

BeeSpace – An Interactive Environment for Analyzing Nature and Nurture in Societal Roles

Bruce Schatz (PI), Gene Robinson, ChengXiang Zhai, Sandra Rodriguez-Zas, Bertram Bruce,
University of Illinois at Urbana-Champaign, Susan Fahrback, Wake Forest University

One of the most important questions in biology is the origin of behaviors: nature or nurture? Biologists of course now recognize this to be a false dichotomy, but the study of behavior can be liberated from “nature-nurture” only by viewing it as the result of inherited and environmental influences acting on the same substrate, the genome. The environment (“nurture”), which includes other individuals, impacts an inherited genome (“nature”) by controlling gene expression during the life of the organism.

We propose to use the honey bee to elucidate the relationship between genes and life in an animal society on an unprecedented, whole genome, scale. Honey bees live in complex societies organized on the basis of division of labor among workers; instead of performing one role for the duration of her life a worker assumes many roles, influenced by heredity and the environment. The bee genome is just being sequenced, normal bee behavior in the field is accessible to molecular analysis, and the bee literature is uniquely comprehensive in discussing the influences of heredity and environment, reflecting a thousand years of natural history due to beekeeping.

The interactive environment we propose to develop will support the integration of molecular description with information from ecology, evolution, behavioral science, and physiology. We will perform a functional analysis of all the roles of a worker honey bee within her society by combining microarray analysis, large-scale brain *in situ* hybridization, and a novel approach to informatics that links all sources from current genome databases to the existing scientific and natural history literatures on honey bees. Using state-of-the-art methods of genomics, molecular biology, and statistics, we will create a database of brain gene expression profiles for all major social roles and localize the expression of many bee genes to precise regions of the brain. Using state-of-the-art informatics techniques, we will develop algorithms that will enable users to navigate diverse databases and sources of literature for hypothesis development and testing.

This approach will enable all who study a bee gene to place this information in an integrative biological context. Our research will take us to the frontier of contemporary biology, where *biotechnology enables sequencing and expression analysis -- and bioinformatics enables functional analysis that is unconstrained by pre-existing categories.*

Achieving these goals requires development of a novel comprehensive software environment for our model organism. We term this software environment **BeeSpace**. The system will enable users to navigate interactively across multiple sources for functional analyses. We will test the prototype in 15 laboratories studying honey bees and related organisms. We will also test BeeSpace by providing research experiences at the graduate, undergraduate, high school, middle school levels with appropriate training in each case and minority outreach at the lower levels.

In summary, we will use genomics and informatics to develop a modern reformulation of the roots of behavior that goes beyond the nature-nurture dichotomy. Addressing these issues with the honey bee, a “lovable” and economically important animal whose behavior has already captured the fancy of the public and scientists alike, ensures this project will have high impact.

The scientific merit is demonstrating an effective analysis environment for understanding the behavior of individuals in a complex society. The broader impact uses the general attractiveness of the honey bee to encourage student participation in biological research.

Understanding Nature and Nurture in Societal Roles

One of the most important questions in biology is the origin of particular behaviors: nature or nurture? Each discovery of a gene influencing behavior elicits two reactions: excitement for the benefits the discovery portends, and anxiety over the implications of attributing control to DNA. Many fear that “biological determinism”, how genes play a dominant role in the regulation of behavior, will diminish appreciation for the role of environment in shaping our actions.

The “nature-nurture” controversy has not gone away in the hearts and minds of the public, even though biologists clearly understand that there are both genetic and environmental determinants of behavioral variation as well as complex genotype by environment interactions. This is certainly more accurate than nature versus nurture, but even this reformulation retains the flavor of a dichotomy that impacts how research is conducted and interpreted.

The study of behavior can be liberated from “nature-nurture” by viewing it as the result of inherited and environmental influences acting on the same substrate, the genome. The environment (“nurture”), which includes other individuals, impacts an inherited genome (“nature”) by orchestrating gene expression during the lifetime of the animal.

For behavior, gene expression in the brain is the “first phenotype,” the initial read-out of the interaction between hereditary and environmental information. Only now is this “first phenotype” accessible to study, due to the advent of genomics.

This proposal harnesses genomics to generate information on the relationship between behavior and gene expression in a single species, the honey bee, on an unprecedented scale. We focus on patterns of gene expression linked to social behavior, because social behavior is by definition uniquely sensitive to environmental influence. A genome-based approach means these studies can be conducted free of *a priori* assumptions of relative importance of specific genes.

An integrative approach to behavioral genomics requires that these studies connect to the extensive literature on organismal biology. Anyone who studies a bee gene expression profile should be able to place this information within evolutionary, ecological, behavioral, neural, physiological, cellular, and molecular contexts.

BeeSpace, the interactive environment we propose to develop, will make this possible. For Biology Research, we will develop brain gene expression profiles for all major roles performed during a worker honey bee’s life in her society, and localize patterns of expression within the brain for a careful selection of genes. For Informatics Research, we will use all sources from current genome databases and the existing scientific literature to integrate molecular description with information from physiology, behavior, neuroscience, and evolution. The corresponding software will go far beyond a searchable database by incorporating statistical analyses of the literature for the purpose of generating hypotheses, enabling functional explanations of relationships between genes and behavior.

The point is this: *when biotechnology enables routine sequencing and expression analysis, bioinformatics must enable routine functional analysis unconstrained by pre-existing categories.*

The first wave of genome sequencing of multicellular organisms focused on organisms selected for genetic tractability in the laboratory, such as the roundworm *C. elegans* and the fruit fly *Drosophila*. The second wave of genome sequencing by the National Human Genome Research Institute (NHGRI) began in May 2002, with the selection of the honey bee and three other animals (chimpanzee, chicken, sea urchin). Our biology lead coPI Robinson was the lead author of the proposal for the honey bee genome and coordinates the project together with sequencers at the Human Genome Sequencing Center at Baylor College of Medicine.

The honey bee was chosen by NHGRI as a model for social behavior, in part because the ancient craft of beekeeping effectively bridges the gap between studies of animal behavior in artificial and natural settings. The unique relationship between humans and honey bees permits scientists to study normal behavior and associated patterns of gene expression as they unfold in the species-typical environment. NSF had already recognized the value of the honey bee, having listed it as a model species some years earlier.

Our project is possible now because of advances in sequencing and computational technology. The descriptions of gene expression will be based on standard techniques, but the scale of the project and grounding of it in environmentally-determined behaviors is unprecedented. Our goal is to complete today a cutting edge project driven by a fundamental question in biology that will show what is possible for everyday biology tomorrow. BeeSpace will be the model for BioSpace, universal infrastructure to navigate all biological knowledge [Schatz,2002a].

BeeSpace will contain a full set of quantitative brain gene expression profiles, linked to our brain maps of gene expression, linked to FlyBase, WormBase and all other databases containing information on genetics, genomics, and molecular biology, linked in turn to the entire pre-existing scientific literature on the biology of the honey bee and related organisms.

This will permit identification of clusters of genes whose expression predicts specific behaviors, genes expressed independent of environmental influences, and behaviors that share patterns of gene expression. Implementation of novel text analysis algorithms will, for the first time, make the description of specific behaviors and the contexts in which they occur as easily retrievable (and as subject to statistical analysis) as DNA sequences.

Modern reformulation of nature-nurture questions concerning behavior requires knowing which genes respond to environmental factors, which genes vary as a result of heredity, and which genes are sensitive to both environment and heredity.

Addressing these issues with the honey bee, a “lovable” and economically important animal whose behavior has already captured the fancy of the public and scientists alike, ensures that this project will have a high impact on our understanding of the roots of behavior. BeeSpace will create new research tools and provide the first post-genome project opportunity for biologists to ask and answer fundamental questions about nature and nurture.

The Model Organism: the Western Honey bee, *Apis mellifera*

Our choice as a model organism for functional analysis of normal behavior at the molecular level is the Western honey bee, *Apis mellifera*. Honey bees live in societies that rival our own in complexity, internal cohesion, and success in dealing with the myriad challenges posed by social life, including those related to communication, aging, social dysfunction and infectious disease.

The honey bee society features striking behavioral variation among individuals, due to systems of division of labor. Bees are individually complex enough so that they can take on a series of different roles at different times in their lives. The roles they perform, and when they perform them, depend both on hereditary predispositions and the needs of their colony [Robinson 2002b]. Moreover, recent microarray analyses [Whitfield et al. 2003] have revealed striking differences in brain gene expression profiles associated with components of two roles: caring for the young (brood) and food acquisition. Honey bees thus represent an excellent model to study how hereditary and environmental influences on the genome act to orchestrate behavioral variation among the individuals in a society.

The life cycle of an adult worker honey bee is highlighted by tending the brood within the hive as a “nurse bee” when young, performing a series of jobs related to hive maintenance and

food processing in middle age, and collecting nectar and pollen outside the hive when old. The change from hive work to foraging is a significant lifestyle change. This change typically occurs around two to three weeks of age, depending on both genotypic and environmental factors.

The salient features of bee biology for this project are as follows. (1) Because the bee is an insect, its behavior is highly stereotyped and rigorously assayable; complete behavioral maturation occurs within a relatively short lifespan (4-6 weeks). (2) Owing to a long and rich association with humans, comprehensive knowledge of bee behavior [Winston 1987] provides a firm foundation upon which to build analyses that integrate molecular biology, neuroscience, ecology, sociobiology, and evolutionary biology. (3) Methods of raising and manipulating bees are well established, due to the ancient close association between bees and humans for honey production. These techniques enable control of both genetic and environmental parameters. Genetic control is a consequence of the structure of the society, where all individuals are the progeny of a single queen; due to haplodiploidy, instrumental insemination of a queen with semen from a single drone results in offspring that are closely related. Environmental control is due to the physical structure within the society, where all individuals base their behaviors within the hive. Thus genetics can be controlled by varying the colony (features of the queen) while environment can be controlled by varying the base (features of the hive). (4) Bees live in large colonies that are maintained economically, making it easy to obtain robust sample sizes. (5) Bees live in tightly structured societies in which an individual's physiological and behavioral status is dependent upon communication with other society members. (6) Bees display a pattern of behavioral maturation that is "vertebrate-like" in richness and complexity, proceeding from hive tasks such as nursing to the cognitively demanding task of foraging. This behavioral development is controlled by neural and endocrine mechanisms similar to vertebrates.

The effectiveness of the honey bee as a model is that it is just the right size to perform a complete functional analysis at the present level of technology, both biological and informational. The genome will soon be sequenced, so the expression products of virtually all of the genes in the brain can be captured. The efficiency of the honey bee as a model is that both genotypic and environmental variation can be controlled, under natural conditions. That is, nature itself can become the laboratory, varying conditions to observe behaviors while maintaining only normal conditions that commonly occur in the field. Capturing both genetic and environmental variation enables integrative analysis of the roles of a bee within its society.

