Dat H. Nguyen

Department of Genetics, Harvard Medical School
77 Avenue Louis Pasteur, Boston, MA 02115
Tel: 617-432-6512 ◆ Fax : 617-432-6513
Email: dnguyen@genetics.med.harvard.edu
Web : http://arep.med.harvard.edu/~dnguyen

Education

2002	Ph.D. in Theoretical Chemistry University of California, Davis
2002	M.S. in Computer Science University of California, Davis
1996	B.S. *summa cum laude* in Chemistry University of California, Davis

Research Experience

2002-present *Alfred P. Sloan and U.S. Department of Energy Postdoctoral Fellow. Computational Systems Biology and Bioinformatics, Harvard Medical School, Boston. Advisor: Professor George M. Church*

Development of mathematical methods for deciphering regulatory principles underlying *Saccharomyces cerevisiae* transcriptional networks; development of mathematical framework for deciphering operating principles in *E. coli* metabolic networks and examining their robustness; development of computational methods, data standards, and distributed software for analyzing proteomic mass spectrometry data.

1996-2002 *Doctoral Student. Theoretical Chemistry and Biophysics, University of California, Davis. Advisor: Professor William H. Fink*

Study of principles underlying physical interactions between small molecules and RNA aptamer; study of the detailed mechanism of biological antifreezes in inhibiting the freezing of water; study of dynamical properties of water freezing process; development of efficient parallel algorithm for molecular simulation of large material and biological systems.

1999-2002 *Pre-doctoral Fellow. Computational Science, Lawrence Livermore National Laboratory. Advisor: Professor Michael E. Colvin*

Development of the linear-scaling and massively parallel algorithm for molecular dynamics simulation; study of principles underlying anti-cancer nitrogen mustard drug-DNA crosslink.

Publications

- **Nguyen, D.H**. and P. D'haeseleer. 2005. Deciphering Principles of Transcription Regulation in Eukaryotic Genomes. Provisionally accepted for publication in *Nature/EMBO Molecular Systems Biology*.
- Ahmad Q.R, **D.H. Nguyen**, M.A. Wingerd, G.M. Church, and M.A. Steffen. 2005. Molecular Weight Assessment of Proteins in Total Proteome Profiles Using 1D-PAGE and LC/MS/MS. *Proteome Science*. **3**:6.
- **Nguyen, D.H.**, K.C. Leptos, L.J. Andrews, and G.M. Church. 2005. Optimal and Extensible XML-Based File Format for Proteomic Mass Spectrometry Data. In revision.
- **Nguyen, D.H.**, M.E. Colvin, Y. Yeh, R.E. Feeney, and W.H. Fink. 2004. Intermolecular Interaction Studies of Winter Flounder Antifreeze Protein Revealing the Existence of the Thermally Accessible Binding State. *Biopolymers*. **75**:109-117.
- **Nguyen, D.H.**, T. Dieckmann, M. E. Colvin, and W.H. Fink. 2004. Dynamics Studies of a Malachite Green – RNA Complex Revealing the Origin of the Red-shift and Energetic Contributions of Stacking Interactions. *Journal of the Physical Chemistry B*. **108**:1279-1286.

Publications (cont')

- Segrè, D., J. Zucker, J. Katz, X. Lin, P. D'haeseleer, W.P. Rindone, P. Kharchenko, **D.H. Nguyen**, M. A. Wright, and G. M. Church. 2003. From Annotated Genomes to Metabolic Flux Models and Kinetic Parameter Fitting. *OMICS*. 7:301-316.
- **Nguyen, D.H.**, S.C. DeFina, W.H. Fink, and T. Dieckmann. 2002. Binding to an RNA Aptamer changes the Electron Distribution of Malachite Green. *Journal of the American Chemical Society*. **124**:15081-15084.
- **Nguyen, D.H.**, M.E. Colvin, Y. Yeh, R.E. Feeney, and W.H. Fink.2002. The Dynamics, Structure and Free Energy Profile of Proline-Containing Antifreeze Glycoprotein. *Biophysical Journal*. **82**:2892-2905.
- Yeh Y., W.H Fink, **D.H. Nguyen**, and R.E. Feeney. 1997. Simulation Studies of the Dynamic Interface between Ice and a Solution of Antifreeze Glycoprotein. *Proceedings of the International Symposium on Theoretical Biophysics and Biomathematics*. ISTB-**97**: 100-107.

