

**Annual Report for Period:**09/2006 - 08/2007**Submitted on:** 06/29/2007**Principal Investigator:** Schatz, Bruce R.**Award ID:** 0425852**Organization:** U of Ill Urbana-Champaign**Title:**  
FIBR: BeeSpace - An Interactive Environment for Analyzing Nature and Nurture in Societal Roles**Project Participants****Senior Personnel****Name:** Schatz, Bruce**Worked for more than 160 Hours:** Yes**Contribution to Project:****Name:** Robinson, Gene**Worked for more than 160 Hours:** Yes**Contribution to Project:****Name:** Fahrbach, Susan**Worked for more than 160 Hours:** Yes**Contribution to Project:****Name:** Rodriguez-Zas, Sandra**Worked for more than 160 Hours:** Yes**Contribution to Project:****Name:** Zhai, ChengXiang**Worked for more than 160 Hours:** Yes**Contribution to Project:****Name:** Bruce, Bertram**Worked for more than 160 Hours:** Yes**Contribution to Project:**

Project Lead for Education and Outreach.

Senior graduate student in education supported on grant, plus summer salary for Biology Teacher at University Laboratory High School.

**Post-doc****Graduate Student****Undergraduate Student****Technician, Programmer****Name:** Littell, Todd**Worked for more than 160 Hours:** Yes**Contribution to Project:**

Chief Programmer developing the software system

**Name:** Buell, Jim**Worked for more than 160 Hours:** Yes

**Contribution to Project:**

project coordinator for user community and education/outreach

**Other Participant****Research Experience for Undergraduates****Organizational Partners****Texas A&M University Main Campus**

computational annotation of the honeybee genome

**CORNELL UNIVERSITY**

fulltext digitalization of the beekeeping literature

**Indiana University**

FlyBase curator site, working to evaluate our analysis software

**Harvard University**

FlyBase curator site, working to evaluate analysis software

**Other Collaborators or Contacts**

Chris Elsik, Department of Animal Sciences, Texas A&M University

Nan Hyland, Mann Biology Library, Cornell University

Kathy Mathews, Department of Biology, Indiana University

Bill Gelbart, Department of Biology, Harvard University

**Activities and Findings****Research and Education Activities: (See PDF version submitted by PI at the end of the report)**

See attached file with details of our substantial efforts into Research (Informatics and Biology) and Education (including Outreach).

**Findings:**

For Biology research, the nature-nurture dissection has been completely planned, the honey bees have been collected from each experimental situation, and the microarray expression pipeline has been established. The expression experiments are now proceeding. Analysis of findings will begin this coming year.

For Informatics research, the first fully fledged analysis environment has been developed. This supports concept navigation of community collections, with all collections and indexing being dynamically computed. Each individual component has been evaluated and published. The integrated system is beginning user testing with early adopters and will receive heavy usage this coming year.

**Training and Development:**

We have been focusing on providing experiences with real science research as appropriate for students at levels ranging from postgraduate to middle school. A small group (10-15) for each level.

Graduate Students. Work closely with postdocs and PhDs for judging results and helping their research.

Building early adopter community at external sites, notably the North Carolina honey bee consortium and the FlyBase curators.

Undergraduate. Developed freshman bioinformatics course based on BeeSpace at Wake Forest. Taught for first time this year, will teach again in coming years. Students produce educational materials for middle school students which we will use in summer workshops.

High School. Paying biology teacher at University Laboratory High School to integrate materials into Field Biology course. Website produced focusing on behavior and ecology.

Middle School. Hosting first summer workshop, using BeeSpace-developed materials. This year focus on evaluating materials with small group, next year will focus on training with large group.

#### **Outreach Activities:**

We support an extensive website of slides and videos, including hosting the campuswide Biomedical Informatics seminar. The educational materials developed are being packaged for external distribution to general audiences.

See <http://www.beespace.uiuc.edu>

We did extensive outreach to Campus Middle School for Girls, including investigator lectures on their research and field trips to bee research facility.

