Evaluation of structural similarity based on reduced dimensionality representations of protein structure

Birgit Albrecht, Guy H. Grant1 and W. Graham Richards

Department of Chemistry, University of Oxford, Central Chemistry Laboratory, South Parks Road, Oxford OX1 3QH, UK

1To whom correspondence should be addressed.
E-mail: guy.grant@chemistry.ox.ac.uk

Protein similarity estimations can be achieved using reduced dimensional representations and we describe a new application for the generation of two-dimensional maps from the three-dimensional structure. The code for the dimensionality reduction is based on the concept of pseudo-random generation of two-dimensional coordinates and Monte Carlo-like acceptance criteria for the generated coordinates. A new method for calculating protein similarity is developed by introducing a distance-dependent similarity field. Similarity of two proteins is derived from similarity field indices between amino acids based on various criteria such as hydrophobicity, residue replacement factors and conformational similarity, each showing a one factor Gaussian dependence. Results on comparisons of misfolded protein models with data sets of correctly folded structures show that discrimination between correctly folded and misfolded structures is possible. Tests were carried out on five different proteins, comparing a misfolded protein structure with members of the same topology, architecture, family and domain according to the CATH classification.

Keywords: dimensionality reduction/molecular similarity/protein comparison/Sammon mapping

Introduction

Similarity measures between small molecules have proved very fruitful. Similarity between proteins has great promise, but it is a much more difficult problem.

Protein modelling and protein bioinformatics are increasingly important disciplines within the pharmaceutical and biotechnology industries, as well as an important area of academic research and commercial software development. However, protein modelling also poses a major challenge. To develop accurate models that account for hundreds of thousands of atoms can be difficult and similarity comparisons with other proteins take considerable computational power and CPU time.

Protein similarity has emerged as an area of interest. Large amounts of detailed structural information have become widely available thanks to advances in crystallographic and spectroscopic techniques, while the function of the identified structure, classification of the proteins or even just the accuracy of the derived coordinates is often unknown. Protein similarity could be used as a tool in all of these problems and can provide additional information on a protein based on its structure.

Several approaches have been taken over the last few years to determine similarities between a large number of protein structures (Holm and Sander, 1996). Maggiora et al. (2001) used spherical Gaussian functions located on single atoms to evaluate the optimal alignment of steric fields of proteins. This allows for a flexible description of the underlying fold geometry, where the focus can be shifted from an atom-like description to a better representation of the general shape of a molecule by adjusting the width of the Gaussian function. Bostick and Vaisman (2003) applied Delaunay-based topological maps to generate three-dimensional arrays representing the global structural topology, which are then used together with an integral scoring scheme to determine pairwise protein similarity. Carugo and Pongor (2002) consider Cα positions as descriptors for protein fold similarity. Large data sets of proteins are compared using distance histograms and the root mean square distances between Cα to evaluate similarity. However, this approach does not consider the diverse properties of different amino acids. Although all these methods have been successful to a certain extent, protein similarity calculations are still not routinely used in structural genomics. As yet, no tool has been developed that considers structural features as well as topological properties and characteristics. The most obvious potential difficulty is the complexity of the problem and hence the computation time needed to obtain results.

Reduced dimensionality representations attempt to simplify structural problems and thus allow enormous increases in speed. Previous work on small molecules by Allen et al. (2001) used Sammon maps to generate two-dimensional maps for three-dimensional structures. It was also concluded that similarity comparisons in three dimensions scale exponentially with the number of structures in the data set, but if two-dimensional representations are used the time factor increases linearly.

