Current Protein and Peptide Science, **2002**, *3*, 441-449

Recent Progress in Protein 3D Structure Comparison

Oliviero Carugo[1,2*] and Sandor Pongor[2]

[1]*General Chemistry Department, Pavia University, viale Taramelli 12, 27100 Pavia, Italy &* [2]*Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, Area Science Park, Padriciano 99, 34012 Trieste, Italy*

Abstract: Quantitation of protein 3-D structure similarity is crucial in such fields as evolutionary studies, structural modeling and prediction of biological function. There are various approaches, many of which are tailored to specific problems. This review summarizes the recent developments in this field with particular interest in two main areas: i) improvements to and statistical interpretation of the root-man-square distance between equivalent atoms, *rmsd*; and ii) methods of protein structural classification based on geometrical features. Special attention is given to fast methods capable of analyzing large structural databases.

I. INTRODUCTION

Recognizing protein pair similarity is of fundamental importance in modern molecular biology. Traditionally, the comparisons of protein three-dimensional (3D) structures have been applied to evolutionary analyses [1-6] and to evaluate structure prediction methods [7-10]. More recently and largely because of the structural genomic initiatives [11], macromolecular 3D structures are increasingly used to predict detailed, biological information [12-16]. The simplest and probably most promising approach to identify the unknown function of a protein whose 3D structure has been determined, is based on the search for similar 3D structures of proteins with known and well characterized function. Although the structural similarity cannot *per se* guarantee that the function is the same, it can nevertheless reduce the spectrum of possibilities. In the trivial case of two proteins with very similar sequences, the structural determination is even unnecessary since the simple analysis of the sequences allows the detection of the homology. On the contrary, in case of distantly related proteins, the sequence comparison hardly allows the identification of similar proteins. The 3D structure therefore becomes necessary: given that protein spatial structures are more conserved in evolution than amino acid sequences, it is much easier and more reliable to detect 3D similarities.

The literature offers a very large number of examples in which some insight into the biological function of proteins has been reached through the analysis of the 3D structures. An excellent example was recently given by Hilgers and Ludwig, who determined the crystal structure of the quorum-sensing protein *LuxS* from *Bacillus subtilis* [17]. The role of this protein was largely unknown, though it was clear that it is involved in intracellular signaling of bacteria [18] and is expressed in most bacterial species for which complete genome sequences are available.

The determination of the 3D crystal structure of *LuxS* allowed to propose that it is an enzyme catalyzing a hydrolytic reaction, despite that it is significantly different from any other protein fold, as evidenced by the DALI of the TOP algorithms [19-21]. *LuxS* was recognized to be a homodimeric protein, by considering the geometry of the protein-protein interface [22] and the degree of sequence conservation at the protein surface. The homodimeric nature of *LuxS* was later confirmed by light scattering experiments. Two metal biosites were recognized at the homodimer interface. Given that the cation first co-ordination sphere was formed by two His, one Cys and a water molecule, it was proposed, and later confirmed by chemical analysis, that the cation is a tetra-coordinated Zn(II) atom, probably functional and not structural, because of the presence of the solvent molecule coordinated to the metal cation [23]. A functional Zn(II) cation is systematically associated with an acid-base reactivity, typically in hydrolytic reactions of peptides and amides. Close to the Zn(II) site, a substrate binding pocket was identified by stereochemical and sequence conservation analysis. The active site is quite small (90 Å^3) and a flexible gate was proposed to allow the entrance of the substrate on the basis of a simple analysis of the crystallographic B-factors, which monitors the degree of fluctuations of the atoms around their equilibrium positions.

In the case of *LuxS*, two main factors were essential to deduce some biochemical insight into the enzyme role: (i) the recognition of the homodimeric nature of the protein and (ii) the identification of a functional Zn(II) biosite. Both aspects were recognized due to the similarity between some features of *LuxS* and of other protein 3D structures. The *LuxS* protein-protein interface was compared to all the known protein-protein interfaces and the metal biosite was identified because of its apparent similarity with many other Zn(II) biosites. The availability of rational databases and descriptions of protein 3D structures were therefore essential in deducing detailed biochemical features of *LuxS*.

Here some recent advances in the methods for comparing overall protein structures are examined. The first section of the review is devoted to the most recent improvements to the

[*] Address correspondence to this author at the International Centre for Genetic Engineering and Biotechnology, Area Science Park, Padriciano 99, 34012 Trieste, Italy; E-mail: carugo@icgeb.trieste.it

1389-2037/02 $35.00+.00

root-mean-square-distance (*rmsd*) between equivalent atoms, computed after optimal superposition of a structure over another structure. Although *rmsd* is certainly the most popular measure of structural similarity, it is often used improperly because its actual meaning is rather obscure. The second part of the review is devoted to the alternative methods, based on the geometrical properties, which have been designed in order to make the browsing of very large protein 3D structure databases possible.

II. SIGNIFICANCE OF *RMSD*

The statistical significance of the *rmsd* values has been investigated several times in the past, generally by analysing distributions of *rmsd* values obtained by comparing selected, representative sets of protein three-dimensional structures [24-27]. For example, it has been proposed that two structural moieties are significantly similar if their *rmsd* value is considerably smaller than a reference value, which discriminates the 1% of the most similar fragment pairs from the remaining 99% of the less similar pairs [28]. Analogous similarity scores, based on the statistical distributions of the *rmsd* values, have also been used subsequently [29, 30]. Alternatively, adopting a totally different approach, which does not at all consider the distributions of the *rmsd* values, it has been proposed [31] that two protein structures must be considered rather similar if their *rmsd* is lower than that obtained when one of them is inverted.

