BIOINFORMATICS

Vol. 19 no. 12 2003, pages 1556–1563
DOI: 10.1093/bioinformatics/btg188

# Pandit: a database of protein and associated nucleotide domains with inferred trees

Simon Whelan[1, 2,*], Paul I. W. de Bakker[3] and Nick Goldman[1, 2]

[1]Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK, [2]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [3]Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge, CB2 1GA, UK

Received on November 25, 2002; revised on February 10, 2003; accepted on February 21, 2003

**ABSTRACT**

**Motivation:** A large, high-quality database of homologous sequence alignments with good estimates of their corresponding phylogenetic trees will be a valuable resource to those studying phylogenetics. It will allow researchers to compare current and new models of sequence evolution across a large variety of sequences. The large quantity of data may provide inspiration for new models and methodology to study sequence evolution and may allow general statements about the relative effect of different molecular processes on evolution.

**Results:** The Pandit 7.6 database contains 4341 families of sequences derived from the seed alignments of the Pfam database of amino acid alignments of families of homologous protein domains (Bateman *et al.*, 2002). Each family in Pandit includes an alignment of amino acid sequences that matches the corresponding Pfam family seed alignment, an alignment of DNA sequences that contain the coding sequence of the Pfam alignment when they can be recovered (overall, 82.9% of sequences taken from Pfam) and the alignment of amino acid sequences restricted to only those sequences for which a DNA sequence could be recovered. Each of the alignments has an estimate of the phylogenetic tree associated with it. The tree topologies were obtained using the neighbor joining method based on maximum likelihood estimates of the evolutionary distances, with branch lengths then calculated using a standard maximum likelihood approach.

**Availability:** The Pandit database is available for browsing and download via its home page at http://www.ebi.ac.uk/goldman-srv/pandit/.

**Contact:** simon@ebi.ac.uk

## INTRODUCTION

The amount of sequence data available for computational analysis has been increasing at a tremendous rate. In turn, to partition and store this data in an ordered manner the number of sequence databases and the variety of their content has also increased. Of particular interest to those studying molecular evolution are databases of homologous sequences. These generally contain multiple sets of aligned protein sequences whose homology has been detected using either automated procedures, for example hidden Markov models trained on a small number of homologous sequences to detect other related sequences, or through laboratory investigation. These homologous sequence databases concentrate on different aspects of biology; for example Pfam (Bateman *et al.*, 2002) is a database of homologous protein domains, HOVERGEN (Duret *et al.*, 1994) contains a selection of homologous vertebrate proteins and HOMSTRAD (Mizuguchi *et al.*, 1998; de Bakker *et al.*, 2001) contains structure-based alignments of homologous protein families used for protein structure prediction (Williams *et al.*, 2001). Currently few databases explicitly concentrate on the evolutionary relationships of the aligned sequences they contain (though see, for example, PALI, Balaji *et al.*, 2001) and as a consequence the variety of databases for which phylogenetic trees are readily available is limited. Where phylogenies are provided the models and approaches used to estimate them are often simplistic and leave room for improvement. Additionally, the majority of databases have concentrated on proteins and very few contain the DNA sequence counterparts of their protein families, although a notable exception to this is HOBACGEN (Perrière *et al.*, 2000).

Researchers who develop novel models of DNA and protein evolution usually present their potential utility by the application of statistical tests on a small selection of datasets. Currently the widespread acceptance of models comes over time after analyses demonstrating their usefulness are independently performed on a wide variety of datasets. A large, standardised set of DNA and protein sequence alignments, with their associated phylogenetic tree estimates, will allow general, objective and statistical

*To whom correspondence should be addressed.

comparison of current models of sequence evolution. It will also allow researchers developing new models of evolution to rigorously test those models by assessing their usefulness via measures of their performance on the different families in database (see Whelan and Goldman, 2001; Goldman and Whelan, 2002). To these ends we present a new database, Pandit (Protein and Associated Nucleotide Domains with Inferred Trees), based on the seed alignments of protein domains contained in the Pfam-A database (Bateman *et al.*, 2002).

