*Nucleic Acids Research, 2004, Vol. 32, Database issue* ***D351–D353***
*DOI: 10.1093/nar/gkh040*

# PHYTOPROT: a database of clusters of plant proteins

**S. Mohseni-Zadeh, A. Louis[1], P. Brézellec and J.-L. Risler***

Laboratoire Génome et Informatique, UMR 8116 and [1]Infobiogen, Tour Evry 2, 523 Place des Terrasses, 91034 Evry Cedex, France

Received July 17, 2003; Revised and Accepted September 18, 2003

## ABSTRACT

**All the protein sequences from plants (including *Arabidopsis thaliana*) available from SwissProt/TrEMBL have been the subject of an all-by-all systematic comparison and grouped into clusters of related proteins. Within each cluster, the sequences have been submitted to pyramidal classification; in the case where two or several subfamilies have been grouped together, the pyramidal tree helps in finding which sequences make the links between subfamilies. In addition, the 'domains' that are common to two or more sequences within a cluster were determined and displayed *à la* ProDom. The resulting graphical representations proved to be quite efficient in pinpointing those protein sequences suffering from a probable error in the annotation of their genes. The clusters can be searched through various criteria and their pyramidal classifications and their domain representations can be displayed by querying http://genoplante-info.infobiogen.fr/phytoprot. The user can also launch a BLAST search of a query sequence against all the clusters.**

## INTRODUCTION

Just as the number of completely sequenced genomes maintains its exhausting pace of growth, the number of (mostly putative) protein sequences increases regularly. This has prompted various projects based on massive all-by-all sequence comparisons, aimed, for example, at predicting functions of proteins, delineating characteristic subsequences, differentiating orthologues from paralogues, building phylogenetic reconstructions (1–9). While functional annotation by homology of protein sequences is certainly efficient—even if not error free, such comparisons and clusterings can also help in pinpointing those conceptual protein sequences that result from probably erroneous genomic annotations. A former study based on ~14 000 proteins from plants (10) indeed showed that artifactual gene fusions or sequencing errors resulting in frameshifts and premature stop codons could easily be detected by mere inspection of the ProDom-like (6) representation of domain arrangements of proteins within the clusters. We present here an extension of this study, where the complete proteome of *Arabidopsis thaliana* and all the available sequences from other plants have been compared and grouped into clusters. The resulting database of clusters, called PHYTOPROT, can be queried at http://genoplante-info.infobiogen.fr/phytoprot.

## CONSTRUCTION OF PHYTOPROT

The protein sequences from *A.thaliana* were retrieved from the EMBL proteome site (http://www.ebi.ac.uk/proteome) and those from other plants from the Swiss-Prot (release 40) and TrEMBL (release 20) data banks (11). All the entries annotated as 'fragment' were discarded. Indeed, as shown below, we think that the main interest of PHYTOPROT lies in its ability to track erroneous genomic annotations. This makes sense only when full-length proteins—hence full-length genes—are compared. The resulting 43 754 sequences were submitted to an all-by-all comparison with the Biofacet software from Gene-IT (12) on a Sun E10000 computer with 48 processors, located at Infobiogen (http://www.infobiogen.fr). All the pairwise comparisons were performed using the Smith–Waterman algorithm (13) with the Z-value being used as the index of similarity (14). The reason for this choice as compared with a more classical BLAST comparison (15) is 4-fold: (i) the Z-value associated with a pair of sequences is totally independent of the size of the data bank, which is not the case for the BLAST E-values; (ii) Z-values are less dependent on the lengths of the sequences than alignment scores (14); (iii) as shown by Comet *et al.* (14) Z-values >8 most probably point to related sequences, thus providing a conservative estimation of the cut-off between 'random' and 'real' sequences; (iv) local Smith–Waterman alignments (13) were preferred to global Needleman–Wunsch alignments (16) because the latter are too often grossly erroneous when the overall similarity between sequences is weak, thus missing the biologically significant short segments of higher similarity. Obviously, however, this is much more demanding in CPU resources than BLAST comparisons. The clusters were built from pairs of sequences with $Z > 14$ using a locally developed algorithm based on the search of maximal cliques (17) that prevents the chain effect resulting from multidomain proteins (a consequence of this algorithm is that

