Software

THoR: a tool for domain discovery and curation of multiple alignments

Nicholas J Dickens and Chris P Ponting

Address: MRC Functional Genetics Unit, University of Oxford, Department of Human Anatomy and Genetics, South Parks Road, Oxford OX1 3QX, UK.

Correspondence: Nicholas J Dickens. E-mail: nicholas.dickens@anat.ox.ac.uk

Published: 23 July 2003

Genome ***Biology*** 2003, **4:**R52

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/8/R52

Received: 20 May 2003
Revised: 17 June 2003
Accepted: 25 June 2003

Abstract

We describe a tool, THoR, that automatically creates and curates multiple sequence alignments representing protein domains. This exploits both PSI-BLAST and HMMER algorithms and provides an accurate and comprehensive alignment for any domain family. The entire process is designed for use via a web-browser, with simple links and cross-references to relevant information, to assist the assessment of biological significance. THoR has been benchmarked for accuracy using the SMART and pufferfish genome databases.

Rationale

Data emanating from the genome-sequencing projects are flooding sequence databases. Without informative annotation, these sequences will not achieve their full potential of facilitating directed experimental research. One approach to annotation is to use efficient and automatic procedures to associate predictions of evolution, structure and function to genes. When annotating sequences from an evolutionary perspective, it is important to characterize sequences in terms of domains, which we define as being compact and spatially distinct protein structures [1]. As domains are often present in different molecular contexts and in different combinations, it is necessary to employ sophisticated resources such as Pfam [2] and SMART [3] for their prediction. Each of these is a database of multiple sequence alignments, and associated hidden Markov models (HMMs) [4], and each exploits the HMMER package (profile HMM software for protein sequence analysis) [5]. A problem for SMART is that the rapid expansion of protein-sequence databases has the consequence that multiple alignments are unable to be updated synchronously with newly deposited sequences. Here we describe THoR (the Thorough Homology Resource), a new tool that addresses this problem by automatically curating and updating multiple alignments for large domain families.

Identification of domain homologs for inclusion in SMART multiple sequence alignments [6] typically uses a combination of two algorithms, PSI-BLAST (position-specific iterative BLAST) [7] and HMMER [5]. PSI-BLAST is a method of choice for identification of homologs with divergent sequences [8]. In this method a query sequence is compared with a sequence database and candidate homologs are detected with Expect (E) values less than a threshold value (E_ψ). $E(x)$ is defined as the number of alignments expected in the database search with scores x or higher purely by chance. In subsequent iterations the database is searched using a position-specific score matrix that is calculated from a multiple sequence alignment of all the homologs detected in previous search rounds.

The main advantage of PSI-BLAST over HMMER is its speed. One of its major disadvantages is that it generates local alignments, rather than the global alignments that are required to delineate full-length domains. Thus, BLAST-derived local

alignments often need to be extended in both directions to span intact domains. Furthermore, its searches are not symmetrical. If a sequence X is used to query a database and a sequence Y is predicted to be its homolog, then X is not necessarily always found to be a homolog when Y is used to search the same database [9]. A final disadvantage is that multiple PSI-BLAST searches can generate a significant number of false-positive predictions. As discussed by Jones and Swindells [10], the results of multiple PSI-BLAST searches can provide error rates as high as one false positive in 13.2% of searches when these are taken to five rounds.

In contrast to PSI-BLAST, HMMER is able to generate accurate global alignments. The HMMER hmmsearch method uses a hidden Markov model (HMM), previously calculated from a multiple sequence alignment, to do a global search of a sequence database for candidate homologs, taken to be sequences detected with *E* values less than a threshold, E_{HMM}. As a result of its heuristics, HMMER is often a better tool for the detection of short repeats or domains than PSI-BLAST.

