Automated clustering of ensembles of alternative models in protein structure databases

Francisco S. Domingues¹, Jörg Rahnenführer and Thomas Lengauer
Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany
¹To whom correspondence should be addressed. E-mail: doming@mpi-sb.mpg.de

Experimentally determined protein structures have been classified in different public databases according to their structural and evolutionary relationships. Frequently, alternative structural models, determined using X-ray crystallography or NMR spectroscopy, are available for a protein. These models can present significant structural dissimilarity. Currently there is no classification available for these alternative structures. In order to classify them, we developed STRuster, an automated method for clustering ensembles of structural models according to their backbone structure. The method is based on the calculation of carbon alpha (Co) distance matrices. Two filters are applied in the calculation of the dissimilarity measure in order to identify both large and small (but significant) backbone conformational changes. The resulting dissimilarity value is used for hierarchical clustering and partitioning around medoids (PAM). Hierarchical clustering reflects the hierarchy of similarities between all pairs of models, while PAM groups the models into the ‘optimal’ number of clusters. The method has been applied to cluster the structures in each SCOP species level and can be easily applied to any other sets of conformers. The results are available at: http://bioinf.mpi-sb.mpg.de/projects/struster/

Keywords: alternative models/clustering/distance matrix/hierarchical clustering/structure classification

Introduction

Different databases are currently available for the classification of proteins according to their structural and evolutionary relationships. In particular, SCOP (Murzin et al., 1995), CATH ( Orengo et al., 1997) and Dali (Holm and Sander, 1996) provide convenient access to the known structures deposited in the Protein Data Bank (PDB) (Berman et al., 2000) and facilitate the understanding of the structural and evolutionary relationships between proteins. In addition, they provide a definition of protein domains based on structural or evolutionary relationship. The SCOP and CATH databases provide a hierarchical classification of the domains in different levels, where the lower levels correspond to a structural domain from a protein of a certain organism (SCOP level Species) or from a set of domains with identical sequence (CATH level Identical). In the Dali database, the protein structures are grouped according to sequence similarity. For each group a representative is selected and used to perform an all-against-all 3D structure comparison.

It is now common that alternative experimental models of the same protein are available in the PDB. These models can display considerable structural differences as, in general, they correspond to different crystal forms, different physical-chemical conditions, different mutants and proteins forming different complexes or binding to different ligands. Currently no database is available that classifies these alternative models according to their structure relationships. In the current SCOP release (1.65, December 2003), 75% of sets at the species level have at least two entries (two or more alternative structural models for a protein domain) and on average there are 6.7 entries per species level. The problem is becoming more relevant as the average number of alternative structures for a given domain is expected to increase.

In recent years, clustering techniques have been applied to classify protein structures. Clustering has been used to automate the classification of proteins in different folds and families or in the analysis of the trajectories from molecular dynamics simulations; see, for example, May (1999), Laboulais et al. (2002) and Choi et al. (2004). The method of Carugo and Pongor (2002), in particular, is based on the comparison of the distributions of Cα coordinates between two proteins to estimate the structural similarity of different proteins in an efficient way. Clustering techniques have also been applied to determine representatives for ensembles of NMR-derived structures (Kelley et al., 1996). The all-atom root-mean-square deviation (rmsd) is used as a dissimilarity measure for hierarchical clustering. The models are grouped into different clusters and a representative for each cluster is given (model closest to the centroid of each cluster). The OLDERADO database (Kelley and Sutcliffe, 1997) provides the results for the NMR ensembles available in the PDB. So far, these methods have not been applied to clustering alternative structure models (determined by protein crystallography or NMR spectroscopy) available for each protein domain available in the PDB.

