Henry (Xiufeng) Wan
Digital Biology Laboratory
Department of Computer Science
University of Missouri-Columbia
Columbia, MO 65211

October 22, 2004

Dear Faculty Search Committee:

I am writing to apply for a tenure tracked faculty position in Biocomplexity Institute, Indiana University, posted on Nature Magazine. I am interested in the Assistant Professor level.

I entered the graduate program at the Vet School of Mississippi State University in 1998 after I earned a Bachelor's degree in Veterinary Medicine (equivalent to D.V.M) and a Master degree in Avian Medicine from China. During that era, we had been surrounding by continuously emerged advances in large-scale sequencing, genomics, and proteomics methods. Bioinformatics, as an emerging field along with these inventions, was shown its giant impact on biological research. I was so eager to enter this field but there was not many schools offering bioinformatics studying opportunities. At that time, bioinformatics was even not well defined. Under the strong support of Dr. G. T. Pharr, my Ph.D. advisor, I began my graduate study in the Department of Computer Science in 2000 along with my Ph.D. study at Vet School. Since then, I have been working in the fields of bioinformatics/computational biology.

After my graduation from MSU in 2002, I obtained more than one year postdoctoral training in the Microbial Ecology and Functional Genomics Group, Environmental Sciences Division, Oak Ridge National Laboratory. During my training in Oak Ridge, I gained much experience for microarray data analysis from head to toe (both bioinformatics and wet lab skills). It is worthwhile to mention that there were more than 20 postdocs in my previous laboratory, and that I learned how to manage research collaboration and productivity as well.

After being familiar with chip technology, I am easer to explore other aspects of bioinformatics beyond microarray data analysis. Thus, I joined Prof. Xu's group at University of Missouri-Columbia (UMC) in November 2003. Prof. Xu and I were familiar with each other since we had collaborations when Prof. Xu was research scientist in the bioinformatics group at ORNL. My current research mainly focuses on three areas: (1) computational modeling of infectious diseases (bird flu, SARS, and HIV); (2) RNA motif prediction; and (3) high-throughput analysis, including microarray and proteomics

data analysis. I have been also involved in several other projects (detailed in Research Statement), such as phylogeny analysis and tiling array data analysis.

Besides doing research works, I have also been involved in writing six different grants, four of them are related to bioinformatics method developments and applications (detailed in C.V.). Although only half of these proposals are funded, we have gained much experience in research idea assembling and proposal preparation. Moreover, I have been invited to review or co-review papers for *FEBS Letters* and various bioinformatics conferences, such as *the IEEE Computational Systems Bioinformatics conference* (2003 & 2004) and *the international conference on intelligent systems for molecular biology* (2004).

Indiana University is a prestige university, which I highly respect. The research carried out at Indiana University is very impressive. I see many opportunities for me to contribute my expertise and collaborate with researchers at your university.

Enclosed you will find my curriculum vitae, research statement, and teaching statement. Thank you very much for your consideration. I am looking forward to hearing from you soon.

Sincerely yours,

Henry (Xiufeng) Wan

RESEARCH STATEMENT

I have been working in the fields of bioinformatics/computational biology since I was a graduate student at the Department of Computer Science, Mississippi State University (MSU). During my study in the Department of Computer Science, I developed an interactive clustering model to discover gene regulation patterns in archaea (Wan *et al. ANNIE*, 12: 753-758, 2002). By using this model, I was able to discover different translation initiation and transcription initiation patterns in different archaeal species (Wan *et al. Extremophiles*, 8: 291-299, 2004). These works were in joint with Prof. Susan M. Bridges, my supervisor for my Computer Science degree, and Prof. John A. Boyle, Head of Department of Biochemistry and Molecular Biology, MSU.

After my graduation from MSU in 2002, I obtained more than one year postdoctoral training in the Microbial Ecology and Functional Genomics Group, Environmental Sciences Division, Oak Ridge National Laboratory (ORNL). My research in Oak Ridge was mainly involved in microarray data analysis, microarray database construction, and oligo design. I was also involved in proteomics data analysis, which was collaboration with Dr. R. L. Hettich at the Chemical Sciences Division, ORNL. To learn the entire procedure of microarray, I did many lab bench works from microarray printing, experimental design, sample preparation, hybridization, image scan and processing. During this training, I gained much experience related to gene expression data analysis from head to toe. I was involved mainly in 3 projects: (1) defining regulons (*fur*, *etrA*, *crp*, and *arcA*) in *Shewanella Oneidensis* MR-1 using gene expression data, which is a subproject of "Environmental Sensing, Metabolic Response and Regulatory Networks in the Respiratory Versatile Bacterium *Shewenella"* funded by *Shewanella* federation. I managed to publish fur paper (Wan *et al. J. of Bacteriology*, in press), and two more papers about *etrA*, *crp* and *arcA* regulons are in preparation. (2) oligo and primer design for microarray and phage display. I have been awarded a coauthor for one paper (Rhee *et al.*, *AEM*, **70**: 4303-4317). (3) codon usage bias analysis. We developed SCUO (Synonymous Codon Usage Orderliness) and the software *CodonO* (http://digbio.missouri.edu/~wanx/cu/codonO/) to measure synonymous codon usage bias (Wan *et al,*. *ANNIE*, 13: 1101-1018, 2003). SCUO is the first codon usage bias measurement allowing comparing codon usage across genomes. I also quantified the relationship between GC content and synonymous codon usage bias (Wan *et al.*, *BMC Evolutionary Biology*, **4**: 19). It is worthwhile to mention that there were more than 20 postdocs in my previous laboratory at ORNL, and I learned how to manage research collaboration and productivity.

