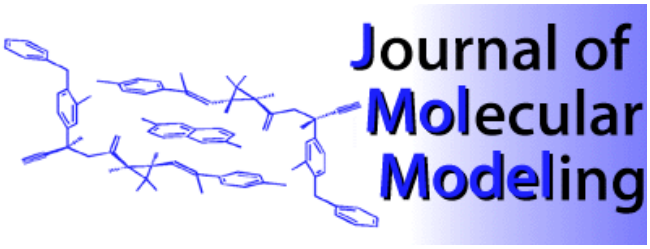
J. Mol. Model. **2000**, *6*, 539 – 549



FULL PAPER

Gaussian-based Alignment of Protein Structures: Deriving a Consensus Superposition when Alternative Solutions Exist

Jordi Mestres

Department of Molecular Design & Informatics, N.V. Organon, 5340 BH Oss, The Netherlands.
Email: j.mestres@organon.oss.akzonobel.nl

Received: 5 August 1999/ Accepted: 25 April 2000/ Published: 17 August 2000

Abstract The use of a Gaussian-based representation of protein structures for evaluating protein-structure similarities and deriving three-dimensional superpositions is presented. The approach, as implemented in the program GAPS, is applied to three pairs of proteins with different topological characteristics (rich α-helix, mixed α-helix/β-strand, and rich β-strand), low sequence identities (10-30%), and recognized difficulties to define a unique optimum alignment. Validation of the GAPS superpositions is done by comparison with superpositions obtained by the TOP, GA_FIT, and ALIGN programs and those directly extracted from the FSSP database. Results suggest that a Gaussian-based methodology offers an objective means to, depending on the Gaussian-based representation, derive a consensus three-dimensional superposition when alternative superposition solutions exist.

Keywords Protein similarity, Protein structure comparison, Protein structure superposition, Consensus alignment

Introduction

The number of protein structures available in the Protein Data Bank[1] has become significant and it is expected to increase rapidly in coming years.[2] Routine comparative analysis of protein structures is thus becoming indispensable for transforming the information inherent in these structures into relational information that can be more useful for predicting and classifying protein folds,[3] deriving evolutionary relationships,[4] or modeling of proteins by homology,[5] among many other aspects.

Three-dimensional comparison of protein structures involves finding the optimum superposition. Over the past few years, a large variety of strategies has been developed to obtain relevant protein-structure superpositions[6-23] and databases of protein-structure superpositions are now directly available from the web.[24-28] However, there are still important ambiguities in defining and characterizing the optimum superposition between protein structures.[29,30] On one hand, because the optimum superposition depends on the particular measure used to quantify the three-dimensional similarity between proteins, each technique will produce an essentially different optimum superposition for the same pair of protein structures.[24,29] On the other hand, depending on the particular topological characteristics of some protein structures, multiple superposition solutions may be identified and hence even the existence of a unique optimum superposition can be questioned.[29,30]

In cases where multiple superposition solutions may exist, the relevance of selecting one of the superpositions as the optimum structural superposition goes beyond the structural level. It will ultimately have implications in all post-superposition analyses, both at a structural level (analysis

of protein domain movements, for example) and at a sequence level (construction of structure-based sequence alignments, for example). Therefore, the selection of the optimum superposition among the different superposition solutions is an important issue. At a sequence level, the normal criteria used are either the maximum number of residues aligned or the minimum root mean square deviation (rmsd) of the residues aligned in the sequence alignments obtained from the different structural superpositions. However, as stated previously,[30] the use of these criteria is ambiguous as it is not clear, for instance, whether alignment of *M* residues to *m* Å rmsd is more significant than aligning *N* (*N*>*M*) residues to *n* (*n*>*m*) Å rmsd. Alternatively, one could try to eliminate ambiguities at the sequence level by deriving a consensus superposition at the structural level that will ultimately lead to a consensus sequence alignment.

This work aims at presenting the use of a Gaussian-based approach to protein-structure similarity as a strategy for deriving a consensus optimum superposition when multiple superposition solutions exist. A Gaussian representation of a protein structure provides a fuzzier means to define the positions of atoms in space and, thus, it is in principle more suitable for obtaining relevant superpositions and to avoid being trapped in marginal superpositions due to the locality of the protein-structure representation. Moreover, the degree of fuzzyness induced by a Gaussian representation can be controlled by using different Gaussian descriptions and this will also have implications in the optimum superposition and its uniqueness. The performance of the present Gaussian-based approach to derive a consensus optimum protein-structure superposition is illustrated in three pairs of proteins showing distinct topological characteristics.

Methodology

A program for Gaussian-based Alignment of Protein Structures, GAPS, has been developed. GAPS is a modified version of the MIMIC program for obtaining ligand superpositions,[31] which has been recently adapted to protein superpositions.[32] In GAPS, every atom *i* in the protein is represented by a Gaussian function, g_i, centered at the atom position, $\boldsymbol{R}_i$, as

$$g_i(\boldsymbol{r}) = \alpha_i \cdot \exp(-\beta_i |\boldsymbol{r}-\boldsymbol{R}_i|^2) \quad (1)$$

where the coefficient, α_i, and the exponent, β_i, determine the value of its maximum height at the origin and its decay, respectively. A Gaussian-based representation of the structure of a protein *A*, P_A, is then defined as

$$P_A(\boldsymbol{r}) = \sum_{i \in A} g_i(\boldsymbol{r}) \quad (2)$$

It has been shown that the regular features of protein secondary structure such as α-helix and β-sheet are clearly defined by the trace of the protein C_α carbons.[33] Therefore, in order to speed up similarity calculations, only C_α carbons were considered in the present study.

Once a Gaussian representation of the protein structure is defined, the structural similarity between two proteins *A* and *B*, S_{AB}, is assessed by evaluating the overlap integral, Z_{AB}, between their respective representations, P_A and P_B, as

$$Z_{AB}(\boldsymbol{t},\theta) = \int P_A(\boldsymbol{r})\, P_B(\boldsymbol{r})\, d\boldsymbol{r} \quad (3)$$

which can be then normalized using a cosine-like index

$$S_{AB}(\boldsymbol{t},\theta) = \frac{Z_{AB}(\boldsymbol{t},\theta)}{(Z_{AA} \times Z_{BB})^{1/2}} \quad (4)$$

The values of S_{AB} in eq. (4) range from 0 to 1. A value of 1 is achieved only in the limiting case of identity. Any dissimilarity between the two proteins will be reflected in a value smaller than 1.

Exploration of the structural similarity between a pair of proteins is performed using a systematic spherical search.[31] Basically, one of the proteins is kept fixed (the reference protein) while the other protein (the target protein) is systematically placed in a number of unique starting orientations about the reference protein. Then, from each starting orientation the structural similarity between the two proteins is optimized in all translational (***t***) and rotational (θ) degrees of freedom using common gradient-seeking techniques. This procedure ensures a wide and uniformly distributed exploration of the similarity landscape defined by the structural characteristics of the two proteins. The sampling of the search depends on the rotational step of the sphere used to define the starting orientations.