Manuscripts in Preparation

- **Nguyen, D.H.** and G.M. Church. FLUXOR: An Open-Source Software Environment for Flux Balance Analysis.
- **Nguyen, D.H.**, M.E. Colvin, and W.H. Fink. The Complete Folding Landscape of a RNA Aptamer upon Packing Malachite Green Molecule in Aqueous Solution in 1 Microsecond.

Fellowships and Honors

- Postdoctoral Fellowship, Alfred P. Sloan Foundation and US Department of Energy (2002-2004)
- Institutional Pre-doctoral Fellowship, Lawrence Livermore National Laboratory (1999-2002)
- Outstanding Graduate Student Award, University of California and Clorox Chemical (1999)
- Citation for the Most Meritorious Undergraduate Achievement in Chemistry, UC-Davis (1996)
- Undergraduate Summer Research Fellowship in Nuclear Chemistry, American Chemical Society (1996)
- Willis Martin Vansell Scholarship, UC-Davis (1995)
- Vivian Bryan Nelson Scholarship, UC-Davis (1995)
- Travel Awards to:
    - "Summer School: Intelligent Extraction of Information from Graphs and High Dimensional Data," Institute for Pure and Applied Mathematics, University of California, Los Angeles (7/2005)
    - "Workshop on Proteomics: Sequence, Structure, Function," Institute for Pure and Applied Mathematics, University of California, Los Angeles (3/2004 - 4/2004)
    - "NEC Lectures on Biophysics: Biological Networks," NEC Research Institute, Princeton (6/2002)
    - "Workshop on Linear Scaling Electronic Structure Methods," Institute for Pure and Applied Mathematics, University of California, Los Angeles (4/2002)
    - "International Workshop on Methods for Macromolecular Modeling," Courant Institute of Mathematical Sciences, New York University, New York (12/2000)
    - "American Conference on Theoretical Chemistry," University of Colorado, Boulder (7/1999)

Invited Seminars

2005 Decipherable Principles of Transcription Regulation are Decipherable with Minimal Knowledge. The Bauer Center for Genomics Research, Harvard University, Cambridge, Massachusetts

Deciphering Principles of Transcription Regulation in Eukaryotic Genomes. Center for Computational Biology, University of California, Merced, California

2004 Malachite Green – RNA Aptamer Complex: the Origin of Red-Shift, the Binding Affinity, and the Free Energy Landscape. American Chemical Society 227th National Meeting, Anaheim, California.

2003 Biological Antifreezes: A Hypothesis for the Mechanism of Function. Symposium on Stress Proteins: From Antifreeze to Heat Shock, Bodega Marine Laboratory, California

Teaching/Mentoring

2003-2004 *Research Advisor, Department of Genetics, Harvard Medical School*
Supervised a master student in developing an XML-based file format for proteomic mass spectrometry data standards.

2001 *Teaching Assistant, Department of Computer Science, University of California-Davis*
Taught computer architecture and computer algorithm courses.

1996-1999
2001 *Teaching Assistant, Department of Chemistry, University of California-Davis*
Taught general chemistry, organic chemistry, physical chemistry, biophysical chemistry, and graduate level in theoretical and computational chemistry courses.

Professional Activities and Affiliations

- Elected member of the Board of Directors, Vietnamese Association for Computing, Engineering, Technology, and Science (2005)
- Referee, OMICS A Journal of Integrative Biology (2004)
- Member of Organizing Committee, West Coast Theoretical Chemistry Conference, Davis (6/2001)
- Member, Society for Industrial and Applied Mathematics (2001)
- Member, Biophysical Society (2001)
- Member, Sigma Xi-The Scientific Research Society (1996)
- Member, American Chemical Society (1996)

Personal Information

- Date of Birth: January 02, 1973
- Citizenship: USA

References

Professor William H. Fink
Department of Chemistry
University of California
One Shields Avenue
Davis, CA 95616
Phone: (530)-752-0935
Email: fink@chem.ucdavis.edu

Professor George M. Church
Department of Genetics
Harvard Medical School
77 Avenue Louis Pasteur
Boston, MA 02115
Phone: (617)-432-7266
Email: g1m1c1@receptor.med.harvard.edu

Professor Yin Yeh
Department of Applied Science
University of California
One Shields Avenue
Davis, CA 95616
Phone: (530)-752-4874
Email: yyeh@ucdavis.edu