Biology Research: Brain Gene Expression Profiles for All Societal Roles

Our biology experiment will generate a molecular signature of all the major roles performed by honey bees in their society. To accomplish this, we will generate brain gene expression profiles for individuals captured in the very act of performing a normal activity. The small size of the bee enables efficient flash freezing in the field, even bees captured in flight. Since current microarrays measure the abundance of mRNA, physically generated within minutes, behaviors representing social roles that last for days will be emphasized. We will also sample some more short-term behaviors that are part of longer lasting social roles.

While the experimental model is an insect, we will use broad categories of social roles that are potentially applicable to higher organisms, including humans. Previous efforts at sociobiology tried to use detailed observations of insects to make predictions about humans [Wilson 1971,1975]; one difficulty with this approach is that it is difficult to infer the evolutionary relationships between behaviors across species based on observation alone.

The current efforts at sociogenomics [Robinson 2002a] have the advantage of pitching the analysis at the molecular level, where the insights are more objective. The power of this approach can be seen in developmental biology, where molecular analysis has shown penetrating insights into both mechanistic and evolutionary aspects of development. Consider the discovery of genes and pathways common to invertebrates and vertebrates involved in development of structures previously thought to have no homology, e.g. Pax-6 and the eye [Callaerts et al. 1997].

The fundamental features of social behavior are unusually well classified for our model, by E. O. Wilson and others [Wilson 1976; Oster & Wilson 1978; Seeley 1982,1985; Robinson 1987]. Observations of honey bees and several other species of social insects have revealed a total repertoire of about thirty behaviors that relate to the basic roles of: caring for the home (nest), generating and caring for the offspring, defending the nest, and acquiring food.

Table 1 gives our master list of the 22 behaviors that we shall study as societal roles. The categories for these roles are intended to be generic, although the roles are specialized to social insects. All roles have been studied by co-PI Robinson, who has 30 years of bee experience.

Home			
Comb build	Remove corpses	Hygienic behavior (remove diseased brood)	
Offspring			
Brood care	Attend queen	Personal reproduction (worker)	
Defense			
Guard	Soldier		
Food			
Forage for nectar	Forage for pollen	Forage for water	Forage for resin
Dance communication: sender	Dance communication: receiver	Process food (nectar to honey)	Scout
Experimentally induced roles to study molecular basis of development of social behavior			
Precocious forager	Normal age nurse	Normal age forager	
Overage nurse	Reverted nurse	Socially isolated	

Table 1. Principal Societal Roles in the Honey Bee Colony

Honey bees are complex social animals. Bee behavior is highly flexible; bees live in an “urban environment,” in which the stimuli that elicit the performance of all behaviors are present throughout their lives. Complex social behavior occurs in response to environmental conditions, such as changing the floral sources exploited by a colony by means of the dance language or repelling intruders by means of kin recognition mechanisms [Winston 1987; Seeley 1995].

Bee behavior is ideally suited to “go beyond nature and nurture,” to analyze hereditary and environmental influences acting on the genome. Both genetic and environmental influences on bee behavior have been identified, e.g. in the case of precocious foraging. When a colony experiences a shortage of foragers or food stores, some of the individuals in the colony respond by accelerating their pattern of maturation and becoming foragers. Precocious foragers may initiate foraging as much as two weeks earlier than usual, an impressive acceleration given their 5-7 week life span. The control of rate of honey bee behavioral maturation is influenced by such factors as inhibitory social interactions with older workers and pheromones produced by the brood and queen. These interact with inherited factors; bees of some genotypes are more likely to become precocious foragers than other bees [Robinson et al. 1989]. Precocious foraging is thus determined in part by colony need and in part by individual predisposition.

Our first microarray study [Whitfield et al. 2003] has shown that precocious foraging is associated with extensive changes in brain gene expression. A microarray study that Robinson's lab is performing with Co-PI Rodriguez has demonstrated that some genes regulated in association with precocious foraging also show genotypic variation. We will build on these results to generate brain gene expression profiles for other behaviors: we will emphasize behaviors responsive to changing conditions and we will also sample for genotypic variation.

These two studies have been performed with the first generation honey bee microarray, a cDNA-based platform which contains approximately half of the genes in the genome. The BeeSpace database will use the next generation microarray, which will be based on the entire annotated genome and fabricated with oligonucleotides, which provides greater efficiency and consistency. Co-PI Robinson has been awarded a grant from USDA to build the new honey bee "whole genome" microarray. Our project will thus be able to use a unique genomic resource.

We will provide an unprecedented foundation for the study of the molecular basis of social behavior by generating a complete set of "first phenotypes" all major social roles in the beehive. We will generate about 660 brain gene expression profiles, for 660 individual bees. This is computed by 22 behaviors * 3 colonies (genotypes/behavior) * 10 bees/colony. Under the guidance of Co-PI Rodriguez we will use state-of-the-art statistical analysis to account for both biological and technical variation [Jin et al. 2001; Whitfield et al. 2003], using both Bayesian and Least Squares analyses.

We will use as many as 2700 microarrays (660 * 4 = 2640 microarrays including reverse labeling, plus pilot studies) because the various profiles will be generated as a result of different sets of experiments that each involves their own setups and controls. This will constitute the largest gene expression database for normal behavior for any animal, but is quite feasible given our past experience and the genomics infrastructure at Illinois.

Bees will be collected performing these behaviors from beehives in the field. As is routine in the Robinson lab [Ben-Shahar et al. 2002; Whitfield et al. 2003] they will be flash-frozen to preserve natural gene expression states. Because division of labor in honey bee colonies is influenced (but not determined) by worker age, all bees will be of known age (done by marking 1-day-old bees with colored numbered tags or paint dots, so we can control for the influence of age on brain gene expression profiles).

In many cases, we will perform precise manipulations on bee colonies in order to study the environmental components that affect brain gene expression. As in previous experiments, we will create colonies that induce precocious foraging, overage nursing, and reverted nursing (foragers "going back in time" to take up brood care). We also will create food shortages to study precocious foraging, manipulate food discovery (to study "scouts," bees that discover a food source first and communicate their findings by means of the dance language, as well as other roles in this communication process), threats to colonies (to study guards and soldiers), threats to colony hygiene (corpse removers and hygienic bees), and threats to colony integrity (temperature manipulations to study ventilators).

Expression profiles will be generated from whole brains to provide a broad and extensive survey of behaviorally related gene expression. Brain expression profiles differ strikingly between nurses and foragers [Whitfield et al. 2003], suggesting that some gene regulation occurs on a global, brain-wide level. Genes showing interesting patterns of expression in relation to some of the social roles to be studied will then be studied in more depth via brain localization.

Feasibility: Statistical Differentiation of Brain Gene Expression Profiles

We will produce the largest gene expression database for normal behavior for any animal, providing a powerful new resource to study the molecular basis of social behavior. A database that is generated from up to 2700 microarrays is eminently feasible given our past experience and the genomics infrastructure at Illinois. CoPI Robinson's lab has developed the first cDNA microarray for honey bees [Whitfield et al. 2002] and recently reported that brain gene expression profiles can be used to distinguish between social roles for individual bees [Whitfield et al. 2003]. The later study involved 72 microarrays. In addition Robinson and Co-PI Rodriguez have just finished analyses of a study involving 124 microarrays. The Keck Center for Comparative and Functional Genomics on our campus has considerable expertise with large-scale NSF-funded microarray projects involving soybeans and *Arabidopsis* [see letter].

We can use our previous analyses to estimate the statistical power of our proposed experiment. Our simulations indicate that the proposed experiment can detect even subtle differences in expression profile between similar behaviors. For each behavior, we will have data from 3 hives, 10 bees per hive, 2 arrays per bee and 2 spots per array, for a total of 120 observations potentially available in a ratio of intensity analysis and 240 observations in an absolute intensity analysis. For 120 observations, and adjusting the degrees of freedom for the estimation of fixed effects (e.g. hives), approximately 100 DF per behavior are available. Assuming that, in the worse scenarios, half of the ratio observations must be discarded, 50 DF would be available. Both scenarios, 100 and 50 degree of freedom, were evaluated for power. Standard errors were assumed to be equal (0.2 units) or larger (0.4 or 0.6 units) than those observed in previous bee expression studies. Three experiment-wise false positive rates, allowing for different multiple testing adjustments, from less to most stringent ($\alpha = 1E-4, 1E-7, 1E-10$), and different magnitudes of expression differences or standardized differences were evaluated.

The results indicate that we will be able to detect even relatively small differences in patterns of mRNA abundance between behaviors. Our computation shows that at the most stringent criteria ($\alpha = 1E-10$) and for a conservative number of observations (100 degrees of freedom), even a one unit log change between behaviors can be detected with a power of 99%. This is equivalent to a two-fold change in expression units.

We can even detect a 0.5 unit fold change with a power of 89% under a more reasonable significance criteria ($\alpha = 1E-7$) for the proposed design and statistical methodology. These power computations are empirically confirmed by previous results from the comparison of brain gene expression profiles between nurses and foragers conducted in co-PI Robinson's lab [Whitfield et al. 2003]. Using the same bee population and microarray techniques we will use here, we were able to detect more than 30 sequences differentially expressed ($P < 1E-7$) with < 2 -fold difference. coPI Rodriguez-Zas has extensive experience with such statistical techniques in expression experiments for many organisms [Rodriguez et al. 2002,2003; Clough et al. 2004].

For the proposed experiments, we will evaluate models that accommodate potential sources of variation. The statistical analysis of gene expression data will be carefully normalized [Cui and Churchill 2003; Whitfield et al. 2003]. These models will permit us to attain the following goals: 1) detection of genes that exhibit differential expression between behaviors, 2) identification of behaviors with related gene expression patterns, 3) identification of genes with correlated expression patterns across behaviors, 4) construction of mathematical functions that can help predict behaviors based on the most informative gene expression patterns.

To accomplish these goals, we will use an experimental design that involves comparisons to a reference sample. The reference sample will be made by combining mRNA from all behavioral groups [Churchill, 2002]. Although we have used pair-wise comparison in a previous project [Whitfield et al. 2003], a reference design (“loop design”) is favored for the present study because of its huge scope. This will facilitate the staging of the studies, i.e., to conduct a set of separate experiments to capture the ideal set of social roles. This also will generate additional statistical power enabling us to identify differences in brain gene expressions for roles less differentiated than foragers versus nurses. For example, preliminary results in Robinson’s lab reveal statistical significance between guards and corpse removers.

Localizing Brain Gene Expression in the Brain: *In situ* hybridization studies

Gene expression profiles will be based on RNA extracted from whole brains of individual bees performing specific behaviors. An additional level of data refinement is needed to understand behaviorally related patterns of gene expression in the brain. We will use *in situ* hybridization to identify specific cell populations within brain regions, which express behaviorally relevant genes identified by microarray analysis. This will ground the project in contemporary neuroscience. coPI Fahrbach has extensive experience with gene localization in insect brains [Withers et al. 1993; Farris et al. 2001]; her project will enable biologists who use the brain expression profile database to:

(1) confirm that the gene expression being studied is neuronal or glial; (2) assess the consistency of patterns of gene expression across brains of bees performing the same behavior; (3) determine whether genes identified as relevant for more than one behavior are expressed in the same or different cell populations in the brain or conversely, (4) determine whether sets of genes identified as relevant for a single behavior are all expressed in the same cell populations; and (5) identify large scale region-specific changes in gene expression.