Using the Pfam database has several appealing properties. It imposes the constraint that the structural domains contained within it may not overlap, eliminating redundancy in the database. The high quality of the manually curated Pfam-A seed alignments makes studies based on these data comparable to those performed on carefully aligned protein sequences created by biologists studying molecular evolution. The broad range of evolutionary distances in each alignment ensures that the evolutionary trees estimated from them will not be swamped by closely related or identical sequences. The large variety of proteins contained within Pfam, and thus Pandit, allows a general indication of the comparative performance of different approaches to phylogenetic inference; for example, inferential methodologies and models of sequence evolution (Goldman and Whelan, 2002; Aloy *et al.*, 2002; Pandit *et al.*, 2002). Finally, the seed database contains fewer sequences per alignment than the automatically generated full alignments in Pfam-A, thus making phylogenetic analyses more practical.

The Pandit database is divided into three sections, each containing multiple sets of aligned sequences and an estimate of the phylogenetic tree describing the evolutionary relationships of those sequences. The families in the first section (Pandit-aa) contain the protein sequences from the Pfam-A seed alignments. Those in the second section (Pandit-dna) contain all the DNA sequences corresponding to the protein sequences of the first section that could be recovered using an automated search procedure. Not all sequences in Pandit-aa have an entry in a nucleotide database that translates directly to the amino acid sequence; therefore Pandit-dna contains fewer sequences than Pandit-aa. The final section of the database (Pandit-aa-restricted) contains families of the protein sequences for which a DNA sequence could be recovered. This section of Pandit will allow direct comparisons of phylogenetic tree estimation methods under models of nucleotide and amino acid evolution. For example, one could investigate how often models describing nucleotide and amino acid evolution give different estimates of sequences' phylogeny. We believe that this is the first time such a large and complete database has been produced explicitly for the study of molecular evolution and the assessment and development of phylogenetic methods.

## MATERIALS AND METHODS

### Sequence data contained within Pandit

The alignments contained within Pandit 7.6 are based on the Pfam-A seed alignments of Pfam version 7.6 (Bateman *et al.*, 2002). Note that Pandit version numbers are taken to match the version of the Pfam database they are based upon. Pandit-aa contains all the families from the Pfam-A seed alignments that contain fewer than 1000 sequences and for which two or more DNA sequences could be recovered (see below). This results in 101 230 sequences of protein domains divided into 4341 gapped alignments, each representing an individual protein domain family.

The individual seed alignments for Pfam are not usually altered substantially between Pfam database updates, making Pandit quite stable to updates of Pfam (minor updates occur nearly monthly and major updates occur yearly; see ftp://ftp.sanger.ac.uk/pub/databases/Pfam/relnotes.txt). This and the non-redundancy of the database helps ensure that comparisons between results using present and future releases of Pandit are comparable and that upkeep of Pandit is straightforward; we intend to update Pandit in line with major updates to Pfam.

For all sequences in the Pfam-A seed alignments, cross-references to the EMBL Nucleotide Sequence Database (Stoesser *et al.*, 2002) were obtained from the SWISS-PROT (Bairoch and Apweiler, 2000; release 40.28, of 19 September 2002) and TrEMBL (Bairoch and Apweiler, 2000; release 21.12 of 13 September 2002) databases. Coding sequences with a clean EMBL cross-reference status identifier ('–') are retrieved from the SRS server at the EMBL-EBI (Zdobnov *et al.*, 2002), thus excluding coding sequences with ambiguous equivalences between the protein and DNA sequences (`ALT_INIT`, `ALT_TERM`, `ALT_FRAME`, `ALT_SEQ`, `JOINED` or `NOT_ANNOTATED_CDS` status identifiers). To recover the 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 minimal sequence identity of 98% is enforced between Pfam protein sequences and translated coding sequences (using the genetic code that maximizes the sequence identity; genetic codes were downloaded from http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html) to allow 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, and when multiple DNA sequences give 100% sequence identity the first occurrence is chosen. 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.

The final section of the database, Pandit-aa-restricted, consists of only those protein sequences in Pandit-aa that have a corresponding DNA sequence in Pandit-dna.