#### **Journal Publications**

Gene Robinson, "Beyond Nature and Nurture", *Science*, p. 397, vol. 304, (2004). Published,

Honey Bee Genome Sequencing Consortium, "The genome sequence of the honey bee, *Apis mellifera*, a highly social animal", *Nature*, p. , vol. 443, (2006). Published,

R. Velarde, G. Robinson, S. Fahrbach, "Nuclear receptors of the honey bee: Annotation and Expression in the Adult Brain", *Insect Molecular Biology*. Honey bee genome special issue., p. , vol. 15, (2006). Published,

A. Hummon, T. Richmond, P. Verleyen, G. Baggerman, J. Huybrechts, M. Ewing, E. Vierstraete, S. Rodriguez-Zas, L. Schoofs, G. Robinson, J. Sweedler, "From the Genome to the Proteome: Uncovering Peptides in the *Apis* Brain", *Science*, p. , vol. 314, (2006). Published,

S. Rodriguez-Zas, B. Southey, C. Whitfield, G. Robinson, "Characterization of unique gene expression trajectories across behavioral maturation in honey bees using a semiparametric model", *Genome Research*, p. , vol. , ( ). Submitted,

C. Whitfield, Y. Ben-Shahar, C. Brillat, I. Leoncini, D. Crauser, Y. LeConte, S. Rodriguez-Zas, G. Robinson, "Genomic dissection of behavioral maturation in the honey bee", *PNAS*, p. , vol. 103, (2006). Published,

X. Ling, J. Jiang, X. He, Q. Mei, C. Zhai, B. Schatz, "Automatically Generating Gene Summaries from Biomedical Literature", *Proceedings of Pacific Symposium on Biocomputing 2006 (PSB'06)*, p. 40-51, vol. , (2006). Published,

J. Jiang, C. Zhai, "Exploiting Domain Structure for Named Entity Recognition", *Proceedings of HLT/NAACL*, p. , vol. , (2006). Published,

Q. Mei, C. Zhai, "A Mixture Model for Contextual Text Mining", *Proceedings 2006 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (KDD'06)*, p. , vol. , (2006). Published,

Q. Mei, C. Liu, H. Su, C. Zhai, "A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs", *Proceedings of the World Wide Web Conference 2006 ( WWW'06)*, p. , vol. , (2006). Published,

Q. Mei, D. Xin, H. Cheng, J. Han, C. Zhai, "Generating Semantic Annotations for Frequent Patterns with Context Analysis", *Proceedings 2006 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*, p. , vol. , (2006). Published,

B. Chee, B. Schatz, "Document Clustering using Small Worlds Communities", *Proceedings 2007 Joint (ACM/IEEE) Conference on Digital Libraries*, p. , vol. , (2007). Published,

Xu Ling, Jing Jiang, Xin He, Qiaozhu Mei, Chengxiang Zhai, Bruce Schatz, "Generating Semi-Structured Gene Summaries from Biomedical Literature", Information Processing & Management, p. , vol. , ( ). Accepted,

Yue Lu, Xin He, Sheng Zhong, "Cross-species microarray analysis with the OSCAR system suggests an INSR->Pax6->NQO1 neuro-protective pathway in ageing and Alzheimer?s disease", Nucleic Acids Research, p. , vol. , ( ). Accepted,

S. Sinha, X. Ling, C. Whitfield, C. Zhai, G. Robinson, "Genome scan for cis-regulatory DNA motifs associated with social behavior in honey bees", PNAS, p. , vol. 103, (2006). Published,

C. Whitfield, Y. Ben-Shahar, C. Brillet, I. Leoncini, D. Crauser, Y. LeConte, S. Rodriguez-Zas, G. Robinson, "Thrice Out of Africa: ancient and recent expansions of the honey bee", Science, p. , vol. 314, (2006). Published,

T. Tao, X. Wang, Q. Mei, C. Zhai, "Language Model Information Retrieval with Document Expansion", Proceedings of HLT/NAACL, p. , vol. , (2006). Published,

G. Robinson, J. Evans, R. Maleszka, H. Robertson, D. Weaver, K. Worley, R. Gibbs, G. Weinstock, "Sweetness and Light: Illuminating the Honey Bee Genome", Insect Molecular Biology, p. , vol. 15, (2006). Published,

### **Books or Other One-time Publications**

#### **Web/Internet Site**

##### **URL(s):**

<http://www.beespace.uiuc.edu>

##### **Description:**

The project website.  
 Contains detailed information on the background behind the project.  
 Contains slides and videos of the lectures given during the project.  
 Contains prototype software system for biology users.  
 Contains scientific publications and educational materials.

#### **Other Specific Products**

##### **Product Type:**

##### **Audio or video products**

##### **Product Description:**

Slides and Videos of Project. Overviews and Training.

##### **Sharing Information:**

Freely available on project website.

