# Mendel-GFDb and Mendel-ESTS: databases of plant gene families and ESTs annotated with gene family numbers and gene family names

**David Lonsdale*, Mark Crowe, Benjamin Arnold and Benedict C. Arnold**

Mendel Bioinformatics Group, John Innes Centre, Colney, Norwich NR4 7UH, UK

Received September 1, 2000; Revised and Accepted October 30, 2000

## ABSTRACT

**There is no control over the information provided with sequences when they are deposited in the sequence databases. Consequently mistakes can seed the incorrect annotation of other sequences. Grouping genes into families and applying controlled annotation overcomes the problems of incorrect annotation associated with individual sequences. Two databases (http://www.mendel.ac.uk) were created to apply controlled annotation to plant genes and plant ESTs: Mendel-GFDb is a database of plant protein (gene) families based on gapped-BLAST analysis of all sequences in the SWISS-PROT family of databases. Sequences are aligned (ClustalW) and identical and similar residues shaded. The families are visually curated to ensure that one or more criteria, for example overall relatedness and/or domain similarity relate all sequences within a family. Sequence families are assigned a 'Gene Family Number' and a unified description is developed which best describes the family and its members. If authority exists the gene family is assigned a 'Gene Family Name'. This information is placed in Mendel-GFDb. Mendel-ESTS is primarily a database of plant ESTs, which have been compared to Mendel-GFDb, completely sequenced genomes and domain databases. This approach associated ESTs with individual sequences and the controlled annotation of gene families and protein domains; the information being placed in Mendel-ESTS. The controlled annotation applied to genes and ESTs provides a basis from which a plant transcription database can be developed.**

## INTRODUCTION

The rapid expansion of gene sequencing has led to an exponential increase in the number of genes being deposited in the sequence databases. Associated with this is the proliferation of idiosyncratic gene names or gene sequence identifiers which provide no useful or meaningful information. In addition the description field of sequence records, in closely related sequences, whether they are from the same or different species, are often inconsistent and can sometimes be incorrect and misleading (1; http://www.bioinfo.de/isb/1998/01/0007). As a consequence and in the absence of any correction or the application of controlled annotation, incorrect descriptions in the sequence databases can seed further incorrectly annotated sequences (1,2). In 1992, the International Society for Plant Molecular Biology initiated an effort to standardise gene nomenclature in higher plants (3). This gave rise to the development of Mendel-GFDb, a database of plant gene families. Sequences which are substantially similar were grouped together. These groups or families contain sequences, which though not necessarily functionally equivalent, provide a basis for the application of controlled annotation. As a consequence, substantially similar sequences or those sharing domains which fall into the same BLAST family are given the same gene family number across all photosynthetic eukaryotes and members of the same gene family within any species are distinguished by their database sequence accession number.

The application of controlled annotation to genes and gene families could also be applied to expressed sequence tags (ESTs). This led to the development of the Mendel-ESTS database, a database of public domain ESTs which have been compared to Mendel-GFDb to identify the genes from which they are derived. In the absence of the cognate gene ESTs can still be associated with gene families based on sequence similarity. Applying gene and gene family annotation to ESTs unifies nomenclature within and across species. This unification of annotation of genes and their products will permit the physical maps of genomes to be compared in a meaningful way.

## THE MENDEL-GFDB DATABASE

Plant sequences, *Viridiplantae*, in the SWISS-PROT family of databases (4,5) are downloaded and those that are <30 amino acids in length are removed. Sequences from different species, which have been merged into a single SWISS-PROT record are demerged. The entire sequence set is then converted to FASTA format and compared to itself using unfiltered gapped-BLAST v2.0 (6,7). Sequences with a pLog value $\geq 1 \times 10^{-11}$ are removed from the BLAST output files and the sequences which form the BLAST family, having pLog values $< 1 \times 10^{-11}$ are removed from the dataset prior to the next BLAST analysis. This reductive BLAST analysis ensures that

