Nucleic Acids Research, 2006, Vol. 34, Database issue ***D327–D331***
doi:10.1093/nar/gkj087

PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees

Simon Whelan*, Paul I. W. de Bakker[1], Emmanuel Quevillon, Nicolas Rodriguez and Nick Goldman

EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK and [1]Department of Molecular Biology, Massachussetts General Hospital, Boston, MA 02114, USA

Received September 15, 2005; Revised and Accepted October 12, 2005

ABSTRACT

PANDIT is a database of homologous sequence alignments accompanied by estimates of their corresponding phylogenetic trees. It provides a valuable resource to those studying phylogenetic methodology and the evolution of coding-DNA and protein sequences. Currently in version 17.0, PANDIT comprises 7738 families of homologous protein domains; for each family, DNA and corresponding amino acid sequence multiple alignments are available together with high quality phylogenetic tree estimates. Recent improvements include expanded methods for phylogenetic tree inference, assessment of alignment quality and a redesigned web interface, available at the URL http://www.ebi.ac.uk/goldman-srv/pandit.

INTRODUCTION

Molecular evolutionary and phylogenetic studies examine patterns of sequence change, observed over large timescales, to investigate the selective and mutational forces acting upon biological sequences (1–4). These forces are informative about how natural selection has acted to shape and adapt molecules to perform specific functions (2,4), and how genetic variation, the raw material that selection acts upon, is generated from mutational processes acting upon the DNA (5,6). There is still much to learn about these forces and progress is dependent on the data and methodology available for investigating them. The evolutionary relatedness of data is often overlooked when it is processed for storage in traditional sequence databases, making large-scale evolutionary analyses difficult. The PANDIT database (7), which is based upon Pfam (8), is designed to assist researchers investigating molecular evolutionary phenomena and developing phylogenetic and comparative genomic methodology. It provides a large database of aligned protein coding sequences, both at the DNA and amino acid level, and good estimates of the evolutionary trees describing the relationship between these sequences. The structure and contents of PANDIT are intended to complement other existing evolution-based databases [e.g. HOBACGEN (9), HOMSTRAD (10), HOVERGEN (11) and TREEBASE (12)]. Its breadth of evolutionary time and the sheer volume of its contents offer researchers a potentially powerful resource.

Molecular evolutionary studies are crucially dependent on the quality of sequence data. During the construction of PANDIT we have taken steps to ensure a high level of sequence and alignment quality. First, all alignments in PANDIT are based on the Pfam-A seed alignments, which are manually curated and therefore comparable with the carefully crafted alignments used to study molecular evolution. Basing PANDIT on Pfam-A seed alignments also has other appealing properties, including the constraint that the structural domains contained within it do not overlap, eliminating redundancy in the database. The broad range of evolutionary distances in each alignment, induced by the procedure used to select sequences for the Pfam-A seed alignment, ensures that the evolutionary trees estimated from them are rarely dominated by closely related or identical sequences, and that many different species are represented. The extensive coverage of known protein domains by the Pfam database means that general conclusions can be drawn about molecular evolution and the inferential methodology associated with it.

Second, new to version 17.0, the columns in alignments are labeled with a measure of their reliability. This quality indicator enables researchers whose methods may suffer a systematic bias resulting from alignment error to exclude problematic regions from their analyses. For example, when examining the selective constraints acting on individual protein sites (2,13), misaligned sequences may be falsely inferred to have undergone adaptive evolution.

*To whom correspondence should be addressed. Tel: +44 1223 494655; Fax: +44 1223 494468; Email: simon@ebi.ac.uk
Present address:
Emmanuel Quevillon, URGI-INRA, 523 Place des Terrasses, 91000, Evry, France

The quality of the phylogenetic tree estimate is also an important consideration for molecular evolutionary studies and can have a direct impact on the biological conclusions drawn from data. The first version of PANDIT (6.2) used only the neighbor-joining clustering algorithm (14) for tree reconstruction. Recent PANDIT releases have improved on this by using several popular methods to estimate evolutionary trees and choosing the best estimate to include in the database, based on its statistical fit to the data. The PANDIT website has also been updated, offering users an improved facility for browsing individual families' sequence alignment and phylogenetic tree estimates, and still offers the option for downloading the whole database in a single, convenient flatfile.

