*Nucleic Acids Research, 2006, Vol. 34, Database issue* ***D349–D353***
*doi:10.1093/nar/gkj100*

# INVHOGEN: a database of homologous invertebrate genes

**Ingo Paulsen* and Arndt von Haeseler[1,2,3,4]**

Department of Bioinformatics, Institute for Computer Sciences, Heinrich-Heine-University Duesseldorf, Universitaetsstrasse 1, 40225 Duesseldorf, Germany, [1]Center for Integrative Bioinformatics Vienna, Dr Bohr-Gasse 9/6, A-1030 Vienna, Austria, [2]University of Vienna, Vienna, Austria, [3]Medical University of Vienna, Vienna, Austria and [4]University of Veterinary Medicine Vienna, Vienna, Austria

Received August 15, 2005; Revised and Accepted October 17, 2005

## ABSTRACT

**Classification of proteins into families of homologous sequences constitutes the basis of functional analysis or of evolutionary studies. Here we present INVertebrate HOmologous GENes (INVHOGEN), a database combining the available invertebrate protein genes from UniProt (consisting of Swiss-Prot and TrEMBL) into gene families. For each family INVHOGEN provides a multiple protein alignment, a maximum likelihood based phylogenetic tree and taxonomic information about the sequences. It is possible to download the corresponding GenBank flatfiles, the alignment and the tree in Newick format. Sequences and related information have been structured in an ACNUC database under a client/server architecture. Thus, complex selections can be performed. An external graphical tool (FamFetch) allows access to the data to evaluate homology relationships between genes and distinguish orthologous from paralogous sequences. Thus, INVHOGEN complements the well-known HOVERGEN database. The databank is available at http://www.bi.uni-duesseldorf.de/~invhogen/invhogen.html.**

## INTRODUCTION

Genome projects (1) are generating an enormous amount of data in molecular and evolutionary biology. One goal of functional genomics is to determine the function of proteins predicted by these sequencing projects (2). To overcome the problem of assigning protein functions to sequences one approach is to classify them into gene families on the basis of the presence of shared features or by clustering using some similarity measures under the assumption that proteins within the same gene family possess similar or identical biochemical functions. To determine the function of new proteins one can infer its function or detect its functional regions by homology to other sequences. (If two proteins share a significant sequence similarity, then one typically concludes that they are probable to have similar function.) However, there are some cases where conserved structures within a protein group do not necessarily imply that these proteins perform the same function (3) owing to low-complexity sequences, multifunctional sequences and gene recruitment (4).

Gene families are generated using sequence clustering. Sequence clustering allows the detection of all pair-wise sequence similarities within a given set of protein sequences. Proteins are then clustered into families based on their sharing of significant sequence similarity patterns. When sequence clustering is performed accurately, proteins within a family may be considered as sharing a common evolutionary history and possibly similar or identical functions (5).

However, within a gene family one has to distinguish between two types of homologies: genes are said to be orthologues in two different species if gene copies originate from a common ancestral gene after a speciation event. Paralogues are genes in a given species pair that diverged after duplication of an ancestral gene (6). The distinction between paralogy and orthology is essential for molecular phylogeny since it is necessary to work with orthologous genes to infer species phylogeny from gene phylogeny. Because the orthologous genes provide the required protein function, paralogous genes are more free to mutate (mutations are under weaker negative selection), possibly yielding genes with new functions. As a result, paralogous genes are often less similar in sequence to a homologue from another organism than are orthologous genes. However, the issue of (7) and (8) is not of interest for the design of a database of homologous gene families.

To address the problem of detecting homologous genes, we built the INVertebrate HOmologous GENes (INVHOGEN) database. This database complements the three homologous databases HOVERGEN (9) devoted to vertebrates, HOBACGEN (10) devoted to prokaryotes and HOGENOM devoted to

*To whom correspondence should be addressed. Tel: +49 211 81 13716; Fax: +49 211 81 15767; Email: paulseni@uni-duesseldorf.de completely sequenced organisms. INVHOGEN contains the available invertebrate protein sequences from UniProt organized into families of homologous genes defined by sequence similarity. For each family INVHOGEN provides a multiple protein alignment, a maximum likelihood based phylogenetic tree and taxonomic information about the sequences.

