JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 15, Number 3, 2008
© Mary Ann Liebert, Inc.
Pp. 317–355
DOI: 10.1089/cmb.2007.0161

Statistics of Random Protein Superpositions: *p*-Values for Pairwise Structure Alignment

JAMES O. WRABL[1] and NICK V. GRISHIN[1,2]

ABSTRACT

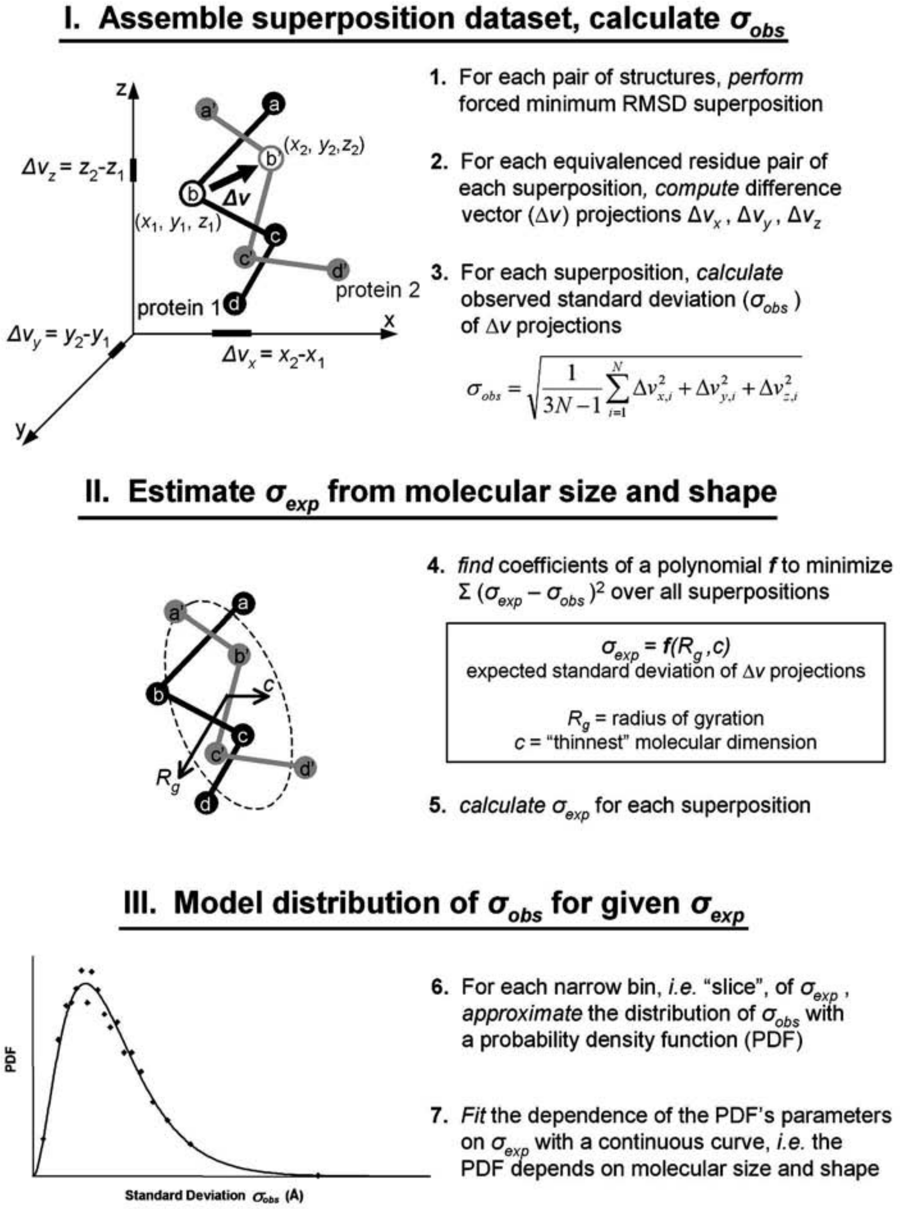
Quantification of statistical significance is essential for the interpretation of protein structural similarity. To address this, a random model for protein structure comparison was developed. Novelty of the model is threefold. First, a sample of random structure comparisons is restricted to molecules of the same size and shape as the superposition of interest. Second, careful selection of the sample and accurate modeling of shape allows approximation of the root mean square deviation (RMSD) distribution of random comparisons with a Nakagami probability density function. Third, through convolution, a second probability density function is obtained that describes the coordinate difference vector projections underlying the random distribution of RMSD. This last feature allows sampling random distributions of not only RMSD, but also *any* similarity score that depends on difference vector projections, such as GDT_TS score, TM score, and LiveBench 3D score. Probabilities estimated from the method correlate well with common measures of structural similarity, such as the Dali Z-score and the GDT_TS score. As a result, the *p*-value for a given superposition can be calculated using simple formulae depending on RMSD, radius of gyration, and thinnest molecular dimension. In addition to scoring structural similarity, *p*-values computed by this method can be applied to evaluation of homology modeling techniques, providing a statistically sound alternative to scores used in reference-independent evaluation of alignment quality.

Key words: protein structure alignment, random model, RMSD, statistical significance, superposition.

1. INTRODUCTION

COMPARISON OF PROTEIN STRUCTURES by analysis of three-dimensional atomic coordinates is one of the most fundamental tasks of structural biology. Identification and quantification of local and global conformational similarities can be compelling evidence for evolutionary relationships between proteins (Shah et al., 2005), for understanding the physical basis of molecular structure (Fitzkee et al., 2005), for

[1]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas.
[2]Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas.


I. Assemble superposition dataset, calculate σ_{obs}
z
x
y
$\Delta v_z = z_2 - z_1$
$\Delta v_y = y_2 - y_1$
$\Delta v_x = x_2 - x_1$
a'
a
b'
b
c
c'
d'
d
(x_2, y_2, z_2)
(x_1, y_1, z_1)
Δv
protein 1
protein 2
II. Estimate σ_{exp} from molecular size and shape
c
R_g
III. Model distribution of σ_{obs} for given σ_{exp}
PDF
Standard Deviation σ_{obs} (A)

1. For each pair of structures, *perform* forced minimum RMSD superposition

2. For each equivalenced residue pair of each superposition, *compute* difference vector (Δv) projections Δv_x, Δv_y, Δv_z

3. For each superposition, *calculate* observed standard deviation (σ_{obs}) of Δv projections

$$\sigma_{obs} = \sqrt{\frac{1}{3N-1}\sum_{i=1}^{N}\Delta v_{x,i}^2 + \Delta v_{y,i}^2 + \Delta v_{z,i}^2}$$

4. *find* coefficients of a polynomial ***f*** to minimize $\Sigma(\sigma_{exp} - \sigma_{obs})^2$ over all superpositions

$\sigma_{exp} = f(R_g, c)$
expected standard deviation of Δv projections

R_g = radius of gyration
c = "thinnest" molecular dimension

5. *calculate* σ_{exp} for each superposition

6. For each narrow bin, *i.e.* "slice", of σ_{exp}, *approximate* the distribution of σ_{obs} with a probability density function (PDF)

7. *Fit* the dependence of the PDF's parameters on σ_{exp} with a continuous curve, *i.e.* the PDF depends on molecular size and shape

FIG. 1. Development of the random model of protein structure superposition. Two four-residue fragments of non-homologous protein chains are shown as spheres, representing alpha-carbon (CA) atoms, connected by virtual bonds. One fragment, "protein 1," is colored black, and the second, "protein 2," is gray. As described in the text, gapless equivalences of CA atoms, depicted by, for example, the labels b–b′ in the figure, were forced regardless of structural similarity, and a minimum root mean square deviation (RMSD) superposition was performed. Such a forced superposition was termed a "random" superposition. Coordinate difference vector projections (DVPs) and their standard deviation, σ_{obs}, were computed from the random superposition. As an example, three projections $\Delta v_{x,y,z}$ (short thick lines on the x-, y-, z-axes) from one vector Δv (arrow) are indicated. This procedure was repeated for all pairs of fragments in the dataset (32,004 pairs). Then, as described in Methods, two size and shape parameters from each superposition, R_g and c, were used in a singular value decomposition (SVD) generalized linear least squares fit to determine coefficients for a polynomial f. Polynomial f (i.e., Equations (6A) or (6B)) estimated the expected standard deviation of projections, σ_{exp}, given the size and shape of a superposition, and σ_{exp} was calculated for all random superpositions in the set. Finally, distributions of σ_{obs} for a narrow range of σ_{exp} (0.2 Å increments)

(caption continues on next page)

assessing the quality of a structure prediction (Vincent et al., 2005), or for prediction of function (Whisstock and Lesk, 2003). Several algorithms have been developed that rapidly calculate optimal superpositions of structures from specified coordinate equivalences (Coutsias et al., 2004; Flower, 1999; Kabsch, 1976, 1978; Theobald, 2005). An even larger number of programs identify similarities between substructures and generate equivalences (i.e., alignments) automatically (Holm and Sander, 1993; Ortiz et al., 2002; Shindyalov and Bourne, 1998; Zhang and Skolnick, 2005).

Despite the widespread use of these software tools, it is sometimes difficult to judge the significance of a potential structural match. One reason for this difficulty is the lack of a realistic random model for structure comparison. The ideal random model should take into account many different aspects of protein structure, including chain length, chain self-avoidance, overall shape, secondary structure composition, globularity, close side-chain packing, and internal satisfaction of hydrogen bonding. However, although many such models have been advanced, in practice they prove to be computationally costly to generate. In addition, no "decoy" mimics all physical properties of proteins well enough to be adopted as an undisputed random model (Jia and Dewey, 2005; Reva et al., 1998; Taylor, 2006).

In the absence of a theoretical random model, others have proposed statistical measures based on analysis of score distributions of experimentally determined protein structures (Holm and Sander, 1994; Levitt and Gerstein, 1998). Although practically useful, they are often applicable only within the context of a particular scoring system or an alignment algorithm from which the distributions were generated (Jia et al., 2004; Ortiz et al., 2002; Shindyalov and Bourne, 1998). A third approach advocates "intrinsic measures of significance," such as mirror-image structures, that depend only on characteristics of the proteins being compared (Maiorov and Crippen, 1994, 1995). While mathematically and geometrically appealing, this alternative approach has not been widely employed.

In this work, an empirical method is proposed to estimate the significance of a structural match between two proteins. Significance is estimated with respect to a random model of gapless, minimum root mean squared deviation (RMSD) superposition of non-homologous structures of similar size and shape, characteristics controlled by a simple function depending on the superposition's radius of gyration and a measure of its "thinnest" molecular dimension. The method is parameterized for either comparison of entire proteins or comparison of substructures (fragments). Parameters for these random models are estimated from statistical analysis of superpositions' coordinate difference vector projections (i.e., the projections onto the coordinate axes of distances between equivalenced atoms).

Importantly, once parameterized, the method is computationally inexpensive in calculating an explicit p-value for a given match, since it uses simple formulae depending on only three inputs: the RMSD of the superposition, and the two size and shape measurements mentioned above. Adding to its utility, the method can generate random baseline scores for potentially *any* scoring system based on coordinate difference vector projections (e.g., TM score) (Zhang and Skolnick, 2004). Distributions of these randomly generated scores could provide estimates of p-values for other scoring systems.

2. OUTLINE OF THE APPROACH

The goal of this work is development of a simple, yet accurate, empirical model of random structure superpositions. The model consists of three parts, as briefly outlined next and then detailed below (Fig. 1). First, a dataset of minimum RMSD superpositions (allowing rigid body translation and rotation) of random protein pairs of various size and shape was assembled. These superpositions were considered "random,"

FIG. 1. (*Continued*) were extracted and approximated with a PDF—Nakagami distribution (Nakagami, 1960), as described in Methods. Parameters for this PDF were therefore dependent on size and shape of the random superpositions contained within the narrow range, or "slice," of σ_{exp}. These parameters were subsequently fit to a continuous curve with σ_{exp} as the independent variable (as described in Methods and shown in Figs. 5a and 5b). The continuous curve allowed construction of a random model, and therefore estimation of a p-value, for a superposition of arbitrary size and shape. As explained in the text but not shown in the figure, PDF parameters describing a distribution of σ_{obs} simultaneously described the corresponding distribution of projections $\Delta v_{x,y,z}$, using a second PDF, the Variance-Gamma distribution (Madan, 1990) (Supplementary Fig. S1, step III).

since participating proteins were not homologous and superpositions were not optimized to highlight local structural similarities. Second, an expression was found to estimate the expected RMSD (actually a proportional quantity, σ_{exp}) of such a random superposition from the size and shape of its constituent molecules. This expression could then be used to "predict" the expected random RMSD, given the molecular size and shape of a superposition of interest. Third, an expression for the probability density function (PDF) of the RMSDs of random superpositions was proposed, conditioned upon the value of expected RMSD obtained on the second step. In other words, the observed distribution of random RMSDs (actually a proportional quantity, σ_{obs}) was modeled for those pairs of molecules with a similar expected random RMSD. Using this PDF, one could estimate a p-value of a superposition of interest under the random model of gapless superpositions of proteins with similar size and shape.

In more detail (Supplementary Fig. S1), the dataset was constructed from minimum RMSD superpositions of non-homologous proteins. Superpositions were generated not for the purpose of finding structural similarities; rather, the alignment was "forced" throughout, matching a randomly chosen residue of the first structure to a randomly chosen residue in the second structure. This forced alignment was continued without gaps until the end of the shorter structure was reached. Then, rigid body translation and rotation to minimize RMSD between the two molecules was used, resulting in a random procedure analogous to that usually applied to non-random superpositions between structurally similar proteins. Since structurally equivalent residues were not found to optimize a certain function, but instead were randomly forced, these minimum RMSD superpositions were defined as "random" for the purpose of this work. To avoid possible, although rare, random matches between homologous segments, homologous pairs of proteins were not considered. Pairwise superpositions of complete proteins and pairwise superpositions of protein fragments were separately analyzed.

Coordinate difference vector projections (i.e., the projections onto the coordinate axes of distances between equivalenced atoms) were focused upon, since they determined the RMSD and other common measures of structural similarity—for example, TM (Zhang and Skolnick, 2004), GDT_TS (Zemla et al., 1999), and 3D (Rychlewski et al., 2003) scores. For instance, as shown in Equations (4) and (5), below, RMSD was proportional to the standard deviation, σ_{obs}, of these projections. For each superposition, projections of individual difference vectors onto the x-, y-, and z-axes were computed from the equivalenced alpha-carbon (CA) coordinates (step I of Fig. 1). These projections were pooled and tabulated along with the corresponding length, radius of gyration, eigenvalues of the coordinate inertia matrix, and RMSD of the molecular pair.

Because distances between equivalenced atoms of a random superposition, and thus the distribution of difference vector projections (DVPs), depended strongly on molecular properties, such as size and shape; it was next desired to model this dependence. As a first step in development of that model, it was demonstrated that a distribution of standardized DVPs taken from many superpositions was approximately Gaussian. For each superposition, the standardization was accomplished with division of each DVP by the standard deviation, σ_{obs}, of the distribution of DVPs for that particular superposition. Subtraction of the mean was not needed for standardization, since the mean of DVPs already equaled zero due to translation of coordinate centers of mass to a common point by the minimum-RMSD algorithm. The resulting Gaussian suggested that a distribution of DVPs x for a sample of superpositions could be modeled by a PDF $F(\sigma) * G(x, \sigma)$, where $*$ denotes a convolution, $G(x, \sigma)$ is a PDF of a Gaussian with zero mean and standard deviations σ, and $F(\sigma)$ is a PDF of standard deviations σ of DVPs in this sample of superpositions.

Clearly, PDF parameters estimated from the data from shorter superpositions were expected to be different from parameters obtained from longer superpositions. To model this dependence of PDF parameters on size and shape, the entire set of superpositions was divided into narrow subsets, or "slices," to incrementally estimate parameters for the entire set. Each slice was constructed such that molecules within a slice shared similar size and shape characteristics, as approximated by a function estimating the expected standard deviation of a superposition (σ_{exp}, a quantity proportional to RMSD, step II of Fig. 1). This function was chosen to be a second degree polynomial of two size and shape measurements: radius of gyration of the molecule, R_g, and the size characteristic along the thinnest dimension of the molecule as measured by the square-root of the smallest eigenvalue of its inertia matrix, c. In other words, c was defined as the standard deviation of the DVPs along the smallest principal component. Basing shape characteristics only on the first molecule of the pair mimicked the case of a structural database search, where the first molecule was "query" and the second molecule was "hit." Alternatively, basing shape characteristics on both molecules of the pair (considering the two sets of superimposed coordinates as a single "molecule") accounted for the different case of an isolated pairwise structural comparison.

The two shape measurements were empirically chosen from a large range of combinations in order to minimize the number of parameters in the function while also minimizing the variance of observed standard deviations within each slice. Ideally, if observed standard deviations of all random superpositions within a slice were exactly the same as expected, the variance of these standard deviations would be equal to zero. In other words, the PDF of random standard deviations would be a delta function placed at the expected standard deviation value σ_{exp}. In such a case, any standard deviation resulting from a non-random comparison would differ from σ_{exp} and thus would have a p-value of zero. Therefore, the function to estimate the expected standard deviation (step II of Fig. 1) was developed to minimize, on average, the variances of observed standard deviations for each narrow bin, or "slice," of expected standard deviation. The smaller these variances were, the more significant p-values would be for non-random superpositions, which would allow for a more sensitive evaluation of similarities between proteins.

Function $F(\sigma)$ (PDF of standard deviations σ of DVPs), with its dependence on the length and shape of superimposed molecules, completely defined the probabilistic model. On the one hand, $F(\sigma)$ gave the PDF of RMSDs for random superpositions of proteins with approximately the same size and shape (since σ was proportional to RMSD); on the other hand, PDF of DVPs could be obtained from it through convolution with a Gaussian. Since it did not seem feasible to find $F(\sigma)$ from theoretical considerations, a data modeling approach was taken. The goal was to choose a simple function that well approximated these data on random superpositions and also had a closed form expression after convolution with a Gaussian. In addition, the function $F(\sigma)$ was required to be zero for negative σ, because standard deviations were non-negative.

Theoretically and historically, the Gamma family of equations has been frequently employed to model distributions of variances, for example, the distribution of variances of samples from a normal distribution follows a Chi-square distribution, which is a special case of Gamma (Johnson et al., 1994). To generalize a two-parametric Gamma density to three parameters, the Generalized Inverse Gaussian (GIG) distribution was chosen (Johnson et al., 1994; Jorgenson, 1982). This distribution is widely used in statistical linguistics (Sichel, 1975), finance (Barndorff-Nielsen, 2001; Eberlein, 1995), and geostatistics (Barndorff-Nielsen, 1977). Its convolution with Gaussian results in a closed form, known as a symmetric case of generalized hyperbolic distribution (Barndorff-Nielsen, 1977). The GIG distribution includes Gamma and Inverse Gamma, Inverse Gaussian and Reciprocal Inverse Gaussian distributions as its special cases. In practice, we found the three-parametric GIG form to be unnecessary, and both two-parametric special cases—Gamma (Johnson et al., 1994) and Inverse Gaussian (Johnson et al., 1994; Tweedie, 1957)—gave statistically acceptable approximations to the data on random superpositions. Approximations did not significantly improve when the third parameter was included. For the purpose of this work, we approximate a distribution of standard deviations of random superpositions with the Nakagami distribution (Nakagami, 1960), a distribution describing the square-root (standard deviation) of a Gamma variable (variance). The statistical quality of its parameter estimates, as assessed by the method of minimum chi-squares, was equal to the best out of several two-parametric forms tested (Supplementary Fig. S2) and did not differ substantially from the quality of three-parametric GIG-based approximation.