Similarity calculations involve finding the optimal alignment of the two molecules to be compared and the subsequent calculation of their overlap or the overlap of their property X. Common expressions for similarity between two molecules are the Tanimoto similarity coefficient, the Carbo Index, the Hodgkin Index and Euclidean distances. The Tanimoto coefficient is a purely two-dimensional measure, which is calculated with structural keys and depends on the presence or absence of specific structural features in the structures being compared (Downs and Willett, 1995):

\[
T_{\text{Tanimoto}}_{AB} = \frac{\sum_{i=1}^{N} X_{A}X_{B}}{\sum_{i=1}^{N} (X_{A})^2 + \sum_{i=1}^{N} (X_{B})^2 - \sum_{i=1}^{N} X_{A}X_{B}}
\]

The Carbo Index (Carbo and Arnau, 1980) is probably the most common similarity descriptor and uses the overlap of the electron densities of the two molecules:

\[
\text{Carbo}_{AB} = \frac{\sum_{i=1}^{N} X_{A}X_{B}}{\sqrt{\sum_{i=1}^{N} (X_{A})^2 + \sum_{i=1}^{N} (X_{B})^2}}
\]
The Hodgkin index is based on the same principle as the Carbo Index with the difference that it accounts not only for the signs of the property such as the electrostatic potentials, but also for their magnitudes (Hodgkin and Richards, 1987):

$$Hodgkin_{AB} = \frac{2 \sum_{i=1}^{N} X_{Ai}X_{Bj}}{\sum_{i=1}^{N}(X_{Ai})^2 \sum_{j=1}^{N}(X_{Bj})^2}$$

Euclidean distances in turn are based solely on the position of atoms in space:

$$Euclidean_{AB} = \sqrt{\sum_{i=1}^{N} (X_{Ai} - X_{Bj})^2}$$

Here we report the development of a new mapping technique to produce two-dimensional maps of protein structures and a new method of calculating similarity between these protein maps.

**Materials and methods**

Protein similarity calculations are a computationally very expensive task. For small molecules it has been demonstrated that transferring a structure into a two-dimensional representation speeds up similarity based superposition. We have applied this approach to proteins and used two-dimensional representations of protein structures to determine their similarity. Two-dimensional maps are generated using a novel algorithm and are related to the original three-dimensional structure via their respective distance matrices. Similarity between two proteins is then calculated introducing the concept of a distance-dependent similarity field. With these tools it is possible to distinguish between misfolded and correct structure at the level of protein families.

**Mapping**

The generation of the two-dimensional protein maps is based on the optimization of distance matrices of two-dimensional coordinates via a Monte Carlo-like technique so that the atom-atom distances in the two-dimensional map are as close as possible to those in the experimental three-dimensional structure. As in the original Monte Carlo approach, this algorithm samples the two-dimensional space by generating random coordinates. Inter-residue distances are calculated and serve as acceptance criteria which determine if the generated coordinates will be accepted or rejected. A flowchart of the program structure is given in Figures 1 and 2 and shows an example of the structural representations during the different steps.

Initial starting coordinates are created by the projection of the original three-dimensional coordinates into a plane along the z-axis. These two-dimensional coordinates are then improved by minimising the differences between the distance matrices of the two-dimensional Cα positions and the initial coordinates in three-dimensional space. This compares all inter-residue distances in the two-dimensional map with the distances in the three-dimensional map. An error value is then calculated as the difference between the distance squares in the two-dimensional map and the original three-dimensional coordinates. This error is used as the seed for the first step of the pseudo-random coordinate generation to generate new Cα positions using the Rand function in C.

Again all distances between all points in two dimensions are calculated and compared to the corresponding three-dimensional distances. The error is calculated as the difference between each corresponding set of inter-residue distances in two and three dimensions. Then each distance error is compared to the error cut-off value. If the errors remain below the cut-off value the coordinates are accepted and the map file is written. If the errors exceed the cut-off value the number of previous coordinate generation cycles is compared to a maximum number of steps. If the maximum number of steps has been reached the map file is written, otherwise a new set of coordinates for every Cα position of the protein is generated using the Rand function. As in the original Monte Carlo procedure there is no temperature dependence, as the generated two-dimensional maps are only a model with no physical meaning and therefore are not dependent on a Boltzmann distribution. Acceptance of coordinates depends only on the difference of atom–atom distances in two and three dimensions.