*To whom correspondence should be addressed at: Cereals Research Department, John Innes Centre, Colney, Norwich NR4 7UH, UK. Tel: +44 1603 450012; Fax: +44 1603 450012; Email: lonsdale@bbsrc.ac.uk sequences are only ever associated with one family and generates a non-redundant set of BLAST families, a proportion of which contain only one sequence. Finally, BLAST families, which contain two sequences or more, are aligned using ClustalW (8) and similar and identical amino acids are identified by differential shading using BOXSHADE (http://www.ch.embnet.org/software/BOX_faq.html). The aligned and shaded BLAST families are then visually inspected to determine whether or not the sequences represent a single family or whether there are two or more groups which can be differentiated by clear and definable sequence differences. Following the removal of any rogue sequences, the groups or families are assigned a unique numeric identifier, the 'Gene Family Number'. Their Mendel database accession number identifies individual members of the family and the SWISS-PROT five character species abbreviation identifies the species from which the gene was isolated. For example, Arath;105;7148 (sp. abv; gene family number; accession no.) represents the sequence associated with SWISS-PROT record P25858 which is a cytosolic NAD-dependent glyceraldehyde-3-phosphate dehydrogenase from *Arabidopsis thaliana*. If authority exists the gene family number can be replaced by a gene family name or gene family abbreviation. In the case of P25858, GapC is the accepted abbreviation. All of this information is placed in Mendel-GFDb. Associated with each Mendel record is information extracted from the sequence records: SWISS-PROT accession number and DE field; gene names/synonyms, EMBL accession numbers; Prosite and Prodoc (9), Pfam (10), DOMO (11), TRANSFAC (12) and MEDLINE information. At the level of the gene family, a unified description is applied. This is an attempt to condense the individual sequence record DE fields, within a family, into a meaningful description which best represents all the sequences in the gene family. It also attempts to control the order of words in the description, so that different families which encode functionally related sequences can be grouped together in spread sheet applications; e.g. ribosomal protein, DNA binding, Photosystem etc. In the absence of any meaningful information, the term 'Unknown function' is applied except in cases of families where a DOMO, Prosite or Pfam information is available. In such instances the domain description is applied, e.g. zf_C2H2 domain. The database can be searched by any of the terms but is best accessed via the BLAST server. For example, a sequence related to gene family 105 (GapC) will identify all sequences in the database which are GapC related. From gene family page, [*ACFTP*] lists all accession numbers of the sequences in the gene family, [*SEQFTP*] returns the FASTA file of the gene family sequences and [*MENDEL- ESTS*] returns all the ESTs which are related to GapC and information relating to those ESTs: EST accession number, species from which EST derived, SWISS-PROT accession number of BLASTX best hit, its Mendel gene name and its relative confidence value (RCV) score (see below, percentage identity), finally the domain associated with the EST; its number, RCV score and domain name.

## THE MENDEL-ESTS DATABASE

Mendel-ESTS is a database of plant EST sequences that have been compared to Mendel-GFDb and other sequence databases, currently including the protein domain database DOMO. Each EST sequence is associated with the following information: its sequence accession number; the species from which the EST was derived and EST library description; the accession numbers and DE field associated with the best-hit from the SWISS-PROT family of databases, Mendel-GFDb, DOMO and a selection of sequenced genomes. The high score and pLog values of the best-hit as well as the self BLAST values for each EST, generated using TBLASTX, are given as well as the coding frame of the EST. All the information relating to any EST library or EST can be recovered in tab-delimited format

The ratio of the best-hit high score to the self-high score of the EST is given for each BLAST analysis. The value is given the terminology 'RCV' and values fall in the range of 0.0 (no homology) to 1.00 (100% homology/identity). For example *Arabidopsis* EST, T88131, identifies SWISS-PROT record P52780, (Luplu;2315;13115, Mendel gene family 2315) with an RCV of 0.68, DOMO domain DM01455 {RCV 0.43} and the *Escherichia coli* gene *glnS* {RCV 0.43}. All results suggest that this EST is derived an *Arabidopsis* gene encoding glutamyl tRNA synthetase even though its best hit is against a Lupin homolog and that reading frame 3 is the coding strand of the EST. All ESTs and associated information, in tab-delimited format, with homology to this plant gene, gene family, to the protein domain or to the *E.coli* homologue is hyperlinked via the appropriate accession number. This approach associates individual ESTs with the controlled annotation of genes and gene families in the Mendel-GFDb database and with the annotation of the domain database, DOMO.