Basing this gene localization study on a comprehensive list of brain expression profiles will allow us to go beyond all other brain *in situ* studies. We will be the first to use gene expression to link specific pathways within the brain with specific behaviors across an animal’s entire behavioral repertoire. Our goal will be to report the distribution of each “behavioral transcript” in the context of the extensive pre-existing literature on the neuroanatomy of the insect brain in general and the honey bee brain in particular [Fahrbach, 2003].

Genes identified as having a strong association with a particular behavior will be screened using *in-situ* hybridization. Prior studies of microarrays based on the UIUC Bee Brain EST Project predict that several hundred genes will fall into this category. We propose the resources to screen 300 genes. Data will be recorded using a checklist of major bee brain regions. The checklist method is preferred for effective condensation of a large amount of neuroanatomical data and avoiding the need for labor- and resource-intensive collection, storage, and publication of electronic images. The goal is not to create an atlas or to duplicate existing resources for the study of the insect brain, but to form dynamic links between neural circuits and behavior.

Gene expression patterns resulting from these studies will link the microarray data to a century of descriptive neuroanatomy on the honey bee brain [Fahrbach, 2003]; importantly, it will also, through recognition of functional correspondences between vertebrate brain regions and insect brain centers, provide an entry to this new database for researchers focused on vertebrates. For example, functional and anatomical comparisons suggest that the insect mushroom bodies are similar to the vertebrate hippocampus [Strausfeld et al. 1998; Capaldi et al. 1999]. Our project will enable this provocative idea to be explored at the molecular level.

The bee is again just the right complexity to perform both whole brain expression and brain region hybridization experiments. One study of a mouse brain estimated that 75 million neurons were present [Williams, 2000]; current estimates of the number of genes encoded by the mouse genome range from 30-35,000. The corresponding figures for the honey bee brain are 750,000 neurons (100-fold less), and 13,000 genes (approximately one-third based on the estimated number of genes in the fruit fly genome). Brain mapping in mouse is requiring huge on-going technology innovation, but we can map expressions in the bee now for the scale of a FIBR grant.

Informatics Research: Interactive Environments for Functional Analysis

PI Schatz was the PI of the flagship project in the NSF National Collaboratory Program that built the Worm Community System (WCS), co-sponsored by NSF BIO. This project helped elevate the interactive analysis of *C. elegans* [Shoman et al. 1995]. The goal of WCS was to capture all the knowledge of the worm community of molecular biologists, within an information system that could be used interactively across the Internet as an analysis environment [Schatz 1997]. <http://www.canis.uiuc.edu/projects/wcs>.

WCS developed a comprehensive collection for a community in molecular biology, with hand-built mapping across sources, running in 50 labs worldwide in 1993. The worm community collection encoded the full range of knowledge, formal and informal, literature and data, within a tightly linked information space. The literature was from MEDLINE and BIOSIS, the data from genome databases and worm sources. Searches could be done across all the sources, including journal articles and community newsletters, and navigations could be done across all the links, including literature and newsletter articles, gene and clone descriptions, physical and anatomical maps.

Ten years ago, the Worm Community System first demonstrated the feasibility of interactive discovery in an analysis environment [lead news in Science 1993]. At that time, databases were sparse and the links were forged manually. Today, the databases are far more complete and the links can be forged largely automatically. It was clear from the WCS experience that automatic mapping across subject domains would require semantic indexing, where the items with “similar” meaning can be mapped together with only limited domain knowledge.

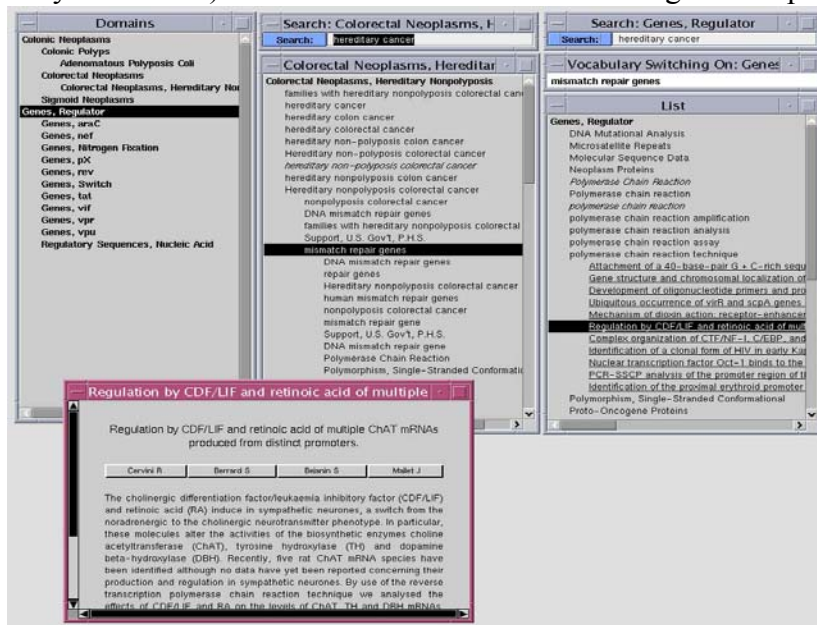
PI Schatz subsequently served as PI of the flagship project in the NSF Digital Libraries Program [Schatz et al. 1999], which built the Interspace Prototype. The “Interspace” is a future generation of the Internet, where the infrastructure directly supports information retrieval rather than data transmission [Schatz 2002b]. The Interspace protocols connect across “spaces” of information, just as the Internet protocols connect across networks of data.

Five years ago, the Interspace Prototype first demonstrated the feasibility of semantic indexing across the entire biomedical literature [lead news in Science 1998]. A large-scale experiment using supercomputers was performed, which parsed all articles in MEDLINE for conceptual phrases and computed relationships between these phrases within community collections [Bennett et al. 1999]. Each community collection became a concept space, enabling users to conceptually navigate across subject boundaries, without need for explicit searches.

Figure 1 shows an example session from the 1998 MEDLINE experiment where 1000 community collections were generated by partitioning the 10M abstracts in MEDLINE [Chung et al. 1999; Schatz 2002b]. Subject domains “Colorectal Neoplasms, Hereditary Nonpolyposis” and “Genes, Regulator” were chosen and their concept spaces displayed in the middle and the right respectively. “Hereditary cancer” was entered as a search term in the first concept space

and lexical variants returned. Navigating in the concept space moved from “hereditary nonpolyposis colorectal cancer” to the related “mismatch repair genes”.

The user then tried to search for this desired term in another community collection, “Genes, Regulator”. A straight text search at top right returned no hits. So Vocabulary Switching was invoked to concept switch from one community to another across their respective concept spaces. The concept switch took the term “mismatch repair genes” and all related terms from its indented co-occurrence list in the source concept space for “Colorectal Neoplasms” (including many not shown) and intersected this set into the target concept space for “Genes, Regulator”.



At middle right, Vocabulary Switching computed a List of concepts suggested as semantically equivalent to “mismatch repair genes” within “Genes, Regulator”. Concept space navigation then located the article displayed at the bottom of the screen. This article discusses a leukaemia inhibitory factor, related to colon cancer in animal models. Note this article was located by concept switching across community collections starting with the term “hereditary cancer”.

Figure 1. Sample session from Interspace Prototype showing interactive navigation across concept spaces via concept switching [Chung et al. 1999; Schatz 2002b].

BeeSpace: Developing Community Collections

To build the BeeSpace, we will first gather “all” relevant information about honey bees, encompassing data and information about nature and nurture. We then can utilize our unique semantic indexing to support interactive navigation across concept spaces. The indexing technology is general purpose and works equally effectively on any collection of suitable scale. A “community collection” is a special literature with constrained subject and terminology. A typical scientific community might be 5000-50,000 articles and 50-500 members. A community might partition the scientific literature by organism, such as the worm community or the bee community, or by methodology, such as behavioral development or neuro anatomy.

The sources to interconnect to form BeeSpace can be found in scientific literature, trade literature, gene descriptions, and database archives. First, we will include genome information from standard archives, such as sequences (e.g. NCBI’s GenBank) and expressions (e.g. EMBL’s ArrayExpress). Next, we will include genomic classification schemes, such as Gene Ontology [Ashburner et al. 2001] and KEGG, to cluster related genes into semantic classes.

The next inclusion is genetics for model genetic systems. For example, the gene description database within FlyBase provides a rich source of functional information for honey bees. Our

BeeSpace project will be closely connected to FlyBase. [see letter] The FlyBase PI, William Gelbart at Harvard University, plans to develop InsectBase to provide a platform for comparative insect genomics that encompasses bees and other insects slated for genome sequencing. We shall also utilize the similar model system databases from other organisms, most notably the worm (ACEDB) and the mouse (MGI), which have similar databases describing gene functions. Our test user labs include some specializing in each of these model systems (fly, worm, mouse).

The most extensive functional descriptions are contained within the scientific literature. Bibliographic databases contain a comprehensive selection of the scientific literature in downloadable format. We will use MEDLINE, which covers the medical literature with over 14 million citations of journal articles and BIOSIS, which covers the biological literature with over 17 million. Similarly, AGRICOLA, AGRIS, CAB Abstracts cover the agricultural literature, with over 11 million abstracts of journal articles among them. These databases are available in public archives sponsored by the US government or in our university archives via site licenses.

Honey bee behavior and its environmental regulation are recorded in the extensive natural history literature. We will obtain complete electronic versions of a wide sampling of the standard reference books. For example, through an arrangement with Harvard University Press [see letter], we will obtain such books as Winston's *The Biology of the Honey Bee* and Seeley's *The Wisdom of the Hive*. We intend to seek other arrangements with important entomology academic presses, such as Cornell University Press. The books will be divided into small conceptual units comparable to bibliographic citations, such as paragraphs or sections.

The key to effective semantic indexing is to appropriately partition the biological literature into community collections of related semantic topics. Previously, we used a general classification scheme for the bibliographic database MEDLINE, the Medical Subject Headings (MeSH), which partitioned the medical literature into 1000 community collections (of at least 10,000 articles each). For BeeSpace, we will use the Gene Ontology (GO) (www.geneontology.org), which provides comprehensive support for gene classifications across model systems. At over 17 thousand classes, as of January 2004, it is roughly the same scale as MeSH. Thus GO can partition the scientific literature into thousands of community collections.