**Similarity**

In this application we established a new method for calculating protein similarity by introducing a distance-dependent similarity
field. Similarity of two amino acids p and q can thus be calculated as the product of a similarity index dependent on the nature of the amino acids and a Gaussian representation of the distance between them:

\[ \text{amino\_acid\_similarity\_field} = \text{sim}_{pq} \times G_{pq} \]

with \( G = \) single Gaussian representation for distance dependence; \( \text{sim}_{pq} = \) similarity index of the pair of amino acids to be compared.

The similarity field of two proteins is then the sum of the amino acid similarity fields:

\[ \text{similarity\_field} = \sum_{p=1}^{l} \sum_{q=1}^{m} \text{sim}_{pq} \times G_{pq} \]

with \( G = \) single Gaussian representation for distance dependence; \( \text{sim}_{pq} = \) similarity index of pair of the amino acids to be compared; \( p = \) index of amino acid in protein 1; \( l = \) number of amino acids in protein 1; \( q = \) index of amino acid in protein 2; \( m = \) number of amino acids in protein 2.

To calculate the overall protein similarity a normalization factor that accounts for the self-similarities of the structures needs to be applied. This factor has the same form as the similarity field, the only difference being that similarities are calculated between the protein and itself:

\[ \text{self\_similarity} = \sqrt{\left( \sum_{p=1}^{l} \sum_{q=1}^{m} \text{sim}_{pp} \times G_{pp} \right) \times \left( \sum_{q=1}^{m} \sum_{q=1}^{m} \text{sim}_{qq} \times G_{qq} \right) / \left( \sum_{p=1}^{l} \sum_{q=1}^{m} \text{sim}_{pq} \times G_{pq} \right)} \]

with \( G = \) single Gaussian representation for distance dependence; \( \text{sim}_{pq} = \) similarity index of the pair of amino acids to be compared; \( p = \) index of amino acid in protein 1; \( l = \) number of amino acids in protein 1; \( q = \) index of amino acid in protein 2; \( m = \) number of amino acids in protein 2.

Hence, the formula to calculate the similarity between two protein structures is the sum of all amino acid similarities divided by a normalization factor:

\[ \text{protein\_similarity} = \frac{\text{similarity\_field}}{\text{self\_similarity}} = \frac{\sum_{p=1}^{l} \sum_{q=1}^{m} \text{sim}_{pq} \times G_{pq}}{\sqrt{\left( \sum_{p=1}^{l} \sum_{q=1}^{m} \text{sim}_{pp} \times G_{pp} \right) \times \left( \sum_{q=1}^{m} \sum_{q=1}^{m} \text{sim}_{qq} \times G_{qq} \right) / \left( \sum_{p=1}^{l} \sum_{q=1}^{m} \text{sim}_{pq} \times G_{pq} \right)}} \]

The similarity indices used were based on both experimentally and theoretically calculated properties: hydrophobicity (George et al., 1990; Riek et al., 1995); residue replacement tables (Cserzo et al., 1994); conformational similarity weight (Kolaskar and Kulkarni, 1992); two different structure-derived correlations matrices (Niefind and Schomburg, 1991), and two tables of cross correlation coefficients of preference factors for the amino acid main chain and side chain (Ou et al., 1993).

To calculate the similarity of two proteins, the two-dimensional maps of the proteins are first aligned by rotating and translating one map with respect to the other. In three dimensions this requires six degrees of freedom: the two-dimensional representation reduces the freedom to translate in the x and y plane and allows only one rotation, which offers considerable increases in the speed of the alignment. The similarity of the two proteins is then calculated by summing the single amino acid similarities using the equation and similarity indices described above.

