**NemaControl FAQ:**

**What are some examples of gene searches that I could do?**

NemaControl contains data from 17 different organisms for which the proteome is known. The helminth and platyhelminth species include: T. spiralis, B. malayi, M. hapla, M. incognita, C. brenneri, C. briggsae, C. elegans, C. japonica, C. remanei, P. pacificus, S. japonicum, and S. mansoni. Given the helminth and platyhelminth species have plant and mammalian hosts, two plant and two mammalian species were included: A. thaliana, G. max, H. sapiens, M. musculus, respectively. Much genetic and biochemical characterization have been done on D. melanogaster and S. cerevisiae, so these genomes were also included to provide more information to the orthologous groups.

Under gene name, the following examples of gene names could be used:

C. elegans 2L52.1

T. spiralis TS\_TRISPI\_Contig0.a.EN.1396

B. malayi 12798.m00022

M. hapla Mh10g200708\_Contig0\_102938\_103347

M. incognita prot\_Minc05158

C. brenneri CBN00001

C. briggsae CBG00001

C. japonica CJA00004

C. remanei CRE00001

P. pacificus GENEPREDICTION\_SNAP300000057131

S. japonicum Sjc\_000020

S. mansoni Smp\_078570

A. thaliana At1g01080.1

G. max Glyma0021s00410.1

H. sapiens ENSP00000375415

M. musculus ENSMUSP00000106833

S. cerevisiae YAL001C

D. melanogaster

**Can I search for multiple species at once?**

Yes, one or multiple boxes can be checked and genes from those species will be output based on the criteria specified in the output section.

**How can I find proteins that are orthologous to my protein of interest and how are orthologs determined?**

Species can be included or excluded from a search and only orthologous groups that comply with the criteria specified will be outputted. [Add in more information here after this is implemented]

With the large number of protein sequences and limited annotation, gene orthologs are very valuable, allowing some extrapolation of gene function, annotation, and structure to orthologs in other species. This sort of annotation is useful for finding potential drug targets in other species and also studying the evolution of genes. Orthologous groups were assigned using OrthoMcl version 2.01. OrthoMCL uses recipical best hits within each genome as recent paralog pairs and recipical best hits between two genomes as orthologous pairs. A Markov Clustering algorithm (MCL) is used to split clusters.

**What sort of functional classification information can I obtain for my genes of interest?**

Functional classification of the protein is based on GO, KO, and InterproID. The GO IDs classify a protein into a set of predefined terms (cellular component, molecular function, and biological process). The GO IDs are extremely useful for determining the function of the protein, which provides the first general hypothesis regarding a potential target lead2

The KEGG Orthology numbers (KO) were assigned by blasting the KEGG version XX [parameters]. KEGG orthology links proteins to a KO identifier, which can map to various types of proteins, including enzymes and signaling proteins. The KO number provides additional functional annotation about the type and function of the protein. The KO numbers can also be placed in metabolic and signaling pathways using KEGG pathways tools3.

InterproIDs4 provide a wealth of information regarding protein families, domains, and regions of proteins. The information is from a combination of different databases, such as ProDom, PROSITE, HAMAP, PRINTS, PANTHER, PIRSF, Pfam, SMART, TIGRFAMs, Gene3D, and SUPERFAMILY to determine the signature of the protein. This can also add in classification of distant proteins and also inferring their function from similar proteins.

 **How is essentiality of the genes determined?**

Essentiality is based on C. elegans RNAi data. Essentiality of genes in other species are inferred from the RNAi phenotype of the orthologous gene in C. elegans. For drug target discovery, the proteins with the most severe phenotype are more desirable candidates.

**How was homology to the PDB determined and how can extract the information from NemaControl?**

A solved crystal structure is a valuable asset when classifying proteins, determining structure and function, and designing drugs. For proteins that do not have a solved crystal structure, homology models can be generated from proteins in the PDB that are similar. BLAST was used to find protein structures within the PDB that had similar sequences to the nematode proteins a cutoff of XXX was used [blast parameters]. The PDB ID can be used to search for genes that have homology to that PDB. Alternatively, the PDB ID can be output if any of the genes in the search have homology to the PDB.