In addition to general GO categories, we will use specific classifications relevant to BeeSpace. Each such classification will generate a community collection within BeeSpace, enabling biologist users to conceptually navigate across that collection. Keying off our biology research, specific classifications will include the master list of societal roles and the master list of brain regions. These classifications can be used to partition the bee literature, of course, but can also be used to partition for other organisms. For example, we will develop community collections for roles and regions for the scientific literature of other insects, such as social bees and solitary bees, wasps and ants. Other collections will be devoted to roles and regions for the model genetic systems, such as worm and fly, mouse and vole. Finally, individual biologists can create their own partitions for particular research. For example, BeeSpace analysis of precocious foraging in bees might involve collections on "behavioral maturation" in rodents.

Functional Analysis using Conceptual Navigations in BeeSpace

To build the BeeSpace environment, we will develop a comprehensive collection of textual sources related to genes, brains, and behavior. Then we will develop a comprehensive system for mining these textual sources for functional phrases and integrating them with genome databases. Finally, we will deploy this information system to an international community of relevant biologists, who will use BeeSpace for conceptual navigation to support functional analysis.

The initial users will be the bee laboratories of coPIs Robinson at University of Illinois and Fahrbach at Wake Forest University. The eventual community during the project will encompass 15 laboratories, including molecular bee labs, both domestic and foreign, and model genetic system labs studying societal roles. The List of User Labs is in the separate section on Supplementary Documentation, along with letters of support, under Sharing the Outcomes. These laboratories have committed to working with BeeSpace during the development period.

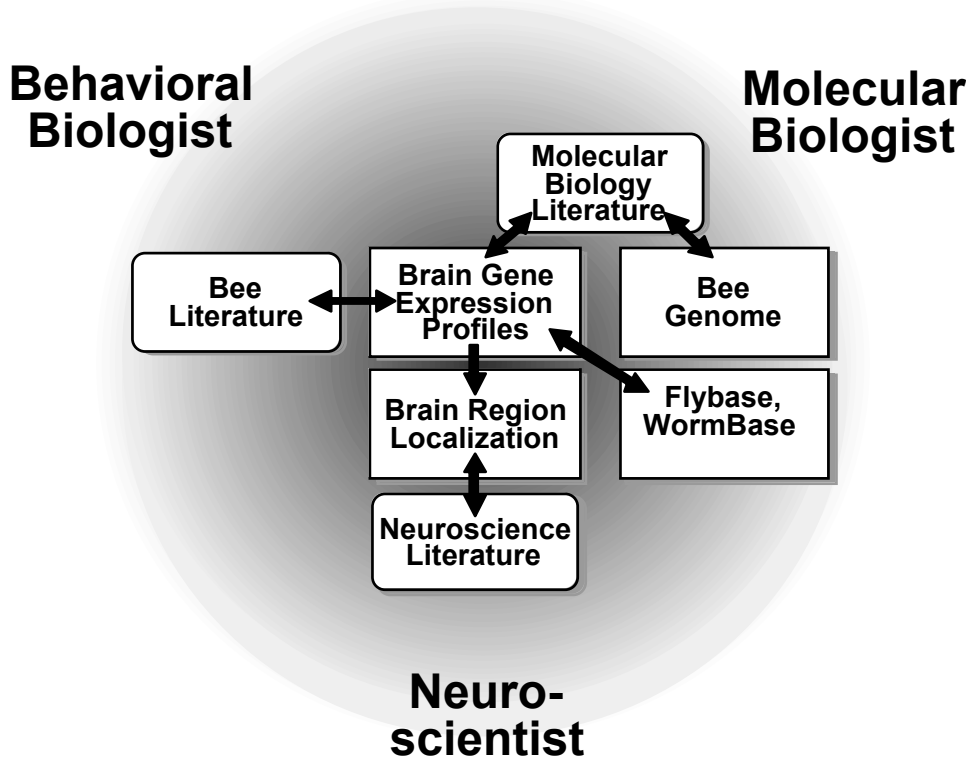


Figure 2. Conceptual Navigation in BeeSpace. Boxes show the text and data sources. Users from different backgrounds will be able to "enter" BeeSpace from the viewpoint of their community literature. They can then navigate across literature, existing data bases and our new databases to understand the relations between genes and social behavior. These

navigations enable integration across evolution, ecology, physiology, neuro and molecular biology.

The conceptual navigations will support different types of functional analysis. In data-driven mode, a biologist will use the brain gene expression profiles and the brain localizations to discover clusters of genes relevant to a particular social role. In this case, the navigations are seeking "molecular function" in the GO sense, likely to be found in the gene descriptions from a model genetic system. Annotation of these genes via the bee genome and FlyBase, plus comparison of clusters of genes for similar behaviors in other organisms, will enable development of hypotheses concerning these genes to be pursued with causal experiments.

In hypothesis-driven mode, a biologist seeking to identify genes involved in a particular social role will collect information on genes from databases and the literature in order to develop sets of candidate genes. In this case, the navigations are seeking "biological process" in the GO sense, likely to be found in the text phrases in some community literature. The candidate genes will be genes implicated in a behavior in another species, suggestive of involvement in the societal role.

coPI Robinson successfully used the candidate gene approach on an ad hoc basis several years ago. With then-graduate student Yehuda Ben-Shahar, he wanted to identify genes involved in honey bee foraging behavior. Just at this time, Osborne et al. [1997] reported the discovery of a gene (*foraging*) involved in causing *Drosophila* to collect food either in a "rover-like" or "sitter-like" manner. Similarities between these behaviors and the behavior of hive bees and foragers

were sufficient to motivate a study, which culminated in the discovery that this gene indeed plays an analogous role in the honey bee [Ben-Shahar et al. 2002].

BeeSpace will vastly increase the power of the candidate gene approach by enabling scientists to creatively and systematically navigate the literature for behavioral analogies (and their related genes), using a variety of semantic relationships to identify putatively similar behaviors. BeeSpace functional analysis will use of new informatics tools described below, which enable a biologist to interactively uncover “similar” behaviors on the basis of semantic similarity.

BeeSpace: Developing the Interactive Environment

The BeeSpace analysis environment is a software system that supports a variety of interactive navigations across information sources. The conceptual navigations fall into two basic classes corresponding to whether they primarily support data-driven research from genes to behaviors or primarily support hypothesis-driven research from behaviors to genes. It is the combination of all these interactive analysis techniques that will give BeeSpace navigation its unique power.

Data-Driven Navigation from bee genes to functional phrases using model genetic databases. These navigations will be implemented by programmers, as standard technology is adequate.

- (1) Use gene sequence similarity (e.g. bee gene to fly gene using GenBank).
- (2) Use common gene classification (e.g. bee genes to fly genes using GeneOntology).
- (3) Use common societal roles (e.g. bee genes to bee literature to fly literature to fly genes).

Hypothesis-Driven Navigation from functional phrases to bee genes using scientific literature. These navigations will be first implemented by graduate students then hardened by our programmers. Research technology is required for this implementation of the Interspace specialized to biology in general (BioSpace) and honey bees in particular (BeeSpace).

- (4) Community to Community Collection, via term-term mapping (concept switching).
- (5) Community to Community Collection, via cluster-cluster mapping (category switching).

The difference between these analytic techniques has to do with what is being switched across collections. Concept switching maps similar terms from one collection to another, enabling “low-level” switching in concept space as illustrated in the Figure 1 screendump. Category switching locates clusters of documents and common term clusters, enabling “high-level” switching in functional theme space. In previous experiments [Chen et al. 1998; Houston et al. 2000], we have used such techniques as Self-Organizing Maps [Kohonen 1997] for clustering biological literature, by finding clusters of “similar” documents within a community collection.

To automatically generate such term-term and clustering-cluster mappings in BeeSpace, we will extend our previous work [Evans & Zhai 1996; Zhai 1997a,b; Chen et. al. 1997; Chung et al. 1999], and propose new probabilistic clustering models specifically designed for BeeSpace. We will first apply part of speech (POS) tagging to identify the syntactic categories of all the words. The POS tagging technology is now fairly mature [Manning & Schuetze 1999; Brill 1995]. Based on the POS tags, we will use the biomedical domain noun phrase parser developed by our longtime Arizona colleagues [Leroy et. al. 2003], which is evolved from our previous experiments [Tolle and Chen 2000] to further identify all the noun phrases. Based on these noun phrases, we will use a robust statistical parsing technique developed by CoPI Zhai to analyze the internal modification structures of noun phrases to obtain meaning subphrases [Zhai 1997a]. The concepts can be improved through biology name extraction techniques [Hirschman et al. 2002].

We have previously used algorithms for concept spaces and concept switching [Chen et al. 1997,1999], by computing semantic similarity between terms based on co-occurrence frequency within community collections. Many other methods exist, including spreading activation [Kozima & Furugori 1993], syntactic context [Grefenstette 1994; Lin 1998], and probabilistic models [Dagan et al. 1992]. We will implement and compare these different methods using the BeeSpace text sources and choose the best method(s) for our final research system. Furthermore, we will exploit lexical resources such as the WordNet [Budanitsky & Hirst 2001] and biology domain thesauri to further improve the accuracy of the mapping [Resnik 1999].

To generate cluster-cluster mappings, we propose a new probabilistic mixture model for simultaneous clustering of terms and documents. Probabilistic mixture models have been very successfully applied to text mining and theme analysis recently [Hofmann 1999; Blei et al. 2002]. The key idea in our approach is to identify term clusters based on similar documents and at the same time to identify the document clusters by exploiting term clusters, across multiple community collections. This seemingly cyclic approach has a well-founded statistical basis, the Expectation-Maximization (EM) parameter estimation algorithm [Dempster et al. 1977]

Formally, Let C_i be a collection of documents in the i -th domain. Let a_1, \dots, a_k be k functional aspects that are known to be across these domains. Let θ_{ai} be a term distribution for aspect a_i , Let θ_{C_i} be the background term distribution for the i -th domain, and let θ_{C_i, a_j} be the residual model for domain i and aspect a_j . We assume that a document from collection C_i is generated by mixing the background model of C_i , the common aspect models and the domain specific residue models. The likelihood of the whole set of collections is

$$\log p(C) = \sum_{i=1}^n \sum_{d \in C_i} \sum_{w \in V} c(w, d) \log \sum_{j=1}^k \pi_{d,j} (\lambda_B p(w | \theta_{C_i}) + \lambda_C p(w | \theta_{a_j}) + \lambda_S p(w | \theta_{C_i, a_j}))$$

where $\pi_{d,j}$ and λ_B , λ_C , and λ_S are additional parameters in the mixture model, V is the vocabulary, and $c(w, d)$ is the count of word w in document d .

By using the EM algorithm, we can fit our mixture model to the literature available to us and estimate all the parameters. Once the parameters are estimated, we can obtain term clusters and document clusters easily. In particular, θ_{ai} would give us term distributions in the common themes across all domains, while θ_{C_i, a_j} would give us special themes within a particular domain. Documents can be easily clustered based on the posterior distribution $p(a_j | d)$. In our preliminary experiments, we tested this model with sample bee literature and fly literature, and successfully generated some interesting cross-organism functional clusters such as “foraging”. Further experiments will be conducted at much larger scale to evaluate how effectively equivalent clusters can be automatically detected across community collections.