**Identification of decoy structures**

Identification of decoy protein structures was used to test the above tools, with recognition of misfolded structures based on structural similarity to proteins of the same domain or family as classified by CATH (Orengo et al., 1997). CATH stands for class, architecture, topology, homologous superfamily: these being different structural categories that are used to classify proteins. In total there are four different classes based on the secondary structure composition and packing within the structure. The 37 different architectures then describe the overall shape of a protein. Connectivity is taken into account on the topology level, which ranks proteins according to their fold families. Homologous superfamilies group proteins with an alleged common ancestor, and hence are assumed to be homologues. A further subdivision of homologues leads to protein domains depending on sequence identity, indicating not only a high structural similarity, but also a high similarity in function.

Misfolded structures were obtained from the Decoys ‘R’ Us Database (V1.1) (http://dd.stanford.edu/). They were generated by superposing an amino acid sequence onto a known incorrect protein construct and minimizing the structure (Holm and Sander, 1992). Twenty data sets based on five decoy protein structures including the misfolded structure as well as correct models for the proteins of the same architecture, topology, homologues superfamily and domain were prepared. Two-dimensional maps for all proteins were generated and similarity was calculated between all proteins within a data set. The decoys used were misfolded structures of 1FDX, 1HIP, 2PAZ, 1P2P and 1LH1. Three-dimensional models for the correct and misfolded structures of these proteins are compared in Figure 3, which shows an example for a misfolded model with the data set to which it was compared.

1FDX is an electron transporting ferredoxin that adopts a mixed alpha beta structure forming a two layer sandwich. 1HIP is an oxidized chromatium high potential iron protein with an unclassified secondary structure and irregular architecture that belongs to the family of high-potential iron–sulfur proteins. 1P2P is a pancreatic phospholipase composed of mainly alpha up and down bundles. 2PAZ is an oxidized native pseudo-azurin with a mostly beta sandwich structure and belongs to the immunoglobulin-like family. 1LH1 is an oxygen transporting leghemoglobin with a mostly alpha helical secondary structure that takes the shape of an orthogonal bundle.
Results and discussion

The two major aims during the development of the code for the dimensionality reduction were computational speed and the reproducibility of results. For the latter a pseudo-random number generation was implemented, which will regenerate the same coordinates for identical proteins under identical conditions. The two-dimensional maps for one protein, generated in different runs but based on the same input, overlap perfectly.

In a further test of the validity of the mapping, the ensemble of structures from an NMR structure determination of 1A03 was downloaded from the Protein Data Bank and maps generated for all structures. Figure 4 shows the overlay of the generated maps for 1A03. There are 20 different structures in the NMR ensemble: all were reduced to identical two-dimensional maps. The same experiments were performed with NMR ensembles for 1A3P, 1A13 and 1A24, in each case producing identical results for all ensemble structures.

The most common method for reduced dimensionality representations to date is Sammon mapping. It is a non-linear mapping technique that plots a set of input points on a plane while trying to preserve the relative distance between the input points. Sammon maps are based on neural networks and hence the resulting maps will strongly depend on the training of the network. Sammon maps have recently been used by Allen et al. (2001) to calculate two-dimensional maps of small molecules. We tried to apply this approach on proteins, but did not manage to produce consistent maps. Thus, we have developed a new Monte Carlo map, which operates on the same principle of preserving the distance matrix, but is independent of training. Even though both techniques rely on the optimization of distance matrices the resulting maps look quite different as shown in Figure 5. The Monte Carlo-like procedure samples a wider spectrum of property space and hence offers an increased chance of finding a global minimum for the final map.

The feature that is common between the initial three-dimensional map and the newly generated two-dimensional maps—for Sammon mapping just as well as Monte Carlo mapping—is the distance matrix. This can be exploited to gain a better understanding of the quality of the maps and to allow for comparisons between the different representations. Figure 6 plots the distances in three dimensions versus distances in two dimensions for both Sammon and Monte Carlo maps. Both maps show a good correlation to the initial three-dimensional distances, with the Monte Carlo map obtaining slightly better results in the mid-ranged distances.