## METHODS

The second release of INVHOGEN (August 2005) has been built from the invertebrate entries in UniProt Release 5.5 (July 19, 2005) (11) consisting of SwissProt Release 47.5 and TrEMBL Release 30.5. The data consist of 284 763 protein entries, 11 702 of them from SwissProt and 273 061 from TrEMBL (12). From both sequence files a total of 174 958 invertebrate protein entries were extracted. The Swiss-Prot/TrEMBL protein entries were used owing to their high level of annotation and integration with other databases, and of their minimal level of redundancy. By following the references in the database cross-reference (DR) field of Swiss-Prot/TrEMBL annotations, the corresponding nucleotide sequences from EMBL (13) were also integrated in the database structure. Nucleotide and protein sequences were organized into two separated ACNUC databases (14). INVHOGEN will be updated four times per year with the latest major release of the UniProt Knowledgebase.

For building the families, the BLASTP2 (15) program was applied to identify common regions between proteins, and to collect related proteins. A similarity search of all proteins against each other was performed by filtering low-complexity regions with SEG (16), and using the BLOSUM62 amino acid similarity matrix (17) and an *E*-value threshold of $1 \times 10^{-4}$.

BLAST output was filtered to remove incompatible high-scoring segment pairs (HSPs) within a global alignment (Figure 1). For complete protein sequences, two sequences in a pair were classified as being in the same family if the remaining HSPs covered at least 80% of the protein length and if their similarity was $\geqslant 50\%$ (two amino acids are considered similar if their BLOSUM62 similarity score is positive). This procedure reduces the risk of mis-assigning proteins with a complex evolutionary history involving gene fissions and fusions, and domain shuffling (3). We used simple transitive links to build families. Once families of complete protein sequences were built, partial sequences were included in the classification. A partial sequence matching, a complete protein was included in a family if it fulfilled the two conditions required for a complete sequence and if its length was at least 100 amino acids (18) or at least half the length of the complete protein.

Gene families were named using a program that parsed the sequence description (DE) and similarity comment fields (SIMILARITY) of the Swiss-Prot/TrEMBL annotations. In the first step DE entries were clustered into subgroups of similar word orders. Each subgroup was named by assigning the most frequent position of every word and by joining these words together. A family description was created by combining all subgroup names considering only those with a large number of non-redundant entries in relation to the other subgroups. In the second step particular families were completed by available similarity comment lines for clarification reasons or if subgroup names were too different among themselves. Manual expertise was used to specify the name for a gene family if both attempts failed to generate a meaningful name.

For each gene family with at least four sequences, a multiple sequence alignment and a phylogenetic tree were built. Protein

