Published online 16 October 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D69–D76***
doi:10.1093/nar/gkm804

CMGSDB: integrating heterogeneous *Caenorhabditis elegans* data sources using compositional data mining

Amrita Pati[1,*], Ying Jin[1], Karsten Klage[2], Richard F. Helm[2], Lenwood S. Heath[1] and Naren Ramakrishnan[1]

[1]Department of Computer Science and [2]Department of Biochemistry, Virginia Tech, Blacksburg, VA 24061, USA

Received August 15, 2007; Revised September 16, 2007; Accepted September 17, 2007

ABSTRACT

CMGSDB (Database for Computational Modeling of Gene Silencing) is an integration of heterogeneous data sources about *Caenorhabditis elegans* with capabilities for compositional data mining (CDM) across diverse domains. Besides gene, protein and functional annotations, CMGSDB currently unifies information about 531 RNAi phenotypes obtained from heterogeneous databases using a hierarchical scheme. A phenotype browser at the CMGSDB website serves this hierarchy and relates phenotypes to other biological entities. The application of CDM to CMGSDB produces 'chains' of relationships in the data by finding two-way connections between sets of biological entities. Chains can, for example, relate the knock down of a set of genes during an RNAi experiment to the disruption of a pathway or specific gene expression through another set of genes not directly related to the former set. The web interface for CMGSDB is available at https://bioinformatics.cs.vt.edu/cmgs/CMGSDB/, and serves individual biological entity information as well as details of all chains computed by CDM.

INTRODUCTION

The availability of high-throughput screens has opened up awareness of the importance of data integration to reveal useful biological insight. For instance, the study of even a focused aspect of cellular activity, such as gene action, now benefits from multiple high-throughput data acquisition technologies, such as microarrays, genome-wide deletion screens and RNAi assays. While enormous quantities of data are available, it remains a major challenge to construe meaningful biological evidence from this data that explains, for example, the role of a biological pathway, the effects of a SNP on disease phenotypes or the regulatory networks or metabolic pathways underlying a cellular state. Two major factors make this process harder. First, high-throughput experiments for a given genome are performed by independent groups of researchers that develop their own naming conventions and schemes for information storage and retrieval. This makes it difficult for scientists to utilize 'all' available data for a genome to draw inferences. Second, even if such integration is accomplished, the possibility of linking data across sources is often restricted to individual entities, such as genes or proteins; it is difficult to track 'sets' of entities, which is the more natural way to interact with such databases.

As a case in point, consider the possibilities of integration opened up by the availability of RNAi screens. Post-transcriptional gene silencing via RNAi was first described in the nematode *Caenorhabditis elegans* (1), and is presently utilized for a variety of functional genomics experiments using RNAi assays. Although Wormbase serves as a centralized repository for *C. elegans* data, the sources of RNAi experiments in *C. elegans* are many, their data representation formats are varied and some information is lost while integrating them into the Wormbase (2) schema.

Here, we present CMGSDB, a database for computational models in gene silencing, where the following goals have been achieved. We have integrated genome annotation data, gene expression data, protein interaction data, gene regulation data, GO (Gene Ontology) annotation data and RNAi data for *C. elegans* into a centralized schema. RNAi experiments and phenotypes have been integrated from independent research groups into a single schema. A common hierarchical structure has been designed to organize the phenotypes from different sources. The hierarchy is available in the form of a web browser. Compositional data mining (CDM) (3) is used to identify relationships among sets of entities across the database schema, where these sets are mined automatically and not defined *a priori*. A detailed web interface that reports all the data and the patterns computed

*To whom correspondence should be addressed. Tel: +1 540 231 7857; Fax: +1 540 231 7040; Email: apati@vt.edu

is available at https://bioinformatics.cs.vt.edu/cmgs/CMGSDB/.

COMPOSITIONAL DATA MINING (CDM)

The basic idea in CDM is to mirror the shift-of-vocabulary as we traverse a database schema in a composition of data-mining algorithms that mine the respective entities and relationships. For instance, consider a multiple stress environment where numerous physiological responses are occurring simultaneously. Efforts to identify a set of *C. elegans* genes [perhaps encoding transcription factors (TFs)] to knock down (via RNAi) in order to ascertain key mechanisms of response might begin by identifying those genes whose knockdown produces phenotypes that modulate survival, and then find one or more TFs that combinatorially control the expression of these genes. This analysis can be modeled as a chain: TFs → genes → phenotypes. Each step in this chain is computed using a data-mining algorithm, so that we first mine the relationship between TFs and genes for concerted (TF, gene) sets called 'biclusters', then mine the relationship between genes and phenotypes to find concerted biclusters of (gene, phenotype) pairs. The biclusters share the gene boundary leading us to investigate if these biclusters approximately match at the gene interface. The projection of the biclusters with an approximate match at one interface is called a 'redescription'. Thus, CDM is a way of problem decomposition (see Ref. (3) for more details) where biclustering and redescription mining algorithms are chained in a way that mirrors the underlying 'join-order' path in the database schema.

As illustrated in Figure 1, we mine biclusters between genes and the TFs that regulate them, mine biclusters between genes and the phenotypes that result when they are knocked down, and relate one side of the first bicluster with one side of the second bicluster. Hence the task of integrating diverse data sources is reduced to composing data-mining patterns computed over each of the sources separately. The advantage of this formulation is that each data source can be mined individually using a biclustering algorithm that is suited for that purpose. For instance, the xMotif (4), SAMBA (5) and ISA (6) algorithms are suited for mining numeric data (e.g. such as gene expression relationships), while *a priori* (7) and CHARM (8) algorithms are suited for mining Boolean data (e.g. graph adjacencies).

