Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems School of Computing and Information Systems

5-2011

BeeSpace Navigator: Exploratory Analysis of Gene Function using Semantic Indexing of Biological Literature

Moushumi Sen SARMA

David Arcoleo

Radhika S. Khetani

Brant CHEE

Xu Ling

See next page for additional authors

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

🗸 Part of the Bioinformatics Commons, and the Computer Sciences Commons

Citation

Sarma, Moushumi Sen, David Arcoleo, Radhika S. Khetani, Brant Chee, Xu Ling, Xin He, Jing Jiang, Qiaozhu Mei, ChengXiang Zhai, and Bruce Schatz. 2011. BeeSpace Navigator: Exploratory Analysis of Gene Function using Semantic Indexing of Biological Literature. Nucleic Acids Research doi:10.1093/nar/ gkr285

This Journal Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

Author

Moushumi Sen SARMA, David Arcoleo, Radhika S. Khetani, Brant CHEE, Xu Ling, Xin He, Jing JIANG, Qiaozhu MEI, ChengXiang ZHAI, and Bruce Schatz

Nucleic Acids Research, 2011, 1–8 doi:10.1093/nar/gkr285

BeeSpace Navigator: exploratory analysis of gene function using semantic indexing of biological literature

Moushumi Sen Sarma¹, David Arcoleo¹, Radhika S. Khetani¹, Brant Chee¹, Xu Ling², Xin He², Jing Jiang², Qiaozhu Mei², ChengXiang Zhai² and Bruce Schatz^{1,2,*}

¹Institute for Genomic Biology and ²Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Received January 28, 2011; Revised March 19, 2011; Accepted April 12, 2011

ABSTRACT

With the rapid decrease in cost of genome sequencing, the classification of gene function is becoming a primary problem. Such classification has been performed by human curators who read biological literature to extract evidence. BeeSpace Navigator is a prototype software for exploratory analysis of gene function using biological literature. The software supports an automatic analogue of the curator process to extract functions, with a simple interface intended for all biologists. Since extraction is done on selected collections that are semantically indexed into conceptual spaces, the curation can be task specific. Biological literature containing references to gene lists from expression experiments can be analyzed to extract concepts that are computational equivalents of a classification such as Gene Ontology, yielding discriminating concepts that differentiate gene mentions from other mentions. The functions of individual genes can be summarized from sentences in biological literature, to produce results resembling a model organism database entry that is automatically computed. Statistical frequency analysis based on literature phrase extraction generates offline semantic indexes to support these gene function services. The website with BeeSpace Navigator is free and open to all; there is no login requirement at www.beespace.illinois.edu for version 4. Materials from the 2010 BeeSpace Software Training Workshop are available at www.beespace.illinois. edu/bstwmaterials.php.

INTRODUCTION

Automatic classification of gene function enables computers to complement curators for new genomes, by developing new pipelines for annotating genomes. As sequencing becomes cheaper, task-oriented genome analysis will become common, studying dynamic expression of differential genes for particular tasks for particular organisms. Fixed microarrays are already being replaced for expression experiments by high-throughput sequencing using RNA-seq methodologies. The source of functional information still primarily resides in the biological literature, but automatic extraction is necessary due to the range of experimental tasks now feasible on organisms with unknown genetics. The original proposal for the BeeSpace project stated testing on honey bees the goal of 'biotechnology enables routine expression analysis and bioinformatics enables functional analysis unconstrained by pre-existing categories'.

The BeeSpace Navigator is a prototype software for interactive analysis of gene function, via semantic indexing of biological literature. The software enables users to create spaces, which are literature collections targeted toward the gene functions of a particular task. Available services analyze gene lists from expression experiments and summarize gene functions using descriptions extracted within specified spaces. These services support automatic analogues of curator processes of reading literature for creating databases, with accuracy high enough for practical utility. Using the analysis environment enables biologists to interactively transform unknown genes into known functions.

The semantic indexing enables interactive sessions to discover gene functions, using statistical operations on user-specified collections task specific for the user's experiments. Indexing details are discussed in separate papers

© The Author(s) 2011. Published by Oxford University Press.

^{*}To whom correspondence should be addressed. Tel: +1 217 244 0651; Fax: +1 217 244 1781; Email: schatz@illinois.edu

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/ by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

for the individual components as described below; this article focuses on their integration into an interactive system for functional analysis on community spaces. See Figure 1 for schematic of the services interacting with the spaces for BeeSpace Navigator. All services uniformly revolve around spaces, which represent a specific point of view for the specific task at hand. The paradigm of space manipulation is explicitly embedded into the user model of creating spaces, with Filter and Cluster, then navigating them to support the functional analysis, with Summarize and Analyze.

COMMUNITY SPACES

The BeeSpace Navigator is an analysis system for creating and traversing spaces. It represents the first fully fledged implementation of the Interspace, the next-generation information infrastructure for the Internet, where the fundamental paradigm is concept navigation of community spaces rather than document search of scientific literature (1). The technologies of the Interspace promise a new level of functionality in building analysis environments for functional genomics (2).

In the BeeSpace system, a '*space*' is a collection that is semantically indexed. Every service operates on spaces. Every collection of documents generated in any way can be saved as a space. The spaces are automatically named for user convenience, but can be renamed for special purposes. For example, a search query generates a space for the user with the search string as name. The spaces are used as the input to the services, for example, the Genelist Analyzer and the Gene Summarizer operates on the selected space.

The easy selection of the underlying space is a major difference between manual and automatic curation processes, between 'base' systems such as FlyBase and 'space' systems such as BeeSpace. This has to do with the underlying collection from which the function facts are extracted. With manual curation, there is a single static collection. In the FlyBase case, this collection is a selected subset of the MEDLINE articles that mention *Drosophila melanogaster*.



Figure 1. Interaction schematic of system usage, services interacting with spaces. A 'space' is set of scientific documents for a particular task. (a) Search (or filter) any system or user space and make new smaller spaces. (b) Use terms of interest as seeds to cluster together related terms, in a space; save the documents associated with a given cluster of terms as a smaller space. (c) Summarize functional information about a gene, in the context of a space. (d) Analyze gene lists in the context of a space to reveal significant 'concepts' and genes, resent; and select concepts and/or genes to make new smaller spaces.

