

Yong Kong
373 Race Hill Rd
Madison, CT 06443, USA
Phone: 1-203-421-8415
Email: ykong@alum.wustl.edu

November 8, 2003

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 Professor van Steveninck,

I wish to apply for the faculty position in biocomplexity as advertised in *Nature* on October 29. Attached are copies of my curriculum vitae and statements of research and teaching interests.


I graduated from Washington University School of Medicine in St. Louis with a Ph.D. in Molecular Biophysics in 1997. Under the guidance of Prof. Jay W. Ponder, I developed the first force field with polarizable multipole electrostatics for flexible biological macromolecules. I also collaborated with Prof. Enrico Di Cera on applying statistical mechanics methods to biochemical models. My X-ray crystallography work was supervised by Prof. Gabriel Waksman.

Since receiving my doctoral degree, I've been working in the Bioinformatics Department of CuraGen Corporation. As a Project Leader in Bioinformatics in CuraGen, I led the group responsible for DNA sequence processing and analysis, including EST clustering and assembly, transcript mapping, and splice variant and regulatory motif analysis. Right now I'm working on the project to develop new algorithms and software for the next-generation whole-genome assembly program for 454 Corporation's high-throughput pyro-sequencing technology, led by Professor Eugene Myers of UC Berkeley (previously at Celera Genomics).

In the meantime, I have carried out some independent research on statistics and lattice models, resulting in several publications by myself. The results have direct applications to biological sequence analysis. One of my manuscripts has just been submitted to a premier statistics journal (the *Journal of the American Statistical Association (JASA)*) and is under formal review. Some new general statistical methods are developed in the manuscript that can be applied directly to analyze biological data. The use of the general methods is illustrated in the manuscript with an application to the proteome of bacterium *Mycoplasma genitalium*.

I look forward to hearing from you regarding my application. If you wish to discuss my educational and research background in further detail, please call me at 1-203- 421-8415, or send email to ykong@alum.wustl.edu. Thank you for your time and consideration.

Sincerely,


Yong Kong

Statement of Research Interests and Career Goal

Yong Kong
ykong@alum.wustl.edu

October 20, 2003

The human genome projects and other large scale sequencing projects are generating vast amount of information at an accelerating rate. New technologies, such as microarrays, are creating unprecedented amount of data that is transforming biology into a high throughput endeavor. The challenge we are facing now is to turn this abundant data into useful biological discovery.

My research interests focus on the development of mathematical, statistical, and computational techniques and tools for the analysis of DNA and protein sequences.

I'm looking for a tenure-track position at an inter-disciplinary academic environment. My career goal is to develop a nationally and internationally recognized research and teaching program with a focus on applying statistics and computational methods to solve biological relevant problems. I hope to work with researchers and students with different backgrounds to explore new ways of approaching important issues at the interface of biology, computer sciences, and mathematics.

1 Past Research

My past research interests and experiences lie in the following four areas: (1) lattice models and run-related statistics, (2) bioinformatics, (3) molecular dynamic simulation, and (4) X-ray crystallography.

1.1 Lattice models, run statistics, and computational biology

In the past few years, I have pursued independent research on the areas of lattice models and applications of statistical mechanical methods to biochemistry. The research not only leads to some interesting results on these areas themselves [1–4], but also yields some quite general results in run statistics, which in turn have direct applications to biological sequence analysis. Some of these statistical results, together with an application to the proteome of the bacterium *Mycoplasma genitalium*, are submitted to the *Journal of the American Statistical Association (JASA)* and is under formal review [5].

1.1.1 Lattice models and statistical mechanics applications to biochemistry

Lattice models have been widely used in physics, biophysics, and biochemistry. I used lattice models to study the ligand-binding problems. Two methods are developed, one

is the recurrence method, and the other is the generating function method.