The approximate matching of biclusters is ensured using a similarity search algorithm or redescription mining approach. This problem, in various guises, has been studied by the database community; see Refs. (9) and (10) for examples. In this article, we utilize a cover-tree approach for fast computation of similar biclusters. The overlap between the sides of biclusters is qualified using the Jaccard's coefficient: the Jaccard's coefficient between two sets X and Y is the ratio:

$$|X \cap Y|/|X^*Y|$$

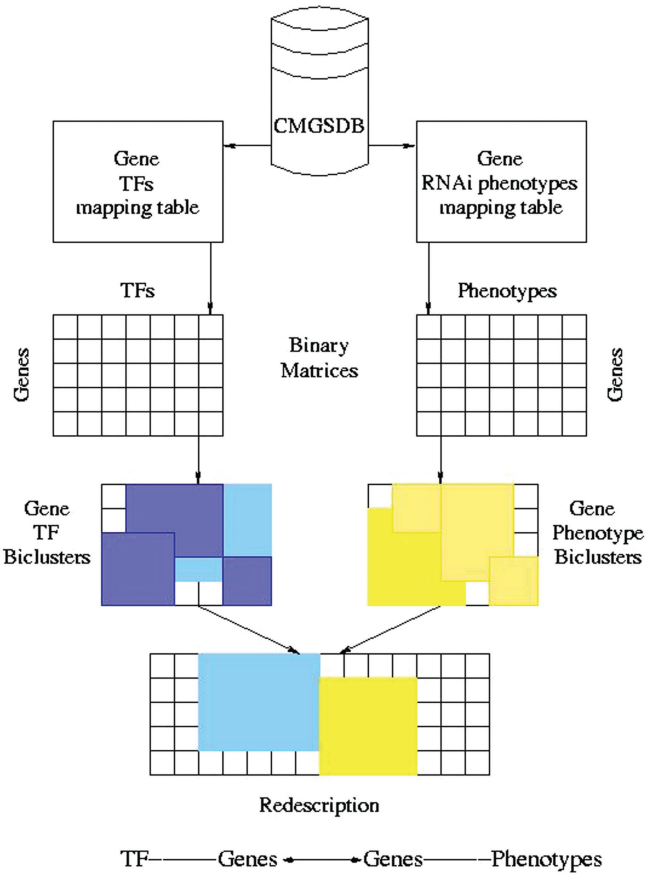

Figure 1. Finding TFs whose knockdown induces improved desiccation tolerance in *C. elegans*. Two biclusters (shaded rectangles) joined at the gene interface using a redescription between their projections. Below that is the CDM schema, displaying the sequence of primitives.

It is zero if the sets are disjoint and one if they are the same. In practice, we use a lax threshold on Jaccard's coefficient such as 0.5 and ensure that all similarities have a *P*-value significance of at least 0.001. Specifically, we use the hypergeometric distribution to assess the likelihood of observing a given Jaccard's threshold (given the sizes of X and Y) and use this probability to derive a *P*-value test.

Given a database schema and two entity sets participating in it, e.g. 'TFs' and 'phenotypes', we first identify the paths between these entity sets in the underlying E/R diagram of the schema. Observe that there can be many paths, including recursive ones (e.g. 'TFs regulate TFs which regulate other genes, contributing to phenotypes, when knocked down.'). Corresponding to each path, we instantiate a sequence of biclusterings and use the cover tree to identify redescriptions that can link them into chains.

CMGSDB DATA SOURCES AND METHODS

We refer to the biological entities captured in CMGSDB as 'biots'. CMGSDB contains exhaustive data about the following biots in *C. elegans*: chromosomes, genes, transcripts and proteins. For genes, extensive annotations (IDs, locations, names, annotations, locus and transcripts)

are complemented by microarray data, RNAi knockout experimental data, interaction data, gene regulatory information and functional categorization using the GO categories. Proteins, besides containing complete annotations, are enhanced by the addition of SwissProt/TrEMBL cross-references, physical structure details and properties and orthology/paralogy information. Finally, groups of all types of biots and biot information are linked together by patterns found by CDM, as described in the CDM section.

Data sources

Genome annotation data (chromosomes, genes, proteins, sequences, transcripts) for *C. elegans* are retrieved from Wormbase (2). Attention has been paid to retaining all transcripts and their respective constituting coding sequences for each gene. These transcripts serve as a link to gene expression data and RNAi transcript information. Gene orthology and paralogy data have also been taken from Wormbase.

Protein sequences and annotations have been obtained from Wormbase, while their physical properties and Protein Data Bank [PDB; (11)] homologs have been obtained from the Structural Genomics of *C. elegans* [SGCE; (12)] project. Protein interaction data and gene regulatory information have been obtained from BioGRID (13). Internal mappings from BioGRID IDs to Wormbase IDs have been generated.

Genome-wide gene expression data for 496 *C. elegans* microarray experiments have been collected from Stanford Microarray Database [SMD; (14)]. Expression values have been related to the genes through gene transcripts.