##### **Product Type:**

##### **Software (or netware)**

##### **Product Description:**

BeeSpace analysis environment supporting concept navigation.

##### **Sharing Information:**

available from website <http://www.beespace.uiuc.edu>

#### **Contributions**

##### **Contributions within Discipline:**

coPI Gene Robinson was elected to the  
National Academy of Sciences  
in the section on Evolutionary Biology

Honey Bee Genome completed and published in Nature  
as part of multi-article multi-journal burst

Informatics Research is pioneering semantic summaries  
of biomedical literature for genomic biology

**Contributions to Other Disciplines:**

Collaboration with FlyBase is leading towards automatic support for biology curation. Important for future of biology as genome sequencing becomes routine and each community must provide their own annotations.

**Contributions to Human Resource Development:**

Established research facility in new Institute for Genomic Biology at University of Illinois, training new generation of integrative biologists across wet lab biology and dry lab informatics.

**Contributions to Resources for Research and Education:**

the various informatics investigators  
(Schatz, Zhai, Rodriguez-Zas) have helped  
establish a new Master's degree program in  
Bioinformatics at the University of Illinois

PI Schatz is the campus leader to develop new interdisciplinary PhD program in informatics at University of Illinois, with special focus on biomedical informatics.

**Contributions Beyond Science and Engineering:**

coPI Gene Robinson published OpEd piece in the New York Times about Nature-Nurture, a popular version of the project topic.

Honey Bee genome research widely reported in popular press, in tandem with scientific publication of the genome sequence.

BeeSpace-catalyzed genomics research being used to study agricultural crisis of honeybee disappearance (CCD Colony Collapse Disorder).

PI Schatz doing industrial outreach, e.g. invited big Tech Talk at Google on BeeSpace-catalyzed semantic searching in July 2007.

**Special Requirements**

**Special reporting requirements:** None

**Change in Objectives or Scope:** None

**Unobligated funds:** \$ 0.00

**Animal, Human Subjects, Biohazards:** None

**Categories for which nothing is reported:**

Any Book

## **BeeSpace Research and Education, 9/1/2006 – 8/31/2007**

This year was the third of five. We made progress on all aspects of Research, both Biology and Informatics, and of Education, both training and outreach. These are outlined below. See the BeeSpace website <http://www.beespace.uiuc.edu> for more information, including software and slides, among other digital materials.

In November 2006, we moved the project into a state-of-the-art facility just opened for the Institute for Genomic Biology at the University of Illinois at Urbana-Champaign. The IGB is the flagship new building for the campus, specifically designed for integrative biology research including laboratory experiments and informatics analysis. BeeSpace has been its flagship project throughout the first 3 years of its existence.

See <http://www.igb.uiuc.edu/about/research%20schematic.html> which shows how BeeSpace is at the center of the building and the core of the activities.

In May 2007, we held our Third Annual Workshop, the first in the new IGB building. Over 80 attendees listened to a morning of overview lectures, participated in an afternoon of interactive demonstrations, and discussed their feedback on our progress and plans during the final session. The participants included local biologists and computer scientists, plus a variety of collaborators. Our project paid the expenses for 12 external participants including curators from FlyBase and programmers from BeeBase.

See [http://beespace.uiuc.edu/pubs\\_talks.php](http://beespace.uiuc.edu/pubs_talks.php) for slides and videos of the talks.

### **Informatics**

The goal of the Informatics Research is to build BeeSpace, an interactive environment for functional analysis. More specifically, to interactively annotate functions for differential expression using concept-based navigation of biological literature and gene-centered summarization analysis. This year, we completed the first fully fledged analysis environment, with spaces explicitly as a paradigm. This new system version 3 is being tested by the initial power users from local laboratories at the University of Illinois, supplemented by friendly users in North Carolina and elsewhere. To run this version, see <http://beespace.igb.uiuc.edu:8080/BeeSpace3/> and also see the documentation.

Last year, we developed the first fully fledged system for concept navigation. We worked closely with our early adopter users, including a postdoc Moushumi Sen Sarma shared between the informatics and the biology research teams. This version v2 is still available on our website. A live demonstration of the new v3 was given at the workshop by the postdoc; the video of this can be viewed with workshop materials on our website.

The system components are being developed by graduate students in computer science and bioinformatics, who are publishing papers on the research aspects of each component. These students are supervised by coPI Zhai and PI Schatz. The system integration and architectural foundations are being developed by our senior research programmer Todd Littell, including optimization of the student developed algorithms.

