

December 3, 2004

Biocomplexity Faculty Search Committee
c/o Prof. Rob de Ruyter van Steveninck
Biocomplexity Institute
Indiana University
Swain Hall West 117
Bloomington IN, 47405-7105

Dear Biocomplexity Faculty Search Committee,

I am responding to the advertisement on Nature Jobs and your website for a junior faculty position in biocomplexity. I am a postdoctoral researcher at Duke University working with Dr. Erich Jarvis in the Department of Neurobiology, where I am applying computational biology techniques to examine gene regulation in the songbird brain. I received my Ph.D. in Biology from Indiana University working with Dr. Meredith West, where I used experimental and computational methods to study social behavior of birds. I received my undergraduate degree in Biology with a Mathematics minor from the College of William and Mary.

My collaborative research approach would complement current work in the Biocomplexity Institute at Indiana University. I approach evolutionary and molecular biological questions using a combination of computational and experimental approaches. I have a strong background in biology, including molecular biology, evolution, and animal behavior. I have extensive experience in computation, including designing network inference algorithms, building genetic algorithms, and developing simulations of complex systems of behavior, brain electrophysiology, and gene regulation.

I am well prepared to develop and direct my research program in computational biology. I am adept at communicating in both biological and computational fields, as is reflected by my papers and presentations in both biological and computational forums. I have experience in raising research funds. Since my undergraduate years, I have been offered a total of eight fellowships and scholarships (see CV). In three years of postdoctoral work, I have produced four publications in the field of computational biology (two as first author) and one more first author manuscript submitted. I am adept at developing independent research efforts, as is demonstrated by the fact that for my doctorate I conducted a computation-intensive research program in a lab which had previously done no computational work. Likewise, in my postdoctoral research, I developed a novel framework for evaluating algorithms for use in a biological setting in a lab that was not previously using such approaches. I also applied algorithms in novel ways, for example, I applied network inference algorithms, commonly used for examining gene regulatory influences, to infer flow of neural signals in the brain. My work has been recognized at international conferences and is being applied by others in the U.S. and abroad.

I am eager to participate in the training of students and postdoctoral fellows in computational biology. I believe that trainees presented today with opportunities to gain a basis in both biology and computation will form a strong core for the future of research in this field. I have experience in teaching from my time as an associate instructor, as well as extensive experience with one-on-one training.

Thank you for your consideration. This document contains my research and teaching statements, and I have uploaded my CV, which contains a URL where select of my publications may be downloaded. Letters from five references, covering both computational and biological expertise, should be arriving independently. If there is any other information you require, please do not hesitate to contact me at the laboratory, (919) 681-1681, at home, (919) 452-7732, or by email, asmith@neuro.duke.edu.

Sincerely,

V. Anne Smith

Duke University Medical Center
Department of Neurobiology, Box 3209
Durham, NC 27710

Research Statement

Victoria Anne Smith

My goal is to perform research at the interface of computation and biology. I work closely with biological researchers, using computational techniques in concert with biological experimentation to elucidate biological function. I have a strong background in biology and extensive experience in computation throughout my research training. In my dissertation research, I combined experimental manipulation with computer simulation to understand social behavior of birds in their complex social environment. In my postdoctoral research, I developed a novel method for evaluating inference algorithms that purport to predict networks representing gene regulatory pathways based on microarray data. I then extended the use of these algorithms to a different field, inferring flow of neural signals based on electrophysiological data from the brain. In my future research, I plan to study the evolution of gene regulatory systems, continuing to form collaborations with both biological and computational scientists in order to produce computational methods that are applicable to real biological data and able to reveal biologically useful information.

Doctoral Research in Complex Biological Systems

In my doctoral research in the laboratory of Drs. Meredith West and Andrew King (Departments of Biology and Psychology), I struck a balance between examining social development in controlled settings—to obtain definitive evidence for factors affecting development—and in more complex settings where it naturally occurs. I performed traditional controlled experiments, where I manipulated the population origin of female companions of young male cowbirds (*Molothrus ater*) and showed that differences in social interaction affected males' vocal development. I then examined these same issues in a more complex setting. I found that a group of captive birds self-organized into an age- and sex-based pattern of association. The social experience of juvenile males over the year correlated with differences in courtship behaviors, suggesting an interaction between social interaction and development in this more complex setting.

I then took a computational approach, using computer simulation to investigate individual rules of behavior within a complex social group. I simulated the birds as modified classifier systems, where a bird used a probabilistic method to "classify" its next behavior based on its current social surroundings. I then evolved these classification rules with a genetic algorithm, optimizing the simulation's match to the observed patterns of association in the previous experiment. The evolved simulations generalized to successfully predict the interaction patterns of two new groups of real birds having different age and sex class compositions. The simulations were generalizable only for those features for which they were evolved: they matched real movement data which were related to patterns of association, such as proportion of approaches by birds to others of different age and sex classes, but not data which could vary independently of association, such as total number of flights. The rules produced by the simulations gave insight into how real birds self-organize and showed that the same rules for behavior can produce different patterns of association in different social contexts.