1. Recurrence method

By using this method ([1]) I proved that the conventional transfer matrix contains more information than needed to describe the linear lattice. Its characteristic function can be factored into two parts. Only one part is needed for linear lattice. This leads to a counter-intuitive conclusion that linear models are in general simpler than the circular models, although the open ends break symmetry in the linear models.

2. Generating function method

A simple method is obtained by the combination of generating function method and matrix method to evaluate the partition functions of linear polymers ([3]). By using this method, some general formulas for run-related statistics are obtained (see below).

1.1.2 Run-related statistics

By using the generating function method mentioned above, some general formulas are developed to systematically study distributions of runs, the longest-runs, and other statistics of runs in samples from multi-letter alphabet systems. Explicit formulas for these distributions can be readily obtained with the help of the explicit expression of an important combinatorial entity, the number of ways to put r_1 red balls, r_2 blue balls, etc, into $n = r_1 + r_2 + \dots$ boxes, *without balls with the same color in adjacent boxes*. Those explicit formulas can be easily implemented in computer programs. A software package of various run-related statistics for general-purpose use with arbitrary precision arithmetic was written in C programming language. To illustrate the applications to computational biology and bioinformatics, the newly developed methods are applied to the proteome of bacterium *Mycoplasma genitalium* [5].

1.2 Bioinformatics

Since receiving my doctoral degree, I've been working in the Bioinformatics Department of CuraGen Corporation. As a Project Leader in Bioinformatics in CuraGen, I led the group responsible for DNA sequence processing and analysis. The projects in which I played a major role include EST and cDNA clustering and assembly, transcript mapping, and splice variant and regulatory motif analysis. Those projects involve algorithm development and implementation, large-scale biological database design and implementation, and constant interaction with bench scientists. For details of these projects, please refer to my *curriculum vitae*. Many of these projects involved extensive computation and massive data storage and retrieve. The experience I gained in the industry may be helpful in an academic environment.

1.3 Molecular dynamics and force field development

My thesis work involved molecular dynamic simulations and force field development. I developed the first force field with polarizable multipole electrostatics for flexible

biological macromolecules. Analytical first and second derivatives were implemented. I also developed general and efficient formulas to calculate reaction field due to off-center point multipoles [6]. A polarizable multipole water model was developed. The force field and the water model are parts of TINKER package for molecular mechanics and dynamics of macromolecules. TINKER is available at <http://dasher.wustl.edu/>.

1.4 Structural biology

Together with colleagues I solved the crystal structure of Klenow fragment of *Thermus Aquaticus* DNA polymerase I complexed with deoxyribonucleoside triphosphates [7].

2 Current Research

2.1 Run-related statistics

The results in [5] are quite general. New run-related statistics can be derived from these general results. Some of the projects I am now working on include:

2.1.1 The second longest run

Results have been obtained for the distribution of the second longest run. I am in the process of simplifying the results, coding the results into the general run-statistic package, and applying the results to biological relevant problems.

2.2 New assembler for pyro-sequencing

As a member of the team led by Professor Eugene Myers of UC Berkeley (previously at Celera Genomics), I wrote major modules of the whole genome assembly program for 454 Corporation's next generation high-throughput whole genome sequencing technology. 454 Corporation is a majority-owned subsidiary of CuraGen Corporation. The modules include OVERLAPPER, an efficient all-against-all homology search program specially designed for the pyro-sequencing technology.

3 Projected Research

On the theory and method side, I'll continue to work on the the run-related statistics. I'll embark on the following projects in the near future:

1. Finish the study on the second longest run distribution. Finish the implementation of the results in the general run-statistic program package.
2. Investigate the power of the various run-related statistics.
3. Apply the general results developed in [5] to the k -tuple statistic that is used in sequence homology studies.