BeeSpace version 3 now implements literature pipelines equivalent to traditional sequence pipelines. For example, a user can summarize a gene from a specified collection, where the system automatically extracts relevant sentences placed into useful categories. A user can also specify a gene list, to summarize the most related concepts within a specialized collection. This enables the analysis environment to functionally analyze the expression experiments from the biology research to dissect nature-nurture, by locating candidate concepts for candidate genes, a new-form of meta-analysis.

The gene summarization is being evaluated by professional curators at FlyBase, with good marks thus far. PI Schatz visited the main FlyBase sites at Indiana University and at Harvard this year, and curators from those sites attended the BeeSpace workshop to continue the collaboration. FlyBase can no longer generate gene summaries from the text of articles, due to curator loads (nor can any of the other GMODs Generic Model Organism Databases). So the general facility in BeeSpace to automatically generate gene summaries against a specified collection is potentially extremely valuable.

All analysis is done interactively relative to a “space”. A space is a collection of documents that contain concepts representing functions. A user can create new spaces very easily then interactively analyze the functions using these new spaces as the base background. Thus the literature pipeline has a similar analysis to the sequence pipeline, with differential expression of foreground against background. Spaces can be created by search (e.g. phrases against Medline or against Insects) and by navigation (e.g. switching key phrases from collection to collection, trying different backgrounds). Space algebra is supported, to intersect different spaces, or subtract one space from another. Any resulting collection can be saved and made into a new space. For example, a user can easily create a background space on foraging behaviors in winged insects to compare against a different background space of foraging behaviors in social insects.

We have developed special new algorithms for clustering collections to break them apart into semantically related subcollections. Each subcollection can be saved as a new space if desired. One of our clustering algorithms is bottom-up, using small worlds to find the natural coherent clusters. Another of our clustering algorithms is top-down, using mixture models to find clusters steered towards specified phrases.

All of the clustering algorithms have been implemented with parallel optimization, so that clustering of special collections (e.g. less than 100K documents) takes place interactively (e.g. less than 10 seconds) on our 4-processor 32-GB-RAM server. Next year we will buy a 16-processor 128-GB-RAM server and be able to handle larger base collections (all of Medline and Biosis and Agris) for larger user populations (our 100 test users and perhaps all of the users referred to us from an arrangement with FlyBase).

So we already have a dynamic system, where users can create new spaces on-the-fly, rather than requiring pre-computation of pre-specified spaces. This approaches our goal of supporting “functional analysis unconstrained by pre-existing categories”. In the coming year, we will be extending the meta-analysis capabilities into question answering.

## Biology

The goal of the Biology Research is to carefully dissect the relative contributions of nature and nurture for social behavior in the honey bee. In particular, we are experimentally measuring brain gene expression for important societal roles during normal behavior, with different experiments varying heredity (nature) and environment (nurture) for different societal roles.

The genome sequence of the honey bee was published in a special issue of *Nature* appearing on October 26, 2006. The special issue included cover art, news coverage, and specially commissioned News & Views pieces on this milestone. There was also an unprecedented burst of coordinated publication, some 50 papers in all, which appeared at the same time in *Science*, *PNAS*, *Genome Research* and *Insect Molecular Biology*. See [http://www.beespace.uiuc.edu/resources\\_genome.php](http://www.beespace.uiuc.edu/resources_genome.php). The event was also widely covered in the popular media. While this is preliminary work for this project, BeeSpace was explicitly mentioned in many of the publications, e.g. the *Nature* news article <http://www.nature.com/nature/journal/v443/n7114/full/443893a.html>.

This year, we have completed our experimental pipeline and the first biology experiments are emerging to be analyzed by the informatics. We tested the meta-analysis software for gene annotation, using previous experimental results with an older honey bee EST array. The bees for the designated nature-nurture dissections have all been collected and stored for microarray processing. We are focusing on Foraging and Defense, as symbolic of animal behavior for food and warfare. Hopefully, this will enable comparisons to higher organisms. The honey bee is a model for natural behavior; we are collecting bees during their normal behavior in the field.

The honey bee genome is now complete; it was generated as part of the second wave of NIH sequencing at Baylor. We have completed a genome annotation, using a computational pipeline based on ortholog comparisons to assign Gene Ontology categories to each generated sequence. This was done by our collaborator Chris Elsik from Texas A&M using BeeBase. The complete microarray based on the genome sequence is now fabricated, and we are using it in our full experiments being carried out in the laboratory of coPI Robinson. The statistical analysis of differential expression for discriminating genes is being done by BeeSpace students of coPI Rodriguez-Zas.

