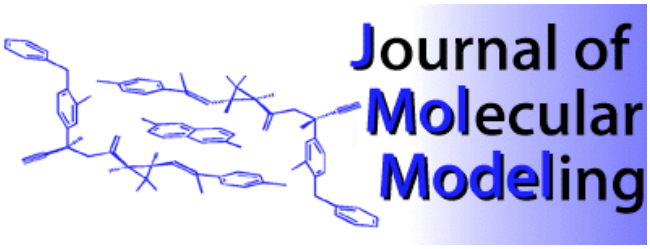
*J. Mol. Model.* **1998**, *4*, 234 – 238




**FULL PAPER**



# Clusterization of P450 Superfamily Using the Objective Pair Alignment Method and the UPGMA Program

**Alexander I. Archakov[1], Andrey V. Lisitsa[1], Victor G. Zgoda[1], Marina S. Ivanova[1], and Luc Koymans[2]**

[1]Institute of Biomedical Chemistry, Moscow 119832, Pogodinskaya 10, Russia. E-mail: fox@ibmh.msk.su

[2]Janssen Research Foundation - Center for Molecular Design, Antwerpsesteenweg 37, B-2350 Vosselaar, Belgium.

*Received: 3 February 1998 / Accepted: 21 April 1998 / Revised: 26 June 1998 / Published: 16 July 1998*

**Abstract** DNA translation to the protein sequences determines the common usage of gene name as the enzyme identifier. The previously constructed single-family-member phylogenetic trees are produced by the pair alignment. The alignments strictly depend upon the user-defined parameters and algorithmic peculiarities, such as but not limited to: homology matrix, initial gap penalty value and gap elongation function. This rises the necessity to create complete clusterization which reflects the protein primary structure relationships. This protein-based clusterization should be made using the objective pair alignment. The standard dynamic alignment procedure is modified in order to discriminate between the suboptimal resulting scores. The special function treats the presence of continuous matching n-tuples as a good property of alignment. Pair alignment is objectified by finding the optimal gap penalty, that allows to get the maximal difference in identity between random and relative sequences. The method is applied to the cytochrome P450 superfamily. Our sample also contained 15 nitric oxide synthases and 30 random sequences. The similarity matrix, obtained by objective pair alignment, is worked up by standard UPGMA method.

**Keywords** Cytochrome P450, Objective pair alignment, Proteins clusterization, Homology estimation

## Introduction

Cytochrome P450 superfamily is now classified to 481 genes and 22 pseudogenes. However, the authoris original assumption, that each gene of P450 almost always produces one protein, is not always correct [1]. Several families and subfamilies, in which one gene is responsible for production of different proteins, are found. For example, in family 1, only two genes are responsible for production of more than 20 different proteins. The same situation is observed in families 19, 52, 51. Besides, the attempts to define the digital frontiers of family and subfamily reveal rather many exceptions: the incorporation of more distant species decrease the identity from 40% to 30-34% and from 55 to 46% for family and subfamily respectively. It is quite obvious, that these values depend both upon the applied method of pair alignment and upon the calculation of the gap penalty value [2]. In our opinion, the proposed objective pair alignment procedure allows to obtain more stable clusterization of proteins and to reduce the number of exceptions. The obtained clusters of proteins should be regarded as objective analogs

*Correspondence to:* A. Lisitsa

of genetic (sub)families. Later on, cytochrome P450 proteins are assumed to be ranked in accordance with their distance from consensus sequences of subfamilies, families and superfamily in hierarchical way.

## System

The program ENTRY is written in C and compiled with the Borland Turbo C 2.0 compiler. It can be executed under MS-DOS 3.x or higher. The minimal requirements are: PC386, 640KB RAM, N2/100 KBytes free disk space (N-number of entries in the sample). Our sample, containing approximately 500 entries, is analyzed on a Pentium Pro 200MHz.

