Nucleic Acids Research, 2004, Vol. 32, Database issue ***D423–D426***
DOI: 10.1093/nar/gkh010

Nematode.net: a tool for navigating sequences from parasitic and free-living nematodes

Todd Wylie[1,*], John C. Martin[1], Michael Dante[1], Makedonka Dautova Mitreva[1], Sandra W. Clifton[1], Asif Chinwalla[1], Robert H. Waterston[1,2], Richard K. Wilson[1] and James P. McCarter[1,3]

[1]The Genome Sequencing Center, Washington University School of Medicine, St Louis, MO 63108, USA, [2]Department of Genome Sciences, University of Washington, 1705 NE Pacific Street, Seattle, WA 98195, USA and [3]Divergence Inc., 893 North Warson Road, St Louis, MO 63141, USA

Received August 13, 2003; Accepted August 21, 2003

ABSTRACT

Nematode.net (www.nematode.net) is a web-accessible resource for investigating gene sequences from nematode genomes. The database is an outgrowth of the parasitic nematode EST project at Washington University's Genome Sequencing Center (GSC), St Louis. A sister project at the University of Edinburgh and the Sanger Institute is also underway. More than 295 000 ESTs have been generated from >30 nematodes other than *Caenorhabditis elegans* including key parasites of humans, animals and plants. Nematode.net currently provides NemaGene EST cluster consensus sequence, enhanced online BLAST search tools, functional classifications of cluster sequences and comprehensive information concerning the ongoing generation of nematode genome data. The long-term goal of nematode.net is to provide the scientific community with the highest quality sequence information and tools for studying these diverse species.

INTRODUCTION

Nematodes or roundworms are members of an ancient phylum that accounts for perhaps four out of every five individual animals in the world (1). Parasitic nematodes infect nearly half the world's human population, resulting in significant morbidity and mortality. Nematodes also parasitize livestock and companion animals and cause over 80 billion dollars in crop damage annually (2,3). Nematode.net is a specialty database that makes accessible the rapidly expanding nucleotide sequence data and related resources from species across this phylum to target audiences including human/mammalian parasitologists, plant nematologists, *Caenorhabditis elegans* biologists and other scientists.

SEQUENCES FROM PARASITIC NEMATODES

Following the completion of the first fully sequenced animal genome, the nematode *C.elegans* (4), increasing efforts have been made to rapidly generate and make public gene sequences from parasitic nematodes of medical and economic importance as a route toward research on new anthelmintic drugs, vaccines, safe pesticides and resistant plants. Initiatives have primarily utilized expressed sequence tags (ESTs), focusing first on the filarial worms responsible for elephantiasis and river blindness (5,6). A collaboration is currently underway involving the Genome Sequencing Center (GSC) at Washington University in St Louis, the Wellcome Trust Sanger Institute, the University of Edinburgh and dozens of participating parasitologists to extend EST-based gene discovery to more than 30 nematode species (7,8). To date, over 295 000 ESTs have been generated from nematodes beyond *C.elegans*, with nearly 220 000 of these sequences provided by the GSC (Table 1).

NemaGene CLUSTERS AND NemaBLAST SEARCHES

While GSC-generated ESTs are immediately deposited in GenBank's database of ESTs (dbEST), no such repository exists for nematode EST cluster consensus sequences, nor are tailored BLAST searches easily performed. Nematode.net began in 2000 by providing these services. NemaGene clustering improves upon EST data by reducing data redundancy, increasing transcript length and improving base accuracy. The NemaGene method uses the Phred/Phrap/Consed suite of analysis programs (10), together with internal supplemental scripts, and has the advantage that clusters can be edited when necessary and tracked by name through multiple builds (11). Clusters can be searched on the nematode.net website by EST name, putative identity and individual contig or cluster name (Fig. 1). Cluster entries provide EST membership with NCBI links, as well as SWIR non-redundant protein database, Sanger Centre and *C.elegans* (Wormpep) homology. Cluster information and sequences can also be downloaded by FTP. NemaGene clusters have so far been generated for 15 species (Table 1). Both NemaGene clusters and individual ESTs can be searched for sequence identity using the online NemaBLAST tool, which utilizes a local WU-BLAST server (12) (http://blast.wustl.edu). Searches can be performed on ESTs from specific species, clades, stages and libraries, in any combination desired by the user.