II.1. *rmsd* and Protein Size

It is only recently that the dependence of the *rmsd* on the protein size has been addressed, although this problem was also implicit in several older papers. As an example, the *rmsd* threshold value able to discriminate the 1% of the most similar fragment pairs from the remaining 99% of the less similar pairs clearly depends on the protein fragment size (see Figure 1b in reference [29]). Maiorov and Crippen [32] discussed the "often neglected fact that *rmsd* is affected by both the conformational similarity and the overall sizes of the proteins being compared". They found an elegant way to standardize the measure of three-dimensional similarity. Once two protein structures have been superposed and their residues have been consequently aligned, a "sum" and a "difference" structures are defined. If structure A is defined by its $\mathbf{a}_i$ positional vectors (i.e. the coordinates of its atoms) and structure B is defined by its $\mathbf{b}_i$ positional vectors, the positional vectors of the "sum" structure are defined as,

$$\mathbf{sum}_i = \frac{\mathbf{a}_i + \mathbf{b}_i}{2} \qquad (1)$$

and the positional vectors of the "difference" structure are defined as,

$$\mathbf{dif}_i = \frac{\mathbf{a}_i - \mathbf{b}_i}{2} \qquad (2)$$

The similarity between the two structures A and B can be measured with ρ(A,B)

$$\rho\,(A,B) = \frac{2R(\text{dif})}{R(\text{sum})} \qquad (3)$$

where R(dif) and R(sum) are the radii of gyration of the "difference" and "sum" structures, respectively. It can be shown that ρ(A,B) can be computed as,

$$\rho\,(A,B) = \frac{2\,rmsd(A,B)}{\sqrt{2R^2(A) + 2R^2(B) - rmsd^2(A,B)}} \qquad (4)$$

where *rmsd*(A,B) is the root-mean-square distance computed after optimal superposition of the two structures A and B, and R(A) and R(B) are the radii of gyration of A and B, defined as

$$R^2 = \frac{\sum_{i=1}^{n} r_i^2}{n} \qquad (5)$$

where r_i denotes the distances from the centre of mass of the i^{th} atom in either structure A or B.

An alternative measure of similarity (the relative *rmsd*), unbiased by the protein dimension, has been proposed by Betancourt and Skolnick [33]. The simple ratio between the *rmsd* and the average *rmsd* for two random proteins with equivalent dimensions is taken as the unbiased similarity score. A rather similar approach to treat the dependence of the *rmsd* on the protein size has been adopted by Carugo and Pongor [34]. A size-corrected *rmsd* was designed viz the $rmsd_{100}$, so that its value is the *rmsd* that would have been observed if the two structures that are compared had 100 equivalent atoms. 180 non-homologous protein structures were selected, with very variable secondary structure content and with very variable number of residues. Each protein structure was superposed to 400,000 of its variants obtained by randomly shuffling its sequence (only the Cα atoms were considered). As expected, the 400,000 *rmsd* values were distributed differently as a function of the length of the protein. Higher rmsd values were more frequently observed for longer proteins (Fig. **1a**). Also the maximal *rmsd* value of the *n* most similar structure pairs was dependent on the protein length (Fig. **1a**). Higher values of such a maximal value were observed for longer proteins. The dependence of such a maximal *rmsd* on the protein length was logarithmic (Fig. **1b**) and the exact shape of this curve was obviously dependent on the arbitrary number *n* of the most similar protein pairs (Fig. **1c**). Nevertheless, by dividing these maximal *rmsd* values by $rmsd_{100}$, which is their value if the protein contains 100 residues, a unique curve is obtained (Fig. **1d**). Eventually, it is possible to transform any *rmsd* into a normalized $rmsd_{100}$ as

$$rmsd_{100} = \frac{rmsd}{-1.3 + 0.5\ln(N)} = \frac{rmsd}{1 + \ln\sqrt{\dfrac{N}{100}}} \qquad (6)$$

where *N* is the number of fitted atoms. $rmsd_{100}$values are, therefore, the *rmsd* values that would be observed for a pair of structures with 100 equivalent residues.