Furthermore, note that protein-structure similarities as computed from eq. (4) depend on the parameters of the Gaussian functions defined by eq. (1). This means that for a given maximum height, α_i, different decays, β_i, will lead to different values of structural similarity. On the one side, for very small β_i values (which could be associated to a low-resolution description of protein structures) every protein structure would look almost alike, whereas on the other side, very large β_i values (which could be associated to a high-resolution description of protein structures) would result in no overlap at all between the structural representations and, thus, every protein structure would be essentially unique. In between these two limiting cases there is a long range of possibilities and, ultimately, β_i values could be user-customizable. In the present study, the coefficients α_i were defined by $\alpha_i = 0.4798 * Z_i^{3.1027}$, where Z_i is the atomic number of atom *i*.[34] Once α_i is defined, the exponent β_i in eq. (1) can be set to give "practically null" values of $g_i(\mathbf{R})$ at radius **R** from the atom centers.[35] Throughout the work, Gaussian representations vanishing at 1, 2, 5, and 10 Å will be referred as G1, G2, G5, and G10, respectively.

Throughout this study, in order to examine the connections between alignment solutions from different Gaussian

representations the following procedure was used. Initially, a G1 representation was selected. Due to the locality of a G1 representation, a large number of local similarity maxima exist. Therefore, to ensure an extensive exploration of the similarity function defined by the topological characteristics of the two protein structures under study, a rotational step of 15 degrees (6384 starting orientations) was applied during the systematic spherical search. Then, each superposition solution *i* obtained using a G1 representation, denoted as (*i*,G1), was systematically reoptimized using the other Gaussian-based protein-structure representations. Thus, the original (*i*,G1) superposition solution converged to a (*j*,G2) superposition solution, which then evolved to a (*m*,G5) superposition solution and finally to a (*n*,G10) superposition

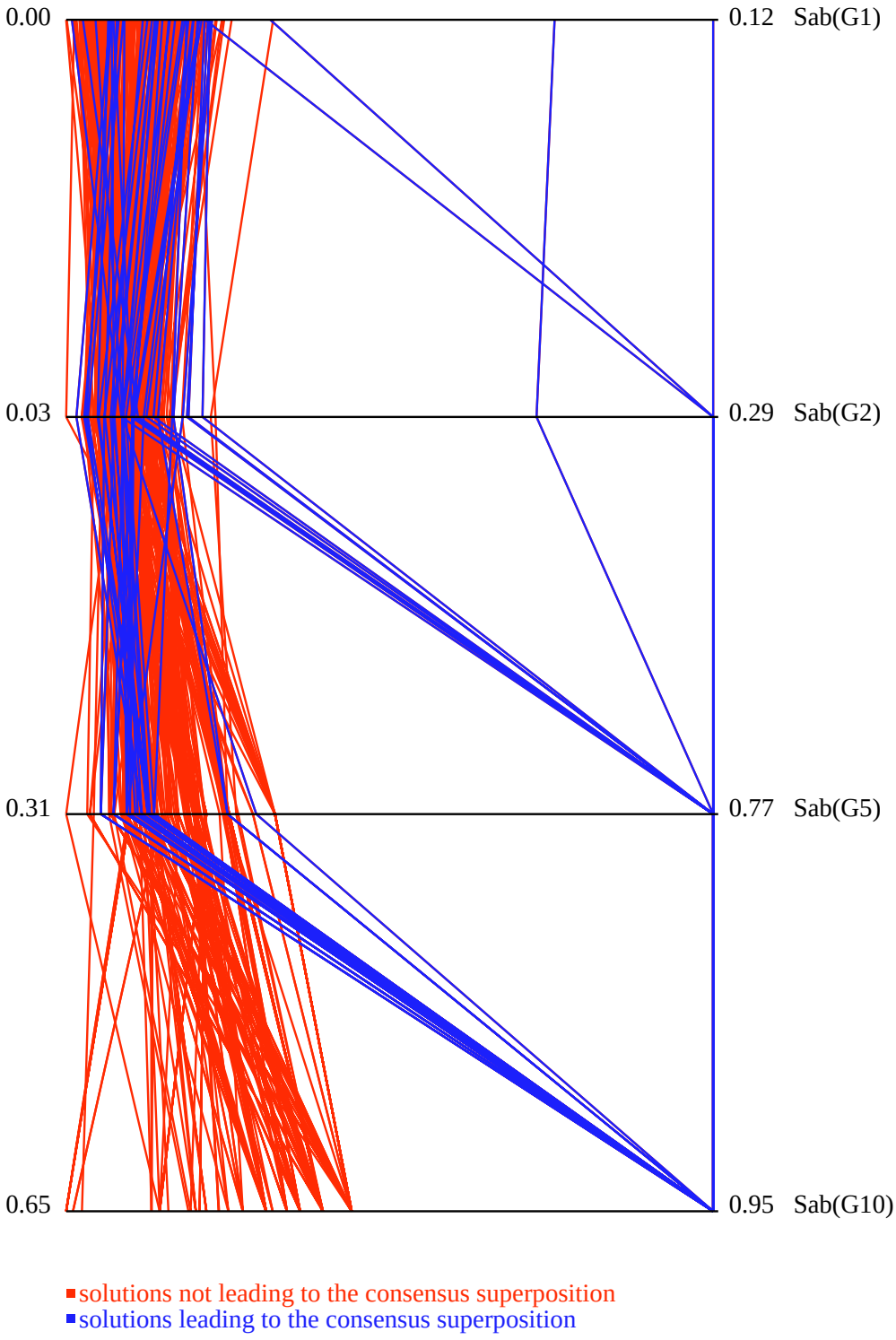


Figure 1 *Connectivity between the superposition solutions obtained from each Gaussian-based representation for the {1GUH,1GSS} pair of proteins (see text). Each parallel coordinate represents the range of similarity values (Sab in eq. (4)) assigned to the superposition solutions obtained from each Gaussian-based representation (G1, G2, G5, and G10)*