**How can I use this database to find potential anthelminth drugs to test in my laboratory?**

There are many currently available drugs that might have new indications for parasitic worm infections. DrugBank5 contains much chemical, pharmacological, and pharmaceutical information, coupled with details regarding the drug target and its sequence. Blast [list parameters here] was used to find targets in DrugBank that had homology to nematode proteins. This output option will provide a link to the DrugBank drug cards entry, as well as cheminformatic information that can be used to evaluate the quality of the drug. The output includes parameters determined by Christopher Lipinski that can be used to evaluate druglikeness and to also determine if the drug will be orally active in humans (absorption, distribution, metabolism, and excretion - ADME)6, including LogP, number of hydrogen bond donors and acceptors, and molecular weight. LogP is the octanol/water partition coefficient, which gives an indication of how hydrophobic the molecule is. According to Lipinski, LogP ≤ 5, number of hydrogen bond donors ≤ 5, number of hydrogen bond acceptors ≤ 10, and the molecular weight ≤ 500 Da have the most potential for yielding good drugs. A more recent study by Ghose et al7 determined that Lipinski’s rules should be expanded to include LogP between -0.4 to 5.6, molecular weight between 160 to 480, and molar refractivity between 40 and 130. Hydrogen-bond donors and acceptors, LogP, and molar refractivity were calculated using ChemViz Plugin (<http://www.rbvi.ucsf.edu/cytoscape/chemViz/>) within Cytoscape8.

The number of rotatable bonds is included in the output (also calculated using ChemViz), as a large number of rotatable bonds can also make a compound less biologically active and much more difficult to synthesize. The SMILES string is also listed, which can be used to find molecules that are similar to the molecule found in the present search. Finding similar molecules allows an SAR (structure activity relationship) study to be conducted, as various programs will search a database using the SMILES string alone using the Tanimoto Index Similarity score. Information regarding the status of the compound (approved, experimental, etc), current indication for the drug, and drug synonyms are also included in this output.

**Is this database useful for finding potential pesticides to test?**

Because NemaControl also includes plant parasites, the chemoinformatic information can also be expanded to pesticide discovery. Different parameters have been found to be specific for finding good agrochemicals: molecular weight ≤ 500; LogP ≤ 5; hydrogen-bond donors ≤ 3; and hydrogen-bond acceptors ≤ 129.

**What search functions can I use to narrow down or expand the number of potential drug targets?**

The number of potential drug targets can be reduced by including or excluding more species in the orthologous group search, specifying the essentiality of the protein, presence of the signal peptide, and whether the target is a druggable target. All of the information, including additional information in output options, can be output and further parsed.

**What is a Hopkins druggable target?**

Hopkins has identified proteins to which drugs that follow Lipinski’s rule of 5 bind. The proteins fall into six main classes: G-protein coupled receptors (GPCRs), serine/threonine and tyrosine protein kinases, zinc metallopeptidases, serine proteases, nuclear hormone receptors, and phosphodiesterases10. The classes are defined by interpro ID. If this box is checked, only proteins that have one of Hopkins “druggable” interpro IDs will be listed.

**How are the epitopes in the vaccine section determined?**

Epitopes are regions of a protein that are recognized by the immune system. Epitopes can be linear and unstructured or have a defined structural motif (the majority). Regions of disorder between 25 and 100 (determined by RONN) were mined to find regions of disorder that could be used for vaccine development. Coiled coils are a structural motif that could also be used for vaccine development. The output of Paircoil2 was mined for coiled coils between X and Y heptad repeats. Zinc-fingers, knottins, animal toxins, EGF modules, and collagens were mined based on a manually curated set of InterPro IDs. These structural domains could be mapped onto a small mini-protein that could be used as a vaccine. There are many examples discussed in a review by Corradin et al11.

**What is InDel information in the Output options section?**

Indels specific to nematodes and not found in other mammals(check) were found12. These indels can serve as unique drug targets for controlling a specific group of nematodes. If these indels are located in proteins essential for the survival of the specific group of nematodes, the drugs targeting these indels can be delivered with high specificity for the nematode protein.

**Why would I be interested in outputting protein-protein interaction data for drug development?**

Proteins rarely act in isolation and often interact with other proteins to fulfill their biological function, which results in complex protein-protein interaction (PPI) networks13. Discovery of important PPIs could lead to the next frontier of drug targets for parasitic nematodes. Despite the PPIs being challenging drug targets, there has been recent success with targeting PPI interactions by small molecules14. Because PPIs are of paramount importance to biological systems, this option allows PPIs for a gene of interest to be output. The PPIs are drawn from the curated MINT15 and IntAct16 databases which have experimental evidence for the interactions in C. elegans, D. melanogaster, and M. musculus. The PPIs from these species are related to other species via orthologous groups of proteins.