Here we propose STRuster, a method for clustering alternative structural models corresponding to different structure determination experiments. The structures are classified according to backbone structure similarity using Cα distance matrices. The dissimilarity measure used for clustering is based on the Euclidean distance for each pair of Cα coordinates. Filters are applied in order to render the method sensitive to a wide range of backbone conformational changes. The method has been applied to each SCOP species level and the results are available online. These results can be useful for guiding further structure determination experiments, in the design and interpretation of mutational experiments, in the selection of models for docking or in the selection of templates for structure prediction. More generally, the results can be useful in the selection of a non-redundant set of protein structures.
Materials and methods

Structure data set

The SCOP classification database, release 1.65 (December 2003), is used. Analysis is restricted to the first seven classes, the true classes. Therefore, we excluded coiled-coil proteins, low-resolution protein structures, peptides and designed proteins, which are not true classes. Of 7705 sets corresponding to the SCOP species level, 4187 with at least three domain structural models (entries) are clustered. For each set, entries are compared with the one with the largest number of residues. Only the entries with at least 80% of the number of residues and at least 90% sequence identity to the largest entry are considered. The program ALIGN (Myers and Miller, 1989) is used to align the sequences and to determine the equivalent residues. The final number of species sets to be clustered is 3716, comprising a total of 36 531 entries. The ASTRAL SCOP 1.65 PDB-style files (Chandonia et al., 2004) are used as the source of the coordinates for each SCOP entry.

Dissimilarity measure

For each SCOP species set, the Cα distance matrices are calculated for all entries. Consider the Cα coordinates for residue i, (xi, yi, zi). The Euclidean distance between the Cα atoms of residues i and j in entry a is

\[ D_{ij}(a) = \sqrt{(x_i-x_j)^2 + (y_i-y_j)^2 + (z_i-z_j)^2}. \]

In order to reduce the influence of differences in large distances associated with extensive conformational changes, a first filter is applied with cut-off F1, resulting in \( D'_{ij}(a) \):

\[ D'_{ij}(a) = \begin{cases} D_{ij}(a), & D_{ij}(a) \leq F_1 \\ F_1, & D_{ij}(a) > F_1 \end{cases} \]

For each pair of entries a, b the absolute difference is then calculated for each residue pair: \( \Delta_{ij}(a,b) = |D'_{ij}(a) - D'_{ij}(b)| \). Only residues that can be aligned to the largest entry in the set (used as reference) are considered. If one of the entries includes an insertion, the corresponding residues in the insertion are not considered. A second filter is then applied with cut-off \( F_2 \) in order to restrict the analysis to ‘significant’ structural differences:

\[ \Delta'_{ij}(a,b) = \begin{cases} 0, & \Delta_{ij}(a,b) \leq F_2 \\ 1, & \Delta_{ij}(a,b) > F_2 \end{cases} \]

In this study, we set \( F_1 = 14.0 \) and \( F_2 = 1.0 \). The matrix \( M(a,b) \) contains the dissimilarity values of all pairs involving the N entries in the set, where \( M(a,b) \) corresponds to the dissimilarity between entries a and b with L aligned residues:

\[ M(a,b) = \sum_{i=1}^{L} \sum_{j=1}^{L} \Delta'_{ij}(a,b) \]

Clustering

The R programming environment for data analysis (version 1.8.1) is used for clustering (http://www.r-project.org/). The dissimilarity matrix \( M \) is used as input for two alternative clustering methods. The first is a hierarchical method, using group average agglomeration (Gordon, 1999). Dendrograms are generated for visualization of the hierarchical dependencies in the data. The second method, partitioning around medoids (PAM), is applied in order to obtain the optimal number of clusters where the entries are grouped in a robust way (Kaufman and Rousseeuw, 1990; Struyf et al., 1997).

PAM is a partitioning algorithm and can be regarded as a generalization of K-means clustering to arbitrary dissimilarity matrices. The goal is to minimize the objective function \( \sum_{i=1}^{N} \min_{j=1,...,k} M(a,m_j) \), where the sum is taken over all entries \( a_1,...,a_N \) in the protein set and \( m_1,...,m_k \) are appropriately chosen representatives (medoids) from the set. The algorithm consists of two steps. In the BUILD step, k initial medoids are sequentially selected. In the SWAP step, the objective function is minimized iteratively by replacing one medoid with another entry. This step is repeated until convergence.