The RNAi component of CMGSDB is one of the chief characteristics that separates CMGSDB from other *C. elegans* resources. The RNAi experiments obtained from Wormbase have been supplemented by RNAi experiments retrieved from Phenobank (15), PhenomicDB (16) and RNAi phenome database (17). The same has been done for RNAi phenotypes. All RNAi phenotypes, thus obtained, have been organized into a hierarchical structure, with body, cell, development, lethal and sterile and miscellaneous as the top phenotypic categories. While Phenobank's experiments test all *C. elegans* genes for their role in the first two rounds of mitotic cell division, RNAi phenome database's experiments are aimed at evaluating the effects of RNAi on genes whose knock down causes embryonic lethality. PhenomicDB is a multi-organism phenotype–genotype database including human, mouse, fruit fly, *C. elegans* and other model organisms. Apart from these web-based RNAi data sources, there are a number of genome-wide RNAi screens in literature that are undocumented in these web-based sources but have been included in CMGSDB (18–36).

Database schema

The key components of CMGSDB are illustrated in Figure 2. Biots are contained in light green boxes, which are represented by one or more relations in CMGSDB.

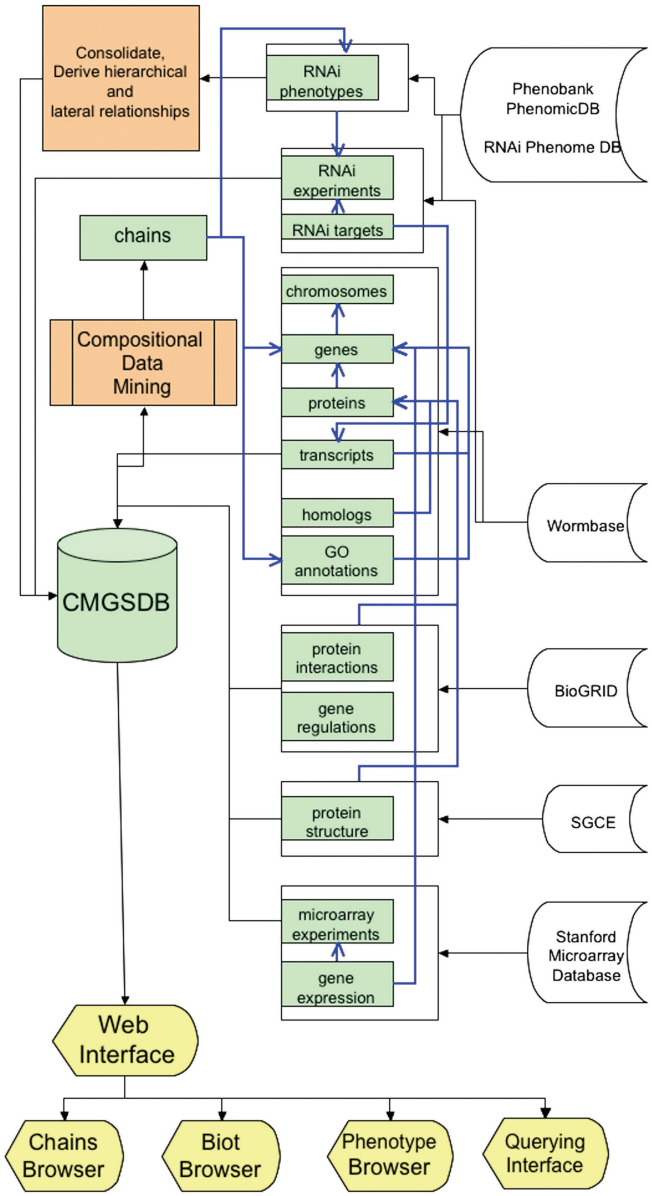

Figure 2. Data integration and analysis in CMGSDB.

Blue arrows represent relationships in CMGSDB. Plain black arrows represent data flow.

Applying CDM to CMGSDB

We applied CDM to CMGSDB as follows. There are a variety of biclustering algorithms that can be applied for mining relationships (37). For the purpose of this study, we utilized CHARM (38) to mine biclusters in binary relationships. For gene expression data, we utilized SAMBA (5) to mine biclusters.

Given a binary 0–1 matrix, the CHARM algorithm identifies sets of rows that show the same bit (0/1) patterns across all columns. The row set is grown to be maximal in size and, together with the columns for which the rows have a '1', defines the bicluster. CHARM identifies

overlapping biclusters, which can be organized alongside a lattice of subset relationships.

The SAMBA algorithm casts biclustering as a problem of finding bicliques in a bipartite graph. Given an edge-weighted graph (e.g. between genes and experiments labeled with expression levels) SAMBA detects dense subgraphs, which are then iteratively improved (using local addition/removal of vertices) in a post-processing phase.

Biclusters are connected if the overlap between the participating entities satisfied a Jaccard's threshold of 0.5. Chains computed in this manner all mediate through the gene entity set, since it serves a central role in CMGSDB (i.e. all relationships involve genes).

Patterns mined by CDM serve many purposes. For instance, they can be used to impute functions and properties to unannotated genes, they can make unexpected connections between upstream and downstream indicators, and they can summarize the distribution of data in the database more succinctly by identifying the sets of entities that dominate in many compositions.

QUERYING CMGSDB

The CMGSDB consists of a web interface and a PostgreSQL database management system. The web interface has been implemented using static and dynamic HTML, PHP, CSS and JavaScript. PostgreSQL is used to store the data described in the previous section and in Figure 2.