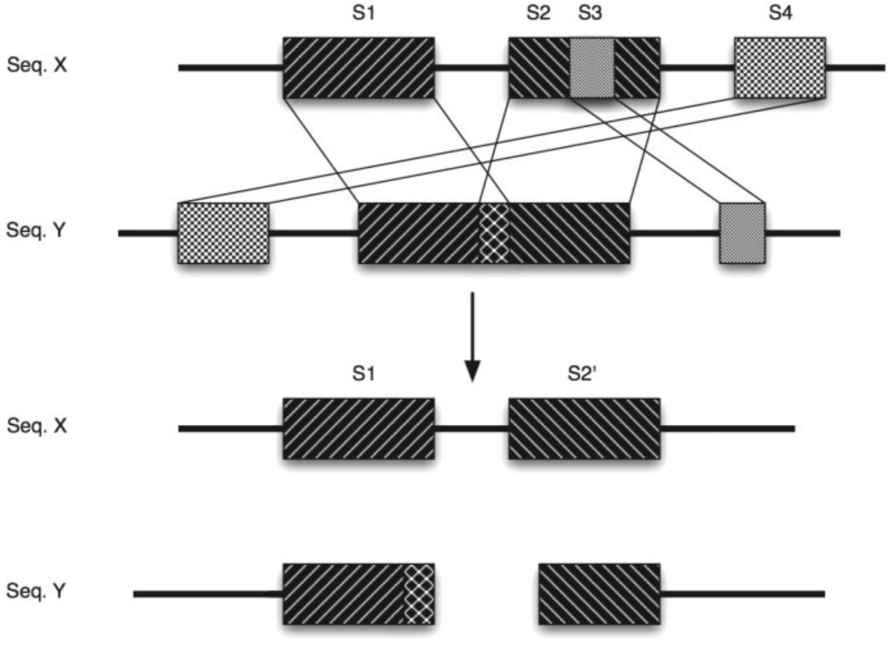


**Figure 1.** Removing incompatible HSPs. For each pair of sequences X and Y that hit each other using BLASTP, HSPs that are not compatible with a global alignment are removed. In this example, hits H1 and H2 are compatible. However H3 and H4 are not compatible. Therefore, only H1 and H2 are considered for further computations on similarity measures. Because H1 and H2 are overlapping, the overlap is allocated to H1 and H2 is shortened accordingly. In a crossing-over situation between H1 and H2 for the sequences X and Y, H1 will be used if length(H1) > length(H2), otherwise, H2 is to take into account.

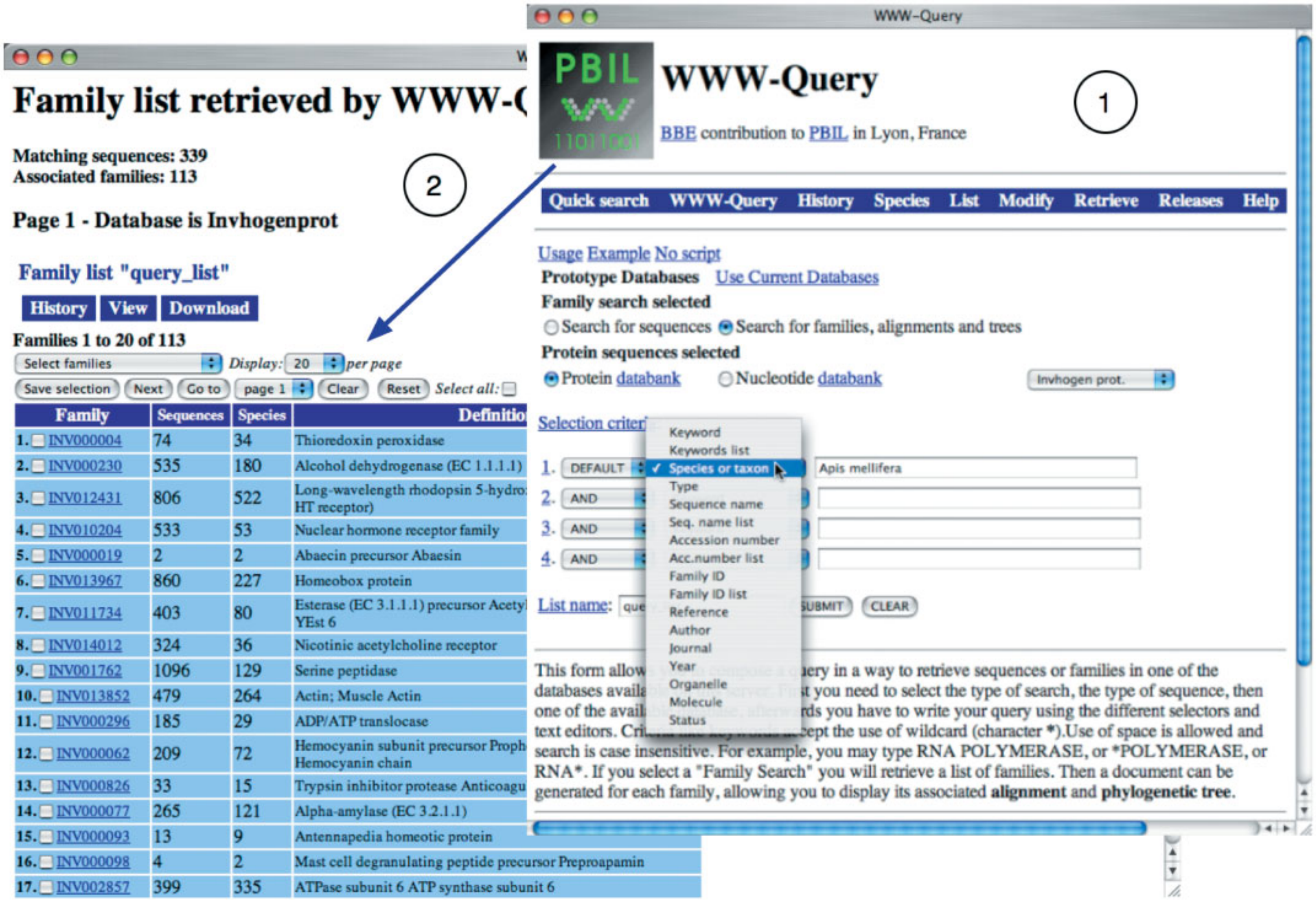
sequences were aligned with CLUSTALW 1.82 (19) using the default parameters. Phylogenetic trees were reconstructed with IQPNNI 2.6 (20) by considering the so-called stopping rule with at most 100 iterations. The stopping rule decides whether it is probable (with a 95% confidence level) that a continuation of the search will lead to no further improvement.