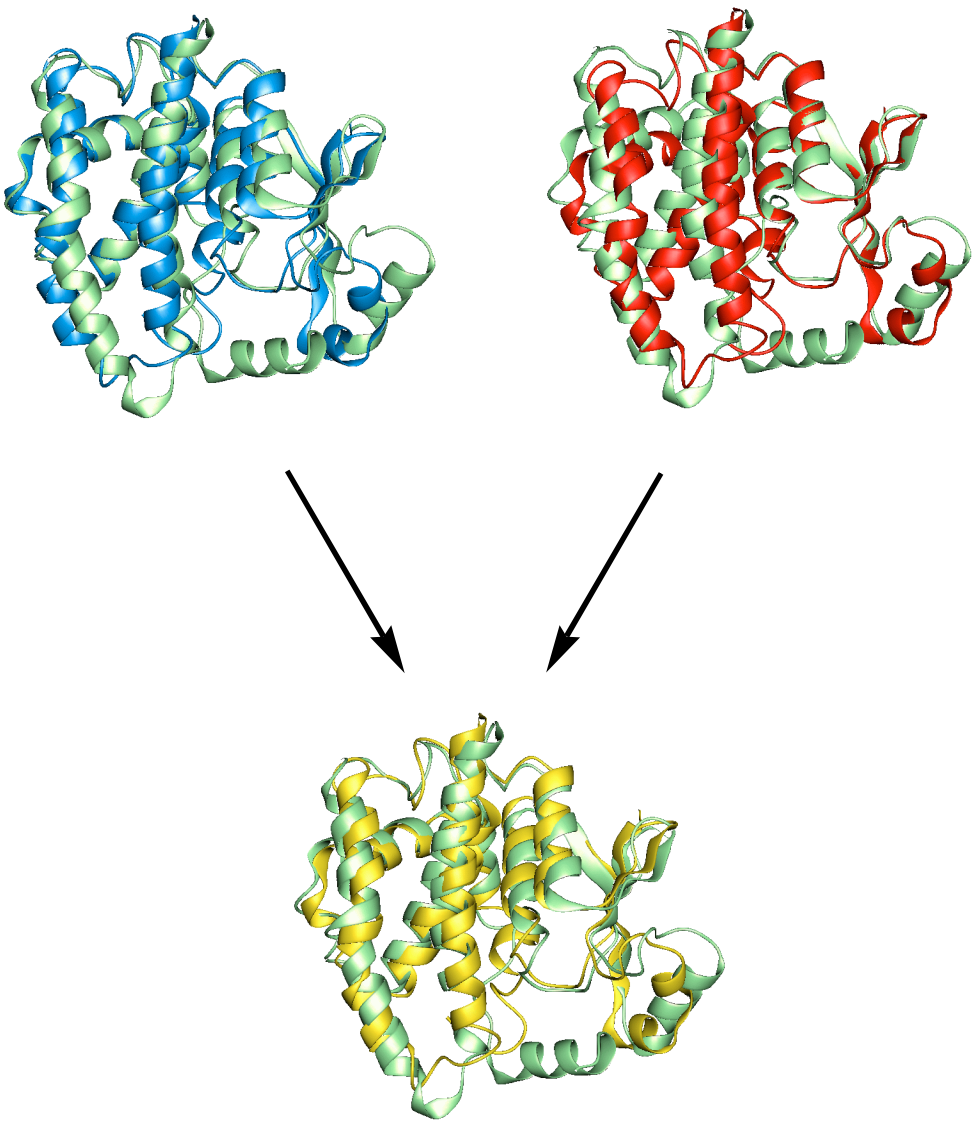


Figure 2 *Convergence of the superpositions (1,G2), top-left, and (2,G2), top-right, to the superposition (1,G5), bottom, for the {1GUH,1GSS} pair of proteins. The reference protein, 1GUH, is always in green, whereas the target protein, 1GSS, is in blue, red, and yellow in the superpositions (1,G2), (2,G2), and (1,G5), respectively*

solution. The increase in smoothness of the Gaussian representation from G1 to G10 leads to a significant reduction in the number of superposition solutions, several of them converging to a single superposition solution at every step of the procedure. Parallel coordinates were used to represent the connectivity between superposition solutions using different Gaussian representations.

Finally, in order to assess the significance of the superposition solutions obtained with GAPS, results were compared to the superposition solutions produced by three other protein-structure similarity programs publicly available from the web, namely, GA_FIT,[15] TOP,[19] and ALIGN[21]. Original default parameters were always used. Due to the stochastic nature of the genetic algorithm used by GA_FIT, 10 runs were always performed. In addition, comparison with structural superpositions extracted from the FSSP[24] database is also included. In all cases, the degree of agreement between the final three-dimensional orientations of the target protein with respect to the reference protein obtained by the different methods was assessed by computing the root mean square deviation (RMSD) of the corresponding C_α carbons for the target protein.

Results and discussion

Three pairs of protein structures were selected as examples to illustrate the degree of difficulty in assessing the optimum superposition in different cases. First, a pair of glutathione S-transferases is taken as an example of members of the same protein family having a rich α-helix topology. Second, flavodoxin and CheY are taken as an example of proteins with mixed α-helix/β-strand topology. And third, bean mottle virus and tumor necrosis factor are taken as examples of proteins of rich β-strand topology.

Table 1 *RMSD values (in Å) between the relative orientations of the target protein, 1GSS(A), with respect to the reference protein, 1GUH(A), obtained from superpositions derived by different approaches*

	GAPS (1,G2)	GAPS (2,G2)	GAPS (1,G5)
GAPS (1,G2)	—		
GAPS (2,G2)	3.6	—	
GAPS (1,G5)	1.5	2.6	—
TOP (1)	0.6	3.2	1.0
TOP (2)	3.0	1.0	1.9
GA_FIT	0.6	3.2	1.0
ALIGN	1.0	2.9	0.6
FSSP	1.1	2.9	0.6

Table 2 *RMSD values (in Å) between the relative orientations of the target protein, 3CHY(A), with respect to the reference protein, 1RCF(A) obtained from superpositions derived by different approaches*

	GAPS (1,G2)	GAPS (2,G2)	GAPS (1,G5)
GAPS (1,G2)	—		
GAPS (2,G2)	3.8	—	
GAPS (1,G5)	2.2	2.8	—
TOP	6.2	5.8	4.6
GA_FIT (1)	1.4	3.6	1.4
GA_FIT (2)	8.1	6.7	6.3
ALIGN	4.7	5.1	4.2
FSSP	2.7	2.2	0.7

Glutathione S-transferase A1-1 (1GUH) and P1-1 (1GSS)

The structure of these two proteins is characterized by two domains covalently connected. Domain I is an α/β structure built up of a mixed β-sheet of four strands together with three α-helices, while domain II is entirely composed by α-helices. The sequence identity between the two proteins is about 30%. Structural superpositions of these two proteins have been reported previously.[36,37] However, the difficulty in deriving relevant structural superpositions for this pair of proteins is manifested by the fact that it was recently proposed to superpose the protein-bound ligands for obtaining the protein-structure superposition instead of using the protein structures themselves.[37]

The spectra of similarity values for the different structural superpositions obtained with each Gaussian-based representation is shown in Figure 1. The connectivity between the superposition solutions at the different Gaussian representations can also be followed. Superpositions leading to the consensus superposition at the G10 protein-structure representation are given in blue. It is important to stress out that the superpositions given in red cannot be strictly considered valid superposition solutions at the sequence level although they are indeed superposition solutions from a pure shape point of view. Most of them correspond to local superpositions of protein substructures. Therefore, for the sake of completeness, they have been also included in Figure 1 to provide an idea of the discrimination power of the actual similarity scoring to retrieve the "correct" structural superposition(s) among the best superposition solutions.