*To whom correspondence should be addressed. Tel: +1 314 286 1114; Fax: +1 314 286 1810; Email: twylie@watson.wustl.edu

Table 1. Nematode EST projects by species

Clade	Nematode species	Host	Total ESTs	GSC ESTs	ESTs clustered	Clusters	Database
V	*Ancylostoma caninum*	Mammal	9331	9331	9286	4020	NemaGene
	Ancylostoma ceylanicum	Mammal	10651	10590	10590	3369	NemaGene
	Caenorhabditis briggsae	Free-living	2424	2424			
	Caenorhabditis elegans	Free-living	215202	388			Wormbase
	Haemonchus contortus	Mammal	21967	14014	5181	1970	NEMBASE
	Necator americanus	Mammal	4766		4766	2298	NEMBASE
	Nippostrongylus brasiliensis	Mammal	1234		1234	750	NEMBASE
	Ostertagia ostertagi	Mammal	7009	6558			
	Pristionchus pacificus	Free-living	8818	8818	4979	2603	NemaGene
	Teladorsagia circumcincta	Mammal	4313				
IVA	*Strongyloides stercoralis*	Mammal	11392	11335	10908	3311	NemaGene
	Strongyloides ratti	Mammal	14822	14822	8618	2941	NemaGene
	Parastrongyloides trichosuri	Mammal	7963	7963	4528	2155	NemaGene
IVB	*Globodera rostochiensis*	Plant	5934	5040	5039	2375	NemaGene
	Globodera pallida	Plant	1832				
	Heterodera glycines	Plant	20114	20109	4307	1790	NemaGene
	Heterodera schachtii	Plant	2662	2662			
	Meloidogyne arenaria	Plant	3519	3519	3321	1866	NemaGene
	Meloidogyne chitwoodi	Plant	10789	10789			
	Meloidogyne hapla	Plant	13869	13869			
	Meloidogyne incognita	Plant	13452	13168	5661	1625	NemaGene
	Meloidogyne javanica	Plant	5600	5578	5574	2598	NemaGene
	Pratylenchus penetrans	Plant	1928	1928	1926	420	NemaGene
	Zeldia punctata	Free-living	391	391	378	195	NemaGene
III	*Ascaris lumbricoides*	Mammal	1822				
	Ascaris suum	Mammal	39242	29960	19280	4262	NemaGene
	Brugia malayi	Mammal	26212	3773	18741	8392	NEMBASE
	Dirofilaria immitis	Mammal	4005	4005			
	Litomosoides sigmodontis	Mammal	873				
	Onchocerca volvulus	Mammal	14971	1230	7911	3504	NEMBASE
	Toxocara canis	Mammal	4889	4370			
	Wuchereria bancrofti	Mammal	2166				
I	*Trichinella spiralis*	Mammal	10 767	10548	10130	3454	NemaGene
	Trichuris muris	Mammal	3063		2125	1322	NEMBASE
	Trichuris vulpis	Mammal	2402	2402			
	Totals		510394	219584	144483	55220	

Nematodes with >100 ESTs are shown. NEMBASE clusters are available at www.nematodes.org. Clades are based upon (9).

FUNCTIONAL CLASSIFICATIONS AND OTHER FEATURES

Nematode.net provides the user with two avenues to explore the putative function of NemaGene clusters. Both are based on extrapolation from homology and must be regarded as providing only a starting hypothesis in studying function. Cluster sequences were used to search the Interpro protein domain database (13) (www.ebi.ac.uk/interpro) with InterProScan. Based on the presence of conserved domains, clusters were then mapped onto the Gene Ontology (GO) classification scheme (14) (www.geneontology.org). GO biological, molecular and cellular classifications are provided at nematode.net with the AmiGO interface. NemaGene clusters have also been mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG) database of biochemical pathways using enzyme commission (EC) numbers as the basis for putative assignment (15) (www.genome.ad.jp/kegg). Additional useful features of nematode.net include summaries of sequence status for all nematode species, cDNA library descriptions, project specifics, >300 organized nematology links and a trace viewer that allows users to examine raw sequence data. Nematode.net is also used to manage requests for clones generated by the project. Since 1999, 377 clones and dozens of plates have been provided to 37 investigators in 14 countries.

SITE AND DATABASE DESIGN

The Nematode.net interface was constructed using the Dreamweaver MX web development application in combination with a Perl CGI/DBI database interface. The GUI-based Dreamweaver MX editor was chosen for HTML design due to ease of use, ability to make rapid site-wide modifications and project tracking features. HTML pages written under Dreamweaver MX are sourced by a GSC Perl module, which has proved to be fast, extensible, and useful for recycling previously written code. Relational databases were initially built in MySQL and are now being replaced by a single, more efficient Oracle database.

Figure 1. A NemaGene Cluster Search query response showing constituents of consensus sequence by contig.

FUTURE DIRECTIONS

Nematode.net is a work in progress with the long-term goal of providing the nematology community with useful, consistent and lasting integrated databases and tools. With over 29 000 unique users in the past year, nematode.net is already providing a useful service, but improvements are envisioned in three areas. First, the site's current databases will be extended to include almost all available nematode species and sequences, expedited by further automation of clustering algorithms. Second, nematode.net will become more closely integrated with the *C.elegans* database Wormbase (16) (www.wormbase.org) and Nembase (www.nematodes.org), a site maintained by our collaborators at the University of Edinburgh that also provides tools for investigating nematode sequences (8). Plans for Wormbase integration include the layering of non-*C.elegans* nematode gene sequences over *C.elegans* homologs using the Distributed Annotation System (DAS) method (17). Currently, 9894 *C.elegans* genes have strong homologs in other nematodes (BLAST score of <1e-20). *C.elegans* information will continue to reside only at Wormbase. Third, in collaboration with Nembase, additional features for navigating nematode sequences will be made available. Databases covering all nematodes will include: postulated amino acid translations of EST clusters; protein domains connected to Pfam (18) and Interpro

including new nematode-specific domains; genes with homologs in *C.elegans* where RNA interference phenotype information is available (19); proteins with predicted signal peptide sequences; and codon usage tables for each species. Other possible additions include the integration of whole-genome information for parasitic nematode species (e.g. *Brugia malayi*) as such data become available.