Parameters s and k of the Nakagami distribution were estimated for samples of standard deviations of DVPs for superpositions of proteins with approximately the same size and shape (step III of Fig. 1). PDFs of samples of DVPs were described by the convolution with a Gaussian, $Nakagami(\sigma, s, k) * G(x, \sigma)$, expressed in closed form by means of a modified Bessel function of the second kind, $K_n(x, s, k)$ (Step III of Supplementary Fig. S1). The parameterization (s and k) was chosen such that the parameters had intuitive meaning: the quantity s corresponded to the standard deviation of DVP distributions as well as to the mean of the standard deviation distributions, while k was proportional to the inverse coefficient of variance squared of the standard deviation distributions and also was a simple function of the excess kurtosis of the DVP distributions.

Because distributions of DVPs and their corresponding standard deviations were related by a convolution and thus depended on the same parameters s and k, it was possible to obtain parameter estimates in one "space," for example, from a sample of standard deviations, and then check the correctness of these parameter values in the second "space," for example, the corresponding sample of DVPs. It was demonstrated that parameter estimates obtained in either space agreed and were statistically reasonable, confirming the validity of our approximations.

Finally, the estimated values of parameters $\{s, k\}$ were themselves approximated by a continuous function depending on the expected standard deviation of superpositions as computed from molecular size and shape (Fig. 1, step III). Then, the expected distribution of DVPs or corresponding standard deviations (i.e., RMSD) for a superposition of arbitrary length and shape could be constructed. These two distributions, conditioned upon molecular size and shape parameters, equally represented the random model of structure superposition for a particular superposition of interest.

Given the RMSD of superposition of interest and the ability to reconstruct the expected random distribution of non-homologous minimum RMSD superpositions of similar shape, the probability of a chance occurrence of an RMSD equal or less than the one of interest could be estimated by integration of the expected distribution's probability density function. A powerful additional feature of this approach was that, in principle, statistical significance of *any* score based on difference vector projections (e.g., TM-score) could be estimated because that underlying distribution could also be reconstructed.

3. RESULTS

Two datasets of minimum RMSD superpositions of randomly drawn pairs of non-homologous proteins, one consisting of protein fragments and one consisting of whole proteins, were constructed as described in Methods. The "fragment" dataset contained 32,004 superpositions, and the "whole" dataset contained 22,561 superpositions. Coordinate DVPs were extracted from each superposition (shown schematically in step I of Fig. 1). Standard deviations, σ_{obs}, of the DVPs from each superposition were also calculated. Subsets (i.e., "slices") encompassing the complete sets of standard deviations and DVPs were subjected to statistical analysis as detailed in Methods, and the results presented directly following.

Figure 2a displays distributions of DVPs for collections of individual superpositions of a given length. Distributions derived from shorter chains exhibited sharper peaks while distributions derived from longer chains were broadened. A necessary prerequisite for the present analysis was standardization of these disparate distributions. When each individual projection was divided by the standard deviation of all projections from the superposition to which it belonged and the results pooled together for each length, each standardized distribution could be well approximated by a Gaussian distribution with mean zero, standard deviation 1 ($0.001 < p < 0.20$) (Fig. 2b), and thus was length-independent.

Statistical modeling might be performed simultaneously on the entire data set, that is, a global fit, or by parameter estimation performed on subsets of the data. The latter route was chosen, but instead of constructing subsets of superpositions by chain length, a procedure was adopted that subsumed chain length into overall molecular shape and size. A second-degree polynomial function based on two shape parameters, radius of gyration and the standard deviation of coordinates along the axis of the smallest spread of coordinates, was constructed (Equations (6A) or (6B) in Methods; shown schematically in step II of Fig. 1). This function estimated the expected standard deviation of projections, σ_{exp}, for the minimum RMSD superposition from overall molecular shape. The complete data was then "sliced" into narrow subsets of observed standard deviation, σ_{obs}, and projections as a function of σ_{exp} (Fig. 3). Each slice was constructed to contain approximately equal numbers of coordinate difference vector projections. To accomplish this, the number of superpositions was randomly decreased as σ_{exp} increased. The task was now reduced to modeling of every individual slice, shown schematically in step III of Figure 1.

The present approach allowed modeling of each slice in two different coordinate spaces: σ_{obs} space (Equation (8)) and difference vector projection space (Equation (10)). The advantage of modeling the data in this way was that parameter estimates obtained in either space could be used in PDFs of the second space and the agreement with data could be evaluated to ensure accuracy and consistency of the results. Figure 4 shows examples of the standard deviation space parameter estimates and projection space tests from two representative slices of the fragment data: Figure 4a displays distributions derived from very short chains (average length $N \sim 5$ residues), and Figure 4b displays distributions derived from longer chains (average length $N \sim 30$ residues). The longer chain distribution was approximated reasonably well ($p = 0.17$) in standard deviation space by a single Nakagami probability density function (Fig. 4b, inset), and the test in projection space (evaluation of Equation (10) with the best estimate $\{s, k\}$ parameters) was

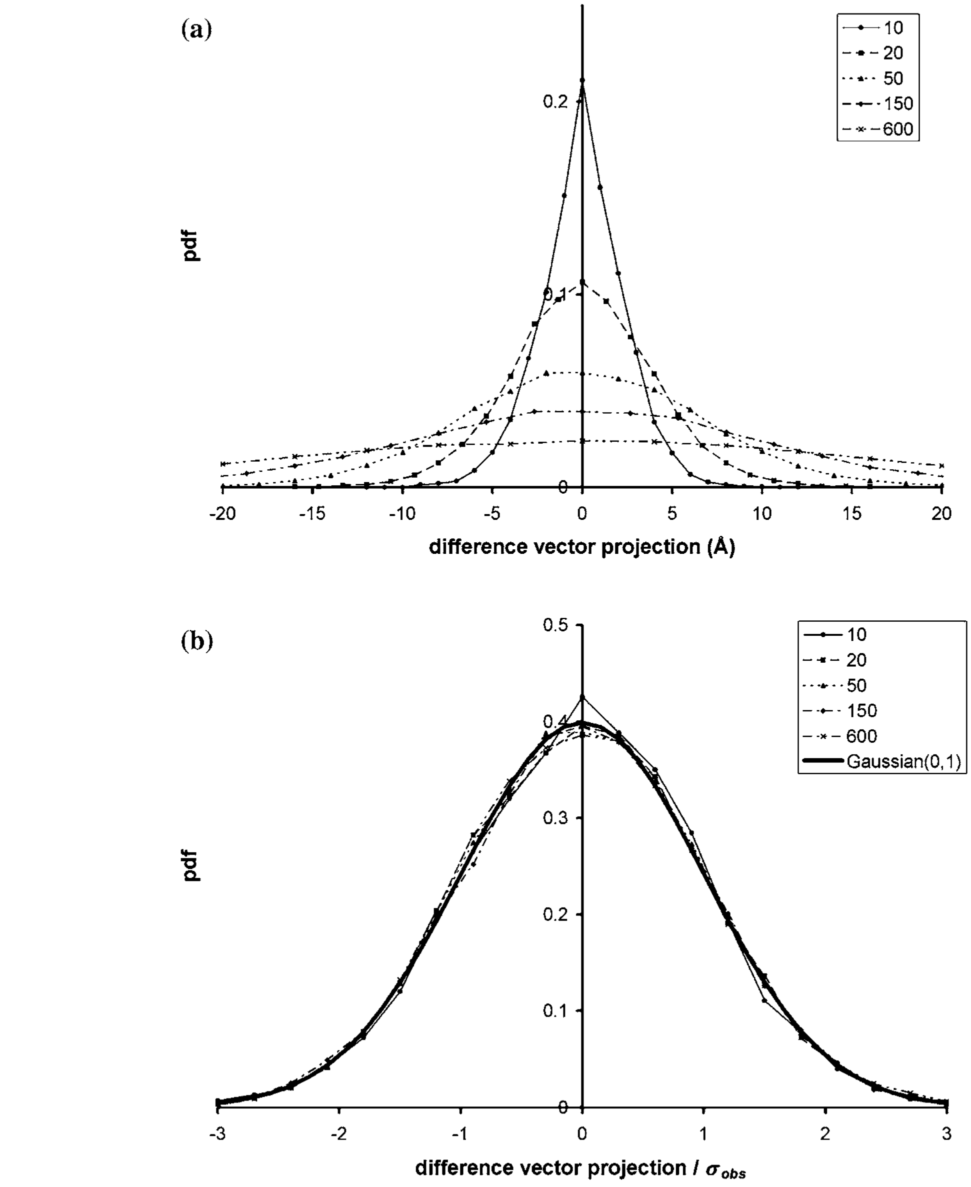


FIG. 2. Standardization of coordinate difference vector projections. (**a**) Probability distribution functions of unstandardized coordinate difference vector projections at fragment superposition lengths of 10–600 residues. Approximately 20,000 projections per length were partitioned into 50 bins of equal width for each curve. Shorter lengths exhibited a sharper peak about zero; longer lengths exhibited broadening. (**b**) Probability distribution functions of coordinate difference vector projections for the same distributions shown in a, where each projection was divided by standard deviation, σ_{obs}, of the superposition from which the projection was taken. A Gaussian probability distribution with mean of zero and standard deviation of 1 is shown for comparison (thick line). Distributions for all lengths were thus statistically the same as standard Gaussian.

statistically acceptable ($p = 0.02$) with a χ^2 value of 87.8 over 47 *d.o.f.* (Fig. 4b, red curve). However, it is stressed that a Gaussian curve with standard deviation equal to that observed from the projection distribution clearly did not describe these data well, with a χ^2 value of 304.5 over 47 *d.o.f.* (Fig. 4b, cyan curve, $p < 10^{-10}$).

In contrast, distributions from short chain lengths (less than ~15 residues) could not be approximated well with a single Nakagami PDF (Fig. 4a, inset, dark blue line). At these short lengths, randomly chosen structure fragments, even from proteins with no global structural similarity, would sometimes result in

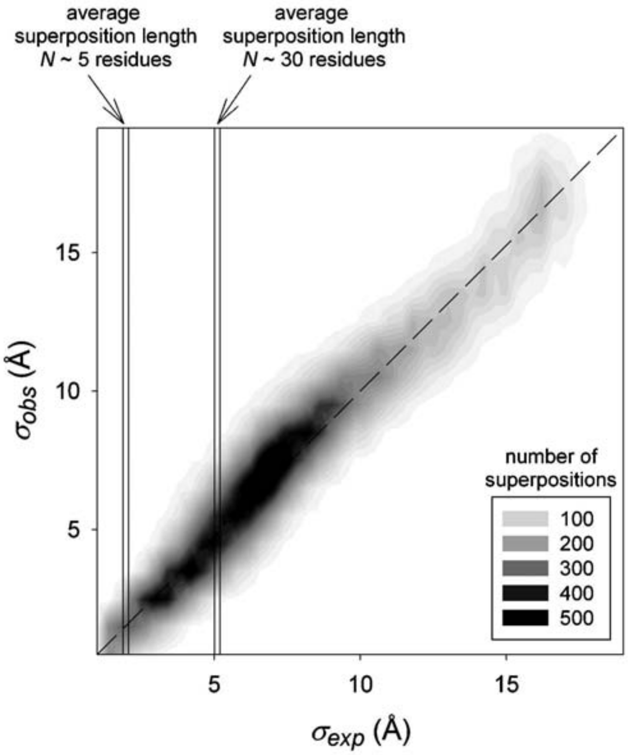


FIG. 3. Partitioning of minimal root mean square deviation (RMSD) superposition data into "slices" using an expected standard deviation function based on molecular size and shape. Dark contours represent 32,004 fragment superpositions, binned by expected (σ_{exp}) and observed standard deviation (σ_{obs}) of coordinate difference vector projections (Equation (6A)). Darker colors denote larger numbers of individual superpositions within a bin. Bin size for this figure was 0.4 Å in expected dimension and 1.0 Å in observed dimension. On average, σ_{obs} was approximately equal to σ_{exp}; a dashed identity line is drawn for comparison. Vertical "slices" of these data were analyzed at σ_{exp} intervals of 0.2 Å, resulting in two distributions for each slice: a distribution of σ_{obs} and a distribution of coordinate difference vector projections. Slices taken at larger values of σ_{exp}, corresponding to longer length superpositions, contained fewer numbers due to equalization of the total number of projections per slice, as discussed in Methods. Slices at σ_{exp} values of 1.6–1.8 Å and 5.0–5.2 Å, corresponding to average superposition lengths N of approximately 5 and 30, respectively, are highlighted. For details on the modeling of the distributions of σ_{obs} and coordinate difference vector projections contained in these two slices, see Figure 4.

nearly exact superpositions of two similar secondary structure elements. The major contributions were believed to come from helix-helix superpositions, but other secondary structure types certainly contributed to varying degrees. These nearly exact superpositions manifested in standard deviation space as a sharp peak near zero, or as a sharp symmetric peak around zero in DVP space. A second Nakagami PDF, based on explicit modeling of homologous superpositions (as described in Methods) was required to obtain acceptable approximations to these projection distributions (Fig. 4a, solid red line).

Estimated parameters $\{s, k\}$ are displayed in Figure 5. Parameter s for whole molecules (Fig. 5a) was approximately identical to the mean σ_{exp} for each slice, increasing proportionally to σ_{exp}. However, substantial curvature was observed at intermediate lengths in the fragment data, and using a linear approximation to estimate sigma turned out to not be statistically justifiable in subsequent work (data not shown). Therefore, an expression composed of a sum of logistic and linear functions (Equation (19)) was chosen to describe s for both fragments and whole molecules, representing a tradeoff between minimizing the number of parameters and quality of fit. Parameter s from homolog fragment data exhibited a curve of qualitatively different shape, rising to a maximum at an approximate chain length of 20 residues.

In contrast, the estimated parameter, k, from fragment data displayed a sigmoid shape as a function of σ_{exp} (Fig. 5b). Although it was not possible to collect whole molecule data that defined the entire curve, the existing data were also consistent with a sigmoid, slightly offset from the fragment data of similar chain

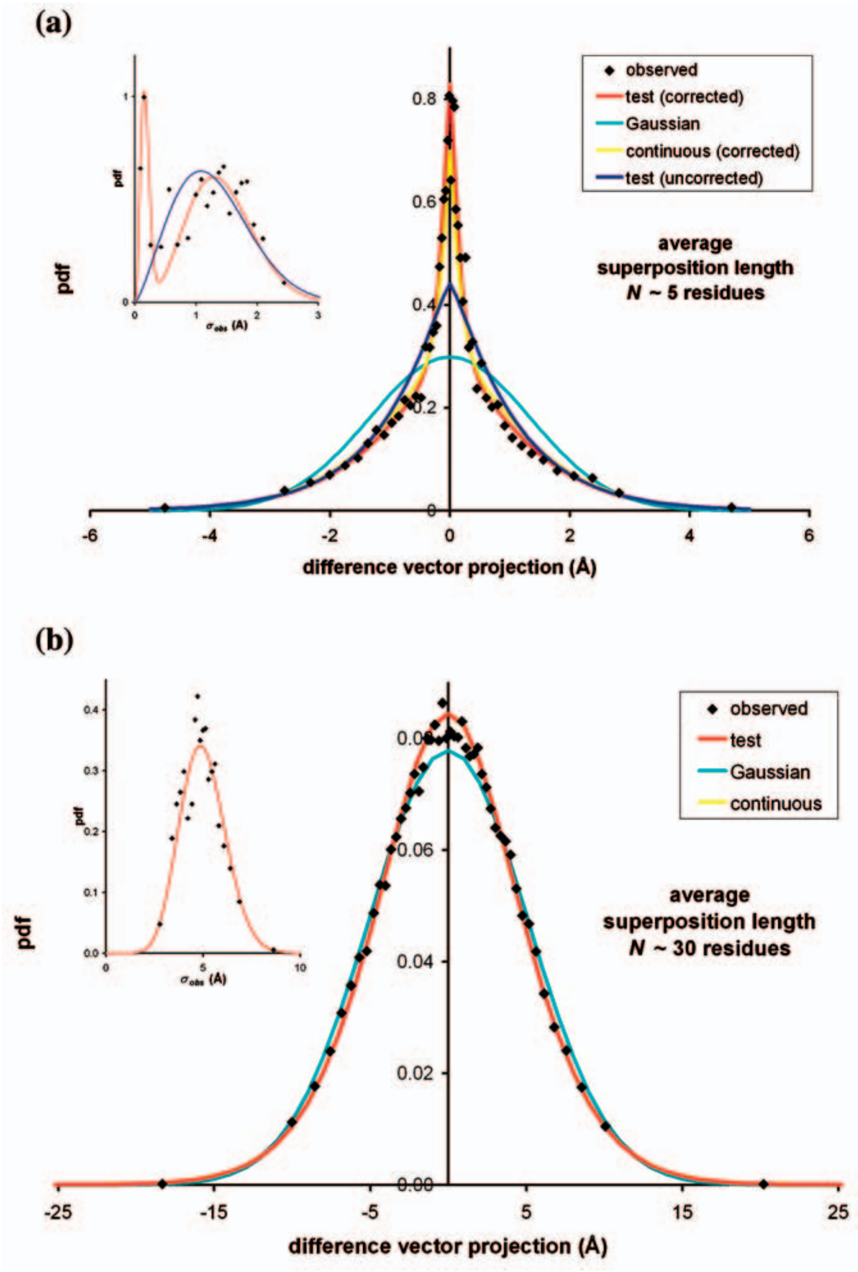


FIG. 4. Estimation of probability density function parameters for "slices" of observed standard deviation distributions and critical evaluation of these parameters using the corresponding coordinate difference vector projection distributions. **(a)** Distributions of coordinate difference vector projections and observed standard deviations, σ_{obs}, from expected standard deviation, σ_{exp}, "slice" 1.6–1.8 Å (average superposition length, ~5 residues). The σ_{obs} distribution, partitioned into equal-count bins whose normalized areas totaled 1, is shown in the inset. Black diamonds indicate observed values of the probability density function (PDF) at bin midpoints. A prominent peak near the σ_{obs} value of zero was due to the presence of short, nearly exact superpositions, as discussed in the text. The minimum χ^2 approximation of these observed values with a single Nakagami PDF is shown as a dark blue line and did not capture the prominent peak ($\chi^2 = 87.2$, 17 *d.o.f.*, $p < 10^{-10}$). The minimum χ^2 approximation of these data with the weighted sum of two Nakagami PDFs, implementing the short-length correction (Equation (18)), is shown as a solid red line and captured the prominent peak ($\chi^2 = 13.6$, 17 *d.o.f.*, $p = 0.52$). The larger figure shows the coordinate difference vector projection distribution corresponding to the inset σ_{obs} distribution, also partitioned into equal-count bins whose normalized areas totaled 1. The dark blue curve, "test (uncorrected)," resulted from evaluation of Equation (10) with single Nakagami PDF $\{s, k\}$ parameters estimated from the σ_{obs} distribution. The χ^2 value was 268.1 for 47 *d.o.f.* ($p < 10^{-10}$), indicating

(caption continues on next page)

length. Examination of the average lengths and radii of gyration (data not shown) associated with different regions of these curves suggested a physical rationalization of the protein conformations in each region. The limb of the sigmoid at short lengths, less than 50 residues, was approximately linear and represented chains that were extended or non-globular. The rapidly ascending middle portion of the curve, of lengths between 50 and 150, was composed of a mixture of extended, partially collapsed, and globular structures. The final linear limb of the curve, at lengths greater than 150 residues, was composed almost exclusively of single domain, globular structures. As the initial and final limbs were of similar slope, the sigmoid was fit with a sum of linear and logistic functions (Equation (21)). Estimated parameter k from homolog data exhibited a curve of qualitatively different shape, rising to a maximum at an approximate chain length of 20 residues.

Parameter estimates giving statistically acceptable approximations of distributions ($0.01 < p < 0.97$, mean $= 0.38$) were obtained from all standard deviation distributions. For all projection vector distributions modeled with parameters estimated from the corresponding standard deviation distributions, reasonable χ^2 values ($\chi^2 < \sim 100$, $10^{-6} < p < 0.25$, mean $= 0.02$) were obtained after the short chain homolog correction was applied (Fig. 6a, red line). Use of a Gaussian to describe difference vector projection distributions was not statistically valid for chain lengths less than ~150 (Figs. 6a and 6b, cyan lines, $p < 10^{-10}$).