Our final goal for research informatics is to support general comparative information retrieval, by combining the different navigation paths into a decision-theoretic framework developed in CoPI Zhai’s dissertation [Zhai 2002]. This software will take full advantage of the information gained by integrating similarity of clusters across many sources, through statistical language modeling techniques successfully applied to general text retrieval [Zhai & Lafferty 2002; Zhai et al. 2003a] and genomic information retrieval [Zhai et al. 2003b]. This statistical integration simulates navigation across collections, and may be useful in discovering behavioral analogies.

Education and Training Plan

Our plan assumes students learn science best when they are engaged in authentic scientific inquiry, making use of the methods and ideas of current science [Dewey 1933; Donovan 1999;

Driver 1985; Krajcik 1994; Minstrell 2000]. It emphasizes the importance of community, whether the learning takes place in a classroom or the larger scientific community [Bruce 2003]. We will involve middle and high school students, undergraduates and graduate students.

The education components of the BeeSpace project will be integrated with the research components. The goal is to target available resources to high school and college students prepared to make use of the opportunity. In addition, we will involve younger students in BeeSpace in the context of a stimulating but nurturing summer camp environment supportive of inquiry and intellectual growth. We will avoid superficial “meet a bee” outreach in favor of sustained development of student competence in modern integrative biology and bioinformatics. This focused approach will ensure that we do not duplicate ongoing efforts targeted at the general public (e.g. the UIUC Bee Research Facility's annual short course in beekeeping), but instead create educational opportunities vital to the success of BeeSpace.

Graduate Component. BeeSpace graduate students will work with each of the biology and informatics co-PIs on both the Illinois and Wake Forest campuses. Another project graduate student, based with the project development team at Illinois, will serve as the user coordinator and visit user labs. This person will foster a BeeSpace community through the annual BeeSpace workshop, where graduate students from the user labs will present the results of their research. In addition, they will be linked through the BeeSpace Community Inquiry Laboratory website.

Undergraduate Component. The undergraduate component will be based at Wake Forest University, a private university located in Winston-Salem, North Carolina, with an enrollment of 3700 undergraduates. coPI Fahrback will design and teach an entirely new BeeSpace-inspired course for advanced undergraduates, called “Bioinformatics for Beginners”, which targets biology majors and neuroscience minors. It will be designed by Fahrback to emphasize a rigorous and interdisciplinary approach to biology as described in *Bio 2010: Transforming Undergraduate Education for Future Research* [National Academy Press]. The anticipated enrollment of 15 students will permit a high degree of one-on-one interactions between teacher and student. Course participants will master modern tools for searching and accessing biological information, and will be the first non-expert users of BeeSpace. This course will exploit the outstanding teaching facilities available at Wake Forest University including a campus network in the process of going wireless and laptops provided for every student and faculty member. It will also tap into a very high degree of interest in neuroscience, one of the most popular minors on campus [see letter from Chair of Biology at Wake Forest University].

Students will be invited to take on the additional challenge of developing teaching materials based on bee biology and BeeSpace for middle school students. This is a natural extension of the Wake Forest emphasis on service learning (in fact, the Wake Forest undergraduate neuroscience program may be the only one in the US with a service learning requirement). Undergraduates choosing the service learning option will travel with Fahrback to UIUC during the summer after the course to serve as teachers and mentors in the middle-school summer camp, with costs covered by our BeeSpace project. The goal is to offer a rigorous and forward-looking course to a small set of well-prepared undergraduates, who in turn share their knowledge with younger students. The undergraduate course and the linked summer camp will be offered in years 2 and 4.

High School Component. BeeSpace will exploit access to the University Laboratory High School (UniHi) at UIUC. This is a small (300 students in five grades), academically-selective public high school that draws students from throughout East Central Illinois. The school is noted for a long tradition of academic excellence, and 100% of Uni students attend college. The UniHi biology teacher, David Stone, received a master’s degree in entomology from UIUC and has

strong ties to faculty on campus [see letter of collaboration]. Stone has been recognized nationally as an outstanding teacher, particularly in use of insects to lead students into key questions of biology, and won the Award from the Entomological Society of America for Outstanding Achievement in Secondary Teaching Using Insects as Educational Material in 1992.

Students at UniHi may elect a field biology and a genetics course during their junior or senior years. These students will have the option of conducting research under the supervision of BeeSpace graduate students. Stone will develop new materials to prepare students for this opportunity, and will assist in matching students to appropriate mentors and monitor the academic progress of each student. The high school students will also participate in the annual BeeSpace workshop. Our focus is exposing high school students to new ways of doing biology, but they will also contribute directly to outreach as follows. Each spring during "Agora Days," normal classes at UniHi are suspended for a week to permit students to teach courses in their own areas of interest. Students electing the BeeSpace-based field biology course in the fall will teach an Agora Days course based on this material for their fellow students in the spring.

Middle School Component. The middle school program places special emphasis on minority students in a low-resource community. In years 2 and 4, BeeSpace will offer two-week-long biology summer day camps for minority middle school students. These will be coordinated with Summer Math, an ongoing project of the UIUC Office for Mathematics, Science, and Technology Education. Summer Math targets 8th-grade students who have not previously been excited about or well-prepared for science and mathematics. Rather than focus on basic skills alone, the workshops emphasize hands-on and high-tech activities, such as I-Movie and Working with Robots. Students will be recruited through organizations such as Don Moyer Boys and Girls Club, an afterschool program offering academic support and mentoring for minority (primarily African-American) children. We will cover all costs for the summer camps (20 students each).

Wake Forest undergraduates will develop educational materials and serve as "counselors" to introduce bee biology and modern research through hands-on activities designed to stimulate inquiry. A BeeSpace education graduate student will develop a science kit for the campers, which will promote continued interest in bee biology and insect behavior. Campers will have access to resources at the Bee Research Facility, and will visit many campus venues to increase their awareness of future educational possibilities. The goal is to capitalize on the interest in bees that is naturally present in many children, and to use it to lead them into an understanding of how preparation for a career in science begins in high school and continues into college. Bruce and Fahrback are experienced camp designers, who will ensure that camp activities are aligned with Illinois Learning Standards in science and mathematics (<http://www.isbe.state.il.us/ils/>).

We will use genomics and informatics to develop a modern reformulation of the roots of behavior that goes beyond the nature-nurture dilemma. Research and Education, in Biology and Informatics, will intertwine in the BeeSpace project. We will bring together two different universities, a large public and a small private, for research and education. Interdisciplinary faculty from international research communities will link with graduate students, graduate students with high school students, undergraduates with middle school students, and science-passionate high school students with their peers. Minority middle school students will be supported in choosing high school course paths towards higher education and careers in biology. Our project will use a compelling, attractive, and economically important animal to address one of the leading issues of the day, the roots of behavior.

References

- Ashburner, et. al. (Gene Ontology Consortium) [2001] Creating the Gene Ontology Resource: Design and Implementation, *Genome Research* 11: 1425-1433.
- Bennett, N., He, Q., Powell, K., Schatz, B. [1999] Extracting Noun Phrases for All of MEDLINE, *AMIA '99 (American Medical Informatics Assoc) Annual Conf*, Washington, DC, Nov, 671-675. won **Best Paper** award.
- Ben-Shahar, Y., Robichon, A., Sokolowski, M., Robinson, G. [2002] Influence of gene action across different time scales on behavior, *Science* 296:741-744.
- Blei, D., Ng, A., Jordan, M. [2002] Latent Dirichlet allocation, *Advances in Neural Information Processing Systems* vol. 14.
- Brill, E. [1995] Transformation-Based Error-Driven Learning and Natural Language Processing, *Computational Linguistics* 21(4): 543-565.
- Bruce, B., Bishop, A. [2002] Using the web to support inquiry-based literacy development, *Journal of Adolescent and Adult Literacy*, 45(8), 706-714.
- Bruce, B. [2003] The role, value, and limits of S&T data and information in the public domain for education, In P. Uhlir (ed), *The role, value, and limits of S&T data and information in the public domain* (pp. 56-59), Washington, DC, National Academy Press .
- Budanitsky, A. and Hirst, G. [2001] Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures, *Workshop on WordNet and Other Lexical Resources, in North American Chapter Association Computational Linguistics (NAACL-2000)*, Pittsburgh.
- Callaerts, P., Halder, G., Genring, W. [1997] PAX-6 in development and evolution, *Ann. Rev. Neurosci.* 20:483-532.
- Capaldi, E., Robinson, G., Fahrback, S. [1999] Neuroethology of spatial learning: The birds and the bees, *Annual Review Psychology* 50: 651-682.
- Chen, H., Martinez, J., Ng, D., Schatz, B. [1997] A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System, *J. American Society Information Science*, 48(1): 17-31 (Jan).
- Chen, H., Houston, A., Sewell, R., Schatz, B. [1998] Internet Browsing and Searching: User Evaluations of Category Maps and Concept Space Techniques, *J. American Society for Information Science*, 49(7):582-603.
- Chen, H., Chung, Y., Houston, A., Li, P., Schatz, B. [1999] Using Neural Networks for Vocabulary Switching. Special Issue on Applications of Intelligent Information Retrieval, *IEEE Expert*.
- Chung, Y., He, Q., Powell, K., Schatz, B. [1999] Semantic Indexing for a Complete Subject Discipline, *4th Int ACM Conf on Digital Libraries*, Berkeley, CA, Aug, 39-48.
- Churchill, G. [2002] Fundamentals of experimental design for cDNA microarrays, *Nature Genetics* 32: 490-495.
- Clough, S., Zou, J., Rodriguez-Zas, S., Vuong, T., Li, M., Vodkin, L. [2004] Expression profiling soybean responses to pathogens, *2nd Int. Conf. Legume Genomics and Genetics*, Dijon, France, June.
- Cui, X., and Churchill, G. [2003] Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* 4(4):210. Epub 2003 Mar 17. Review.
- Cutting, D., Kupiec, J., Pedersen, J., Sibun, P. [1992] A practical part-of-speech tagger, *Proc. Third Conference Applied Natural Language Processing*, pp 133-140, Trento, Italy.
- Dagan, I., Marcus, S., Makovitch, S. [1992] Contextual word similarity and estimation from sparse data, *Proc 30th Annual Meeting Association Computational Linguistics*, pp. 164-171.
- Dempster, A., Laird, N., Rubin, D. [1977] Maximum likelihood from incomplete data using the EM algorithm, *Journal of the Royal Statistical Society*, 39(B), pp 1-38.
- Dewey, J. [1933 (1910)] *How We Think* , Lexington, MA: D.C. Heath.
- Donovan, M., Bransford, J., Pellegrino, J. [1999] *How People Learn: Bridging Research and Practice*, Washington, DC: National Academy Press.
- Driver, R., Guesni, E., Tiberghiem, A. [1985] *Children's Ideas in Science*, Philadelphia: Open University Press.
- Evans, D. and Zhai, C. [1996] Noun phrase analysis in unrestricted text for information retrieval, *Proc. 34th Ann. Meeting Assoc. Computational Linguistics*, pp 17-24, Santa Cruz, USA.
- Fahrback, S. [2003] A taste for learning?, *J. Comparative Neurology* 465:164-167.
- Farris, S., Robinson, G., Fahrback, S. [2001] Experience- and age-related outgrowth of intrinsic neurons in the mushroom bodies of the adult worker honey bee. *J. Neuroscience* 21: 6395-6404.
- Grefenstette, G. [1994] *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Publishers.