## Phylogenetic trees contained within Pandit

Much information about the relative performance of phylogenetic models may be obtained from good estimates of the evolutionary tree (e.g. Yang *et al.*, 1994, 1995; Sullivan *et al.*, 1996; Yang *et al.*, 1998; Adachi *et al.*, 2000; Yang *et al.*, 2000; Whelan and Goldman, 2001). Having a good estimate of the branch lengths for a tree topology may also prove useful in phylogenetic analyses, for direct use or as good starting points in the optimisation procedure for assessing different models (for example see Whelan and Goldman, 2001). Each of the alignments in the three sections of the Pandit database has two phylogenetic trees presented with it: one inferred phylogeny including estimates of all branch lengths, and one detailing only the topology.

Because of the large number of sequences contained within some of the families in the database it is not feasible to use the statistically preferred method of maximum likelihood (ML; see Whelan *et al.*, 2001) to estimate the tree topology. Instead, we use an approach similar to that of Whelan and Goldman (2001) to estimate good phylogenetic tree topologies. All calculations regarding the estimation of pairwise distances or the estimation of branch lengths and model parameters on given tree topologies (see below) use a ML approach; the methods for performing these calculations are well documented elsewhere (Felsenstein, 1981; Swofford *et al.*, 1996). All pairwise distance calculations are performed using purpose written software, neighbor joining (NJ) analyses (Saitou and Nei, 1987) are performed using the PHYLIP software package (as implemented in the neighbor program of PHYLIP version 3.5c; Felsenstein, 1993) and calculations using ML on given tree topologies are performed using PAML version 3.13 (Yang, 1997).

## Calculations for amino acid alignments

For each family of Pandit-aa and Pandit-aa-restricted, all calculations are performed using the WAG+F model, which has been shown to be a good description of the evolutionary process (Whelan and Goldman, 2001). This model contains 189 fixed parameters that were estimated empirically from a large database and describe the propensity of different amino acids to replace each other, and 19 free parameters describing the equilibrium frequencies of the amino acids. Within each family, pairwise distances are estimated with frequency parameters estimated independently for that family by counting the amino acids observed in the alignment. A phylogenetic tree topology for the family is estimated using NJ on this set of distances. The branch lengths from this topology estimation are discarded, and branch lengths (in units of expected amino acid replacements per site) for each tree are re-estimated using the standard ML approach on the estimated tree topology (Swofford *et al.*, 1996). We expect our methods of phylogeny reconstruction to provide more reliable tree estimates than are currently available via the Pfam website (see descriptions at http://www.sanger.ac.uk/Software/Pfam/help/alignments.shtml and http://www.sanger.ac.uk/Software/analysis/quicktree): the use of more sophisticated models to estimate initial distances and the re-estimation of branch lengths under ML may be expected to provide improvements over both NIFAS (Storm and Sonnhammer, 2001) and QuickTree (Howe *et al.*, 2002), and the NJ tree estimation method used in Pandit-aa is widely considered a more sophisticated and reliable approach than that employed in NIFAS.

## Calculations for nucleotide alignments

For each family of Pandit-dna, all calculations are performed using the HKY model (Hasegawa *et al.*, 1985). This model contains four free parameters for each family: $\kappa$, describing the tendency for transitions (A $\leftrightarrow$ G, C $\leftrightarrow$ T) to occur at a higher rate than transversions (A,G $\leftrightarrow$ C,T), and three free parameters describing the equilibrium frequencies of the nucleotides. This model is thought provide a reasonable description of the process of nucleotide substitution in DNA (for a discussion of models see Swofford *et al.*, 1996; Whelan *et al.*, 2001).

The parameter $\kappa$ normally may be estimated directly from the observed data when using a standard ML approach on a full tree topology (Swofford *et al.*, 1996). The estimation of $\kappa$ for the calculation of pairwise distances is more complex. Ideally, a single $\kappa$ would be estimated simultaneously with all pairwise distances using a ML approach but due to the large size of some of the families in Pandit-dna this approach is not feasible. A multi-step approach for estimating $\kappa$ for each family is used instead. Initially, $\kappa_i$ and its variance, var($\kappa_i$), are estimated for each pairwise comparison $i$ of sequences in a family using the formulae of Kimura (1983). An average value for the family is then calculated as the mean of the pairwise $\kappa_i$ weighted by the inverse of their variances:

$$\bar{\kappa} = \sum_i \kappa_i \frac{1/\mathrm{var}(\kappa_i)}{\sum_j 1/\mathrm{var}(\kappa_j)}. \quad (1)$$

