D674–D679 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 28 November 2006*
doi:10.1093/nar/gkl990

GeneSpeed: protein domain organization of the transcriptome

Alecksandr Kutchma, Nayeem Quayum and Jan Jensen*

Barbara Davis Center for Childhood Diabetes, University of Colorado at Denver and Health Sciences Center,
1775 North Ursula Street, Mail Stop B140, PO Box 6511, Denver, CO, 80045, USA

Received August 15, 2006; Revised October 27, 2006; Accepted October 30, 2006

ABSTRACT

The GeneSpeed database (http://genespeed.uchsc.edu/) is an online database and resource tool facilitating the detailed study of protein domain homology in the transcriptomes of *Homo sapiens, Mus musculus, Drosophila melanogaster* and *Caenorhabditis elegans*. The population schema for the GeneSpeed database takes advantage of HOWARD^{TM} parallel cluster technology (http://www.massivelyparallel.com/) and performs exhaustive tBLASTn searches covering all pre-assigned PFAM domain classes in all species (currently 7973 domain families) against the respective Unigene EST databases of the selected four transcriptomes. The resulting database provides a complete annotation of presumed protein domain presence for each Unigene cluster. To complement this domain annotation we have also performed a custom transcription factor-family curation of all Pfam domains, incorporated the Gene Ontology classifications for these domains as well as integrated the Novartis SymAtlas2 dataset for both human and mouse which provides rapid and easy access to tissue-based expression analysis. Consequently, the GeneSpeed database provides the user with the capability to browse or search the database by any of these specialized criteria as well as more traditional means (gene identifier, gene symbol, etc.), thereby enabling a supervised analysis of gene families through a top-down hierarchical basis defined by domain content, all directly linked to an optimized gene expression dataset.

INTRODUCTION

Our aim was to create a tool that would allow the construction of gene lists with homology to any specified protein domain, coupled with the capability to analyze the expression levels of these genes as provided by gene microarray experiments. Our motivation to develop such a resource originated from the general lack of such interactivity between existing highly used online resources. We specifically intended to develop a resource that could rapidly provide exhaustive gene-type classification information, concomitantly with expression information of the entire set of members within a specific species type. Such a working profile suits numerous areas of investigation, as it is often the case that a priori knowledge of gene family involvement in a given biological process is available, but the individual gene member is unknown. However, if exhaustive gene family lists could be generated with ease, the coupling of such lists to an extensive gene expression base, should help in narrowing down family members playing tissue-specific roles. If such a tool was available, we reasoned it would benefit investigators in a diverse range of disciplines. Here we have created such a tool and named it the GeneSpeed database.

The core knowledgebase of GeneSpeed is the comprehensive homology categorization of all Pfam (1) protein domain homology within the Unigene transcriptomes (2) of human, mouse, worm and fly, without which its development would have been impossible. This dataset facilitates the generation of gene lists with homology to any specified protein domain by first allowing the user to identify a single protein domain that may characterize a certain protein family of interest and then searching the transcriptome of one of these species for all translations containing homology to this domain. The current version of GeneSpeed only allows queries of single domains; however, we plan to implement the ability for multiple domain searches in a future version. The resultant list will contain all genes in the transcriptome with homology to the protein domain of interest (displayed with decreasing similarity). In order to assist the user in identifying an initial protein domain of interest, GeneSpeed provides several straightforward entry points. The first takes advantage of a custom/manually curated transcription factor characterization of all domains within Pfam (currently 7973 domains) so that an investigator may browse by transcription factor type. Second, we incorporated the Gene Ontology (GO) characterization of domains (3), which allows browsing by the GO organizational tree and the ability to generate a gene list from any specified GO node. The third technique allows a more traditional query by gene/domain name, symbol or accession identifier. Finally, the capability to search the

*To whom correspondence should be addressed. Tel: +1 303 724 6844; Fax: +1 303 724 6830; Email: jan.jensen@uchsc.edu