There are two evident results that can be immediately extracted from Figure 1. On one hand, the best structural superposition at the G1 protein-structure representation consistently converges to the best structural superpositions at G2, G5, and G10. On the other hand, the best superposition is clearly discriminated from the lower-ranking solutions obtained within each Gaussian representation. This is an interesting outcome because it reflects the fact that, despite the low sequence identity between these two proteins, they still can clearly identify each other at a structural level. Another aspect to remark from Figure 1 is that the number of superposition solutions is significantly reduced when going from a more local Gaussian-based representation, as G1, to a more global representation, as G10. This is the typical function-smoothing effect, which compiles several close similarity maxima when using a more local protein-structure representation into a single similarity maximum when a more diffuse representation is used. Due to this smoothing effect, multiple low-ranking solutions at G1, G2, and G5 ultimately converge to a unique consensus solution at G10 (in blue in Figure 1).

Although in Figure 1 the best superposition is unique at G10 and clearly discriminated from the other solutions at G1, G2, and G5, an alternative superposition solution is revealed at the more local G1 and G2 representations. This alternative solution finally converges to the best superposition solution at G5 and G10. The convergence of (1,G2) and (2,G2) (the two best superpositions in blue at G2 in Figure 1) to (1,G5) (the best superposition in blue at G5 in Figure 1) is illustrated in Figure 2. As can be observed, the (1,G2) solution superposes domains II of the two proteins, causing a slight misalignment of domains I. In contrast, the (2,G2) solution superposes domains I of the two proteins, resulting in a poorer fit of domains II where α-helices are not well superposed but parallel to each other. The final (1,G5) consensus superposition provides a balance between the two more local alternative superpositions (1,G2) and (2,G2). Thus, with respect to (1,G2), the superposition of domains I in (1,G5) is improved at expenses of a poorer fit of domains II, and vice versa with respect to (2,G2). The RMSD between the relative orientations of the target protein (1GSS) with respect to the reference protein (1GUH) obtained from the (1,G5) superposition shown in Figure 2 and the final (1,G10) superposition is 0.1 Å.

Comparison of the Gaussian-based superpositions presented in Figure 2 with the superpositions obtained by other programs is given in Table 1. In all cases a Gaussian-based superposition is found to be close to a superposition obtained

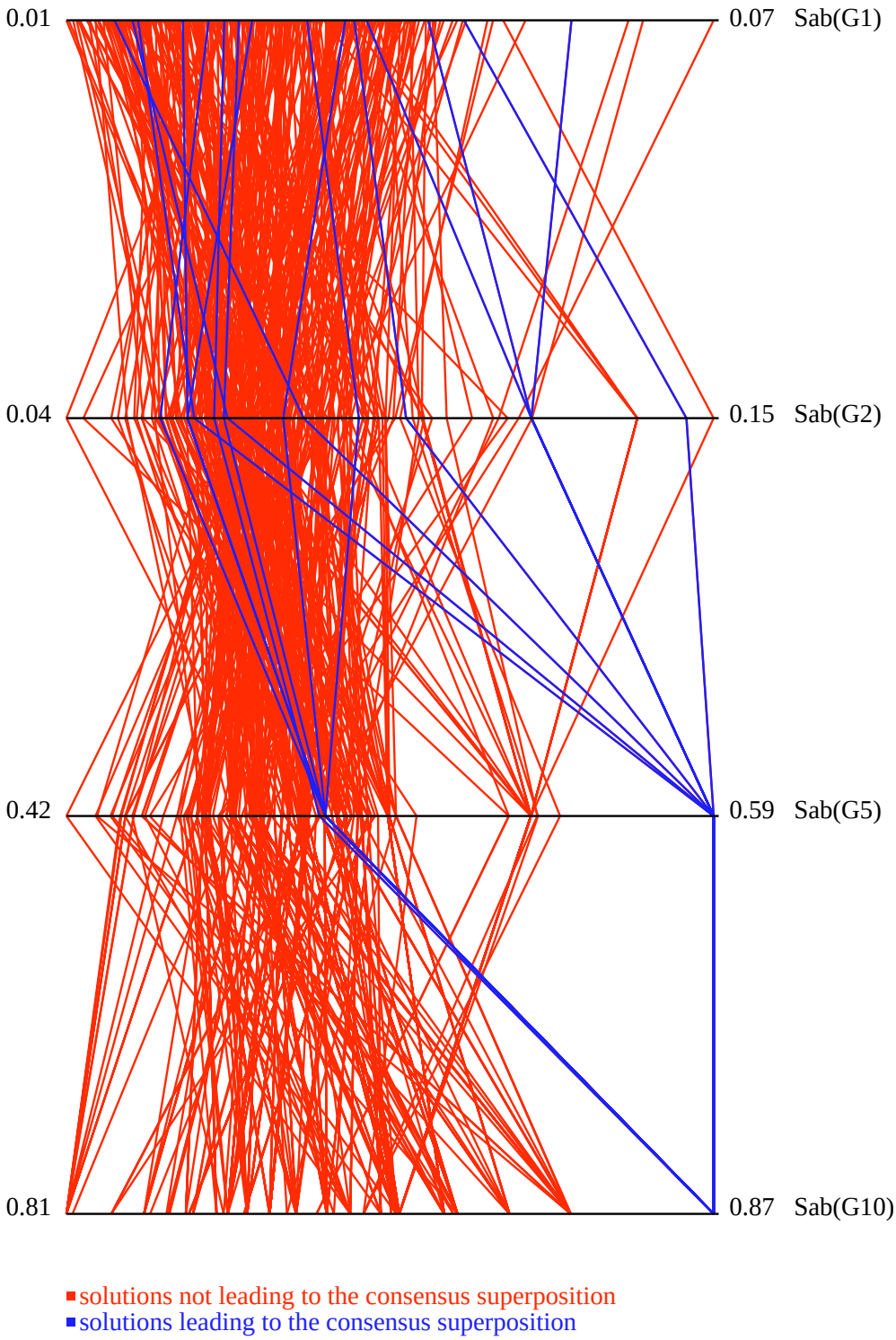


Figure 3 *Connectivity between the superposition solutions obtained from each Gaussian-based representation for the {1RCF,3CHY} pair of proteins (see text). Each parallel coordinate represents the range of similarity values (Sab in eq. (4)) assigned to the superposition solutions obtained from each Gaussian-based representation (G1, G2, G5, and G10)*

by other means. Interestingly, the TOP program identifies two alternative superposition solutions similar to the two local solutions generated by GAPS at the G2 protein-structure representation. In contrast, GA_FIT, ALIGN, and FSSP retrieve one single superposition. All GA_FIT runs converged to a superposition in close agreement with the (1,G2) solution, whereas the ALIGN superposition and the superposition extracted from the FSSP database are closer to the more diffuse (1,G5) solution. In summary, for the {1GUH,1GSS} pair of proteins all programs produce a structural superposition within 1.0 Å RMSD of the proposed consensus optimum superposition by GAPS.