CONSTRUCTION OF PANDIT

Obtaining sequence data

Each PANDIT database release is numbered to correspond with the Pfam release (7,8) from which it is derived, and is divided into families using the family designations in Pfam. Each family contains three alignments: (i) PANDIT-aa—containing exactly the corresponding Pfam-A seed protein sequence alignment; (ii) PANDIT-dna—containing all the DNA sequences corresponding to the protein sequences in PANDIT-aa that could be recovered using an automated search procedure (see below); (iii) PANDIT-aa-restricted—containing only those protein sequences for which a DNA sequence could be recovered, allowing, for example, direct comparisons of phylogenetic tree estimation methods under models of nucleotide and amino acid evolution. The PANDIT-dna sequence alignment is formed using the corresponding PANDIT-aa alignment as a template; in other words, PANDIT-dna relies on back-translation of the manually curated Pfam protein sequence alignments and makes no attempt to realign using DNA sequence alignment software. Not all sequences in PANDIT-aa have an automatically detectable entry in a nucleotide database that translates to the amino acid sequence; therefore, PANDIT-dna contains fewer sequences than PANDIT-aa.

To retrieve DNA for the protein sequences in PANDIT-aa, cross-references to the EMBL Nucleotide Sequence Database (15) are obtained from the SWISS-PROT and TrEMBL databases that are now incorporated in UniProt (16). Coding sequences are retrieved using the SRS server at the EMBL-EBI (17) and EMBOSS software (18). To create the PANDIT-dna alignment reliably from the Pfam alignment, coding sequences are first translated and aligned to their corresponding protein sequences, taking into account frame shifts specified by the /codon_start qualifier in the feature table of the EMBL entry. This ensures that our DNA alignment correctly reflects the Pfam protein alignment without solely relying on the sequence numbering correspondence between SWISS-PROT/TrEMBL and EMBL, which we have found is sometimes incorrect. A minimum sequence identity of 98% is enforced between Pfam protein sequences and translated coding sequences, allowing for small numbers of mismatches due to annotation inconsistencies. In the case of multiple cross-references to the EMBL database, the sequence with the highest sequence identity between the Pfam protein sequence and translated coding sequences is chosen. DNA sequence alignments are then generated by custom-written software, using the corresponding Pfam-A seed alignment as a template. SWISS-PROT/TrEMBL and EMBL identifiers and accession numbers, protein sequence identifiers and statistics describing the fit to the various genetic codes are all stored.

Alignment quality

Molecular evolutionary studies generally assume the evolutionary homology of amino acid residues or nucleotide bases contained in alignment columns. The Pfam-A seed alignments used as templates for the PANDIT-aa, PANDIT-dna and PANDIT-aa-restricted alignments necessarily vary in the quality of their reconstruction of homology, both within and between alignments. This is a consequence of the varying difficulty of the alignment problems encountered and the alignment methods used in the construction of Pfam [e.g. allowing evolutionarily distinct insertion elements in different sequences to be aligned to one another (19)]. When alignment quality is at a premium, it is commonplace in molecular evolutionary studies to remove all columns containing gaps. In many families in PANDIT this results in the majority of aligned columns being removed and a dearth of data from which to make inferences. In PANDIT 17.0, we implement a more liberal quality control procedure that identifies the more reliable columns in an alignment. Each PANDIT-aa alignment is analyzed using the hmmbuild program of the HMMER package (20) (see http://hmmer.wustl.edu) to generate a profile hidden Markov model (pHMM) describing the alignment (21). The alignment columns inferred to be derived from a 'match' state are taken to be reliable homologs; columns derived from insert/delete states are not considered reliable. This information is recorded in the format of an additional sequence line defining a 'mask', in which each site is recorded with characters 'x' and '.' denoting reliable and unreliable sites, respectively. This information is projected onto the PANDIT-dna and PANDIT-aa-restricted alignments in the obvious manner.

