DEPARTMENT OF HEALTH & HUMAN SERVICES

Public Health Service

National Institutes of Health
Bethesda, Maryland 20892

www.nih.gov

December 29, 2003

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

Re:

Faculty appointment in Computational Biology

From:
Vincent VanBuren, Ph.D.
National Institutes of Health
National Institute on Aging
Laboratory of Genetics
Developmental Genomics and Aging Section
333 Cassell Drive - Suite 3000
Baltimore, MD 21224
vincent_vanburen@nih.gov
410-558-8410 (office)
410-558-8331 (fax)

Dear Search Committee,

I am writing to be considered for an appointment as **Assistant Professor of Computational Biology** with the **Biocomplexity Institute.** A statement of career goals that also serves as an introduction to my previous and present work accompanies this letter. I believe you will see that my experiences and interests (as described in the statement and in my C.V.) are a very nice fit to the description provided for this position. For example, my experience and interests include computational approaches to fundamental genetic mechanisms, genomics oriented computational biology, computational developmental genetics/genomics, and computer simulations of the kinetics and biophysics of microtubule assembly. Furthermore, although my work now is entirely "dry lab" or computer-oriented, my graduate training in molecular biology included work at the bench and opportunities to instruct advanced laboratories, giving me the rare ability to engage both sides of the table in an effort to communicate with both computer scientists and bench scientists.

I plan to build an internationally recognized lab group where each of the members participates in some collaborative effort with other groups at the University. I look forward to hearing from you so we may begin some discussion of the potential for collaborative efforts. Please call me at 1-410-558-8410 or send an email to *vincent_vanburen@nih.gov* to arrange an interview or if something in my application requires clarification. Thank you for your kind consideration.

Best regards,

Vincent VanBuren, Ph.D.
National Institutes of Health

DEPARTMENT OF HEALTH & HUMAN SERVICES

Public Health Service

National Institutes of Health
Bethesda, Maryland 20892

www.nih.gov

Statement of Research and Teaching Interests

Vincent VanBuren
National Institutes of Health

My present position as a Postdoctoral Fellow at the National Institutes of Health has provided me with rich opportunities for collaboration with bench scientists, and I have seized those opportunities whenever it appeared that the results of collaboration would be productive. I have broad experiences in a leadership role in Bioinformatics and Computational Biology, and my work of late has focused on novel approaches towards an understanding of the relationship between measured fluorescence intensity from DNA microarray experiments and absolute transcript abundance, essential knowledge for the reconstruction and simulation of biochemical networks (see below). I would like to continue this work while organizing the infrastructure and team of individuals necessary for an internationally recognized research group in computational biology and bioinformatics. Below I will offer a description of some my research experiences and discuss some of the general principles applied to my work.

The value of computational approaches to biology may be summarized in three categories: (1) *interpretation*, or the analysis of biological data for the purpose of uncovering hidden relationships using an accepted model, (2) *prediction*, or using models to formulate computational hypotheses that may be further explored with bench experiments, and (3) *model generation*, or producing a new model by fitting known constraints on an *in silico* system in order to generate known results and new hypotheses. Interpretation and prediction using computational approaches requires some computational knowledge, including programming skills for novel approaches, and requires some knowledge of the biological model employed. Attempts at model generation using computational approaches, however, require both computational skills and a thorough understanding of the biological system of interest.

My past and current research efforts address all of these approaches. My graduate work at Lehigh University addressed the dynamic properties of microtubule assembly and disassembly. Microtubules are essential for cell division, and are the target of several anti-cancer treatments. The approach to simulation used Monte Carlo simulations in an effort to generate a new pseudo-mechanical model for the assembly, disassembly and the rates of switching between these two states, the first work by anyone that considers mechanical effects during microtubule assembly. This effort produced many fruitful results, including predictions for the lateral and longitudinal bond strength between tubulin dimers, a prediction for the mechanical energy of strain within the microtubule dimers buried in the microtubule lattice, theoretical support for a model of XMAP215's action in binding to microtubules, as well as other predictions. This first part of my work was published in PNAS and I am currently preparing a manuscript that will describe a more computationally-intensive approach at simulating microtubules that gives a full three-dimensional account for the mechanical effects during assembly. This model is able to account for many of the forms microtubules are known to take in assembly and disassembly, including ram's horns, frayed ends, sheets, extensions, and blunt ends. This 3D model of microtubule assembly and disassembly is the first of its kind, offering a comprehensive picture of both the structure and kinetics of microtubules. My collective work on microtubule simulation should provide insights into rational drug design for the treatment of cancer by providing predictions for the effects of targeting either the disruption or strengthening of lateral and longitudinal dimer associations, thereby interfering with the dynamicity of microtubules required for cell division.