For each superposition scenario (i.e., fragments, whole molecules, and homologous fragments), the estimated values for each parameter $\{s, k\}$ were fit by a continuous curve. A total of six continuous curves are displayed as solid lines in Figures 5a and 5b, representing these cases in which the size and shape of only the first molecule was taken into account. Thus, for a given slice, its midpoint σ_{exp} value defined on these curves values of $\{s, k\}$ that would allow construction of a PDF (in either standard deviation or projection space). These PDFs represented the expected random model of structure superpositions for the slice. To ensure that the expected random models implied by these continuous curves were reasonable ones, their PDFs could be compared to the observed projection distributions for each slice. Critical evaluation of χ^2 values (Fig. 6, yellow line) between PDFs and observed projection distributions, using parameters taken from the continuous curves for the PDFs, resulted in χ^2 values of similar magnitude ($\chi^2 < \sim 100$) to those resulting from PDFs using parameters estimated directly from the corresponding standard deviation distribution for each slice (Figs. 6a and 6b, red lines). Similarity of χ^2 values provided confidence that the continuous curves were accurate descriptions of the estimated parameters. In particular, χ^2 values for the shortest fragments, where the homolog correction had the greatest effect, were also in the same range as the tests at longer chain lengths.

Given these continuous curves, distributions of either standard deviations or DVPs of random superpositions could be reconstructed from a particular σ_{exp} value. Significances of arbitrary superpositions of interest could then be estimated with respect to the random distribution of standard deviation expected

FIG. 4. (*Continued*) a statistically unreasonable test of the fitted parameters. The red curve, "test (corrected)," resulted from evaluation of Equation (30) with $\{s, k, s_h, k_h, w_h\}$ parameters from the weighted sum of two Nakagami PDFs approximating the σ_{obs} distribution. The χ^2 value was 46.0 for 47 *d.o.f.* ($p = 0.47$), indicating a statistically reasonable test of the fitted parameters when the short-length correction was applied. The cyan curve, "Gaussian," resulted from evaluation of a Gaussian PDF whose standard deviation was set equal to the standard deviation of the difference vector projection distribution; its χ^2 value was 918.7 over 47 *d.o.f.*, indicating a statistically unreasonable test ($p < 10^{-10}$). The yellow curve, "continuous (corrected)," resulted from evaluation of the continuous approximation (Equations (20)–(24) to the estimated $\{s, k, s_h, k_h, w_h\}$ parameters; its χ^2 value was 81.9, indicating a statistically reasonable test ($p = 0.008$) and validity of the short-length correction. **(b)** Distributions of coordinate difference vector projections and standard deviations, σ_{obs}, from expected standard deviation, σ_{exp}, "slice" 5.0–5.2 Å (average superposition length ~30 residues). The σ_{obs} distribution is shown in the inset; the larger figure shows the corresponding projection distribution. The red solid curve, "test," resulted from evaluation of Equation (10) with single Nakagami PDF $\{s, k\}$ parameters estimated from the σ_{obs} distribution. The χ^2 value was 87.8 for 47 *d.o.f.* ($p = 0.003$), indicating a statistically reasonable test of the fitted parameters. The cyan curve, "Gaussian," resulted from a normal distribution with the standard deviation equal to the standard deviation of the observed difference vector projection distribution, exhibited a χ^2 value of 304.6 over the same *d.o.f.* ($p < 10^{-10}$), indicating a statistically unreasonable test. The yellow curve, "continuous," resulted from evaluation of the continuous approximation (Equations (20)–(24)) to the estimated $\{s, k, s_h, k_h, w_h\}$ parameters; its χ^2 value was 95.5 ($p = 0.001$), indicating a statistically reasonable test. This curve overlaps the "test" curve completely.

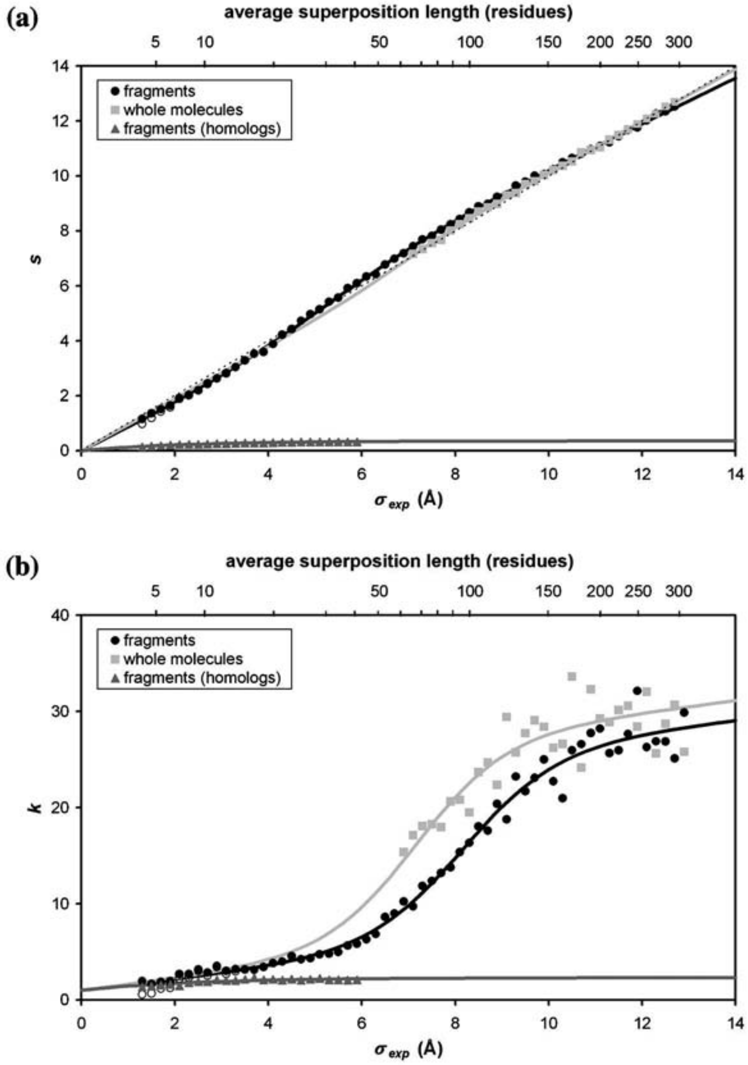


FIG. 5. Estimated and continuous parameters of Nakagami probability density function resulting from minimum χ^2 approximations to observed standard deviation distributions. **(a)** Plot of estimated s as a function of expected standard deviation of projections (σ_{exp}). Fragment data are plotted as dark filled circles; a linear/logistic fit of Equation (20) to these points is shown as a dark solid line. Parameters estimated for a single Nakagami probability density function (PDF) uncorrected for short chain length, and therefore not used for the continuous fit, are shown as unfilled circles. Whole molecule data are plotted as light gray filled squares; a linear/logistic fit of Equation (20) to these points is shown as a light gray line. Homolog fragment data are plotted as dark gray triangles; a saturating exponential growth fit of Equation (22) to these points is shown as a dark gray line. A dashed identity line is provided for comparison; the non-linearity of the fragment data is apparent. In contrast, the whole molecule data is more nearly linear, and the homolog fragment data is qualitatively different than either curve. **(b)** Plot of estimated k as a function of expected standard deviation of projections (σ_{exp}). Fragment data are plotted as dark filled circles; a linear/logistic fit of Equation (21) to these points is shown as a dark solid line. Parameters estimated for single Nakagami PDFs uncorrected for short chain length, and therefore not used for the continuous fit, are shown as unfilled circles. Whole molecule data are plotted as light gray filled squares; a linear/logistic fit of Equation (21) to these points is shown as a light gray line. Homolog fragment data are plotted as dark gray triangles; a saturating exponential growth fit of Equation (23) to these points is shown as a dark gray line. The fragment and whole molecule curves are sigmoid, with the homolog fragment data qualitatively different from either. Protein structural characteristics are attributed to the linear and logistic phases of the sigmoid, as described in the text.

from molecules of similar size and shape, using Equation (27) as described in Methods. Examples of such estimates based on the shape parameters of both molecules are displayed in Figure 7. Figure 7a plots significances derived from structure alignments (performed by Dali), and Figure 7b plots significances derived from sequence alignments linked to structure (performed by COMPASS). Although there was scatter for the data belonging to the most similar (homologous) structural comparisons, a good correlation existed between the Dali Z-score and the estimated probability with respect to the random distribution developed in this work (Fig. 7a). A similar result obtained for sequence comparisons when GDT_TS scores were used to measure the structural similarities of the superimposed, aligned sequence fragments (Fig. 7b).

Significances for a second situation, where the size and shape of only one molecule in the superposition was considered, were computed. This situation mimicked the case of a database search, for example, where the shape of only the query molecule was known in advance. These values, while expected to be less significant than values based on the two-molecule case, correlated highly with those from the two-molecule case (Fig. 8). The degree of correlation and the amount of additional significance were similar regardless of whether whole molecules or fragments were considered.

Another application of the expected random models was explored, involving the DVP distribution. Four measures of structural similarity based on difference vector projections, RMSD, TM (Zhang and Skolnick, 2004), GDT_TS (Zemla et al., 1999), and 3D (Rychlewski et al., 2003) were studied (Equations (31)–(34). These four measures were computed for 1000 random fragment superpositions of lengths 4–600. These observed values were then compared to predicted values for each superposition, where the predicted values were computed using DVPs drawn from the expected random distribution appropriate to the shape and size of each individual superposition, as described in Methods. Results from these tests are displayed in Figures 9a–9d. Predicted values generally correlated with observed values, with Pearson correlation coefficients $r > 0.5$ for all four measures of structural similarity. RMSD exhibited the best correlation, $r = 0.93$, possibly because deviations between observed and randomly drawn projections were reduced when the square root was taken. Although 3D-score exhibited the worst correlation, $r = 0.54$, because deviations between observed and randomly drawn projections were amplified by the exponential, the absolute values of these predicted scores were nonetheless all within a narrow range, as expected for random scores in this scoring system. For several measures, saturation was noted at the highest observed values, corresponding to the longest superposition lengths. This was due to the eventual breakdown of the continuous approximations, as evidenced by the highest χ^2 values at the longest fragment lengths in Figure 6a. For these scoring systems based on DVPs, it was concluded that random distributions specific to the particular score system could be quickly approximated by the procedure and equations outlined above.

4. DISCUSSION

An empirical method was developed to estimate statistical significance of a pairwise protein structural superposition. Once parameterized, the method required as input only the superimposed coordinates of the two molecules, from which two shape parameters and the observed standard deviation (RMSD) were calculated. Output was a numerical probability estimate of the statistical significance of the superposition with respect to a random distribution of minimal RMSD superpositions of similar size and shape. Figure 10 displays a schematic showing the essential steps in the process; parameterized values for variables in the schematic are reported in Table 1. Besides its potential utility to quantitatively rank hits from searches of structural databases, it is expected that this method will find application as a reference-independent evaluator of structural alignment quality. Another useful feature of the method was its ability to compute random baselines for other scoring systems based on coordinate difference vector projections (Figs. 9a–9d). It must be kept in mind that in this latter application, the random baseline was always with respect to minimal RMSD superposition, regardless of the particular scoring system simulated. For practical use, however, differences were expected to be negligible.

In agreement with the conclusions of other reports (Jia and Dewey, 2005; Reva et al., 1998), it was found that the RMSD threshold was rather high for significance of a superposition. For example, considering a superposition between two fragments approximately 50 residues in length, an RMSD of approximately 6 Å would achieve a p-value at the 0.05 level. This result, while underscoring the vast conformational space that

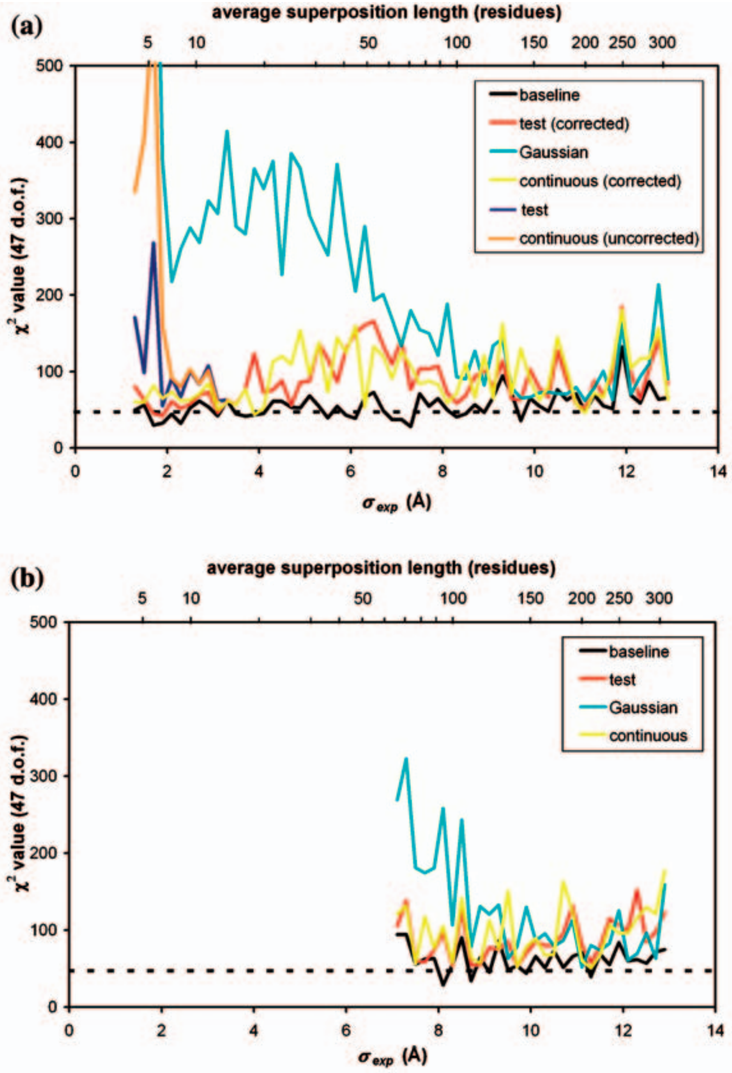


FIG. 6. Assessment of statistical validity of probability density function (PDF) parameter estimates using critical evaluations of PDFs against coordinate difference vector projection distributions. **(a)** χ^2 values for slices of fragment superposition data as a function of expected standard deviation, σ_{exp}. A dashed black line at $\chi^2 = 47$ marks an expected theoretical measure of statistical validity for a system with 47 *d.o.f.* The solid black line, “baseline,” represents an estimate of minimum possible χ^2, as described in Methods. This line hovers around $\chi^2 = 47$, creeping upward at $\sigma_{exp} > \sim 11$. The solid red, cyan, and yellow lines correspond to χ^2 results of the observed data compared with evaluation of Equation (30) with different $\{s, k\}$ parameters, as described in the legend to Figure 4. The dark blue and orange lines at the left of the plot correspond to tests performed without a short-chain correction, as described in the legend to Figure 4 and in the text. χ^2 values for red and yellow curves, generally values of 50–100, indicated that estimated and continuous parameters were statistically reasonable. In contrast, the cyan curve indicates that a simpler approximation of these observed data with Gaussian distributions was statistically unreasonable for superposition lengths less than ~150 residues, and that approximation of these data with one Nakagami distribution was statistically unreasonable for superposition lengths less than ~15 residues. **(b)** χ^2 values for slices of whole molecule superposition data as a function of expected standard deviation, σ_{exp}. The descriptions of the solid and dashed black lines are given in a. The solid red, blue, and yellow lines correspond to χ^2 results of these observed data compared with evaluation of Equation (10) with different $\{s, k\}$ parameters, as described in the legend to Figure 4. χ^2 values for solid red and yellow curves, generally between 50 and 100, indicate that estimated and continuous parameters were statistically reasonable. In contrast, the cyan curve indicates that approximation of these data with Gaussian distributions was statistically unreasonable for superposition lengths less than ~100 residues. No short-chain correction was necessary for whole molecule data.

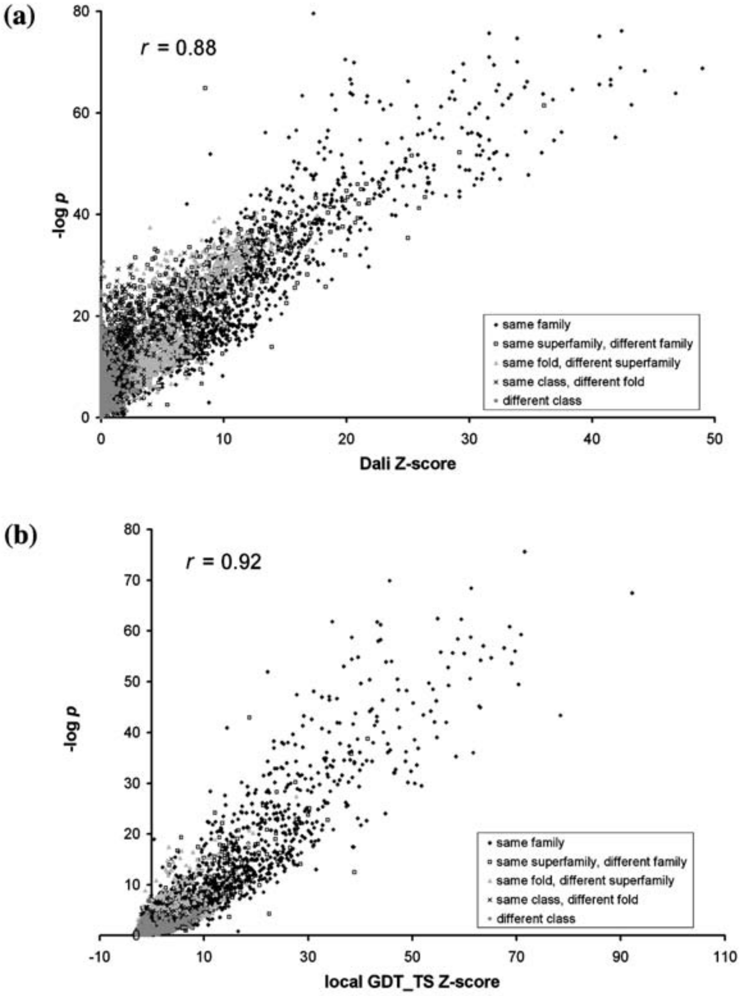


FIG. 7. Superposition p-values correlate with common measures of structural similarity. **(a)** Scatterplot of superposition p-values calculated for Dali (Holm and Park, 2000) structure alignments versus Dali Z-score. Approximately 5000 randomly selected pairs of structures with known relationship in the SCOP hierarchy (Andreeva et al., 2004) were selected for each of five possible relationship classifications. An optimal pairwise structural alignment was obtained from Dali, and its Z-score was plotted against the p-value. P-values were calculated with Equation (27), and parameterized for either whole molecule or fragment comparisons depending on the nature of the superposition, as detailed in Methods. Values were transformed by $-\log$ as indicated so that more significant probabilities had higher values on the y-axis. A strong correlation of Pearson correlation coefficient $r = 0.88$ was observed, indicating that p-values were good surrogates for Dali Z-scores. Substantial variability of p-values at Z-score $<\sim 2$ suggests that probabilities may be a more accurate assessment of structural similarity when the similarity is not extensive. **(b)** Scatterplot of superposition p-values versus GDT_TS score (Zemla et al., 1999), calculated from structure-linked sequence alignments. Approximately 5000 randomly selected pairs of structures with known relationship in the SCOP hierarchy were selected for each of five possible relationship classifications. An optimal pairwise alignment was obtained for profile comparisons using COMPASS (Sadreyev and Grishin, 2003), and the local GDT_TS score was calculated for the minimal RMSD superposition of the structural fragments equivalenced according to the COMPASS alignment. The local GDT_TS score was transformed to a Z-score according to the procedure Qi (2007), as described in Methods. The Z-score measured the number of standard deviations above average the GDT_TS score of the superposition was, as compared to structurally unrelated fragments of similar length. P-values were calculated with Equation (27), parameterized for either whole molecule or fragment comparisons depending on the nature of the superposition, as detailed in Methods. A strong correlation of Pearson correlation coefficient $r = 0.92$ was observed, indicating that p-values were good surrogates for the GDT_TS Z-scores.