Exhaustive PSI-BLAST searches that use multiple individual sequences as queries, have been shown to detect a high proportion of homologues in databases [8,9]. However, the task of manually cross-referencing these multiple PSI-BLAST results to extract all predicted candidate homologs is extremely onerous and time-consuming. It is made even more difficult for populous domain families and for domains that are repeated often within single proteins by the large amount of data involved in the analyses. The system for easy analysis of lots of sequences (SEALS) package [11] eases this task by providing PSI-BLAST output parsers linked to sequence retrieval. However, SEALS can be difficult to install and cannot accurately retrieve full-length domain sequences from BLAST-derived local alignments.

The software package described here takes advantage of the speed and sensitivity of PSI-BLAST, together with the global alignment benefit of HMMER. This 'thorough homology resource' (THoR) provides a convenient web-based interface to the cross-referenced results of exhaustive PSI-BLAST database searches. The package takes a multiple sequence alignment as input and generates an updated and extended global alignment of all domain homologs, defined as those predicted by both the PSI-BLAST and HMMER methods. It also provides the benefits of convenient access to the search results and cross-referencing against the National Center for Biotechnology Information (NCBI) protein sequence and taxonomy databases.

The THoR process is represented in Figure 1 as a flow diagram. In the initial step, it takes as its input a multiple protein sequence alignment A_n that is dismantled into its constituent sequences $A_{seq\ n}$. Multiple PSI-BLAST searches, using these sequences as queries, are initiated employing a nonredundant database D_{nr}. These searches are continued to a user-specified number of iterations, or else to convergence, using a supplied value of E_ψ. All significant ($E < E_\psi$) high-scoring pairs (HSPs) resulting from these PSI-BLAST searches (HSP-ψ) are extracted and stored in plain-text format. These may include multiple HSPs in single sequences that are indicative of domain repeats. In a next step, the complete sequences of all proteins that contain one or more HSPs are extracted from D_{nr} and are reconstituted as a flat-file database of homologs, D_H.

Subsequently the multiple alignment A_n is used to generate an HMM_n. High-scoring global alignments against HMM_n are then calculated for all sequences in D_H and all HMMER HSPs (HSP-H) with *E*-values less than a threshold value (E_{HMM}) are collected (Figure 1).

The domain homologs predicted by this protocol are defined as the intersection between the HMMER aligning regions, HMMER high-scoring pairs (HSP-H), and the PSI-BLAST HSPs (HSP-ψ). An HSP-H (with initial and final amino-acid coordinates A_i, A_j) is considered to be equivalent to an HSP-ψ (with initial and final amino-acid coordinates B_i, B_j) if:

$$n < \frac{\min(A_j, B_j) - \max(A_i, B_i)}{A_j - A_i} \times 100$$

where, by default, n = 60%, although this can be supplied by the user. However, values of n < 50% are more suitable for the identification of motifs, rather than domains.

Typically, $E_\psi << E_{HMM}$, as these values are determined using different approaches and because PSI-BLAST is the method of choice for homolog detection. $E_\psi = 5 \times 10^{-3}$ and $E_{HMM} = 1$ are recommended values used in THoR for comparison against current protein-sequence databases. Higher values of E_ψ result in increased false-positive rates, whereas higher E_{HMM} values generate less accurate global alignments, albeit of likely true homologs predicted by PSI-BLAST.

All domain homologs identified using this approach are then realigned against the original sequence alignment A_n, using hmmalign (Figure 1) with the model HMM_n. There is a series of intermediate iterations of this process, where an HMM, constructed from the new alignment $A_{n,i}$, is then used to research the database of PSI-BLAST results. Where new domain homologs are revealed, that satisfy both E_{HMM} and E_ψ thresholds, these are appended to the alignment in an iterative manner until convergence, resulting in a revised multiple sequence alignment A_{n+1}. The entire process can then be repeated using A_{n+1} as the input alignment for the subsequent run of THoR. In order to reduce redundancy, an option is provided to purge the alignment of one of a pair of sequences that are greater than a threshold percentage identity. At this point it is advisable that the user assesses the quality of the alignment A_{n+1} for minor misalignments and fragments resulting from gene mispredictions or incomplete protein sequences.