The web interface of CMGSDB can be used for querying. The user can search against all *C. elegans* biots. Genes, for example, can be searched using names, loci, transcript IDs and annotations. A biot page, apart from displaying basic information about that biot, also displays relationships with other biots that have been captured within CMGSDB. For instance, the phenotype page not only displays phenotype description, ID and source, but also shows existing relationships with other phenotypes, GO categories associated with the phenotype, RNAi experiments in which the phenotype was observed, genes whose knockdown resulted in the phenotype, and chains in which the phenotype participates. Biot pages are closely interlinked through biot IDs. As far as possible, biots are hyperlinked on pages. A biot page also contains hyperlinks to Wormbase and GO wherever applicable. Figure 3 illustrates the page for the *gpr-1* gene through a screenshot.

Chains, as described before, are available for searching and browsing. Chains can be queried by participating genes, number of common genes among all biclusters and number of biclusters. A chain with three biclusters containing gene *glp-1* is shown in Table 1.

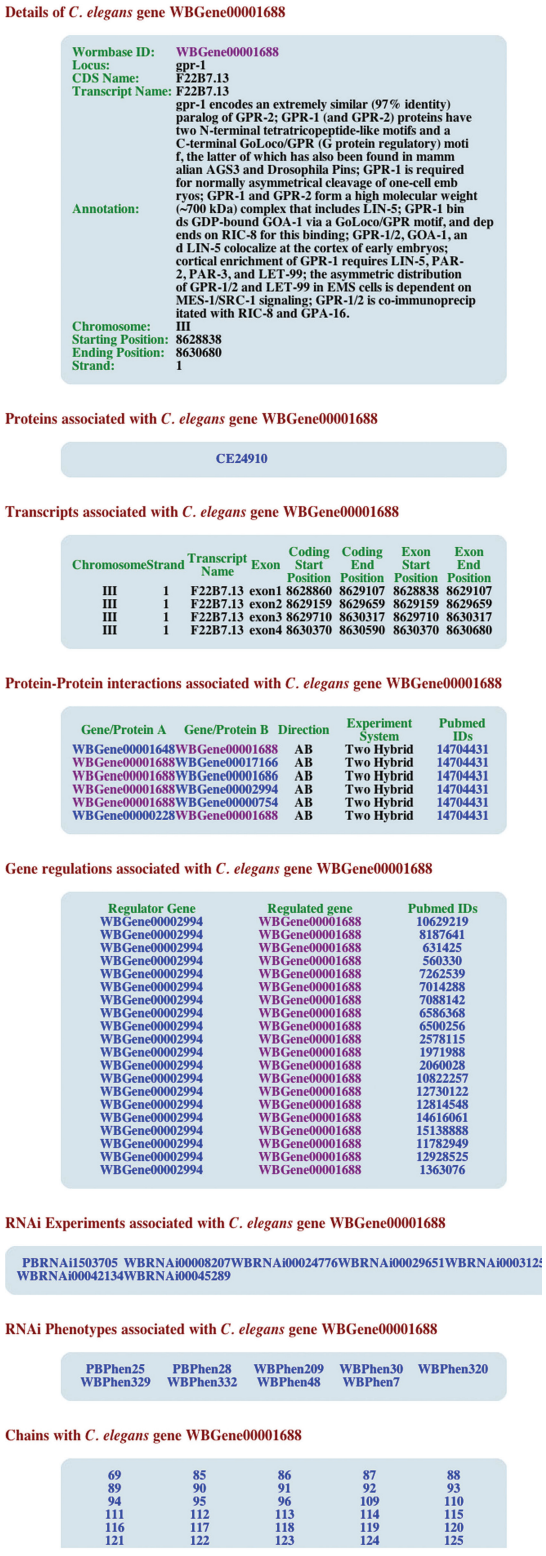

Figure 3. Screenshot of the gene page.

LIN-12/Notch signaling

In *C. elegans*, the LIN-12/Notch protein family mediates cell–cell interactions. *Glp-1* and *lin-12* encode two proteins in the LIN-12/Notch pathway, which is conserved in mammalian development. The two general cell–cell interactions that determine cell fate and involve these proteins are lateral specification and induction. Querying CMGSDB for *glp-1* gives two chains (chain 153 and chain 154). Table 1 illustrates chain 153, which demonstrates a chain of three (two non-trivial) biclusters. The biclusters with the GO categories and RNAi phenotypes

Table 1. Summary of chain 153 containing gene *glp-1*

Bicluster	Type	Set 1	Set 2
1	Gene-phenotype	*nmy-1, par-1*	PBPhen25 (Asymmetry of division), WBPhen30 (Embryonic lethal), WBPhen301 (Protruding vulva), WBPhen320 (Sterile), WBPhen326 (Sterile progeny), WBPhen7 (Asymmetry of division abnormal)
2	Gene-GO	*apx-1, glp-1, nmy-1, par-1*	GO:0002119 [Larval dev. (sensu Nematoda)], GO:0044464 (Cell part), GO:0009987 (Cellular process), GO:0048856 (Anatomical structure dev.), GO:0007389 (Pattern specification process), GO:0009790 (Embryonic dev.), GO:0009791 (Post-embryonic dev.)
3	Gene-gene	*glp-1, par-1*	*glp-1, par-1*

suggest that genes in this chain contribute to the structural aspects of cell division such as pattern specification leading to asymmetry of division, and these might be important to avoid embryonic lethality, protruding vulva and sterile progeny. Furthermore, this set of genes is likely to be self-regulated.