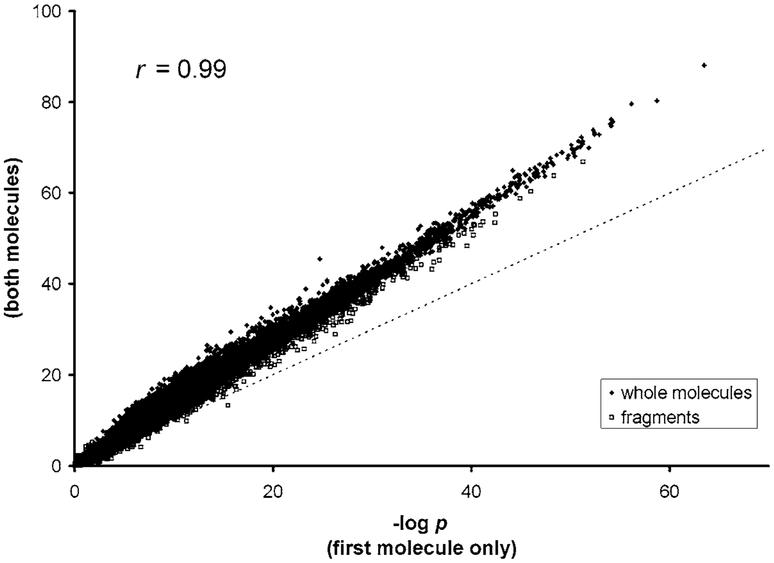


FIG. 8. P-values calculated from size and shape parameters of both molecules are more significant than, yet highly correlated with p-values calculated from shape parameters of one molecule only. Each data point represents a structural superposition based on a Dali (Holm and Park, 2000) structure alignment (whole molecules, filled diamonds) or a COMPASS (Sadreyev and Grishin, 2003) sequence-based structure alignment (fragments, open squares), as described in the legend to Figure 7. P-values were calculated for the approximately 5000 superpositions described in the legend to Figure 7, using Equation (27) parameterized for cases I and II of Table 1 (x-axis) or parameterized for cases III and IV (y-axis). For cases I and II, size and shape parameters of only one molecule were considered in the calculation, but for cases III and IV, parameters of both molecules in the superposition were considered. Values were transformed by $-\log$ as indicated; more significant probabilities had higher values. The Pearson correlation coefficient between single and both molecule probabilities was $r = 0.99$, and the slope of the best linear fit was approximately 1.3, with a y-intercept of approximately zero. A dotted identity line is shown as a guide.

was available to the structural random model, demonstrated that superpositions of only modest similarity would exhibit enormously significant absolute values of p. This also indicated that in practical work (comparisons involving structurally similar fragments), p-values would critically depend on the behavior of the tail of the probability density function used to describe σ_{obs}. Indeed, it was also observed that different functional forms of the PDF (e.g., Reciprocal Inverse Gaussian instead of Nakagami) could give very different, albeit non-linearly correlated, absolute p-values for such superpositions (Supplementary Fig. S3), despite exhibiting similar χ^2 values for their respective statistical parameter estimates (Supplementary Fig. S2). This effect was due to the difference in heaviness of the tails of the PDFs. Caution was therefore advised when interpreting p-values of superpositions between homologous proteins in an absolute, instead of a relative, sense.

Four different cases of structural comparison were parameterized: whole molecules or fragment comparisons, and database searches or isolated pair comparisons (Table 1). In general, probabilities for whole molecule comparisons were more significant than probabilities for fragment comparisons of similar size and shape, due to the wider conformational diversity of fragments. Because more shape information was known *a priori* for comparisons involving both molecules of a pair, probabilities for that case were generally more significant than probabilities for comparisons in which the shape information of only one molecule was considered.

Several design choices and approximations were made in parameterization of the method. The underlying philosophy for these choices always was accurate modeling of the data with the simplest equations and fewest variables. However, the results of analysis and the conclusions drawn appeared insensitive to the mathematical details. In particular, modeling of variance as well as standard deviation distributions was attempted with several different equations from the same functional family, Generalized Inverse Gaussian (GIG) (Johnson et al., 1994). As expected, essentially identical statistical parameter estimates to those reported in Figure 6 were obtained with a two-parameter Gamma function applied to variance distributions

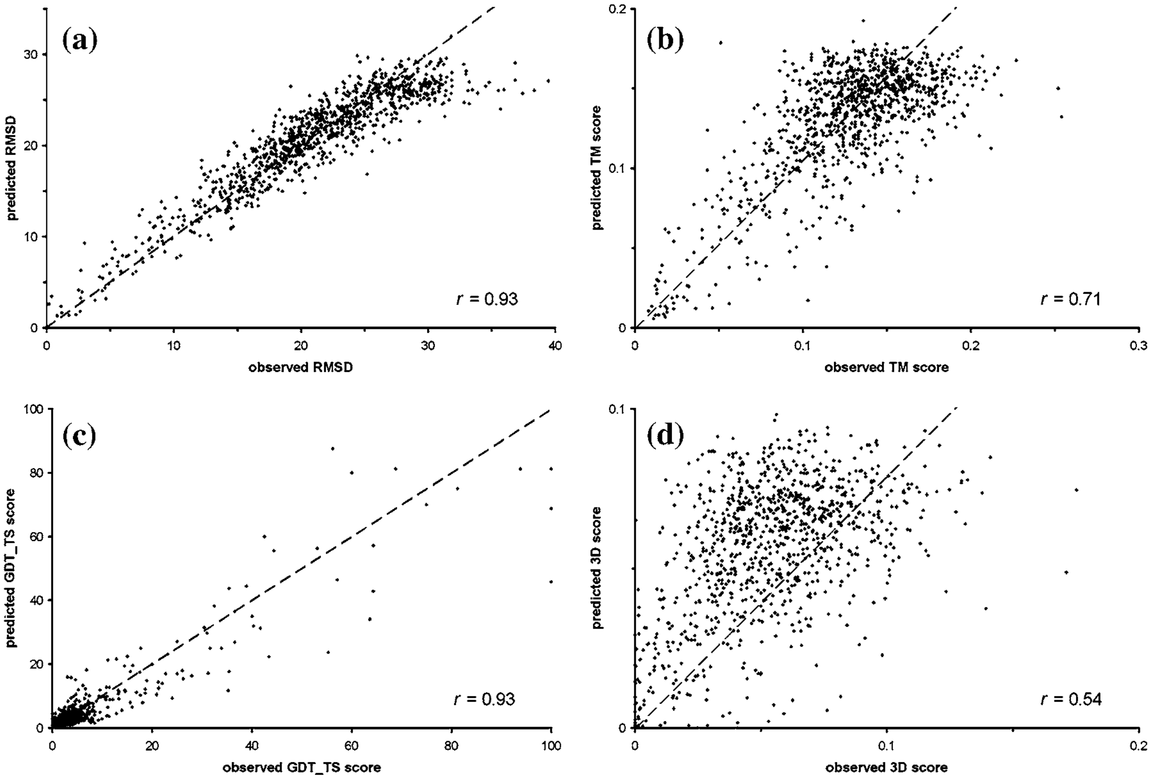


FIG. 9. Observed scores from four structural superposition scoring systems correlated with predicted scores drawn from expected random coordinate difference vector projection (DVP) probability density functions (PDFs). 1000 optimal root mean square deviation (RMSD) superpositions of structurally unrelated fragments of lengths 4–600 were generated, and observed values for four scoring systems; RMSD, TM (Zhang and Skolnick, 2004), GDT_TS (Zemla et al., 1999), and LiveBench 3D score (Rychlewski et al., 2003) were tabulated. Predicted values were computed by uniform random drawing of difference vector projections from expected random distributions appropriate to size, shape, and length of the superposition, as described in Methods. Dashed identity lines are shown for comparison. Strong correlations suggested that random values for many scoring systems based on coordinate difference vector projections could be estimated using the present approach. **(a)** A strong correlation of Pearson $r = 0.93$ was observed for RMSD. **(b)** A modest correlation of Pearson $r = 0.71$ was observed for TM score. **(c)** A strong correlation of Pearson $r = 0.93$ was observed for GDT_TS score. **(d)** A modest correlation of Pearson $r = 0.54$ was observed for 3D score.

(Supplementary Fig. S2). In addition, modeling with a three-parameter generalized GIG equation resulted in only negligible improvements of the parameter estimates and did not alter the conclusions obtained with the simpler two-parameter Nakagami distribution.

One unavoidable complication was the introduction of a second Nakagami distribution to describe random superpositions of fragments of shortest chain length ($L < {\sim}15$ residues). This second curve was necessary because a single function, representing a random component, did not accurately describe the observed distributions (Fig. 4a). Short, nearly exact, fragment superpositions taken from homologous proteins were used to parameterize this second curve. However, the second curve, although simple in conception and successful in statistically improving the goodness-of-fit, probably did not represent an accurate physical picture of all the molecular species present. More likely a continuum of species contributed, ranging from exact to partial structural identity. Examination of the standard deviation distribution in the inset of Figure 4a, for example, revealed density at a standard deviation of approximately 0.5 Å, not well approximated by either the homolog component or the random component of Equation (25). However, a more sophisticated correction was judged to not be worth the modeling effort, as the magnitude of the crude correction was small and essentially had no effect at chain lengths greater than approximately 15 residues (i.e., the correction was not applicable to whole molecule superpositions of naturally occurring proteins).

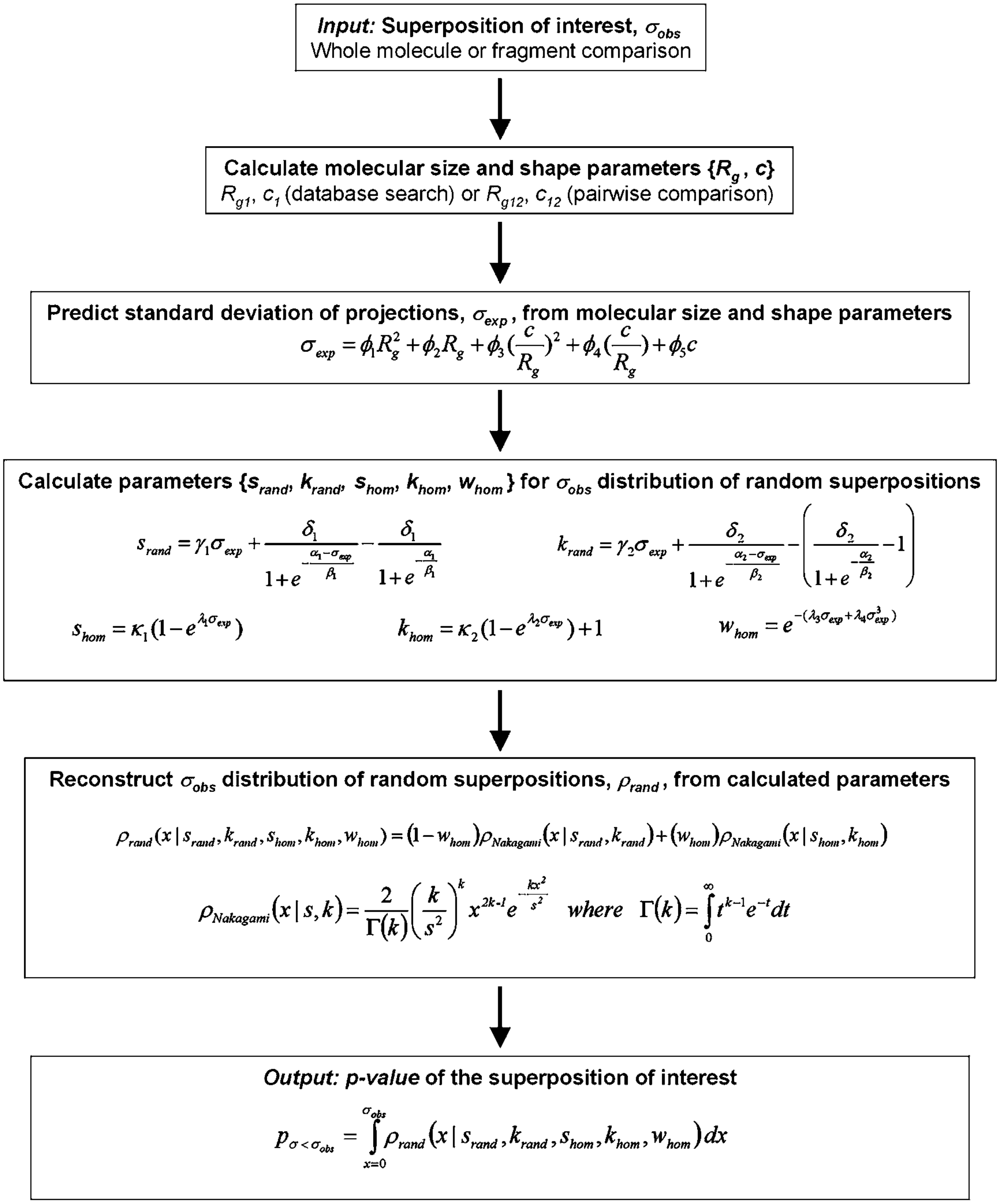

Input: Superposition of interest, σ_{obs}
Whole molecule or fragment comparison
Calculate molecular size and shape parameters $\{R_g, c\}$
R_{g1}, c_1 (database search) or R_{g12}, c_{12} (pairwise comparison)
Predict standard deviation of projections, σ_{exp}, from molecular size and shape parameters
$\sigma_{exp} = \phi_1 R_g^2 + \phi_2 R_g + \phi_3 (\frac{c}{R_g})^2 + \phi_4 (\frac{c}{R_g}) + \phi_5 c$
Calculate parameters $\{s_{rand}, k_{rand}, s_{hom}, k_{hom}, w_{hom}\}$ for σ_{obs} distribution of random superpositions
$s_{rand} = \gamma_1 \sigma_{exp} + \frac{\delta_1}{1+e^{-\frac{\alpha_1 - \sigma_{exp}}{\beta_1}}} - \frac{\delta_1}{1+e^{-\frac{\alpha_1}{\beta_1}}}$
$k_{rand} = \gamma_2 \sigma_{exp} + \frac{\delta_2}{1+e^{-\frac{\alpha_2 - \sigma_{exp}}{\beta_2}}} - \left(\frac{\delta_2}{1+e^{-\frac{\alpha_2}{\beta_2}}} - 1 \right)$
$s_{hom} = \kappa_1 (1 - e^{\lambda_1 \sigma_{exp}})$
$k_{hom} = \kappa_2 (1 - e^{\lambda_2 \sigma_{exp}}) + 1$
$w_{hom} = e^{-(\lambda_3 \sigma_{exp} + \lambda_4 \sigma_{exp}^3)}$
Reconstruct σ_{obs} distribution of random superpositions, ρ_{rand}, from calculated parameters
$\rho_{rand}(x \mid s_{rand}, k_{rand}, s_{hom}, k_{hom}, w_{hom}) = (1 - w_{hom}) \rho_{Nakagami}(x \mid s_{rand}, k_{rand}) + (w_{hom}) \rho_{Nakagami}(x \mid s_{hom}, k_{hom})$
$\rho_{Nakagami}(x \mid s, k) = \frac{2}{\Gamma(k)} \left(\frac{k}{s^2} \right)^k x^{2k-1} e^{-\frac{kx^2}{s^2}}$ *where* $\Gamma(k) = \int_0^\infty t^{k-1} e^{-t} dt$
Output: p-value of the superposition of interest
$p_{\sigma < \sigma_{obs}} = \int_{x=0}^{\sigma_{obs}} \rho_{rand}(x \mid s_{rand}, k_{rand}, s_{hom}, k_{hom}, w_{hom}) dx$

FIG. 10. Summary flowchart of significance estimation procedure. Parameters for each variable in different structural alignment cases (whole molecules versus fragments with or without short length correction, and database searches from a single query versus an isolated comparison of two molecules) are given in Table 1.

An interesting result was the interpretation of the estimated parameter, k, in terms of physical properties of the protein chain. A single sigmoid curve was fit to this parameter (Fig. 5b), where the initial limb of the sigmoid was observed to correspond to mainly extended conformations and the final limb was observed to correspond to more compact conformations. It was further noted that the observed sigmoid behavior of k contrasted with that expected from a simpler approximation built on the hypothesis that, for a sample of superpositions of chain length N, the corresponding difference vector projection distribution would be described by a Gaussian of mean zero and standard deviation s. This approximation yields the

Table 1. Estimated Values of Parameters to be Used in Probability Density Functions (PDFs) for Different Superposition Cases

Case I. Database search, fragment superposition.					
Calculate R_{g1} and c_1 from equivalenced coordinates of query molecule.					
Equation (6A):	$\phi_1 = 0.001687$	$\phi_2 = 0.282767$	$\phi_3 = -4.500119$	$\phi_4 = 0.719107$	$\phi_5 = 0.845287$
Equation (20):	$\alpha_1 = 5.27974$	$\beta_1 = 1.33866$	$\delta_1 = -2.01648$	$\gamma_1 = 0.827294$	
Equation (21):	$\alpha_2 = 8.17503$	$\beta_2 = 1.01879$	$\delta_2 = -20.1167$	$\gamma_2 = 0.570707$	
Equation (22):	$\kappa_1 = 0.35321$	$\lambda_1 = -0.443825$			
Equation (23):	$\kappa_2 = 1.2956$	$\lambda_2 = -0.409048$			
Equation (24):	$\lambda_3 = 1.0117$	$\lambda_4 = 0.1780335$			
Case II. Database search, whole molecule superposition.					
Calculate R_{g1} and c_1 from equivalenced coordinates of query molecule.					
Equation (6A):	$\phi_1 = -0.009484$	$\phi_2 = 0.695814$	$\phi_3 = 5.181274$	$\phi_4 = -7.140836$	$\phi_5 = 0.552282$
Equation (20):	$\alpha_1 = 6.80757$	$\beta_1 = 0.704487$	$\delta_1 = -0.645042$	$\gamma_1 = 0.94617$	
Equation (21):	$\alpha_2 = 7.16386$	$\beta_2 = -0.969174$	$\delta_2 = 21.6219$	$\gamma_2 = 0.609171$	
Case III. Isolated comparison, fragment superposition.					
Calculate R_{g12} and c_{12} from equivalenced coordinates of both molecules.					
Equation (6B):	$\phi_1 = 0.001047$	$\phi_2 = 0.296637$	$\phi_3 = -2.829648$	$\phi_4 = -1.824280$	$\phi_5 = 0.883476$
Equation (20):	$\alpha_1 = 5.42186$	$\beta_1 = 1.65218$	$\delta_1 = -2.13484$	$\gamma_1 = 0.826381$	
Equation (21):	$\alpha_2 = 7.83231$	$\beta_2 = 1.06344$	$\delta_2 = -26.2321$	$\gamma_2 = 0.869058$	
Equation (22):	$\kappa_1 = 0.338755$	$\lambda_1 = -0.551165$			
Equation (23):	$\kappa_2 = 1.21182$	$\lambda_2 = -0.538371$			
Equation (24):	$\lambda_3 = 1.89722$	$\lambda_4 = 0.1082705$			
Case IV. Isolated comparison, whole molecule superposition.					
Calculate R_{g12} and c_{12} from equivalenced coordinates of both molecules.					
Equation (6B):	$\phi_1 = -0.006106$	$\phi_2 = 0.692760$	$\phi_3 = 4.869105$	$\phi_4 = -7.476495$	$\phi_5 = 0.438529$
Equation (20):	$\alpha_1 = 10.6237$	$\beta_1 = 0.358413$	$\delta_1 = 0.169605$	$\gamma_1 = 1.01401$	
Equation (21):	$\alpha_2 = 7.00908$	$\beta_2 = -0.681195$	$\delta_2 = 32.8527$	$\gamma_2 = 0.620304$	

Nakagami density with $k = N/6$, where N equaled the chain length. In such a model, that probability density function would describe a random variable, ζ:

$$\zeta = \sqrt{\frac{\sum_{i=1}^{3N} \xi_i^2}{3N}} \tag{1}$$

In Equation (1), ξ_i are N independent Gaussian random variables with mean zero and standard deviation s (i.e., the s parameter of the Nakagami density, Equation (8), below). In the context of the present work, ξ_i are coordinate difference vector projections taken from a minimum RMSD superimposed pair of theoretical protein fragments of chain length N, and ζ is approximately the superposition's σ_{obs}. When the average superposition length in each slice of σ_{exp} was used as N, this simple approximation reproduced the k parameter for fragment lengths less than 150 residues (Supplementary Fig. S4). However, at longer fragment lengths, the k of the approximation increased faster than the observed k, suggesting that real proteins did not conform to the simpler model. One reason for this observed behavior might be the presence of multiple domains at longer chain lengths. Although the statistically reliable data reported here only reached a maximum chain length not much longer than the length of an average globular domain, it was speculated that data from longer, multi-domain proteins would result in multiple sigmoids. In this hypothesis, each additional sigmoid would therefore represent another level of structural transition, analogous to hydrophobic collapse, where multiple globular domains, joined by flexible linkers, would associate to varying degrees. Unfortunately, the numbers of known multi-domain structures were currently too sparse to test this speculation.