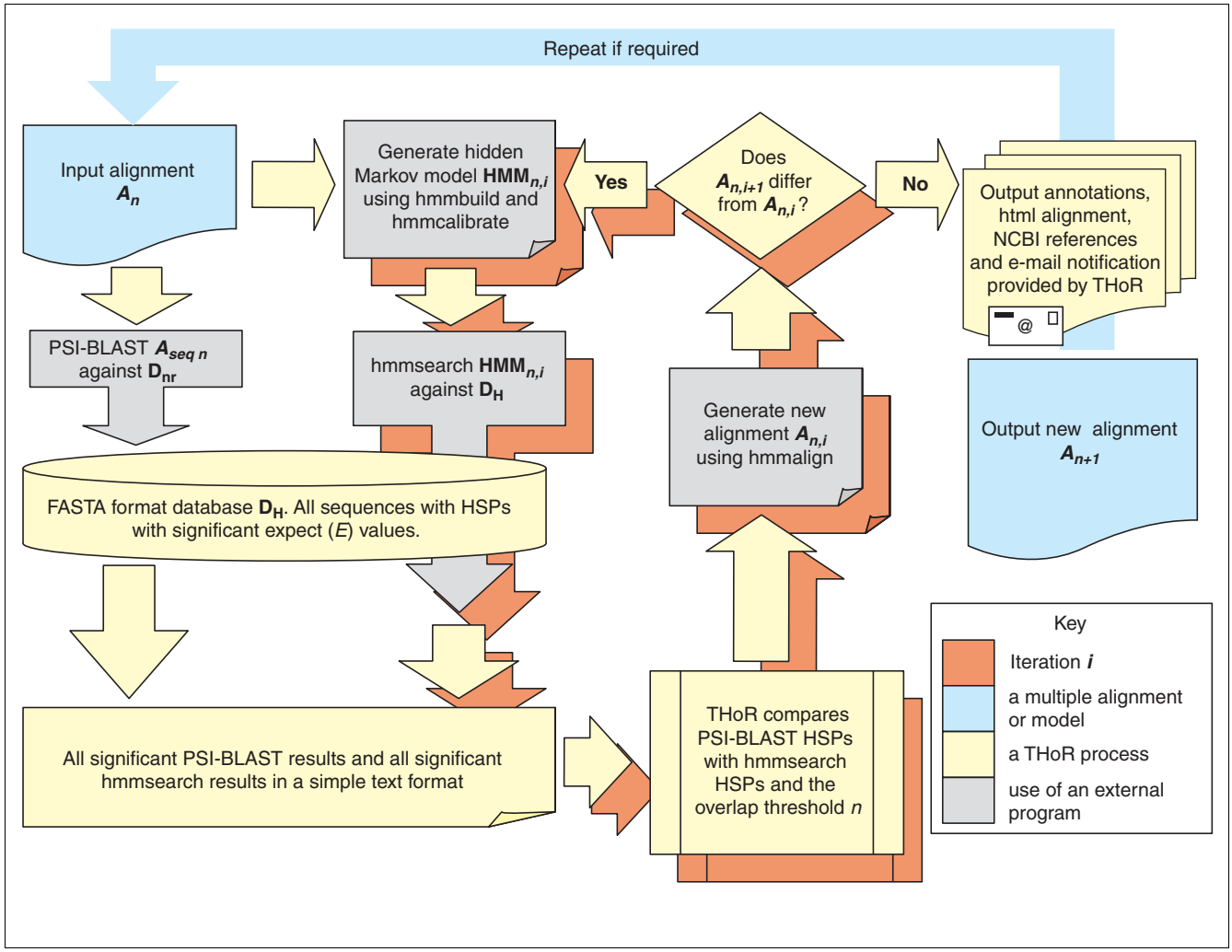

Figure 1
Flow diagram that illustrates the procedures used by THoR. A starting alignment A_n is broken into its constituent sequences and each of these is compared against the nonredundant database of choice (D_{nr}) using PSI-BLAST and a user-specified number of iterations. The full-length sequence for each of the high-scoring pairs (HSP-ψ) from the PSI-BLAST results are deposited in a temporary FASTA database of homologs (D_H) in a nonredundant fashion. When this database is complete, a hidden Markov model (HMM) of the alignment A_n is generated and compared with D_H using hmmsearch. Subsequently, the PSI-BLAST HSPs (HSP-ψ) and HMMER HSPs (HSP-H) are compared. They are considered equivalent if the HSP-ψ and HSP-H overlap significantly, overlap ≥ *n* (see text). During iterations *i* intermediate alignments ($A_{n,i}$) and HMMs ($HMM_{n,i}$) are produced. The latter are used to re-search the database (D_H) in order to accumulate and append additional sequences to the alignment. This iterative step is repeated until no new entries are added to the alignment when compared with the previous search ($A_{n,i-1}$). At this stage results are written to the THoR output file and the final alignment ($A_{n,i+1}$) is produced.

Advice for the construction and manual editing of multiple sequence alignments is given by Bork and Gibson [12], and Ponting and Birney [13].

The THoR analyses are optimized so that for any subsequent runs of THoR for the given alignment (A_{n+1}), only those sequences that are novel when compared with the previous alignment (A_n) will be subjected to PSI-BLAST searches and only results from these searches will be appended to those emanating from the previous runs of THoR.

Benchmarking

The following benchmarking tests were used to investigate the performance of THoR.

Domain homolog identification

An HMM can be used to find domain family homologs using the HMMER program hmmsearch. This method has been used extensively for the annotation of new genomes [14,15] and as such provides an excellent benchmark against which the success of THoR can be assessed. For the purposes of

comparison, a set of 50 domain family multiple sequence alignments were selected from the SMART domain database. To prevent any domain-based bias in the searches the domains were chosen in such a way that there were representatives of all the different domain types; that is enzymes and non-enzymes, large and small in size, found broadly and narrowly in diverse taxa, with many and few homologs, and domains found in various cellular compartments (nuclear, cytoplasmic and secreted). Among the domains found in narrowly diverse taxa were domains that were not expected to be found within the pufferfish genome; these served as negative controls.