Four genes characterize the two chains: *par-1*, *apx-1*, *nmy-2* and *glp-1*. *Par-1* encodes a serine threonine kinase, which is required for the spatial regulation of GLP-1 asymmetry (39). *Par-1* is connected to *glp-1* through the GO and gene regulation blocks. *Apx-1* encodes a ligand homolog to the Delta protein of *Drosophila*. Both proteins contribute to the establishment of the dorsal-ventral axis in the early *C. elegans* embryo (40). Chains 153 and 154 suggest an interaction between *par-1* and *apx-1*. The likelihood of this prediction is further strengthened by the computational prediction of interaction between the same pair of genes (or their products) by Zhong and Sternberg (41). A putative gene in the Notch pathway is *nmy-2*, which encodes a maternally expressed non-muscle myosin II. The corresponding protein is linked through the phenotype bicluster containing *par-1*. The function of NMY-2 and PAR-5 is to together establish polarization in the *C. elegans* zygote along the anterior–posterior axis 23. In summary, *glp-1* and *par-1* interaction was already suggested, while *apx-1* and *nmy-2* represent new potential interactions with *glp-1* in the LIN-12/Notch pathway, uncovered through CDM.

Wnt pathway

The *Wnt* signal transduction pathway regulates diverse processes including cell proliferation, migration, polarity, differentiation and axon outgrowth in *C. elegans*. The signaling is composed of two pathways, the canonical *wnt*/BAR-1 pathway and the non-canonical *wnt*/WRM-1 pathway. A common component in both pathways is the HMG box containing protein POP-1, which is a member of the TCF/LEF family of TFs. The *wnt*-signaling pathway regulates the activation of the latter (42,43). CMGSDB reported 32 chains containing *pop-1*, the common target of the two *wnt*-pathways. These 32 chains suggested 18 new gene candidates (*daf-2*, *par-2*, *par-3*, *par-5*, *par-6*, *pkc-3*, *pkc-6*, *ooc-3*, *gpa-16*, *mbk-2*, *mes-1*, *csn-3*, *pgl-1*, *egl-46*, *tac-1*, *rab-5*, *tba-2*, *uri-1*) for the pathway. Of these, only *par-5* (chains 234, 236, 240) has been confirmed as a regulator of *pop-1* (44). *pop-1* is connected to *par-2* (chains 204, 206, 210, 212) through a regulatory network (45,46). Consistent with the results from CMGSDB, Zhong and Sternberg (41) predicted interactions between *par-2*, *mes-1* (chains 246, 248), a gene encoding a tyrosine kinase-like protein that is required for unequal cell division (47), *ooc-3* (chains 222, 224), encoding a protein required to establish asymmetrical anterior–posterior cortical domains and spindle orientation (48), and *gpa-16* (chains 234, 236), encoding a member of the G-protein *alpha*-subunit family of heterochromatic GTPase that effects spindle position and orientation (49). It can be hypothesized that PAR-2 is regulated by POP-1 over PAR-5. Further evidence shows that PAR-2 is regulated independently from the *wnt*-pathway, as it is not regulated by MOM-5 and MOM-2, the *wnt*-receptor and *wnt*-ligand, respectively (50). From the above gene list of 18 genes, CMGSDB suggests an interaction of *wnt*-proteins with the tyrosine kinase receptor DAF-2, which is involved in longevity and insulin signaling. This can be a potential link between *daf*-proteins and *wnt*-pathway proteins, indicating a possible connection between insulin and *wnt* signaling.

Some database statistics

In this section, we describe some basic statistics about the data in CMGSDB, especially focusing on data related to RNAi experiments and phenotypes and chains. Figure 4 illustrates some of the statistics of chains. Chains consisting of 3, 4 and 5 biclusters, number 2054, 1654 and 426, respectively. Figure 4 examines the distribution of the total number of genes in a chain and the number of common genes among all biclusters in a chain.

CMGSDB stores 81 722 RNAi experiments and 565 RNAi phenotypes. This includes 145 028 relationships between 21 222 unique *C. elegans* gene transcripts and the above 565 phenotypes.

PHENOTYPE BROWSER

In CMGSDB, phenotypes from several different sources have been organized into a common hierarchy. This hierarchy is available for browsing via a phenotype browser available at https://bioinformatics.cs.vt.edu/cmgs/CMGSDB/Treeview/index.php. The viewer has