On the application side, I am interested in applying the run-related statistics and other statistical, computational methods to study the randomness/non-randomness of biological sequences, to detect the genetic elements that control and regulate gene expression, and to identify alternatively spliced exons and their enhancers/silencers. A few related research areas I plan to consider include:

1. Combine the various newly developed run-related statistics with other content measures of the biological sequences, such as base composition, codon usage, oligomer frequency, autocorrelation, and open reading frame, etc., to improve the current computation tools.
2. Use these tools to detect regulatory sequences.
3. Use these tools to distinguish sequences from different species.
4. Extend the re-grouping of the natural alphabets. In [5], only two groups have been tried. Other combinations might give biological relevant insights that are previously unnoticed.
5. Apply the methods used in [5] to proteomes of other species. Apply the methods to genome sequences to detect regulatory and other biologically relevant motifs.

For longer term, I'm interested in applying these tools to high order structures of biological sequences. In particular, I'm very interested in the relation between the three-dimensional structures of proteins and their underlying genetic structures.

I'm also interested in taking some of the newly developed statistical and machine learning tools, such as Bayesian networks and support vector machines, and applying them to analyze microarray data.

For molecular dynamic simulations, I'm interested in developing new and efficient methods to calculate free energy difference and ligand binding affinity.

4 Summary

Determining the genome sequences and collecting expression data are just the first step in understanding the function of the genomes. To fully understand the complexity of biological system and their evolution history, experimental technologies have to be combined with mathematical and computational methods. It's an exciting time for inter-disciplinary research. I enjoy working closely with researchers with various backgrounds and trainings. In addition to theoretical interests on run-related statistics, lattice models, and molecular dynamic simulations, I have a wide range of research interests in applications of mathematics, statistics, and computer sciences to biology relevant problems. Developing mathematical results and implementing the results into computational tools are my strength. I am passionate with my research interests and hope to pursue my interests further.

References

- [1] Y Kong. General recurrence theory of ligand binding on three-dimensional lattice. *Journal of Chemical Physics*, 111:4790–4799, 1999.
- [2] Y Kong. Ligand binding on ladder lattices. *Biophysical Chemistry*, 81:7–21, 1999.
- [3] Y Kong. A simple method for evaluating partition functions of linear polymers. *Journal of Physical Chemistry B*, 105:10111–10114, 2001.
- [4] Y Kong. A note on the quantitative properties of McGhee-von Hippel model. *Biophysical Chemistry*, 95:1–6, 2002.
- [5] Y Kong. Distribution of runs and longest runs: A new generating function approach with applications to the proteome of *Mycoplasma genitalium*. Submitted, 2003.
- [6] Y Kong and J W Ponder. Calculation of the reaction field due to off-center point multipoles. *Journal of Chemical Physics*, 107:481–492, 1997.
- [7] Y Li, K Yong, S Korolev, and G Waksman. Crystal structures of the Klenow fragment of *Thermus Aquaticus* dna polymerase I complexed with deoxyribonucleoside triphosphates. *Protein Science*, 7:1116–1123, 1998.

Statement of Teaching Philosophy and Interests

Yong Kong
ykong@alum.wustl.edu

October 12, 2003

1 Teaching Philosophy

My teaching will be structured around the following principles, which I derive from my own learning and teaching experience:

1.1 Cultivate active learning and independent thinking

The teacher's responsibilities are not only to disseminate the existing knowledge to students, but also to teach students how to think independently. In this fast changing world, the information that is spoon-fed to students today will become obsolete quickly tomorrow. However, if students are equipped with an inquisitive mind, an independent way of thinking, and the courage to explore novel ideas and solutions, they will have the skills and experiences needed to identify and solve problems they will face in the future.

I am committed to the kind of teaching that not only gives students the "fish", but also teaches students "how to fish". I will design course projects around relevant real-world problems, and encourage students to apply the theory taught in class to these real problems. I want to help my students integrate the theories with the subject material I teach. I will encourage students to formulate the problem in different ways and to suggest multiple approaches to solving the same problem. Students learn best when they are active participants in their education. I will encourage active student participation in the classroom. I will design lectures so that a lot of "how" and "why" will be asked to spark the curiosity of the students.