The pufferfish (*Takifugu rubripes*) genome [16] was chosen for benchmarking the THoR method for two reasons. First, this would substantially reduce any sequence-based bias in the searches, as there are few alignments in SMART that contain pufferfish sequences. If the human genome had been chosen instead, a bias in the searches would have arisen from the high representation of human and mammalian sequences in SMART alignments. Second, there are relatively few pufferfish sequences represented in the NCBI nonredundant (nr) database that will be part of the THoR search process. This maintains the nonredundancy of the database, which is important for the efficiency of both the PSI-BLAST and HMMER searches.

Speed

Typical timings for a complete round of THoR against the nr plus pufferfish database were taken for domains of different sizes. Two other factors that increase the search speed are the size of the target database and the frequency of occurrence of the domain within that database, simply because domains that occur more frequently produce more results that are analyzed and compared by the THoR program.

Stability of the search process

Domains that are highly represented within the database that is searched by THoR produce large files of results to be analyzed. Often a single domain family search produces gigabytes of PSI-BLAST results. Consequently it is essential that all searches are completed without encountering memory problems. All 50 of the benchmarking domains completed without these errors and successfully generated a subsequent alignment. However, the 'ATPases associated with diverse cellular activities' (AAA) domain family produced an alignment with 14,279 members which could not generate an HMM owing to a limitation in hmmbuild on the benchmarking machine. The WD40 repeat also encountered similar problems. One way to circumvent this limitation is described in the discussion.

For each domain family, lists of the results that were found only by THoR and those that were found only by hmmsearch were generated and then each item in these lists was examined manually in order to discover false-positive and false-negative results. This analysis combined PSI-BLAST and subsequent hmmsearch searches using the closest homologs for each hit as queries. Examination of the results that were unique to THoR revealed that only one putative false positive was present among the 50 resulting alignments. The only cases where hmmsearch identified homologs that THoR did not were domain alignments less than 40 amino acids in length (see Results and discussion).