This method of analysis of ESTs compares well to data generated by more rigorous and time-consuming methodology that requires the use of clean data sets and contig analysis (13,14; http://www.bioinfo.de/isb/1999/01/0018). By comparing ESTs to databases such as Mendel-GFDb and DOMO, both of which use controlled annotation, suppression of poor or misleading annotation associated with individual sequence records is achieved. It permits clustering and analysis of ESTs by a number of different criteria placing ESTs into a framework from which function can be predicted.

## CITATION AND AVAILABILITY

Users are requested to cite this article and the database including the version number: Mendel-GFDb (v7.0), Mendel Bioinformatics Group, John Innes Centre, Norwich NR4 7UH, UK. The databases and associated information files are freely available to users from non-profit organisations. Users from commercial organisations are requested to contact the database manager (corresponding author) for further information. Mirror sites are at UK CROPNET (http://ukcrop.net/) (15) and at USDA-ARS Center for Bioinformatics and Comparative Genomics, Cornell University (http://genome.cornell.edu/).

## FUTURE PERSPECTIVES

Current funding will allow the databases to be updated more regularly, the aim being to provide a major update every 3 months. Future developments will aim to expand and control the annotation of genes and ESTs. The expression profiles of genes based on published data, EST libraries and micro-array data will be incorporated. This information will be interfaced with an electronic plant, the life cycle of which will be described by a hierarchical annotation describing organ, tissue and cell development within a temporal framework of development. This will effectively create a plant transcriptome database.

## ACKNOWLEDGEMENTS

We would like to thank Virginia Barnard for designing the web pages. B.A. and B.C.A. and the development of Mendel-GFDb were funded by a European Community Framework IVgrant (BIO4 CT950247) and M.C. by a BBSRC grant (IGF/124242). The group would also like to acknowledge support from the John Innes Centre.

## REFERENCES

1. Galperin,M.Y. and Koonin,E.V. (1998) Sources of systematic error in functional annotation of genomes: domain arrangement, non-orthologous gene displacement, and operon disruption. *In Silico Biol.*, **1**, 55–67.
2. Lonsdale,D.M. (1999) Nomenclature regulation. *Nature*, **39**, 118
3. Lonsdale,D.M., Price,C.A. and Reardon,E.M. (1996) A guide to naming sequenced plant genes. *Plant Mol. Biol.*, **30**, 225–227.
4. Bairoch,A. and Apweiler,R. (1996) The SWISS-PROT protein sequence databank and its new supplement TrEMBL. *Nucleic Acids Res.*, **24**, 21–25.
5. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48
6. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
7. Altschul,S.F., Madden,T.L., Schaffer,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 programmes. *Nucleic Acids Res.*, **25**, 3389–3402.
8. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
9. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
10. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.
11. Gracy,J. and Argos,P. (1998) Automated protein database classification: I. Integration of compositional similarity search, local similarity search and multiple sequence alignment. II. Delineation of domain boundaries from sequence similarities. *Bioinformatics*, **14**, 164–187.
12. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Mienhardt,T., Pruss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 281–283.
13. Rounsley,S.D., Glodek,A., Sutton,G., Adams,M.D., Somerville,C.R., Venter,J.C. and Kerlavage,A.R. (1996) The construction of Arabidopsis expressed sequence tag assemblies. A new resource to facilitate gene identification. *Plant Physiol.*, **112**, 1177–1183.
14. Ewing,R., Poirot,O. and Claverie,J.-M. (1999) Comparative analysis of the Arabidopsis and rice expressed sequence tag (EST) sets. *In Silico Biol.*, **1**, 0018. http://www.bioinfo.de/isb/1999/01/0018
15. Dicks,J., Anderson,M., Cardle,L., Cartinhour,S., Couchman,M., Davenport,G., Dickson,J., Gale,M., Marshall,D., May,S., McWilliam,H., O'Malia,A., Ougham,H., Trick,M., Walsh,S. and Waugh,R. (2000) UK CropNet: a collection of databases and bioinformatics resources for crop plant genomics. *Nucleic Acids. Res.*, **28**, 104–107.