With this new technique, the time required to generate a two-dimensional map depends only on the size of the protein and the acceptance criteria. With identical acceptance criteria and proteins of similar size the computation time needed rises exponentially with the size of the protein as shown in Figure 7.

The results of a similarity comparison of randomly chosen proteins, shown in Figure 8, showed that the similarity of a
protein with itself was correctly predicted at 100% (protein similarity = 1). Also, a wide range of different similarity values was achieved for a set of random proteins, which indicates that all, or at least a vast majority, of similarity space gets sampled with this new proposed method. The lowest similarity value obtained was 0%, the average similarity of the data set was 29% and the median 26% similarity. Excluding self-similarity, the highest similarity within this data set was 99%.

Similarity calculations based on CATH classification of proteins showed that similarity for a more specific classification such as domains was on average higher than less specific classifications like architectures. Results for 1LH1 are shown in Figure 9 and illustrate that there is still the possibility of high similarities even in relatively loose classifications as at the architecture level some proteins might still be very similar. Comparing both levels we see that there are a number of proteins that show only low similarity at the level of architecture, whereas in domains the majority of proteins have considerable similarity with each other.

Figure 10 shows that the average similarity of all correct structures increases with a more accurate classification from architectures, through topologies, homologous families and reaches a maximum for proteins from the same domain. The average deviations, which indicate the difference between the actual results and the average values, are very similar for all four different levels of classification. The results show that at the level of architecture the calculated value for similarity falls within the average deviation range of the correct structures, and hence differentiation between correct and incorrect structures is not possible. At the topology level it is already possible to make a distinction, and for families and domains the differences in similarity between correct and misfolded structures are reliable enough to distinguish decoy models.

Even though there was good differentiation between overall and average similarity values between the different levels of classification for one protein, it is difficult to extrapolate to proteins of different classifications. Figure 11 compares the similarity tables of 1HIP and 1LH1, both at the architecture level. Again different ranges of similarities are observed, with occasional highly similar protein pairs being seen. However, 1HIP shows more similar protein structure on the architecture level compared to 1LH1, and thus also a higher average similarity.

Similarity calculations for sets of proteins that included one misfolded structure showed that the use of different similarity indices has a great influence on the magnitude of the similarity results, as well as the level of discrimination. Indices based on cross correlation factors (Figure 12) and conformational similarity (Figure 13) generally produce a more representative value for similarity between the proteins and were better at discriminating between correctly and incorrectly folded structures. Residue replacement tables (Figure 14), which are often used in mutagenesis studies, produced only a slight differentiation between the decoy structures and correct structures. As all similarities produced with this coefficient fell within a very narrow range of values, definite discrimination between correctly and misfolded structures is difficult. Hydrophobicity scoring tables (Figure 15) showed similar effects. Although correct structures do have higher similarities, the difference in similarity values between correctly and incorrectly folded structures is small.
Conclusion

Structural alignment and similarity comparisons of two proteins are difficult and computationally intensive tasks, to the extent that three-dimensional screening of individual structures against libraries containing hundreds of protein folds is currently not feasible. By reducing the representation of the fold database to two dimensions, this task can be speeded up enormously. The provision of such a tool would be of enormous benefit to the structural proteomics community.

A new method for the generation of two-dimensional maps has been developed. Two-dimensional coordinates are based on pseudo-random numbers and optimized using a Monte Carlo approach. As this approach is not truly random, the resulting maps are reproducible. Hence, running an experiment twice under identical conditions will produce identical results. Map generation is fast and depends only on the size of the protein. As this method is based on a Monte Carlo approach it explores a greater proportion of property space than minimization-based methods, such as Sammon mapping, which have been used previously. Hence, the chances of finding the global minimum for the two-dimensional structure instead of a local minimum are greatly increased.

