J Mol Model (2001) 7:140–142
DOI 10.1007/s008940100023

ORIGINAL PAPER

Alexander Archakov · Andrey Lisitsa · Semen Gusev
Luc Koymans · Paul Janssen

Inventory of the cytochrome P450 superfamily

Received: 29 March 2000 / Accepted: 22 March 2001 / Published online: 22 May 2001

Abstract A non-redundant database of 325 protein sequences of the P450s has been subjected to cluster analysis. Thresholds for (sub)families have been optimized to minimize the differences between the obtained clusters and nomenclature adopted for the P450s. At the given thresholds, approximately 80% of the systematic nomenclature is reproduced by the cluster analysis. The differences primarily occur among the CYP4 and CYP6 families, which include cytochromes P450 of mammalian and insect origin. Conflicts are also encountered among plant families and among P450s of *Mycobacterium tuberculosis*.

Keywords Inventory · Cluster analysis · Cytochrome P450 · Nomenclature

Introduction

Existing classifications attempt to group together enzymes sharing a similar function. However, for many recently sequenced proteins the details of their function are unknown. Expert-provided annotations are ought to be replaced by the automatic ones.

The preliminary characterization of the proteins is carried out using BLAST. [1] This local alignment search tool successfully inserts the novel structures into the correct superfamilies. However, to obtain the interior order for the proteins within a given superfamily, a cluster analysis is required.

Cluster analysis starts from the matrix of pairwise cross-similarities. Similarity between biological sequences is assessed by pairwise alignment. [2] Usually the results of cluster analyses are represented as a dendrogram. One can cut off the dendrogram, thus obtaining a number of groups of proteins. In other words, the process of cluster agglomeration can be stopped at a given threshold. One of the methods proposed for threshold determination [3] is the comparison of the clusters obtained with experts' opinions.

In this context such comparisons seem to be useful for inventorying the cytochrome P450 superfamily. This superfamily principally follows the nomenclature conventions developed by the P450 Committee. [4,5] The proteins are classified into several families; those members of the family that exhibit particular resemblance to each other are grouped into subfamilies. Each P450 isozyme has a unique systematic name assigned in accordance with its family and subfamily.

Initially, the P450 nomenclature was based on the functional principals, e.g. (sub)families united the proteins that have catalyzed the same reactions. As the number of P450s increased, Nebert and co-workers [4] proposed separating out (sub)families on the basis of structure. Thus, those cytochromes P450 that are at least 40% identical constitute a family. Mammalian proteins are grouped into a subfamily if they are at least 54% identical, whereas for incorporation of other vertebrates this value is reduced to 46%.

The goal of the present research is to find out how well the official P450 nomenclature is reproduced by a cluster analysis and what are the optimal thresholds for (sub)families. The mismatches are summarized and discussed.

Materials and methods

The amino acid sequences of the cytochromes P450 were obtained from the Cytochrome P450 Database, release 1998 (http://cpd.ibmh.msk.su). Only 325 of 712 sequences were selected for the analysis using the NRDB90 program, [6] which discarded more then 90% identical sequences. The non-redundant database includes 78 families and 136 subfamilies.

A. Archakov · A. Lisitsa (✉) · S. Gusev
Institute of Biomedical Chemistry,
Pogodinskaya 10, Moscow, Russia
e-mail: fox@ibmh.msk.su

L. Koymans · P. Janssen
Center for Molecular Design,
Antwerpsesteenweg 37, Vosselaar, Belgium

The similarity between every pair of P450 sequences was assessed by pairwise alignment under the optimized gap penalty [7] and unit/blosum62 substitution matrix. The obtained matrix of pairwise cross-similarities was handled by UPGMA (unweighted pair grouping using mean arithmetics [8]) cluster analysis method.

If agglomeration procedure is halted at some threshold T, then two types of mismatches between clusters and (sub)families will occur. First, a "together" mismatch is observed when the members of different (sub)families are found within the same cluster. Second, an "apart" mismatch happens when members of the same (sub)family are distributed over several clusters. The number of "together" and "apart" mismatches is referred to as M^T and M^A, respectively.