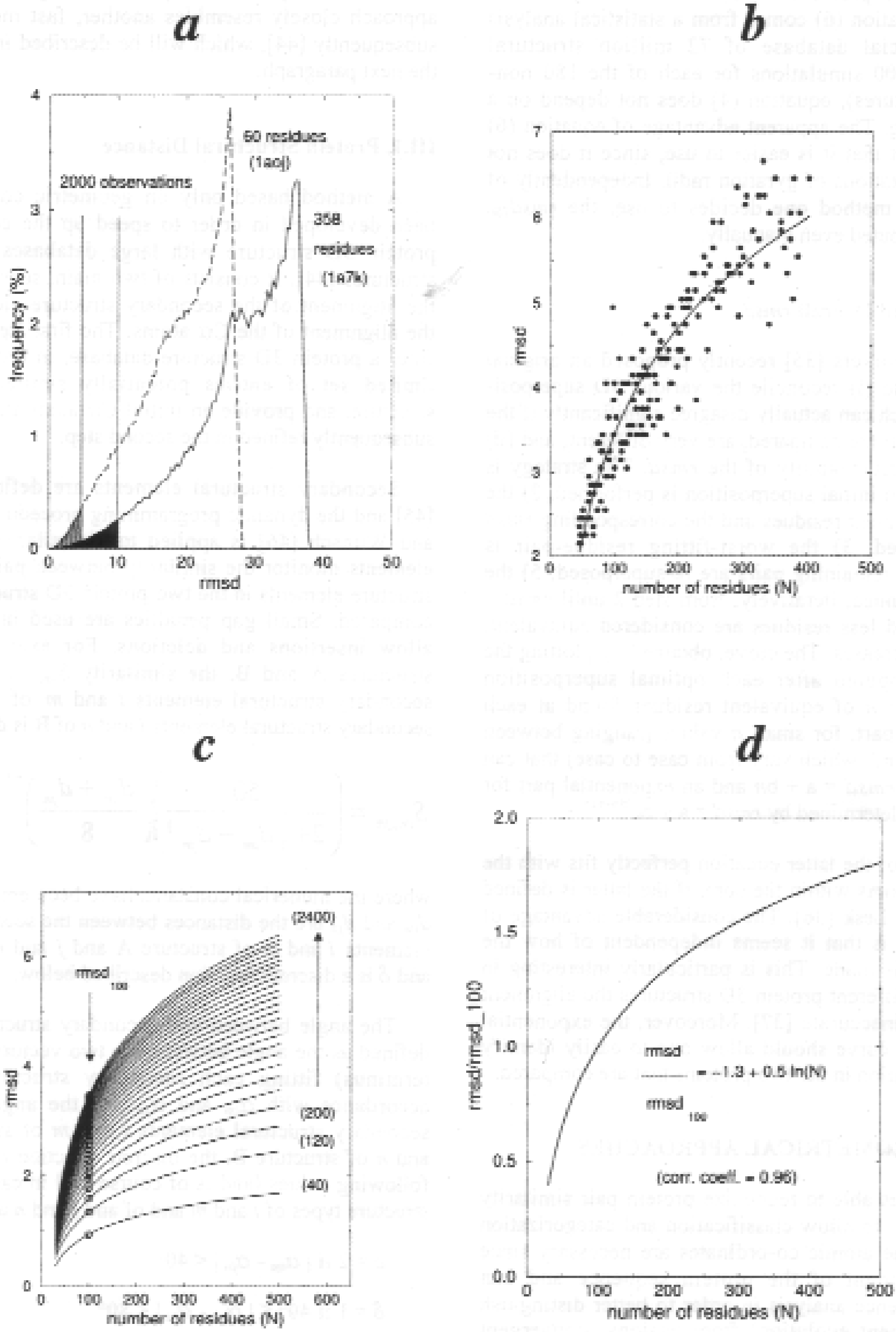

Fig. (1). *a*) Typical distribution of the *rmsd* values obtained by comparing a protein structure with 400,000 of its variants obtained by random shuffling its sequence. Larger values are observed more frequently than small values. The distribution depends on the dimension of the protein. Larger *rmsd* values are observed more frequently for larger proteins. The maximal *rmsd* value of the 2,000 best fitting comparisons are indicated. Also this maximal *rmsd* value depends on the dimension of the protein. Larger maximal *rmsd* values are associated with larger proteins. *b*) Dependence on the protein dimension of the maximal *rmsd* value of the 2,000 best fitting comparisons. *c*) Dependence on the protein dimension of the maximal *rmsd* value of the X best fitting comparisons, with X ranging from 40 to 2,400, as indicated in parentheses. The $rmsd_{100}$ values are indicated with small empty circles. *d*) The curves shown in (*c*) collapse into a single curve when they are divided by the $rmsd_{100}$ values. The curve can be fitted by an exponential function that allows to compute the $rmsd_{100}$ analogue of any *rmsd* value, see equation (6).

The approaches presented above lead to very similar results. While equation (6) comes from a statistical analysis of a large artificial database of 72 million structural alignments (400,000 simulations for each of the 180 non-homologous structures), equation (4) does not depend on a statistical sampling. The apparent advantage of equation (6) consists in the fact that it is easier to use, since it does not require the computations of gyration radii. Independently of the superposition method one decides to use, the $rmsd_{100}$ values can be computed even manually.

II.2. Core *rmsd* and Overall *rmsd*

Lesk and co-workers [35] recently proposed an original procedure aimed to (i) reconcile the various 3D superposition methods, which can actually disagree significantly if the two proteins, which are compared, are very different; and (ii) improve the understandability of the *rmsd*. The strategy is rather simple: 1) an initial superposition is performed; 2) the number *n* of equivalent residues and the corresponding *rmsd* value are recorded; 3) the worst-fitting residue-pair is eliminated; 4) the remaining pairs are re-superposed; 5) the procedure is continued, iteratively, from step 2 until $rmsd <$ 0.2 Å. As less and less residues are considered equivalent, the *rmsd* value decreases. The curve, obtained by plotting the *rmsd* values computed after each optimal superposition versus the number *n* of equivalent residues found at each step, has a linear part, for small *n* values (ranging between the values *n1* and *n2*, which vary from case to case) that can be determined by $rmsd = a + bn$ and an exponential part for $n > n2$, that can be determined by $rmsd = a + be^{d(n-n1)}$

The exponent of the latter equation perfectly fits with the *rmsd* of the Cα atoms within the core, if the latter is defined as in Chothia and Lesk [36]. The considerable advantage of such a procedure is that it seems independent of how the initial alignment is made. This is particularly interesting in the case of very different protein 3D structures the alignment of which can be inaccurate [37]. Moreover, the exponential part of the fitting curve should allow one to easily identify the similar core region in the two proteins that are compared.