*To whom correspondence should be addressed. Tel: +33 1 01 60 87 38 67; Fax: +33 1 01 60 87 38 97; Email: risler@genopole.cnrs.fr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

 

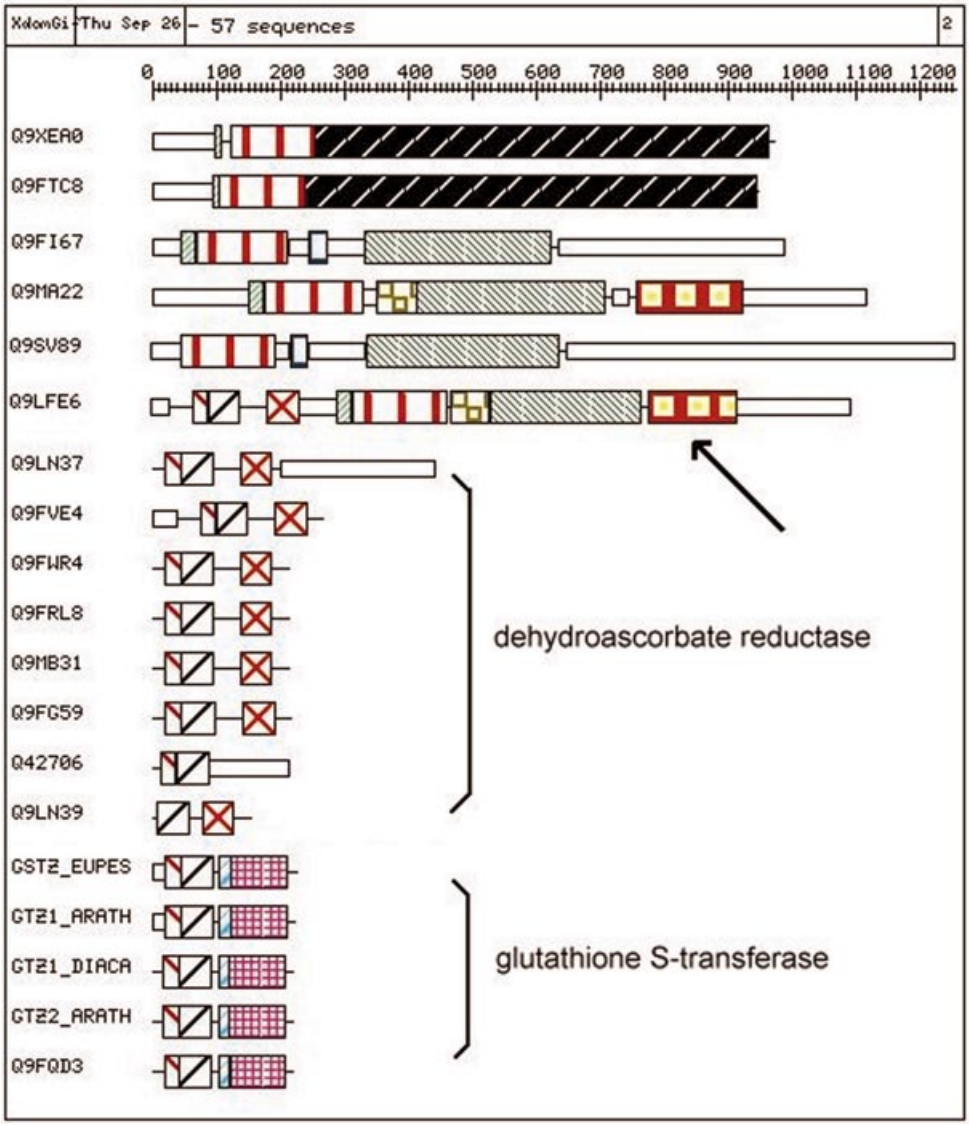


**Figure 1.** Part of the XDOM output for cluster 435 (edited for the purpose of clarity). The cluster was retrieved by using the word 'leucyl' as query. The top six proteins are aminoacyl-tRNA synthetases belonging to the class-I subfamily, namely leucyl-, isoleucyl- and valyl-tRNA synthetases. In the same cluster are found much shorter proteins such as dehydroascorbate reductases and glutathione *S*-transferases, which is surprising. Examination of the drawing shows that protein Q9LFE6 (arrow) is responsible for this grouping: it possesses one domain shared by the five other tRNA synthetases, and other domains shared by the shorter proteins. In this particular case, it is highly probable that the 'gene' corresponding to Q9LFE6 results in fact from an artificial and erroneous fusion between one gene encoding a dehydroascorbate reductase and one gene encoding an aminoacyl-tRNA synthetase. It is clear also that some predicted genes encoding certain putative dehydroascorbate reductases (Q9LN37, Q9LN39, etc.) are worth re-examination.

a given sequence may belong to two or more clusters). This resulted in 4053 clusters containing from 2 to 1788 proteins and 5185 singletons. As already noted (18) the largest cluster(s) are built up mainly by kinases. For each of the 3982 clusters comprising <500 members, the 'domains' of similarity shared by two or more sequences were calculated and displayed with the program XDOM (19). For each of the 3913 clusters comprising <255 members, a pyramidal classification was calculated and displayed. As shown by Aude *et al.* (20) the pyramidal representation can prove to be useful in delineating subfamilies and in pinpointing those sequences that make the link between subfamilies. The XDOM and pyramidal representations were not calculated for the largest clusters for two reasons: (i) the CPU requirements become prohibitive, and (ii) the graphical representations become so large that they are of very limited practical use. Finally, the composition of the clusters, their pyramidal classifications and their decompositions into domains were stored in a relational database (Oracle).