Both M^T and M^A depend upon the threshold T (Fig. 1). The optimal threshold T^O is the one where mismatches of the two types are balanced: $M^T(T^O)=M^A(T^O)$. At the optimal threshold the correspondence between clusters and (sub)families is calculated as:

$$\frac{\text{Number of (sub)families exactly reproduced by cluster analysis}}{\text{Total number of (sub)families}} \times 100\%$$

Note the optimal threshold T^O determined as described above does not guarantee maximal correspondence between clusters and (sub)families. It only minimizes the number of mismatches. Such an approach is justified because P450 (sub)families are unevenly represented due to artificial reasons. The inequality of (sub)families' sizes is neutralized because M^T stands for the number of wrong (sub)families whereas M^A counts the wrong clusters.

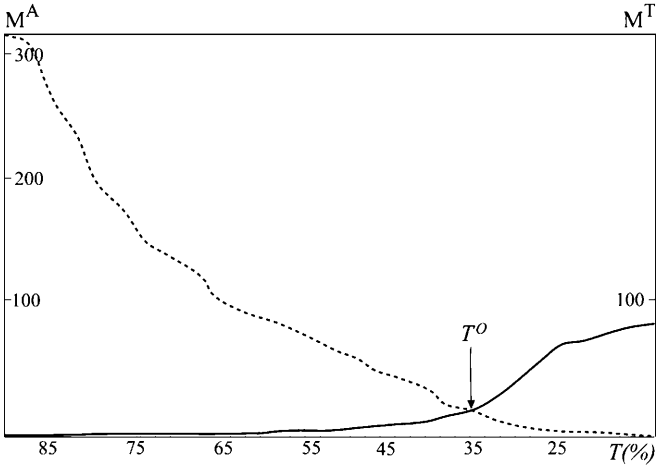

Fig. 1 T threshold-dependence of "together" (M^T, solid) and "apart" (M^A, dotted) mismatches. The arrow indicates the optimal threshold, T^O, for cytochrome P450 families

Table 1 The thresholds for families and subfamilies

	Families		Subfamilies	
Substitution matrix	Unit	Blosum	Unit	Blosum
Optimal threshold	35%	44%	47%	57%
Thresholds of Nelson *et al.* [5]	40%		46%	
Correspondence with nomenclature	82%	82%	83%	85%

Results and discussion

The results are summarized in Table 1. The nomenclature of the cytochromes P450 is reproduced quite well by means of the cluster analysis. Protein clusters coincide with the families in 82% of the cases. Such coincidence is achieved at 35% threshold, which is lower than the threshold proposed by Nelson and co-workers.

The threshold for subfamilies (47%) slightly exceeds that of Nelson et al. The correspondence between subfamilies and clusters is also remarkably high – 83%. It is worth mentioning that the substitution matrix promotes the formation of subfamilies, but it is not effective in revealing weak relatedness within P450 families.

Essential differences revealed by comparison of P450 families with UPGMA clusters are accumulated in Table 2. "Apart" mismatches are encountered in families CYP4 and CYP6. Family CYP6 contains exclusively cytochromes P450 of insect origin whereas CYP4 incorporates both insect and mammalian sequences.

The antique family CYP4 plays an important role in endogenous metabolism of fatty acids in insects as well as in mammals. At the same time, significant diversity of their primary structures is revealed. It probably means that some members of CYP4 have acquired new functional properties in addition to the metabolism of fatty

Table 2 Mismatches between families and clusters

"Apart" mismatches			
Family	Clusters		
CYP4	CYP4M	CYP4C CYP4D CYP4E	CYP4A CYP4B CYP4F
CYP6	CYP6A CYP6C	CYP6A	CYP6C
"Together" mismatches			
CYP93 CYP75 CYP92			
CYP89 CYP77			
CYP123 CYP124 CYP125			

acids. It is most likely that insect cytochromes of the family CYP4 have evolved to provide insecticide resistance. [9] Such functional specification is certainly unusual for mammals. Therefore they form a separate cluster. The remaining members of CYP4 are of insect origin. They stay together, with the exception of CYP4M2 from the tobacco hornworm. This exception may indicate that there is no uniformity even among the insect P450s of CYP4. The observed structural variety can be explained by a wide range of different insecticide agents or by the stepwise metabolism of xenobiotics.

The same processes probably take place in the CYP6 family, which falls apart into three clusters. The evolutionary peculiarities of insect P450s may be confirmed by the fact that different clusters include enzymes of the same origin (house fly). Presumably, CYP6 collects the genes that diverged very early and since that moment have evolved independently because they are oriented to different functions. In all likelihood CYP6 gathers representatives of different actual families. The same is true of the insect portion of CYP4.