Finally, similarities and differences between the described approach to significance estimation and two previously published methods were assessed. The first method used large numbers of structurally dissimilar decoys of similar size and shape to members of a set of nonhomologous real proteins to obtain distributions of minimal RMSD superpositions (Reva et al., 1998). The distributions of RMSDs from the superpositions were approximated by Gaussians whose means were dependent on chain length and whose standard deviations were observed to be invariant. The probability of chance occurrence of the RMSD of a superposition of interest with respect to minimal RMSD superpositions of compact nonhomologs could be obtained by integration of the Gaussian appropriate to chain length. The second method was similar in spirit to the present work, analyzing minimal RMSD superpositions of structurally dissimilar protein fragments (Jia and Dewey, 2005). From those superpositions, a quantity termed the "reduced RMSD," essentially RMSD divided by radius of gyration expected from random superpositions, was approximated by an extreme value distribution. The reduced RMSD brought superpositions of arbitrary length and size to an identical scale. The probability of chance occurrence of the RMSD of a superposition of interest with respect to minimal RMSD superpositions of nonhomologs could be obtained by integration of the extreme value distribution.

Probabilities computed using the present approach and the two existing approaches were compared for approximately 5000 Dali superpositions (Fig. 11), where each pair of proteins had a known evolutionary relationship reported by the SCOP database, as described in Methods. Strong correlations were noted between p-values calculated from all three approaches. Probabilities calculated from the compact decoy approach were correlated with (Pearson $r = 0.90$), but generally two-fold log units lower than, those of the present approach (Fig. 11a). Probabilities calculated from the reduced RMSD approach were also correlated to, but generally five-fold log units greater than, those of the present approach (Fig. 11b). However, the correlation was weaker (Pearson $r = 0.73$) than that of the first approach, and many off-diagonal points were observed (Fig. 11b, lower right corner). This region of the plot contained points whose calculated significance was high according to the reduced RMSD approach, but low according to the present approach. These points corresponded to structurally compact superpositions of less than 100 residues with low RMSD values, ~1 Å (data not shown).

It was hypothesized that the off-diagonal points were the result of technical differences between the two methods that were sensitive to the small, compact, and low RMSD characteristics of these particular superpositions. Specifically, the p-values from the reduced RMSD approach were relatively size-independent due to the dependence of p-value on only the ratio of RMSD to radius of gyration (Equation (39)). In contrast, the p-values calculated in both the present approach and that of Reva et al. depended on both the RMSD (i.e., σ_{obs}) and the location (i.e., the mean of the PDF) of the random model for its particular size and shape (Equations (27) and (38), respectively). For small, compact, low RMSD superpositions, the reduced RMSD of the former method would be relatively small, resulting in a high significance. In contrast, the random models of the latter methods for the same small, compact structures would be relatively close in RMSD to the RMSD of the superposition of interest, resulting in a low significance. Despite these differences, it is again emphasized that p-values calculated from all three approaches were correlated with each other, and with common measures of structural similarity such as Dali Z-scores (Supplementary Fig. S5) and GDT_TS scores.

5. METHODS

5.1. Preparation of protein structure set

A set of 7290 protein structure coordinate files based on the ASTRAL_40 compendium (version 1.69) (Chandonia et al., 2004) was used. This set had been processed (B.H. Kim, unpublished data) from the original ASTRAL_40 files to create alpha-carbon (CA) only files where all non-standard amino acid types were changed to alanine and residue numbering was standardized, that is, coordinates of a structure with N residues were uniquely identified such that residue numbers ran consecutively from 1 to N. 294 structures classified as multidomain or membrane proteins—SCOP (Andreeva et al., 2004) classes e and f—were removed. 104 structures that Dali (Holm and Park, 2000) did not run for technical reasons (B.H. Kim, unpublished data, mostly extremely short chain lengths) were removed. The most elongated 5% of the remaining structures were removed as follows. The degree of elongation was measured by the ratio of a^2, defined as the largest eigenvalue of the inertia matrix (variance of coordinates), to c^2, the smallest eigenvalue. Thus, structures with an a^2/c^2 ratio > 9.73 were removed. The final set was composed of

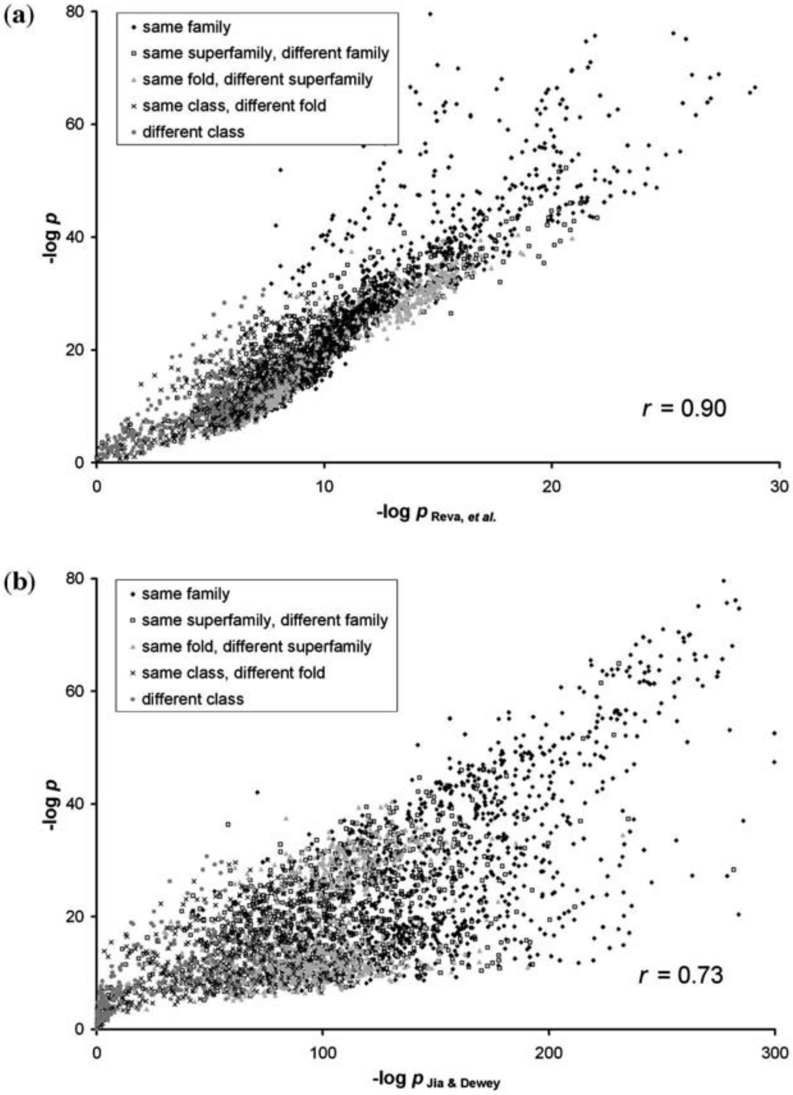


FIG. 11. Superposition probabilities correlate with those computed from two previously published approaches. Approximately 1000 randomly selected pairs of structures with known relationship in the SCOP hierarchy (Andreeva et al., 2004) were selected for each of five possible relationship classifications. An optimal pairwise structural alignment was obtained from Dali (Holm and Park, 2000), and the chance probability of observing a superposition of identical root mean square deviation (RMSD) was computed, either by the present approach or by an existing approach, as indicated. Probabilities of the present approach were calculated with Equation (27), parameterized for either whole molecule (case IV of Table 1) or fragment comparisons (case III of Table 1) depending on the nature of the superposition. Values were transformed by $-\log$ as indicated so that more significant probabilities have higher values. Probabilities of existing approaches were calculated according to the published procedures, as described in Methods. Superpositions of length less than 50 residues were not considered, as the existing approaches were developed based on analysis of lengths greater than approximately 50 residues. **(a)** Existing approach based on a random model of structurally compact decoys (Reva et al., 1998). A strong correlation of Pearson correlation coefficient $r = 0.90$ was observed. The present approach resulted in absolute significances approximately 2-fold log units greater than those of the existing approach. **(b)** Existing approach based on a random model of structurally dissimilar fragment superpositions (Jia and Dewey, 2005). A modest correlation of Pearson correlation coefficient $r = 0.73$ was observed. The present approach resulted in absolute significances approximately 5-fold log units less than those of the existing approach. Technical differences resulted in some superpositions exhibiting relative probabilities that differed greatly between the two approaches (lower right hand corner of b), as detailed in Discussion.

6551 structures. Statistics of chain length were: minimum = 31 residues, maximum = 1074, mode = 105, mean = 176.2, standard deviation = 107.2.

5.2. Selection of non-homologous protein pairs

For whole molecule structure comparisons, a pair of likely non-homologous proteins was selected as follows. Two non-identical proteins were chosen at random from the set described above. If the chain lengths of this pair were not within 10% of each other, the pair was discarded and a new pair selected. Otherwise, the pair was subjected to a support vector machine (SVM) homology filter (Qi, 2007). Briefly, the input for this SVM was five scaled (Z-scored) structure or sequence scores based on comparison of the pair of chosen proteins: the Dali Z-score (Holm and Park, 2000), the FAST (Zhu and Weng, 2005) score, the local GDT_TS (Zemla et al., 1999) score based on TM (Zhang and Skolnick, 2005) alignment, coverage over the length of the shorter protein based on FAST (Zhu and Weng, 2005) alignment, and the normalized BLOSUM62 score based on the Dali (Holm and Park, 2000) alignment. These input scores were precomputed for all possible pair comparisons in the protein structure set (B.H. Kim, unpublished data). An SVM score less than −0.6 indicated that the pair was not structurally similar, was likely non-homologous, and was therefore retained. The shorter protein was aligned without gaps starting at a randomly chosen residue position within the longer protein of the pair to allow a complete chain match, resulting in a gapless set of coordinate equivalences equal to the length of the shorter protein.

Fragment structure comparisons were accomplished in a slightly different manner. A superposition length was selected at random between 4 and 600 residues, inclusive. Two non-identical proteins were chosen at random from the set described above. If the chain length of either member of the pair was less than the selected superposition length, the pair was discarded and a new pair selected. Otherwise, the pair was subjected to the SVM homology filter as described above and only non-homologous pairs were retained. An alignment start position was randomly chosen in each member of the pair that allowed a gapless set of coordinate equivalences equal to the selected superposition length.

5.3. Minimum RMSD superposition of gapless aligned structures and preprocessing of data

Each non-homologous protein pair or protein fragment pair was aligned without gaps (i.e., the first residue of the first chain was equivalenced with the first residue of the second chain and the alignment gaplessly continued until the end of the shorter chain). To eliminate orientational bias, the first structure was given a rotation for a random angle around a random axis passing through the molecule's center of mass and the minimum RMSD superposition of the pair was calculated using a quaternion algorithm (Coutsias et al., 2004). A total of 180,000 superpositions were generated from whole proteins and a total of 220,000 superpositions were generated from protein fragments. A radius of gyration for both the first molecule of the pair (R_{g1}) and the pair itself (R_{g12}), coordinate difference vector projections ($\Delta v_{\{x,y,z\},i}$) for the optimal superposition of the pair, the standard deviation of coordinate difference vector projections for the optimal superposition of each pair (σ_{obs}), and the pair root-mean-square deviation ($RMSD_{obs}$) were calculated:

$$R_{g1} = \sqrt{\frac{\sum_{i=1}^{N}(x_{1,i} - c_{1,x})^2 + (y_{1,i} - c_{1,y})^2 + (z_{1,i} - c_{1,z})^2}{N}} \tag{2A}$$

$$R_{g12} = \sqrt{\frac{\sum_{i=1}^{N}(x_{1,i} - c_{1,x})^2 + (y_{1,i} - c_{1,y})^2 + (z_{1,i} - c_{1,z})^2 + (x_{2,i} - c_{2,x})^2 + (y_{2,i} - c_{2,y})^2 + (z_{2,i} - c_{2,z})^2}{2N}} \tag{2B}$$

$$\Delta v_{xi} = x_{1,i} - x_{2,i} \qquad \Delta v_{y,i} = y_{1,i} - y_{2,i} \qquad \Delta v_{z,i} = z_{1,i} - z_{2,i} \tag{3}$$

$$\sigma_{obs} = \sqrt{\frac{1}{3N-1}\sum_{i=1}^{N}(\Delta v_{x,i} - \langle \Delta v_{x,i}\rangle)^2 + (\Delta v_{y,i} - \langle \Delta v_{y,i}\rangle)^2 + (\Delta v_{z,i} - \langle \Delta v_{z,i}\rangle)^2} \tag{4}$$

$$RMSD_{obs} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(\Delta v_{x,i})^2 + (\Delta v_{y,i})^2 + (\Delta v_{z,i})^2} \tag{5}$$

In Equations (2)–(5), N is the number of residues in the gapless superposition, $(x_{1,i}, y_{1,i}, z_{1,i})$ are the coordinates of the CA atom of residue i in molecule 1, $(x_{2,i}, y_{2,i}, z_{2,i})$ are the coordinates of the CA atom of residue i in molecule 2, and $(c_{1,x}, c_{1,y}, c_{1,z})$, $(c_{2,x}, c_{2,y}, c_{2,z})$ denote the centers of mass of molecules 1 and 2, respectively. Coordinate difference vector projections onto the x-, y-, and z-axes were pooled to form a single dataset. It is noted that the traditional measure of similarity, RMSD, is directly proportional to the standard deviation of DVPs. In Equation (4), the averages (in brackets) over Δv_x, Δv_y, and Δv_z are all zero, so the ratio of Equations (4) and (5) is simply $N/(3N-1)$.

Visual inspection of the radius of gyration distribution of the sfragment data revealed evidence of undersampling or oversampling concentrated at the longest and shortest lengths (Supplementary Fig. S6). Undersampling was due to the sparseness of longer length structures (>450 residues). At shorter lengths (<10 residues), superpositions of individual helices and strands extracted at random from otherwise structurally dissimilar proteins were apparent. In addition, bias in length selection (oversampling of the most probable lengths) dominated each dataset. For these reasons, and to ensure statistical significance of parameter estimates (described below), the number of points (coordinate difference vector projections) in each dataset was reduced and approximately equalized as a function of radius of gyration. This was accomplished by controlled random reduction of the number of pairs with increasing radius of gyration (Supplementary Fig. S6). After equalization, the data set of whole molecule superpositions comprised 22,561 pairs and the set of fragment superpositions comprised 32,004 pairs.

5.4. Estimation of expected standard deviation based on size and shape parameters

The RMSD of a superposition, and thus the standard deviation of the DVP distributions derived from that superposition, depended on the molecular size and shape (Jia and Dewey, 2005; Maiorov and Crippen, 1994; Reva et al., 1998). A function was therefore sought that estimated the standard deviation of DVPs from molecular size and shape of the non-homologous superpositions. A second-degree polynomial was selected for this purpose. The polynomial depended on two shape parameters: R_g, radius of gyration and c, the square-root of the smallest eigenvalue of the inertia matrix of molecular coordinates (Manly, 1986). The value of c described the average size of a particular molecule along its "thinnest" dimension, as simply the standard deviation of coordinates along that dimension. Thus, R_g was related to the size and shape of the molecule and c characterized its shape (deviation from a sphere). For a database search, where the size and shape parameters of only the query molecule were known in advance, the equation was:

$$\sigma_{exp} = \phi_1 R_{g1}^2 + \phi_2 R_{g1} + \phi_3\left(\frac{c_1}{R_{g1}}\right)^2 + \phi_4\left(\frac{c_1}{R_{g1}}\right) + \phi_5 R_{g1}\left(\frac{c_1}{R_{g1}}\right) \tag{6A}$$

In Equation (6A), σ_{exp} refers to expected standard deviation of DVPs for the minimum RMSD superposition (both first and second molecules), R_{g1} refers to radius of gyration for the first molecule, and c_1 is the square root of the smallest eigenvalue of the first molecule's inertia matrix of molecular coordinates. c_1 was normalized by R_{g1} to minimize correlation between the two variables. An analogous equation was used for the case of an isolated pairwise comparison between two molecules, where the sizes and shapes of both were known in advance:

$$\sigma_{exp} = \phi_1 R_{g12}^2 + \phi_2 R_{g12} + \phi_3\left(\frac{c_{12}}{R_{g12}}\right)^2 + \phi_4\left(\frac{c_{12}}{R_{g12}}\right) + \phi_5 R_{g12}\left(\frac{c_{12}}{R_{g12}}\right) \tag{6B}$$

Many other parameters describing shape and size were tried, e.g., superposition length, eigenvalue corresponding to the "thickest" molecular dimension, and values of RMSD to the axes along the "thinnest" and "thickest" dimensions, but R_g and c/R_g were empirically observed to result in minimum χ^2 fits and thus resulted in better estimation of standard deviation from molecular shape. In general, although higher degree polynomials were tried, second degree polynomials of two variables were found to give the best tradeoffs between number of parameters and goodness-of-fit. Coefficients ϕ in Equations (6A) and (6B) were separately determined using singular value decomposition (SVD) (Press et al., 1992) by solving the following general overdetermined matrix equation for x:

$$b = A \cdot x \tag{7}$$

In Equation (7), A denotes the matrix of shape parameters (R_{g1} and c_1 for R_{g12} and c_{12}) for each superimposed pair, x denotes a vector of coefficients ϕ, and b denotes a vector of observed standard deviations (σ_{obs}) of DVPs for each superimposed pair. The SVD procedure (Press et al., 1992) determined the coefficients ϕ in Equations (6A) and (6B) as those that minimized the sum over all superpositions of $(\sigma_{exp} - \sigma_{obs})^2$ (step II of Fig. 1).