In my postdoctoral training at the National Institute on Aging, I have worked on diverse bioinformatics and computational biology problems, including: leading an effort in annotation and quality control for the assembly of cDNA clones into a large collection representing libraries derived from early embryos, stem cells, and newborn organs in the mouse (published in Genome Research), development of in-house Web-based applications to speed research efforts, and participation in development of a mouse gene index that includes many genes

unique to early mouse development (submitted for publication). My current emphasis is to better understand the relationship between probe hybridization mean intensity on microarrays and the absolute abundance of the target transcript being measured. Knowledge of absolute transcript abundance is invaluable for any effort to reconstruct genetic networks. Presently, microarrays are used to measure the *relative* abundance of transcript given two collected samples, while the *absolute* abundance is obscured. At least two things contribute to clouding the relationship between probe intensity and absolute transcript abundance: probe quality and target amplification when the samples used are rare (such as preimplantation embryos). Probe quality is related to the physical properties of the probe, which include factors such as GC content, probe length, position of the probe within the target, and the energy of hairpin formation. Target amplification may skew the relationship of mean intensity and absolute abundance in a way that is dependent on the number of rounds of amplification. To both provide a useful tool and better understand how abundance might be estimated from mean intensity, I created a Web-based application that automates probe design for large-scale applications, including DNA microarrays and large-scale *in situ* hybridization (ISH). This program has different modes for each of these large-scale applications, and as there are presently much more microarray data than ISH data, the probe design tool is better optimized for microarrays. The probe scoring algorithm uses a neural network to score the complex relationship between probe quality and average mean intensity, and thus selects the optimal probe (best scoring) for a target-of-interest. This should greatly facilitate the construction of custom arrays (providing an easier design process) and generally improve DNA microarray studies by providing more probes with measured intensities significantly different from background intensity (providing more data).

By incorporating further microarray data with multiple probes against different regions of the same gene or spike-ins of known quantities of target, it should be possible to adapt the tools I constructed for probe design to estimate absolute abundance from measured mean intensity on microarrays. An estimate of absolute transcript abundance is a necessary ingredient in Monte Carlo simulations of metabolic networks. Monte Carlo methods applied to biochemical processes are simulations in which biochemical changes are treated as stochastic processes, where there is some probability that a given molecule will participate in a given reaction. An alternative to this practice, and the more standard approach, is to use differential equations to model this type of system. Differential equations are deterministic and the major caveat of using differential equations is that these methods require an assumption that generally false for biological systems: it is assumed that the process modeled by the differential equation is in equilibrium. Equilibrium is an operational description of what a cell reaches when it is dead, so Monte Carlo simulations are intuitively more appealing because they do not assume equilibrium conditions and instead simulate systems at the level of the individual molecule. The major disadvantage of Monte Carlo is that these methods are slower. Hybrid models that incorporate both differential equations and Monte Carlo methods may also be used, thus off-setting some of the disadvantages of each of these approaches (assumptions of differential methods and the speed of Monte Carlo methods). Presently, there are already many measures and estimates for the concentrations of many cellular components, and Monte Carlo-style simulations are nothing new to the field of metabolic simulations. The key idea here is that with knowledge of transcript abundance from microarray experiments, a new window will have opened revealing a large amount of information about metabolic constituents, thus presenting the opportunity to build rich models that will more accurately describe the underlying biochemical processes of the cell.

In the next 3-5 years, I am planning to (1) apply my efforts in nucleic acid probe design towards creating an application that will calculate absolute transcript abundance from measured intensities in microarray experiments, (2) use data mining techniques and Monte Carlo and differential methods to reconstruct metabolic pathways from microarray data and other high-yield data sets, including yeast two-hybrid data, (3) apply bioinformatics methods and data mining strategies towards a better understanding of the cellular mechanisms involved in RNAi, and (4) refine methods of simulating microtubules and interactions between microtubules and microtubule associated proteins (MAPs), as well as the interaction of microtubules and small molecules that interact with them. I expect be able to fully engage these research goals while exploring collaborations and teaching undergraduate and graduate courses. The involvement of undergraduates, graduates, and postdoctoral fellows, as well as collaborative involvement will all contribute to the vibrancy of this research plan. Undergraduates may contribute toward Web-based applications, including interface design, connectivity to the CGI (main program) from the interface, and towards developing efficient algorithms for analyzing probe sequences and calculating abundances. Graduate students may participate in any of the above research goals to the extent that they have an interest and aptitude for the work, but it seems likely that (1) and (3) above are the

most likely candidates for graduate students, while (2) and (4) are more suitable for advanced graduate students or postdocs (or myself).

Teaching is a great way to keep an active understanding of both relatively broad subjects (undergraduate) and advanced topics (graduate), and shapes future science by sharing insights and guiding new scientists. Presently, my strongest interests for undergraduate teaching would be courses in Introductory Bioinformatics, Introductory Genomics, Advanced Cell Biology, or Genetics. At the graduate level, my hope is that there is some flexibility in designing special topics classes. Some of my interests in this regard would be subjects like: advanced sequence analysis, computational discovery of *cis* regulatory elements, modeling biochemical systems, modeling complex biochemical systems with compartmentalization, modeling the biophysical/biochemical properties of macromolecules, reconstruction of cellular biochemical pathways from transcriptome and proteome data, large-scale biology, and the cytoskeleton. I am also capable of teaching any of the core courses of a graduate program in Molecular Biology.

The teaching interests and research described above demonstrate an aptitude and appetite for diversity. As it is research from the laboratory bench that has inspired all of my work, I am committed to building relationships and collaborations with bench scientists whenever possible (as with my graduate and postdoctoral work). In addition to cultivating collaborations with traditional molecular biologists, it has been important to my work to build relationships with specialists outside the field of molecular biology. In my previous work, I have formed close collaborations with chemical and mechanical engineers (graduate work) and a statistician (postdoctoral work), all of which yielded fruitful results. These are my greatest strengths: to conduct productive research in a leadership role with diverse approaches and to synthesize diverse and complex ideas while effectively communicating, collaborating and building relationships with experts from diverse fields of study.