This estimate is not directly comparable to that of the

HKY model because it does not take into account the nucleotide composition of the sequences. The estimate of the parameter used for the calculation of pairwise distances needs to be adjusted using the equation

$$\kappa = \bar{\kappa}\frac{(\pi_A + \pi_G)(\pi_C + \pi_T)}{\pi_A\pi_G + \pi_C\pi_T} \quad (2)$$

with $\pi_A$, $\pi_C$, $\pi_G$, $\pi_T$ representing the equilibrium frequencies of the nucleotides A, C, G and T, respectively, estimated for each family by counting the nucleotides observed in the alignment.

Pairwise distances using HKY can then be estimated using this estimate of $\kappa$ and the equilibrium frequencies estimated from the entire family by counting. A phylogenetic tree topology for each family is estimated by applying NJ to the resulting set of distances. Branch lengths (in units of expected nucleotide substitutions per site) are then re-estimated using a standard ML approach on this tree topology under the HKY model, with $\kappa$ simultaneously re-estimated from all the sequence data for each family. We give a brief comparison of the two methods for estimating values of $\kappa$ in the next section; this indicates that the pairwise method used prior to pairwise distance calculation and topology estimation performs reasonably well.

## RESULTS AND DISCUSSION

### Web access to database

The Pandit database is available for browsing and download via its home page at http://www.ebi.ac.uk/goldman-srv/pandit/. Browsable pages include notes on Pandit and index pages listing all the families in the database and allowing browsing by family name or by Pfam accession number. Each family has a dedicated page detailing the size of its three (Pandit-aa, -dna and -aa-restricted) alignments and with links to those alignments, their associated phylogenetic trees, and the Pfam entry for that family. These pages also are linked to and from their corresponding Pfam families. It is also possible to visualise phylogenetic trees via the ATV tool (Zmasek and Eddy, 2001), which we have modified to include links directly from the tree visualisation to the EMBL and SWISS-PROT database entries for the sequences in the tree. Also available for those who would like a local copy of the entire database is a flatfile containing all the information within Pandit. This makes downloading of large numbers of estimated phylogenetic trees more convenient than is possible with Pfam.

### Database and associated statistics

We now demonstrate that Pandit contains a wide range of sequence data useful to those researching molecular evolution and representative of the type of data sets that biologists may encounter in their studies. Figure 1a shows a comparison of the number of sequences and alignment length for each family in Pandit-aa. The majority of families contain between two and 50 sequences (mean = 23.3; min = 2; max = 972; inter-quartile range = 15) and vary in length between 50 and 500 amino acid residues (mean = 246.6; min = 5; max = 1780; inter-quartile range = 221). The data contained in Pandit-aa-restricted follows a very similar pattern: the mean number of sequences reduces slightly to 19.0 (min = 2; max = 929; inter-quartile range = 13), and the distribution of alignment lengths remains the same because Pandit contains the same alignment information in all its three subsections. Figure 1b shows the distribution of mean pairwise sequence identity for the individual families for the three sections of the database. Pandit-aa and Pandit-aa-restricted show very similar distributions of divergence, as expected because one is a restricted subset of the other, and this demonstrates there has been no obvious bias in the identification of DNA sequences. The markedly greater mean pairwise sequence identities of the Pandit-dna families is caused by each amino acid residue being represented by a triplet of DNA nucleotides in a codon. Each observed amino acid replacement between two protein sequences may be due to only a single nucleotide replacement. The occurrence of synonymous substitutions, usually at the third codon position, does not offset this effect.

The upper graph in Figure 2 gives an indication of the total amount of phylogenetic information within each alignment in Pandit-dna in terms of the total expected number of nucleotide substitutions. To obtain this the tree length (sum of all branch lengths in a phylogeny) is multiplied by the length of the alignment and is plotted against the number of sequences for each family. This may be of interest to those developing new evolutionary models who wish to ensure that sufficient substitutions have occurred in the evolutionary history of an alignment to permit robust parameter estimation. For example, the estimation of empirical models of evolution requires alignments that have sufficient numbers of observed changes (e.g. Dayhoff *et al.*, 1978; Jones *et al.*, 1992; Whelan and Goldman, 2001). For those studying and developing models incorporating heterogeneity of evolutionary process amongst sites, for example codon models describing positive selection (Yang and Bielawski, 2000; Yang *et al.*, 2000), the lower graph in Figure 2 plots the number of sequences in an alignment against tree length for Pandit-dna. This provides an indication of the expected amount of phylogenetic information per site in an alignment.

