Fine-Grained Structure-Function Clustering of Pfam Protein Domain Families: A case study Using CYP450

Nicolás Denis [1, 2], David W. Ritchie [2]
1. University Paris 7 – Denis Diderot, Paris, France
2. INRIA Nancy – Grand Est, 615 Rue du Jardin Botanique, Villers-lès-Nancy, France

Introduction

Until recently, it has been very computationally costly to compare multiple three-dimensional (3D) protein structures. Current protein structure databases such as SCOP [1] and CATH [2] provide extremely useful resources. However, it is generally very expensive to search these databases using structure-based queries. Also, these databases should be updated regularly thanks to the growing number of new structures solved by structural genomics efforts, but this is also a very expensive procedure. In order to help tackle such problems, a new protein structure alignment and comparison tool called “Kpax” (kpx.keria.fr) [3] has recently been developed. This allows us to perform all-against-all structural alignment and comparison calculations rapidly and reliably. Despite its speed, Kpax is limited to only perform structure-based comparisons through similarity scores. Thereby, we used these scores to obtain distance values in order to build distance matrices. With them, we were able to make a rapid clustering of proteins domains shapes and defined subgroups inside a family by using a hierarchical classification approach.

As a first step, we wanted to test our approach on a known family. We chose the cytochrome P450 family (Cyp450) [4] because of the significant number of different structures and the already known data on the InterPro [5] database. Therefore, we developed a series of Python scripts to retrieve redundant domains and to extract automatically information from InterPro. The aim was to perform a variety of shape-based clustering calculations using R scripts. We found that the global grouping of these domains seems consistent with the classification from InterPro. We also found that Cyp450 can be clustered into 8 significant structural groups. We are currently working to extend our approach to be able to apply it automatically to all Pfam families [6].

Kpax

Kpax is a protein structure alignment program which exploits the special tetrahedral geometry around each Cα atom (Figure 1). It uses dynamic programming (Needleman and Wunsch, 1970) to optimize the local (Figure 2) and global (Figure 3) Cα environments in order to find the best superposition and to calculate a similarity score between two proteins. In fact, Kpax calculates two similarity scores, the "K-score" (before superposition) and the "G-score" (after superposition). It also calculates a superposition "T-score" using the formula of the "TM-Align" structure alignment program.

Methods

Data filtering

In order to obtain clusters based on the structures shapes of protein domains, we used Kpax to firstly get a distance matrix between them. Distance values of the matrix are calculated with the formula below:

\[
\text{Distance} = \sqrt{\left| A - B \right|^2 - 2 \times \text{K-score} \times \left| A \right| \times \left| B \right|}
\]

where A and B are the 3D protein vectors and J-score is the normalised K-score.

Clustering

In order to obtain clusters based on the structures shapes of protein domains, we used Kpax to firstly get a distance matrix between them. Distance values of the matrix are calculated with the formula below:

\[
\text{Distance} = \sqrt{\left| A - B \right|^2 - 2 \times \text{K-score} \times \left| A \right| \times \left| B \right|}
\]

where A and B are the 3D protein vectors and J-score is the normalised K-score.

3D viewing

We observed several 3D superpositions of proteins to have a better understanding of the similarity and dissimilarity between them. Thus, we were able to observe proteins with high (Figure 5), medium (Figure 6) and low (Figure 7) similarity between them.

Conclusions

We used the 3D structural information from Pfam and InterPro to compare our results with already known data. By using both the K-score and T-score scoring functions provided by Kpax, we were able to measure the similarity between proteins and build distance matrices. We used this distance matrix to build a hierarchical classification of protein domains. From this classification, we found that more than 50% of our results are very similar to the InterPro classification of cytochrome P450 (CYP450). This classification is based on structural information and expert analyses provided by several different databases.

This methodology could therefore be applied to other Pfam families to detect their general organization and 3D folds. Our next objective is to carry out this on a large scale.

Kpax

Kpax is a protein structure alignment program which exploits the special tetrahedral geometry around each Cα atom (Figure 1). It uses dynamic programming (Needleman and Wunsch, 1970) to optimize the local (Figure 2) and global (Figure 3) Cα environments in order to find the best superposition and to calculate a similarity score between two proteins. In fact, Kpax calculates two similarity scores, the "K-score" (before superposition) and the "G-score" (after superposition). It also calculates a superposition "T-score" using the formula of the "TM-Align" structure alignment program.

Results

We used the 3D structural information from Pfam and InterPro to compare our results with already known data. By using both the K-score and T-score scoring functions provided by Kpax, we were able to measure the similarity between proteins and build distance matrices. We used this distance matrix to build a hierarchical classification of protein domains. From this classification, we found that more than 50% of our results are very similar to the InterPro classification of cytochrome P450 (CYP450). This classification is based on structural information and expert analyses provided by several different databases.

This methodology could therefore be applied to other Pfam families to detect their general organization and 3D folds. Our next objective is to carry out this on a large scale.

References