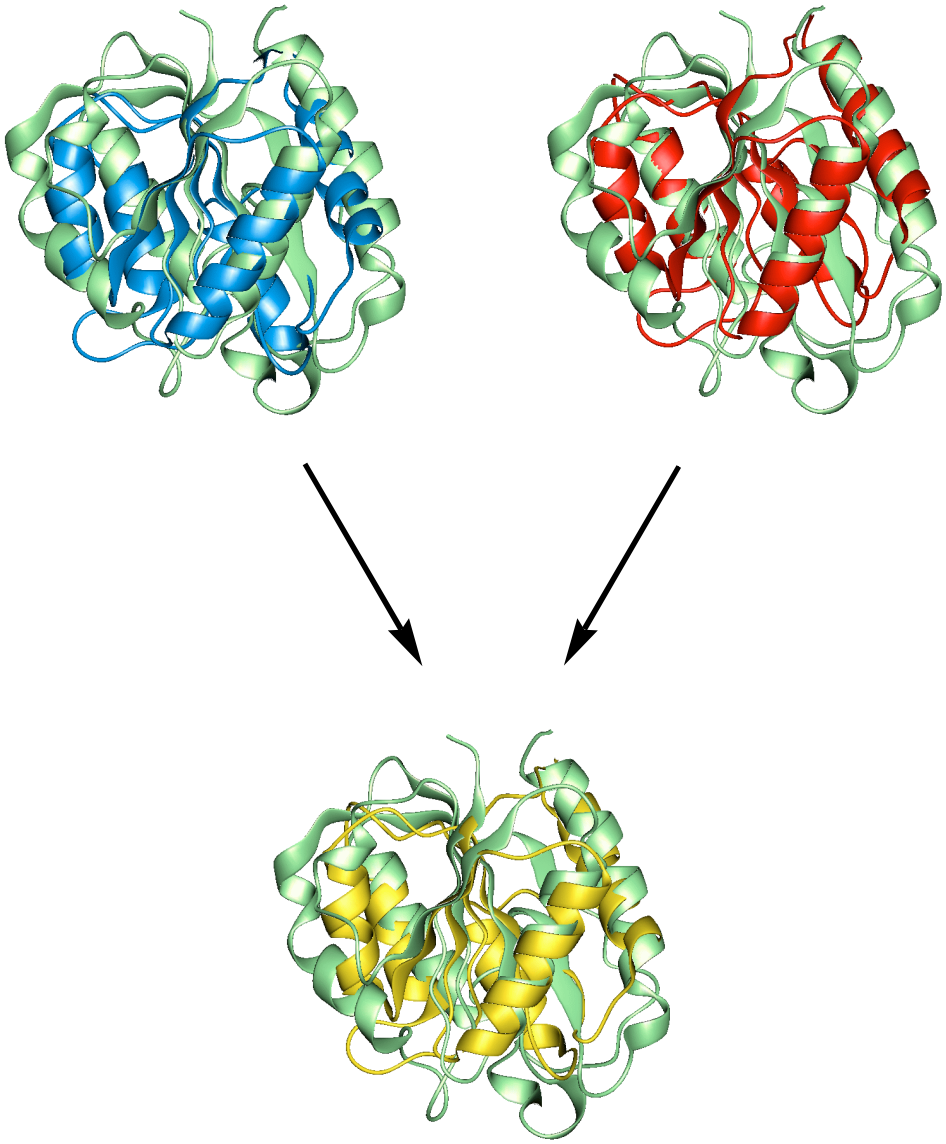


Figure 4 *Convergence of the superpositions (1,G2), top-left, and (2,G2), top-right, to the superposition (1,G5), bottom, for the {1RCF,3CHY} pair of proteins. The reference protein, 1RCF, is always in green, whereas the target protein, 3CHY, is in blue, red, and yellow in the superpositions (1,G2), (2,G2), and (1,G5), respectively*

Flavodoxin (1RCF) and CheY (3CHY)

The overall structure of these two proteins consists of a five-stranded parallel β-sheet core, flanked by five α-helices. Their sequence identity is about 15%, which is practically random, and there is no evidence for their homology. As stated previously,[29] these two proteins probably represent an example of structurally convergent evolution, where the same structural solution was independently reached by two distinct protein families. The ambiguities in deriving a unique structural superposition for this pair of proteins have been already recognized.[30]

For the {1RCF,3CHY} pair of proteins, the spectra of similarity values for the structural superpositions obtained with each Gaussian-based representation is given in Figure 3. Compared with results obtained above for {1GUH,1GSS}, three main differences can be underlined (see Figures 1 and 3). First, in contrast to the behavior observed for the pair {1GUH,1GSS}, the best structural superposition solutions at the G1 and G2 levels do not converge to the best solutions at the G5 and G10 levels for {1RCF,3CHY}. Instead, lower-ranking solutions at the G1 and G2 levels are the ones ultimately leading to the unique consensus superposition at G10 (in blue in Figure 3). This is a clear example of the ability of protein-structure similarities using the more local representations (G1 and G2) to get trapped in local superpositions where some of the dominant secondary-structure elements may be perfectly superposed despite a poor overall fit. When more diffuse protein-structure representations are used (G5 and G10), the global structural superposition is directly identified as the best superposition solution. Second, note that the similarity value of the best solution at each Gaussian-