ACKNOWLEDGEMENTS

Sequence generation has been aided by numerous collaborators in the nematology community, cDNA library creation by Claire Murphy and Brandi Chiapelli, and the dedicated members of the Darwin EST laboratory at the GSC. Wormbase efforts at the GSC are headed by John Spieth. We would like to thank our collaborators at NemBase, Mark Blaxter and John Parkinson, and others involved in Wellcome-Trust-funded nematode sequencing at the University of Edinburgh and the Sanger Institute. Additional feedback on website development was provided by Ben Oberkfel and Mike Nhan. Nematode.net and the parasitic nematode EST sequencing at the GSC is supported by US National Institute for Allergy and Infectious Disease grant AI46593 to R.H.W. and R.K.W. and National Science Foundation Plant Genome award 0077503 to S.W.C. and David M.Bird. J.P.M. was a Helen Hay Whitney/Merck Fellow.

REFERENCES

1. Platt,H.M. (1994) Foreword. In Lorenzen,S. (ed.), *The Phylogenetic Systematics of Free-Living Nematodes*. The Ray Society, London, pp. i–ii.
2. Blaxter,M. and Bird,D. (1997) Parasitic Nematodes. In Riddle,D.L., Blumenthal,T. Meyers,B.J. and Priess,J.R. (eds), *C. elegans II*. Cold Spring Harbor Laboratory Press, Plainview, NY, pp. 851–878.
3. Barker,K.R., Hussey,R.S., Krusberg,L.R., Bird,G.W., Dunn,R.A., Ferris,V.R., Freckmann,D.W., Gabriel,C.J., Grewal,P.S., Macguidwin,A.E., Riddle,D.L., Roberts,P.A. and Schmitt,D.P. (1994) Plant and soil nematodes—societal impact and focus for the future. *J. Nematol.*, **26**, 127–137.
4. The *Caenorhabditis elegans* Genome Sequencing Consortium (1998) Genome sequence of *Caenorhabditis elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.
5. Williams,S.A., Lizotte-Waniewski,M.R., Foster,J., Guiliano,D., Daub,J., Scott,A.L., Slatko,B. and Blaxter,M.L. (2000) The filarial genome project: analysis of the nuclear, mitochondrial and endosymbiont genomes of *Brugia malayi*. *Int. J. Parasitol.*, **30**, 411–419.
6. Unnasch,T.R. and Williams,S.A. (2000) The genomes of *Onchocerca volvulus*. *Int. J. Parasitol.*, **30**, 543–552.
7. McCarter,J.P., Clifton,S., Bird,D.M. and Waterston,R.H. (2002) Nematode gene sequences, update for June 2002. *J. Nematol.*, **34**, 71–74.
8. Parkinson,J., Mitreva,M., Hall,N., Blaxter,M. and McCarter,J.P. (2003) 400 000 nematode ESTs on the Net. *Trends Parasitol.*, **19**, 283–286.
9. Blaxter,M.L., De Ley,P., Garey,J.R., Liu,L.X., Scheldeman,P., Vierstraete,A., Vanfleteren,J.R., Mackey,L.Y., Dorris,M., Frisse,L.M. *et al.* (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature*, **392**, 71–75.
10. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
11. McCarter,J.P., Mitreva,M.D., Martin,J., Dante,M., Wylie,T., Rao,U., Pape,D., Bowers,Y., Theising,B., Murphy,C.V. *et al.* (2003) Analysis and functional classification of transcripts from the nematode *Meloidogyne incognita*. *Genome Biol.*, **4**, R26.
12. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
13. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
14. Ashburner,M. and Lewis,S. (2002) On ontologies for biologists: the Gene Ontology—untangling the web. *Novartis Found. Symp.*, **247**, 66–90, 244–252.
15. Kanehisa,M. (2002) The KEGG database. *Novartis Found. Symp.*, **247**, 91–103, 119–128, 244–252.
16. Harris,T.W., Lee,R., Schwarz,E., Bradnam,K., Lawson,D., Chen,W., Blasier,D., Kenny,E., Cunningham,F. and Kishore,R. (2003) WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res.*, **31**, 133–137.
17. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The Distributed Annotation System. *BMC Bioinformatics*, **2**, 7.
18. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
19. Kamath,R.S., Fraser,A.G., Dong,Y., Poulin,G., Durbin,R., Gotta,M., Kanapin,A., Le Bot,N., Moreno,S., Sohrmann,M. *et al.* (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, **421**, 231–237.