## THE PHYTOPROT WEBSITE

Any protein (or group of proteins) can be searched through its Swiss-Prot/TrEMBL ID or AC, or through words that appear in the description line (DE) such as cytochrome P450 or lactate dehydrogenase. The result of the query consists in the list of the cluster(s) containing one or more proteins that matched the query, the size of the cluster(s) and their

identifiers. This first step prevents the involuntary display of the largest clusters. The content of one particular cluster is then displayed by selecting its identifier [an alternative way to obtain the content of one cluster (or family) is to enter its number directly in the query form]. All the proteins belonging to the selected cluster are displayed with their ID/AC, the description and keyword lines in their entry and the organism they come from. Any Swiss-Prot/TrEMBL entry can be retrieved through a wgetz (SRS) call by selecting its AC. Two buttons are of particular interest here: 'View Pyramid' and 'View XDOM'. The first enables the display of the pyramidal classification of the proteins within the cluster while the second gives access to the graphical representation of the domain arrangements in the family. An example of an XDOM display is depicted in Figure 1, which shows how some probably erroneous genomic annotations can be easily detected. Finally, the PHYTOPROT interface allows the launch of a BLAST search of a query sequence against all those in the database. Upon completion, the program will return the cluster(s) where one or more hits were observed (E-value < $10^{-6}$), the sequence alignments and the pyramidal classification of the cluster(s) to which the query sequence was added (due to CPU constraints, the pyramidal classification is recalculated only for those clusters containing <250 proteins).

## UPDATES OF PHYTOPROT

Before the end of 2003 a new set of comparisons will be added to PHYTOPROT. It will consist of the *A.thaliana* proteome (released by the TIGR Institute in July 2002) compared against itself, which should be useful for the study of the numerous multigenic families in this plant. Another all-by-all comparison of all the proteins from plants, comprising the *A.thaliana* proteome and more than 32 000 sequences from other plants is well under way.