With automatic curation, there can be multiple dynamic collections. That is, different collections can be used for extraction for different tasks. For example, a gene can be described with the Gene Summarizer using a genetic model (*Drosophila melanogaster*), the organism being studied (*Apis mellifera*) or the behavior being studied (early foraging). The gene will match different sentences and different functions may result. Thus, task-dependent curation can be achieved, using the automatic extraction from different collections.

The current master database is all MEDLINE abstracts, but other biological or scientific literature would work equally well. For example, previous versions of BeeSpace Navigator had substantial collections from Biological Abstracts and from beekeeping books. The statistical technologies for semantic indexing have been utilized across a variety of subjects over many years with good results (3,4); they rely upon consistent terminology for contextual redundancy.

Figure 2 illustrates the nascent BioSpace within the system, with spaces across all of biology. There are spaces derived from the Universe, which is currently all of MEDLINE. Some of these spaces focus on organisms, such as *Apis mellifera* or *Tribolium castaneum*, whose gene expression is being analyzed. Some of these spaces focus on behaviors, such as foraging, which represent specific tasks being studied. Note the foraging space includes this behavior within both insects and mammals. Finally, some of these spaces are broad categories, such as insect or mouse, primarily useful as standard platforms to leverage new spaces. These spaces are moderated by the administrator; user spaces can be migrated into system spaces as appropriate.

All system services utilize spaces, which are collections of documents generated by internal operations. This concept is a more general research version of the currently popular commercial concept of smart folders automatically generated with simple searches. The basic operations enable users to create spaces in various ways, and then use these spaces to analyze gene functions. A common user session is thus to generate an appropriate space describing the task at hand and then operate on this space with appropriate services. For example, search for foraging within insect collection then cluster to find the set of documents relating to foraging within social insects. After saving this set as a space, use that space to analyze the functions of genes from an expression experiment concerning honey bee foraging.

SPACE CREATION

Building new spaces from old is an important part of the new paradigm. As mentioned, users can subsearch within their previous spaces to create smaller spaces or utilize space arithmetic to merge together several spaces into a combined one. But users can also leverage the spaces found by other users. This feature represents a powerful form of information sharing. No login is required, since spaces are default to the guest account, but users can login to create spaces that are categorized within their own

BeeSpace Navigator, versio	on 4 - Mozilla Firefox					
Eile Edit View History Boo	okmarks Iools Help					
🔇 Der C X 🟠 🗋 http://workerbee.igb.uiuc.edu/8200/						Զ 🧟
BeeSpace Navigator, ver	Restance Navigator werein 4 +					
Logged in as schatz & Logged in as schatz	og out	Í				
Space Panel		?	Filter Cluster Summarize	Analyze		
Authors		<u>^</u>	Search:	R [Query expand]		
System	System	=	Author 🔺	Title	Source	Pub Date
abrudan	jenica		BYHThong : : KP Leong : : H	Insect venom hypersensitivity; experience in a clinical immunology/alleray service in Singapore.	Singapore medical journal	2005 Oct
Arcoleo	David		CHWu::MFLee::SCLiac	Sequencing analysis of cDNA clones encoding the American cockroach Cr-PI allergens. Homology with insect hemolymph	The Journal of biological chemistry	1996 Jul
beretta	ottavio		G J Bolter : : W Chefurka	Extramitochondrial release of hydrogen peroxide from insect and mouse liver mitochondria using the respiratory inhibitors of	Archives of biochemistry and biophysics	1990 Apr
BioSpace	BioSpace		D Jones : : H Turner : : V Chhol	Altered hexamerin regulation in precupal Trichoplusia ni pseudoparasitized by Chelonus sp. near curvimaculatus.	Archives of insect biochemistry and physiology	1996
Brown	Sue		D Knebel : : H Lubbert : : W Do	The promoter of the late p10 gene in the insect nuclear polyhedrosis virus Autographa californica: activation by viral gene p	The EMBO journal	1985 May
Burke	Pat		E A Abdallah : : M E Eldefrawi :	Pharmacologic characterization of muscarinic receptors of insect brains.	Archives of insect biochemistry and physiology	1991
chandrasekaran	sriram	-	ET Schmidtmann : : J E Llovd	Suppression of mosquito (Diptera: Culicidae) and black fly (Diptera: Simuliidae) blood feeding from Hereford cattle and poni	Journal of medical entomology	2001 Sep
•	III	- F	J Balzarini : : B Degreve : : S H	The multifunctional deoxynucleoside kinase of insect cells is a target for the development of new insecticides.	Molecular pharmacology	2000 Apr
Spaces		?	JL Dimarco : : D Zachary : : J/	Insect immunity: expression of the two major inducible antibacterial peptides, defensin and diptericin, in Phormia terranova	The EMBO journal	1990 Aug
A Remove	(Onerations		H Lisa R Scully : : Michael J Bido	An alternative insect pathogenic strategy in an Aspergillus flavus auxotroph	Mycological research	2009 Feb
			B N M Anthony :: J B Harrison ::	GABA receptor molecules of insects	EXS	1993
Id little		5120	PETeal:: RJ Nachman	Prolonged pheromonotropic activity of pseudopectide mimics of insect pyrokinin neuropectides after topical application or in	Regulatory peotides	1997 Oct
654 Universe	1	8874780	B Molina I M Lohse I Niet	Evaluation of a topical solution containing 65% permethrin against the sandfly (Phlebotomus permiciosus) in dogs	Veterinary therapeutics research in applied yet	2001
715 Universe : Apis mellifera	a 2	186	Ralf Heermann : Thilo M Fuch:	Comparative analysis of the Photorhabdus luminescens and the Yersinia enterocolitica genomes: uncovering candidate ge	BMC genomics	2008
716 Universe : Drosophila m	nelanogaster 6	8432	Richard G Voot : : Natalie E Mil	The insect SNMP gene family.	Insect biochemistry and molecular biology	2009 Jul
717 Flybase : Reference	4	7570	Author: Richard G Vogt : : Natal	lie E Miller :: Rachel Litvack :: Richard A Fandino :: Jackson Sparks :: Jon Staples :: Robert Friedman :: Joseph C Dicke	ens	
718 Universe : Tribolium cas	staneum 8	182	Title: The insect SNMP gene fan Source: Insect biochemistry and	nily. 1 malecular biology Google Scholar Pubmed		
719 Universe : Foraging	3	767	Pub Date: 2009 Jul	molecular unregy course contrat runned		
720 Universe : Insect	6	9154	Abstract: SNMPs are membrane	a proteins observed to associate with chemosensory neurons in insects: in Drosonhila melanoraster. SNMP1 has been show	in to be essential for the detection of the oberomo	ne ciswaccenvl acetate (CVA)
721 Universe : Mouse 424766		24766	SNMPs are one of three insect g	ene clades related to the human fatty acid transporter CD36. We previously characterized the CD36 gene family in 4 insect	Orders that effectively cover the Holometabola, or	some 80% of known insect
724 Universe : Mouse Mice		023807	species and the 300 million year. Diptera (D. melangaster, Drosoj topography within the SNMP clas. D. melanogaster and Ae. aegypti orthologues in the B. mori and C. the dipteran and lepidopteran line functional proteins. We observed Chemicals: insect Proteins : Re Genes: Keywords:	is of evolution since tims ameage enveyer. Leptopteria (e.g. comby): mon, Arthbriase polypherius, Mandica soda, Halohts high periodococcus, Adeva asyyti, Anophesis ganabae, Clark polymo, signadescianty, Hymericent Dys maliferta, Si high periodococcus, Adeva asysti, Anophesis ganabae, Clark polymo, signadescianty, Hymericent Dys maliferta, Si La well as IRAC derived genomics sequences from Ae, asystia as models for proposing corrected arguments of rothologues pipers og genomes. We then used these sequences to analyze the SNMP clark of the insect CDG gene family, support sages, and pausibly throughout the Holometabola and across a boad evolutionary time scale. Wel ⁴ ysent indirect evidence expansion of the <u>SNMP toxic-clark</u> in C. pipers q, and T. castaneous suggesting that the <u>SNMP time clark</u> and acquare captors, Phenomone	vreescens, reincoverpa assuita, Helicoverpa arms di o Coleoptera (Tholbium castaneum). This préviou e used cDNA sequences of <u>SINIP1</u> and <u>SINIP2</u> in the D. pseudobscura and An. gambaie genor g the existence of two well supported sub-clades, based on evolutionary selection (dNdS) that the c d functional role in these species.	ere, Marmestra brassicae); s study suggested a complex from various Lepidoptera species, les, and for identifying SIMIP2 and SIMIP2, throughout lipteran <u>SIMIPs</u> are expressed as
			B Rodrigo A Velarde : : Gene E R	Nuclear receptors of the honey bee: annotation and expression in the adult brain.	Insect molecular biology	2006 Oct
			🔳 Sigrid D Roosendaal : : Jan M \	Delipidation of insect lipoprotein, lipophorin, affects its binding to the lipophorin receptor, LpR: implications for the role of Lp	Insect biochemistry and molecular biology	2009 Feb
			🔳 Suparna Taneja-Bageshwar : : /	Biostable agonists that match or exceed activity of native insect kinins on recombinant arthropod GPCRs.	General and comparative endocrinology	2009 May
			🔳 Xiaowei Wang : : Lanzhu Ji : : C	[Potential effects of elevated carbon dioxide on forest leaf-feeding insects]	Ying yong sheng tai xue bao = The journal of ap	2006 Apr
			M d Dans The states			