## Algorithm

It is known, that standard dynamic programming may produce several suboptimal alignments [3]. Hence, first of all, the program chooses one of such isomorphic results by means of the F function. From suboptimal alignments the one, that shows maximal F is selected. The function F is calculated after the alignment is completed. It analyses the pair alignment consensus, which contains the matching amino acids, mismatches and gaps. The function F evaluates the coinciding n-tuples instead of the separate amino acids. It is believed, that the greater number and/or length of n-tuples is a positive criterion for the alignment quality.

For the given consensus of length L one can calculate the probability $p_0$ of observing two amino acids without gaps between them as p0=(N-1)/(L-1), where N represents the number of matches. In other words, p0 is the probability of the appearance of duplet in the alignment consensus. Thus, the function is calculated by the formula:

$$F = \log \sum n(l) \cdot p0^{(l-1)} \ , \ 1< l <L$$

where n(l) is the number of n-tuples of length l. If there is only single duplet (i.e. l=2, n(l)= 1) in the consensus, then F=log p0. The extension of n-tuple leads to linear increase of

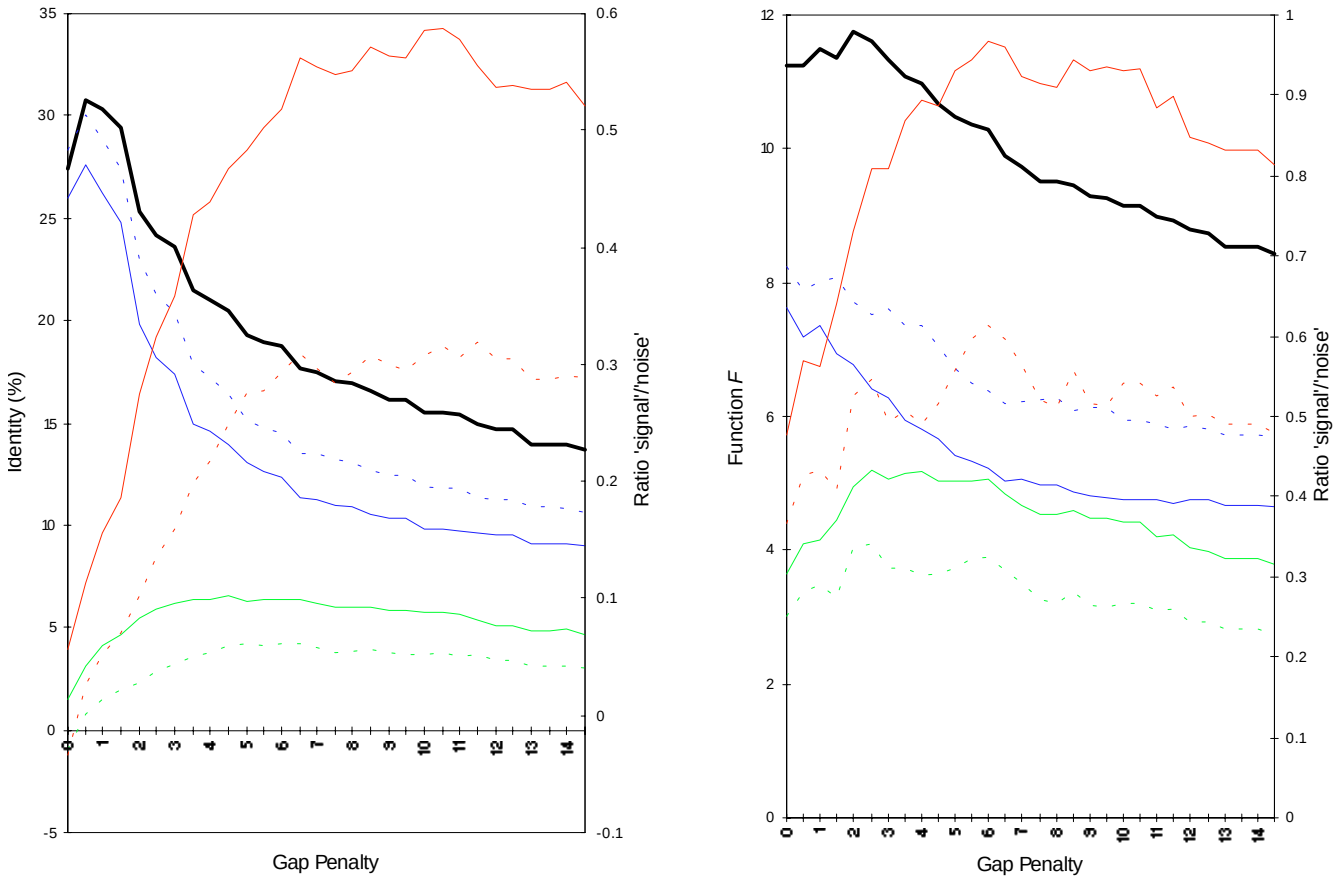



**Figure 1** *Dependence of Identity (left box) and* F *(right box) on gap penalty. Solid black line - Identity and* F *for distant relatives; solid lines - e-randoms, dotted lines - c-randoms; blue - Identity and* F *for random sequences, green - difference (*d*Identity and* d*F), red - ratio "signal"/"noise" (*r*Identity and* r*F).*