- Grozinger, C., Sharabash, N., Whitfield, C., G. Robinson [2003] Pheromone-mediated gene expression in the honey bee brain, *Proc National Academy of Sciences* Suppl. 2: 14519-25.
- Hirschman, L., Park, J., Tsujii, J., Wong, L., Wu, C. [2002] Accomplishments and challenges in literature data mining for biology, *Bioinformatics* 18 (12) 1553-1561 (2002).
- Hofmann, T. [1999] Probabilistic latent semantic indexing, *Proc 22nd ACM-SIGIR Int. Conference Research and Development in Information Retrieval*, pp. 50-57.
- Houston, A., Chen, H., Schatz, B., Sewell, R., Doszkocs, T., Ng, D. [2000] Exploring the Use of Concept Space and Category Map Techniques to Improve Medical Information Retrieval, *Decision Support Systems*, Special Issue on Decision Support for Health Care in a New Information Age, 30(2): 171-186 (Dec).
- Jamison, C., Mills, B., Schatz, B. [1996] ENQUIRE: An Extensible Network Query Unification System for Biological Databases, *Bioinformatics* 12(2): 145-150 (Apr).
- Jin, W., Riley, M., Wolfinger, R., White, P., Passador-Gurgel, G., Gibson, G. [2001] The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*, *Nature Genetics* 29; 389-395.
- Kohonen, T. [1997] *Self-Organizing Maps*. 2nd edition, Springer Series in Information Sciences, Vol. 30, NY: Springer-Verlag.
- Kozima, H. and Furugori, T. [1993] Similarity between words computed by spreading activation on an English dictionary, *Proc 6th Conference European Chapter Assoc Computational Linguistics*, pp. 232-239.
- Krajcik, J., Blumenfeld, P., Marx, R., Soloway, E. [1994] A Collaborative Model for Helping Middle Grade Science Teachers Learn Project-based Instruction, *The Elementary School Journal*, 94(5), 483-497.
- Leroy, G., Chen, H., Martinez, J. [2003] Shallow Parser Based on Closed-Class Words to Capture Relations in Biomedical Text, *Journal Biomedical Informatics*, vol. 36, pp 145-158, June 2003.
- Lin, D. [1998] Automatic retrieval and clustering of similar words, *Proc COLING/ACL-98*, pp. 768-774, Montreal.
- Manning, C. and Schuetze, H. [1999] *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA.
- Minstrell, J., Van Zee, E. (Eds) [2000] *Teaching in the Inquiry-Based Science Classroom*, Washington, DC: American Association for the Advancement of Science.
- Mizunami M, Weibrecht J.M., Strausfeld N.J. [1993] A new role for the insect mushroom bodies: Place memory and motor control. In R. Beer, R. Ritzman and T. McKenna (ed.): *Biological Neural Networks in Invertebrate Neuroethology and Robotics*, Academic Press, Inc., New York: pp. 199-225.
- Osborne, K., Robichon, A., Burgess, E., Butland, S., Shaw, R., Coulthard, A., Sokolowski, M. [1997] Natural behavior polymorphism due to a cGMP-dependent protein kinase of *Drosophila*, *Science* 277: 763-4 (Aug 8).
- Oster, G. and Wilson, E. [1978] *Caste and Ecology in the Social Insects*, Princeton University Press.
- Resnik, P. [1999] Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language, *Journal Artificial Intelligence Research*, 1999.
- Robinson, G. [1987] Regulation of honey bee age polyethism by juvenile hormone, *Behavioral Ecology and Sociobiology* 20: 329-338.
- Robinson, G., Page, R., Strambi, C., Strambi, A. [1989] Hormonal and genetic control of behavioral integration in honey bee colonies, *Science* 246: 109-112.
- Robinson, G. [2002a] Sociogenomics takes Flight, *Science* 297:204-5 (12 Jul).
- Robinson, G. [2002b] Genomics and integrative analyses of division of labor in honey bee colonies, *American Naturalist* 160: S160-S172.
- Rodriguez-Zas, S. [2002] Comparison of statistical methods to study cDNA microarray data, *J. Dairy Sci.* 80 (Suppl. 2):10.
- Rodriguez-Zas, S., Band, M., Everts, R., Southey, B., Liu, Z., Lewin, H. [2003] Analysis of gene expression patterns in the cattle digestive system, *J. Animal Sci.* 81(Suppl. 1):628.
- Schatz, B. [1997] Information Retrieval in Digital Libraries: Bringing Search to the Net, *Science* 275: 327-334 (Jan). Cover article for special issue on Bioinformatics.
- Schatz, B., Mischo, W., Cole, T., Bishop, A., Harum, S., Johnson, E., Neumann, L., Chen, H., Ng, D. [1999] Federated Search of Scientific Literature, *IEEE Computer*, Special Issue on Digital Libraries, 32: 51-59 (Feb).
- Schatz, B. [2002a] Building Analysis Environments: Beyond the Genome and the Web, *IEEE Intelligent Systems* 17: 70-73 (May/June). Special issue on Mining Information for Functional Genomics.
- Schatz, B. [2002b] The Interspace: Concept Navigation across Distributed Communities, *Computer* 35(1): 56-62 (Jan). Information Infrastructure article for annual Trends issue.
- Science [1993] Networking the Worm, 261: 842, sidebar within lead news article Beyond E-mail and Databases.
- Science [1998] Taming MEDLINE with Concept Spaces, 281:1785, sidebar within lead news article Assembling the World's Biggest Library on your Desktop.

- Shoman, L., Grossman, E., Powell, K., Jamison, C., Schatz, B. [1995] The Worm Community System, Release 2.0 (WCSr2), in H. Epstein & D. Shakes (eds), *C. elegans: Modern Biological Analysis of an Organism*, chap 26, *Methods in Cell Biology* 48: 607-625. <http://www.canis.uiuc.edu/projects/wcs>
- Seeley, T. [1982] Adaptive significance of the age polyethism schedule in honey bee colonies, *Behavioral Ecology and Sociobiology* 11: 287-293.
- Seeley, T. [1985] *Honey bee Ecology: A Study of Adaptation in Social Life*, Princeton University Press.
- Seeley, T. [1995] *The Wisdom of the Hive: Social Physiology of Honey Bee Colonies*, Harvard University Press.
- Strausfeld, N., Hansen, L., Li Y., Gomez, R., Ito, K. [1998] Evolution, discovery, and interpretations of arthropod mushroom bodies, *Learning and Memory* 5:11-37.
- Tolle, K. and Chen, H. [2000] Comparing noun phrasing techniques for use with medical digital library tools, *J. American Society Information Science* 51(4): 352-70 (Mar).
- Thakkar, U., Yeh, M., Bruce, B., Jakobsson, E. [2000] Experts using Biology Workbench: A case study and challenges of open-world environments for education, *Proc ED-MEDIA 2000, World Conf Educational Multimedia, Hypermedia, Telecommunications*, Charlottesville, VA, Assoc Advancement Computing Education.
- Whitfield, C., Band, M., Bonaldo, M., Kumar, C., Liu, L., Pardinis, J., Robertson, H., Soares, M., Robinson, G. [2002] Annotated Expressed Sequence Tags and cDNA Microarrays for Studies of Brain and Behavior in the Honey Bee, *Genome Research* 12: 555-566. Cover article.
- Whitfield, C., Cziko, A., Robinson, G. [2003] Gene expression profiles in the brain predict behavior in individual honey bees. *Science* 302:296-299.
- Williams, R. [2000] Mapping genes that modulate brain development: a quantitative genetic approach, In: *Mouse brain development*, Goffinet, A., Rakic P., (eds), Springer-Verlag, New York, pp 21-49.
- Wilson, E. [1971] *The Insect Societies*, Harvard University Press.
- Wilson, E. [1975] *Sociobiology: A New Synthesis*, Harvard University Press.
- Wilson, E. [1976] Behavioral discretization and the number of castes in an ant species, *Behavioral Ecology and Sociobiology* 1(2): 141-154.
- Winston, M. [1987] *The Biology of the Honey Bee*, Harvard University Press.
- Withers, G., Fahrback, S., Robinson, G. [1993] Selective neuroanatomical plasticity and division of labour in the honey bee. *Nature* 364: 238-240.
- Zhai, C. [1997a] Fast statistical parsing of noun phrases for document indexing, *5th Conf on Applied Natural Language Processing (ANLP-97)*, pp 312-319, Apr 1997.
- Zhai, C. [1997b] Exploiting context to identify lexical atoms – a statistical view of language context, *Proc International Conf on Modeling and Using Context (CONTEXT-97)*, pp119-129.
- Zhai, C. [2002] *Risk Minimization and Language Modeling in Text Retrieval*, Ph.D. dissertation, Department of Computer Science, Carnegie Mellon University.
- Zhai, C. and Lafferty, J. [2002] Two-stage Language Models for Information Retrieval, *Proc ACM SIGIR 2002*, pp. 49-56.
- Zhai, C., Cohen, W., Lafferty, J. [2003a] Subtopic Retrieval – Methods and Evaluation Metrics, *Proc ACM SIGIR 2003*, pp 10-17.
- Zhai, C., Tao, T., Fang, H., Shang, Z. [2003b] Improving Robustness of Language Models – UIUC TREC-2003 Genomics and Robust Experiments, *Proceedings of TREC-2003*.
<http://trec.nist.gov/pubs/trec12/papers/uillinois-uc.robust.genomics.pdf>

Key Personnel

Bruce R. Schatz, PI and informatics lead, is Director of the Community Architectures for Network Information Systems (CANIS) Laboratory at the University of Illinois at Urbana-Champaign. He served as Principal Investigator of the NSF/DARPA/NASA Digital Libraries Initiative project, a \$4M flagship effort in the Federal Program in National Information Infrastructure. This project built a large-scale testbed of structured documents from scientific journals, still being served by the Engineering Library for federated search to 2500 users. He then served as PI of the \$4M flagship project in the DARPA Information Management Program, which built a prototype analysis environment to support community repositories (Interspace). The Interspace Prototype was used to semantically index all of MEDLINE, winning Best Paper at the 1999 annual meeting of the American Medical Informatics Association.

He is Professor in Library and Information Science, Computer Science, Neuroscience, Biomedical and Health Information Sciences. He is Senior Research Scientist at the National Center for Supercomputing Applications (NCSA), serving as scientific advisor for information systems. He has served in this role since 1989, including the period during which NCSA developed Mosaic, the browser that catalyzed the Web. He holds a PhD in computer science from Arizona, with postdoctoral training in molecular biology, a MS in artificial intelligence (computational neuroscience) from MIT, a BA in mathematical sciences from Rice University.