Figure 2. System spaces provide standard starting collections for new task spaces within the BioSpace.

accounts, enabling them to save and manage their spaces. The privacy settings permit spaces to be hidden, but encourage openness for scientific progress. Indeed, building new research upon old research is the paradigm of science; this is directly represented within the space paradigm using community spaces as the knowledge unit.

The system services for space creation are general. Searching is a mechanism for partitioning spaces via the Filter service, there are also powerful features for clustering via the Cluster service. These have been optimized via parallel computation on shared memory servers, to operate interactively with appropriate-sized collections. The Natural Clusterer works from the bottom-up, partitioning on the basis of phrases occurring in the documents. Using small worlds algorithms, it operates interactively on collections in the range of 50-100 K documents (5). The Steerable Clusterer works from the top-down, partitioning on the basis of phrases weighted by words specified by the user for each cluster. Using language models algorithms, it operates interactively on collections in the 5-10 K document range (6). Individual clusters can be saved as spaces by the user to feed into other services for functional analysis.

The real-time interactive clustering is possible only because of modern computing technology. Clustering involves computing distance in an *n*-dimensional space, so speed is proportional to size of the memory. The production system for BeeSpace Navigator is hosted on a \$70 K parallel computer with 128GB memory at http:// workerbee.igb.uiuc.edu. This is double the memory of the \$7 M super computer that hosted Interspace Prototype for the first semantic indexing of MEDLINE a decade ago (7). The project server has fewer processors than the supercomputer did, 16 versus 128, but the software is largely memory bound on the shared-memory machine. The implementation gains its interactive speed by locking the pre-computed indexes into memory.

GENE LIST ANALYSIS

Genomics has progressed from single genes to multiple genes, as the wet lab technologies have improved with microarrays and next-generation sequencing. The dry lab technologies have to correspondingly progress from single genes to multiple genes, to support functional analysis. The Genelist Analyzer is the automatic computational analogue of a genetic classification database, just as the Gene Summarizer discussed next is the automatic computational analogue of a model organism database.

The Gene Ontology (GO) is a hierarchical classification for biological topics used by human curators to specify the functions of genes (8). The major model organism databases are committed to GO and use the GO categories to specify additional functions for the curated genes. Functional analysis for microarray experiments is commonly supported using GO, by translating differentially expressed genes into the closest model organism using ortholog computations and then looking up the major classifications of the corresponding genes in the classification.

The Gene Ontology has the same limitations as other biomedical classifications, being dependent on the literature curated by the model organism database projects. For example, the BeeSpace project performed large-scale microarray experiments on social behavior in honey bee. There was thus difficulty in obtaining good results from GO, since the closest model organism was fruit fly, an insect but one with little social behavior. Recent analyses of GO nodes have also shown that annotation quality varies across model organisms (9) and that functional coherence of hierarchical terms is not uniform (10).