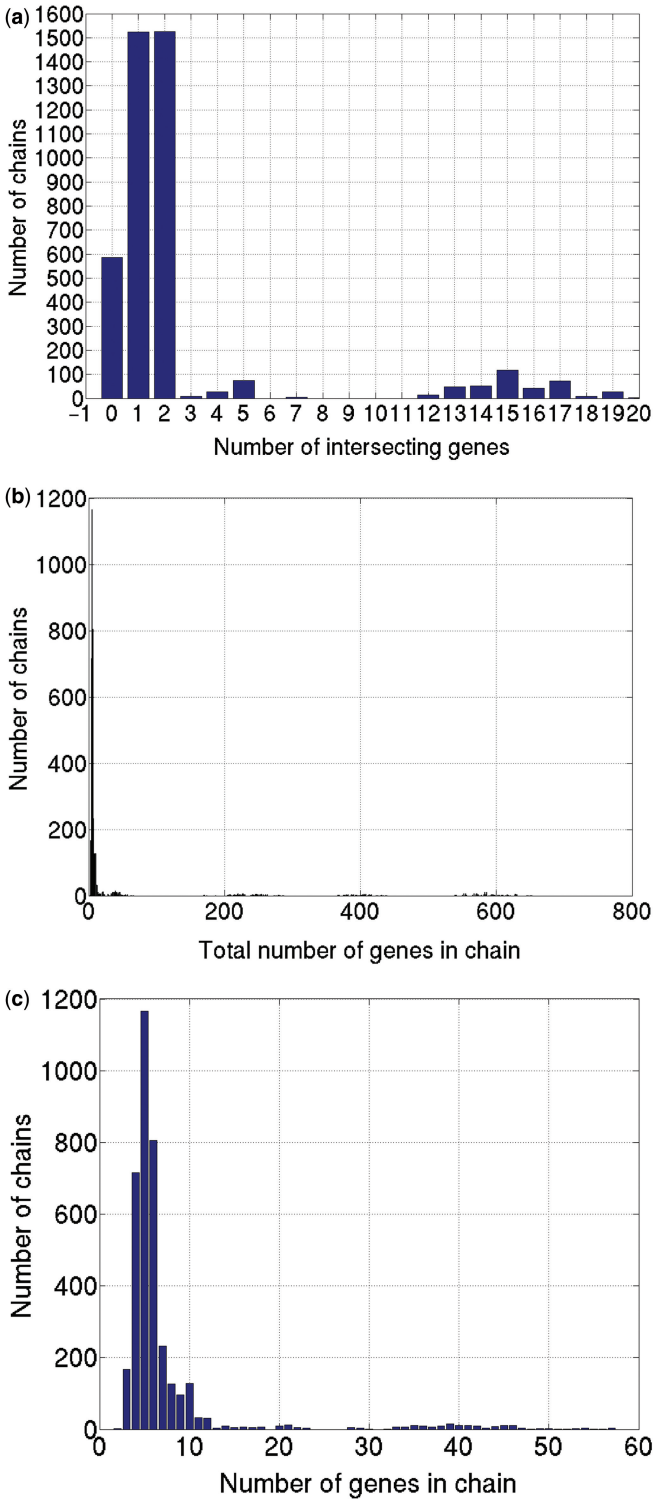

Figure 4. Statistics of chains. (**a**) Distribution of number of common genes in a chain. (**b**) Distribution of total number of genes in a chain. (**c**) Subset of (**b**).

been implemented using the PHP TreeView class and is dynamically linked to individual phenotype pages and to other biots. Figure 5 illustrates the phenotype browser with the tree view on the left.

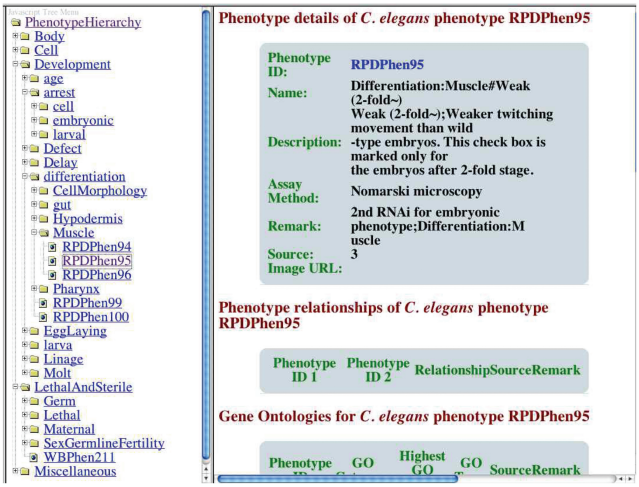

Figure 5. Screenshot of the phenotype browser.

DOWNLOADS

We have made the CMGSDB schema, scripts and raw data freely available under the GPL. Only the software for computing chains is not included. The download package is available at https://bioinformatics.cs.vt.edu/cmgs/CMGSDB/download.php.

Using this package, a user with proper hardware and software resources (including PostgreSQL and Perl) can locally set up an exact replica of CMGSDB's back end. The data is downloaded at runtime dynamically over the Internet. Scripts prepare the data and populate the database. This includes the integration of phenotypes from various sources.

All data in CMGSDB (except data related to chains) is available for download as flat files in download page.

CONCLUDING REMARKS

The integration of RNAi data and the application of data mining within CMSGDB provides the user with enhanced abilities to interpret raw *C. elegans* data. Unlike existing *C. elegans* resources, CMGSDB integrates RNAi data from multiple discrete sources. Using chains, users can discover new associations and relationships in the data that can be tested experimentally. A very meaningful future direction is to further consolidate the phenotypes to support alternate sets of phenotypes. This could be done by identifying very similar phenotypes as the same or by choosing a level of specialization in the phenotype tree. During the next 2 years of the CMGS project, additional data mining and modeling capabilities will be added.

ACKNOWLEDGEMENTS

This work is supported by NSF-ITR Grant-0428344 for the CMGS project. We also thank the two anonymous reviewers for their very useful suggestions. Funding to pay

the Open Access publication charges for this article was provided by NSF-ITR Grant-0428344.

Conflict of interest statement. None declared.