© 2006 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

GeneSpeed database by fold expression level or by tissue specificity. This is made possible through the incorporation of the Novartis SymAtlas2 dataset (4), containing expression data from 280 microarrays of 61 mouse and 79 human tissues. These integrated expression data also serve an exploratory role, in that the genes within gene lists generated by the other query techniques mentioned above may be analyzed for fold-change from median, tissue specificity as well as absolute expression level. All resulting datasets and tools were organized into a LAMP (Linux/Apache/MySQL/PHP) based dynamic website named GeneSpeed and is available at http://genespeed.uchsc.edu/.

DATABASE CONSTRUCTION

The process for populating the GeneSpeed Database is illustrated in Figure 1, and is nearly fully automated. (i) The first step involves obtaining the full set of Pfam domains from the Pfam database (currently representing 7973 domain families). (ii) A single domain family is selected out of the total of 7973 (in this example the C2H2-Zinc Finger Domain will be used). (iii) The 'full' Pfam alignment file for the C2H2-Zinc Finger Domain (containing 32 874 distinct C2H2-Zinc Finger protein sequences from a multitude of different species) is generated. (iv) Utilized in a batch (iterative) tBLASTn process. This process involves using the tBLASTn algorithm (5) to search for homology for each of the 38 874 C2H2-Zinc Finger domains within the Unigene EST databases for *Mus musculus, Homo sapiens, Caenorhabditis elegans* and *Drosophila melanogaster*. In order to preserve protein ID/aa$^{\mathrm{from}}$-aa$^{\mathrm{to}}$/*E*-score value, we imposed an upper *E*-score cutoff at 1×10^{-2}. Custom PERL and Python scripts were utilized to perform all tBLASTn homology searches using the BLAST PbH cluster maintained at Massively Parallel Technologies Inc. (v) The tBLASTn output for this Pfam domain is subsequently parsed for expectancy score (*E*-score), Unigene Id, GenBank Id, Pfam Id, domain range and domain size. These results are sorted first by *E*-score and then by Unigene Id and subsequently all redundancy was eliminated by retaining only the most significant e-score containing record for each unique homologous Unigene hit. (vi) All extracted information was deposited into a custom MySQL relational database. (vii) This process is then repeated for each of the 7973 Pfam domains. (viii) In addition to the results of the tBLASTn population process, several other sources of data