III. PURELY GEOMETRICAL APPROACHES

Fast procedures able to recognize protein pair similarity and, consequently, to allow classification and categorization on the basis of the atomic co-ordinates are necessary since they are independent of the protein sequence and can complement sequence analysis in order to better distinguish homology (divergent evolution) from analogy (convergent evolution). A number of techniques have been developed to solve that task and research in the field is still very popular because of the need to explore large databases of macromolecular 3D structures. Some approaches were developed 20-30 years ago when only a few protein 3D structures were known [26, 38-40]. There are now more than 17,000 macromolecular 3D structures in the Protein Data Bank [41, 42]. This number will surely increase to a great extent in the near future as a consequence of the numerous ongoing structural genomics initiatives [11]. Quite a substantial progress has recently been reported in the automatic comparison of protein 3D geometry [43]. The approach closely resembles another, fast method proposed subsequently [44], which will be described in some detail in the next paragraph.

III.1. Protein Structural Distance

A method based only on geometric consideration has been developed in order to speed up the comparison of a protein 3D structure with large databases of protein 3D structures [44]. It consists of two main, successive steps, (i) the alignment of the secondary structure elements, and (ii) the alignment of the Cα atoms. The first step allows one to filter a protein 3D structure database, in order to fish out a limited set of entries potentially similar to the query structure, and provide an initial Cα atom alignment. This is subsequently refined in the second step.

Secondary structural elements are defined with DSSP [45] and the dynamic programming procedure of Needleman and Wunsch [46] is applied to a similarity matrix whose elements monitor the similarity between pairs of secondary structure elements in the two protein 3D structures which are compared. Small gap penalties are used in order to easily allow insertions and deletions. For example, given two structures A and B, the similarity $S_{im,jn}$ between the two secondary structural elements *i* and *m* of A and the two secondary structural elements *j* and *n* of B is defined as

$$S_{im,jn} = \left(\frac{50}{2 + | d_{im} - d_{jn} |} \right) \left(\frac{d_{im} + d_{jn}}{8} \right)^{-1.7} \delta \qquad (7)$$

where the numerical constants have been empirically chosen, d_{im} and d_{jn} are the distances between the secondary structural elements *i* and *m* of structure A and *j* and *n* of structure B, and δ is a discrete function described below.

The angle between two secondary structural elements is defined as the angle between the two vectors (from N- to C-terminus) fitting each secondary structural element. In accordance with α_{im} and α_{jn} ,i.e. the angles between the secondary structural elements *i* and *m* of structure A and j and *n* of structure B, the discrete function δ can assume the following values (and is of course null in case the secondary structure types of *i* and *m* and of and *j* and *n* are different)

$$\begin{aligned} &\delta = 2 \text{ if } | \alpha_{im} - \alpha_{jn} | < 40° \\ &\delta = 1 \text{ if } 40° \leq | \alpha_{im} - \alpha_{jn} | < 80° \\ &\delta = 0 \text{ if } | \alpha_{im} - \alpha_{jn} | \geq 80° \end{aligned} \qquad (8)$$

The distance between two secondary structural elements *i* and *m* (or *j* and *n*) containing l_i and l_m residues, respectively, is given by

$$d_{im} = \frac{\sum_{k=1}^{l_i} \min(a_{k1}, a_{k2}, \ldots, a_{kl_m})}{l_i} \qquad (9)$$

where a_{xy} is the distance between residue *x* of the secondary structural elements *i* and residue *y* of the secondary structural elements *m*.

Once the equivalences among the secondary structural elements of the two structures have been found, the Cα atoms are aligned with an algorithm similar to STAMP [47] which is a combination of dynamic programming and rigid-body superposition, performed with the very popular algorithm of Kabsch [48, 49]. The initial Cα atoms alignment, which must be as accurate as possible in order to ensure the convergence towards the final alignment, is obtained from the initial alignment of the secondary structural elements, by assuming that the small subset of residues with similar solvent accessibility within the aligned secondary structural elements are equivalent. The two structures are then superposed according to such an alignment and the similarity between residues is defined as

$$S_{ij}(\mathrm{n}+1) = S_{ij}(\mathrm{n}) + \frac{50}{\max(d_{ij},1)} \qquad (10)$$

where d_{ij} is the distance between the Cα-s of residues *i* and *j*, S_{ij}(n) is the initial similarity between residues *i* and *j* and S_{ij}(n+1) is the similarity between the same two residues after rigid-body superposition. By processing such a similarity matrix with the Needleman-Wunsch algorithm [46], a new alignment is obtained, a new rigid-body superposition is performed, the new similarity scores S_{ij} are computed and re-analyzed by dynamic programming. About ten iterations are needed until a convergence of the *rmsd* is reached.

Given that the procedure described above consists in two steps, i.e. the alignment of the secondary structural elements and the alignment of the *Cα* atoms, the overall similarity between two protein 3D structures, the protein structural distance (*PSD*), is given by a mixture of the goodness-of-fits of the two alignment types

$$PSD(\mathrm{A,B}) = \left\{ -\frac{\log\left[\left(\dfrac{a}{\max(a,b)}\right)\left(\dfrac{s(\mathrm{A,B})}{s(\mathrm{A,A})}\right)\right]}{\log(x)} \right\}^2 + \left(\frac{rmsd}{y}\right)^2 \qquad (11)$$

where *a* and *b* are the number of secondary structural elements of proteins A and B, respectively, *s*(A,B) is the secondary structural element alignment score for comparing protein A to protein B, *s*(A,A) is the analogous score for comparing protein A with itself, and *x* and *y* are adjustable parameters that should be close to 3 and 5, respectively. Their role is to weight the relative importance of the two addenda which define the PSD, the first measuring the alignment of the secondary structural elements and the second reflecting the alignment of the Cα atoms. The ratio *s*(A,B)/*s*(A,A), which ranges between 0 and 1, normalizes the alignment score with respect to one of the two protein which are compared. The ratio *a*/max(*a*,*b*) makes the PSD symmetrical, so that *PSD*(A,B) ≅ *PSD*(B,A).