In contrast, the BeeSpace Navigator uses a computational approach to gene classification based upon current terminology contained in biological literature. Just as full-text search can extract the most specific and most current documents, automatic curation can extract the most specific and most current categories. Since discriminating concepts are uniformly extracted from whatever terms are represented within the specified literature, there is no bias from existing classifications. The Genelist Analyzer thus trades accuracy for currency: it uses the free text of specified documents to extract concepts that frequently appear with genes.

The Analyzer transforms gene names into concept names (11). The concepts identified are those with differential expression against the background, just as the statistical model for differential expression in microarray experiments. But concepts are identified, rather than genes, and the background is the specified space that describes the current task being analyzed. The ability in BeeSpace for space-oriented computations enables some of the accuracy loss in automatic extraction to be overcome, since the concepts are extracted only from the documents within that space. Just as different backgrounds produce different expressions with microarray experiments, different spaces produce different lists of concepts given the same lists of genes.

BeeSpace Navigator is an interactive system with selectable spaces, where the background for differential expression can be rapidly changed, providing support for dry lab experiments. Since every document collection created within the system can be transformed into a space, sets of concepts can be sweep selected and the documents associated with these concepts saved into a space. The saved space is then fully usable within the system and can be filtered or clustered to produce relevant subsets. For example, the most frequent concepts or the most frequent genes that are differentially expressed from a microarray experiment can be saved into a space, and then this space clustered using Steerable Cluster to find the documents most relevant to the behavior task. This dynamically created space can then be used as background for gene list analysis.

GENE SUMMARIZATION

Once a gene list is analyzed within a space, individual genes can be described using functional information contained within the articles of that space. BeeSpace began with automatic reproduction of FlyBase summarization (12). In FlyBase, gene descriptions are generated from database entries specified manually by human curators, who read literature articles to extract facts. The automatic curation process in BeeSpace Navigator was developed into a production service called the Gene Summarizer (13). This takes a gene name as input and produces relevant sentences describing the gene's functions as output.

During early system usage, it became clear that within an interactive system, showing the sentences from which the facts were extracted was more useful than simple summaries, since it enabled the biologist users to fully interpret the document context. The context for utility for biologists is necessarily more detailed than that for curators since they are generating hypotheses rather than extracting facts; the language processing is necessarily deeper than interactive support for FlyBase curators (14).

The categorization for the summarizer relied upon a semi-structured machine learning algorithm, derived in part from those for newspaper summarization. Development required new algorithms for deriving and training to classify sentences into functional categories (15). These categories were initially taken from FlyBase summaries, later refined by the curators themselves through our collaboration with the FlyBase project. The categories of gene function are: Protein Domain/ Structure, Homologs/Orthologs, Expression Patterns, Regulatory Elements, Phenotype/Function, Genetic/ Physical Interactions and Population Genetics.

The original plan was to have FlyBase curators provide training sets through our collaboration. But in comparing the recognitions of the professional curators to those of the practicing biologists who were working closely with us on developing the analysis system, we discovered that they differed substantively. Curators wanted only unambiguous facts, whereas biologists wanted useful observations. For example, a sentence speculating about a function was eliminated by the curators as uncertain but marked by the biologists as interesting.

We chose those biologists closest to our projected user population, graduate students and postdocs in entomology and neuroscience, as trainers to judge the categories of sentences. Semantic indexing using the trained recognizers was performed during 8 hour batch jobs on the National Science Foundation Cloud Computing Testbed hosted at the University of Illinois at Urbana-Champaign, making BeeSpace its largest internal user. Later, we were able to develop separate software for question answering (4), not included in version 4, which extracted entities and relations to generate biological facts, so closer in semantics to human biologist curators.

EXPLORATORY ANALYSIS OF GENE FUNCTION

The software has gone through four versions over 6 years and has been used by over 100 biologists, including local researchers, workshop trainees and meeting attendees at Arthropod Genomics Symposia. The production Version 4 was early released in January 2010 and fully released in June 2010.

Training materials are available at www.beespace.uiuc. edu/bstwmaterials.php, including overview talks and sample sessions from the Training Workshop held in April 2010. The software itself is a mix of C, Python and Java. The user interface runs within a web browser using ExtJS javascript, an AJAX toolkit. A recent version of Mozilla Firefox or Google Chrome is recommended for best performance. No login is required for usage as guest, but if the user logs in, then generated spaces are saved under their login for future utilization.

This interface enables all the services to be used in concert with each other, demonstrating the utility of a space-based system. Every service can create a new space during the interactive session, whenever a relevant collection of documents is created. Thus the whole of the system is greater than the sum of the parts. In particular, community sharing of discovered knowledge about gene function is an integral part of the system, since biologists can reuse existing spaces.

As an example session, following are representative screendumps of the analysis performed by the first author, a BeeSpace biologist postdoc, for her gene expressions on behavioral maturation in honey bee (16). Only a few key screens are shown; for a full step-by-step usage, see the sample sessions in the Supplementary Data, with each step annotated.

The gene list extracted from these experiments and used below in the session can be found at www.canis.uiuc.edu/ ~schatz/GAtest.txt. The gene list must consist of record numbers from model organism databases, here FlyBase (FB), but also including Mouse Genome Informatics (MGI) and Saccharomyces Genome Database (SGD).

In the session, the biologist did multiple Filter operations with space arithmetic to build a task space on behavior maturation containing 46 340 of the 69 154 documents for insect. She created separate spaces for social insects (sociality within insect space) and for behavioral maturation (maturation within animal space), then intersected these custom spaces into a task space.

Now that the background is selected, the exploratory analysis of gene function can commence. Figure 3 shows the Analyze service where the gene list from the expression experiment has been entered into the box and the concepts within the background space analyzed. The most discriminating concepts are rank ordered in the left lower pane. These are the terms that differentially occur most often in documents with genes from the list mentioned as opposed to terms in other documents. The genes themselves are rank ordered in the right pane, with names translated from the FlyBase ortholog numbers specified as input, as described elsewhere (11).