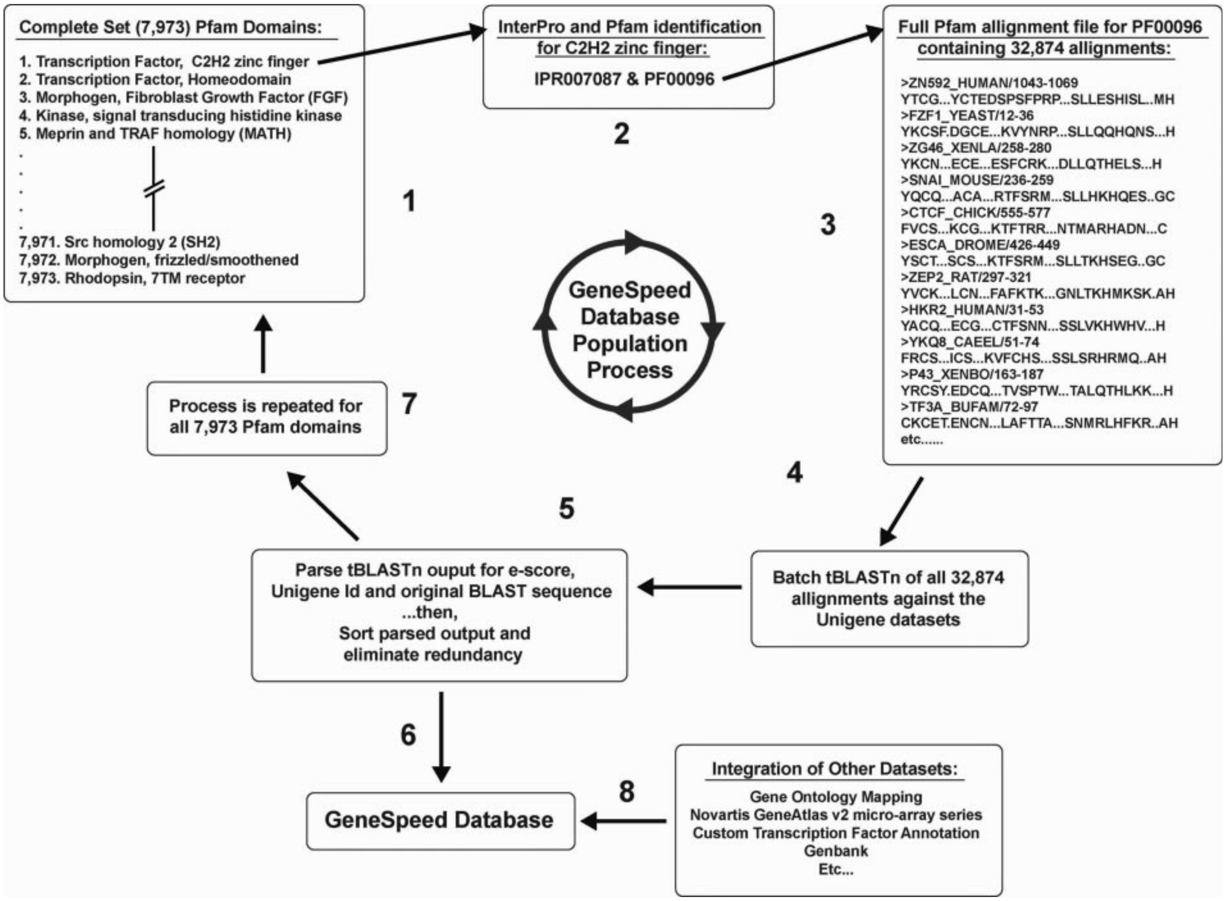
Complete Set (7,973) Pfam Domains:
1. Transcription Factor, C2H2 zinc finger
2. Transcription Factor, Homeodomain
3. Morphogen, Fibroblast Growth Factor (FGF)
4. Kinase, signal transducing histidine kinase
5. Meprin and TRAF homology (MATH)
7,971. Src homology 2 (SH2)
7,972. Morphogen, frizzled/smoothened
7,973. Rhodopsin, 7TM receptor
InterPro and Pfam identification for C2H2 zinc finger:
IPR007087 & PF00096
2
Full Pfam allignment file for PF00096 containing 32,874 allignments:
>ZN592_HUMAN/1043-1069
YTCG...YCTEDSPSFPRP...SLLESHISL..MH
>FZF1_YEAST/12-36
YKCSF.DGCE...KVYNRP...SLLQQHQNS...H
>ZG46_XENLA/258-280
YKCN...ECE...ESFCRK...DLLQTHELS...H
>SNAI_MOUSE/236-259
YQCQ...ACA...RTFSRM...SLLHKHQES..GC
>CTCF_CHICK/555-577
FVCS...KCG...KTFTRR...NTMARHADN...C
>ESCA_DROME/426-449
YSCT...SCS...KTFSRM...SLLTKHSEG..GC
>ZEP2_RAT/297-321
YVCK...LCN...FAFKTK...GNLTKHMKSK.AH
>HKR2_HUMAN/31-53
YACQ...ECG...CTFSNN...SSLVKHWHV...H
>YKQ8_CAEEL/51-74
FRCS...ICS...KVFCHS...SSLSRHRMQ..AH
>P43_XENBO/163-187
YRCSY.EDCQ...TVSPTW...TALQTHLKK...H
>TF3A_BUFAM/72-97
CKCET.ENCN...LAFTTA...SNMRLHFKR..AH
etc......
1
3
GeneSpeed Database Population Process
7
Process is repeated for all 7,973 Pfam domains
4
5
Parse tBLASTn ouput for e-score, Unigene Id and original BLAST sequence ...then, Sort parsed output and eliminate redundancy
Batch tBLASTn of all 32,874 allignments against the Unigene datasets
6
8
Integration of Other Datasets:
Gene Ontology Mapping
Novartis GeneAtlas v2 micro-array series
Custom Transcription Factor Annotation
Genbank
Etc...
GeneSpeed Database