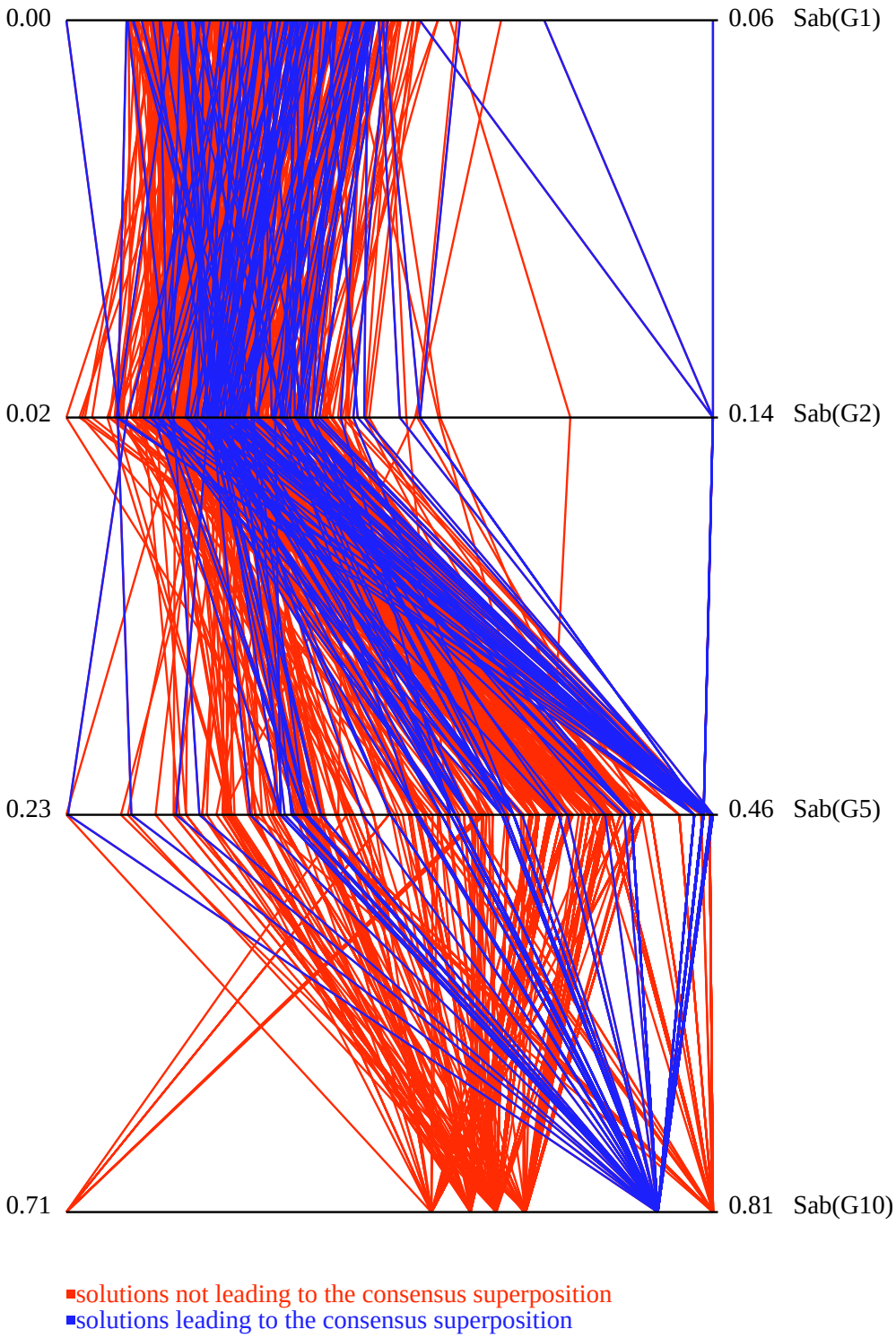


Figure 5 *Connectivity between the superposition solutions obtained from each Gaussian-based representation for the {1BMV,1TNF} pair of proteins (see text). Each parallel coordinate represents the range of similarity values (Sab in eq. (4)) assigned to the superposition solutions obtained from each Gaussian-based representation (G1, G2, G5, and G10)*

based representation for {1RCF,3CHY} is always smaller than for {1GUH,1GSS}. This is an indication of the poorer overall protein-structure similarity of the {1RCF,3CHY} pair of proteins with respect to {1GUH,1GSS}. And third, the discrimination between the best superposition and the rest of low-ranking solutions at the G5 and G10 levels for {1RCF,3CHY} is not as clear as found previously for {1GUH,1GSS}. For instance, at the G10 level, the gap between the similarity score of the best and the second best structural superpositions is 0.159 and 0.014 for the {1GUH,1GSS} and {1RCF,3CHY} pairs of proteins, respectively. This reflects the fact that, from a pure shape point of view, some arrangements of secondary structures in proteins are more discriminative than others.

In order to inspect the nature of some of the structural superpositions obtained for the {1RCF,3CHY} pair of proteins visually, the convergence of (1,G2) and (2,G2) (the two best superpositions in blue at G2 in Figure 3) to (1,G5) (the best superposition in blue at G5 in Figure 3) is illustrated in Figure 4. As can be observed, the (1,G2) solution nicely aligns the β-sheet cores of the two proteins, resulting in a poorer fit of the pairs of α-helices. In contrast, the (2,G2) solution provides a better superposition of the α-helices of the two proteins, distorting the fit of the two β-sheet cores. The final (1,G5) consensus superposition compromises the superposition of the β-sheet cores (overemphasized in (1,G2) at the expense of the fit of the α-helices) with the superposition of the α-helices (overemphasized in (2,G2) at the expense of the fit of the β-cores). The RMSD between the relative orientation of the target protein (3CHY) with respect to the reference protein (1RCF) obtained from the (1,G5) superposition shown in Figure 4 and the final (1,G10) superposition is 0.7 Å.

Comparison of the Gaussian-based superpositions presented in Figure 4 with the superpositions obtained by other programs is given in Table 2. Contrary to the situation found previously for {1GUH,1GSS} where all programs basically agreed with a similar structural superposition, there is much more diversity in the superpositions obtained from different programs when applied to the {1RCF,3CHY} pair of proteins. On the one hand, the GA_FIT program is able to identify two alternative superposition solutions. However, only one of them, GA_FIT (1), can be considered close to one of the superpositions obtained by GAPS. The other superposition solution, GA_FIT (2), is different from any other alignment obtained by the other programs and, thus, it is essentially unique. On the other hand, the superpositions derived with the TOP and ALIGN programs appear to be more similar to each other than to the rest of the superpositions produced by other programs. Finally, a good correspondence is found between the superposition extracted from the FSSP database and the GAPS (1,G5) superposition. In summary, for the