## RESULTS

The INVHOGEN interface is based on a client/server architecture originally developed for HOVERGEN (9) and HOBACGEN (10). To query the database the FamFetch (21) application needs to be installed or one can use the web interface (Figure 2). FamFetch has a graphical interface that allows users to easily access and see the list of the families available in the database, the protein or nucleotide sequences of the genes in the families, the corresponding multiple protein sequence alignments and the maximum likelihood based phylogenetic trees computed with these alignments.

The present version of INVHOGEN contains a total of 174 958 protein sequences (and 159 922 nucleic sequences) classified into 15 389 families. Among all the proteins included in this release 132 556 (75.8%) are classified into 15 389 families containing at least two sequences, and 42 402 (24.2%) partial proteins are not assigned to any family (so-called singletons). Table 1 shows the distribution of families in INVHOGEN grouped by family size in comparison with HOVERGEN. Table 2 displays the 10 largest families for both databases. These families consist of genes coding for proteins (or protein subunits) involved in protein translation, nucleotide biosynthesis, tissue development and glycolysis. Cytochrome *c* oxidase polypeptide I, Cytochrome *b* and NADH dehydrogenase subunit 1 are the only gene families that occur in both databases in the list of the top 10.

Table 3 presents the invertebrate and vertebrate species for which the greatest number of genes have been sequenced. Not surprisingly, species that are completely sequenced (e.g. *Drosophila melanogaster* and *Caenorhabditis elegans*) are the most frequent. They take up 44.5% of 132 556 protein sequences in INVHOGEN and 64.3% of the 214 379 sequences in HOVERGEN. Moreover, the distribution of all 22 053 species in INVHOGEN among all families is non-uniform. The first three species from Table 3 are each over-represented by at least 10 000 occurrences in number of sequences and appearance in families. However, 11 162 species only contribute a total of one sequence (data not shown).



WWW-Query

BBE contribution to PBIL in Lyon, France

Usage Example No script

**Prototype Databases** Use Current Databases

**Family search selected**

○ Search for sequences ● Search for families, alignments and trees

**Protein sequences selected**

● Protein databank ○ Nucleotide databank Invhogen prot.