### Estimates of $\kappa$

As described above, for each family in the Pandit-dna database there are two estimates of the parameter $\kappa$ of the HKY model (Hasegawa *et al.*, 1985), which describes transition/transversion bias of nucleotide substitutions

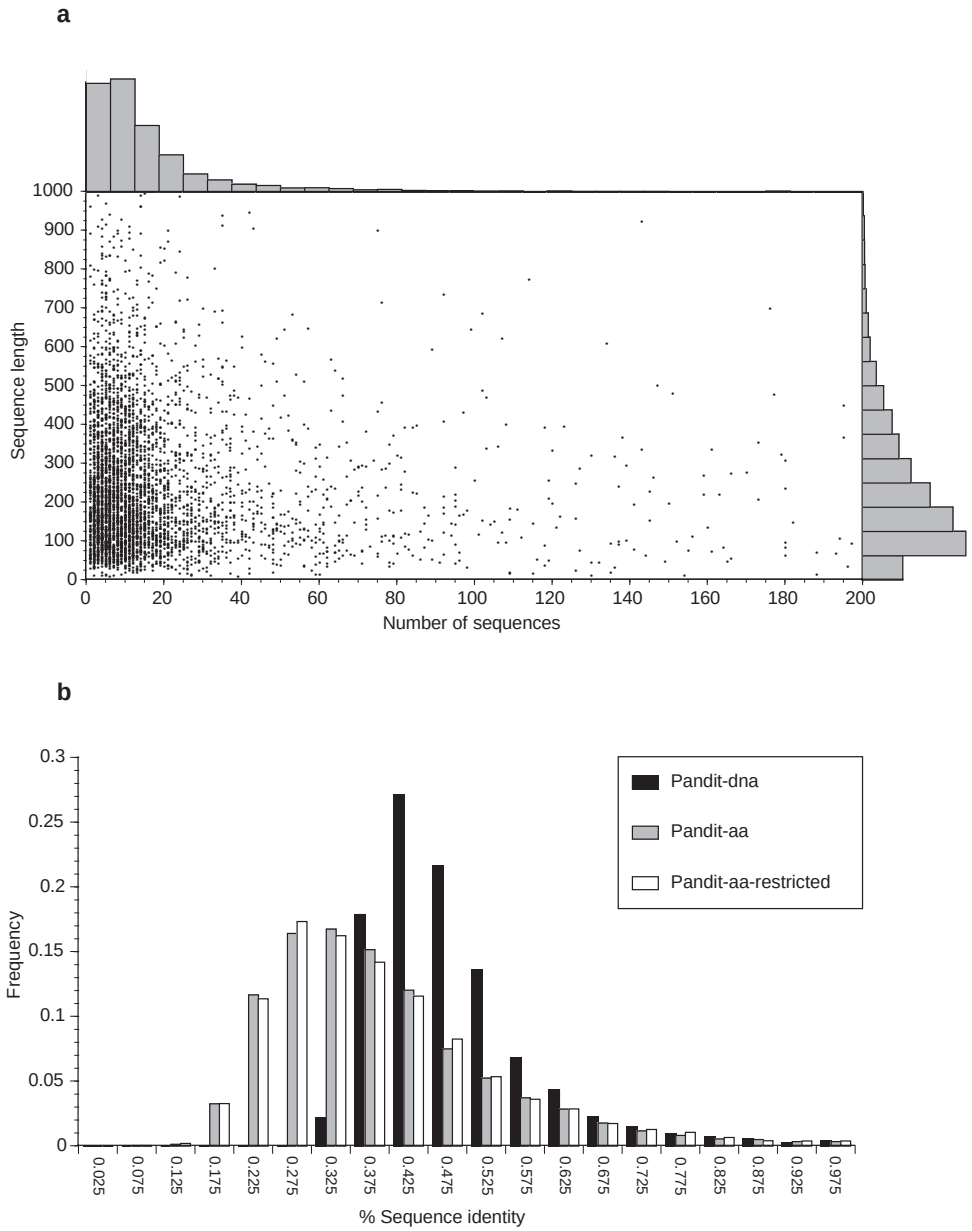


**Fig. 1.** Pandit database statistics. (**a**) Sequence length plotted against number of sequences in a family for Pandit-aa. The histograms located on the top and right-hand side of the scatter plot indicate the density of points along the *x*- and *y*-axes, respectively. Note that for clarity the axes do not contain the whole range of the data; 81 points (1.9%) are excluded. (**b**) Distributions of family mean pairwise sequence identities for the different sections of the Pandit database. Pandit-aa: mean = 0.38, min = 0.09, max = 1.00, inter-quartile range = 0.16; Pandit-aa-restricted: mean = 0.38, min = 0.09, max = 1.00, inter-quartile range = 0.17; Pandit-dna: mean = 0.48, min = 0.31, max = 1.00, inter-quartile range = 0.11.

during evolution. In order to assess the quality of the estimate made using our pairwise approach it can be compared to the ML estimate for the whole family, which can be expected to be a reasonable estimate (Edwards, 1972). Figure 3 shows a strong correlation between the two estimates, with the pairwise method having some tendency to overestimate $\kappa$. This graph and simulation results (not shown) suggest that the pairwise approach provides a reasonable method for the estimation of $\kappa$ in most cases. There are a few outliers where the ML estimate of $\kappa$ is considerably higher than the pairwise estimate. These are cases where the sequences are very