My current research in Prof. Xu's laboratory focuses on four projects: (1) RNA motif identification, which is a subproject of the Genome to Life project "Carbon Sequestration in *Synechococcus* sp. From Molecular Machines to Hierarchical Modeling" from Department of Energy. I developed Rnall, a new algorithm for RNA local secondary structure prediction (Wan and Xu, *Journal of Molecular Biology*, submitted; http://digbio.missouri.edu/~wanx/Rnall). We have also developed a new strategy for intrinsic terminator prediction based on Rnall (Wan and Xu, *J. of Computer Science and*

Technology, in press). Currently, we are collaborating with Prof. Ying Xu at University of Georgia for pathway modeling using the new algorithm we developed. (2) Microarray data analysis. By collaborating with my previous laboratory at ORNL, I am developing new approaches to construct gene regulatory network using double mutant strategy. I have also been involved in tiling array data analysis and antisense gene expression characterization in *Arabidopsis*, which project is in collaboration with Dr. Curt Palm at Stanford Genome Technology Center and Dr. Gary Stacy at Plant Sciences Unit, UMC. (3) Proteomics data analysis. I assisted in developing of a new score schema and statistical model for Mass Spec fingerprinting data analyses (Ganapathy *et al., EMBS:* 3051-3054, 2004). (4) Phylogenetic analysis. I characterized several avian influenza virus strains isolated in 2003-04 Asian flu pandemic (Wan *et al., Archives of Virology*, submitted). I am also involved in a new strategy for phylogeny construction using whole genome sequences based on complete composition vectors. This project is in collaboration with a bioinformatics group in Canada. A paper related to this research has been submitted recently (Wu *et al.,* submitted). Besides the above four aspects, I did some computational research in emerging infectious diseases and remote homolog search. Two book chapters describing these works will appear (Wan *et al.,* in *Progress in Bioinformatics,* in press; Xu *et al.,* in *Handbook of Computational Molecular Biology*, in press).

In addition to the above research experience, I took part in writing six different grants, four of which were related to bioinformatics method developments and applications. Although only half of these proposals are funded, I have gained much experience in research idea assembling and proposal preparation. It is worthwhile to mention that one project related to microbial community research was funded by Department of Energy (Development and use of integrated microarray-based genomic technologies for assessing microbial community composition and dynamics, CO-PI with Dr. J. Zhou). My role was to provide bioinformatics support (microarray data analysis, oligo design, database construction) to explore microbial community diversity using oligo array. Although the money in this proposal was not transportable, this has given me an opportunity to gain experience for succeeding in assembling a big proposal. Moreover, I have been invited to review or co-review papers for *FEBS Letters* and various bioinformatics conferences, such as *the IEEE Computational Systems Bioinformatics conference* (2003 & 2004) and *the international conference on intelligent systems for molecular biology* (2004).

Based on my previous research experience and my own research interests, I would like to perform researches in the following directions.

Gene Expression Data Analysis and Regulatory Network Construction

My research interest would be mainly focused on the areas of gene expression data analysis and regulatory network construction. My vision is to bridge the gap between traditional biology research and high-throughput biomolecular data with the help of bioinformatics. While massive high-throughput data are being generated and contain rich biological information, they are difficult to analyze due to their sizes and associated

noises. Hence, traditional biology research often cannot take full advantage of these high-throughput data, such as genomic sequence, microarray and proteomics data. My previous researches have been involved in regulon characterization using microarray data and proteomics data. Part of my current research is to develop a new strategy for regulatory network construction using double mutant strategy. I am especially interested in inferring the gene regulatory network using gene deletion microarray data. New tools and computational models (especially Bayesian approaches) will be developed for these purpose. I am very also interested in identifying the cancer biomarkers using microarray and proteomics data. I will seek new collaborators at Indiana University. NSF, DOE, and NIH have various funding opportunities in quantitative and systematic approaches for biological research, which is fit for this research direction.

Identification and Functional Analyses of RNA Motifs

RNA structural motifs may function in various biological processes. For example, the hairpin-loop structure at the end of an mRNA can act as intrinsic terminators, which serves as an economic transcriptional termination machinery in bacteria. Many of conserved local secondary structures are also found in viral RNAs. Thus, predicting RNA structural motifs can shed some light on the biological mechanisms of viruses and may rapidly provide information for virus control and treatment, especially for emerging viral diseases. I am interested in identifying and functionally characterizing RNA motifs in genomic scale, such as various RNA motifs in viruses, intrinsic terminators, RNAi, and splicing sites. My current research in RNA motif identification has provided me a strong basis for this research direction. I will expand the functions of Rnall software packages for this research purposes and perform different biological applications. I would like to seek collaborators in Indiana University to validate the functions of RNA motifs that I will identify. The evolutionary relationship of these motifs across species will be my research interest as well.