Selection criteri

1. DEFAULT Species or taxon Apis mellifera
2. AND
3. AND
4. AND

List name: que

This form allow ery in a way to retrieve sequences or families in one of the databases availa t you need to select the type of search, the type of sequence, then one of the avail rds you have to write your query using the different selectors and text editors. Cri ept the use of wildcard (character *).Use of space is allowed and search is case insensitive. For example, you may type RNA POLYMERASE, or *POLYMERASE, or RNA*. If you select a "Family Search" you will retrieve a list of families. Then a document can be generated for each family, allowing you to display its associated **alignment** and **phylogenetic tree**.

Family list retrieved by WWW-(

**Matching sequences: 339**
**Associated families: 113**

**Page 1 - Database is Invhogenprot**

**Family list "query_list"**

**Families 1 to 20 of 113**

Display: 20 per page

Select all: □

| Family | Sequences | Species | Definitio |
|---|---|---|---|
| 1. □ INV000004 | 74 | 34 | Thioredoxin peroxidase |
| 2. □ INV000230 | 535 | 180 | Alcohol dehydrogenase (EC 1.1.1.1) |
| 3. □ INV012431 | 806 | 522 | Long-wavelength rhodopsin 5-hydro HT receptor) |
| 4. □ INV010204 | 533 | 53 | Nuclear hormone receptor family |
| 5. □ INV000019 | 2 | 2 | Abaecin precursor Abaesin |
| 6. □ INV013967 | 860 | 227 | Homeobox protein |
| 7. □ INV011734 | 403 | 80 | Esterase (EC 3.1.1.1) precursor Acetyl YEst 6 |
| 8. □ INV014012 | 324 | 36 | Nicotinic acetylcholine receptor |
| 9. □ INV001762 | 1096 | 129 | Serine peptidase |
| 10. □ INV013852 | 479 | 264 | Actin; Muscle Actin |
| 11. □ INV000296 | 185 | 29 | ADP/ATP translocase |
| 12. □ INV000062 | 209 | 72 | Hemocyanin subunit precursor Proph Hemocyanin chain |
| 13. □ INV000826 | 33 | 15 | Trypsin inhibitor protease Anticoagu |
| 14. □ INV000077 | 265 | 121 | Alpha-amylase (EC 3.2.1.1) |
| 15. □ INV000093 | 13 | 9 | Antennapedia homeotic protein |
| 16. □ INV000098 | 4 | 2 | Mast cell degranulating peptide precursor Preproapamin |
| 17. □ INV002857 | 399 | 335 | ATPase subunit 6 ATP synthase subunit 6 |

**Figure 2.** The web interface for querying gene family databases. Window 1 allows to perform queries for different kinds of search criteria. In this example INVHOGEN is asked to search for all gene families containing *Apis mellifera* (honey bee). The resulting gene families for his query are listened in window 2.

**Table 1.** Distribution of families in INVHOGEN Release 2 and HOVERGEN Release 46

| Family size | No. of families INVHOGEN | | No. of families HOVERGEN | |
|---|---|---|---|---|
| 2 | 8567 | 55.7% | 3219 | 24.5% |
| 3 | 2257 | 14.7% | 1788 | 13.6% |
| 4 | 1210 | 7.8% | 1369 | 10.5% |
| 5–9 | 2093 | 13.6% | 3677 | 28.0% |
| 10–19 | 693 | 4.5% | 1928 | 14.7% |
| 20–49 | 358 | 2.3% | 832 | 6.3% |
| 50–99 | 116 | 0.8% | 182 | 1.4% |
| ⩾100 | 95 | 0.6% | 149 | 1.1% |
| Total | 15 389 | 100% | 13 144 | 100% |