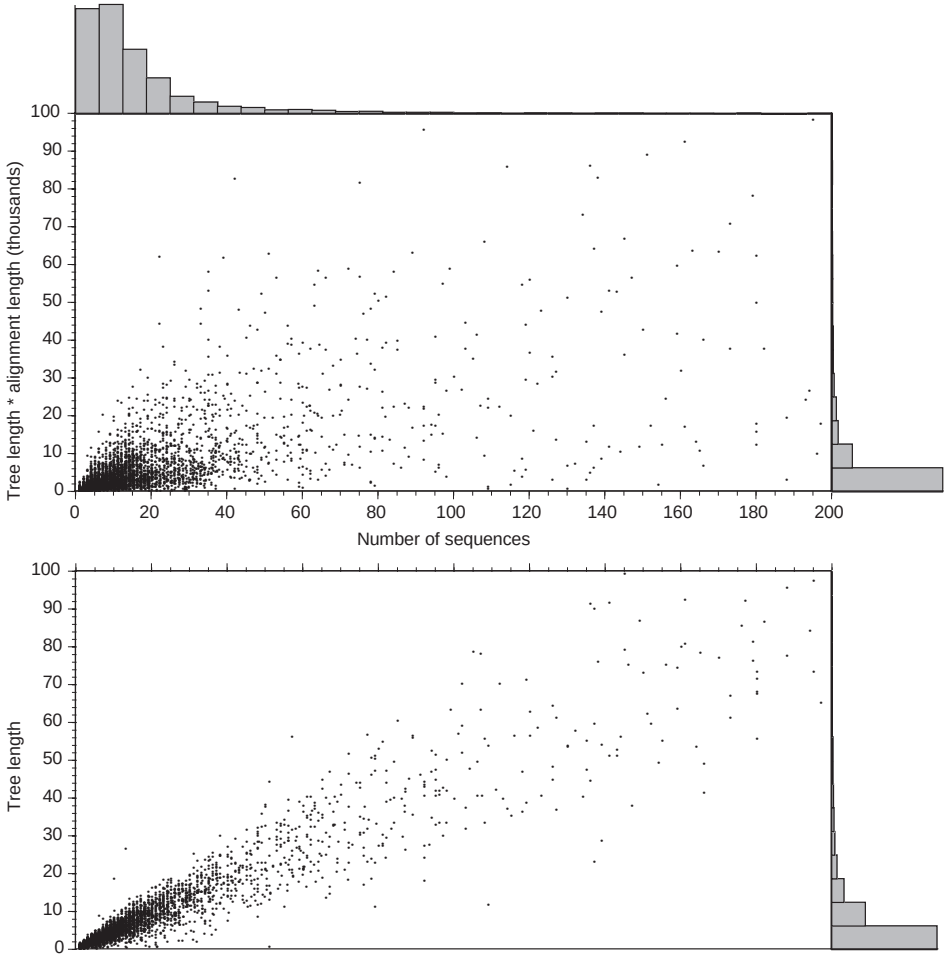


**Fig. 2.** Summary of phylogenetic information contained within Pandit-dna families. The upper and lower graphs give indications of the total information and information per site in each family, respectively (see text for details). Histograms indicate the densities of points along the *x*- and *y*-axes. Note that for clarity the axes do not contain the whole range of the data; 62 points (1.4%) are excluded from the upper graph and 54 points (1.2%) are excluded from the lower graph.

similar and there are very few observed changes from which to calculate the pairwise estimate. The histograms in Figure 3 show the distributions of estimates of $\kappa$ over families in Pandit-dna. We know of no previous studies describing these distributions over a large number of DNA sequence alignments, and they may be of use to applied researchers who wish to see whether their estimates of $\kappa$ fall within the usual range and to researchers devising prior distributions on $\kappa$ for Bayesian analyses (e.g. Huelsenbeck *et al.*, 2001).

## CONCLUSION

We present the Pandit database, version 7.6, which contains 4341 families of protein domain sequences. Pandit contains three sub-databases of sequence alignments: Pandit-aa, comprising amino acid alignments taken from Pfam-A seeds; Pandit-dna, comprising the DNA sequences coding for the Pandit-aa alignments whenever it