of Energy) and R01GM55418 (U.S. National Institute of Health). We thank L. Holm and C. Sander for making their misfolded models publicly available, M. Shenderohivh for helpful comments, and A. Maiorov and S. Silverstein for reading the manuscript.

References

1. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F. & Weng, J. (1987). Protein data bank. In *Crystallographic databases-information content, software systems, scientific applications.* (Allen, F.H., Bergerhoff, G. & Sievers, R., eds), pp. 107-132. Data Commission of the International Union of Crystallography, Bonn, Germany.
2. Herzberg, O. & Moult, J. (1991). Analysis of the steric strain in the polypeptide backbone of protein molecules. *Proteins* **11**, 223-229.
3. Karplus, P.A. (1996). Experimentally observed conformation-dependent geometry and hidden strain in proteins. *Protein Sci.* **5**, 1406-1420.
4. Janin, J. (1990). Errors in three dimensions. *Biochimie* **72**, 705-709.
5. Branden, C. & Jones, T.A. (1990). Between objectivity and subjectivity. *Nature* **343**, 687-689.
6. Kleywegt, G.J. & Jones, T.A. (1995). Where freedom is given, liberties are taken. *Structure* **3**, 535-540.
7. Hendrickson, W.A. (1985). Stereochemically restrained refinement of macromolecular structures. *Methods Enzymol.* **11**, 252-270.
8. Brunger, A.T. & Nilges, M. (1993). Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy. *Q. Rev. Biophys.* **26**, 49-125.
9. MacArthur, M.W., Laskowski, R.A. & Thornton, J.M. (1994). Knowledge-based validation of protein structure coordinates derived by X-ray crystallography and NMR spectroscopy. *Curr. Opin. Struct. Biol.* **4**, 731-737.
10. Kleywegt, G.J. & Jones, T.A. (1996). Phi/psi-chology: Ramachandran revisited. *Structure* **4**, 1395-1400.
11. Ponder, J.W. & Richards, F.M. (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775-791.
12. Huang, E.S., Subbiah, S. & Levitt, M. (1995). Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.* **252**, 709-720.
13. Gregoret, L.M. & Cohen, F.E. (1990). Novel method for the rapid evaluation of packing in protein structures. *J. Mol. Biol.* **211**, 959-974.
14. Pontius, J., Richelle, J. & Wodak, S.J (1996). Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J. Mol. Biol.* **264**, 121-136.
15. Vriend, G. & Sander, C. (1993). Quality control of protein models: directional atomic contact analysis. *J. Appl. Crystallogr.* **26**, 47-60.
16. Subramaniam, S., Tcheng, D.K. & Fenton, J.M. (1996). A knowledge-based method for protein structure refinement and prediction. In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology.* (States, D.G., Agarwal, P., Gaasterland, T., Hunter, L. & Smith, R.F., eds), pp. 218-229, The AAAI Press, Menlo Park, CA.
17. Colovos, C. & Yeates, T.O. (1993). Verification of protein structure: patterns of nonbonded atomic interactions. *Protein Sci.* **2**, 1511-1519.
18. Holm, L. & Sander, C. (1992). Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.* **225**, 92-105.
19. Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283-291.
20. Laskowski, R.A., Rullmann, J.A., MacArthur, M.W., Kaptein, R. & Thornton, J.M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8**, 477-486.
21. Carson, M., Buckner, T.W., Yang, Z. & Narayana, S.V.L. (1994). Error detection in crystallographic models. *Acta Crystallogr.* **D50**, 900-909.
22. Novotny, J., Bruccoleri, R. & Karplus, M. (1984). An analysis of incorrectly folded protein models: implications for structure predictions. *J. Mol. Biol.* **177**, 787-818.
23. Novotny, J., Rashin, A.A. & Bruccoleri, R. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins* **4**, 19-30.
24. Sippl, M.J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355-362.
25. Wodak, S.J. & Rooman, M.J. (1993). Generating and testing protein folds. *Curr. Opin. Struct. Biol.* **3**, 247-259.
26. Kocher, J.-P.A., Rooman, M.J. & Wodak, S.J. (1994). Factors influencing the ability of knowledge-based potential to identify native sequence-structure matches. *J. Mol. Biol.* **235**, 1598-1613.
27. Bryant, S.H. & Altschul, S.F. (1995). Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.* **5**, 236-244.
28. Jones, D.T. & Thornton, J.M. (1996). Potential energy functions for threading. *Curr. Opin. Struct. Biol.* **6**, 210-216.