**Table 2.** Ten largest families of INVHOGEN Release 2 and HOVERGEN Release 46

| Family name INVHOGEN | Sequences | | Family name HOVERGEN |
|---|---|---|---|
| Cytochrome *c* oxidase polypeptide I | 22 287 | 22 616 | Cytochrome *b* |
| Cytochrome *c* oxidase polypeptide II | 6192 | 8480 | NADH dehydrogenase subunit 4 |
| Cytochrome *b* | 3229 | 5987 | Family 1 of G-protein-coupled receptors |
| Elongation factor-1α | 3124 | 3608 | Class I histocompatibility antigen |
| NADH dehydrogenase subunit 1 | 1586 | 2990 | ATP synthase subunit 6 |
| NADH dehydrogenase subunit 5 | 1568 | 2291 | ATP synthase subunit 8 |
| WNT family | 1528 | 2090 | Cytochrome *c* oxidase polypeptide I |
| Serine peptidase | 1096 | 1657 | NADH dehydrogenase subunit 1 |
| Homeobox protein | 860 | 1499 | Zinc finger protein |
| Histone H3 | 836 | 1314 | NADH dehydrogenase subunit 6 |
| Total | 42 306 | 52 532 | |

**Table 3.** The top 10 species in INVHOGEN Release 2 and HOVERGEN Release 46

| Species INVHOGEN | Sequences | | Species HOVERGEN |
|---|---|---|---|
| *Drosophila melanogaster*[a] | 17 348 | 56 932 | *Homo sapiens*[a] |
| *C.elegans*[a] | 16 604 | 46 693 | *Mus musculus*[a] |
| *C.briggsae*[a] | 10 704 | 9066 | *Rattus norvegicus*[a] |
| *Anopheles gambiae* PEST[a] | 8423 | 7577 | *Danio rerio*[a] |
| *Schistosoma japonicum* | 2143 | 5392 | *Xenopus laevis*[a] |
| *Drosophila simulans* | 998 | 3258 | *Gallus gallus* |
| *Anopheles gambiae*[a] | 894 | 3038 | *Bos taurus* |
| *Bombyx mori*[a] | 689 | 2790 | *Sus scrofa* |
| *Drosophila yakuba* | 608 | 1720 | *Macaca fascicularis* |
| *Ixodes scapularis* | 538 | 1325 | *Oryctolagus cuniculus* |
| Total | 58 949 | 137 791 | |

[a]The organisms where the complete genomic sequence is published (Genomes OnLine Database, August 11, 2005)

**Table 4.** Distribution of the main classified invertebrate groups in INVHOGEN Release 2 and from the literature (20)

| Invertebrate groups | Species/fraction from literature | | Species/fraction in INVHOGEN | | Sequences/ fraction in INVHOGEN | |
|---|---|---|---|---|---|---|
| Arthropods | 900 000 | 85.86% | 16 681 | 77% | 81 896 | 62.36% |
| Urochordates | 3000 | 0.29% | 65 | 0.30% | 910 | 0.69% |
| Echinoderms | 7000 | 0.67% | 326 | 1.50% | 2718 | 2.07% |
| Poriferans | 9000 | 0.86% | 112 | 0.52% | 398 | 0.30% |
| Nematodes | 15 000 | 1.43% | 348 | 1.61% | 29 630 | 22.56% |
| Platyhelminths | 20 000 | 1.91% | 369 | 1.70% | 4296 | 3.27% |
| Cnidarians | 9000 | 0.86% | 448 | 2.07% | 1629 | 1.24% |
| Molluscs | 70 000 | 6.68% | 2930 | 13.52% | 8088 | 6.16% |
| Annelids | 15 000 | 1.43% | 369 | 1.70% | 1041 | 0.79% |
| Hemichordates | 100 | 0.01% | 3 | 0.01% | 74 | 0.06% |
| Cephalochordates | 25 | 0% | 8 | 0.04% | 608 | 0.46% |
| Ctenophorans | 150 | 0.01% | 6 | 0.03% | 31 | 0.02% |
| Total | 1 048 275 | 100% | 21 665 | 100% | 131 319 | 100% |