**Why would I be interested in whether my protein has a transmembrane or signal peptide and how are these features determined?**

The signal peptide and transmembrane region were assigned by running Phobius [http://phobius.sbc.su.se/] 17 on the complete proteome of all the species. Proteins with signal peptides provide an indication of proteins that may be secreted by the nematodes, allowing the nematodes to interact and survive within different environments. Proteins secreted by the nematode provide insight into host-parasite interactions which may yield important insights regarding how the parasite evades the immune system of the host. Transmembrane proteins include groups of chemoreceptors, ion channels, immune receptors, and proteins involved in energy metabolism. Several current anthelminth drugs target transmembrane proteins, making them an important for discovering new therapies. By checking this box, transmembrane and signal peptide information will be displayed for each protein using the following nomenclature [XXXX]

**Why would I be interested in learning about whether my protein has a coiled coil region and how are these features determined?**

Coiled coils are common structural motifs, consisting of 2 or more alpha-helices intertwined, found in all organisms. The heptad repeat is the hallmark of coiled coils, which consists of a repeating pattern a,b,c,d,e,f,g, where a and d are hydrophobic residues and e and g are charged residues. The motif is important for mediating protein-protein interactions, oligomerization, membrane fusion, and biological processes such as regulation of gene expression through various transcription factors. It is an important additional annotation step for analyzing genomes. Paircoil218 was used to find coiled-coil regions in all the nematode genomes in this study. Paircoil2 is a sequence-based method for identifying coiled coils.

**Why should I bother to look at the protein’s secondary structure?**

Many nematode proteins do not have X-ray crystal structures solved and also are not homologous to proteins in the PDB, so secondary structure prediction programs can aid to begin to understand the protein fold and provide a starting place for understanding and modeling tertiary and quaternary structure. In particularly divergent sequences, a fold may be very similar to a known fold, but the sequence may be too divergent to recognize this similarity from normal sequence comparison. Function may be elucidated based on information about a protein’s structure. Several secondary structure prediction programs were run because the programs are bias based on the data set on which the protein was trained.

**Are regions of disorder important? I’ve always just been interested in well folded proteins.**

Many proteins or regions of proteins that are unfolded in their native state have been found to be involved in molecular recognition, often undergoing disordered-to-ordered transitions to fold and form complexes with other proteins19. These interactions are typically highly specific, but have weak binding affinity. Accurately identifying disordered regions can aid when designing drugs that target a protein20 and also when determining the structure via X-ray crystallography. Checking the box provides the output from RONN21, a disorder prediction program [list parameters used to run]. The output is displayed as a raw file and in graphical format as an EPS file.

**REFERENCES**

**1 Chen, F., Mackey, A. J., Vermunt, J. K. & Roos, D. S. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One* 2, e383 (2007).**

**2 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25-29 (2000).**

**3 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28, 27-30 (2000).**

**4 Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res* 37, D211-215 (2009).**

**5 Wishart, D. S. *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 36, D901-906 (2008).**

**6 Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 46, 3-26 (2001).**

**7 Ghose, A. K., Viswanadhan, V. N. & Wendoloski, J. J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J Comb Chem* 1, 55-68 (1999).**

**8 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13, 2498-2504 (2003).**

**9 Delaney, J., Clarke, E., Hughes, D. & Rice, M. Modern agrochemical research: a missed opportunity for drug discovery? *Drug Discov Today* 11, 839-845 (2006).**

**10 Hopkins, A. L. & Groom, C. R. The druggable genome. *Nat Rev Drug Discov* 1, 727-730 (2002).**

**11 Corradin, G., Villard, V. & Kajava, A. V. Protein structure based strategies for antigen discovery and vaccine development against malaria and other pathogens. *Endocr Metab Immune Disord Drug Targets* 7, 259-265 (2007).**

**12 Wang, Z. *et al.* Systematic analysis of insertions and deletions specific to nematode proteins and their proposed functional and evolutionary relevance. *BMC Evol Biol* 9, 23 (2009).**

**13 Berg, T. Modulation of protein-protein interactions with small organic molecules. *Angew Chem Int Ed Engl* 42, 2462-2481 (2003).**

**14 Fletcher, S. & Hamilton, A. D. Targeting protein-protein interactions by rational design: mimicry of protein surfaces. *J R Soc Interface* 3, 215-233 (2006).**

**15 Ceol, A. *et al.* MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res* 38, D532-539 (2010).**

**16 Aranda, B. *et al.* The IntAct molecular interaction database in 2010. *Nucleic Acids Res* 38, D525-531 (2010).**

**17 Kall, L., Krogh, A. & Sonnhammer, E. L. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res* 35, W429-432 (2007).**

**18 McDonnell, A. V., Jiang, T., Keating, A. E. & Berger, B. Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics* 22, 356-358 (2006).**

**19 Shimizu, K. & Toh, H. Interaction between intrinsically disordered proteins frequently occurs in a human protein-protein interaction network. *J Mol Biol* 392, 1253-1265 (2009).**

**20 Cheng, Y. *et al.* Rational drug design via intrinsically disordered protein. *Trends Biotechnol* 24, 435-442 (2006).**

**21 Yang, Z. R., Thomson, R., McNeil, P. & Esnouf, R. M. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 21, 3369-3376 (2005).**