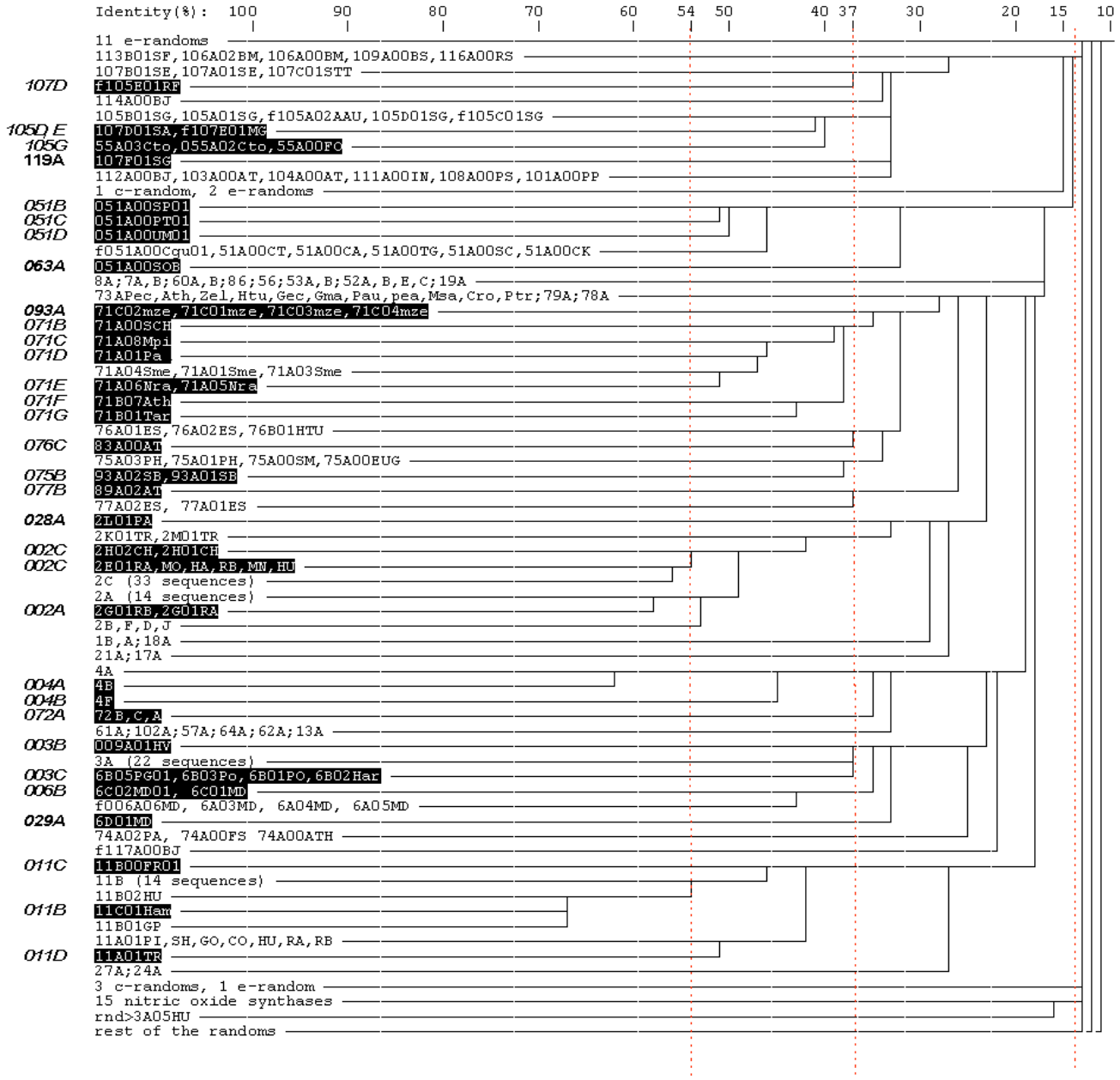



**Figure 2** *Comparison of objective protein clusterization with common genetic classification (GC). The differences between objective protein clusterization and common genetic classification are selected using black background. Proposed new protein names are at the left of scheme in italic. All bacterial cytochromes (exceptions are 102A and 117A, which appear among fungi and plants) form the separate families at the beginning of the scheme. The distinctions with GC are: 105E joins 107 family as 107D subfamily, former 107E, D joins 105 family. The identity analysis of 55A proteins undoubtedly shows, that they also belong to 105 family. Besides, 107F is treated as new 119 family. On contrary to GC family 51 is separated to four subfamilies and, moreover, 51A00SOB constitutes the other family - 63. Plant proteins (73, 76, 83, 75, 93, 89, 72) follow GC. 71C splits off as family 93, former 93A joins to 75 family and 89 joins 77. Most of animal family 2 members (2L subfamily probably represents the separate family 28) really belongs to one family. 2H and 2E appeared to be a part of large 2C subfamily. 2A and 2G are also combined to single subfamily. Rest subfamilies of family 2 are preserved. Subfamilies 72A,B,C are unified to single 72A subfamily. With small exceptions (6D1MD makes up new family 29, 9A and 6B join family 3, families 4 and 11 are internally rearranged) the rest of clusterization fits GC.*

proposed function: F~(l-1). The increase of n-tuples quantity leads to logarithmic function increase: F~log(n(l)*p0).

The described behavior of F function reflects two obvious assumptions. First, the statistical probability of random coincidence of many n-tuples in different parts of the alignment is rather low, so the prize is logarithmically increased. Second, the probability of the random occurrence of the long n-tuple is decreasing with the growth of its length, so the prize is linearly increased. Thus, function F treats the existence of long-stretched n-tuples as more valuable property of an alignment than the great amount of short n-tuples.

In spite of function F determines the optimal alignment, one should realize, that F is calculated a posteriori, so it doesnit drive the alignment procedure. This restricts the usage of F function as alternative estimate to the identity.

Further step in creation of the objective pair alignment is the choice of optimal gap penalty. We consider the optimal alignment parameter as the one, that allows the largest difference in identity between random generated and relative sequences. Two types of random sequences are used. The random sequences of first type (e-randoms) are generated on the basis of equiprobable amino acid composition. The random sequences of second type (c-randoms) are generated using the amino acid average composition of the cytochromes superfamily.