The percentages of different classified species in the 12 main invertebrate groups and their representation in INVHOGEN are reported in Table 4. It is remarkable that the proportions of molluscs, echinoderms and cnidarians in INVHOGEN are at least twice higher than the proportions reported in the literature (22). On closer examination, the proportion of sequences in INVHOGEN for nematode sequences (22.56%) is disproportionately high—owing to the completety sequenced genomes of *C.elegans* and *Caenorhabditis briggsae*—in comparison with the relative abundance of nematode species reported in the literature (1.43%) and in INVHOGEN (1.61%), respectively.

## DISCUSSION

INVHOGEN allows rapid selection of gene families according to various criteria. First, one can select homologous sequences for a user-defined set of taxa. The colour graphical interface provides easy access to all the data (multiple alignments, phylogenetic trees, taxonomic data and sequence annotations) required to interpret homology relationships between genes and thus to distinguish orthologues from paralogues. Thus, INVHOGEN is a useful tool for comparative genomics, phylogeny or molecular evolutionary studies for invertebrates.

In the process of analysing animal phylogenetic relationships, we have extended the approach used to structure the available vertebrate sequence data in a database (HOVERGEN) to collect all available invertebrate sequences. This work shows that under the assumption of 1.1 million known animals (97% of them are invertebrates) (23) only a small number of invertebrate species have proteins sequenced—and within these species, a dozen contribute the majority of the invertebrate sequences to this database. INVHOGEN has been built in the same way as HOVERGEN (to have a starting point for comparative analysis). Thus, in the near future it will be easily possible to merge both collections into a database for homologous gene families for all known metazoans.

However, we also note, that further work is needed to define homologous gene families. Different approaches exist that have not yet been fully exploited for the applications suggested here. Some approaches classify proteins into families using structural similarities (24) [structures available in PDB (25)], or grouping them into families on the basis of the presence of shared domains or similar domain architecture (26) using domain databases like Prodom (27), Pfam (28) and InterPro (29).

Apart from classification methods based on sequence alignments and motifs, statistical learning methods applying support vector machines (SVM) (30) are useful for classifying diverse protein sequences. SVM and related approaches will complement sequence similarity and clustering methods. Another approach is adopted by ontology-driven systems to build families of specific proteins (31). Ontologies are also useful for pre-processing BLAST searches presenting a weighted list of Gene Ontology (32) terms associated with similar sequences to give information about potential functions of unknown proteins (33).

However, it remains to be seen how approaches like OntoBlast can be utilized to reconstruct more reliable gene families. We hope that more sophisticated algorithms using all available methods will substantially reduce the number of gene families with only very few members (Table 1). Additionally the discrepancy between few often sequenced species and many infrequent sequenced species should be kept in mind when generating gene families. Moreover sequence sampling is biased towards a few model organisms. This may also be the reason for a lot of gene families with few species. Thus for a better understanding of the evolution of gene families one should sequence genes from a wide variety of taxa and not only from a few well-known model organisms.

## ACKNOWLEDGEMENTS

We would like to thank Jochen Kohl, Dominic Mainz, Le Sy Vinh and Jennifer A. Garner for valuable help and fruitful discussions. We especially thank Laurent Duret and Simon Penel for integrating INVHOGEN in the ACNUC database structure on their server at Pôle Bioinformatique Lyonnais (PBIL) and for making it accessible via the World Wide Web. We also thank them for many helpful discussions. We also thank the referees for valuable hints. This work was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG: SFB-TR1). Funding to pay the Open Access publication charges for this article was provided by the Deutsche Forschungsgemeinschaft (DFG).

*Conflict of interest statement.* None declared.