Because the PANDIT database is intended for the study of molecular evolution, only families with two or more recovered DNA sequences (and thus some evolutionary information) are retained. PANDIT 17.0 contains 7738 families, compared with 7868 families in Pfam 17.0. There are 181 448 sequences in PANDIT-aa, yielding 174 760 sequences in PANDIT-dna (and PANDIT-aa-restricted): a hit-rate of 96.3%. PANDIT-aa contains a total of 1 821 853 alignment columns, of which 1 703 482 (93.5%) are considered reliable according to the HMMER pHMMs.

Tree estimation

The tree estimates provided in PANDIT are a crucial part of the database. Since PANDIT's inception, we have progressively improved the tree estimation procedure. Version 17.0 uses five different methods for tree estimation to produce a list of candidate trees for each of the three datasets in each family. Three methods [Neighbor-Joining (14), BioNJ (22) and Weighbor (23)] use clustering algorithms to produce a phylogenetic tree estimate from a pairwise distance matrix produced using maximum likelihood-estimated distances computed with in-house software. FastME (24) uses a

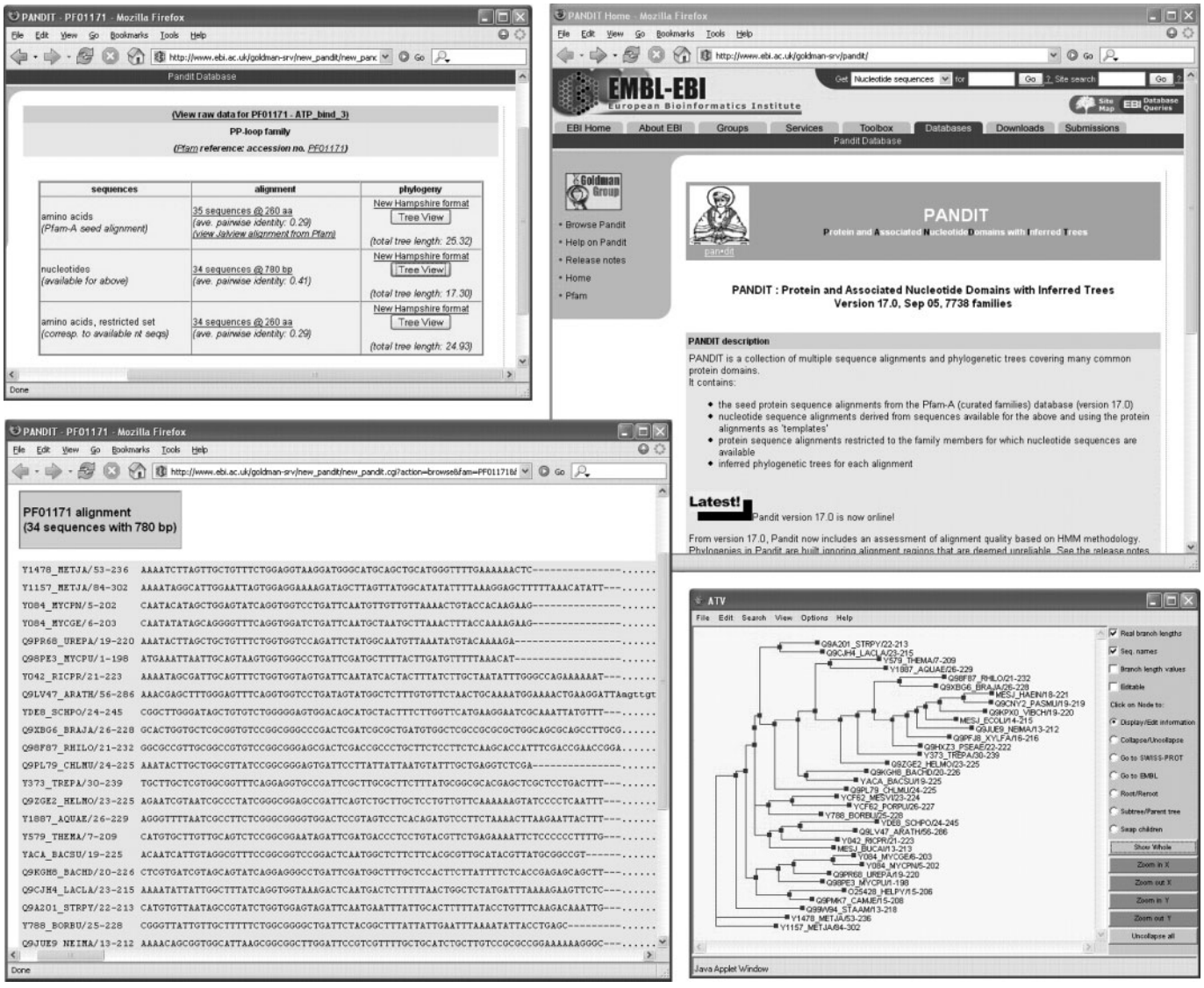

Figure 1. Example of the contents of the PANDIT website, including (anticlockwise from top right) the homepage, the page describing the data available for PF01171 (PP-loop family), the nucleotide sequence alignment for PF01171 and its associated phylogenetic tree.

minimum evolution criterion with local tree-rearrangements to estimate a tree, and Phyml (25) uses maximum likelihood with local tree searching. All procedures requiring probabilistic models of sequence evolution (pairwise distance calculation, likelihood-based tree searching and comparison) use WAG (26) [with amino acid frequencies estimated directly from the data (27)] for amino acid sequences or HKY (28) for corresponding nucleotide data. A list of candidate trees is produced by pooling the results from all five methods, and their likelihoods are then computed and compared using the codeml (amino acids) and baseml (nucleotides) applications of the PAML package (29). For each alignment, the single tree with the highest likelihood is included in the database. We are currently working to expand the repertoire of tree estimation software to further improve this aspect of the database.

Some sequence alignments in PANDIT are quite short, and in combination with large numbers of members in a family this may make some tree estimates unreliable. Nevertheless, these families may still be of value in studies of evolutionary processes, and we therefore retain these in PANDIT. Researchers may wish to remove short alignments from their studies.

The PANDIT website