The silhouette width value is a measure of cluster validity (Rousseeuw, 1987) and is used to select the best number of clusters obtained with the PAM algorithm. Assume that we have a clustering of \( N \) protein entries into \( k \) clusters, such that an entry \( a \) belongs to cluster \( C \) of size \( r \). The average dissimilarity between \( a \) and all other entries in cluster \( C \) is

\[ c(a) = \frac{1}{r-1} \sum_{b \in C \setminus a} M(a,b) \]

The average dissimilarity of \( a \) to all entries \( b \) that belong to another cluster \( U \neq C \) of size \( t \) is

\[ g(a,U) = \frac{1}{t} \sum_{b \in U} M(a,b) \]

The dissimilarity between \( a \) and the closest cluster that is different from \( C \) can be defined as

\[ v(a) = \min_{U \neq C} g(a,U) \]

The silhouette width \( s(a) \) for entry \( a \) and the average silhouette \( \bar{s} \) for the set are defined as

\[ s(a) = \begin{cases} \frac{v(a) - c(a)}{\max\{c(a),v(a)\}}, & r \neq 1 \text{ and } r \neq N \\ 0, & r = 1 \text{ or } r = N \end{cases} \]

\[ \bar{s} = \frac{1}{N} \sum_{a=1}^{N} s(a) \]

Silhouette values lie in the range \([-1, 1]\). Entries with a silhouette value \( s(a) \) close to 1.0 are well clustered, in the sense that the average distance to entries in the same cluster is small, compared to average distance to the closest other cluster. If the silhouette value is smaller than 0, the entry is misclassified. PAM clustering is applied for all numbers of clusters \( k \) between 1 and \( N - 1 \) and the corresponding average silhouette values \( \bar{s}(k) \) are calculated. The best clustering corresponds to \( k^* \) number of clusters:

\[ k^* = \arg\max_k \bar{s}(k) \]

Results

The size distribution of the SCOP sets is analyzed first. The dissimilarity measure used for clustering is then compared with the rmsd after superposition. Three clustering examples are presented. In the first example the clustering is based on small structural differences. In the second example, there is both a large conformational change and a local structural difference. The third example illustrates how the silhouette width can be used as a measure of cluster quality.
Distribution of set size and dissimilarity values

Figure 1 shows the histogram of sizes of the protein sets that have been clustered. Most of the sets (76%) have \( \leq 10 \) entries. The percentage of sets with \( >40 \) entries is small (2%). The largest set, with 309 entries, results from the extensive protein engineering work on bacteriophage T4 lysozyme, with SCOP unique identifier (sunid) 53983. All these sets have been clustered and the results are available online at http://bioinf.mpi-sb.mpg.de/projects/struster/.

Table I gives the distribution of the maximum dissimilarity values \( M(a,b) \) observed in each of the 3716 sets clustered. For 10% of the sets there is no dissimilarity. Almost 50% of the cases have considerable dissimilarity (100–1000) and very high dissimilarity (\( >1000 \)) is observed in 10% of the sets.

Dissimilarity versus rmsd

The \( \text{rmsd} \) after rigid-body superposition is a popular measure for expressing the structure similarity between proteins. Let \( d \) be the distance between each pair of equivalent atoms in two optimally superposed structures. The \( \text{rmsd} \) over \( n \) equivalent atoms is defined as

\[
\text{rmsd} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} d_i^2}
\]

In the present work, the structural models are clustered based on a measure of dissimilarity \( M(a,b) \) between structures. This measure is sensitive to both large and small (but still significant) backbone conformational differences. It reflects the extent of significant differences (\( >1 \text{Å} \)) in short- to medium-range intramolecular distances between the two structures \( a \) and \( b \). The value for normalized dissimilarity \( R(a,b) = M(a,b)/L^2 \) is easier to interpret. It is independent of the number of equivalent residues \( L \) in the set and its value lies in the interval between 0 (identity) and 1 (maximum difference).