He also spent five years at the University of Arizona, where he was PI of a \$1.5M NSF National Collaboratory project which built the Worm Community System in molecular biology that served as a foundation of current Bioinformatics databases and was highlighted as the future model for science information systems in news articles in *Science* and NRC reports.

He is a AAAS fellow and was an NSF Young Investigator. He has published 2 cover articles in *Science*, the leads in special issues on Bioinformatics, and 3 lead articles in *Computer*, the computing equivalent to *Science*. For more than a decade, his unique science information systems with research technologies and real users have been regularly featured in lead news articles in *Science* and *Nature* (11 separate news articles with screendumps and quotes).

Gene E. Robinson, coPI and biology lead, is the G.W. Arends Professor of Integrative Biology. He is Director of the Neuroscience Program, and Professor of Entomology, Cell & Structural Biology, and Animal Biology. He received his Ph.D. from Cornell University in 1986 and was an NSF Postdoctoral Fellow at Ohio State University before joining the faculty in 1989.

Robinson's research group studies the regulation of social behavior, using the honey bee. The research is integrative, involving analyses of behavior, communication, hormones, brain chemicals, brain structure, and expression of genes in the brain. The model system is the bee colony's division of labor, based on an intricate process of individual behavioral maturation that results in age-related changes in occupation.

He has authored or co-authored over 140 papers including over 100 peer-reviewed original reports, including 15 in *Nature*, *Science* and *PNAS*. He has pioneered the application of genomics to the study of social behavior, spearheaded the effort to gain approval from NIH for the sequencing of the honey bee genome, and heads the Honey Bee Genome Sequencing Consortium. His honors include: University Scholar of University of Illinois; AAAS Fellow; Certificate of Distinction from International Congress of Entomology; Burroughs Wellcome Innovation Award in Functional Genomics; the Founders Memorial Award from Entomological Society of America; a Fulbright Senior Research Fellowship; a Guggenheim Fellowship.

Susan E. Fahrbach, neuro anatomy, is Professor of Entomology and of Neuroscience. She works on development encompassing the whole life span. In the past decade her focus has been on the structural plasticity of the adult honey bee brain.

She is a member of the Editorial Board of the *Journal of Comparative Neurology*, the leading neuroanatomical journal. Since 1992, she has collaborated with G.E. Robinson on studies of honey bee brain and behavior. Their demonstration in 1993 (*Nature* 364: 238-240) that forager bees have a larger volume of neuropil associated with the mushroom bodies than do hive bees represented the first application of modern stereological techniques to the insect brain. The techniques pioneered by Fahrbach are now widely used to study neuroplasticity in insect brains. She is a AAAS Fellow and a University Scholar of the University of Illinois.

She has also been deeply involved in undergraduate education. In addition to personally mentoring 41 undergraduate researchers since 1988, she has served as PI of an NSF REU Site on Honey Bee Brains and Behavior, and is currently PI on an NSF Undergraduate Mentoring in Environmental Biology Program. Since 1998, she has served as the Director of the Illinois' HHMI-sponsored program for undergraduate education in biology, in which role she directs both an undergraduate research program and leads three science outreach programs (BOAST, for K-5 students; Prairie Flowers, for middle school science teachers; and BEOP, a biotechnology program for high school teachers).

Her desire to develop further her skill in undergraduate teaching has led her to accept a new position as Reynolds Professor of Developmental Neuroscience in the Department of Biology at Wake Forest University in Winston-Salem, North Carolina. Her laboratory will move to Wake Forest University in August 2004, where she will be during our FIBR project. She will be given considerable resources at Wake Forest, including a state-of-the-art laboratory designed specifically for insect neuroanatomy and close ties to the Bowman Gray School of Medicine.

ChengXiang Zhai, text analysis, is Assistant Professor of Computer Science, with a joint appointment in Information Science. He holds a Ph.D. in Computer Science from Nanjing University and a Ph.D. in Language and Information Technologies from Carnegie Mellon University. He has extensive research experience on natural language text analysis and information retrieval from both academia and industry. His accomplishments include developing effective algorithms for noun phrase analysis and personalized information filtering. He is a recent recipient of an NSF CAREER award from IDM-CISE for information retrieval.

His recent work on applying statistical language models to information retrieval represents a new generation of models for searching text. He developed a new general framework for information retrieval based on Bayesian decision theory, which facilitates modeling complex retrieval problems and automatic tuning of performance. Recently, working with biologists, he developed several algorithms for biological data analysis, including algorithms for predicting the function of an unknown protein motif by mining the gene ontology annotations and new algorithms for clustering microarray data with order constraints and statistical significance.

He has also significant experience with developing information management software. While working in Clairvoyance Corp., he developed a commercial toolkit underlying the ConceptBase software product, which won the "Software of the Year" award in Japan. He also led a team working on personalized information filtering system. The filtering techniques he developed consistently perform well in TREC – the premier international text retrieval evaluation workshop sponsored by NIST. These techniques resulted in four US patents. While at Carnegie-Mellon University, he was the major architecture designer and implementor of Lemur, an information retrieval and language model toolkit that is now used worldwide for both research and education.

Sandra Rodriguez-Zas, data analysis, is Assistant Professor of Animal Science and of Statistics. She received her MS and Ph.D. degrees at the University of Wisconsin-Madison. Her research program is a continuous quest to understand the genetic architecture of health, social behavior, agricultural and other complex characteristics through statistical genomics and bioinformatics. This quest is directed by a methodical process of evaluating complementary approaches and identifying parsimonious models to study complex characteristics in humans, laboratory animals, insects, livestock and plants including pain, social behavior, milk production, and meat quality. Outcomes of her research help to discriminate nature from nurture and enhance the opportunities to cure and prevent diseases, and to understand complex biological systems.

She is the PI on a \$1M NIH NIGMS funded project with G. Robinson on multifactorial gene expression in honey bees. She also collaborates across campus in a wide variety of gene expression projects using microarrays to study: nutritional effects in dairy cows (USDA), meat quality in pork (USDA), development in *Drosophila*, and growth in soybean, among others.

As an executive committee member of the Agricultural Genome Science and Public Policy Training Program (<http://www.ansci.uiuc.edu/AGGENOME/>), she fosters the recruiting of graduate students with interest in the areas of genomics, communication and policy making. Likewise, Dr. Rodriguez-Zas is a faculty on the GK-12 EdGrid Graduate Teaching Fellows Program (NSF DGE), to assist high school teachers with bioinformatics tools in their classrooms.

Bertram (Chip) Bruce, education and training, is Professor of Library and Information Science, Curriculum and Instruction, and Bioengineering. Before moving to the University of Illinois in 1990, he was faculty in Computer Science at Rutgers University and Principal Scientist in at Bolt Beranek and Newman (1974-90). He received a BA in Biology from Rice University and a Ph.D. in Computer Sciences from University of Texas at Austin in 1971.

His central interest is in learning, the constructive process whereby individuals and organizations develop as they adapt to new circumstances. This work draws on ideas such as John Dewey's theory of inquiry, as well as on action research and situated studies. Much of it has focused on changes in the nature of knowledge, community, and literacy, as discussed in his new book, *Literacy in the Information Age: Inquiries into meaning making with new technologies*. He has long been a prominent researcher in the use of computers in curriculum development, as witnessed by his books a decade ago on *Network-based classrooms: Promises and realities* and *Electronic Quills: A situated evaluation of using computers for writing in classrooms*.

He has been a principal in several major NSF Educational Technology projects for science education. These include: the Distributed Knowledge Research Collaborative, the Biology Student Workbench, Plants Pathogens & People, Physics Outreach, and SEARCH. His research has also contributed to the development of environments for inquiry-based learning. These include the Inquiry Page, which is a website, a set of communities, and a research project on inquiry; Biology Student Workbench, an interactive resource to support investigations in molecular biology and evolution; Chickscope, a project in which students engage with a community of inquiry as they use remote instrumentation to study chicken embryology; Quill, tools and environments to support literacy learning; Statistics Workshop, an interactive system for learning statistical reasoning; and Discoveries, a series of CD-ROM-based multimedia environments for supporting students' inquiries in science and social studies.

Bruce is the lead for Education and Outreach for a new NSF Science and Technology Center on Advanced Materials for Water Purification led by the University of Illinois at Urbana-Champaign. He is a Fellow of the National Conference on Research in Language and Literacy.

Management Plan: Organization and Timeline

The investigators for the BeeSpace project have been described above. The Project Organization naturally flows out of the skills and interests of the individual investigators. This is a fully integrated team, who understand and appreciate each other's expertise and viewpoints.

The investigators have been meeting regularly for long lunches and monthly sessions, during the nearly two years required for preparation of this proposal. This form of interaction has worked well, with periodic meetings to share general progress supplemented by special meetings for specific purposes. To collaborate with our diverse international user community, we will employ a full-time user coordinator. We also will host an annual workshop for all project members and all user laboratories, with all costs covered, which will support a formal mechanism for sharing results and gaining feedback. This structure of informal meetings and formal workshops served well in previous projects that built research systems with real users.

PI Schatz is an *experienced manager* of systems projects, particularly those which develop systems with research technologies and deploy these to working scientists. During the past decade, he served as the PI of the NSF BIO flagship at the University of Arizona that developed the Worm Community System and deployed WCS to an international community of molecular biologists (30 registered labs). He also served as the PI of the NSF CISE flagship at the University of Illinois that developed the Illinois Digital Library and deployed this digital library to thousands of faculty and students around the University and Big Ten (3000 registered users).

The BeeSpace project will be Schatz's primary systems research during the period of the grant. As cost matching for this proposal, the Institute for Genomic Biology (IGB) will be paying half of his salary, to provide a substantial time buyout [see letter of support]. This half-time management position will enable Schatz to spend the time necessary to carry out this substantial systems project. A similar half-time position at the National Center for Supercomputing Applications (NCSA) enabled Schatz to successfully carry out the Illinois Digital Library project, matching a similar NSF flagship grant (\$5M including supplements for the 5-year period from 1994 to 1998).

A complex project such as BeeSpace would not be possible without the full range of skills and interests. The *diversity of the team* is a key feature of the project, and critical to its successful completion. Two universities are represented – the large public University of Illinois and the small private Wake Forest University.

The bulk of the project is at the University of Illinois, a great research institution with a significant allotment of NSF funds. The major Colleges are represented, e.g. Engineering and Agriculture are the north and south pillars at Illinois. Two promising junior faculty represent these Colleges -- Zhai is in the Department of Computer Science where he was recently awarded an NSF CAREER award in Information Management, while Rodriguez-Zas is in the Department of Animal Sciences where she was recently awarded an NIH grant in Mathematical Biology.