The PANDIT database is available for browsing and download via its home page at http://www.ebi.ac.uk/goldman-srv/pandit/. Recent improvements permit improved searching and browsing of index pages according to family accession number or name; once a family of interest has been located, a dedicated page details the size of its three (PANDIT-aa, -dna and -aa-restricted) alignments and contains links to those alignments, their associated phylogenetic trees and the associated Pfam (8) and InterPro (30) entries. Phylogenetic trees can be displayed either in the Newick (New Hampshire) format (see http://evolution.genetics.washington.edu/phylip/

newicktree.html) or graphically via a modified version of the ATV tool (31). Figure 1 shows a screen-shot illustrating the content of pages from the PANDIT website.

All of the information contained in PANDIT is available for download as a single flatfile (PANDIT 17.0: 205 Mb, or 57 Mb in .gz compressed format) to allow researchers to produce datasets relevant for their needs quickly and efficiently. Links to PANDIT are now included in the Pfam (8) and InterPro (30) databases. These reciprocal links will be of value to researchers studying the evolution and function of proteins and their constituent domains.

CURRENT APPLICATIONS OF PANDIT AND FUTURE DIRECTIONS

PANDIT is predominantly intended as a tool for researchers investigating the evolutionary process. It provides a valuable resource for those developing and testing novel methodology for examining sequence evolution, allowing generalized conclusions to be drawn about the evolution of proteins and the DNA sequence encoding them. PANDIT has recently been used to demonstrate that proteins with multiple interactions, located in intracellular components, and/or involved in complex processes and functions are generally more conserved than average and less likely to undergo adaptive evolution (32). The observation that large-scale nucleotide substitutions spanning multiple adjacent bases contribute significantly to the evolution of proteins (6) has also been demonstrated using PANDIT. We hope that high quality studies using PANDIT will continue and the ongoing improvements in the database will open the door to new research opportunities.

PANDIT is a developing resource, with the contents of the database and the methodology used to produce it adapting as new data become available and new applications are found for it. As an evolution-centric database, PANDIT will include only limited annotation. By building upon the information held in Pfam, PANDIT users have an established and high quality set of annotation available for all families if required, while cross-links with other established databases such as Pfam (8) and InterPro (30) offer further detail. In future versions, we will further expand PANDIT to include information useful to those performing evolutionary analyses, both in response to comments made from users and according to what we believe will prove useful for prospective analyses. Future additions are likely to include gene ontology (33) terms for each family, species identifiers for sequences, expanded numbers of tree estimation methods, support values for branches in the estimated trees and further quality control and improvement with respect to the Pfam-A seed alignments.