Professor Michael E. Colvin
School of Natural Sciences
University of California
P.O. Box 2039
Merced, CA 95344
Phone: (209)-724 4364
Email: mcolvin@ucmerced.edu

Research Statement

◆◆◆

Transcription is the first step in the universal pipeline of the biological information flow from genome, where all genetic programs are stored, to proteome, through which these programs are executed. Accordingly, the regulation of transcription is critical for the development, complexity, and homeostasis of all living organisms[1,2]; its dysfunction can have far-reaching biological consequences such as developmental defects and cancer. Although transcription can be regulated at different levels[3] (e.g., chromosome structure level[4]), at the most fundamental level, which is based on the model proposed by Jacob and Monod,[5] the production of transcripts of a given gene is governed by complex combinatorial interplays of cis-regulatory elements (or motifs) present in the gene's promoter region and associated regulatory proteins (transcription factors (TFs)) present in the cell. Therefore, because TFs are gene products themselves, transcription of a gene is fundamentally regulated by the motif set present in its promoter, and the set of functions describing the dependency of motif strength, i.e., the quantitative level of motif's influence on gene expression, on its sequence/geometric context constitutes the set of principles of transcription regulation.

In spite of major efforts aimed at identifying motifs in different species using a variety of approaches and analyzing their precise influence on gene expression,[6-17] little is known about the principles in which a gene's motifs translate into an expression level. In other words, quantitative effects of motifs on gene expression as a function of their promoter context remain poorly understood. One of the obstacles to the study of such principles has been the absence of a rigorous definition of how to quantify the level of motifs' contribution to gene expression, and subsequently, necessary mathematical framework for deriving them.

With the support of the Alfred P. Sloan and U.S. DOE postdoctoral fellowship to help my career transition from theoretical chemistry and computer science to theoretical/computational molecular systems biology and bioinformatics, one of my major research focuses has been aiming at the study of principles underlying motif's quantitative behaviors as a function of promoter context. I developed the first deterministic mathematical method, Motif Expression Decomposition[18] (MED), which provides a framework for deriving principles of transcription regulation at the single gene level of resolution. The main feature of the MED method is that, unlike other metrics used to measure the effect of motif on gene expression,[6,16] MED provides a metric for quantifying both the level of each motif's influence on the expression of each gene with which it is associated, and the level of global activities of each TF under a set environmental conditions from genome-wide expression data. In addition, it operates on all genes in a genome without requiring any *a priori* knowledge of gene cluster/module membership or manual tuning of parameters. Applying MED to yeast *Saccharomyces cerevisiae* transcriptional networks, I have shown[18] that motif strength can have a complex dependence on the motif's geometry—one of the attributes of promoter context—such as distance from the translation start site (Fig. 1). My work has led to the identification of four classes of regulatory principles, all of which were validated by expression data (Fig. 1); these include length-dependent and orientation-dependent effects of motifs with respect to gene expression profiles. In addition, my work has discovered a novel mechanism demonstrating that nature has used motif geometry as the means for amplifying gene expression levels in lieu of motif-motif functional interactions.

In the new era of systems biology, the availability of genomes of many species along with the massive amount of corresponding expression data has created exciting, but enormous challenges for understanding how genomes encode and execute transcriptional programs. These challenges can largely be met by efforts targeted at identifying regulatory motifs, understanding how they quantitatively affect the production of gene's transcripts given their promoter context, and examining their combinatorial interplays with TFs, so that the dynamics underlying the transcriptional genetic circuit can be studied more accurately. To this end, my laboratory will address these challenges using the MED method to study two fundamental questions: (1) how evolution has designed and selected regulatory principles across species, from prokaryotes such as *E. coli* to eukaryotes such as yeast *S. cerevisiae* and *S. pombe*, by cataloging transcriptional regulatory principles in each species and comparing them, and (2) how TFs interpret a genome's transcriptional programs via interactions with regulatory motifs,[19-21] and how such interpretation is affected by protein post-translational modifications. Specifically, in the first objective, my laboratory will derive regulatory functions like those shown in Fig. 1 for all motifs present in the yeast *S. cerevisiae* genome as recently determined experimentally by chromatin immunoprecipitation[22], along with regulatory functions involving other attributes of promoter context such as exact motif sequence and motif-motif co-occurrence. I foresee that this kind of dataset can potentially be useful to the synthetic biology community, for whom one major goal is to construct biological systems with desired properties using *a priori* knowledge of constituent parts.[23] My laboratory will also study the evolution of regulatory principles across kingdoms and phyla using bacterial genomes such as *E. coli* and other yeast genomes such as *S. pombe* as model organisms. In addition, my laboratory will pursue the elucidation of genetic, biochemical, and biophysical mechanisms by which motif context controls transcription, for instance, by examining protein domains of TFs or structures of promoters — e.g., the level of stress-induced duplex destabilization[24] – that might be responsible for determining the geometry at which a motif has maximal effect. In the second objective, since the quantitative level of influence (or activity) of each TF on gene expression under a set of environmental conditions can also be derived from MED, my laboratory will study how the relationship between such levels of influence of TF and its concentration is affected by protein post-translational modifications, such as phosphorylations or co-factors, etc.; and how such relationship affects gene expression. Finally, on the theoretical side, my laboratory will pursue the improvement of the MED method to (a) allow the explicit treatment of motif-motif functional interactions, an important biological feature of eukaryotic genomes, (b) make it even more flexible vis-à-vis the incorporation of experimental data, and (c) develop user-friendly software for the community at large.