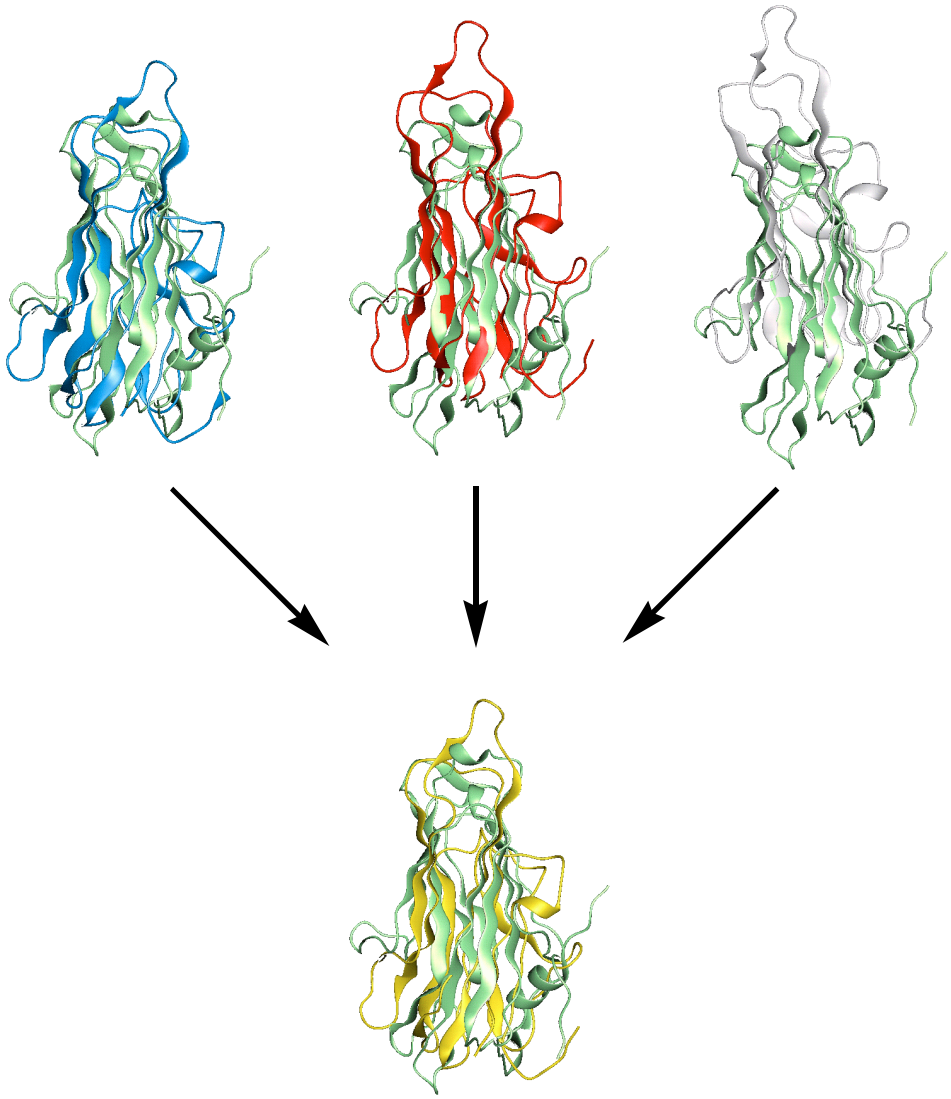


Figure 6 *Convergence of the superpositions (1,G5), top-left, (2,G5), top-center, and (3,G5), top-right, to the superposition (1,G10), bottom, for the {1BMV,1TNF} pair of proteins. The reference protein, 1BMV, is always in green, whereas the target protein, 1TNF, is in blue, red, white, and yellow in the superpositions (1,G5), (2,G5), (3,G5) and (1,G10), respectively*

Table 3 *RMSD values (in Å) between the relative orientations of the target protein, 1TNF(A), with respect to the reference protein, 1BMV(A), obtained from superpositions derived by different approaches.*

	GAPS (1,G5)	GAPS (2,G5)	GAPS (3,G5)	GAPS (1,G10)
GAPS (1,G5)	—			
GAPS (2,G5)	6.2	—		
GAPS (3,G5)	11.9	6.4	—	
GAPS (1,G10)	1.7	6.0	12.0	—
TOP (1)	3.2	3.5	9.6	2.7
TOP (2)	5.6	2.8	7.1	5.8
GA_FIT (1)	3.8	3.4	9.7	3.3
GA_FIT (2)	6.3	0.5	6.5	6.1
GA_FIT (3)	12.4	6.8	0.7	12.5
ALIGN	5.3	1.6	7.5	5.3
FSSP	3.2	3.8	10.0	2.8

{1RCF,3CHY} pair of proteins only the superpositions obtained from GA_FIT and the FSSP database are found within 1.5 Å RMSD of the proposed consensus optimum superposition by GAPS.

Bean mottle virus (1BMV) and tumor encrosis factor (1TNF)

These two proteins have in common a ten-stranded β-sheet topology. It has already been recognized that proteins having this kind of architecture are likely to permit alternative structural superpositions.[30] Their sequence identity is about 10%, which puts them deep into the twilight zone.

The set of similarity values for the structural superpositions obtained with each Gaussian-based representation is given in Figure 5. There are two main aspects to note when comparing results for the {1BMV,1TNF} pair of proteins (Figure 5) with those obtained above for {1GUH,1GSS} (Figure 1) and {1RCF,3CHY} (Figure 3). First, a larger number of local structural superpositions at the G1, G2, and G5 levels of protein-structure representation finally converge to the consensus optimum superposition at G10 (in blue in Figure 5). Second, although the best superposition solutions at the G1, G2, and G5 representations lead to the consensus optimum superposition at G10, the latter is not the best solution found. From a pure shape point of view, it is actually the second best solution at G10. Note also that the similarity values corresponding to the best solution at each Gaussian-based representation are the lowest among the three pairs of proteins studied (compare Figures 1, 3, and 5). Both aspects reveal the poor discriminative power of this kind of structural architecture and confirm previously reported ambiguities in deriving a unique optimum superposition for this pair of proteins.[30]

The convergence of (1,G5), (2,G5), and (3,G5) (the three best superpositions in blue at G5 in Figure 5) to (1,G10) (the best superposition in blue at G10 in Figure 5) is illustrated in Figure 6. In this case, the difference between the alternative superpositions found at the G5 level of representation is not due to the preferential superposition of a type of domain or structural characteristic (see Figures 2 and 4) but to a shift in the protein-structure superposition (*vide infra*). As can be observed, the two proteins are slightly shifted away (by approximately two amino acids) when going from (1,G5), to (2,G5), and to (3,G5). Because GAPS is based on the steric overlap between protein structures (eq. (3)), the more diffuse the Gaussian-based representation used, the less accessible superposition solutions with large non-overlapping regions become. This is the reason why the most shifted superpositions at G5, (2,G5) and (3,G5), collapse together with (1,G5) to a final consensus optimum superposition (1,G10).

Comparison of the Gaussian-based superpositions presented in Figure 6 with the superpositions obtained by other programs is given in Table 3. Note that the three GAPS solutions using a G5 representation have consecutively a RMSD of ca. 6 Å, which reflects the approximate two amino acid shift mentioned above. The GA_FIT program also produces three alternative superpositions with a similar 6 Å RMSD gap between them. Interestingly, each of the GA_FIT superpositions can be essentially associated with one GAPS superposition. The TOP program identifies two alternative superpositions with resemblance to the (1,G10) and (2,G5) superpositions from GAPS. In summary, all superpositions produced by the different programs can be clustered in three general groups: one group composed of the superpositions GAPS (1,G5), GAPS (1,G10), TOP (1), GA_FIT (1), and FSSP; a second group formed by the superpositions GAPS (2,G5), TOP (2), GA_FIT (2), and ALIGN; and a third group containing the GAPS (3,G5) and GA_FIT (3) superpositions.

Conclusions

The ability of a Gaussian-based approach to protein-structure similarity, as implemented in the program GAPS, to identify relevant structural superpositions has been illustrated for three pairs of proteins with different topological characteristics and very low sequence identities. For the sake of validation, the superpositions obtained by GAPS were compared with those produced by other programs (TOP, GA_FIT, and ALIGN) or directly extracted from a database (FSSP). The comparative analysis revealed the resemblance between some of the superpositions generated but also the differences between the alternative superpositions identified by a variety

of methods, thus confirming the ambiguities in defining a unique optimum superposition. The present Gaussian-based methodology offers a means to, depending on the Gaussian-based representation used for evaluating protein-structure similarities, derive a consensus optimum superposition objectively when alternative superposition solutions exist.