The *PSD* score's ability to identify structural similarities when a query protein 3D structure is compared to a structural database has been analyzed statistically. An all-versus-all comparison of 1226 structures of the SCOP collection of protein domains [5] was very satisfactory and reasonably fast. About 20 seconds were necessary to compare myoglobin (153 residues) to the representative set of 1226 proteins and the amount of errors, i.e. the entries incorrectly declared similar or dissimilar to the query 3D structure, was quite modest. Very severe discrepancies with SCOP were observed only in 0.1% of the comparisons. There were nevertheless significant differences between the classification of protein structure of SCOP and the classifications obtained on the basis of the *PSD* scores. Amongst the various justifications for these inconsistencies, one is of primary relevance. While as the SCOP hierarchical organization is based on both sequence and 3D structure information, *PSD* derives from a purely geometrical analysis. Moreover, while SCOP requires some human decisions, which can be variable from case to case, *PSD* results from an automatic procedure. It must also be noted that while SCOP assumes the discontinuity from fold to fold, so that there are regions of the fold space, which cannot be occupied by any stable protein, the *PSD* scores apparently suggest that there is some continuity in the conformational space [44].

III.2. Comparison Without Alignment, PRIDE

Traditionally, the similarity between two protein 3D structures is evaluated by an analysis of inter-atomic distances [19, 40, 50] or by rigid-body superposition [51-53]. These methods have both advantages and disadvantages. On one side, when the co-ordinates of the Cα atoms are considered, they allow the 3D alignment of the residues, with a significant improvement in the sequence alignment quality. This is particularly important given that only the alignment based on the 3D structures allows one to find correct residues equivalencies in case of very distantly related proteins. About twice as many distant homologous protein pairs were detected by Levitt and Gerstein with the 3D alignments rather than with sequence alignments [54]. Analogously, only 10% of known relationships in PDB [41, 42] were detected with BLAST [55] by Brenner *et al.* and many of them were undetectable even with the more sophisticated PSI-BLAST procedure [56], which builds and uses profiles for a certain family [57].

Nevertheless, the 3D alignment of a pair of protein 3D structures becomes computationally rather expensive and partially unreliable when the two proteins are very different in shape and size [37]. Sippl and co-workers have recently discussed this point in great detail [58]. They compared the results of the rigid-body superposition program ProSup [59] with those of several, other procedures like VAST [60], DALI [19], and CE [37]. 10 well-documented "difficult" protein pairs [61, 62] were examined. In all cases, significant differences were observed amongst at least two procedures. The extreme discrepancy was observed in comparing the canine granulocyte colony-stimulating factor (PDB identification code 1bge, chain B)[63] with the human granulocyte-macrophage colony-stimulating factor (PDB identification code 2gmf, chain A) [64]. ProSup found 87 equivalent residues that were totally different from those found by VAST, DALI, and CE which found alternative 71, 94, and 107 equivalent residues, respectively.

A novel method to compare protein 3D structure pairs without superposing them or aligning their equivalent residues has recently been proposed (Fig. 2) [65]. The PRIDE score (PRObability of IDentity) is based on a simplified protein 3D structure representation. The $C\alpha(i)$-$C\alpha(i+n)$ distances are computed and their distributions are stored. The integer n varies from 3 to 30 and indicates the sequence distance between the residues i and $i+n$. Each 3D structure is ,therefore, characterized by 28 histograms, each describing the distribution of the distances between $C\alpha$ atoms separated by 3,4,....30 residues. When two protein 3D structures are compared, each pair of these 28 histograms is compared using contingency table analysis [66], a very robust and model-free statistical tool, resulting in 28 "probability of identity" values. Their average is then defined as PRIDE; its value ranges from 0 (totally different 3D structures) to 1 (identical 3D structures). Through an all-versus-all comparison, the PRIDE scores have been shown to be able to recognize the level of similarity among the protein domains classified in the CATH database [6].

Two important features of PRIDE are noted. (i) It is intuitively more understandable then other measures of similarity, since it has both an upper and a lower limit (1 and 0, respectively) and it indicates a probability, a concept easily appreciable also to non-specialists. (ii) It is very fast. Once the 28 histograms per each structure have been computed and stored, like one must do when scanning a large database of protein 3D structures, 40,000 structural comparisons can be performed in about 37 seconds with a simple SGI R10000 processor.

Another relevant feature of PRIDE is that it is a metric in the mathematical sense. Given three structures A, B, and C, the following properties are crucial if PRIDE is to be used in automatic structure classification:

$$\mathrm{PRIDE}(AA)=1 \qquad \text{(identity)} \qquad (12)$$

$$\mathrm{PRIDE}(AB)\geq 0 \qquad \text{(positivity)} \qquad (13)$$

$$\mathrm{PRIDE}(AB)=\mathrm{PRIDE}(BA) \qquad \text{(symmetry)} \qquad (14)$$

$$(1-\mathrm{PRIDE}(AC)) \leq (1-\mathrm{PRIDE}(AB))+(1-\mathrm{PRIDE}(BC)) \qquad \text{(triangular inequality)} \qquad (15)$$

Since protein similarity/distance measures are often used in conjunction with cluster analysis and other multivariate statistical techniques, it is worth while to point out that PRIDE (or its complementary, 1-PRIDE) possesses the metric properties (equations 12-15) required for such calculations. These metric properties are not always found in other 3D similarity scores. For example, the *rmsd* has been shown to obey the triangular inequality [67], but this is true only for very similar structures that can be superposed unambiguously.

