From: Dmitri Papatsenko [mailto:dap5@nyu.edu]
Sent: Wednesday, December 10, 2003 7:30 PM
To: De Ruyter, Robert R.
Subject: application_papatsenko

Bioxomplexity Faculty Search Committee
c/o Prof. Robert de Ruyter
Department of Physics
727 east Third Street
Swain West 117
Indiana University
Bloomington, IN 47405-7105

Dear Committee Members,

I wish to apply for the advertised Biocomplexity Faculty Position on Department of Physics. I currently hold non-tenure track Principal Investigator position in New York University and conduct research related to gene expression and development in Drosophila. My independent research program is focused on computational analysis of transcription regulatory regions in higher eukaryotes. I have one NIH and one NSF grant proposals in submission:
http://homepages.nyu.edu/~dap5/Prsnl/Papatsenko_RO1.pdf
http://homepages.nyu.edu/~dap5/Prsnl/Papatsenko_NSF.pdf
I also maintain a database of Drosophila cis-regulatory modules, which is available from my web site: http://homepages.nyu.edu/~dap5/PCL/appendix2.htm
Below you will find a link to my CV and my Statement of Research Interests:
http://homepages.nyu.edu/~dap5/Prsnl/D.Papatsenko.pdf
The following people have agreed to write recommendation letters on my behalf:

Claude Desplan, Silver Professor
Dept. Biology, New York University
Phone: (212) 998 8218
E-mail: cd38@nyu.edu

Stephen Small, Margaret and Herman Sokol Professor
Dept. Biology, New York University
Phone: (212) 998 8244
E-mail: sjs1@nyu.edu

Mark Borodovsky, Regents' Professor
Schools of Biology and Biomedical Engineering
Georgia Institute of Technology
Phone: (404) 894 8432
E-mail: mark.borodovsky@biology.gatech.edu

Michael B. Eisen, Scientist
Lawrence Berkeley National Laboratory
Phone: (510) 486 5214
E-mail: mbeisen@lbl.gov

Eugene Nudler, Associate Professor
New York University, School of Medicine
Phone: (212) 263 7431
E-mail: nudlee01@popmail.med.nyu.edu

Thank you for considering my application
Sincerely, Dmitri A. Papatsenko
Associate Research Scientis
New York University

Statement of research interests

My research focuses on problems related to transcriptional regulation and gene expression in higher eukaryotes with the emphasis on development. I implement pattern recognition techniques to identify transcription regulatory signals in eukaryotic genomes and I use methods of molecular genetics to validate my computational findings.

At the initial stages of my career (1991-1995, Engelhardt Institute) I participated in development of DNA-protein crosslinking *in vivo*, a predecessor of the "ChIP" (chromatin immunoprecipitation) technology. Using UV-induced DNA-protein crosslinks, I identified transcription factor's binding sites in yeast rRNA genes and in some other model systems[12-14]. At the same time, I was involved in investigation of conformational stability of nucleosomes in actively transcribed genes[10].

For my postdoctoral training I went to molecular genetics field (1996-1999, Rockefeller University) and investigated transcriptional regulation of rhodopsin genes in *Drosophila*. I identified a new rhodopsin gene (*rh5*), performed computational and experimental analysis of transcriptional signals in *Drosophila* rhodopsin promoters, and generated a number of models explaining differential rhodopsin gene expression in photoreceptor cells[4, 5, 8, 9].

In 1998, I undertook my first independent computational project aimed at extraction of binding motifs for transcription factors from enhancers of *Drosophila* developmental genes. I have shown that low and high affinity binding sites for the same transcription factor form dense clusters of similar words that can be identified computationally[7]. The direct consequence of this finding was construction of 'homotypic' clustering models and recognition of *cis*-regulatory modules (CRM) in genome of *Drosophila* using a set of known binding motifs[6]. 'Homotypic' clustering models describe distribution of each binding motif in a transcription regulatory region independently. Natural promoter and enhancer regions contain binding sites for several transcription factors. I identify them using a combination of the corresponding homotypic clustering models. This approach brings together information concerning local site density and relative site affinity and allows construction of very flexible, but highly selective recognition models. An example of computational CRM recognition using 'homotypic' clustering models and experimental validation of the results is shown in **Figure 1**. My computational techniques proved to be helpful not only for genome-wide CRM recognition, but also for interpretation of gene response to gradients of developmental transcription factors[1].

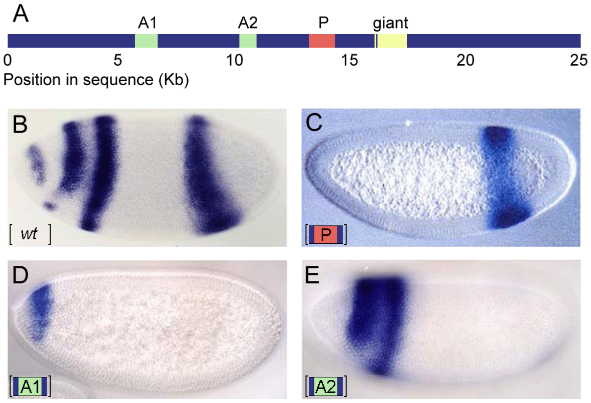

Figure 1. Recognition of CRMs in *Drosophila*
(*A*) Green bars (A1 and A2) indicate positions of clusters of Bicoid binding motif identified in the locus of *giant*. (*B-E*) Expression patterns of the identified CRM sequences in transgenic flies. (*B*) Endogeneous expression of *giant*, (*C*) Expression pattern produced by posterior enhancer of *giant* (known CRM); (*D*, *E*) expression patterns produced by the two predicted CRMs (A1 and A2) correspond to anterior expression domains of *giant*.