## REFERENCES

1. Bernal,A., Ear,U. and Kyrpides,N. (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.*, **29**, 126–127.
2. Eisenberg,D., Marcotte,E.M., Xenarios,I. and Yeates,T.O. (2000) Protein function in the post-genomic era. *Nature*, **405**, 823–826.
3. Henikoff,S., Greene,E.A., Pietrokovski,S., Bork,P., Attwood,T.K. and Hood,L. (1997) Gene families: the taxonomy of protein paralogs and chimeras. *Science*, **278**, 609–614.
4. Orengo,C.A., Todd,A.E. and Thornton,J.M. (1999) From protein structure to function. *Curr. Opin. Struct. Biol.*, **9**, 374–382.
5. Heger,A. and Holm,L. (2000) Towards a covering set of protein family profiles. *Prog. Biophys. Mol. Biol.*, **73**, 321–337.
6. Fitch,W.M. and Margoliash,E. (1970) The usefulness of amino acid and nucleotide sequences in evolutionary studies. *Evol. Biol.*, **2**, 67–109.
7. Lynch,M., O'Hely,M., Walsh,B. and Force,A. (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics*, **159**, 1789–1804.
8. He,X. and Zhang,J. (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, **169**, 1157–1164.
9. Duret,L., Perrière,G. and Gouy,M. (1999) HOVERGEN: database and software for comparative analysis of homologous vertebrate genes. In Letovsky,S. (ed.), *Bioinformatics and Systems*. Kluwer Academic Publishers, Boston, pp. 13–29.
10. Perrière,G., Duret,L. and Gouy,M. (2000) HOBACGEN: database system for comparative genomics in bacteria. *Genome Res.*, **10**, 379–385.
11. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
12. 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.
13. Kanz,C., Aldebert,P., Althorpe,N., Baker,W., Baldwin,A., Bates,K., Browne,P., van den Broek,A., Castro,M., Cochrane,G. *et al.* (2005) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **33**, D29–D33.
14. Gouy,M., Gautier,C., Attimonelli,M., Lanave,C. and di Paola,G. (1985) ACNUC–a portable retrieval system for nucleic acid sequence databases: Logical and physical designs and usage. *Comput. Appl. Biosci.*, **1**, 167–172.
15. 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 programs. *Nucleic Acids Res.*, **25**, 3389–3402.
16. Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.*, **266**, 554–571.
17. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
18. Nei,M. (1996) Phylogenetic analysis in molecular evolutionary genetics. *Annu. Rev. Genet.*, **30**, 371–403.
19. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
20. Le Sy,V. and von Haeseler,A. (2004) IQPNNI: moving fast through tree space and stopping in time. *Mol. Biol. Evol.*, **21**, 1565–1571.
21. Dufayard,J.F., Duret,L., Penel,S., Gouy,M., Rechenmann,F. and Perrière,G. (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, **21**, 2595–2603.
22. Hedges,S.B. (2002) The origin and evolution of model organisms. *Nature Genet.*, **3**, 838–849.
23. May,R.M. (2000) The Dimensions of Life on Earth. In Raven,P.H. (ed.) *Nature and Human Society: The Quest for a Sustainable World*, Chapter 1 Defining Biodiversity. The National Academy of Sciences, Washington DC, pp. 30–45.
24. Holm,L. and Sander,S. (1996) Mapping the protein universe. *Science*, **273**, 595–602.
25. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
26. Geer,L.Y., Domrachev,M., Lipman,D.J. and Bryant,S.H. (2002) CDART: protein homology by domain architecture. *Genome Res.*, **12**, 1619–1623.
27. 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.
28. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
29. 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) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
30. Burges,C.J.C. (1998) A tutorial on Support Vector Machine for pattern recognition. *Data Min. Knowl. Disc*, **2**, 121–167.
31. Wolstencroft,K., McEntire,R., Stevens,R., Tabernero,L. and Brass,A. (2005) Constructing ontology-driven protein family databases. *Bioinformatics*, **21**, 1685–1692.
32. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
33. Zehetner,G. (2003) OntoBlast function: from sequence similarities directly to potential functional annotations by ontology terms. *Nucleic Acids Res.*, **31**, 3799–3803.