In addition, I collaborated with a postdoctoral researcher (Dr. David White) to use computer simulation to investigate the efficacy of different techniques he was using for sampling assortment patterns in animals. I developed a simulation to sample data from a group of assorting animals in the same way that a researcher would sample data from real animals. We found that some sampling techniques had particular biases depending on the properties of the animals. For example, group-based measures tended to overestimate differences in association if the animals formed large groups. Thus, our results provided a guide for sampling techniques to use in different circumstances.

Postdoctoral Research in Computational Biology

In my postdoctoral research in the laboratory of Dr. Erich Jarvis (Department of Neurobiology), I developed a novel method for evaluating inference algorithms that purport to predict networks representing gene regulatory pathways based on microarray data. The difficulty in evaluating such algorithms lies in the inability to perform in reasonable time the necessary gene manipulations to validate even a single network, especially in vertebrate systems. To overcome this limitation, I developed a biologically plausible simulation from which I sampled data. I then applied algorithms to this sampled data to evaluate their effectiveness at recovering the known underlying pathways. The power of this approach lies not only in the ability to evaluate algorithms, but also in the ability to evaluate the data collection techniques used to produce the data for the algorithms.

I have thus far concentrated on evaluating Bayesian network inference algorithms. I evaluated the performance of NetworkInference, an algorithm developed by Dr. Hartemink (Computer Science Department), with my simulation framework. Enabling NetworkInference to use dynamic Bayesian networks, which can take into account the ordered quality of time-series data, improved its performance. I then investigated how variations in data collection and network topology affected performance of the Bayesian network algorithm. I found that some features of network topology, such as multiple-gene control of a single gene, are more difficult to recover using biologically realistic amounts of data, while others, such as feedback loops, are relatively easy. Additionally, the rate of data sampling had a large impact on the recovery success of the algorithm, thus emphasizing that experimental design should be considered with the computational analysis in mind.

I have also applied network inference algorithms in novel ways. In collaboration with a postdoctoral researcher (Dr. Tom Smulders) and a graduate student (Jing Yu) in the Jarvis laboratory, I used a Bayesian network inference algorithm to infer information flow in the brain from electrophysiological activity. Electrophysiological recordings in the brain produce very large amounts of data, making them more amenable to functional network inference than gene expression, where data is often limited. In addition, the anatomical networks through which information flow must occur is known for some brain systems, enabling evaluation of an algorithm's performance against known biology. We applied a dynamic Bayesian network inference algorithm to recordings of the cerebral auditory areas of zebra finches (*Taeniopygia guttata*) hearing natural and synthetic stimuli. We recovered information flow only along known anatomical pathways and in patterns that matched biological expectations, giving credence to the belief that we were recovering the path of electrical signal flow in the brain. We revealed differences in information flow when the animals were presented with different auditory stimuli. With further validation, we believe that this approach will become an important tool for understanding information flow in the brain during behavior.

Experience in Collaborative Research

I am enthusiastic about forming collaborations both with experimental biologists and computational researchers to apply computational techniques to biological questions. As the above research summary indicates, I am experienced in working collaboratively. In my graduate research, I initiated a collaboration with a postdoctoral researcher in my laboratory when I saw that a question with which he was struggling, how to measure social assortment differences, could be addressed using a computer simulation. In my postdoctoral work, I am involved in a collaborative research team between an experimental neurobiologist (Jarvis) and a computational scientist (Hartemink), where I serve as the link between the two. Here, I again initiated the contact, which has resulted in a long-standing collaboration, including several joint papers and grant applications between our labs. I serve as the coordinator for my most recent collaborative project, applying network inference algorithms to brain electrophysiology, which involves a biologist now at the University of Newcastle in Britain (Smulders) and a computational student in Duke's Department of Electrical and Computer Engineering (Yu).

Future Goals: Understanding Biological Function

In my future research, I plan to further investigate gene regulatory pathways in cells. Gene regulation has broad implications across many areas of biology, including basic developmental biology, medical applications, and evolution. Understanding how and why different genes are expressed in different cells can help reveal how a body plan develops, what triggers a disease state, and why two species are different. In particular, I hypothesize that adjustments in gene regulatory networks are a major proximal mechanism for biological evolution. Mutations which adjust the shape of a network or the strength of a regulatory interaction may provide both finer and more modular substrates for selection than mutations which adjust the function of a protein. I believe both that understanding the regulatory network underlying differential gene expression is an important step for biological knowledge and that computational approaches can contribute highly to such understanding. My research approach will be to develop computational methods to reveal gene regulatory networks, validating them both with simulations and biological intervention, and then to use these tools to perform biological experiments investigating the role of gene regulation in biological evolution.