1.2 Prepare course materials carefully and thoroughly

The importance of preparation in teaching should never be overstated. In preparation for the class, the teacher has to find the best way to present the materials. Each problem can be handled with different methods, and the teacher has to be familiar with the alternative solutions. When a particular method is chosen, the teacher should also be aware of and sometimes present to students the solution's advantage and disadvantage. S/he also has to link various concepts into a coherent logical framework. To give students a "big picture", the teacher has to cover the materials with a historical perspective.

For each class, I will prepare the class handouts carefully. Not only I will choose the textbook carefully, I will also do a thorough literature search on the particular topic. After the materials are gathered, I will adequately prepare and rehearse the lecture to give a clear and concise presentation.

1.3 Integrate independent research and latest development

Since the fields I am interested in teaching are evolving at a fast rate, and some of the subjects are not covered in standard textbooks, I would like to use a combination of textbook and journal articles in my classes. And the proportion of articles relative to textbooks will increase with the difficulty of the course as journal articles are often the only way to introduce students to the frontiers of our fields. I will encourage students to review the ideas in the literature critically, and propose their alternative approaches. I will also like to teach students how to do literature search, and how to follow and stay current with the advances for a particular research area.

It's the experience of a lot of people that the best way to teach students how to approach a problem is by example. I will encourage students to try to work out examples. To help students fully understand the materials covered in the class, I will integrate as much as hands-on computer simulation laboratory as possible.

1.4 Respect different learning styles

Different students have different learning styles. Some prefer to learn through reading, other prefer to learn through listening and discussion. I'll respect these different learning preferences, and try to meet the diverse requirements from students by giving out carefully prepared handouts, and engaging students into discussions in the classrooms.

1.5 Use new communication technologies

The recent advance in communication technologies, such as the Internet, makes it easier and faster to transfer information. I will utilize these technologies to communicate with students. I will put class handouts in the Internet, and will put them in formats that do not depend on proprietary software. Some of the formats I will like to use are PDF and HTML. Any student who can access the Internet can use these materials. I will also give my email to students, so that we can have fast and reliable communications outside classroom.

2 Teaching Expertise

Because my research interests and approaches are broad, I am prepared and interested to teach a wide range of courses. Some of the courses that I look forward to teaching are described below.

2.1 Computational Biology

An introductory course at senior undergraduate or graduate level, the course will cover the “classical” materials of computational biology. Topics will include:

Databases and literature access on the web; pattern matching; sequence alignment; rapid similarity search; multiple sequence alignment; secondary structure prediction; machine learning: such as support vector machines, HMM, neural networks, Bayesian network, etc.; phylogenies; sequence blocks, motifs, and profiles; gene expression data analysis and clustering methods; pharmacogenetics: SNPs and haplotypes; statistical analysis of biological sequences; new sequencing technology.

2.2 Computational Biochemistry

An introductory course at senior undergraduate or graduate level, the course will cover:

Molecular structures; statistical mechanics; molecular mechanics and molecular dynamics; classical force fields; electrostatics and solvation; parameterization; optimization; Monte Carlo method; simulated annealing; genetic algorithms; calculation of free energy differences; prediction of affinity: QSAR, CoMFA, Validate, etc.; protein structure prediction: homology modeling, fold recognition and threading, and *ab initio* protein folding.

2.3 Molecular Biophysics

The topic of this course will include:

Structure and function of protein, DNA, and RNA; energetics; linkage and cooperativity in biomolecular systems; enzyme kinetics; transition state theory; protein crystallography and NMR; protein folding.

2.4 Other courses

I can also teach any introductory computer programming (such as C, Perl, or FORTRAN), mathematics or statistics courses at the undergraduate levels.

3 Summary

Passion for learning is infectious, and helps motivate students to learn. Students are engaged by faculty who are engaged. I consider teaching to be an integral part of an academic career, and I will share my enthusiasms of knowledge and research with students.