5.5. Parameter estimates for modeling distributions of observed standard deviations and coordinate difference vector projections corresponding to "slices" of expected standard deviation

For each non-homologous superposition, an expected standard deviation (σ_{exp}) of coordinate difference vector projections was computed from its molecular size and shape parameters using either Equations (6A) or (6B). Data from all superpositions were then grouped by σ_{exp} into narrow "slices" at intervals of 0.20 Å. This slice width was empirically chosen to ensure a large enough number of observations in each slice while retaining the statistical significance of parameter estimates (described below). Two distributions of observed data were extracted for each slice of σ_{exp}: one consisting of standard deviations (σ_{obs}) of DVPs and one consisting of the projections themselves.

A probability density function (PDF) was selected to approximate each type of distribution, and parameters were independently estimated for each PDF, and thus for the slice, using the general method of χ^2 minimization. Several three- and two-parameter probability density functions belonging to the generalized inverse Gaussian (GIG) (Jorgenson, 1982) family were evaluated for this purpose. It was empirically observed that one parameter of the three-parameter GIG consistently refined to zero (Supplementary Fig. S7), suggesting that a third parameter was unnecessary to adequately describe these data. Moreover, a value of zero for the third parameter reduced the GIG to a special case of Gamma (Johnson et al., 1994). Indeed, both two-parameter Gamma and Nakagami probability density functions were observed to describe variance and standard deviation distributions, respectively, equally well as did the three parameter GIG (Supplementary Fig. S2). The Nakagami PDF (and not Gamma) was used throughout this work because standard deviation was mathematically more directly related to the commonly used measure of RMSD (Equations (4) and (5), above) than was variance.

The PDF for each σ_{obs} distribution was thus based on a two-parameter Nakagami probability density function (Nakagami, 1960):

$$\rho_{Nakagami}(x \mid s, k) = \frac{2}{\Gamma(k)} \left(\frac{k}{s^2}\right)^k x^{2k-1} e^{\frac{k}{s^2} x^2}$$

$$(x > 0, s > 0, k > 1) \tag{8}$$

In Equation (8), x is the independent variable, σ_{obs}, standard deviation of DVPs expressed in Ångstroms (defined in Equation (4)), $\{s, k\}$ are the two parameters to be estimated, and $\Gamma(k)$ refers to the Euler gamma function (Press et al., 1992):

$$\Gamma(k) = \int_0^\infty t^{k-1} e^{-t} dt \tag{9}$$

A Variance-Gamma (Madan, 1990) PDF for each DVP distribution resulted from the convolution of Nakagami PDF (Equation (8)) with a Gaussian:

$$\begin{aligned}\rho_{\mathit{Var\text{-}Gamma}}(y \mid s, k) &= \int_0^\infty \rho_{\mathit{Nakagami}}(x \mid s, k) \frac{1}{x\sqrt{2\pi}} e^{\frac{y^2}{2x^2}} dx \\ &= \frac{2^{\frac{3}{4}-\frac{k}{2}}}{\sqrt{\pi}\Gamma(k)} \left(\frac{k}{s^2}\right)^{\frac{1}{4}+\frac{k}{2}} |y|^{k-\frac{1}{2}} K_{k-\frac{1}{2}} \left(\sqrt{2\frac{k}{s^2}}|y|\right) \\ &\qquad (-\infty \le y \le \infty, s > 0, k > 1) \end{aligned} \tag{10}$$

In Equation (10), $K_{k-1/2}$ is a special case of the modified Bessel function of the second kind, $K_n(z)$, where $n = k - 1/2$ (Abramowitz and Stegun, 1972; Press et al., 1992; Weisstein, 2008):

$$K_n(z) = \frac{\Gamma\left(n + \frac{1}{2}\right)(2z)^n}{\sqrt{\pi}} \int_0^\infty \frac{\cos(t)}{(t^2 + z^2)^{n+\frac{1}{2}}} dt \tag{11}$$

In Equation (10), y denotes $\Delta v_{\{x,y,z\},i}$, coordinate difference vector projection (defined in Equation (3)), in units of Ångstroms. Ideally, the estimated values of parameters $\{s, k\}$ from a standard deviation distribution and its corresponding DVP distribution would be identical in Equations (8) and (10). The closeness of parameter values estimated from the sample of standard deviations on the one hand and from the sample of DVPs of the other hand, was crucial to the overall approach, because this allowed estimation and critical evaluation of values (akin to training and testing) in the linked coordinate spaces of standard deviations and projections (step III of Supplementary Fig. S1).

It was noted that $\{s, k\}$ in this particular parameterization could be expressed as simple combinations of moments of the distributions. For example, given a standard deviation distribution described by a Nakagami probability density function (Equation (8)), s would theoretically correspond to the mean of the distribution, and k would be proportional to the inverse coefficient of variation, squared:

$$s = \frac{1}{N}\sum_{i=1}^{N} \sigma_{obs,i} \tag{12}$$

$$k = \left(\frac{\sigma}{4\mu}\right)^{-2} = 4\left(\frac{\frac{1}{N}\sum_{i=1}^{N}\sigma_{obs,i}}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\sigma_{obs,i} - \frac{1}{N}\sum_{i=1}^{N}\sigma_{obs,i}\right)^2}}\right)^2 \tag{13}$$

In Equations (12) and (13), N is the number of superpositions in the distribution, $\sigma_{obs,i}$ is the standard deviation of DVPs of the ith superposition, defined in Equation (4), σ stands for the standard deviation over all $\sigma_{obs,i}$ in the distribution, and μ stands for the mean of all $\sigma_{obs,i}$ in the distribution.

The simple relations of s and k to the moments of distributions suggested an attempt to reduce the number of free parameters from two to one. Ideally, s would equal the expected standard deviation, σ_{exp}, of the projection distribution, leaving only one variable, the shape parameter k, to be estimated in Equations (8) and (10). Unfortunately, s was empirically found to only be approximately linear for the observed data. The deviations from linearity, although small, resulted in worse agreement of the PDF with the data as evidenced by doubling of the χ^2 value in χ^2 tests (Equation (14), below) when s was fixed at σ_{exp}. Therefore, s was explicitly estimated in addition to k. This slight, but statistically significant, non-linearity was hypothesized to be due to imperfection of the second-degree polynomial approximation of expected standard deviation from molecular shape (Equations (6A) and (6B)).

Using Equations (8) and (10), parameter estimates and tests of these estimates were performed on distributions of the observed standard deviations and DVPs corresponding to each slice of expected standard deviation. Each slice contained approximately 150,000 projections and approximately 300 standard deviations, arising from approximately 300 superimposed pairs of proteins.

Parameter estimates were performed by the general method of χ^2 minimization using custom-written routines in *Mathematica*, version 5.2 (Wolfram Research, Inc., Champaign, IL). The routines binned observed standard deviation data into $N = 20$ bins of approximately equal numbers of points, $n^{obs}_{i=1,\ldots,20}$, with at least 10 points per bin. Over these defined bin boundaries, the χ^2 value (Equation (14)) was minimized:

$$\chi^2 = \sum_{i=1}^{J} \frac{(n_i^{obs} - n_i^{exp})^2}{n_i^{exp}} \tag{14}$$

In Equation (14), $J = 20$ bins, n_i^{obs} was the number of observed data points in standard deviation bin i, and n_i^{exp} was the number of expected data points in standard deviation bin i, calculated as follows from Equation (15):

$$n_i^{exp} = N \int_{lower_i}^{upper_i} \rho_{Nakagami}(x \mid s, k) dx \tag{15}$$

In Equation (15), N was the total number of points in the slice, $lower_i$ was the lower boundary of bin i, $upper_i$ was the upper bound for bin i, and $\rho_{Nakagami}(x \mid s, k)$ was defined in Equation (8), above. The estimates of s and k were found as values that minimized the χ^2 value given by Equation (14).

The values of parameters estimated from each standard deviation distribution were tested by computing the χ^2 value between the corresponding observed projection distribution and the PDF of the expected projection distribution, evaluated using said parameter values. These tests served as semi-independent checks on the validity of the best estimate parameters, since the parameters were obtained from one distribution and evaluated against another. These tests were performed using similar custom-written *Mathematica* routines. The routines binned difference vector projection data into 50 bins of approximately equal numbers of points, with at least 10 points per bin. Over these defined bin boundaries, the χ^2 value (Equation (14)) was evaluated using predetermined parameters s and k taken from the corresponding best estimates of that data slice in standard deviation space. For tests using Equation (14), $J = 50$ bins, n_i^{obs} was the number of observed data points in projection bin i, and n_i^{exp} was the number of expected data points in projection bin i, calculated as follows from Equation (16):

$$n_i^{exp} = N \int_{lower_i}^{upper_i} \rho_{Var\text{-}Gamma}(y \mid s, k) dy \tag{16}$$

In Equation (16), N was the total number of points in the slice, $lower_i$ was the lower boundary of bin i, $upper_i$ was the upper bound for bin i, and $\rho_{Var\text{-}Gamma}(y \mid s, k)$ was defined in Equation (10), above.

Larger values of χ^2 indicated that a particular PDF estimated from a standard deviation distribution and tested on the corresponding projection distribution was less statistically probable. For the purposes of this work, a χ^2 value of between one and two times the 47 degrees of freedom, *d.o.f.*, resulted in a visually acceptable approximation to these data, although some of such approximations were technically of marginal statistical significance (Press et al., 1992). All tests of estimated parameters were generally within this range of χ^2 values (Fig. 6), demonstrating that estimated parameters s and k were reasonable. For an additional assessment of statistical significance, p-values were also computed for these tests:

$$p = 1 - P(\chi^2 \mid \nu) \tag{17}$$

In Equation (17) $P(\chi^2 \mid \nu)$ was the cumulative distribution function of the χ^2 distribution evaluated for $\nu = 47$ degrees of freedom (Press et al., 1992). P-values ranged from zero to one, with larger values indicating a more statistically probable model.

To estimate the lower bound of χ^2 values obtainable for the difference vector projection distributions, χ^2 values were calculated using Equation (14) from the positive half of these projection data, binned into

$i = 1, \ldots, 50$ approximately equal-sized bins n_i^{obs}, against the absolute values of the negative half of these projection data, binned into $i = 1, \ldots, 50$ bins n_i^{exp} using the bin boundaries determined from the positive half. Estimates were computed for each slice of expected standard deviation.

5.6. *Correction for short, nearly-exact superpositions using homologous fragments*

For the shortest fragment lengths ($L < \sim 15$, $\sigma_{exp} < \sim 3.3$), it was found that a single Nakagami PDF did not fit the observed distributions well. A second Nakagami PDF was parameterized to account for the excess density contained in these shortest fragment distributions. For each superposition length L from $4 \leq L \leq 50$ residues, 1000 protein pairs were randomly selected from the ASTRAL_90 compendium (version 1.69) (Chandonia et al., 2004). Each member of the pair was within 10% in length and contained >40% identity to the other member of the pair over a BLAST (Altschul et al., 1997) alignment. A gapless fragment of length L, equivalenced according to a Dali superposition of the complete structures, was excised and its minimum RMSD superposition was calculated. Generation of these data from homologous fragments proceeded identically to that described above for non-homologous pairs; that is, DVPs for the superposition of the pair, the standard deviation of DVPs for the superposition of each pair, and the pair RMSD were calculated.

Standard deviation and projection data were partitioned using the identical σ_{exp} function as parameterized for random data, described above in Equations (6A) and (6B). Minimum χ^2 parameter estimates of $\{s_h, k_h\}$ from these distributions were attempted, but the distributions exhibited long tails that made statistically reasonable automatic fits impossible (data not shown). Therefore, to focus only on the sharp peak of nearly exact superpositions, $\{s_h, k_h\}$ were estimated using Equations (12) and (13), above, from the observed mean and standard deviation of homolog distributions truncated at $\sigma_{obs}^2 > 2.0$ Å.

The $\{s_h, k_h\}$ estimated from homolog data were used to construct probability density functions made of two-component mixtures. One component of the mixture contained data from homologous superpositions, and the other component contained data from random superpositions. These mixtures were used to correct parameter estimates for the fraction of nearly exact matches contained in the shortest chain distributions. Because the homolog σ_{obs} peak of the mixture was very narrow relative to the random peak, and broadened somewhat as chain length increased, equal binning by numbers of data points did not result in accurate parameter estimates for the mixtures (data not shown). Therefore, a modified procedure was developed in which the homolog component of the observed data, defined as standard deviations less than the 90% cutoff of values in the pure homolog distribution, was binned with narrower bin widths than the random component of the observed data. Bin widths for the homolog component were adjusted such that each bin contained an approximately equal number of at least 10 points. Again, to focus only on the sharp peak of nearly exact superpositions, only the first half of these bins were considered in the fitting routine. The random component of the observed data was divided into bins containing approximately equal numbers of points such that the sum of the number of bins over the homolog and random components of the observed data totaled 20. These corrected parameter estimates were performed only over slices of $\sigma_{exp} < 3.3$, because zero homolog component (defined as substantial density at $\sigma_{obs}^2 < 2.0$ Å), was observed above this cutoff.

Given an $\{s_h, k_h\}$ determined from the pure homolog distribution corresponding to each slice, the automatic procedure just described was used to estimate parameters for the following mixture PDF to obtain a corrected $\{s, k\}$ for the random component, as well as the fraction of short, nearly-exact matches present, w_h:

$$\rho_{rand}(x \mid s, k, s_h, k_h, w_h) = (1 - w_h)\rho_{Nakagami}(x \mid s, k) + (w_h)\rho_{Nakagami}(x \mid s_h, k_h) \qquad (18)$$

5.7. *Modeling parameters for arbitrary values of σ_{exp}*

Continuous curves $\{s_{rand}, k_{rand}\}$ were fit to the previously estimated modeling parameters $\{s, k\}$. A combination of logistic and linear functions, generally represented by Equation (19), was chosen for this purpose. In Equation (19), parameters α, β, δ are part of the logistic term (α is the position of the infection point while β and δ characterize the width and height of the sigmoid, respectively), γ is the slope of the linear term, and ε is an intercept. This single equation could well capture the sigmoid shape of k, yet also describe the small non-linearity of s. The intercepts in Equations (20) and (21) were derived to coincide with the expected lower bounds of s_{rand} and k_{rand}: $1 \leq k_{rand} \leq \infty$, $0 \leq s_{rand} \leq \infty$.

$$y = \frac{\delta}{1 + e^{\frac{x-\alpha}{\beta}}} + \gamma x + \varepsilon \tag{19}$$

$$s_{rand} = \gamma_1 \sigma_{exp} + \frac{\delta_1}{1 + e^{\frac{\alpha_1 - \sigma_{exp}}{\beta_1}}} - \frac{\delta_1}{1 + e^{\frac{\alpha_1}{\beta_1}}} \tag{20}$$

$$k_{rand} = \gamma_2 \sigma_{exp} + \frac{\delta_2}{1 + e^{\frac{\alpha_2 - \sigma_{exp}}{\beta_2}}} - \left(\frac{\delta_2}{1 + e^{\frac{\alpha_2}{\beta_2}}} - 1 \right) \tag{21}$$

The continuous curves $\{s_{hom}, k_{hom}, w_{hom}\}$ were fit to the previously estimated parameters $\{s_h, k_h\}$. These parameters appeared to saturate as σ_{exp} increased, so the simple function chosen for both parameters was a single exponential rising to a maximum. The intercepts in Equations (22) and (23) were derived to coincide with the expected lower bounds of s_{hom} and k_{hom}: $k_{hom} \geq 1$, $s_{hom} \geq 0$:

$$s_{hom} = \kappa_1 (1 - e^{\lambda_1 \sigma_{exp}}) \tag{22}$$

$$k_{hom} = \kappa_2 (1 - e^{\lambda_2 \sigma_{exp}}) + 1 \tag{23}$$

Conversely, the function selected for w_{hom} exhibited a very rapid decay, reflecting the observed absence of nearly exact matches at average fragment superposition lengths greater than 15 residues:

$$w_{hom} = 1 - e^{-(\lambda_3 \sigma_{exp} + \lambda_3 \sigma_{exp}^3)} \tag{24}$$

Continuous parameters for the random component of the probability density were combined with continuous parameters for the homolog component of the probability density, resulting in a mixture of two Nakagami distributions to describe the complete probability density:

$$\begin{aligned} &\rho_{rand}(x \mid s_{rand}, k_{rand}, s_{hom}, k_{hom}, w_{hom}) \\ &\quad = (1 - w_{hom}) \rho_{Nakagami}(x \mid s_{rand}, k_{rand}) + (w_{hom}) \rho_{Nakagami}(x \mid s_{hom}, k_{hom}) \end{aligned} \tag{25}$$

Tests of these continuous parameters for each slice of expected standard deviation were performed using custom-written *Mathematica* routines. The routines binned difference vector projection data into 50 bins of approximately equal numbers of points, with at least 10 points per bin. Over these defined bin boundaries, the χ^2 value (Equation (14)) was computed using predetermined continuous parameters $\{s_{rand}, k_{rand}, s_{hom}, k_{hom}, w_{hom}\}$ evaluated at the corresponding average value of expected standard deviation, σ_{exp}. In Equation (14), $J = 50$ bins, n_i^{obs} was the number of observed data points in projection bin i, and n_i^{exp} was the number of expected data points in projection bin i, calculated as follows from Equation (26):

$$n_i^{exp} = N \int_{lower_i}^{upper_i} \rho_{rand}(x \mid s_{rand}, k_{rand}, s_{hom}, k_{hom}, w_{hom}) dx \tag{26}$$

In Equation (26), N was the total number of points in the slice, $lower_i$ was the lower boundary of bin i, $upper_i$ was the upper bound for bin i, and $\rho_{rand}(x \mid s_{rand}, k_{rand}, s_{hom}, k_{hom}, w_{hom})$ was defined in Equation (25), above.

5.8. Estimation of gapless superposition probabilities from continuous modeling parameters

Given dependencies of parameters s and k on σ_{exp} described by Equations (20)–(24), an expected random distribution of standard deviations (i.e., RMSD) for a specific molecular shape could be constructed using Equation (25). Thus, the probability of observing an equal or better minimum RMSD superposition due to chance could be simply calculated by:

$$p_{\sigma<\sigma_{obs}} = \int_{x=0}^{\sigma_{obs}} \rho_{rand}(x \mid s_{rand}, k_{rand}, s_{hom}, k_{hom}, w_{hom})dx$$

$$= (1 - w_{hom})\left(1 - Q\left[k_{rand}, \frac{k_{rand}\sigma_{obs}^2}{s_{rand}^2}\right]\right) + (w_{hom})\left(1 - Q\left[k_{hom}, \frac{k_{hom}\sigma_{obs}^2}{s_{hom}^2}\right]\right) \quad (27)$$

where σ_{obs} was the standard deviation of DVPs of the superposition of interest, and $Q[a, b]$ was the regularized incomplete gamma function (Press et al., 1992):

$$Q[a, b] = \frac{\Gamma(a, b)}{\Gamma(a)} \quad (28)$$

where the denominator, the Euler gamma function, was defined in Equation (9), above, and the numerator was the incomplete gamma function (Press et al., 1992):

$$\Gamma(a, b) = \int_{b}^{\infty} t^{a-1}e^{-1}dt \quad (29)$$

In Equation (27), for the cases where whole molecules, not fragments, were involved in the superposition, w_{hom} was set to zero and the term relating to the short chain correction was ignored and thus did not contribute to the random model.

5.9. Comparison of calculated probabilities to corresponding Dali Z-scores of structural superpositions

Optimal structural superpositions generated by Dali (Holm and Park, 2000) were sampled and their corresponding Z-scores, a widely used estimator of significance, were compared to the probabilities calculated for the same superpositions using Equation (27). For these comparisons, only the CA coordinates equivalenced by Dali were used to compute the minimum RMSD superpositions and the prediction of molecular size and shape in Equation (6B). The Dali superpositions involved pairs of proteins chosen at random from the 6551 member structure set described above. Approximately 1000 pairs from each level of the SCOP (Andreeva et al., 2004) hierarchy were selected: pairs belonging to the same family, pairs belonging to the same superfamily but different families, pairs belonging to the same fold but different superfamilies, pairs belonging to the same class but different folds, and pairs belonging to different classes. Dali superpositions involving at least 90% of both chains, whose total lengths were within 10% of each other, were treated as whole molecule superpositions. Thus, parameters $\{s_{rand}, k_{rand}\}$ derived from whole molecule superpositions were used to compute probabilities, and the short-chain correction was ignored. Otherwise, the superposition was treated as fragments and the short-chain correction was included. In all cases, the size and shape parameters of both molecules were considered in the calculations.