I will approach the further development of algorithms to recover gene regulatory networks from multiple theoretical directions. I propose that the application of algorithms with different theoretical bases to the

same network inference task can be considered conceptually as multiple independent estimates of the underlying network. Where estimates concur, it increases confidence of that portion of the network. Where estimates differ, the estimate most likely to be correct can be evaluated with knowledge of each algorithm's strengths. While all algorithms are likely to have weaknesses, the strengths of multiple algorithms applied to the same data may be complementary. I will identify strengths and weaknesses of multiple types of algorithms using my simulation framework. In addition to Bayesian algorithms, I will investigate linear and pair- or triple-wise network inference algorithms. Preliminary results with a triple-wise algorithm I designed indicate that it may identify subsets of variables involved in subnetworks. Thus, a possible scenario for algorithm combination is: a triple-wise algorithm identifies variables in subnetworks; a linear algorithm recovers connections within these subnetworks; and a Bayesian network algorithm double checks the subnetworks and finds connections among them.

Additionally, I believe that inference from mRNA alone may not be sufficient to gain a full understanding of gene regulation, in particular where regulation occurs at the level of protein translation rather than RNA transcription. Thus, I plan to develop computation techniques which can take this into account, using both mRNA and protein expression data. I plan to develop my tools to take into account other types of information where it is available as well, such as transcription binding sites and protein-protein interactions. I will incorporate such information into already developed algorithms, for example adding to a Bayesian network prior beliefs of suggested interactions, as well as design new types of algorithms for new information, for example designing an algorithm to use changes in gene expression from manipulation experiments.

With a suite of validated algorithms, I will apply them to perform research on network evolution. I will use a model system, such as yeast, that has a wealth of genomic information available, has short generation time, and is amenable to intervention experiments. I will recover gene regulatory networks in several closely related species and from artificial selection experiments (for example, growth in hostile media and under extreme environmental conditions). I will look for differential regulation across species or due to artificial selection, and will examine genomic DNA for evidence of mutations in areas such as transcription binding sites to account for these differences. From this beginning, I will build an integrative research program combining computational genomic sciences with evolutionary biology.

Overall, understanding biological function is the core aim of my research interest. I believe in using computation and experimentation in concert, evaluating computational tools with both computer simulation and biological intervention so that a level of confidence can be placed on results obtained. I enjoy working with other biologists to ensure that they understand how computational results are produced and the implications and limitations therein. I believe in producing computational tools that are capable of being used to reveal underlying biology in real situations, by handling the inherent noise and error present in biological experiments and by using data that are practical to collect.

Statement of Teaching Interests

Victoria Anne Smith

I am eager to participate in the training of students in computational biology. I believe that students presented today with opportunities to gain a basis in both biology and computation will form a strong core for the future of research in this field. Such training can occur at both the undergraduate and graduate levels.

I have experience in undergraduate teaching from my time as an associate instructor during my graduate studies. I assisted in a writing intensive course, where I met individually with students and helped them formulate papers over the semester. I have performed group teaching in interactive lectures to my laboratory on computational techniques and scientific writing. I also have extensive experience with one-on-one training. As a graduate student, I trained junior graduate students in procedures of behavioral observation and animal handling. Additionally, I trained a postdoctoral researcher in computer programming and simulation. As a postdoctoral researcher myself, I have participated in the training of and worked closely with a computational biology graduate student jointly advised by Dr. Jarvis and Dr. Paul Wang, a professor in Electrical and Computer Engineering. I will be supervising a first-year graduate student in a rotation project next semester. My interdisciplinary position has also given me additional teaching experience; I have taken on a teaching role to help communicate computational ideas to biologists and biological concepts to computational scientists.

In the classroom, I aim to communicate concepts clearly and allow for discussion and questions. I am particularly intent to transmit the excitement of scientific inquiry to students. I plan to include primary scientific papers as reading material for my courses. I believe this cannot be started too early—I was first introduced to scientific papers as a college sophomore and remember how much clarity this exposure gave the scientific process. In more advanced courses, I plan to assign small research projects, which, due to the relative newness of computational biology, can often be performed on unanswered questions and/or using novel approaches. This brings students in direct contact with the excitement of original research and can sometimes yield valuable new information—a priceless experience for a student.

As a mentor for graduate students, I believe training needs to be individualized to the needs of the student. I will guide students through the process of learning to become an independent scientist, involving learning the basic knowledge in the field, developing the ability to formulate good questions and design good experiments, and preparing clear presentations of one's research for the scientific community.

With my biological and computational background, I am ideally suited to teach introductory courses in biology for computational students and in computation for biological students. I would be able to teach introductory courses in statistics, bioinformatics/computational biology, and courses with intensive writing requirements. I am also qualified to develop more advanced courses with topics such as gene expression analysis, functional network inference, and complex behavioral systems.