Computational modeling of emerging and re-emerging infectious diseases

Influenza A viruses is a negative strand RNA virus with 8 genomic fragments (HA, NA, PA, PB1, PB2, NP, NS, and M). The genetic shift and genetic drift (reassortment between different genomic fragments) lead to a rapid emergence of novel genotypes of the avian influenza viruses during their evolution. However, there has not been an available efficient and effective approach to characterize genetic reassortments between Influenza A viruses. I am very interested in developing and then applying new quantitative methods to define and identify the genetic reassortment in the influenza A viruses. Beyond this, I would like to explore the connection between genotype, phenotype, and epidemiology, especially, in influenza A viruses. To carry out these research goals, different research steps would need to be proposed: (1) Construct the quantitative methods for new genotype identification related to gene reassortment and gene mutations; (2) Computationally model the pathogenesis and the emerging genotypes of avian influenza viruses; (3) Computationally model the relationship between the epidemiology and the emerging genotypes; (4) Incorporate geographic information system into the data analyses. The huge amount sequence and surveillance data about recent H5N1 avian influenza pandemics in Asia provide us a good resource for these

research tasks. Besides the current knowledge about influenza A viruses, I may need to seek collaborators for experiment analyses, for example, pathogenesis analysis. The collaboration will be found in Indiana University or based on my previous connection in the flu field (in HKU, e.g. Dr. Y. Guan or Dr. J. Peritis). Through this project, different bioinformatics tool in addition to a computational framework (from molecular level to population level) for emerging and re-emerging infectious diseases will be developed.

My previous research experience in avian influenza viruses (I worked in the avian influenza field since 1995 when I was a graduate student in South China Agricultural University) and my strong background at Veterinary Medicine (both clinical and experimental), Molecular Biology and Biochemistry, and Bioinformatics/Computational Biology provide me a high confidence for this ambitious research project. Due to the challenges of influenza A viruses to both the animal and public health, there are a number of potential funding opportunities from NIH (e.g. NIAID and NLM), Department of Homeland Security (DHS), USDA, and NSF, for instance, NIH (PA-04-119 and PA 03-178), Bioinformatics and Assays Development Program (BIAD) from DHS. The proposed research directions can be conducted in HIV, SARS-Cov, West Nile viruses, or other emerging infectious diseases. RNA recombination and mutation may generate new genotypes although there are not any reassortment events in these viruses.

Besides the above three directions, I am willing to collaborate with the researchers in Indiana University in any area of bioinformatics/computational biology. I believe my various training background, from sequence analysis, structural prediction, microarray, to proteomics, in the areas of bioinformatics/computational biology (as demonstrated in my publications) will provide a strong basis for these collaborations. More important, I had a very strong background in both biological science and computer science, which will smooth the communications during my collaborations with the researchers in your university.

TEACHING STATEMENT

During my graduate studies and postdoc trainings, I have been given different opportunities in the practices of teaching. As early as when I was a graduate student in China, I taught the course of Poultry Diseases for the undergraduate students. My teaching was highly evaluated by the student in the classroom. When I was graduate student in Mississippi State University, I also participated in training some junior graduate students for basic laboratory techniques in Vet School and teaching the biological concepts for the graduate students in the Department of Computer Science. Since I joined my current postdoctoral training at the Department of Computer Science, University of Missouri-Columbia, I have been given the opportunities to co-teach three courses with Prof. D. Xu, Introduction to Bioinformatics (fall of 2003), Algorithm (spring of 2004), and Computational Methods in Bioinformatics (fall of 2004, http://digbio.missouri.edu/~wanx/cs7010). My lectures have been awarded a good evaluation from the students. In addition, I am also directing three graduate students and one undergraduate student on four different bioinformatics projects. Besides the above teaching experience, I attended the campus intensive writing workshop at University of Missouri, which is designed specially to train the new faculty to teach and communicate with the students in the classroom setting. I believe all of these experiences will provide me some initial bases for my teaching career in your department.

I would be very interested in teaching two courses, including *Bioinformatics* and *Gene Expression Data Analysis*. I believe my researching experience in bioinformatics and my training in both biology and computer science will easily form the contents of these two courses.

Bioinformatics: I would like to cover different concepts of bioinformatics, such as biological sequence comparison, phylogenetic tree construction, protein structure comparison, protein structure prediction, gene finding, DNA regulatory binding motif search, bioinformatics tool development, protein-protein interaction, and so on.

***Gene Expression Data Analysis*:** I would like to cover microarray and proteomics data analysis strategy from experimental design, statistical and computational data analysis approaches towards data interpretation using different biological concepts as well as public database and data mining tools.

Besides these two courses, I think a course on “systematic approaches in biology” with the recent research advances as well as future perspective in systems biology might be interesting. These courses may have been offered at your university. In such case, I can work with other faculty members together to expand the scope of the existing courses.