To examine the results, the top genes are selected, taking the 10 most frequent after skipping the first two as artifacts. A save is then issued, to create a space of the documents containing these genes within the background space. The new space resulting from the top genes differentiating the gene expressions within the maturation task can now be used for further exploration. Figure 4 shows the Cluster service operating on the 2314 documents. At this scale, the biologist can partition semantically along desired lines using the Steerable Clusterer. Here the user partitions along the lines of the top concepts discovered during analysis, namely 'shock', 'diapause', 'repair', and adding the important 'hormone'. Note diapause was discovered automatically, and is relevant to behavioral maturation in response to environmental conditions. Within the 'hormone' cluster, underlined terms such as JH for Juvenile Hormone were

BeeSpace Navigator, version 4 - Mozilla Fir	refox						
Eile Edit View History Bookmarks Tools	i <u>H</u> elp						
C X & C http://workerbeeigb.uiuc.edu/8200/						습 - 🛛 - Yahoo	۹ 🥵
BeeSpace Navigator, version 4 +							
🐣 Logged in as schatz 🤱 Log out 🛫 Seni	d feedback						
Space Panel		? Filter Cluster Summarize	Analyze				
Matthews Kathy		List View Table View					
Munoz-Torres Monica		2					
Newman Katie							
Ronge Matt		Genes:					
Sanders Barry		FBgn0028373 FBgn0031453					
Sanogo Osee		FBgn0052048					
Schatz Bruce		E FRan0037121					
Toth Amy							
Woodard Sarah		-					
۲. III				Analyze n	Tabulate		
Spaces		?		- d	<i>y</i> <u> </u>		
CREMOVE T Rename Operations		Significant Concepts	Significant Concepts Genes				
ld Title	Size	Create space			Create space		
241 Universe : fish	85163	- concept	significance -	cluster	name	gene	found 👻
242 Universe : fish : behavior	4304	shock	2000	6	* chic	FBan0000308	1000
244 Universe : fish : behavior : aggression	225	polyacrylamide	2000	8	b	FBqn0000153	440
245 Universe : fish : aggression	289	leishmaniasi	2000	4	Sulf1	FBqn0040271	383
246 Apis Mellifera : forage	30	hsp	888	6	Ance	EBan0012037	333
422 Mouse Mice Mus musculus : aggression	attack fig 85	diapause	884	6	east	FBgn0010110	308
423 Drosophila melanogaster : foraging	89	= sodium	813	8	Co25C	FBqn0000299	243
424 Universe : forage	3756	electrophoresis	761	8	18w	FBgn0004364	220
588 Mouse Mice Mus musculus : risk aggres	sion 18333	dsma	750	5	Hsc70-4	FBgn0001219	204
589 Universe : fish : behavior: Mouse Ortholog	g Short 13	heparan	702	5	Hsp70Aa	FBgn0013275	204
648 Insect : behavioral maturation	1393	tchsp70	592	6	Rh6	FBgn0019940	156
649 Insect : merge(behavioral maturation)	46340	repair	550	6	Hsp83	FBgn0001233	152
650 Flybase : Reference : maturation	21914	photoreceptor	540	0	Tsf1	FBgn0022355	111
651 Insects : maturation (Query expand)	27779	mai	522	5	ACXD	FBgn0040507	98
652 Insects : maturation (Query expand) : top	Concept: 1222	repressor	492	5	gro	FBgn0001139	58
653 Insects : maturation (Query expand) : top	Concept: 34	stress	492	6	Shal	FBgn0005564	54
668 Universe : insects : maturation	511	simulium	477	4	Нор	FBgn0024352	48
669 Universe : insect insects insecta	87545	courtship	476	5	fru	FBgn0004652	41
670 Universe : insect insects insecta : insect	insecta 69811	N N Page 1 of 1 b	20	Displaying 1 - 500 of 500	IN N Page 1 of 1 b	N	Displaying 1 - 50 of 50
671 Universe : insect insects insecta : insect	insecta: 13028	The standard of the Pl	~	supplying 1 500 01 500	In a ruger out p	71 10	supplying 1 be of be

Figure 3. Analyze transforms a gene list into a list of discriminating concepts within a space.