is possible to recover them by the automated procedure described above; and Pandit-aa-restricted, comprising the amino acid sequences of Pandit-aa for which a DNA sequence could be recovered for inclusion in Pandit-dna. For each of the 13020 (4340 $\times$ 3) alignments, a phylogenetic tree topology is estimated using the NJ method based on pairwise distances estimated by maximum likelihood. The branch lengths for each tree are then estimated using ML. Pandit's unique combination of large numbers of protein and nucleotide sequences with inferred evolutionary trees will be of value to those studying molecular phylogenetics.

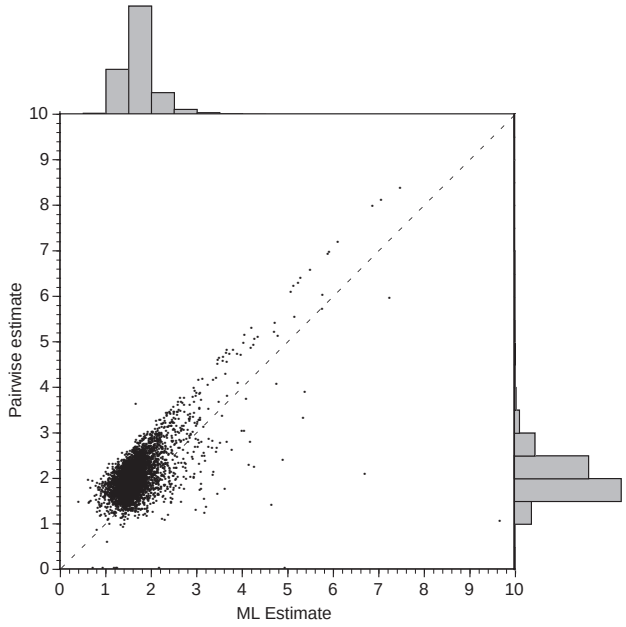


**Fig. 3.** Relationship between the ML estimate and pairwise estimate of $\kappa$ for each family of Pandit-dna. Note that 9 extreme points (0.2%) have been omitted for clarity. Histograms indicate the density of points along the *x*- and *y*-axes.

## ACKNOWLEDGEMENTS

SW and NG are supported by the Wellcome Trust. PIWDB is funded by the Cambridge European Trust and the Isaac Newton Trust. Jenny Wang made modifications to the ATV code used for phylogeny visualisation. We would like to thank Rolf Apweiler, Alex Bateman, Tim Massingham and Ziheng Yang for their useful comments.

## REFERENCES

Adachi,J., Waddell,P.J., Martin,W. and Hasegawa,M. (2000) Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J. Mol. Evol.*, **50**, 348–358.