CONCLUSIONS

This review describes recent trends in assessing the similarity of protein 3D structures. The dependence of root mean square distance, *rmsd*, on the protein dimension has received considerable attention. Three independent alternative similarity measures, the so-called ρ (equation 4, [32]), the relative *rmsd* [33], and the $rmsd_{100}$ (equation 6,[34]) scores have been proposed to make the similarity measure independent of the dimension of the proteins compared. Furthermore, a simple and elegant procedure has been designed to distinguish the overall *rmsd* from the *rmsd* of the protein core [35]. This is particularly important when the two proteins, which are compared, are considerably different from each other, since in this case the overall similarity can greatly vary depending on the method of comparison, and also, totally different three-dimensional alignments are possible. The possibility to reliably measure the core *rmsd*, on the other hand, should help in detecting low similarity degrees.

Considerable interest has been devoted to the development of methods, which are based on purely geometrical criteria and may therefore allow an automated classification or categorization of protein 3D structures independent of the amino acid sequences. These methods are particularly needed in developing protein domain databases that can be used in homology modeling or fold recognition procedures [68]. The PSD score [44] is computed in two steps, an initial alignment of secondary structural elements followed by an alignment of the $C\alpha$ positions. The PRIDE score [65] is based on the comparison of the length distribution of the $C\alpha$-$C\alpha$ distances. This comparison is very fast but does not provide a 3D alignment of discrete residue pairs. Despite their conceptual simplicity, both PSD and PRIDE have been shown to compare well with knowledge-based protein fold classifications like CATH [6] and SCOP [5].