REFERENCES

1. Fire,A., Montgomery,M.K., Kostas,S.A., Driver,S.E. and Mello,C.C. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, **391**, 806–811.
2. Girard,L.R., Fiedler,T.J., Harris,T.W., Carvalho,F., Antoshechkin,I., Han,M., Sternberg,P.W., Stein,L.D. and Chalfie,M. (2007) WormBook: the online review of *Caenorhabditis elegans* biology. *Nucleic Acids Res*., **35**, D472–D475.
3. Jin,Y., Murali,T.M. and Ramakrishnan,N. (2007) Compositional Mining of Multi-Relational Biological Datasets. Tech. Report., TR-07-29, Computer Science, Virginia Tech. http://eprints.cs.vt.edu/archive/00000988/
4. Grothaus,G. (2005) Biologically-Interpretable Disease Classification Based on Gene Expression Data. Master's thesis, 2005, etd-05272005-145543, Virginia Polytechnic Institute and State University, http://scholar.lib.vt.edu/theses/available/etd-05272005-145543/.
5. Tanay,A., Sharan,R. and Shamir,R. (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18**, S136–S144.
6. Bergmann,S., Ihmels,J. and Barkai,N. (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E*., **67**, 031902.
7. Agrawal,R., Gehrke,J., Gunopulos,D. and Raghavan,P. (2005) Automatic subspace clustering of high dimensional data. *Data Mining Knowledge Discov*., **11**, 5–33.
8. Hsiao,C.J. and Zaki,M.J. (2005) Efficient algorithms for mining closed item sets and their lattice structure. *IEEE Trans. Knowledge Data Eng*., **17**, 462–478.
9. Nanopoulos,A. and Manolopoulos,Y. (2002) Efficient similarity search for market basket data. *VLDB J*., **11**, 138–152.
10. Sarawagi,S. and Kirpal,A. (2004) Efficient set joins on similarity predicates. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'04)*, Paris, France, ACM, New York, pp. 743–754.
11. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
12. Symersky,J., Zhang,Y., Schormann,N., Li,S., Bunzel,R., Pruett,P., Luan,C.H. and Luo,M. (2004) Structural genomics of *Caenorhabditis elegans*: structure of the BAG domain. *Acta Crystallogr. D Biol. Crystallogr*., **60**, 1606–1610.
13. Stark,C., Breitkreutz,B.J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*., **34**, D535.
14. Ball,C.A., Awad,I.A., Demeter,J., Gollub,J., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Matese,J.C., Nitzberg,M. *et al*. (2005) The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res*., **33**, D580–D582.
15. Sonnichsen,B., Koski,L.B., Walsh,A., Marschall,P., Neumann,B., Brehm,M., Alleaume,A.M., Artelt,J., Bettencourt,P. *et al*. (2005) Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. *Nature*, **434**, 462–469.
16. Groth,P., Pavlova,N., Kalev,I., Tonov,S., Georgiev,G., Pohlenz,H.D. and Weiss,B. (2007) PhenomicDB: a new cross-species genotype/phenotype resource. *Nucleic Acids Res*., **35**, D696–D699.
17. Maeda,I., Kohara,Y., Yamamoto,M. and Sugimoto,A. (2001) Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi. *Curr. Biol*., **11**, 171–176.
18. Cho,S., Rogers,K.W. and Fay,D.S. (2007) The *C. elegans* glycopeptide hormone receptor ortholog, FSHR-1, regulates germline differentiation and survival. *Curr. Biol*., **17**, 203–212.
19. Cram,E.J., Shang,H. and Schwarzbauer,J.E. (2006) A systematic RNA interference screen reveals a cell migration gene network in *C. elegans*. *J. Cell Sci*., **119**, 4811–4818.
20. Curran,S.P. and Ruvkun,G. (2007) Lifespan regulation by evolutionarily conserved genes essential for viability. *PLoS Genet*., **3**, e56.
21. Frand,A.R., Russel,S. and Ruvkun,G. (2005) Functional genomic analysis of *C. elegans* molting. *PLoS Biol*., **3**, e312.
22. Govindan,J.A., Cheng,H., Harris,J.E. and Greenstein,D. (2006) G-alpha o/i and G-alpha s signaling function in parallel with the MSP/Eph receptor to control meiotic diapause in *C. elegans*. *Curr. Biol*., **16**, 1257–1268.
23. Hamilton,B., Dong,Y., Shindo,M., Liu,W., Odell,I., Ruvkun,G. and Lee,S.S. (2005) A systematic RNAi screen for longevity genes in *C. elegans*. *Genes Dev*., **19**, 1544–1555.
24. Hansen,M., Taubert,S., Crawford,D., Libina,N., Lee,S.J. and Kenyon,C. (2007) Lifespan extension by conditions that inhibit translation in *Caenorhabditis elegans*. *Aging Cell*, **6**, 95–110.
25. Lamitina,T., Huang,C.G. and Strange,K. (2006) Genome-wide RNAi screening identifies protein damage as a regulator of osmoprotective gene expression. *Proc. Natl Acad. Sci. USA*, **103**, 12173–12178.
26. Lettre,G., Kritikou,E.A., Jaeggi,M., Calixto,A., Fraser,A.G., Kamath,R.S., Ahringer,J. and Hengartner,M.O. (2004) Genome-wide RNAi identifies p53-dependent and -independent regulators of germ cell apoptosis in *C. elegans*. *Cell Death Differ*., **11**, 1198–1203.
27. Nollen,E.A., Garcia,S.M., van Haaften,G., Kim,S., Chavez,A., Morimoto,R.I. and Plasterk,R.H. (2004) Genome-wide RNA interference screen identifies previously undescribed regulators of polyglutamine aggregation. *Proc. Natl Acad. Sci. USA*, **101**, 6403–6408.
28. Pothof,J., van Haaften,G., Thijssen,K., Kamath,R.S., Fraser,A.G., Ahringer,J., Plasterk,R.H. and Tijsterman,M. (2003) Identification of genes that protect the *C. elegans* genome against mutations by genome-wide RNAi. *Genes Dev*., **17**, 443–448.
29. Poulin,G., Dong,Y., Fraser,A.G., Hopper,N.A. and Ahringer,J. (2005) Chromatin regulation and sumoylation in the inhibition of Ras-induced vulval development in *Caenorhabditis elegans*. *EMBO J*., **24**, 2613–2623.
30. Saleh,M.C., van Rij,R.P., Hekele,A., Gillis,A., Foley,E., O'Farrell,P.H. and Andino,R. (2006) The endocytic pathway mediates cell entry of dsRNA to induce RNAi silencing. *Nat. Cell Biol*., **8**, 793–802.
31. Schmitz,C., Kinge,P. and Hutter,H. (2007) Axon guidance genes identified in a large-scale RNAi screen using the RNAi-hypersensitive *Caenorhabditis elegans* strain nre-1(hd20) lin-15b(hd126). *Proc. Natl Acad. Sci. USA*, **104**, 834–839.
32. Sieburth,D., Chong,Q., Dybbs,M., Tavazoie,M., Kennedy,S., Wang,D., Dupuy,D., Rual,J.F., Hill,D.E. *et al*. (2005) Systematic analysis of genes required for synapse structure and function. *Nature*, **436**, 510–517.
33. Stein,K.K., Davis,E.S., Hays,T. and Golden,A. (2007) Components of the spindle assembly checkpoint regulate the anaphase-promoting complex during meiosis in *Caenorhabditis elegans*. *Genetics*, **175**, 107–123.
34. Suzuki,Y. and Han,M. (2006) Genetic redundancy masks diverse functions of the tumor suppressor gene PTEN during *C. elegans* development. *Genes Dev*., **20**, 423–428.
35. van Haaften,G., Vastenhouw,N.L., Nollen,E.A., Plasterk,R.H. and Tijsterman,M. (2004) Gene interactions in the DNA damage-response pathway identified by genome-wide RNA-interference analysis of synthetic lethality. *Proc. Natl Acad. Sci. USA*, **101**, 12992–12996.
36. Vastenhouw,N.L., Fischer,S.E., Robert,V.J., Thijssen,K.L., Fraser,A.G., Kamath,R.S., Ahringer,J. and Plasterk,R.H. (2003) A genome-wide screen identifies 27 genes involved in transposon silencing in *C. elegans*. *Curr. Biol*., **13**, 1311–1316.
37. Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Comput. Biol. Bioinform*., **1**, 24–45.
38. Zaki,M.J. and Hsiao,C.J. (1999) CHARM: an efficient algorithm for closed itemset mining. *SIAM Int. Conf. Data Mining*, Tech. Report., RPI, 1999, 457–473, http://citeseer.ist.psu.edu/zaki99charm.html.
39. Crittenden,S.L., Rudel,D., Binder,J., Evans,T.C. and Kimble,J. (1997) Genes required for GLP-1 asymmetry in the early *Caenorhabditis elegans* embryo. *Dev. Biol*., **181**, 36–46.