Figure 2 shows the relationship between \( \text{rmsd} \) and both the dissimilarity measure \( M(a,b) \) (Figure 2A) and the normalized dissimilarity \( R(a,b) \) (Figure 2B). There is considerable positive correlation in both cases (0.813 for \( A \) and 0.815 for \( B \)).

The positive correlation is significant considering that \( \text{rmsd} \) and \( M \) measure different levels of structure similarity. In particular, \( \text{rmsd} \) is applicable for measuring small (all atom) to medium structural differences, while \( M \) captures the extent of backbone conformational changes, including large changes in conformation. In contrast, \( \text{rmsd} \) values spread considerably when large conformational differences are observed. It can

### Table I. Maximum dissimilarity in the clustered sets

<table>
<thead>
<tr>
<th>Dissimilarity</th>
<th>0</th>
<th>0–100</th>
<th>100–1000</th>
<th>&gt;1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sets</td>
<td>373</td>
<td>1220</td>
<td>1748</td>
<td>375</td>
</tr>
<tr>
<td>% of sets</td>
<td>10</td>
<td>33</td>
<td>47</td>
<td>10</td>
</tr>
</tbody>
</table>

Fig. 1. Distribution of the number of entries on the SCOP sets corresponding to the species level. For convenience, only sets with \( <40 \) entries are displayed. The largest set has 309 entries.

Fig. 2. (A) Dissimilarity \( M(a,b) \) versus \( \text{rmsd} \) on a logarithmic scale. For each of the 3716 sets, the \( \text{rmsd} \) was calculated for the pair of entries with largest value of \( M(a,b) \). A smoothed quantile line is shown, corresponding to a sliding window of 1% quantiles of the \( \text{rmsd} \) values, calculated over a window size of 0.1 on a logarithmic scale. Evaluations are made at 100 equidistant points and the resulting quantiles are smoothed with a lowess function (local linear scatter plot smoother).

(B) Analogously, normalized dissimilarity \( R(a,b) \) versus \( \text{rmsd} \) on a logarithmic scale.
also be observed from Figure 2 that for a given dissimilarity value, there is a clear lower limit for \( \text{rmsd} \). A smoothed 1% quantile line is shown for both plots giving the lower \( \text{rmsd} \) limit for given values of \( M(a,b) \) and \( R(a,b) \).

D-2-Deoxyribose-5-phosphate aldolase

The structure models for *Escherichia coli* D-2-deoxyribose-5-phosphate aldolase provide a first example of the clustering analysis. Figure 3 shows the structural superposition of the six structures found in the corresponding SCOP species level (sunid 69395) and the dendrogram obtained from hierarchical clustering. The models present small structural differences. Nevertheless two clusters, \( C_1 \) and \( C_2 \), can be distinguished in the dendrogram. The same two clusters are produced by PAM clustering, with an average silhouette width of 1.0, indicating two well-differentiated clusters. Larger numbers of clusters produce lower average silhouette width values.

It is noticeable from Figure 3A that there are no significant differences in the \( \text{C}\alpha \) atom conformations, except for two residues in the loop before the C-terminal helix (residues 238–239). In this region the structures d1ktna_ and d1ktnb_ (matching cluster \( C_1 \)) have small but significant backbone deviations from the remaining models (for residue 238, \( \text{C}\alpha \) distances of 1.7 Å in the superposition). The dendrogram reflects these small backbone differences. It is noticeable that the models in \( C_1 \) (but not in \( C_2 \)) include the substrate D-2-deoxyribose-5-phosphate attached to the binding pocket. In fact, residue 238 makes a contact to the substrate in the \( C_1 \) models (Heine *et al.*, 2001). As the substrate is not present in the \( C_2 \) models, there is a small conformational change in this last loop. This translates into a more open conformation at the entrance to the binding pocket.