BeeSpace Navigator, version 4 - 1	Mozilla Firefox							
Eile Edit View History Bookman	arks Iools Help							
🔇 💽 - C 🗙 🏠 🗋 http	tp://workerbee.igb.uiuc.edu:82	00/	☆ - Vah	🟠 - 🞯 - Yahoo 👂				
BeeSpace Navigator, version #	4 *							
🐣 Logged in as schatz 🤱 Log out	Send feedback							
Space Panel		? Filter Cluster Summarize Analy.	lyze					
Matthews	Kathy	Steerable Natural						
Munoz-Torres	Monica							
Newman	Katie	17						
Ronge	Matt	Cluster 1: shock						
Sanders	Barry	Cluster 2: diapause						
Sanogo	Osee	Charles 2						
Schatz	Bruce	= Cluster 3: repair						
Toth	Amy	Cluster 4: hormone						
Woodard	Sarah	+ Cluster 5:						
•		Cluster 6:						
Spaces		7						
				Cluster				
Kendine W Op	perduons	Create space		Author Title	Source Pub Date			
Id Title	Size			B Eddides 11 L Insecticidal expectice of constically on	sincered broudering. Applied and aminones. 1992 Mar			
683 Insect	69154	Clusters		Kara II as :: I Exercision of anistroom (Collingator)	and us malt inhibit. The Journal of experim 2002 Jan			
683 Insect 69154		shock, heat, hsp70, hsp90, hsc70, stress, temperature, promoter		Rara 5 Lee : . 1 Expression of crustacean (caninectes Review Manager	sapidus) molt-innibit The Sournal of experim 2002 Jan			
696 Universe : Anic mollifera : fora	111eiiiieia 2215 20	diapause, diapausing, prepupae, upregulated, cold, upregulation, pupae, scorpion		Shinii Nagata : Cloning and characterization of ecuryou	Shinii Nagata : Cloning and characterization of ectiveone receptor and ultr: Annals of the New Yor 2005 Apr Shinii Nagata : Cloning and characterization of ectiveone recentor and ultr: Annals of the New Yor 2005 Apr			
687 Linsarea : Anis mallifara : fora	age Ullaiveree Anie 2186	repair, atpase, na,k, queens, workers, ouaba	repair, atpase, na,k, queens, workers, ouabain, antifreeze, ie1		C Tang : WI Melanular characterization and everyonian of mandifular or The Biochemical journ. 1999 Oct			
201 Universe : Apis mellineral : torage O Universe : Apis 2100		hormone, juvenile, ecdysone, hormones, jh, p	hormone, juvenile, ecdysone, hormones, jh, pith, ecdysteroid, steroid		M Hathaway : Characterization of bayamarin proteins and their mPNAs in Comparathe biochemic 2009 Ney			
701 Universe : Insect : maturation (Query expand) 27779		bacillus, thunngiensis, west, dorsal, bactenal, aedes, ainca, com		D Schelling : Eunctional identification of the transmi	D Schelling : Functional identification of the transcription start site and the Biochemical and bioch 1995 Sen			
703 Linkerse : Insect : maturation	Ton Concente 1222	slik, electron, lipid, tatty, collagen, odorant, p	pheromone, temales	Containing Functional inclution of the database	S Shou M T A DNA-binding motein tint involved in inventie homoneure. Insert biochemistry on 2006 San			
704 Universe : Insect : maturation:	Top Genes 5904	enzyme, ace, kinase, angiotensin, converting	ig, peptides, inhibitors, peptidyl	Author: S Zhou : : M Tejada : : G R Wyatt : : V K Walk	Author: S Zhou :: M Tejada :: G R Wyatt :: V K Walker			
765 Universe : Insect : maturation. Top Genes 5904		alpha, receptors, beta, receptor, modopsin, n	memorane, toil, like	Title: A DNA-binding protein, tfp1, involved in juvenile ho Source: Insect biochemistry and molecular biology	Title: A DNA-binding protein, tfp1, introled in juvenile homone-regulated gene expression in Locusta migratoria. Source: insect biochemistry and molecular biology Pub Date: 2005 Sep Abstract: A pantially paindomic 15-nt. sequence upstream from a juvenile homone-regulated gene (<u>hp21</u>) was previously identified in the Advance migratoria. This sequence was proposed as a juvenile homone (<u>Hp1</u> response element (<u>HERE</u>) and a protein that hourd to it, as a transcription factor (<u>ED</u>) Ayaest stram was constity free domains for undernet copies of the JHER and after transferion with a ADVA bibary made to fat			
767 Insert hebaioral maturation 1303		nost, populations, plant, plants, tube, populat	ation, variation, races	Pub Date: 2006 Sep				
767 Insect behavioral maturation" adult invente 10101		alpha, collagen, subunit, 2, procollagen, proly	ayi, beta, anopheles	Abstract: A partially palindromic 15-nt, sequence upstro				
769 Insect : "behavioral maturation" adult usenile : "bet 8762		wrus, pol, nov, skin, neterochromatin, swine,	, viral, rar	previously identified in the African migratory locust, Loc				
772 Insect : maturation adult adult juvenile : Del 0102		malaria, control, tsetse, disease, capsid, res	sistance, lever, insecticides	was constre ¹ ted containing four tandem copies of the J				
773 Insect : maturation adult : leishmania sc hsn19.5 n. 360		toli, immune, pathway, senne, antimicrobial, innate, recognition, gram		bodies from vitellogenic females, yeast one-hybrid expe	bodies from vitellogenic females, yeast one-hybrid experiments yielded sequences for four putative binding proteins.			
774 Insect : maturation adult : fru.	ACXD Tsf1 18w St 646			encodes a 35kDa protein. This was designated the	appears to have a leucine zipper motif and a lipid-binding			
776 Universe : Insect : maturation	(Query expand): chic 2175	12		motif. <u>Recombinant tfp1</u> bound to <u>JHRE</u> in electrophoret the binding reaction resulted in the disappearance or sh	ic mobility shift experiments and addition of tfp1 antibody in ift of TF. We suggest that JH induces the association of			
777 Flybase : Reference : maturat	tion (Query expand): 1 1691			pre-existing proteins, including tro1, to form an active co	mplex, which binds to the JHRE upstream from jhp21 and			
778 Insects : merge(behavioral m	aturation): Hsp83 Rh 3647			regulates its transcription. Chemicals: DNA-Binding Proteins : Juvenile Hormones	: Transcription Factors			
incige(bendvioral me	turation): Tan Canan 2214			14 4 Page 1 of 6 > > 3	Topics 1 - 25 of 102			

Figure 4. Cluster partitions a space into groups of related documents that each can become a space.

automatically recognized as genes during the semantic indexing (17).

If the biologist wishes to discover the function of genes discovered during the exploratory analysis, she can summarize what knowledge is known within the literature. Figure 5 shows the Summarize service, where it has been implicitly invoked by clicking on the gene name underlined as juvenile hormone. The Gene Summarizer then simulates the curation process for the specified gene on the specified collection. In this case, the specified collection was the task space that resulted from exploratory analysis of the experiment for gene expression. The computed categories of gene function include gene phenotypes and gene interactions. The automatic curation enables a task-centered analysis of gene function rather than an organism-centered, which fits better with experiments such as this on dynamic expression for animal behavior.

The process of concept navigation can be continued by opening a document such as highlighted one on *Manduca sexta* and clicking on an underlined gene name, such as EcR for Ecdysteroid Receptor. Now that the biologist has explored from bees to moths, she could re-analyze the expression gene list within the new context of ecdysteroid receptors by simply creating a space from the summarizer results and clicking the Analyze tab to reinvoke the Analyze service with the same gene list but on the new space. Exploratory analysis of gene function is an iterative process of concept navigation through biological literature.

As another example, another BeeSpace postdoc was investigating gene expression of maternal behavior in

paper wasp (18). Maternal behavior is little studied in insects so running Analyze against *Drosophila* space was little help. After trying analysis with fly orthologs from the wasp genes, she instead generated mouse orthologs from the same wasp genes. A task-oriented space was created from the mouse literature relevant to maternal behavior using Filter of 'maternal' query against mouse space. Then Cluster was done on the 9618 documents within Mouse Maternal space, using the Natural Clusterer this time since the functional category was not known. The 33 resulting documents from one relevant cluster included an article dealing with regulation by a heat shock factor gene of maternal expression in the oocyte (19).

The postdoc insect biologist found this article in the mouse literature of significant interest. Her GO analysis did not identify heat shock as discriminating term for her gene list, possibly due to the wide distribution of heat shock entries across their conceptual hierarchy. The more focused automatic analysis discovered this concept relation. Typically, navigations with appropriate orthologs against task-oriented spaces produce novel documents of interest to the biologist. Besides insects, the BeeSpace Navigator has helped biologists working on fish, pigs and plants.