Results and discussion

The results of the benchmarking test (Table 1) demonstrate that the simultaneous application of both the PSI-BLAST and HMMER methods provides a thorough search procedure that generates a more complete set of results than does either of the two methods individually. The HMM searching method will not necessarily identify all the homologs of a domain family, and although PSI-BLAST alone will often identify many homologs it will not necessarily align over the complete domains. In addition, there is a possibility that multiple PSI-BLAST searches will provide false-positive results with multiple searches at iterations of five or above. The results of multiple PSI-BLAST searches can provide error rates as high as one false positive in 13.2% of searches to five rounds [10]. However, the application of the THoR process significantly reduces the probability of false-positive hits, as a result of the cross-validation of PSI-BLAST and HMMER outputs. It is unlikely that a false-positive sequence identified by PSI-BLAST local alignment statistics would also be identified with significance by the HMMER global alignment method.

Only one candidate false positive was predicted in our benchmarking test. Out of a total of 3,519 domain homologs found by THoR, one sequence (SINFRUP00000075480 amino acids 159 to 242) could not be validated independently as a PINT domain through the use of PSI-BLAST and SMART. Similarly, it could not be proved that this sequence was not a homolog. The alignment of this sequence and SMART PINT domains is provided as additional data available with the online version of this paper (see Additional data file).

Comparison of hmmsearch results and THoR results revealed false negatives, defined as domains identified using the SMART hmmsearch but not found by THoR, only when the starting alignment was shorter that 35 to 40 amino acids (Table 1). Short repeats are often difficult to identify using PSI-BLAST because of its inappropriate use of optimal alignment statistics for suboptimal alignments [17] as well as the general problem of low alignment scores. As THoR is reliant on PSI-BLAST for homolog detection, it is not surprising that it shares its limitations. Consequently, THoR is only appropriate for domains or repeats with lengths exceeding 40 amino acids. For short repeats or domains hmmsearch would prove to be a better tool for domain homolog identification.

The caveats of short alignments and large sets of results highlight two principles that must be applied in order to gain the

Table 1

A comparison of the number of pufferfish hits by hmmsearch results versus the pufferfish database both before and after the THoR process

Domain name (SMART name)	N(SMART)	N(THoR)	N(THoR) - N(SMART)
14-3-3 homologs (14_3_3)	9	9	0
Domains in Ataxins and HMG-containing proteins (AXH)	6	6	0
Breast cancer carboxy-terminal domain (BRCT)	31	39	8
Bromo domain (BROMO)	89	89	0
Bulb-type mannose-specific lectins (B_lectin)	1	2	1
Chromatin organization modifier domain (CHROMO)	62	69	7
Calpain-like thiol protease family (CysPc)	31	32	1
Tandem repeat (DM15)	6	6	0
Endothelin (END).	5	5	0
Exonuclease (EXOIII)	10	12	2
Receptor for Ubiquitination Targets (FBOX)	34	45	11
Formin homology 2 domain (FH2)	20	35	15
Fibronectin type 1 domain (FN1)	49	49	0
High mobility group (HMG)	82	84	2
Homeodomain (HOX)	319	323	4
Protein kinase C-related kinase homology region 1 homologs (HR1)	19	19	0
Short calmodulin-binding motif containing conserved Ile and Gln residues (IQ)	228	226	-2
Kyprides, Ouzounis, Woese motif (KOW)	12	12	0
Kringle (KR)	33	34	1
Zinc-binding domain present in Lin-11, Isl-1, Mec-3 (LIM)	204	214	10
Pleckstrin homology (PH)	373	436	63
Zinc finger (PHD)	216	303	87
Phosphoinositide 3-kinase, region postulated to contain C2 domain (PI3K_C2)	10	12	2
Motif in proteasome subunits, Int-6, Nip-1 and TRIP-15 (PINT)	16	17	1
Phosphatidylinositol phosphate kinases (PIPKc)	14	15	1
Domain found in a protein subunit of human RNase MRP and RNase P ribonucleoprotein complexes and archaeal proteins (POP4)	1	1	0
Domain found in Plexins, Semaphorins and Integrins (PSI)	116	119	3
Domain with conserved PWWP motif (PWWP)	27	29	2
Guanine nucleotide exchange factor for Rho/Rac/Cdc42-like GTPases (RhoGEF)	99	111	12
Src homology 2 domains (SH2)	142	153	11
Src homology 3 domains (SH3)	358	373	15
Staphylococcal nuclease homologs (SNc)	3	6	3
Domain in short gastrulation protein and chordin (SOG)	3	3	0
snRNP Sm proteins (Sm)	18	18	0
TopoisomeraseII (TOP2c)	3	3	0
Tetratricopeptide repeats (TPR)	573	552	-21
Tudor domain (TUDOR)	25	44	19
Domain present in VPS-27, Hrs and STAM (VHS)	15	14	-1