Figure 1. GeneSpeed database population process. In sequential order as given in the figure: (1) all domain information and alignments in the Pfam database are downloaded from Pfam; (2) a single domain is selected (C2H2-Zinc Finger in this example); (3) the 'full' alignment (in this case the C2H2-Zinc Finger domain 'full' alignment contains 32 874 distinct protein sequences) is used in a (4) batch NCBI-based tBLASTn using the Massively Parallel PbH BLAST Server; (5) tBLASTn output is parsed and redundancy eliminated; (6) non-redundant data are banked into a custom MySQL database; (7) the process is repeated for all domain families in Pfam (currently 7973); (8) other datasets are integrated with the database.

were incorporated into the GeneSpeed Database. Information from external databases including Unigene (2), InterPro (3), GO (6) and the Novartis SymAtlas2 (4) were included. From the most recent build of the Unigene database we incorporated the Unigene ID, gene symbol, gene name, Entrez Gene, chromosome and cytoband location. In addition, through retired Unigene information tables, we implemented a tool that automatically checks for Unigene IDs that have been retired from the database. When uploading a custom gene list into GeneSpeed, this script scans the list for all retired or changed Unigene identifiers. Once a retired Unigene ID is discovered, the most recent identifier(s) is displayed for the user. From InterPro, we obtained all InterPro Ids and Gene Ontology (GO) information relating domain families to GO categories. This information allowed the development of a GO browsing tool to search for genes containing domains that have an associated GO classification (3,6). As a result, these incorporated GO classifications may allow for gene-to-gene inference based on an original GO-member domain. Thus, the GeneSpeed database may be used as a GO-prediction tool. However, care should be exercised, as protein homology does not always equate functional similarity, especially at the lower GO-node levels.

As mentioned above, array data for the 79 human and 61 mouse tissues (in duplicate) of the Novartis SymAtlas2 dataset (4) were included into the GeneSpeed database. These array data were downloaded from Novartis (http://wombat.gnf.org/index.html) in its MAS5 (Affymetrix©) normalized form. In addition to incorporating the absolute expression normalized values for each probeset into the database, we also pre-computed the fold-change differential from Median as well as performing an ANOVA statistical test to determine a *P*-value for tissue specificity. This ANOVA computation compared the mean expression value of a gene for one tissue against the mean expression value of the same gene in all other tissues. Such a calculation was repeated for all genes within each tissue type for both human and mouse tissues. All computations were performed within the R statistical environment (http://www.r-project.org/) running on a Linux Mandrake 10.2 operating system (http://www.mandriva.com/).