Aloy,P., Oliva,B., Querol,E., Aviles,F.X. and Russell,R.B. (2002) Structural similarity to link sequence space: new potential superfamilies and implications for structural genomics. *Protein Sci.*, **11**, 1101–1116.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Balaji,S., Sujatha,S., Sai Chetan Kumar,S. and Srinivasan,N. (2001) PALI-a database of Phylogeny and ALIgnment of homologous protein structures. *Nucleic Acids Res.*, **29**, 61–65.

de Bakker,P.I.W., Bateman,A., Burke,D.F., Miguel,R.N., Mizuguchi,K., Shi,J., Shirai,H. and Blundell,T.L. (2001) HOMSTRAD: adding sequence information to structure-based alignments of homologous protein families. *Bioinformatics*, **17**, 748–749.

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.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*, Vol. 5 Suppl. 2, National Biomedical Research Foundation, Washington DC, pp. 345–352.

Duret,L., Mouchiroud,D. and Gouy,M. (1994) HOVERGEN, a database of homologous vertebrate genes. *Nucleic Acids Res.*, **22**, 2360–2365.

Edwards,A.W.F. (1972) *Likelihood*. Cambridge University Press, Cambridge.

Felsenstein,J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Felsenstein,J. (1993) *PHYLIP (Phylogeny Inference Package) Version 3.5c*. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Goldman,N. and Whelan,S. (2002) A novel use of equilibrium frequencies in models of sequence evolution. *Mol. Biol. Evol.*, **19**, 1821–1831.

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.

Howe,K, Bateman,A and Durbin,R. (2002) QuickTree: building huge neighbour-joining trees of protein sequences. *Bioinformatics*, **18**, 1546–1547.

Huelsenbeck,J.P., Ronquist,F., Nielsen,R. and Bollback,J.P. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *CABIOS*, **8**, 275–282.

Kimura,M. (1983) *The neutral theory of evolution*. Cambridge University Press, Cambridge.

Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.

Pandit,S.B., Gosar,D., Abhiman,S., Sujatha,S., Dixit,S.S., Mhatre,N.S., Sowdhamini,R. and Srinivasan,N. (2002) SUPFAM—a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. *Nucleic Acids Res.*, **30**, 289–293.

Perrière,G., Duret,L. and Gouy,M. (2000) HOBACGEN: database system for comparative genomics in bacteria. *Genome Res.*, **10**, 379–385.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V. *et al.* (2002) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **30**, 21–26.

Storm,C.E. and Sonnhammer,E.L. (2001) NIFAS: visual analysis of domain evolution in proteins. *Bioinformatics*, **17**, 343–348.

Sullivan,J., Holsinger,K.E. and Simon,C. (1996) The effect of topology on estimation of among site rate variation. *J. Mol. Evol.*, **42**, 308–312.

Swofford,D.L., Olsen,G.J., Waddell,P.J. and Hillis,D.M. (1996) In Hillis,D.M., Moritz,C. and Mable,B.K. (eds), *Molecular systematics*, 2nd edn, Sinauer, Sunderland, MA, pp. 407–514.

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.

Whelan,S., Liò,P. and Goldman,N. (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.*, **17**, 262–272.

Williams,M.G., Shirai,H., Shi,J., Nagendra,H.G., Mueller,J., Mizuguchi,K., Miguel,R.N., Lovell,S.C., Innis,C.A., Deane,C.M. *et al.* (2001) Sequence-structure homology recognition by iterative alignment refinement and comparative modelling. *Proteins*, Suppl. 5, 92–97.

Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS*, **13**, 555–556.

Yang,Z. and Bielawski,B. (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.*, **15**, 496–503.

Yang,Z., Goldman,N. and Friday,A. (1994) Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. *Mol. Biol. Evol.*, **11**, 316–324.

Yang,Z., Goldman,N. and Friday,A. (1995) Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst. Biol.*, **44**, 384–399.

Yang,Z., Nielsen,R., Goldman,N. and Pedersen,A.-M.K. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449.

Yang,Z., Nielsen,R. and Hasegawa,M. (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.*, **15**, 1600–1611.

Zdobnov,E.M., Lopez,R., Apweiler,R. and Etzold,T. (2002) The EBI SRS server—new features. *Bioinformatics*, **18**, 1149–1150.

Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.