Several advantages can be foreseen in using a Gaussian-based approach to protein-structure similarity. (i) The generation of the structural superposition is not biased by the use of sequence or secondary structure information of the proteins. This having been said, such an unbiased approach results in longer computing times. In this particular study, using a G10 representation with a limited 90-degree rotational step (24 starting orientations) computing times for the {1GUH,1GSS}, {1RCF,3CHY}, and {1BMV,1TNF} were 210, 67, and 162 seconds, respectively, in an SGI/R10000. (ii) The degree of locality of the superposition can be tuned depending on the Gaussian-based representation used to evaluate protein-structure similarities. Alternative solutions can be found when using more local Gaussian-based representations but they converge to a consensus superposition solution when more diffuse Gaussian representations are used. (iii) Because of the smoothness of a Gaussian-based protein-structure representation, the structural superpositions produced are neither dependent on the resolution nor on the completeness of the protein structures. And (iv) the methodology is simple and general. Although the present work has focused on pairwise comparisons of rigid protein structures, it can be easily extended to pairwise flexible superpositions as well as to optimizing the mutual superposition of multiple proteins.[32]

Acknowledgments The present work originated from lively discussions with Gerry Maggiora and Doug Rohrer (Computer-Aided Drug Discovery, Pharmacia & Upjohn) on the possibility of using a Gaussian-based approach to protein-structure similarity.

References

1. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F. J.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535.
2. Brenner, S. E.; Chothia, C.; Hubbard, T. J. P. *Curr. Opin. Struct. Biol.* **1997**, *7*, 369.
3. Orengo, C. *Curr. Opin. Struct. Biol.* **1994**, *4*, 429.
4. Overington, J. P. *Curr. Opin. Struct. Biol.* **1992**, *2*, 394.
5. Hilbert, M.; Bohm, G.; Jaenicke, R. *Proteins* **1993**, *17*, 138.
6. Sutcliffe, M. J.; Haneef, I.; Carney, D.; Blundell, T. J. *Prot. Eng.* **1987**, *1*, 377.
7. Taylor, W. R.; Orengo, C. A. *J. Mol. Biol.* **1989**, *208*, 1.
8. Sali, A.; Blundell, T. J. *J. Mol. Biol.* **1990**, *212*, 403.
9. Rose, J.; Eisenmenger, F. *J. Mol. Evol.* **1991**, *32*, 340.
10. Vriend, G.; Sander, C. *Proteins* **1991**, *11*, 52.
11. Alexandrov, N. N.; Takahashi, K.; Go, N. *J. Mol. Biol.* **1992**, *225*, 5.
12. Russell, R. B.; Barton, G. J. *Proteins* **1992**, *14*, 309.
13. Grindley, H. M.; Artymiuk, P. J.; Rice, D. W.; Willett, P. *J. Mol. Biol.* **1993**, *229*, 707.
14. Holm, L.; Sander, C. *J. Mol. Biol.* **1993**, *233*, 123.
15. May, A. C. W.; Johnson, M. S. *Prot. Eng.* **1994**, *7*, 475. GA_FIT: http://www.btk.utu.fi/molmol/programs/
16. Diederichs, K. *Proteins* **1995**, *23*, 187.
17. Falicov, A.; Cohen, F. E. *J. Mol. Biol.* **1996**, *258*, 871.
18. Alexandrov, N. N.; Fischer, D. *Proteins* **1996**, *25*, 354.
19. Lu, G. *PDB Quarterly Newsletter* **1996**, *78*, 10. TOP: http://alfa.mbb.ki.se:8000/TOP/
20. Carugo, O.; Eisenhaber, F. *J. Appl. Cryst.* **1997**, *30*, 547.
21. Gerstein, M.; Levitt, M. *Prot. Sci.* **1998**, *7*, 445. ALIGN: http://bioinfo.mbb.yale.edu/align/
22. Lehtonen, J. V.; Denessiouk, K.; May, A. C. W.; Johnson, M. S. *Proteins* **1999**, *34*, 341.
23. Verbitsky, G.; Nussinov, R.; Wolfson, H. *Proteins* **1999**, *34*, 232.
24. Holm, L.; Ouzounis, C.; Sander, C.; Tuparev, G.; Vriend, G. *Prot. Sci.* **1992**, *1*, 1691. FSSP: http://www2.ebi.ac.uk/dali/fssp
25. Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. *J. Mol. Biol.* **1995**, *247*, 536. SCOP: http://www.bio.cam.ac.uk/scop
26. Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. *Structure* **1997**, *5*, 1093. CATH: http://www.biochem.ucl.ac.uk/bsm/cath
27. Sowdhamini, R.; Burke, D. F.; Huang, J.-F.; Mizuguchi, K.; Nagarajaram, H. A.; Srinivasan, N.; Steward, R. E.; Blundell, T. L. *Structure* **1998**, *6*, 1087. CAMPASS: http://www-cryst.bioc.cam.ac.uk/~campass/
28. Mizuguchi, K.; Deane, C. M.; Blundell, T. L.; Overington, J. P. *Prot. Sci.* **1998**, *7*, 2469. HOMSTRAD: http://www-cryst.bioc.cam.ac.uk/~homstrad/
29. Godzik, A. *Prot. Sci.* **1996**, *5*, 1325.
30. Zu-Khang, F.; Sippl, M. J. *Folding & Design* **1996**, *1*, 123.
31. Mestres, J.; Rohrer, D. C.; Maggiora, G. M. *J. Comput. Chem.* **1997**, *18*, 934.
32. Results presented at the 12th European Symposium on Quantitative Structure-Activity Relationships, Copenhagen, 1998. Mestres, J.; Rohrer, D. C.; Maggiora, G. M. In *Molecular Modeling and Prediction of Bioactivity*, Gundertofte, K. (Ed.); Kluwer: Dordrecht, 2000, p. 83.
33. Oldfield, T. J.; Hubbard, R. E. *Proteins* **1994**, *18*, 324.
34. Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*, Oxford University Press: Oxford, 1990.
35. An arbitrary "vanishing" value of 0.1 was selected in this case. Under this constraint, using a Gaussian representation vanishing at **R**=2 Å, the Gaussian parameters for a C_α carbon atom are α_i=124.5748 and β_i=1.7819.
36. Sinning, I.; Kleywegt, G. J.; Cowan, S. W.; Reinemer, P.; Dirr, H. W.; Huber, R.; Gilliland, G. L.; Armstrong, R. N.; Ji, X.; Board, P. G.; Olin, B.; Mannervik, B.; Jones, T. A. *J. Mol. Biol.* **1993**, *232*, 192.
37. Koehler, R. T.; Villar, H.; Bauer, K. E.; Higgins, D. L. *Proteins* **1997**, *28*, 202.