15 cytochrome P450 sequences, which showed the identity less than 30% , are selected for the analysis. The procedure runs from zero gap penalty to 15 with step 0.5. Each step includes:

• calculation of all possible pair alignments between 15 relative sequences ( (15·15-15)/2=105 alignments);

• calculation of the same number of pair alignments for e-randoms;

• calculation of the same number of pair alignments for c-randoms;

• calculation of the arithmetic averages of the score, identity and function F for relatives, e-randoms and c-randoms;

The unit substitution matrix is applied to produce the pair alignments. The gap penalty for the next insertion number i is calculated by the formula[4]:

$$GapPenalty(i) = InitialGapPenalty \cdot e^{-0.1 \cdot (i-1)}$$

Every time the random sequence is used it is generated de novo.

The case of comparison of two random sequences is considered as a statistical or evolutionary "noise". On the other hand, the comparison of two relative sequences is considered as a mixture of "signal" and "noise". The supreme goal of the described procedure is to detect the gap penalty, that allows the maximal domination of the "signal" over the "noise". The difference (*d*) between the "signal"+"noise" and "noise" is calculated for each of the alignment outputs (*d*Identity and *d*F). In the above terms, d is the pure "signal" without "noise". Taking into the account, that the preferable situation is characterized by high "signal" and low "noise", the ratio "signal"/"noise" (*r*) is calculated (*r*Identity and *r*F).

Figure 1 shows the dependence of *d*Identity and *d*F on the gap penalty value. It can be seen, that *d*Identity is a function of penalty value. The difference is small when the values of gap penalty are less than 2, maximum is attained at 4.9_1.1 for e-randoms and even more (6.6_1.1) in the case of c-random sequences. One interesting peculiarity is revealed here: the "signal"/"noise" ratio is much higher, when the random sequences are generated from the equiprobable composition (*r*Identity = 0.6), compared to the ratio in the case of c-randoms (*r*Identity = 0.3). This fact shows, that protein relations are revealed in amino acid composition as well as in sequence. Later on, determining the identity, we used the initial gap penalty equal to 6.5 because in all cases we observe the highest possible difference between random sequences and P450s and ratio Èsignali/inoisei is 0.55.

*r*F also exhibits the maximum in the range of gap penalty value from 5.5 to 6.5. The above conclusion, that the relationship is contained in the amino acid composition itself, is once again verified: rFe-randoms > rFc-randoms. The ratio Èsignali/inoisei reaches 1.0 and 0.6 for e- and c-randoms respectively. That is significantly higher in comparison with the ratio values for the identity (0.6 for e-randoms and 0.3 for c-randoms).

The similarity matrix for all proteins of superfamily is built using the alignment tuned in the above mentioned way. Except for P450s and NOS, 10 c-randoms and 20 e-randoms are included to sampling. Similarity matrix is processed by classic procedure UPGMA (Unweighted Pair Group Mean Arithmetic) [5], according to which the distance DAB between cluster A of n elements and cluster B of m elements is calculated by the formula:

$$D_{AB} = \sum \frac{D_{ij}}{n \cdot m}$$

$D_{ij}$ is the distance between i-element of cluster A and j-element of cluster B. Clusters spaced at a minimal distance are combined into one at each step of algorithm.

The agreed-upon values, 40% and 55% for family and subfamily respectively, were the starting point in finding of the cut-off limits [6]. Analysis of the obtained results displayed, that the best agreement between genetic classification [1] and our clusterization is observed, when identity limit is decreased from 40% to 37% for families and from 55% to 54% for subfamilies. Under this conditions 56 clusters constituting various P450s families and 105 clusters of subfamilies were constantly revealed here. The dependence of clustersi quantity on the identity percentage sharply decreases at 96%. It allows to define 96% as the identity value for proteins probably produced by allelic genes.

## Results and discussion

Program. In the most simple case the program demands only the file with protein sequences in FASTA format. The package isnit devoted only for the gap penalty choice by "signal"/"noise" analysis of identity. The module, that compares sequences, is designed as external program, which can be re-

placed by the user-defined one. The existing modules for identity and F calculation provide the interface between ENTRY and standard the Needleman-Wunsch algorithm. It is possible to compute more modules of this kind, which will serve as gates to widely-used programs. The number and type of parameters are also the variable subjects.