A similarity field has been introduced and applied to calculate the similarity between two proteins based on their two-dimensional maps. A similarity field is a distance-dependent field, which is calculated by a similarity index and a Gaussian function. The similarity index can be based on different properties and hence similarity can be calculated placing the focus on specific molecular characteristics. Properties that have been considered include cross correlation factors, conformational similarity weights, residue replacement tables, and hydrophobicity scoring. The magnitude of the similarity values obtained depends strongly on the similarity index used, which indicates that some of these indices still have to be optimized to give true representative similarity values. New optimized similarity indices could either be derived by an exact calculation of amino acid similarities or by combining the existing tables using neural networks. In an analogous way to methods used for choosing molecular descriptors, a neural net could be trained to find an optimum combination of the similarity indices used. Considering different levels of classification, differences in average similarity can be observed. Although it is not possible to give typical average similarities for different levels of classification, the distinction between different levels of one family is feasible. Unfortunately residue replacement tables and
Fig. 11. Comparison of proteins from the HIP (left) and LH1 architecture classification. The axes represent the list of different proteins. A higher correlation within the HIP architecture is observed, with more similar protein pairs and higher maximum similarity values (excluding self similarity). The similarity index used was conformational similarity weight. Proteins in the LH1 architecture are the misfolded versions of HIP and structurally correct versions of A36, AF7, AHU, BM, CBM, CK, DUB, HR, HY, E, UD, LU, EA, MPG, TD and NS. Proteins in the HIP architecture are the misfolded versions of HIP and structurally correct version of AAF, BA3, B5, CKM, FLE, FRE, GZ, IA, QQP, TV, EC, OC, SP, R04 and MT2.

Fig. 12. Similarity table for LH1 using cross correlation tables for the main chain as similarity index. Similarities of the LH1 decoy structure to correctly folded members of the LH1 family are shown in grey. Similarities of a correct structure are shown in black. Proteins in the LH1 family are the misfolded version of LH1 and structurally correct versions of ASH, AB, DL, CE, E, W6, FL, GDI, H97, HBG, HM, IT, LS, MA, SC and VB.

Fig. 13. Similarity table for LH1 using conformational similarity weight as similarity index. Similarities of the LH1 decoy structure to correctly folded members of the LH1 family are shown in grey. Similarities of a correct structure are shown in black. Proteins in the LH1 family are the misfolded version of LH1 and structurally correct versions of ASH, AB, DL, CE, E, W6, FL, GDI, H97, HBG, HM, IT, LS, MA, SC and VB.

Fig. 14. Similarity table for LH1 using residue replacement tables as similarity index. Similarities of the LH1 decoy structure to correctly folded members of the LH1 family are shown in grey. Similarities of a correct structure are shown in black. Proteins in the LH1 family are the misfolded version of LH1 and structurally correct versions of ASH, AB, DL, CE, E, W6, FL, GDI, H97, HBG, HM, IT, LS, MA, SC and VB.

Fig. 15. Similarity table for LH1 using hydrophobicity scoring tables as similarity index. Similarities of the LH1 decoy structure to correctly folded members of the LH1 family are shown in grey. Similarities of a correct structure are shown in black. Proteins in the LH1 family are the misfolded version of LH1 and structurally correct versions of ASH, AB, DL, CE, E, W6, FL, GDI, H97, HBG, HM, IT, LS, MA, SC and VB.
hydrophobicity scoring tables give only poor similarity results. Although correctly folded structures consistently score higher similarity values within a data set than misfolded structures, the differences are not large. The magnitudes of the similarity values are consistently very low, even for relatively similar structures, which indicates that those similarity indices need further optimization. This could also improve the discrimination between correct and incorrect structures. Other similarity indices, namely cross correlation factors for the amino acid main chain and side chain as well as a conformational similarity index, show good discrimination between decoy structures and correct structures. Thus it can be concluded that similarity calculations using two-dimensional maps of protein structures have real potential for the identification of misfolded structures within a set of protein structures.

Acknowledgements
This work is supported by the National Foundation for Cancer Research and a Royal Society equipment grant.

References

Received February 24, 2004; revised May 17, 2004; accepted May 21, 2004

Edited by Valerie Daggett