CONCLUSION

As the examples above show, the effective process of exploring function via concept navigation leverages existing manual classifications and existing biologist descriptions. The model organism databases are essential

BeeSpace Navigator, version 4 - Mozi	illa Firefox			D X
Eile Edit View History Bookmarks	Iools Help			
🔇 🖂 - C 🗙 🏠 🗋 http://w	vorkerbee.igb.uiuc.edu:8200	0/	☆ - Øt- Yahoo	P 😫
BeeSpace Navigator, version 4				-
🐣 Logged in as schatz 🤱 Log out 🤿	7 Send feedback			
Space Panel		? Filter Cluster Summarize	e Analyze	
Matthews Kath	ıy	Search: invenile hormone	 ج (م x)	
Munoz-Torres Mon	lica			
Newman Kati	e	Protein Info: 3 results		
Ronge Matt		Homolog Info: 0 results		
Sanders Barr	У	Expression Pattern: 0 rest	alts	
Sanogo Ose	e	Regulatory Element Info:	to construct the second s	
Schatz Bruc	:e	Regulatory clement into.	i resus	
Toth Amy	1	Phenotype and Function In	ifo: 59 results	
Woodard Sara	ah	🛫 id	sentence	
< III		▶ 3 10	The juvenile hormone (JH) binding activity of the extracts was measured using a hydroxyapatite assay and the purification progress was monitored by native gel chromatography and sodium dod	Jecyl sulfate-
Spaces		· · · · · · · · · · · · · · · · · · ·	A prohormone (P1) of locust adipokinetic hormone I (AKH I) is shown here to be a homodimer of a 41 residue subunit called the A-chain.	
		B 12	Juvenile hormone binding proteins (JHBPs) were extracted from the haemolymph of adult desert locusts, Schistocerca gregaria, and Mediterranean field crickets, Gryllus bimaculatus.	
Viendve L Rename W Operau	Unis	H 13	Prothoracicotropic hormone (PTTH) of the silkworm, Bombyx mori, was purified and its primary structure determined for the most part.	-
Id Title	Size	3 14	Crustacean molt-inhibiting hormone (MIH), a polypeptide produced by neurosecretory cells in eyestalk ganglia, suppresses the synthesis of ecdysteroid molting hormones by paired Y-organs.	
692 Insect	47570	▲ 15	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis analysis of proteins from infected cells revealed several intracellular proteins and two major secreted proteins which reacted with antib	odies to auth
663 Insect	69154	⊞ 16	Eclosion hormone (EH) is a moderate-sized peptide that acts on the insect nervous system to release the stereotyped behaviors involved in shedding the old skin (eclosion) as well as to activate	e new reflex
696 Universe Anis mellifers (forest	2215	3 17	The equilibrium dissociation constants for juvenile hormone I and II were approximately 6 x 10(-10) M at 4 degrees C, while racemic juvenile hormone III displayed an equilibrium dissociation con:	stant of 1.9 >
600 Onverse . Apis mellifera . forage	30		In order to study the potential role of the steroid molting hormone (20-hydroxyecdysone) in regulating molt-induced claw muscle atrophy, full-length cDNAs encoding relinoid-X receptor (GI-RXR)	and E75 earl
701 Universe : Apis mellitera : torage U	Universe : Apis 2100	⊞ 19 ⊳	In the model insect Manduca sexta the production of ecdysteroids by the prothoracic gland is acutely controlled by a brain neurohormone, prothoracicotropic hormone (PTTH).	
701 Universe Insect Instantion	500	B 20	Allatectomized animals treated with juvenile hormone III (JH) or methoprene caused the appearance of vitellogenin in the hemolymph, but had no effect on hexamerin levels.	
702 Universe : Insect : maturation (Que	(yexpand) 2///9	3 21	We obtained a full-length cDNA encoding diapause hormone (DH) and pheromone biosynthesis activating neuropeptide (PBAN) in Samia cynthia ricini based on both reverse transciptase-PCR (I	RT-PCR) and
703 Oniverse : Insect : maturation: Top	Concepts 1222	3 22	Gonadotropin releasing hormone (GnRH)-immunolike peptides were not identified, but progress is being made in the isolation and characterization of factors which stimulate cAMP production by	y the ovary.
704 Oniverse : Insect : maturation. Top	Genes 5904	H 23	Prothoracicotropic hormone of the silkworm, Bombyx mori: amino acid sequence and dimeric structure.	
765 Universe : Insect : maturation Top C	Joncepts 1094	3 24	Within the 2.2 kb region between hsp23 and gene 1 of the small heat shock gene locus 67B1 of Drosophila melanogaster, an approximately 1 kb perturbation of the chromatin architecture has p	previously ber
767 Insect : behavioral maturation	1393	B 25	The naturally occurring enantiomer of juvenile hormone III (10R) was converted to juvenile hormone III acid and then covalently attached to aminohexyl-Sepharose 4B.	
768 Insect . Benavioral maturation add	In juvernie 10101	3 26	Moreover, the prewandering peak of the plasma enzyme juvenile hormone esterase (JHE) was blocked by CsPDV injection.	
769 Insect : benavioral maturation adu	Jit juvenile : Der 0/02	Interaction Info: 55 result	5	
772 Insect : maturation adult	0020	Population Constice Info-	A results	
773 insect - maturation adult - tersman	na se nspia.5 p 360	ropulation ocnetics into a	r solico	
776 Universe : Insect : maturation (Que	p, ren, low, or 040			
777 Character Background Code	Ty expandy. Cinc 21/5			
779 Insects manage behavioral maturation (C	tion) Here? Db 2647			
700 Insects : merget benavioral matural	aony, rispos, kn 3647			
102 insect : merge(benavioral maturation	onj: top Genes 2314			

Figure 5. Summarize describes the gene functions of specified gene within a task-specific.

to turn expressed genes into gene names, which can then be correlated within the biological literature to extract gene functions. BeeSpace is the first major system to build spaces upon bases, to utilize scalable semantic indexing to automatically extract functional concepts and categories. Space creation is a uniform operation, enabling community sharing to build new collections of related documents from old.