N(SMART) is the number of domains found in the predicted set of pufferfish proteins using hmmsearch with SMART thresholds. N(THoR) is the number of domains found in pufferfish using hmmsearch with SMART thresholds using the alignment created by THoR. N(THoR) - N(SMART) is the difference between the THoR results and the SMART results. The SMART domain families COLIPASE, ChW, CheW, Galanin, IL10, IL2, LIGANc, POLIIIc, POX and REC were used for the benchmarking as negative controls. None of these domains was expected to provide positive hits to the pufferfish database, because they are prokaryote-specific or mammal-specific domains; indeed, no pufferfish homologs were detected by THoR. The domains AAA and WD40 were both searched by THoR with only one round of PSI-BLAST, because they were known to contain many members and a full search of five rounds would require an unnecessarily lengthy period of time to complete. They are not shown because they encountered memory-allocation errors with hmmbuild and their search iterations did not complete.

Table 2

Locations from which the programs and files required by THoR can be downloaded

Program	Location
Perl modules	http://search.cpan.org/
Blast 2	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/
Hmmer-2.1.1	ftp://ftp.genetics.wustl.edu/pub/eddy/hmmer/
Chroma	http://www.lg.ndirect.co.uk/chroma/
Apache 1.3.x	http://httpd.apache.org/
PHP4	http://www.php.net/downloads.php
The NCBI nr database	ftp://ftp.ncbi.nih.gov/blast/db/nr.tar.gz
THoR downloads	http://smart.ox.ac.uk/thor/

greatest benefit from the use of THoR. First, the quality of the starting alignment greatly affects the quality of the results that are received, which is also the case for hmmsearch searches. Results from alignments containing many sequences require a long time to complete because there are more PSI-BLAST searches to perform and because these require significantly more HSP-H to HSP-ψ comparisons. The extra information gained from having more members in the starting alignment is, in most cases, offset by the extra time required for calculations. Therefore, it is suggested that in most cases an alignment that is nonredundant at a level of 50-80% sequence identity will be sufficient to produce an improved set of results. This approach reduces the likelihood of encountering memory errors, as for the cases of the AAA domain and WD40 repeat. Similarly, it is important that the target database is as nonredundant as possible [18]. When analyzing domains that are known to have large numbers of homologs within the database (such as AAA) it may even be necessary to use a database that is nonredundant at a level of 50% identity, in order to generate results that are of a size that can be processed within a reasonable period of time.

The THoR package provides an easy-to-use interface to both of the search methods and the results are provided in a simple, intuitive format. The web interface has been designed to provide easy access to as much information as possible, including annotation and taxon information. In order to facilitate the evaluation of large amounts of data, all the results for each entry in the new alignment are displayed in a condensed format, which contains the essential information such as sequence identifiers and HSP-H and HSP-ψ locations. Access to the complete dataset, such as all of the PSI-BLAST results for each entry and all of the queries that identified that sequence, is through a series of links that either open hidden sections of the page for viewing or open a new, smaller window. This is a very effective method for data management, which provides straightforward and transparent quality control of sequence information.

Materials and methods

THoR uses PSI-BLAST and HMMER algorithms and Perl scripts, integrated within a PHP-based interface supplied through the Apache web-server [19]. Although the THoR package has been designed to operate on the Linux platform, it is likely to run on most Unix-based operating systems and possibly on a Windows-based system with some minor modifications.