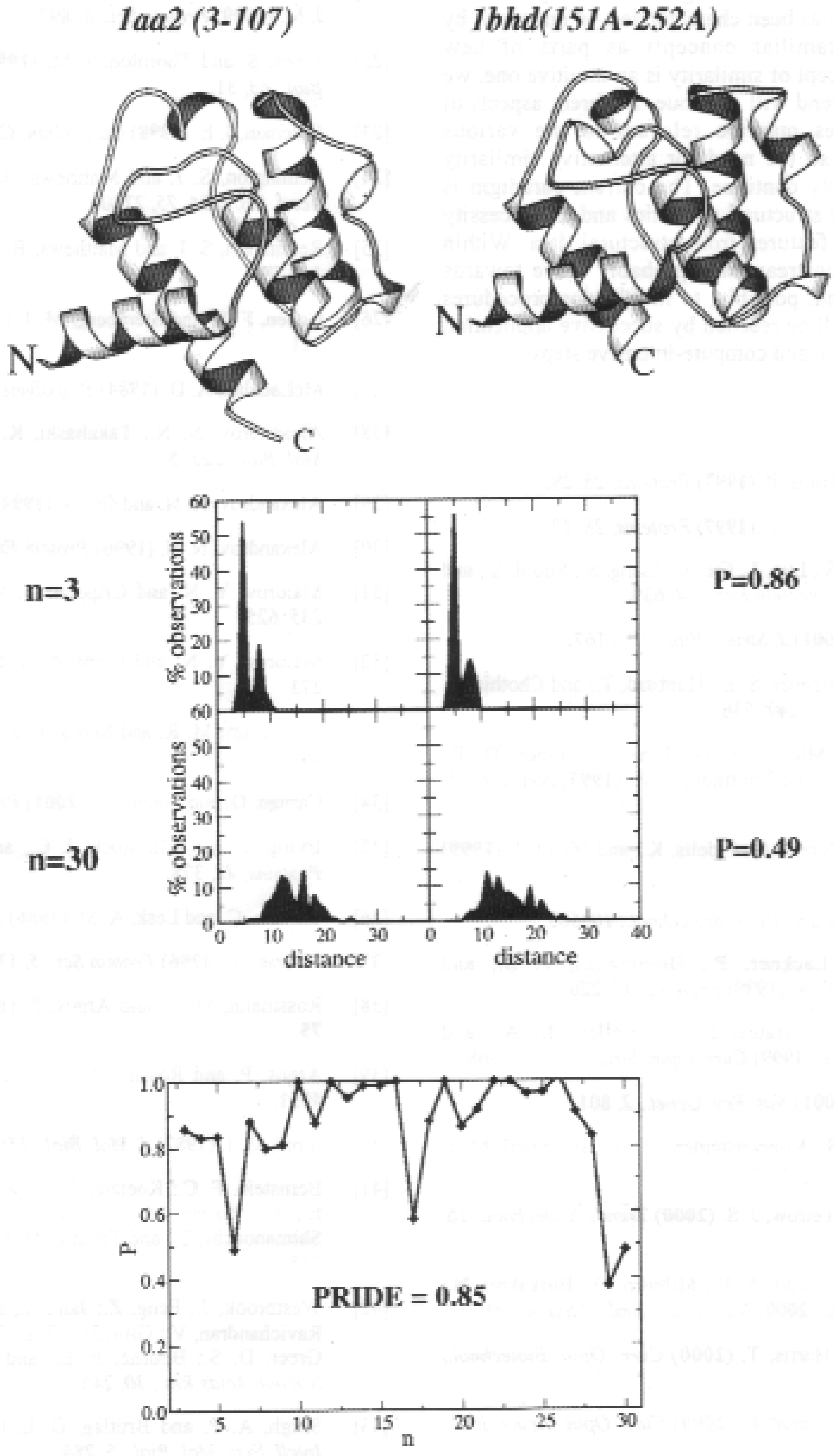

Fig. (2). An example of the procedure that compares two protein structures through the analysis of the distributions of the Cα(*i*)-Cα (*i*+*n*) distances. The calponin homology domains of the human beta spectrin (1aa2, residues 3-107) [69] and of the human utrophin (1bhd, residues 151-252 of the chain A) [70] are considered (top; figure prepared with Molscript [71]. The distances Cα(*i*)-Cα (*i*+*n*) are computed for 3 ≤ *n* ≤ 30. Two distributions of these distances are shown (middle) for *n* = 3 and *n* = 30. By contingency table analysis it is possible to estimate that the pair of distributions associated with *n* =3 have 86% of probability to be identical. The probability that the other two distributions are identical is lower, 49%. All the 28 probability values for 3 ≤ *n* ≤ 30 are computed. They are plotted versus *n* (bottom). Their mean value, PRIDE, is 0.86. The two CH domains have therefore 85% of probability to be identical.

The recent progress in assessing the similarity between protein 3D structures has been characterized, in our view, by a reemergence of familiar concepts as parts of new strategies. As the concept of similarity is an intuitive one, we can expect that this trend will continue. Different aspects of protein 3D structures may be relevant to the various biological problems, so the need for alternative similarity measures will probably continue. The current paradigm is largely determined by structural genomics and the necessity to predict biological features from structural data. Within this scenario, the mainstream will probably move towards fast methodologies and, possibly, to multi-steps procedures in which the target will be reached by successive application of increasingly selective and compute-intensive steps.

REFERENCES

[1] Carugo, O. and Argos, P. (**1997**) *Proteins, 28,* 29.

[2] Carugo, O. and Argos, P. (**1997**) *Proteins, 28,* 10.

[3] Carugo, O.; Lu, S.; Luo, J.; Gu, X.; Liang, S.; Strobl, S.; and Pongor, S. (**2001**) *Protein Eng., 14,* 639.

[4] Grishin, N. V. (**2001**) *J. Struct. Biol., 134,* 167.

[5] Murzin, A. G.; Brenner, S. E.; Hubbard, T.; and Chothia, C. (**1995**) *J. Mol. Biol., 247,* 536.

[6] Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; and Thornton, J. M. (**1997**) *Structure, 5,* 1093.

[7] Venclovas, C.; Zemla, A.; Fidelis, K.; and Moult, J. (**1999**) *Proteins, Suppl.,* 231.

[8] Moult, J. (**1999**) *Curr. Opin. Biotechnol., 10,* 583.

[9] Sippl, M. J.; Lackner, P.; Domingues, F. S.; and Koppensteiner, W. A. (**1999**) *Proteins, 37,* 226.

[10] Sternberg, M. J.; Bates, P. A.; Kelley, L. A.; and MacCallum, R. M. (**1999**) *Curr. Opin. Struct. Biol., 9,* 368.

[11] Brenner, S. E. (**2001**) *Nat. Rev. Genet., 2,* 801.

[12] Domingues, F. S.; Koppensteiner, W. A.; and Sippl, M. J. (**2000**) *FEBS Lett., 476,* 98.

[13] Skolnick, J. and Fetrow, J. S. (**2000**) *Trends Biotechnol., 18,* 34.

[14] Thornton, J. M.; Todd, A. E.; Milburn, D.; Borkakoti, N.; and Orengo, C. A. (**2000**) *Nat. Struct. Biol., 7 Suppl.,* 991.

[15] Shapiro, L. and Harris, T. (**2000**) *Curr. Opin. Biotechnol., 11,* 31.

[16] Moult, J. and Melamud, E. (**2000**) *Curr. Opin. Struct. Biol., 10,* 384.

[17] Hilgers, M. T. and Ludwig, M. L. (**2001**) *Proc. Natl. Acad. Sci. USA, 98,* 11169.

[18] Bassler, B. L. (**1999**) *Curr. Opin. Microbiol., 2,* 582.

[19] Holm, L. and Sander, C. (**1993**) *J. Mol. Biol., 233,* 123.

[20] Gilbert, D.; Westhead, D.; Nagano, N.; and Thornton, J. (**1999**) *Bioinformatics, 15,* 317.

[21] Westhead, D. R.; Slidel, T. W.; Flores, T. P.; and Thornton, J. M. (**1999**) *Protein Sci., 8,* 897.

[22] Jones, S. and Thornton, J. M. (**1995**) *Prog. Biophys. Mol. Biol., 63,* 31.

[23] Coleman, J. E. (**1998**) *Curr. Opin. Chem. Biol., 2,* 222.

[24] Remington, S. J. and Matthews, B. W. (**1978**) *Proc. Natl. Acad. Sci. USA, 75,* 2180.

[25] Remington, S. J. and Matthews, B. W. (**1980**) *J. Mol. Biol., 140,* 77.

[26] Cohen, F. E. and Sternberg, M. J. (**1980**) *J. Mol. Biol., 138,* 321.

[27] McLachlan, A. D. (**1984**) *Biopolymers, 23,* 1325.

[28] Alexandrov, N. N.; Takahashi, K.; and Go, N. (**1992**) *J. Mol. Biol., 225,* 5.