Parametric alignments. The considerable disagreements about how to weight and penalize the alignment matches and gaps involves the conception of the optimal alignment, which is expected at some definite parameters. In absence of strict theory the empirical methods can be used. The idea realized in parametric alignments is to partition the parameter space into regions so, that in each region the alignment is optimal throughout [7]. The maximal score is considered as the optimal one. The analysis of statistical significance of similarities obliges to reject the absolute values in favor of the related to random ones [8]. The combination of parametric and statistical approaches is the key feature of proposed procedure.

Homology estimate. Homology is the qualitative conception, which is assessed by different quantitative methods. The evaluation of local scores and analysis of low-complexity regions [9] are suitable for searching the query sequence homologue through large database. This is explained by the extreme diversity of analyzed information. Cytochrome P450 superfamily is characterized by long and weak homology, that is why the global alignment is used. As far as our task is to determine the boundaries of families and subfamilies quantitatively, we used the identity as protein homology inference. It assists the comparison of the obtained results with the published ones [1].

## Conclusion

A database containing 424 sequences (sequence fragments consisting of less then 350 amino acids were omitted) of cytochrome P450 (297-animal, 99-plant and yeast, 28-bacterial), which make up 117 subfamilies and 62 families according to [1] was used for the analysis. The database contains also 15 sequences of nitric oxide synthases.

As one can see from Figure 2, each of the random sequences forms a separate cluster. The identity between randoms doesn't exceed 15%. Under these conditions, there is no essential difference between sequences generated from the average statistic 1/20 amino acid composition and sequences generated from the average composition of the cytochrome P-450. Nitric oxide synthase forms a cluster, which splits off about 15% of identity, but on the basis of the clusterization data one cannot conclude whether this cluster is a member of the cytochrome superfamily or not, because the cluster of cytochrome family 51 splits off at approximately the same level.

The obtained clusterization is stable and does not change when the sequences got mixed up. The clusters' position has changed one relative to another only, but no changes and redistributions of sequences among the family clusters or within the subfamilies of one and the same cluster have occurred. In fact, a major advantage of obtained clusterization is that it is quite objective, well reproducable and can be achieved in fully automatized mode. The main problems of its application are the disagreements between our clusters and genetic classification, which is well known and widely used in this field. Thus, despite the advantages, its application will be impeded. But how may be resolved this contradiction? In our opinion, the clusterization we suggest may be the basis of hierarchical proteins' ranking with consequent assignment of three positional codes [10]. These codes are projection of the multidimensional space of pair similarities to the single axis, with incorporation of biologically significant information about families and subfamilies. These codes, together with the genetic names can be used for the automatic designation of each cytochrome P450, avoiding mistakes in their names.

## References

1. 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. and Nebert, D.W. *Pharmacogenetics* **1996**, *6*, 1-42.
2. Gotoh, O. *Bull. Math. Biol.* **1990**, *52*, 359-373.
3. Pearson, W.R. and Miller, W. *Methods in Enzymology* **1992**, *210*, 575-601.
4. Moereels, H.; De Bie, L. and Tollenaere, J.P. *J. Comput.-Aided Mol. Design* **1990**, *4*, 131-145.
5. Sneath, P.H.A. and Sokal, R.R. Numerical Taxonomy. *The principles and practice of numerical classification*; W.H. Freeman and Co, San Francisco: 1973.
6. Dayhoff, M.O.; Schwartz, R.M. and Orcutt, B.C. *National Biomedical Research Foundation* **1978**, *5*, 345-352.
7. Gusfield, D. and Stelling, P. *Methods in Enzymology* **1996**, *266*, 481-496.
8. Pearson, W.R. *Methods in Enzymology* **1996**, *266*, 227-258.
9. Altschul, S.F.; Boguski, M.S.; Gish, W. and Wootton, J.C. *Nature Genetics* **1994**, *6*, 119-129.
10. Archakov, A. and Degtyarenko, K. *Biochemistry and molecular biology international* **1993**, *31*, 1071-1080.