The prerequisites of THoR installation are a Linux platform with at least a 2.2 kernel and Perl 5, which is installed by default in most Linux distributions. The only Perl modules that are not in default installations are String::CRC, Mail::Mailer and Getopt::Long. The other software requirements are Blast 2 [7], HMMER 2.1.1 [5], CHROMA [20], and the Apache web-server (version 1.3.x) with the php4 module installed. All these programs are distributed under the terms and conditions of the GNU public licence [21] or are distributed freely under a similar license. For the purpose of analyses, THoR also requires a protein-sequence database formatted as NCBI database index files containing GenInfo (GI) numbers. The most suitable database for most analyses is the NCBI nr database. Databases that do not contain GI numbers need to be converted into the appropriate format by adding artificial GI numbers to their accession lines. The locations from which these files can be downloaded are shown in Table 2.

The THoR package is provided with installation instructions and Perl installation scripts on the Oxford SMART website [22]. THoR will run on a Linux box capable of supporting the software and at least a 10 Gb hard disk drive. The whole package has been successfully installed and run on a single-processor 300 MHz machine with 128 Mb RAM. However, as the PSI-BLAST searches are computer-intensive and the results can be large, it is recommended that the package is installed on a dual-processor Linux machine with at least 512 Mb RAM and 40 Gb of hard disk storage space.

Additional data files

The alignment of SMART PINT domains is available with the online version of this article (Additional data file 1).

Acknowledgements

We would like to thank Leo Goodstadt for helpful discussions concerning the optimization of Perl code.

References

1. Ponting CP, Russell RR: **The natural history of protein domains.** *Annu Rev Biophys Biomol Struct* 2002, **31:**45-71.
2. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2002, **30:**276-280.
3. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P: **Recent improvements to the SMART domain-based sequence annotation resource.**

Nucleic Acids Res 2002, **30:**242-244.
4. Hofmann K: **Sensitive protein comparisons with profiles and hidden Markov models.** *Brief Bioinform* 2000, **1:**167-178.
5. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-763.
6. Ponting CP, Schultz J, Copley RR, Andrade MA, Bork P: **Evolution of domain families.** *Adv Protein Chem* 2000, **54:**185-244.
7. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
8. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C: **Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods.** *J Mol Biol* 1998, **284:**1201-1210.
9. Salamov AA, Suwa M, Orengo CA, Swindells MB: **Combining sensitive database searches with multiple intermediates to detect distant homologues.** *Protein Eng* 1999, **12:**95-100.
10. Jones DT, Swindells MB: **Getting the most from PSI-BLAST.** *Trends Biochem Sci* 2002, **27:**161-164.
11. Walker DR, Koonin EV: **SEALS: a system for easy analysis of lots of sequences.** *Proc Int Conf Intell Syst Mol Biol* 1997, **5:**333-339.
12. Bork P, Gibson TJ: **Applying motif and profile searches.** *Methods Enzymol* 1996, **266:**162-184.
13. Ponting CP, Birney E: **Identification of domains from protein sequences.** *Methods Mol Biol* 2000, **143:**53-69.
14. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
15. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al.*: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
16. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, *et al.*: **Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*.** *Science* 2002, **297:**1301-1310.
17. Andrade MA, Ponting CP, Gibson TJ, Bork P: **Homology-based method for identification of protein repeats using statistical significance estimates.** *J Mol Biol* 2000, **298:**521-537.
18. Holm L, Sander C: **Removing near-neighbour redundancy from large protein sequence collections.** *Bioinformatics* 1998, **14:**423-429.
19. **The Apache Software Foundation** [http://www.apache.org]
20. Goodstadt L, Ponting CP: **CHROMA: consensus-based colouring of multiple alignments for publication.** *Bioinformatics* 2001, **17:**845-846.
21. **The Free Software Foundation** [http://www.gnu.org]
22. **Thorough Homology Resource** [http://smart.ox.ac.uk/thor]