The middle of campus is dominated by the College of Arts and Sciences and the various professional schools. Two prominent senior faculty represent these colleges – Robinson is in the School of Integrative Biology where he was recently awarded an endowed chair, while Schatz is in the School of Library and Information Science where he runs a campuswide facility (CANIS) for information systems projects. It should be noted that the Graduate School of Library and Information at the University of Illinois has long been rated the number 1 nationally in this

professional field. coPI Bruce has been a longtime faculty in the College of Education, another professional school, but recently moved to Library and Information Science.

Robinson is also Director of the campuswide interdisciplinary Neuroscience Program, whose affiliated faculty include Fahrbach and Schatz. Fahrbach has been a longtime faculty in Integrated Biology, but will be moving this summer to assume an endowed chair at Wake Forest University, to better pursue her significant interests in undergraduate education.

The responsibilities of the investigators naturally follow their skills and interests. Most of the research will be performed by *graduate students*.

Three Library and Information Science students will be supervised by PI Schatz and coPI Bruce: a collection librarian will organize the scientific literature used as the core for the BeeSpace (Schatz), a education coordinator will organize the education and outreach activities, including the high school course and the middle school minority workshop (Bruce), a user coordinator will support the biologists using the BeeSpace in their laboratories (joint).

Two Biology students will be supervised by coPI Robinson: a specimen preparer will collect the bees during performance of the appropriate behavior, a functional analyst will use the BeeSpace environment to identify functional information about the expressed genes. A third Biology student, based at Wake Forest University, will be supervised by coPI Fahrbach for anatomical dissections. A professional technician will carry out the actual biology experiments, for expressions at Illinois under Robinson and for localizations at Wake Forest under Fahrbach.

Two Statistics students will be supervised by coPI Rodriguez-Zas: one will process the raw data from the microarray expressions into a standard database, one will analyze the gene products to determine which are differentially expressed for the particular behavior.

Two Computer Science students will be supervised by coPI Zhai: one will develop natural language processing software for extracting noun phrases from the literature documents, one will develop concept switching algorithms for comparing phrases across text collections. Each student's role is to develop algorithms and software for concept navigation, which will later be incorporated by the programmers into the production BeeSpace environment. The professional programmers will be supervised by Schatz. These students will work with the programmers, just as Zhai and Schatz work together, to develop algorithms and deploy systems.

The BeeSpace project will be housed in the *Institute for Genomic Biology* (IGB) at the University of Illinois at Urbana-Champaign (UIUC). This is a new Institute established by the University to pursue pioneering research at the intersection between biological science and computational engineering. Its new building in the heart of campus next to the University Library will be opening in 2006, during our FIBR project. The IGB is modeled after the University's highly successful Beckman Institute, as a central facility where interdisciplinary faculty can carry out pioneering projects in science and technology. See Facilities section.

BeeSpace would be IGB's very first major project. As such, UIUC in general and the IGB in particular are prepared to provide major institutional support. The letter of support from the founding IGB Director explicitly promises office space for all project members, including desktop computers, plus all administration support for computers and monies. Our Biology lead Robinson is the Head of the one of the already approved IGB Themes (on Neurogenomics).

To further insure that our BeeSpace project has the maximum impact on the biological community, we have formed an *Advisory Committee* of internationally prominent biologists and informaticians. [see email letters] The Advisory Committee members are:

William Gelbart, insect neurobiology, PI of FlyBase at Harvard University
 John Hildebrand, insect neurobiology, Director of Neuroscience at University of Arizona
 Chris Fields, genome informatics, New Mexico State University,
 former Chief Scientist at TIGR and at PE Informatics (parent company of Celera)
 Stanley Watson, genome biology, endowed chair neuroscience University of Michigan
 PI of the largest gene expression project on human brains (\$40M Pritzker Network)
 E. O. Wilson, sociobiology, endowed chair natural history Harvard University

Timelines of Tasks

The Biology Research has two major tasks: generating Expressions for each societal role and generating Localizations for each expression experiment. Expressions must be done before the Localizations. These biology experiments must then be analyzed via Statistics to determine the differentiating gene clusters regionally expressed in societal roles.

The Informatics Research has two major tasks: developing Collections that contain functional information and Indexing these collections to support functional analysis. The software for indexing can be developed in parallel with the gathering of collections, but the actual indexing must take place after the collections are assembled. These indexes must then be integrated into the interactive BeeSpace environment.

BeeSpace Version One will be an initial prototype, developed to get implementation experience and deployed to get user feedback from the early adopter labs. The interactive environment will then be completely redesigned. *BeeSpace Version Two* will be developed as a fully fledged implementation and deployed to all the committed user labs.

The Education and Outreach has a major task at each level of educational maturity: Graduate, Undergraduate, High School, Middle School. The Graduate task will be using the research system already discussed. The Undergraduate task will develop a bioinformatics course at Wake Forest University. The High School task will integrate BeeSpace into the Field Biology course at University Laboratory High School. The Middle School task will utilize the Undergraduates as mentors for minority students at a summer outreach camp.

At the end of the 5-year project, we will be ready to present the BeeSpace environment as a working model for functional analysis in the post-genome era of biology. Then we will be ready to propose a project to develop and deploy the *BioSpace* environment as a biological infrastructure that can support all biological knowledge for all biological disciplines.

Task	Year 1	Year 2	Year 3	Year 4	Year 5
<i>Expressions</i>	First Half	Completed			
<i>Localizations</i>	Samples	Quarter	Half	Completed	
<i>Statistics</i>	Samples	Quarter	Quarter	Half	Completed
<i>Collections</i>	Databases	Literatures	Books	Re-Partition	Completed
<i>Indexing</i>	Parser	Indexer	Switcher	Tuning	Completed
<i>BeeSpace</i>	Servers	Collections	Version One	Re-Design	Version Two
<i>Education</i>	Planning	Undergraduate	HighSchool	Undergraduate	HighSchool
<i>Outreach</i>	Planning	SummerCamp	Science Kits	SummerCamp	Evaluation

Sharing the Outcomes of the Research

The BeeSpace project will develop an interactive environment for functional analysis and deploy this system to an international community of biologist users. As a deployed system with research technology, the results of our research will be widely shared.

The User Labs are listed in the Table below. Each of these laboratories has committed to using the BeeSpace in their research to evaluate its effectiveness. See attached letters for details.

We will host an Annual User Workshop, covering all costs at project expense. This will enable the project developers to present the current system and discuss on-going research, and the user laboratories to discuss their needs and describe their reactions. A day and a half is planned each year at our central site in the Institute for Genomic Biology. Our Advisory Committee of prominent biologists will be invited to attend this workshop.

After the end of the NSF funding, we hope to establish BeeSpace as a permanent resource. Similar efforts were successful in the PI's NSF Digital Libraries project, with the institutional support of the University Library, in bridging and propagating before permanent external funding was garnered. The University Library is now establishing a program for Institutional Archives, and we are discussing with them the possibility of such long-term support.

In any case, we will make the BeeSpace collections and softwares freely available to the scientific community. The databases generated from the biology research will be archived, both the expression recordings and the anatomical slides, as well as the statistical interpretations. The text collections that can be made freely available will be. But note that FIBR project resources, by themselves, will not be adequate to adapt the system to particular needs of particular communities. Further resources will be necessary to adapt to communities beyond bees.

The Inquiry Page <http://inquiry.uiuc.edu> project was developed to foster a growing bioinformatics education community. For example, the NCSA Biology Workbench is widely recognized as a significant bioinformatics resource providing a suite of interactive tools. (The first version was developed using PI Schatz's Worm Community System [Jamison 1996].) The Inquiry Page supports incorporation of the Workbench into day-to-day educational activities, in a way that encourages an inquiry-based approach to teaching and learning [Bruce 2003; Thakkar et al. 2000]. All units posted to the Inquiry Page can be searched by keyword, and units can be viewed by the public. Additionally, Inquiry project participants can give feedback on others' pages, and develop and post their own Inquiry Units, to encourage collaboration at all levels.

Education lead Bruce also leads education and outreach activities for a new NSF S&T Center at UIUC on Advanced Materials for Water Purification. For this NSF Center, the Inquiry system again supports interactive discovery, beyond traditional online documentation or curricular materials, for education and outreach at a wide variety of levels and purposes.

We plan to develop a BeeSpace environment for learning by making use of the Inquiry Page and Community Inquiry Labs tools for collaboration within our project and for easier availability to the scientific community. All materials developed will be made freely available. This is part of the way that our educational activities will bring many groups of persons together. BeeSpace will link diverse faculty with graduate students, graduate students with high school students, science-passionate high school students with their peers, and undergraduates with minority middle school students. The education will also bring representatives of the two participating campuses together, with interested members of the scientific community.

List of User Laboratories

Our goal for initial BeeSpace deployment is to focus on a carefully selected set of labs that all employ molecular analyses of either behavior, social life, or both. We especially selected productive laboratories using the honey bee as a model for sociogenomic study of nature and nurture. We wanted laboratories who would commit to using our systems for their research, and work with us during the lengthy development period. We will be collaborating with these scientists intensively, through our user coordinator and user workshops.

These laboratories represent only a small fraction of those who will ultimately use BeeSpace. As indicated in the successful White Paper for the Sequencing of the Honey Bee Genome www.genome.gov/Pages/Research/Sequencing/SeqProposals/HoneyBee_Genome.pdf, there are ca. 150 bee labs in the world (genetics, physiology, pathology, behavior; 30 molecular biology) and ca. 350 social insect/Hymenoptera labs worldwide. As indicated by the letters of users, we expect that many members of the 3000+ *Drosophila* research community will be heavy users of BeeSpace. coPI Robinson initiated a new Gordon Research Conference series on Genes and Behavior and chaired the first one in February 2004, which emphasized molecular approaches to understanding social behavior. This GRC had 135 participants and was oversubscribed.

Fifteen (15) User Laboratories have been chosen as enough to test our prototypes carefully, but few enough to enable our developers to concentrate on improving functionality rather than providing continuity. For example, with twice as many labs, the PI's Worm Community System (WCS) had a major support issue, diluting the resources. The particular User Laboratories were chosen to distribute across place (geographical location) and topic (scientific concentration).

The User Labs fall into 3 roughly equal categories: Domestic Bee Labs, Foreign Bee Labs, Model Genetic Systems. Experience with WCS showed that often the most valuable usage was from related organism labs, such as *Drosophila*.

Domestic Bee Labs

Robert Page, University of California at Davis
Diana Wheeler, University of Arizona
Gene Robinson, University of Illinois
Susan Fahrback, Wake Forest University, North Carolina
Jay Evans, USDA, Maryland

Foreign Bee Labs

Martin Beye, Martin Luther Universitaet, Germany
Guy Bloch, Hebrew University, Israel
Takeo Kubo, Tokyo University, Japan
Zilá Paulino Luz Simões, Brazil
Alison Mercer, University of Otago, New Zealand
Ryszard Maleszka, The Australian National University

Model Genetic Systems Labs

John Carlson, Yale University, *Drosophila*
Marla Sokolowski, University of Toronto, Canada, *Drosophila*
Mario DeBono, Cambridge University, England, *C. elegans*
Larry Young, Emory University, Georgia, *Mus musculus*