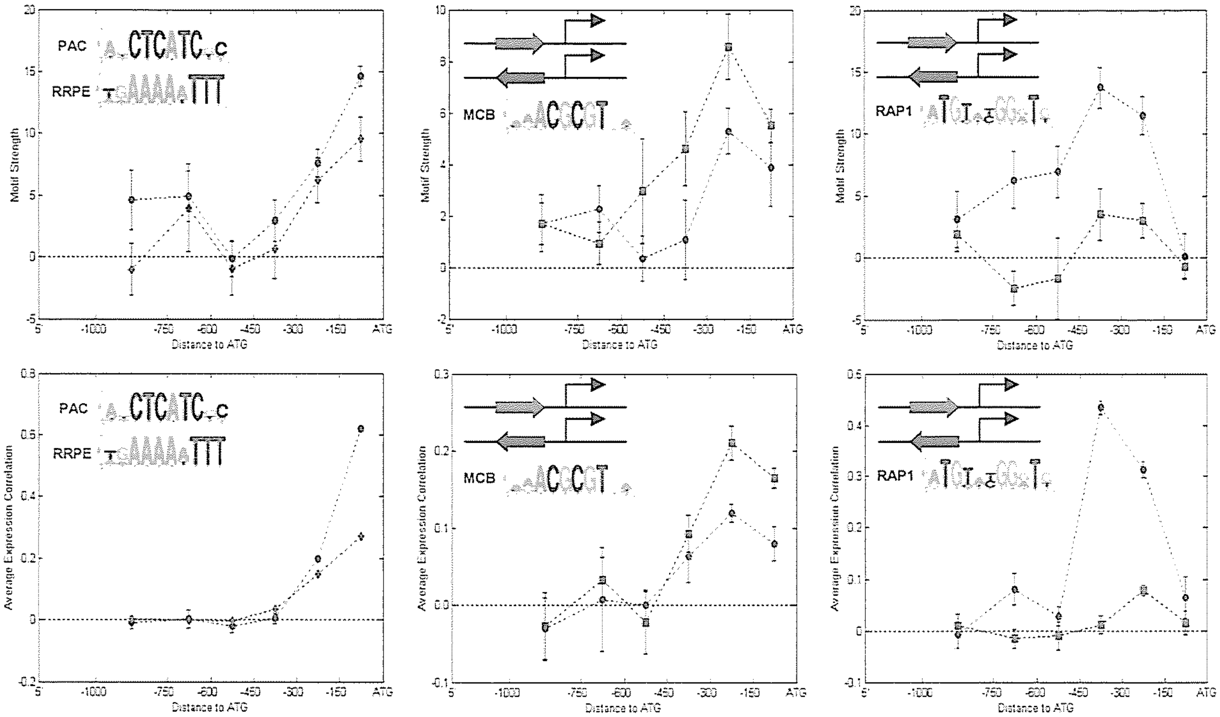

Figure 1: Illustration of four principles of transcription regulation derived from the MED method. Shown are short-range, mid-range, long range, and orientation dependent regulatory modes represented by the PAC/RRPE, MCB, and RAP1 motifs, respectively. Top panels are functions that describe the dependency of motif strength — the quantitative level of motif's influence on gene expression level — on the distance measured in the number of base pairs from the translation start site denoted as ATG. Bottom panels are the corresponding degree of gene co-expression as measured by the average pairwise expression correlation derived from experimental data, which agree well with MED's prediction. These figures are discussed in great detail in reference 18.