40. Mello,C.C., Draper,B.W. and Priess,J.R. (1994) The maternal genes apx-1 and glp-1 and establishment of dorsal-ventral polarity in the early *C. elegans* embryo. *Cell*, **77**, 95–106.
41. Zhong,W. and Sternberg,P.W. (2006) Genome-wide prediction of *C. elegans* genetic interactions. *Science*, **311**, 1481–1484.
42. Lin,R., Thompson,S. and Priess,J.R. (1995) pop-1 encodes an HMG box protein required for the specification of a mesoderm precursor in early *C. elegans* embryos. *Cell*, **83**, 599–609.
43. Siegfried,K.R. and Kimble,J. (2002) POP-1 controls axis formation during early gonadogenesis in *C. elegans*. *Development*, **129**, 443–453.
44. Lo,M.C., Gay,F., Odom,R., Shi,Y. and Lin,R. (2004) Phosphorylation by the beta-catenin/MAPK complex promotes 14-3-3-mediated nuclear export of TCF/POP-1 in signal-responsive cells in *C. elegans*. *Cell*, **117**, 95–106.
45. Morton,D.G., Roos,J.M. and Kemphues,K.J. (1992) par-4, a gene required for cytoplasmic localization and determination of specific cell types in *Caenorhabditis elegans* embryogenesis. *Genetics*, **130**, 771–790.
46. Watts,J.L., Etemad-Moghadam,B., Guo,S., Boyd,L., Draper,B.W., Mello,C.C., Priess,J.R. and Kemphues,K.J. (1996) par-6, a gene involved in the establishment of asymmetry in early *C. elegans* embryos, mediates the asymmetric localization of PAR-3. *Development*, **122**, 3133–3140.
47. Berkowitz,L.A. and Strome,S. (2000) MES-1, a protein required for unequal divisions of the germline in early *C. elegans* embryos, resembles receptor tyrosine kinases and is localized to the boundary between the germline and gut cells. *Development*, **127**, 4419–4431.
48. Basham,S.E. and Rose,L.S. (1999) Mutations in ooc-5 and ooc-3 disrupt oocyte formation and the reestablishment of asymmetric PAR protein localization in two-cell *Caenorhabditis elegans* embryos. *Dev. Biol.*, **215**, 253–263.
49. Afshar,K., Willard,F.S., Colombo,K., Siderovski,D.P. and Gonczy,P. (2005) Cortical localization of the Galpha protein GPA-16 requires RIC-8 function during *C. elegans* asymmetric cell division. *Development*, **132**, 4449–4459.
50. Lee,J.Y., Marston,D.J., Walston,T., Hardin,J., Halberstadt,A. and Goldstein,B. (2006) Wnt/Frizzled signaling controls *C. elegans* gastrulation by activating actomyosin contractility. *Curr. Biol.*, **16**, 1986–1997.