Through the improvements detailed here and other future progress, we hope that PANDIT will become a widely used tool for developing new approaches to studying sequence evolution and casting light on how proteins evolve.

AVAILABILITY

The PANDIT database is freely available on the web via the URL http://www.ebi.ac.uk/goldman-srv/pandit. The entire database is also available for download as a flatfile from this website.

ACKNOWLEDGEMENTS

S.W., N.R. and N.G. are funded by the Wellcome Trust. Tree estimations are performed on a 200-CPU cluster kindly provided by IBM to the Research Program of the European Bioinformatics Institute. Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
2. Yang,Z. (2003) Adaptive molecular evolution. In Balding,D., Bishop,M. and Cannings,C. (eds), *Handbook of Statistical Genetics*. 2nd edn. Wiley, NY, pp. 229–254.
3. The ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **5696**, 636–640.
4. Sawyer,S.L., Wu,L.I., Emerman,M. and Malik,H.M. (2005) Positive selection of primate TRIM5-alpha identifies a critical species-specific retroviral restriction domain. *Proc. Natl Acad. Sci. USA*, **102**, 2832–2837.
5. Hardison,R.C., Roskin,K.M., Yang,S., Diekhans,M., Kent,W.J., Weber,R., Elnitski,L., Li,J., O'Connor,M., Kolbe,D. *et al.* (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.*, **13**, 13–26.
6. Whelan,S. and Goldman,N. (2004) Estimating the frequency of events that cause multiple nucleotide changes. *Genetics*, **167**, 2027–2043.
7. Whelan,S., de Bakker,P.I.W. and Goldman,N. (2003) PANDIT: a database of protein and associated nucleotide domains with inferred trees. *Bioinformatics*, **19**, 1556–1563.
8. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L., Studholme,D.J., Yeats,C. and Eddy,S.R. (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
9. Perriere,G., Duret,L. and Gouy,M. (2000) HOBACGEN: database system for comparative genomics in bacteria. *Genome Res.*, **10**, 379–385.
10. Stebbings,L.A. and Mizuguchi,K. (2004) HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. *Nucleic Acids Res.*, **32**, D203–D027.
11. Duret,L., Mouchiroud,D. and Gouy,M. (1994) HOVERGEN, a database of homologous vertebrate genes. *Nucleic Acids Res.*, **22**, 2360–2365.
12. Morell,V. (1996) TreeBASE: the roots of phylogeny. *Science*, **273**, 569.
13. Massingham,T. and Goldman,N. (2005) Detecting amino acid sites under positive selection and purifying selection. *Genetics*, **169**, 1753–1762.
14. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
15. 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.
16. 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.
17. Zdobnov,E.M., Lopez,R., Apweiler,R. and Etzold,T. (2002) The EBI SRS server—recent developments. *Bioinformatics*, **18**, 368–373.
18. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
19. Löytynoja,A. and Goldman,N. (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl Acad. Sci. USA*, **102**, 10557–10562.
20. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
21. Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.

22. Gascuel,O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.*, **14**, 685–695.
23. Bruno,W.J., Socci,N.D. and Halpern,A.L. (2000) Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.*, **17**, 189–197.
24. Desper,R. and Gascuel,O. (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.*, **9**, 687–705.
25. Guindon,S. and Gascuel,O. (2003) A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. *Syst. Biol.*, **52**, 696–704.
26. Whelan,S. and Goldman,N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.
27. Goldman,N. and Whelan,S. (2002) A novel use of equilibrium frequencies in models of sequence evolution. *Mol. Biol. Evol.*, **19**, 1821–1831.
28. Hasegawa,M., Kishino,H. and Yano,T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**, 160–174.
29. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
30. 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**, D315–D318.
31. Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
32. Aris-Brosou,S. (2005) Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. *Mol. Biol. Evol.*, **22**, 200–209.
33. Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.