## ACKNOWLEDGEMENTS

We are indebted to E. Viara, L. Pereira and Génoplante-Info for their help in setting up the Oracle database, to Infobiogen for help in using its multi-threaded machine and to J. Gouzy for his generous gift of the XDOM program. This work was sponsored by Génoplante, contracts Bi1999051 and Bi2001071, and by Ministère de l'Industrie (ASG 124, No. 01 4 90 6093).

## REFERENCES

1. Sasson,O., Vaaknin,A., Fleischer,H., Portugaly,E., Bilu,Y., Linial,N. and Linial,M. (2003) ProtoNet: hierarchical classification of the protein space. *Nucleic Acids Res.*, **31**, 348–352.
2. Perrière,G., Duret,L. and Gouy,M. (2000) HOBACGEN: database system for comparative genomics in bacteria. *Genome Res.*, **10**, 379–385.
3. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
4. Kriventseva,E.V., Servant,F. and Apweiler,R. (2003) Improvements to CluSTr: the database of SWISS-PROT+TrEMBL protein clusters. *Nucleic Acids Res.*, **31**, 388–389.
5. Krause,A., Haas,S.A., Coward,E. and Vingron,M. (2002) SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. *Nucleic Acids Res.*, **30**, 299–300.
6. Servant,F., Bru,C., Carrere,S., Courcelle,E., Gouzy,J., Peyruc,D.and Kahn,D. (2002) ProDom: automated clustering of homologous domains. *Brief. Bioinform.*, **3**, 246–251.
7. Frishman,D., Mokrejs,M., Kosykh,D., Kastenmuller,G., Kolesov,G., Zubrzycki,I., Gruber,C., Geier,B., Kaps,A., Albermann,K. *et al.* (2003) The PEDANT genome database. *Nucleic Acids Res.*, **31**, 207–211.
8. Hoersch,S., Leroy,C., Brown,N.P., Andrade,M.A. and Sander,C. (2000) The GeneQuiz web server: protein functional analysis through the Web. *Trends Biochem. Sci.*, **25**, 33–35.
9. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
10. Louis,A., Ollivier,E., Aude,J.-C. and Risler,J.-L. (2001) Massive sequence comparisons as a help in annotating genomic sequences. *Genome Res.*, **11**, 1296–1303.
11. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370
12. Glémet,E. and Codani,J.-J. (1997) LASSAP, a large scale sequence comparison package. *Comput. Appl. Biosci.*, **13**, 137–143.
13. Smith,T.F. and Waterman,M.S. (1981) Comparison of biosequences. *Adv. Appl. Math.*, **2**, 482–489.
14. Comet,J.-P., Aude,J.-C., Glémet,E., Risler,J.-L., Hénaut,A. and Slonimski,P.P. (1999) Significance of Z-value statistics of Smith–Waterman scores for protein alignments. *Comput. Chem.*, **23**, 317–331.
15. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang, J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
16. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.*, **48**, 443–453.
17. Mohseni-Zadeh,S., Brézellec,P. and Risler,J.-L. (2003) Large-scale clustering of proteins with sequence similarity based on the extraction of maximal cliques. In Spang,R., Béziat,P. and Vingron,M. (eds), *Currents in Computational Molecular Biology*, RECOMB 2003, pp. 73–74. (http://recomb2003.molgen.mpg.de:9090/poster-02-039/)
18. Yona,G., Linial,N. and Linial,M. (1999) ProtoMap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins*, **37**, 360–378.
19. Gouzy,J., Eugène,P., Greene,E.A., Kahn,D. and Corpet,F. (1997) XDOM, a graphical tool to analyse domain arrangements in protein families. *Comput. Appl. Biosci.*, **13**, 601–608.
20. Aude,J.-C., Diaz-Lazcoz,Y., Codani,J.-J. and Risler,J.-L. (1999) Applications of the pyramidal clustering method to biological objects. *Comput. Chem.*, **23**, 303–315.