Such systems will be used to support hypothesis generation to rapidly compare gene function using different spaces to discover new classifications. Currently, most gene lists are generated via microarrays specialized to particular organisms. But the rapid increase in sequencing power is changing this paradigm to one where gene lists for specific tasks can be generated with general RNA-seq (20). Even an early system such as BeeSpace is just as effective for RNA-seq as for microarray, since it deals only with gene names and gene lists. Exploratory analysis by knowledgeable biologists leveraging existing databases is the viable path towards understanding gene function in the new era of dynamic expression of organism behavior. The interconnected space of biological knowledge will be created by functional analysis from insects such as fly and bee, beetle and wasp, then move to InsectSpace and ArthropodSpace then AnimalSpace and PlantSpace en route to creating the BioSpace.

The future system will change its name to BioSpace Navigator, reflecting this more general mission. As noted, the software has no embedded knowledge of bees or insects, merely general text mining. Future plans are to integrate the navigator functionality described in this production version into a fully fledged gene function discovery environment. Enhancements will include enhanced speed for services and periodic updating of sources, importing gene lists and exporting document sets, computing orthologs and linking to genome browsers.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online, giving a step by step interactive session with BeeSpace Navigator.

ACKNOWLEDGEMENTS

The Informatics team thanks the Biology team, especially leader Gene Robinson and his laboratory including Amy Toth and Nyla Ismail. Other Computer Science graduate students who contributed include Yuanhua Lv, Yanen Li and Yue Lu. Other Institute for Genomic Biology (IGB) professional staff on the project contributed include Matt Ronge, Chris Fields and Jim Buell. We thank the administration at IGB especially the computer support led by Dan Davidson. We thank the Cloud Computing Testbed in the Department of Computer Science, especially lead faculty Indranil Gupta. The administrators of the insect databases provided expertise and support in many ways, and we especially thank William Gelbart of FlyBase at Harvard, Susan Brown of BeetleBase at Kansas State, Christine Elsik of BeeBase at Georgetown. Creating this new system with new research and development would have been impossible without all these people and many others.

FUNDING

National Science Foundation (0425852) in the Frontiers Integrative Biological Research (FIBR) program from September 2004 to August 2010, entitled 'BeeSpace: An Interactive Environment for Analyzing Nature and Nurture in Societal Roles'. Additional funding for software development and open access provided by United States Department of Agriculture National Research Initiative Arthropod Biology Tools (2009-35302-05285, to B.S. PI as with NSF FIBR grant).

Conflict of interest statement. None declared.

REFERENCES

- 1. Schatz, B. (2002) The Interspace: concept navigation across distributed communities. *IEEE Comp.*, **35**, 54–62.
- Schatz, B. (2002) Building analysis environments: beyond the genome and the web. Special Issue on Mining Information for Functional Genomics. *IEEE Intel. Sys.*, 17, 70–73.
- 3. Chen, H., Martinez, J., Ng, T. and Schatz, B. (1997) A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system. J. Am. Soc. Inform. Sci., 48, 17–31.
- 4. He,X., Li,Y., Khetani,R., Sanders,B., Lu,Y., Ling,X., Zhai,C. and Schatz,B. (2010) BSQA: integrated text mining using entity relation semantics extracted from biological literature of insects. *Nucleic Acids Res.*, 38, W175–W181.
- Chee,B. and Schatz,B. (2007) Document clustering using small worlds communities. *Proc 7th ACM/IEEE Jt. Conf. Digit. Lib.*, 53–62.
- Ling,X., Mei,Q., Zhai,C. and Schatz,B. (2008) Mining multi-faceted overviews of arbitrary topics in text collections. *Proc 17th ACM Conf. Knowl. Disc. Data Mining*, 497–505.
- Chung, Y., He, Q., Powell, K. and Schatz, B. (1999) Semantic indexing for a complete subject discipline. *Proc 4th Int. ACM Conf. Digit. Lib.*, 39–48.

- Harris, M., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C. et al. (2004) The Gene Ontology (GO) database. *Nucleic Acids Res.*, 32, D258–D261.
- 9. Buza, T., McCarthy, F., Wang, N., Bridges, S. and Burgess, S. (2008) Gene Ontology annotation quality analysis in model eukaryotes. *Nucleic Acids Res.*, **36**, e12.
- 10. Chagoyen, M. and Pazos, F. (2010) Quantifying the biological significance of gene ontology biological processes—implications for the analysis of systems-wide data. *Bioinformatics*, **26**, 378–384.
- He,X., Chee,B., Sen Sarma,M., Zhai,C. and Schatz,B. (2010) Identifying overrepresented concepts in gene lists from literature: a statistical approach based on Poisson mixture model. *BMC Bioinformatics*, 11, 272–284.
- Drysdale, R., Crosby, M. and FlyBase Consortium. (2005) FlyBase: genes and gene models. *Nucleic Acids Res.*, 33, D390–D395.
- Ling,X., Jiang,J., He,X., Mei,Q., Zhai,C. and Schatz,B. (2006) Automatically generating gene summaries from biological literature. *Pac. Symp. Biocomput.*, **11**, 40–51.
- Karamanis, N., Lewin, I., Seal, R., Drysdale, R. and Briscoe, E. (2007) Integrating natural language processing with FlyBase Curation. *Pac. Symp. Biocomput.*, **12**, 245–256.
- Ling,X., Jiang,J., Mei,Q., He,X., Zhai,C. and Schatz,B. (2007) A study of semi-structured summarization: generating gene summaries from biological literature. *Inform. Process. Manage.*, 43, 1777–1791.
- Sen Sarma, M., Whitfield, C. and Robinson, G. (2007) Species differences in brain gene expression profiles associated with adult behavioral maturation in honey bees. *BMC Genomics*, 8, 202.
- Jiang, J. and Zhai, C. (2007) An empirical study of tokenization strategies for biomedical information retrieval. *Inform. Retriev.*, 10, 341–363.
- Toth,A., Varala,K., Newman,T., Miguez,F., Hutchison,S., Willoughby,D., Simons,J., Egholm,M., Hunt,J., Hudson,M. *et al.* (2007) Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science*, **318**, 441–444.
- Christians, E., Davis, A., Thomas, S. and Benjamin, I. (2000) Embryonic development: maternal effect of Hsfl on reproductive success. *Nature*, 407, 693–694.
- Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10, 57–63.