5.10. Comparison of calculated probabilities of structure-linked sequence alignments to corresponding local GDT_TS Z-scores

PSI-BLAST (Altschul et al., 1997) profiles (E-value cutoff 0.01, 5 iterations) were generated, as query using each of the sequences contained in the 5000 SCOP pairs selected for the Dali superpositions. Each of the 5000 profile-profile alignments was computed using COMPASS, version 2.4 (Sadreyev and Grishin, 2003). CA equivalences were extracted from the structures corresponding to the COMPASS alignments, and these equivalences were used to compute minimum RMSD superpositions, shape parameters, σ_{exp} (Equation (6B)) and probability estimates, Equation (27). As above, superpositions involving at least 90% of both chains, whose total lengths were within 10% of each other, were treated as whole molecule superpositions. Thus, parameters $\{s_{rand}, k_{rand}\}$ derived from whole molecule superpositions were used to compute probabilities, and the short-chain correction was ignored. Otherwise, the superposition was treated as fragments and the short-chain correction was included. In all cases the size and shape parameters of both molecules were considered in the calculations.

These probabilities were compared to local GDT_TS scores calculated for the same superpositions. To remove systematic effects on the local GDT_TS score due to differences in fragment lengths, each score was converted to a Z-score using mean and standard deviation values observed from minimal RMSD superpositions of non-homologous proteins of identical length (Qi et al., 2007). Larger values of the local GDT_TS Z-score indicated a closer structural match between the sequence-aligned structure fragments.

5.11. *Prediction of random scores for other scoring systems*

1000 non-homologous minimal RMSD gapless fragment superpositions of randomly chosen lengths $L = 4$ to $L = 600$ were generated. (This set was independent of the 32,004 superpositions used to parameterize the model.) For each superposition, an expected random distribution of DVPs based on its predicted size and shape was constructed using Equations (6A) and (30), the latter being a mixture of PDFs, analogous to Equation (25), that included homolog, as well as random, components:

$$\rho_{rand}(x \mid s_{rand}, k_{rand}, s_{hom}, k_{hom}, w_{hom}) \\ = (1 - w_{hom})\rho_{\textit{Var-Gamma}}(x \mid s_{rand}, k_{rand}) + (w_{hom})\rho_{\textit{Var-Gamma}}(x \mid s_{hom}, k_{hom}) \tag{30}$$

Difference vector projections p_i were randomly drawn from the distribution until a total of $3 * L$ values were selected. Then, a predicted score for one of four different scoring measures—RMSD, TM (Zhang and Skolnick, 2004), GDT_TS (Zemla et al., 1999), or 3D score (Rychlewski et al., 2003)—was calculated according to the following equations:

$$RMSD_{pred} = \sqrt{\frac{1}{L}\sum_{i=1}^{3L} p_i^2} \tag{31}$$

$$TM\_Score_{pred} = \frac{1}{L}\sum_{i=1}^{3L} \frac{1}{1 + \left(\frac{\sqrt{p_i^2 + p_{i+1}^2 + p_{i+2}^2}}{d_0}\right)^2} \tag{32}$$

$$GDT\_TS - Score_{pred} = \frac{(N_1 + N_2 + N_4 + N_8)}{4L} \tag{33}$$

$$3D - Score_{pred} = \sum_{i=1}^{3L} e^{(-\ln 2)\left(\frac{\sqrt{p_i^2+p_{i+1}^2+p_{i+2}^2}}{d_0}\right)^2} \tag{34}$$

In Equations (31)–(34), i is an index over all coordinate difference vector projections, and p_i is the randomly drawn value of the ith projection.

In Equations (32) and (34), d_0 was defined as follows:

$$d_0 = (1.24 - (L - 15)^{\frac{1}{3}}) - 1.8 \qquad L \geq 15 \tag{35}$$

$$d_0 = 0.5 \qquad L < 15 \tag{36}$$

Although the original implementation of 3D score used a constant d_0 of 3 Å (Rychlewski et al., 2003), it was empirically found that agreement of predicted with observed 3D scores was substantially increased when a length-dependent d_0 was used, thus that change was employed here.

In Equation (33), N_1, N_2, N_4, and N_8 were the number of distances d between equivalenced atoms less than 1, 2, 4, and 8 Å, defined as i ran from 1 to $3 * L$:

$$d = \sqrt{p_i^2 + p_{i+1}^2 + p_{i+2}^2} \tag{37}$$

5.12. Calculation of probabilities from two existing methods

Probability p of chance occurrence of a superposition of RMSD r was computed according to the method of Reva et al. (1998):

$$p = \int_{-\infty}^{r} \frac{e^{-\frac{(x-\langle R\rangle)^2}{2sd^2}}}{sd\sqrt{2\pi}} dx \tag{38}$$

In Equation (38), $sd = 2.0$ Å and $\langle R\rangle = 3.333N^{1/3}$, where N was the length of the superposition in residues.

Probability p of chance occurrence of a superposition of RMSD r was computed according to the method of Jia and Dewey (2005):

$$p = exp\left(-exp\left(-\frac{\frac{r}{M^{0.32}} - 3.37}{0.48}\right)\right) \tag{39}$$

In Equation (39), M was the length of the superposition in residues. Superpositions were taken from the same 5000-pair set used above in the Dali Z-score and GDT_TS score comparisons. However, superpositions of length less than 50 residues were not considered, as the methods of Reva et al. (1998) and Jia and Dewey (2005) were not parameterized on lengths less than approximately 50 residues.

5.13. Calculation of Pearson correlation coefficient

The Pearson correlation coefficient r between two lists of values was calculated as (Press et al., 1992):

$$r = \frac{\sum_{i=1}^{N}(x_i - \langle x\rangle)(y_i - \langle y\rangle)}{\sqrt{\sum_{i=1}^{N}(x_i - \langle x\rangle)^2}\sqrt{\sum_{i=1}^{N}(y_i - \langle y\rangle)^2}} \tag{40}$$

In Equation (40), x_i, y_i were the ith values in the first and second lists, respectively, and $\langle x\rangle$, $\langle y\rangle$ were the averages over the number N values in each list. An r value of 1 indicated a perfect positive correlation between the two lists; a value of -1 a perfect anticorrelation.

ACKNOWLEDGMENTS

We thank Ruslan Sadreyev for critical reading of the manuscript and for helpful suggestions. This work was supported in part by NIH grant GM-67165 to N.V.G.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Abramowitz, M., and Stegun, I. 1972. *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*. Dover, New York.

Altschul, S.F., Madden, T.L., Schaffer, A.A., et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Andreeva, A., Howorth, D., Brenner, S.E., et al. 2004. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* 32, D226–D229.

Barndorff-Nielsen, O. 1977. Exponentially decreasing distributions for the logarithm of particle size. *Proc. R. Soc. Lond. A Math. Phys. Sci.* 353, 401–419.

Barndorff-Nielsen, O.E., and Shephard, N. 2001. Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics. *J. R. Statist. Soc. B Statist. Method.* 63, 167–241.

Chandonia, J.M., Hon, G., Walker, N.S., et al. 2004. The ASTRAL Compendium in 2004. *Nucleic Acids Res.* 32, D189–D192.

Coutsias, E.A., Seok, C., and Dill, K.A. 2004. Using quaternions to calculate RMSD. *J. Comput. Chem.* 25, 1849–1857.

Eberlein, E., and Keller, U. 1995. Hyperbolic distributions in finance. *Bernoulli* 1, 281–299.

Fitzkee, N.C., Fleming, P.J., Gong, H., et al. 2005. Are proteins made from a limited parts list? *Trends Biochem. Sci.* 30, 73–80.

Flower, D.R. 1999. Rotational superposition: a review of methods. *J. Mol. Graph Model.* 17, 238–244.

Holm, L., and Park, J. 2000. DaliLite workbench for protein structure comparison. *Bioinformatics* 16, 566–567.

Holm, L., and Sander, C. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233, 123–138.

Holm, L., and Sander, C. 1994. The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.* 22, 3600–3609.

Jia, Y., and Dewey, T.G. 2005. A random polymer model of the statistical significance of structure alignment. *J. Comput. Biol.* 12, 298–313.

Jia, Y., Dewey, T.G., Shindyalov, I.N., et al. 2004. A new scoring function and associated statistical significance for structure alignment by CE. *J. Comput. Biol.* 11, 787–799.

Johnson, N.L., Kotz, S., and Balakrishnan, N. 1994. *Continuous Univariate Distributions, Volume 1*. John Wiley & Sons, New York.

Jorgenson, B. 1982. *Statistical Properties of Generalized Inverse Gaussian Distribution*. Springer, New York.

Kabsch, W. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Cryst. A* 32, 922–923.

Kabsch, W. 1978. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crys. A* 34, 827–828.

Levitt, M., and Gerstein, M. 1998. A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci. USA* 95, 5913–5920.

Madan, D.B., and Seneta E. 1990. The variance gamma (V.G.) model for share market returns. *J. Business* 63, S11–S24.

Maiorov, V.N., and Crippen, G.M. 1994. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *J. Mol. Biol.* 235, 625–634.

Maiorov, V.N., and Crippen, G.M. 1995. Size-independent comparison of protein three-dimensional structures. *Proteins* 22, 273–283.

Manly, B.F.J. 1986. *Multivariate Statistical Methods: A Primer*. Chapman and Hall, London.

Nakagami, M. 1960. *The M-Distribution, A General Formula of Intensity of Rapid Fading*. Pergamon Press, New York.

Ortiz, A.R., Strauss, C.E., and Olmea, O. 2002. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.* 11, 2606–2621.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., et al. 1992. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, New York.

Qi, Y., Sadreyev, R.I., Wang, Y., et al. 2007. A comprehensive system for evaluation of remote sequence similarity detection. *BMC Bioinform.* 8, 314.

Reva, B.A., Finkelstein, A.V., and Skolnick, J. 1998. What is the probability of a chance prediction of a protein structure with an rmsd of 6 A? *Fold. Des.* 3, 141–147.

Rychlewski, L., Fischer, D., and Elofsson, A. 2003. LiveBench-6: large-scale automated evaluation of protein structure prediction servers. *Proteins* 53, Suppl 6, 542–547.

Sadreyev, R., and Grishin, N. 2003. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.* 326, 317–336.

Shah, P.K., Aloy, P., Bork, P., et al. 2005. Structural similarity to bridge sequence space: finding new families on the bridges. *Protein Sci.* 14, 1305–1314.

Shindyalov, I.N., and Bourne, P.E. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* 11, 739–747.

Sichel, H.S. 1975. On a distribution law for word frequencies. *J. Am. Statist. Assoc.* 70, 542–547.

Taylor, W.R. 2006. Decoy models for protein structure comparison score normalisation. *J. Mol. Biol.* 357, 676–699.

Theobald, D.L. 2005. Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. *Acta Cryst. A* 61, 478–480.

Tweedie, M. 1957. Statistical properties of inverse Gaussian distributions I. *Ann. Math. Statist.* 28, 362–377.

Vincent, J.J., Tai, C.H., Sathyanarayana, B.K., et al. 2005. Assessment of CASP6 predictions for new and nearly new fold targets. *Proteins* 61, Suppl 7, 67–83.

Weisstein, E.W. 2008 Modified Bessel function of the second kind [from MathWorld]. Available at: *http://mathworld.wolfram.com/ModifiedBesselFunctionoftheSecondKind.html*. Accessed February 1, 2008.

Whisstock, J.C., and Lesk, A.M. 2003. Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.* 36, 307–340.

Zemla, A., Venclovas, C., Moult, J., et al. 1999. Processing and analysis of CASP3 protein structure predictions. *Proteins* S3, 22–29.

Zhang, Y., and Skolnick, J. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins* 57, 702–710.

Zhang, Y., and Skolnick, J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* 33, 2302–2309.

Zhu, J., and Weng, Z. 2005. FAST: a novel protein structure alignment algorithm. *Proteins* 58, 618–627.

Address reprint requests to:
Dr. Nick V. Grishin
HHMI/Biochemistry
Univ. Texas Southwestern Medical Center
6001 Forest Park Blvd., Rm. ND10.124A
Dallas, TX 75390-8816

E-mail: grishin@chop.swmed.edu

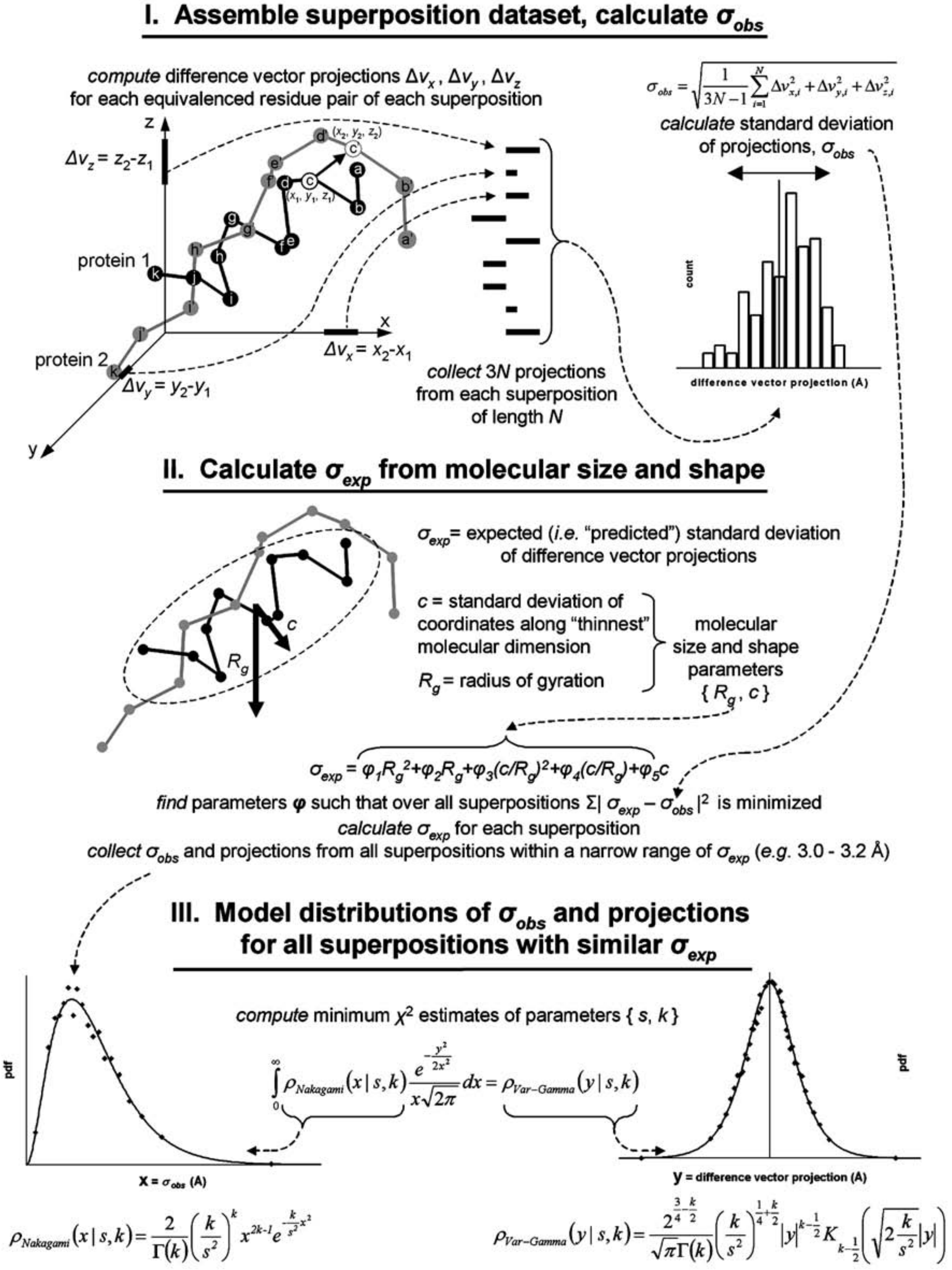

I. Assemble superposition dataset, calculate σ_{obs}
compute difference vector projections Δv_x, Δv_y, Δv_z for each equivalenced residue pair of each superposition
z
$\Delta v_z = z_2 - z_1$
(x_2, y_2, z_2)
(x_1, y_1, z_1)
a b c d e f g h i j k
protein 1
protein 2
$\Delta v_y = y_2 - y_1$
y
x
$\Delta v_x = x_2 - x_1$
collect 3*N* projections from each superposition of length *N*
$\sigma_{obs} = \sqrt{\frac{1}{3N-1}\sum_{i=1}^{N} \Delta v_{x,i}^2 + \Delta v_{y,i}^2 + \Delta v_{z,i}^2}$
calculate standard deviation of projections, σ_{obs}
count
difference vector projection (Å)
II. Calculate σ_{exp} from molecular size and shape
c
R_g
σ_{exp} = expected (*i.e.* "predicted") standard deviation of difference vector projections
c = standard deviation of coordinates along "thinnest" molecular dimension
R_g = radius of gyration
molecular size and shape parameters $\{R_g, c\}$
$\sigma_{exp} = \underbrace{\varphi_1 R_g^2 + \varphi_2 R_g + \varphi_3 (c/R_g)^2 + \varphi_4 (c/R_g) + \varphi_5 c}$
find parameters φ such that over all superpositions $\Sigma | \sigma_{exp} - \sigma_{obs} |^2$ is minimized
calculate σ_{exp} for each superposition
collect σ_{obs} and projections from all superpositions within a narrow range of σ_{exp} (*e.g.* 3.0 - 3.2 Å)
III. Model distributions of σ_{obs} and projections for all superpositions with similar σ_{exp}
compute minimum χ^2 estimates of parameters $\{s, k\}$
$\int_0^\infty \underbrace{\rho_{Nakagami}(x \mid s,k)} \frac{e^{-\frac{y^2}{2x^2}}}{x\sqrt{2\pi}} dx = \underbrace{\rho_{Var\text{-}Gamma}(y \mid s,k)}$
pdf
$\mathbf{X} = \sigma_{obs}$ (Å)
pdf
$\mathbf{y}$ = difference vector projection (Å)
$\rho_{Nakagami}(x \mid s,k) = \frac{2}{\Gamma(k)}\left(\frac{k}{s^2}\right)^k x^{2k-1} e^{-\frac{k}{s^2}x^2}$
$\rho_{Var\text{-}Gamma}(y \mid s,k) = \frac{2^{\frac{3}{4}-\frac{k}{2}}}{\sqrt{\pi}\Gamma(k)}\left(\frac{k}{s^2}\right)^{\frac{1}{4}+\frac{k}{2}} |y|^{k-\frac{1}{2}} K_{k-\frac{1}{2}}\left(\sqrt{2\frac{k}{s^2}}|y|\right)$

FIG. S1. Detailed development of the random model of protein structure superposition. This is a more informative version of main text Figure 1 detailing the source of coordinate difference vector projections (Step I), displaying the expression for σ_{exp} (Step II), and explicitly showing the mathematical relationship (Step III) between the probability density functions used to describe distributions of σ_{obs} (Nakagami (Nakagami, 1960) PDF, main text Equation (8)) and distributions of coordinate difference vector projections (Var-Gamma (Madan, 1990) PDF, main text Equation (10)).