Given personal interests in the regulatory DNA-binding factors, we also incorporated a custom transcription factor classification of all 7973 Pfam protein domains. Modeling the outline of a previous classification scheme (7), domains within the Pfam database were manually evaluated for their 'relatedness' with transcription factors and classified into three hierarchies of Superclass, Class and Subclass. 'Relatedness' in this case refers to any domains characteristic of or associated with known transcription factors (i.e. associated with text queries of 'transcription factor' or 'DNA binding domain' in the Pfam and InterPro databases). 'Relatedness', however, does not imply an absolute transcription factor classification for any particular domain. To clarify this point with an example, the user may browse the GeneSpeed *M.musculus* division of the database for the transcription factor Superclass of 'Helix Turn Helix' (HTH) and Class of 'Fork Head/Winged Helix'. This yields a list of domains including the Fork Head domain as well as 13 other domains and a total of 480 genes. The Fork Head domain is the signature domain that definitively characterizes the HTH and Fork Head/Winged Helix transcription factor; however, 13 other domains have also been found to be associated with these transcription factor proteins. Thus, if an InterPro-based sub-search of this list is performed for all genes in GeneSpeed that only contain the Fork Head domain (PF00250, IPR001766), the gene list shrinks from 480 to 38. In translation, 480 genes were identified that contain domains (albeit some with very low similarity) that have been associated with Fork Head/Winged Helix transcription factors and 38 genes contain a definitive Fork Head domain. The complete transcription factor classification of domains is available in the 'Background' section of the GeneSpeed website.

At the time of this submission, updating of the GeneSpeed Database is scheduled quarterly. An update consists of running the complete database generation process from the most recent versions of all information/database sources.

NAVIGATION AND USE

A general navigation overview for the GeneSpeed Database is illustrated in Figure 2. In order to permit users to save custom gene lists while searching the GeneSpeed database, all users of the database are provided with a free personal account. After logging in, they may browse GeneSpeed, perform queries, create custom gene lists and save these lists into their account. Alternatively, a guest user does not have to log into the database. Guest users have full access to all database content; however, to save gene lists, a free account must be established. There is also an upload tool implemented allowing users to upload their own gene lists into the database. This tool receives (as upload from user) a Unigene gene list, scans the list for all retired or changed Unigene identifiers and updates all retired Unigene identifiers to the most recent build of the Unigene database. Any discrepancies, such as retired or divided clusters, will be displayed to the user. Once the user's external list has been uploaded into the database, all the tools of GeneSpeed may be used on the new uploaded list. The primary rationale for including the import tool is to provide rapid domain assignment and other annotation within the GeneSpeed environment for gene lists obtained from external experiments. This is particularly useful for lists of unknown genes originating from custom microarray experiments. Users may log back into the database at any time to access their previously saved custom gene lists. Each user's gene list information is protected by login name and password and may not be observed by any other user. If using GeneSpeed as a guest user, only a single gene list may be created and manipulated at a given time.

A new search (Figure 2, A. Search) starts with selecting the organism and then the type of search, which includes (i) Keyword, search the 'gene name' and 'gene symbol' fields of the database. (ii) Id or Accession, search by Unigene, Entrez Gene, Chromosome or InterPro. (iii) GO, browse GO categories. (iv) Transcription Factor Classification, to browse GeneSpeed's custom transcription factor classification hierarchy. (v) Microarray Expression data, allowing the user to query the database by fold-change from median as well as tissue specificity (ANOVA *P*-value). Once a search has produced a new list (or a previously saved list is selected from the account page) the user is prompted to select what