Despite CYP6 and CYP4 falling apart at the 35% threshold, they both look like clusters on the dendrogram, suggesting a common ancestor. However, from the functional point of view, proteins within each of these families have accumulated enough mutations to obtain specificity either in the substrate or the reaction or both.

"Together" mismatches occur among cytochromes P450 of plant origin. Thus, families CYP93, CYP75 and CYP92 are united into a single cluster. Another cluster unites CYP89 and CYP77. The reason for the appearance of these clusters is not quite clear because these families are poorly investigated. It is known that CYP75 catalyzes the reaction on the anthocyaninsynthetic pathway. [10] Probably the same function should be expected for families CYP93 and CYP92. In general, the plant P450s were not thoroughly inventoried because a large portion of *Arabidopsis thalania* enzymes was ignored.

Finally, mismatches are observed among bacterial P450s. The cluster contains CYP123, CYP124 and CYP125 obtained from the *Mycobacterium tuberculosis* genome. This group of enzymes shows a resemblance with CYP107A1, which is known to be involved in erythromycin biosynthesis. It can be hypothesized that CYP123–125 also produce some original *Mycobacterium tuberculosis* antibiotics.

The proposed inventory provides the generalized insight onto cytochrome P450 nomenclature. The great part of the nomenclature is confirmed by the cluster analysis although the family threshold ought to be shifted. The reasons for mismatches are:

1. lack of functional data, which leads to erroneous classification of the enzymes just because there is not enough experimental evidences to prove the experts' decision. "Together" mismatches point to the presence of such errors.
2. shortcomings of the cluster analysis, which assumes a molecular clock [11] during evolution remodeling. The concept of a molecular clock assumes an equal rate of evolution through all the branches of the evolutionary tree. This assumption turns out to be false in the case of the P450 superfamily in general. It is most likely that some clans [12] of cytochromes P450 are evolving differently from the others. Most of the P450s analyzed are of mammalian origin; therefore the mismatches mainly occur among insect, plant and bacterial enzymes. The proposed inventory revealed in the superfamily those groups that need special analysis with more advanced computational methods. "Apart" mismatches may be caused by the peculiarities of clustering algorithms mentioned above.

The inventory of the cytochrome P450 superfamily is useful for an adequate annotation of new enzymes. Having established the thresholds, one can easily incorporate the novel structure into the relevant cluster. Thus, the inventory is a step towards automatic preliminary classification of the proteins.

References

1. Muller, A.; MacCallum, R. M.; Sternberg, M. J. *J. Mol. Biol.* **1999**, *293(5)*, 1257–1271.
2. Gotoh, O. *Adv. Biophys.* **1999**, *36*, 159–206.
3. Milligan, G. W.; Cooper M. C. *Psychometrika* **1985**, *50(2)*, 159–179.
4. Nebert, D. W. *et al. DNA Cell Biol.* **1991**, *10(1)*, 1–14.
5. Nelson, D. R.; Koymans, L.; Kamataki, T.; Stegeman, J. J.; Feyereisen, R.; Waxman, D. J.; Waterman, M. R.; Gotoh, O.; Coon, M. J.; Estabrook, R. W.; Gunsalus, I. C.; Nebert, D. W. *Pharmacogenetics* **1996**, *6(1)*, 1–42.
6. Holm, L.; Sander, C. *Bioinformatics* **1998**, *14(5)*, 423–429.
7. Archakov, A. I.; Lisitsa, A. V.; Zgoda, V. G.; Ivanova, M. S.; Koymans, L. *J. Mol. Model.* **1998**, *4*, pp. 234–238.
8. Sneath, P. H. P.; Sokal, R. *Numerical Taxonomy*; Freeman: San Francisco, 1977.
9. Fogleman, J. C.; Danielson *Chem Biol. Interact.* **2000**, *125*, 93–105.
10. Holton, T. A. et al. *Nature* **1993**, *366(6452)*, 276–279.
11. Kimura, M. *The neutral theory of molecular evolution*; Cambridge University Press: Cambridge, 1983.
12. Nelson, D. R. *Arch. Biochem. Biophys.* **1999**, *369(1)*, 1–10.