Binding motif 'cluster' is a simplest formal description of transcription regulatory region, but it lacks architectural information, such as site arrangement within clusters, distances between binding sites and the binding site orientation. Using signal-processing techniques I explored the presence of periodic signals in binding site distribution in the developmental genes of *Drosophila*. The most striking of the detected periodic signals (10-11 base pairs) suggest preferential positioning of some binding site combinations on the same side of the DNA helix. Based on these and related findings I proposed models for several developmental composite elements - structured combinations of several binding sites that perform similar functions in distinct transcription regulatory regions[3]. Composite elements might represent an intermediate level in hierarchy of transcriptional signals (see **Figure 2**) and their identification in developmental genes is one of my prospective goals. I also plan to incorporate the architectural information such as binding site periodicity into my current 'clustering' recognition models.

Identification of transcription regulatory regions using clustering models as well as exploration of architectural information requires *a priori* information about binding motifs for transcription factors. To overcome this limitation I invested my efforts into development of *ab initio* promoter (and CRM) recognition algorithms based on statistical analysis of word distributions in DNA sequences. Many of current promoter-recognition techniques are aimed to identify sequence segments containing *sets of words*, specifically present in transcription regulatory regions. My strategy is based on recognition of sequence segments containing *sets of word frequencies* (pattern frequency distribution analysis) specifically present in CRMs. To validate this novel approach I compared results of CRM recognition using local word frequency analysis with distribution of conserved non-coding regions in loci of developmental genes of *Drosophila*. In most cases, I obtained high correlation between the two independent methods[2] (*CC* =0.8, see **Figure 3**). My further goal is to develop a related recognition technique based on analysis of *word periodicity* in regulatory regions.

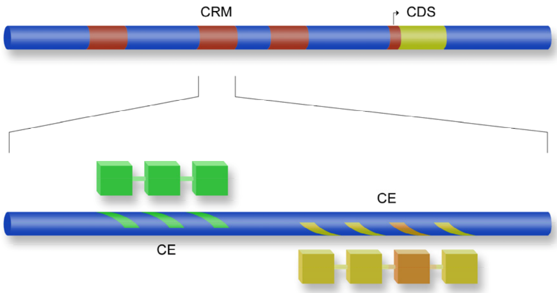

Figure 2. Hierarchy of transcriptional signals
Several periodically spaced binding sites (marks on the bottom blue bar) combine a composite element (CE). Each composite element is responsible for the formation of cognate protein-protein complex (the connected blocks). A set of *independent* composite elements bound by different protein complexes (shown by color) comprises a functional *cis*-regulatory module (CRM).

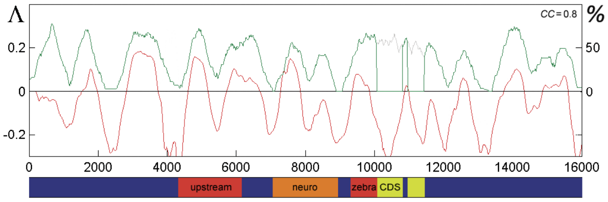

Figure 3. Word frequency analysis
Comparison of profile obtained from local word frequency analysis (red line, Λ) and conservation profile, obtained from Berkeley Drosophila Genome Project (green line, %) constructed for locus of *ftz*. Map of functional CRM regions (deletion analysis data) is given on the bottom. Red bars mark positions of known CRMs, yellow bars correspond to *ftz* coding region.

Ongoing projects and prospective goals

1 Mapping *cis*-regulatory modules and identification of new developmental genes in *Drosophila* genome (in collaboration with Dr. S. Small, NYU)

I propose to identify genes participating in the developmental pattern formation by finding CRMs containing binding motifs for transcription factors encoded by maternal, gap and some pair-rule genes. To perform this task, I take advantage of my clustering algorithms and tools as well as my training data sets from my 'Interactive CRM Database' (http://homepages.nyu.edu/~dap5/PCL/appendix2.htm). Further screening of the identified candidate genes and their *cis*-regulatory modules will be performed with the help of functional BDGP (Berkeley Drosophila Genome Project) annotations. At the final steps, I will carry out functional validation of candidate CRMs *in vivo*. Full text of this proposal is available at: http://homepages.nyu.edu/~dap5/Prsnl/Papatsenko_RO1.pdf

2 Architecture of transcriptional signals in *Drosophila* (in collaboration with Dr. M. Borodovsky, Georgia Institute of Technology)

Major goal of this project is to develop algorithms and software tools for reliable identification of CRM sequences in *Drosophila* genome using pattern frequency distribution analysis and pattern periodicity distribution analysis. The methods take into consideration the presence of *word arrangements* (word density and word periodicity) specific to *Drosophila* CRMs, rather than the presence of specific words themselves. Full text of this proposal is available at: http://homepages.nyu.edu/~dap5/Prsnl/Papatsenko_NSF.pdf

3 Identification of developmental genes and their *cis*-regulatory sequences in vertebrates

My prospective goal is to apply similar computational and experimental strategies to the exploration of developmental transcriptional signals in mammalian genomes. My specific aims will include identification of early developmental genes an their regulatory sequences in mouse and human genomes and interpretation of biological functions of these developmental genes. My experimental background will help me to involve transgenic mice in testing my computational predictions.

4 Education and teaching

Over the years I acted in supervisory capacity to several undergraduate and graduate students. My interdisciplinary approaches will allow me to involve both biology and computer students into the research process in my future independent group. My close connections with several labs in US and abroad will provide creative environment for collaborative efforts and international exchange. I also plan to develop an original graduate/advanced undergraduate course dedicated to eukaryotic transcription and structure of transcription regulatory regions.

[*] - References are numbered according to publication list in Curriculum Vitae.