In summary, one of the grand challenges of 21st century biology is to quantitatively understand cellular physiology in terms of its underlying genetic regulatory networks, and to employ such knowledge to engineer new biological systems with desired novel properties. To achieve this, it is necessary to understand the operating principles determining the dynamics of transcriptional regulatory networks. This is the path that my laboratory will undertake. Using MED as a starting point, and given my interdisciplinary training in mathematics, physics, chemistry, computer science, and recently molecular biology/genetics, my laboratory is in a unique position to tackle such grand challenge of biology, which is to understand not only how genomes encode properties of organisms, but also how genetic information becomes dynamic and is translated in a regulated and coherent way using quantitative knowledge of the transcriptional process learned from this proposed research.

References

1. Davidson, E. H. *Genomic Regulatory Systems: Development and Evolution* (Academic Press, San Diego, 2001).
2. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147-151 (2003).
3. Neidhardt, F. C. & Savageau, M. A. in *Escherichia coli and Salmonella: Cellular and Molecular Biology* (ed. Neidhardt, F. C.) 1310-1324 (ASM Press, Washington, DC, 1996).
4. Hatfield, G. W. & Benham, C. J. DNA topology-mediated control of global gene expression in Escherichia coli. *Annu Rev Genet* **36**, 175-203 (2002).
5. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**, 318-56 (1961).
6. Pilpel, Y., Sudarsanam, P. & Church, G. M. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet* **29**, 153-9 (2001).
7. Sudarsanam, P., Pilpel, Y. & Church, G. M. Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in Saccharomyces cerevisiae. *Genome Res* **12**, 1723-31 (2002).
8. Siggia, E. D. Computational methods for transcriptional regulation. *Curr Opin Genet Dev* **15**, 214-21 (2005).
9. GuhaThakurta, D. et al. Identification of a novel cis-regulatory element involved in the heat shock response in Caenorhabditis elegans using microarray gene expression and computational methods. *Genome Res* **12**, 701-12 (2002).
10. Guhathakurta, D., Schriefer, L. A., Hresko, M. C., Waterston, R. H. & Stormo, G. D. Identifying muscle regulatory elements and genes in the nematode Caenorhabditis elegans. *Pac Symp Biocomput*, 425-36 (2002).
11. Xie, X. et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338-45 (2005).
12. McGuire, A. M. & Church, G. M. Predicting regulons and their cis-regulatory motifs by comparative genomics. *Nucleic Acids Research* **15**, 4523–4530 (2000).
13. McGuire, A. M., Hughes, J. D. & Church, G. M. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res* **10**, 744-57 (2000).
14. Tompa, M. et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23**, 137-44 (2005).
15. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16-23 (2000).
16. Beer, M. A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185-98 (2004).
17. Bussemaker, H. J., Li, H. & Siggia, E. D. Regulatory element detection using correlation with expression. *Nat Genet* **27**, 167-71 (2001).
18. Nguyen, D. H. & D'Haeseleer, P. Deciphering Principles of Transcription Regulation in Eukaryotic Genomes. *Provisionally accepted for publication in Nature/EMBO Molecular Systems Biology Journal* (2005).
19. Hlavacek, W. S. & Savageau, M. A. Subunit structure of regulator proteins influences the design of gene circuitry: analysis of perfectly coupled and completely uncoupled circuits. *J Mol Biol* **248**, 739-55 (1995).
20. Wall, M. E., Hlavacek, W. S. & Savageau, M. A. Design principles for regulator gene expression in a repressible gene circuit. *J Mol Biol* **332**, 861-76 (2003).
21. Savageau, M. A. *Biochemical Systems Analysis: a Study of Function and Design in Molecular Biology* (Addison-Wesley, Reading, MA, 1976).
22. Harbison, C. T. et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99-104 (2004).
23. McDaniel, R. & Weiss, R. Advances in synthetic biology: on the path from prototypes to applications. *Curr Opin Biotechnol* **16**, 476-83 (2005).
24. Benham, C. J. & Bi, C. The analysis of stress-induced duplex destabilization in long genomic DNA sequences. *J Comput Biol* **11**, 519-43 (2004).