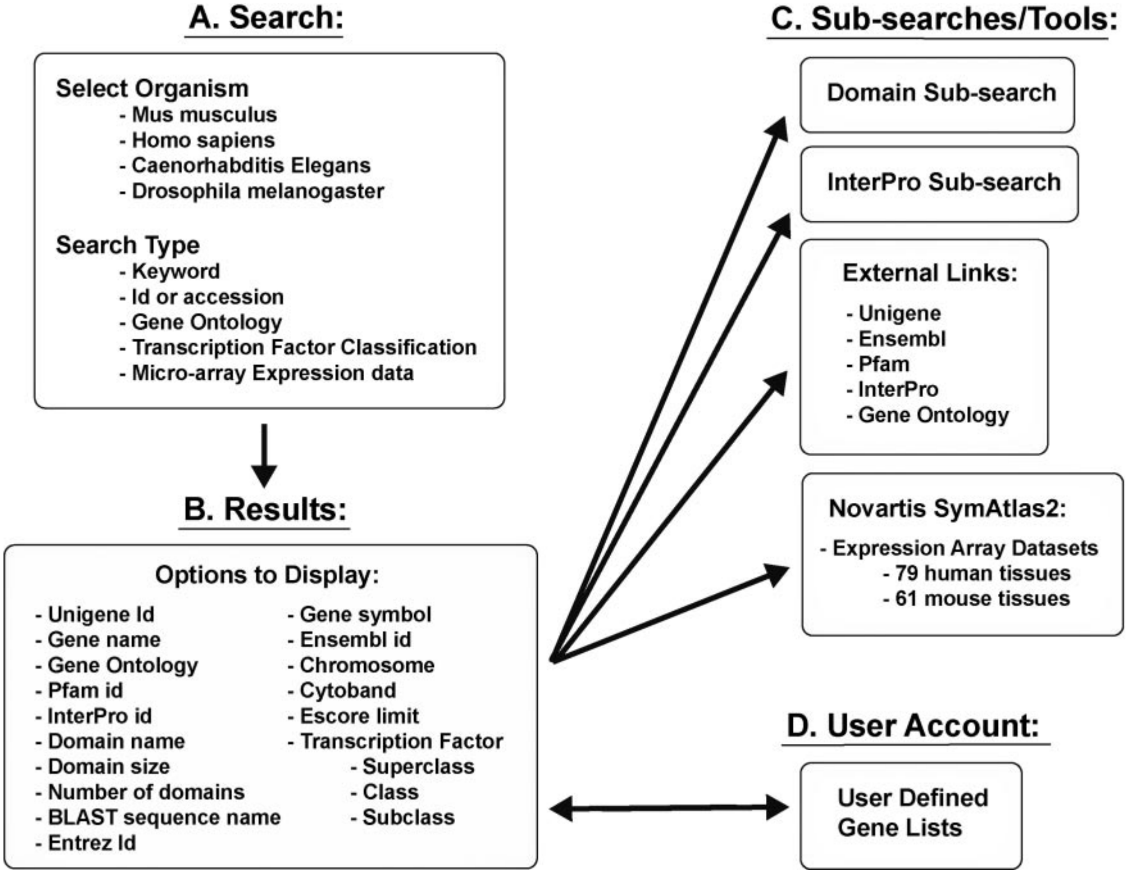

Figure 2. GeneSpeed database navigation diagram. (A. Search) A search is started by first selecting the organism and then the search type. (B. Results) The results page will display the number of Unigene clusters (genes) found and the user then has the option to select additional specific information they would like displayed concerning these genes. (C. Sub-searches/Tool) Other sub-searches/tools are available to refine the original search including Domain Sub-search (finds all domains in a specific gene), InterPro Sub-search (finds all genes with a specific domain), External Links (www links to outside databases) and Novartis SymAtlas2 (investigate the expression level of any number of genes with the data contained within the Norvartis SymAtlas2 dataset including 79 human and 61 mouse tissues). (D. User Account) Each user is provided with a private account, in which they may store any number of user-specified gene lists to keep for analysis at any later time point. Upload tools are also provided, allowing users to analyze any gene list not generated within the GeneSpeed database.

information to display for the genes in the resulting list (Figure 2, B. Results). Output choices include various identifiers (Unigene, Pfam, InterPro, Ensembl, Entrez, GO), gene descriptors (gene symbol, gene name), domain descriptors (name, size, number), mapping information (chromosome, cytoband), transcription factor classification (Superclass, class, subclass) and BLAST homology information (*E*-score, GenBank Id and amino acid range of the original BLAST query sequence). The domain descriptors refer to the name and size of each domain as well as the number of different domains containing homology in each gene sequence. The GenBank Id and amino acid range of all original BLAST queries are also provided so that homology with any domain may be traced back to its original source. The *E*-score field provides a drop down menu of different *E*-score values allowing the user to set the *E*-score cutoff for the resultant output.