Serum transferrin

Transferrins are responsible for sequestering and solubilizing iron. In particular, serum transferrin binds Fe(III) in the blood and transports it to cells, where it is released at low pH into the endosome. The iron-free apotransferrin is then recycled back to circulation. Vertebrate transferrins consist of a single polypeptide chain with a twofold internal repeat, resulting in two homologous lobes (N- and C-lobes). The lobes contain similar iron-binding sites located in a deep cleft between two \( \alpha/\beta \) subdomains. A single Fe(III) is bound in this cleft to four amino acid side chains and to a \( \text{CO}_2^3- \) ion. The SCOP species level of human serum transferrin (sunid 53899) includes 20 entries for the N-lobe.

Figure 4 shows the corresponding dendrogram and the structure comparison of the different models. Two major clusters separated by a large dissimilarity value are visible in the dendrogram, \( C_1 \) and \( C_2 \) (Figure 4A). In fact, these two clusters correspond to the iron-free apo form (\( C_1 \)) and to the iron-binding holo form of transferrin (\( C_2 \)). The apo form is transformed into the holo form by a large (63°) rigid-body rotation of one of the \( \alpha/\beta \) subdomains relative to the other, resulting in the opening of the iron-binding cleft (see Figure 4B).

Cluster \( C_2 \) is further subdivided (\( C_{21} \) and \( C_{22} \)). These two clusters correspond to structural models obtained from two different crystal forms, one in the tetragonal \( P4_12_12_2 \) space

---

**Fig. 3.** (A) Dendrogram for hierarchical clustering of SCOP PDB entry domains for the species level of deoxyribose phosphate aldolase from *Escherichia coli* (SCOP sunid 69395). One can observe two clusters \( C_1 \) and \( C_2 \). Horizontal scale in dissimilarity units \( M(a,b) \). Entries are identified by the SCOP sid code. (B) Multiple structure comparison of the six models displayed as \( \text{C}\alpha \) trace. Entries from \( C_1 \) are in light gray, entries from \( C_2 \) in black. There is a deviation of the backbone conformation for serine 238 (next to the labels). This residue makes a main-chain contact with the substrate (coordinates from substrate in djcla_ are shown in space fill). All multiple structure comparisons are generated with MULTIPROT (Shatsky *et al.*, 2002) and the molecular graphics are produced with RasMol (Sayle and Milner-White, 1995).
group ($C_{21}$ cluster) and the other in the orthorhombic $P2_12_12_1$ space group ($C_{22}$ cluster). Significant differences between the structures in the two clusters can be observed around a surface loop (residues 136–145) (see Figure 4C). Residues in this loop form a crystal contact in the tetragonal form ($C_{21}$ cluster) but not in orthorhombic form ($C_{22}$ cluster). Other differences can also be observed in the C-terminal region (328–331) and in the 307–309 region; see the original publication for details (MacGillivray et al., 1998).

Within each of these clusters ($C_1$, $C_{21}$ and $C_{22}$), the backbone differences are small and the different models correspond to single residue mutants, different crystallization conditions or different expression systems. In particular, within the $C_1$ cluster, one can observe two subclusters with low dissimilarity. They match almost identical structures (1btj and 1bp5 PDB entries) derived from two closely related crystal forms (Jeffrey et al., 1998).

Figure 5 shows the values of average silhouette width for PAM clustering from one to 19 clusters. Best clustering is achieved with three clusters with a high average silhouette width (0.939) indicating a clear separation between the clusters. The three clusters correspond to $C_1$, $C_{21}$ and $C_{22}$. The representatives are d1bp5d, d1a8f and d1n84a for $C_1$, $C_{21}$ and $C_{22}$, respectively.

**Glucose dehydrogenase**

The examples so far demonstrate ensembles which can be clearly classified into different types of structures. It is also possible that the structural neighborhood of a model (the other closely related models) is less clearly defined. This is the case when the structural differences in the set are more continuous or when the structural neighborhood of a given model varies along the polypeptide chain. These cases are associated with a lower silhouette width. The clustering of glucose dehydrogenase structures from *Bacillus megaterium* provides such an
example. The protein consists of a tetramer with four identical subunits. The corresponding SCOP species level (sunid 51785) includes four sequence-identical entries from the same PDB file (Yamamoto et al., 2001).