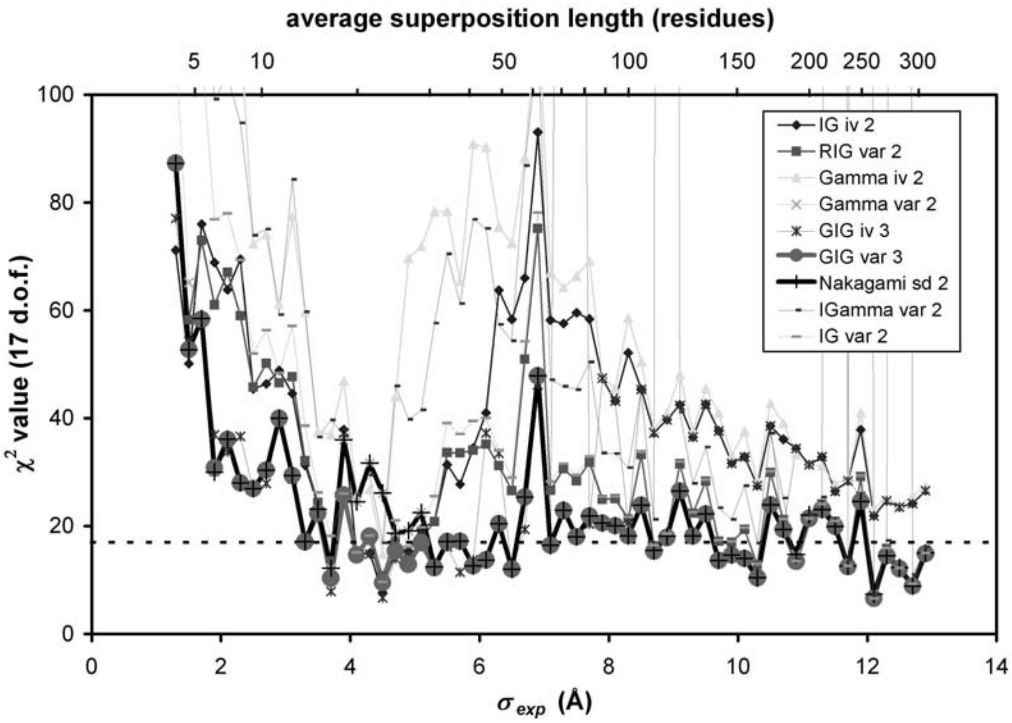


FIG. S2. Assessment of statistical validity of parameter estimates from several probability density functions. Parameter estimates were performed by minimum χ^2 method against nine probability density functions, listed in the figure. Distributions of σ_{obs} were extracted for "slices" of fragment superposition data, sorted by increasing value of σ_{exp} (main text Equation (6A)), as described in Methods. Goodness of fit of a particular probability density function was assessed using the χ^2 value, main text Equation (14). Smaller values of χ^2 indicate the particular PDF is a better approximation of the observed distribution. A dotted line indicates a χ^2 value of 17, reflecting an expected value for the 17 degrees of freedom. Each of nine PDFs in the figure legend is referred to in a shorthand notation: the first term refers to the functional form of the PDF, the second term refers to the independent variable of the PDF (i.e. standard deviation = σ_{obs}, variance = σ^2_{obs}, or inverse variance = σ^{-2}_{obs}), and the third term refers to the number of free parameters in the PDF (two or three). For example, "Nakagami sd 2", the two-parameter Nakagami PDF (main text Equation (8)), with an independent variable of standard deviation, resulted in statistically valid approximations to most distributions analyzed. In particular, the Nakagami χ^2 values were essentially as good as those of a more complex three-parameter PDF, the Generalized Inverse Gaussian (GIG, Equation (S2), below). Abbreviations for the other PDF's are: IG, Inverse Gaussian; RIG, Reciprocal Inverse Gaussian; IGamma, Inverse Gamma.

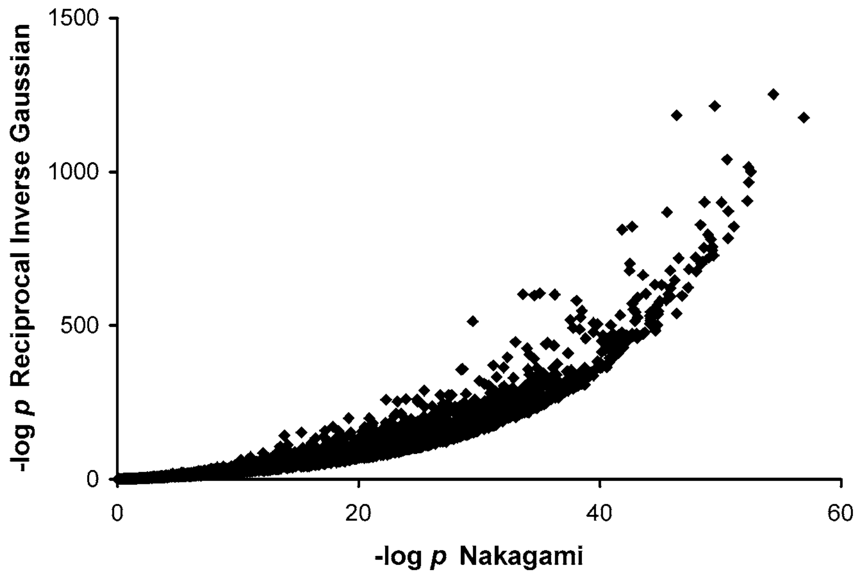


FIG. S3. Comparison of p-values resulting from two different probability density functions parameterized from identical protein data. Approximately 5000 randomly selected pairs of structures were selected from proteins classified in the SCOP database, as described in Methods and the legend to main text Figure 7. For each pair, an optimal pairwise structural alignment was obtained from Dali and its probability of chance occurrence was calculated with main text Equation (27), parameterized for either whole molecule or fragment comparisons depending on the nature of the superposition, as detailed in Methods. Main text Equation (27) was developed with either Nakagami PDFs as described in the text (Equation (8)), or with Reciprocal Inverse Gaussian PDFs (RIG, Equation (S1), development not shown) (Johnson et al., 1994):

$$\rho_{RIG}(x \mid a, b) = \frac{b}{\sqrt{2\pi x}} e^{-\frac{b^2 x}{2} + a - \frac{a^2}{2b^2 x}} \qquad (x > 0,\, a > -1,\, b > 0) \tag{S1}$$

In Equation (S2), x is variance $= \sigma^2_{obs}$ and $\{a, b\}$ are free parameters estimated from the observed variance distributions in a manner similar to the $\{s, k\}$ parameters of the Nakagami distribution previously described. P-values were transformed by $-$log as indicated so that more significant probabilities had larger values. These data demonstrate that different two-parameter PDFs can result in very different absolute significances: p-values based on the RIG PDF are vastly more significant than those based on the Nakagami PDF for the same superposition. This result is due to the relatively lighter tail of the RIG PDF, and the effect is emphasized as the significance increases.

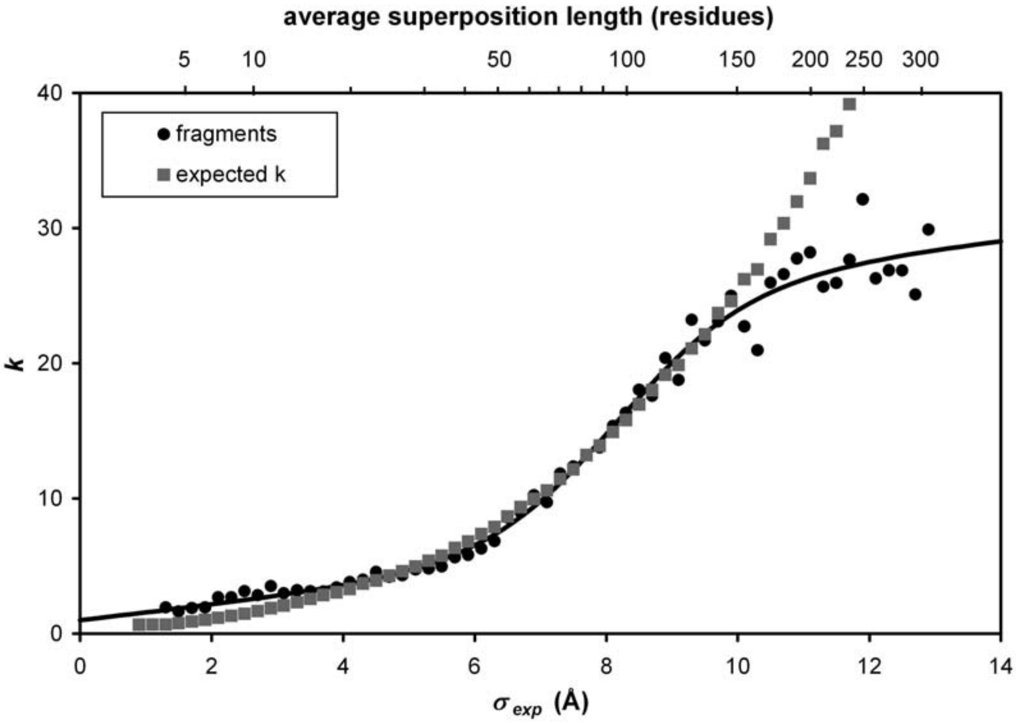


FIG. S4. Comparison of k parameter expected from a simpler, Gaussian, model to estimated k parameter from Nakagami PDF. Dark circles with continuous curve indicate k parameter estimated from observed protein fragment data, repeated from main text Figure 5b. Gray squares indicate an expected from a Gaussian k given by $k = N/6$, where N is the observed average length of the superposition in residues. (Details of the origin of the expected k are given in the Discussion, main text.) The expected k from the simple model deviates from the estimated k from the Nakagami distribution at average superposition lengths greater than ~150 residues.

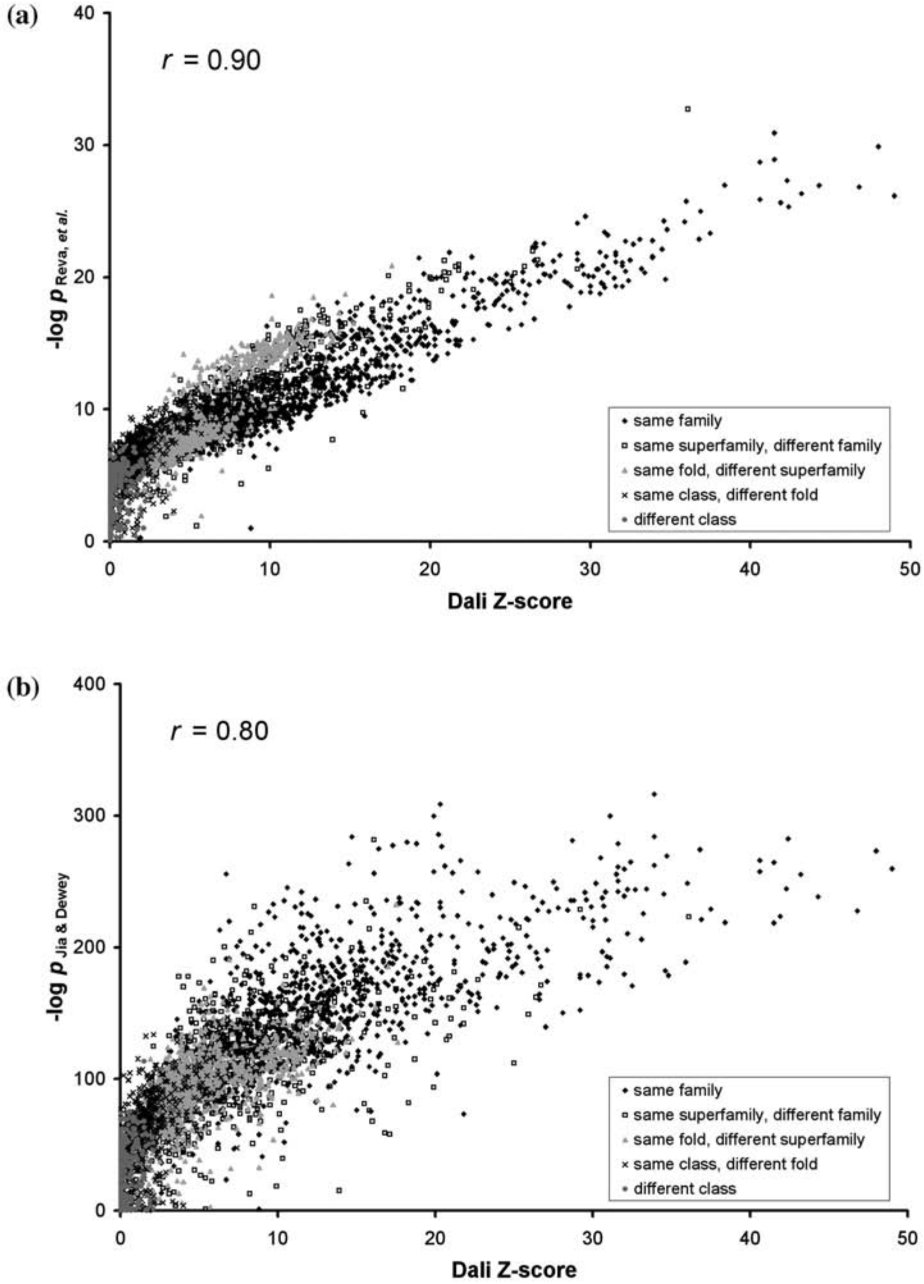


FIG. S5. Dali Z-scores correlate with superposition p-values computed from two previously published approaches. Approximately 1000 randomly selected pairs of structures with known relationship in the SCOP (Andreeva et al., 2004) hierarchy were selected for each of five possible relationship classifications. An optimal pairwise structural alignment was obtained from Dali (Holm and Park, 2000) and the chance probability of observing a superposition of identical RMSD was computed by one of two existing approaches, as indicated. P-values were transformed by −log so that more significant probabilities have higher values. Probabilities of existing approaches were calculated according to the published procedures, as described in Methods. Superpositions of length less than 50 residues were not considered, as the existing approaches were developed based on analysis of lengths greater than approximately 50 residues. (a) Existing approach based on a random model of structurally compact decoys (Reva et al., 1998). A strong correlation of Pearson correlation coefficient $r = 0.90$ was observed. (b) Existing approach based on a random model of structurally dissimilar fragment superpositions (Jia and Dewey, 2005). A strong correlation of Pearson correlation coefficient $r = 0.80$ was observed.

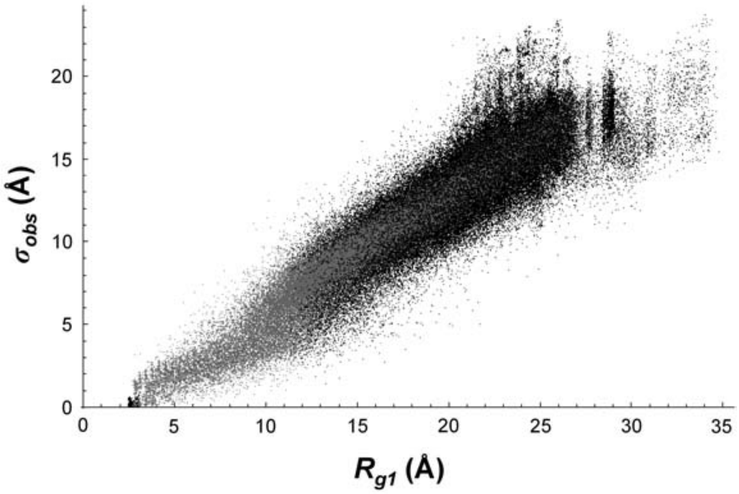


FIG. S6. Scatterplot of σ_{obs} vs. radius of gyration for fragment superpositions. Each point represents one random fragment superposition; 220,000 superpositions from the complete data set are shown in black and 32,004 superpositions from the processed set actually analyzed in this work are shown in gray. Gray points are a randomly chosen subset of the black points selected such that every 0.1 Å wide bin of R_{g1} contains a number of superpositions equal to approximately 60,000 difference vector projections. Thus, as R_{g1} increased, proportionally fewer superpositions were chosen for analysis. (Contours in main text Figure 3 were based on the density of gray points.) Radius of gyration, R_{g1}, was calculated from main text Equation (2A) using the equivalenced CA atoms of the first molecule of each superposition, σ_{obs} was calculated from main text Equation (4) using the coordinate difference vector projections of the equivalenced CA atoms of the complete superposition. Sparseness of data is apparent at longer chain lengths (i.e., $R_{g1} > 25$ Å) as pronounced vertical clusters of points. Superpositions of individual helices and strands result in several smaller but regularly spaced vertical clusters at $R_{g1} < 5$ Å.

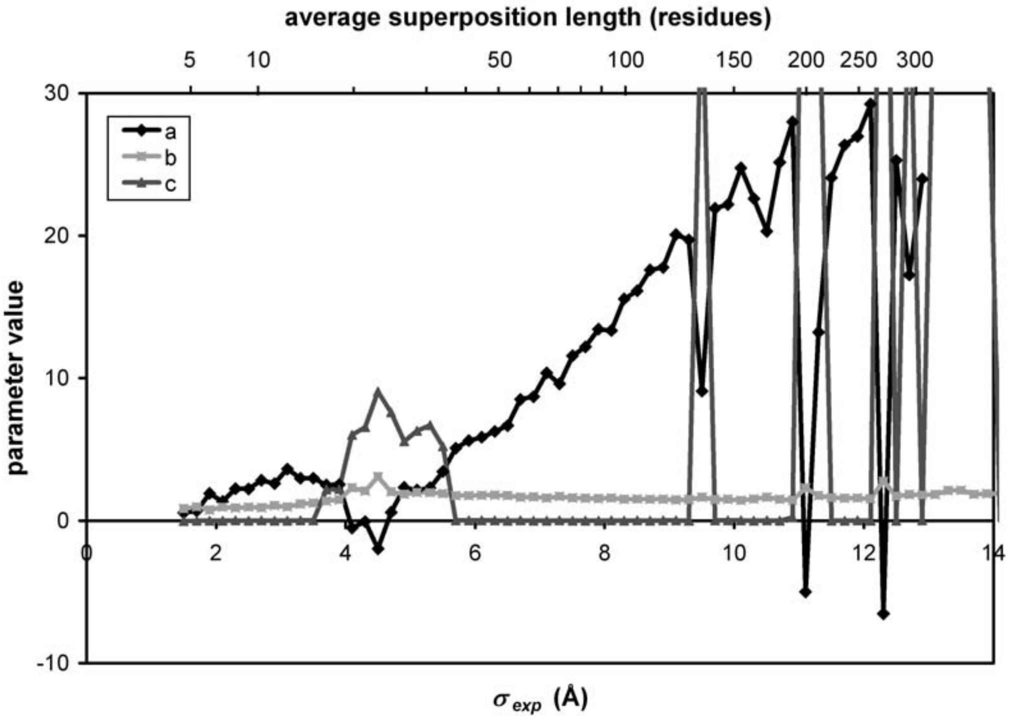


FIG. S7. Estimates of parameter values for generalized inverse Gaussian (GIG) probability density function. Parameter estimates were performed by minimum χ^2 method against the GIG probability density function: (Johnson et al., 1994; Jorgenson, 1982)

$$\rho_{GIG}(x \mid a, b, c) = \frac{|c|^{-a}}{\sqrt{2\pi|b|}K_a\left(\frac{|c|}{|b|}\right)}(c^2 + x^2)^{-\frac{1}{4}+\frac{a}{2}}K_{\frac{1}{2}-a}\left(\frac{\sqrt{c^2+x^2}}{|b|}\right) \tag{S2}$$

In Equation (S2), x = variance, and K is a special case of the modified Bessel function of the second kind, $K_n(z)$ (Equation (10), in the main text). Distributions of σ_{obs} were extracted for "slices" of fragment superposition data, sorted as a function of σ_{exp} (main text Equation (6A)), as described in Methods. Goodness of fit was assessed using the χ^2 value, main text Equation (14). Parameter a generally increased with superposition length, parameter b was relatively constant, but parameter c mostly refined to zero in these best estimates. The latter observation suggested that a three parameter PDF was unnecessary and a simpler two-parameter PDF would be adequate.