29. Abagyan, R.A., Totrov, M.M. & Kuznetsov, D.A. (1994). ICM: a new method for structure modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comp. Chem.* **15**, 488-506.
30. Borchert, T.V., Abagyan, R., Kishan, K.V.R., Zeelen, J.Ph. & Wierenga, R.K. (1993). The crystal structure of an engineered monomeric triosephosphate isomerase, monoTIM: the correct modeling of an eight-residue loop. *Structure* **1**, 205-213.
31. Eisenmenger, F., Argos, P. & Abagyan, R.A. (1993). A method to configure protein sidechains from the mainchain trace in homology modeling. *J. Mol. Biol.* **231**, 849-860.
32. Totrov, M.M. & Abagyan, R.A. (1994). Detailed *ab initio* prediction of lysozyme-antibody complex with 1.6 Å accuracy. *Nat. Struct. Biol.* **1**, 259-263.
33. Abagyan, R.A., Frishman, D. & Argos, P. (1994). Recognition of distantly related proteins through energy calculations. *Proteins* **19**, 132-140.
34. Cardozo, T., Totrov, M. & Abagyan, R. (1995). Homology modeling by the ICM method. *Proteins* **23**, 403-414.
35. Maiorov, V.N. & Abagyan, R.A. (1997). A new method for modeling large-scale rearrangements of protein domains. *Proteins* **27**, 410-424.
36. Momany, F.A., McGuire, R.F., Burgess, A.W. & Scheraga, H.A. (1975). Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *J. Phys. Chem.* **79**, 2361-2381.
37. Nemethy, G., Pottle, M.S. & Scheraga, H.A. (1983). Energy parameters in polypeptides. 9. Updating of geometric parameters, nonbonded interactions and hydrogen bond interactions for the naturally occurring amino acids. *J. Phys. Chem.* **87**, 1883-1887.
38. Nemethy, G., *et al.*, & Scheraga, H.A. (1992). Energy parameters in polypeptides. 10. Improved geometric parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J. Phys. Chem.* **96**, 6472-6484.
39. Abagyan, R.A. & Totrov, M.M. (1994). Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **235**, 983-1002.
40. Brayer, G.D. & McPherson, A. (1993). Refined structure of the gene 5-DNA binding protein from bacteriophage FD. *J. Mol. Biol.* **169**, 565-570.
41. Skinner, M.M., *et al.*, & Terwilliger, T.C. (1994). Structure of the gene V protein of bacteriophage *f1* determined by multiwavelength X-ray diffraction on the selenomethionyl protein. *Proc. Natl Acad. Sci. USA* **91**, 2071-2075.
42. Abagyan, R. & Totrov, M. (1997). Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. *J. Mol. Biol.* **268**, 678-685.
43. Abagyan, R., Batalov, S., Cardozo, T., Totrov, M. & Zhou, Y. (1997). Homology modeling with internal coordinate mechanics: deformation zone mapping and improvements of models via conformational search. *Proteins* **30**, 1-9.
44. Thanki, N., *et al.*, & Schliebs, W. (1997). Protein engineering with monomeric triosephosphate isomerase (monoTIM): the modeling and structure verification of a seven-residue loop. *Protein Eng.* **10**, 159-167.
45. Borchert, T.V., Abagyan, R.A., Jaenicke, R. & Wierenga, R.K. (1994). Design, creation, and characterization of a stable, monomeric triosephosphate isomerase. *Proc. Natl Acad. Sci. USA* **91**, 1515-1518.
46. Katerelos, N.A., Taylor, M.A., Scott, M., Goodenough, P.W. & Pickersgill, R.W. (1996). Crystal structure of a caricain D158E mutant in complex with E-64. *FEBS Lett.* **392**, 35-39.
47. Pickersgill, R.W., Harris, G.W. & Garman, E. (1992). Structure of monoclinic papain at 1.60 Å resolution. *Acta Crystallogr.* **B48**, 59-62.
48. Molsoft, L.L.C. (1996). ICM software manual. Version 2.6.
49. Abagyan, R.A. (1997). Protein structure prediction by global energy optimization. In *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications.* Vol 3. (van Gunsteren, W.F., Weiner, P.K. & Wilkinson, A.J., eds), Kluwer Academic Publishers, London.
50. Abagyan, R.A. & Batalov, S.V. (1997). Do aligned sequences share the same fold? *J. Mol. Biol.* **273**, 355-368.
51. Georgiadis, M.M., *et al.*, & Hendrickson, W.A. (1995). Mechanistic implications from the structure of a catalytic fragment of Moloney murine leukemia virus reverse transcriptase. *Structure* **3**, 879-892.
52. Abagyan, R.A. & Mazur, A.K. (1989). New methodology for computer-aided modelling of biomolecular structure and dynamics. 2. Local deformations and cycles. *J. Biomol. Struct. Dyn.* **6**, 833-845.
53. Abagyan, R.A. & Argos, P. (1992). Optimal protocol and trajectory visualization for conformational searches of peptides and proteins. *J. Mol. Biol.* **225**, 519-532.
54. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers* **22**, 2577-2637.

Because *Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published on the internet before being printed. The paper can be accessed from http://biomednet.com/cbiology/fad – for further information, see the explanation on the contents pages.