The Cα conformation is very similar in these four entries. The only significant difference is located in a flexible surface loop (Arg39–Asp44) (see Figure 6A). Models d1gcoe_ and d1gcof_ (with a very similar backbone conformation) differ from the other models in this region. In particular, the differences between d1gcoe_ and d1gcof_ on the one hand and d1gcoa_ on the other are clear for residues 41 and 44, but the respective differences to 1gcob_ are only significant for residue 44. From the structural comparison it is clear that d1gcoe_ and d1gcof_ should belong to the same cluster and d1gcoa_ to a different one, whereas d1gcob_ is an intermediate case.

The dendrogram in Figure 6B provides one solution, where d1gcob_ clusters with d1gcoa_, but the poor cluster assignment of d1gcob_ is not evident in this type of representation. PAM results (Figure 6C) give a best partition for two clusters, with d1gcob_ and d1gcoa_ clustering together in agreement with the dendrogram. The poor cluster assignment of d1gcob_ is reflected in the lower average silhouette value for d1gcob_.

Discussion

We have provided a method for clustering of protein structures that uses differences in intramolecular distances for the calculation of a dissimilarity measure. Two filters are applied in this calculation. The first introduces a cut-off for large distances. This is especially important in the case of a rigid-body motion of two subdomains as illustrated by the transferrin example. Here, the two subunits maintain the same conformation, only the relative orientation changes. Considering the differences over large distances would otherwise result in very high dissimilarity values, which does not reflect the fact that the subunits have the same conformation. Therefore, the large change in the orientation of the two subunits can mask other smaller structural differences. The second filter places a threshold on the distance differences, therefore discarding small differences that can result in additional noise. The comparison of the clustering results obtained with and without filters illustrates this effect. We use the human serum transferrin set (SCOP sunid 53899) again as an example. With the filters set to $F_1 = 14.0$ and $F_2 = 1.0$ as described before (Figure 4A), one can observe two main clusters (C1 and C2) and two subclusters (C21 and C22). Figure 7A shows the resulting dendrogram when these filters are not applied. In this case the subclusters C21 and C22 are not evident. This observation is confirmed by the PAM clustering results, as the maximum average silhouette width is obtained for two clusters (C1 and C2). Similar PAM results are obtained when rmsd is used as a dissimilarity measure. Figure 7B gives the corresponding hierarchical clustering results. The subclusters are now more noticeable than in Figure 7A, but still not as evident as when the filters are applied (see Figure 4A), as in the method proposed.

Two clustering techniques have been applied, hierarchical and PAM clustering. These different methods provide complementary views on the relationships observed. The former method reflects the hierarchy of structure similarity (or dissimilarity) between the entries, where the dendrogram provides an overall view of the structure relationships. The latter method
(PAM) provides an objective partitioning of the entries into the optimal number of clusters. In addition, the silhouette width value provides a measure of the quality of the clustering. Other clustering approaches could be tested: single-linkage or complete-linkage hierarchical clustering, divisive hierarchical clustering or maximum-likelihood methods, for example. Full cluster optimization is outside the scope of the current work, but we intend to test further these additional methods and dissimilarity measures. In particular, the stability of the resulting classifications over a jackknife procedure should be investigated.

The current method only takes into account the Cα conformations. Including the coordinates of the side-chain atoms would make the procedure more sensitive to small structural changes, which might be important for more detailed structural analysis. In addition, we will also investigate how to include temperature factors and occupancy information in the method.

Acknowledgements

We thank Ingolf Sommer and Mario Albrecht for helpful discussions and Andreas Kämper and Oliver Sander for their comments and suggestions. Financial support was provided to J.R. by BMBF grant No. 031U117. This work is part of the BioSapiens project. The BioSapiens project is funded by the European Commission within its FP6 Programme, under the thematic area ‘Life sciences, genomics and biotechnology for health’, contract number LHSG-CT-2003-503265.

References


Received June 18, 2004; accepted August 3, 2004

Edited by Andrej Sali