For Defense, we will dissect nature via different races of bees (European, African) and nurture via different levels of threat (manipulations of alarm pheromones). For Foraging, our primary focus will be on when a bee transitions her societal role from nurse to forager (age of onset of foraging). We will again dissect nature via different races of bees and dissect nurture via social manipulations (e.g. manipulating food supply in the hive to create precocious foragers or overage nurses) and physiological manipulations (e.g. manipulating NPF with vitellogenin or JH Juvenile Hormone with octopamine).

We are also doing anatomical localization of gene expression, using in situ hybridization of whole bee brains. This plan is being executed at Wake Forest University in the laboratory of coPI Fahrbach. Since the expression experiments have been delayed due to the delay of the genome sequence, we are instead using a list of genes encoding neuropeptides with key roles in regulation of behavior. We have separately identified 36 neuropeptide-encoding genes as the basis for our BeeSpace in situ hybridization studies.

We have developed a standardized method for probe design and a “bank” of information on construction of probes for bee tissue, using vitellogenin as a universal “positive control” for hybridization of honey bee tissues. A graduate student who worked on neuropeptide mapping in the Robinson lab at Illinois will be transferring to the Fahrbach lab at Wake Forest as a postdoc to coordinate these anatomical localization projects.

## **Education**

We are training students at many levels, concentrating on giving research experiences to small groups of focused persons. Our investigators and researchers are directly involved in education and outreach. Extensive photos and videos of our research activities are available on our website.

This year we held a separate workshop on Education and Outreach in June 2007 to plan the summer workshops and other future activities. Our overall strategy is to create a unique and useful collection of educational materials on biology and informatics, which are targeted at different audiences and which are freely available on our website.

At the Graduate level, we support 10 students in Biology and in Informatics. Their department affiliations include entomology and neuroscience, computer science and animal science. The interdisciplinary interactions are facilitated by our new Institute for Genomic Biology. We also work closely with our early adopter community of biologists at other universities and research institutions. Our project coordinator Jim Buell is supporting the user community and also the educational activities; he is concurrently working on his dissertation in Educational Psychology under coPI Chip Bruce.

At the Undergraduate Level, we have developed a new Bioinformatics for Beginners course based on BeeSpace. This is taught as a freshman honors class at Wake Forest University by coPI Fahrbach. The first such course was taught in Fall 2006, with great success, and will be repeated over the next several years. The undergraduate students produce educational materials on nature-nurture in honey bees, which are intended for middle school students. For example, one good project this year was BeeLand, a board game, where the players move through the societal roles of honey bees with realistic science. The produced materials are planned for use in our summer workshops.

Each summer, we are paying the biology teacher at University Laboratory High School to develop Beespace-related materials for use in his Field Biology course. The materials developed span a wide range of biological topics, and can be found on our website.

We will be field testing our summer workshop summer 2007, with a limited number of University High School students, at the entering freshman level. The week-long half-day workshop will include research visits to entomology labs of our two National Academy members (Robinson, Berenbaum), evaluation of the Bioinformatics for Beginners educational materials, and use of the BeeSpace software to investigate the popular topic of CCD Colony Collapse Disorder. We are employing a recent PhD in entomology to teach the workshop, who previously taught a month of insect biology at Campus Middle School, as part of an Illinois NSF GK-12 training grant. These evaluations should debug the workshop for a fully fledged version next summer with 20 middle school students.

For outreach, we used our arrangement with the Campus Middle School for Girls, a private school located on the University of Illinois campus. One of our female biology students (actually the same one who was judging the informatics results) described her neuroscience dissertation research to middle school girls in a hands-on interaction. Previously, our project coordinator had lectured on bee social behavior with slides and videos, with a field trip to the Bee Research Facility with lecture by coPI Robinson.

### **Special Requirements**

Our funds from NSF were frontloaded, in that more money actually arrived in the first two years than we requested. Because of this, we have underspent the funds and wish to push them forward into later years. The PI (Schatz) has significant experience in running large NSF systems projects and this pattern is quite typical. The planning years underspend, but the development years overspend, as the project ramps up with real systems and real users. So we appreciate the flexibility of NSF in permitting us to responsibly spend the funds when the needs of the project dictate, including the possibility of no-cost extensions after the formal end of the project in August 2009.