[29] Alexandrov, N. N. and Go, N. (**1994**) *Protein Sci., 3,* 866.

[30] Alexandrov, N. N. (**1996**) *Protein Eng., 9,* 727.

[31] Maiorov, V. N. and Crippen, G. M. (**1994**) *J. Mol. Biol., 235,* 625.

[32] Maiorov, V. N. and Crippen, G. M. (**1995**) *Proteins, 22,* 273.

[33] Betancourt, M. R. and Skolnick, J. (**2001**) *Biopolymers, 59,* 305.

[34] Carugo, O. and Pongor, S. (**2001**) *Protein Sci., 10,* 1470.

[35] Irving, J. A.; Whisstock, J. C.; and Lesk, A. M. (**2001**) *Proteins, 42,* 378.

[36] Chothia, C. and Lesk, A. M. (**1986**) *EMBO J., 5,* 823.

[37] Godzik, A. (**1996**) *Protein Sci., 5,* 1325.

[38] Rossmann, M. G. and Argos, P. (**1976**) *J. Mol. Biol., 105,* 75.

[39] Argos, P. and Rossmann, M. G. (**1979**) *Biochemistry, 18,* 4951.

[40] Sippl, M. J. (**1982**) *J. Mol. Biol., 156,* 359.

[41] Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. E.; Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; and Tasumi, M. (**1977**) *J. Mol. Biol., 112,* 535.

[42] Westbrook, J.; Feng, Z.; Jain, S.; Bhat, T. N.; Thanki, N.; Ravichandran, V.; Gilliland, G. L.; Bluhm, W.; Weissig, H.; Greer, D. S.; Bourne, P. E.; and Berman, H. M. (**2002**) *Nucleic Acids Res., 30,* 245.

[43] Singh, A. P. and Brutlag, D. L. (**1997**) *Proc. Int. Conf. Intell. Syst. Mol. Biol., 5,* 284.

[44] Yang, A. S. and Honig, B. (**2000**) *J. Mol. Biol., 301,* 665.

[45] Kabsch, W. and Sander, C. (**1983**) *Biopolymers, 22,* 2577.

[46] Needleman, S. B. and Wunsch, C. D. (**1970**) *J. Mol. Biol., 48,* 443.

[47] Russell, R. B. and Barton, G. J. (**1992**) *Proteins, 14,* 309.

[48] Kabsch, W. (**1976**) *Acta Crystallogr., A32*, 922.

[49] Kabsch, W. (**1978**) *Acta Crystallogr., A34*, 827.

[50] Taylor, W. R. and Orengo, C. A. (**1989**) *J. Mol. Biol., 208*, 1.

[51] May, A. C. and Johnson, M. S. (**1994**) *Protein Eng., 7*, 475.

[52] Diederichs, K. (**1995**) *Proteins, 23*, 187.

[53] Zuker, M. and Somorjai, R. L. (**1989**) *Bull. Math. Biol., 51*, 55.

[54] Levitt, M. and Gerstein, M. (**1998**) *Proc. Natl. Acad. Sci. USA, 95*, 5913.

[55] Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; and Lipman, D. J. (**1990**) *J. Mol. Biol., 215*, 403.

[56] Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; and Lipman, D. J. (**1997**) *Nucleic Acids Res., 25*, 3389.

[57] Brenner, S. E.; Chothia, C.; and Hubbard, T. J. (**1998**) *Proc. Natl. Acad. Sci. USA, 95*, 6073.

[58] Lackner, P.; Koppensteiner, W. A.; Sippl, M. J.; and Domingues, F. S. (**2000**) *Protein Eng., 13*, 745.

[59] Feng, Z. K. and Sippl, M. J. (**1996**) *Fold Des., 1*, 123.

[60] Gibrat, J. F.; Madej, T.; and Bryant, S. H. (**1996**) *Curr. Opin. Struct. Biol., 6*, 377.

[61] Shindyalov, I. N. and Bourne, P. E. (**1998**) *Protein Eng., 11*, 739.

[62] Fischer, D.; Eloggsson, A.; Rice, D. W.; and Eisenberg, D.; Proceedings of the 1st Pacific Symposium on Biocomputing, World Scientific Publishing, Singapore. (**1996**), pp 300.

[63] Lovejoy, B.; Cascio, D.; and Eisenberg, D. (**1993**) *J. Mol. Biol., 234*, 640.

[64] Rozwarski, D. A.; Diederichs, K.; Hecht, R.; Boone, T.; and Karplus, P. A. (**1996**) *Proteins, 26*, 304.

[65] Carugo, O. and Pongor, S. (**2002**) *J. Mol. Biol., 315*, 887.

[66] Dowdy, S. and Wearden, S. *Statistics for Research*, Wiley, New York, (**1991**).

[67] Kaindl, K. and Steipe, B. (**1997**) *Acta Crystallogr., A53*, 809.

[68] Hirst, J. D.; Brown, W. E., G. C. Howard, Ed.; Computer Modeling of Protein Structures, CRC, Boca Raton, (**2002**).

[69] Djinovic-Carugo, K.; Banuelos, S.; and Saraste, M. (**1997**) *Nat. Struct. Biol., 4*, 175.

[70] Keep, N. H.; Norwood, F. L.; Moores, C. A.; Winder, S. J.; and Kendrick-Jones, J. (**1999**) *J. Mol. Biol., 285*, 1257.

[71] Kraulis, P. (**1991**) *J. Appl. Cryst., 24*, 946.