Elaborating on the *E*-score further, when designing the GeneSpeed Database we sought to allow a flexible and open environment allowing the user to have precise control over various selection criteria. Because GeneSpeed is a similarity-driven database, the most important criterion for the evaluation of domain sequences in the transcriptome is the expectation-score (*E*-score). The *E*-score describes the number of hits that one can expect to encounter by chance when searching by homology in a database. Consequently, the score is influenced by the size of the database, and by the size of the query sequence. For any particular hit, as the *E*-score approaches 0 (becomes smaller), the hit exhibits a more significant degree of similarity with the query. The purpose of an *E*-score cutoff is to exclude hits generated from the similarity search that do not have sufficient resemblance to the original domain of interest and thus do not represent a homologous domain. Unfortunately, due to the dramatic size differences between domain families, and the variations in conserved residues even between similarly sized domains, there is not a clearcut *E*-score cutoff that may be used for all domains and their corresponding hits in the database. Rather, we leave such critical decisions up to the user because both size and content characteristics of each domain must be taken into consideration when evaluating an appropriate *E*-score cutoff. Despite these complications, we offer some general guidelines for choosing this *E*-score cutoff value. In most cases, an inverse correlation exists between the domain size and *E*-score; therefore, a suitable cutoff may be selected using these criteria. In other less common situations, domain size may not be a suitable tool for selecting the *E*-score cutoff value. Factors such as domain content or taxonomy may have to be studied and evaluated. Explanations and recommendations for choosing an appropriate *E*-score cutoff may be found on the GeneSpeed website in the 'Background' section.

After the results have been displayed, the user has a choice of using other links and tools to further investigate the genes in their list (Figure 2, C. Sub-searches/Tools): (i) Domain Sub-search identifies all the domains with homology in a specified gene. (ii) InterPro Sub-search identifies all other genes in the GeneSpeed database within the currently selected species displaying homology to the specified domain. (iii) External Links provide links to other biological database sites. (iv) Novartis SymAtlas2, when working in the human or mouse section of the database, the average absolute expression, fold change above median and ANOVA tissue specificity statistical information pre-computed from the SymAtlas2 dataset may be accessed and visualized for genes contained within a gene list.

We have provided several additional resources on the GeneSpeed website to help users navigate and use the database with success. These include a section of 'Real Biological Scenarios', where we present possible scientific questions and full descriptions on how GeneSpeed may answer these questions. Similarly, we have included an FAQ page as well as a glossary of helpful definitions and descriptions. In addition, we have also included a tutorial section that provides several short tutorials on how to effectively use the tools within the GeneSpeed database.

DISCUSSION

We here describe the construction and use of GeneSpeed, a database providing an extensive homology compilation of all protein domains within the transcriptomes of human, mouse, worm and fly. At present, we have restricted this domain compilation to these four organisms. Although Unigene databases exist for many species (at present 76), the transcriptome sequencing effort for most is not approaching the saturated level of the four chosen organisms, and incorporating these additional species, although not limited in a practical sense, was deemed premature. These can later be added, as the sequencing effort and their respective genomics platforms become expanded and further utilized.

The homology assemblage employed in the creation of GeneSpeed is based on a computationally demanding batch tBLASTn population process, which to our knowledge has not been used previously for this purpose. When developing this population process it was necessary to choose an efficient homology detection algorithm. Others have shown that a hidden Markov model (HMM) based approach achieves better performance than gapped BLAST techniques for homology identification (8,9). We, on the other hand, have observed that by using a very large number of domains (each used as an individual query) derived from all types of species in the similarity search, a greater number of true hits is possible with the BLAST algorithm than HMM-based techniques (results to be published elsewhere). By using a large batch of query sequences from a diversity of organisms all representative of a particular protein domain family, it may be possible to more accurately reflect the natural diversity and evolution of amino acid combinations within that family (details of these experiments and their results are available on the GeneSpeed website).

Included with the domain analysis of these transcriptomes, GeneSpeed provides a toolset for the exploration of these domains, their protein families and the genes in which they reside. We have incorporated detailed information on each domain including a manual curation of their 'relatedness' with known transcription factors, GO classification, full name, average size, BLAST domain data and various protein domain identifiers (Pfam, InterPro and GenBank). Moreover, with the incorporated BLAST *E*-score cutoff tool, the user has control governing the amount of homology displayed when analyzing query results in order to accommodate any domain type or the specific biological question at hand. This is both a unique and indispensable feature of the database, as an appropriate cutoff may significantly fluctuate depending on the size of a domain, type of domain or even the taxonomic limitation/category of a particular protein domain. We have found it highly useful to be able to follow low levels of homology as this often carries certain functional relevance, which may link very distantly related members and often allows bridging between protein superfamilies. Such information is normally not available at other domain-based sites as only near-perfect matches are curated as true members (additional discussion on this topic is available at the GeneSpeed website). In essence, GeneSpeed provides users with a very 'domain-centric' online suite for analyzing the transcriptome. It is this domain focus that affords the GeneSpeed database its many unique benefits, such as providing the user the ability to infer protein type classification of a gene based on the composition and homology of its protein domain content. This is particularly noticeable in almost any GeneSpeed query result, where many of the resultant hits are genes that have not been classified in any of the common databases (GenBank, Entrez Gene, Unigene, etc.). As an example, a GeneSpeed query for the homeobox domain in human yields 217 genes with significant e-scores. Of these, 24 have not been characterized as homeoboxes and 16 have no previous classification associated with them. In such instances the domain content of these genes may be quickly and easily analyzed in the GeneSpeed environment, thereby revealing distinct clues about the possible classification of these unknown genes.

To our knowledge, few online biological knowledgebases are dedicated to the task of helping investigators define all genes in a particular proteomic category, which is both species specific and non-redundant, as well as linking this information to searchable gene expression information at this level of detail. Only recently have large databases such as NCBI and EMBL started to interlink genetic, functional and expression datasets. Even so, there are still no tools allowing the analysis of large user defined gene lists or the ability to save this information in a private account within the database of interest. Furthermore, there is no current resource where an investigator may study protein domain families in a non-redundant manner and within a select organism of choice. We have developed GeneSpeed in an attempt to fulfill these needs within a single database and provide access from a globally accessible website. We have used the GeneSpeed database with significant success in our own research not only to identify particular protein families based on domain content, but also to analyze these gene lists for expression level in our tissue of interest in order to define candidates for follow up in our laboratory. The success by which GeneSpeed has facilitated this process for our research should exemplify its potential value for other investigators.

ACKNOWLEDGEMENTS

Licensing: The GeneSpeed database is covered under a GNU Lesser General Public License (http://www.gnu.org/licenses/lgpl.html). We would like to thank the Chicago Project (www.chicagoproject.org) in helping defining the overall objectives of the GeneSpeed database. J.J. is supported by the American Diabetes Association. This work was also supported by NIH grant number P30 DK57516 (Diabetes and Endocrinology Research Center). Critical support was provided by the BCBC (Beta Cell Biology Consortium), DK-61248-03. Funding to pay the Open Access publication charges for this article was provided by NIH grant number P30 DK57516.

Conflict of interest statement. None declared.

REFERENCES

1. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
2. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
3. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
4. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
5. Ye,J., McGinnis,S. and Madden,T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res.*, **34**, W6–W9.
6. Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.*, **34**, D322–D326.
7. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
8. Madera,M. and Gough,J. (2002) A comparison of profile-hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.*, **30**, 4321–4328.
9. Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Hubbard,T. and Chothia,C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologs as pairwise methods. *J. Mol. Biol.*, **284**, 1201–1210.