GongXin Yu, Ph.D

1217 University City BLVD S-183. Blacksburg. VA 24060
Phone: (540) 953- 5567. E-mail: yu07_2000@yahoo.com

Research Interests

Biological Data Mining: Developing novel computational algorithms for deriving biological knowledge from large-scale biological data and apply such knowledge to solve biological problems including but not limited to *1* Inferring functions for hypothetic proteins, predicting components for host-pathogen interactions; *2* Developing novel strategies for genome comparative analysis to identify targets for therapeutics, diagnosis and drug and vaccine development; *3* Predicting protein function sites for substrate-binding specificities, regulatory and immunogenic control.

Development of Large Genome Annotation and Analysis System: The application of high-throughput whole genome sequencing technology provides increasing number of sequence data. I am interested in developing database, and robust computational systems for biological data management and analysis including procedures for automatic genome annotation and manual curation, data integration, genome comparative analysis and visualization.

Professional Experience

Senior Computational Biologist since beginning of 2005, Virginia Bioinformatics Institute, Virginia Tech.

- o Involved in a large scale of genome curation project. in which, we systematically curate bacterial and viral genomes and develop data mining methodology to identify distinguished genetic components and important virulence factors for drug and vaccine candidates to microbial pathogens of tremendous health problems to plants, animals and humans.
- o Development of analysis and annotation pipe line for Microarray analysis (Affymetrix GeneChip and cDNA array data). Analysis tools included are basic statistics procedure in BioConductor and R such as Anova, RMA, SAM, and Marray. Additional tools are BioPerl for annotation and sequence analysis.

Research Scientist in bioinformatics and data mining from 2002 to 2004 Mathematics and Computer Science Division. Oak Ridge National Laboratory, Oak Ridge. TN.

- o Developed series of novel computation algorithms to identify function-determining protein surface patches including Surface Patch Ranking (SPR) and Shuffling algorithm. decision tree-based data mining algorithm for the identification of cooperative residues that contribute to substrate specificity and other biologically important protein functional sites.
- o Created KeyGeneMiner to identify key genes for specific biochemical processes. an essential step toward understanding their genetic and biochemical basis of these processes.
- o Designed database and user-interfaces for Genome-to-life (DOE) data server.
- o Involved in building protein interaction networks based on an integrated analysis procedures including pathway template. similarity analysis. structure neighborhoods, and gene clustering information (algorithm of hierarchical clusters and self-organizing maps).

Research Fellow in bioinformatics and data mining from 2000 to 2002. Mathematics and Computer Science Division. Argonne National Laboratory, Argonne, IL.

- o Developed **TAR-GET** Database. This work provides key bioinformatics support for the structure determination of over 30 unique proteins in the Midwest Center for Structural Genomics (MCSG) at Argonne National Laboratory.
- o Devised WIT3. a new generation information system for genome annotation and metabolic reconstruction (http://www-wit.mcs.anl.gov/Gongxin/).

Research Assistant in Genetics from 1994 to 1999, Interdepartmental Genetics Program, Iowa State University, Ames. IA.

- o Studied the *Pca* resistance gene cluster and its interaction with *Puccinia coronata* f. sp. *avenae*. the causal agent of crown rust in cultivated *Avena*.

Teaching experience
- o "PATRIC system for genome annotations and curations" training students, software developers and scientists in VBI, Virginia Tech 2005.
- o "Introduction of experimental design and statistics", Huazhong Agri. University from 1986 - 1993.
- o Experimental course in Rice genetics research. Huazhong Agri. University from 1986 - 1993.

Education
- o **Ph.D.** Genetics GPA: 3.88/4.00. Iowa State University, Ames, IA. 1994 –1999.
- o **M.S.** Plant Genetics and Breeding. Huazhong Agri. Univ., Wuhan, Hubei, P. R. China 1983 –1986.
- o **B.S.** Plant Genetics and Breeding. Huazhong Agri. Univ., Wuhan, Hubei, P. R. China 1979 –1983.

Awards and Honors
- o Feature cover stories by United States Department of Agriculture as major Scientific Publications Featuring NRI-funded Research 2002 No. 8 (Paper published in Molecular Plant Microbe Interactions 14(12):1376-1383).
- o The KeyGeneMiner paper is listed as one of the 20 best papers selected among more than 200 submitted papers in 2003 IEEE bioinformatics conference.
- o Stadler Fellowship Competition (Outstanding research) awarded by Iowa State University for "Microsynteny of the *Mla* gene in barley with rice bacterial artificial chromosomes" 1996.
- o Scientific and Technical Progress Awards by Education Committee of China for research in "Resistance inheritance and gene location of rice", 1991, 1992.

Recent Publications
- o **Yu, G.X.**, Z.J. Fei, K. William and O. R. Crasta. Annotation of vaccine strain *Brucella abortus* S19 and Compared to *Brucella abortus* biovar 1 str. 9-941 and Other Sequenced *Brucella* Genomes. Accepted by Brucellosis 2005.
- o **Yu, G.X.**, A. Purkayastha, M. Hance, E.E. Snyder, M. Czar, J. Setubal and O.R. Crasta. PATRIC: PathoSystems Resource Integration Center. Proceedings of IUMS 2005
- o **Yu, G.X.**, G., Geist, A., Samatova, N.F. (2005) A Conservation-Independent Algorithm for the Identification of functionally important Surface Residue Clusters. **Protein Science** (Submitted).
- o **Yu, G.X.**, Park, H., Chandramohan, P., Munavalli, R., Ostrouchov, G., Geist, A., Samatova, N.F. (2005) In-silico Discovery of Enzyme-Substrate Specificity-Determining Residue Clusters. **Journal of molecular Biology** 352, 1105–1117
- o **Yu, G.X.**, Marland, E., Karonis, NT and Maltsev, N. (2005) A Knowledge-Based Voting Algorithm for Automated Protein Functional Annotation. **PROTEINS: Structure, Function, and Bioinformatics**, 61:907–917.
- o **Yu, G.X.** Byung-Hoon Park, Praveen Chandramohan, Al Geist, Nagiza F. Samatova 2005. A novel evolutionary scheme to identify CO_2/O_2 specificity determinant factors for ribulose 1,5-bisphosphate carboxylase/oxygenase. **Protein Engineering Design and Selection**. 2005 Dec; 18(12):589-96. Epub 2005 Oct 24.
- o **Yu, G.X.**, B.H Park, P. Chandramohan, A. Geist, N. F. Samatova 2004. In-silico Prediction of Surface Residue Clusters for Enzyme-Substrate Specificity. Proceedings of the IEEE Bioinformatics Conference, Stanford, CA, 2004.
- o S. Moy, L. Dieckman1, M. Schiffer, N. Maltsev, **G.X. Yu**, and F.R. Collart. Genome-scale expression of proteins from *Bacillus subtilis*. **Journal of Structural and Functional Genomics** 5: 103–109, 2004.
- o **Yu, G.X.** (2004) RuleMiner: a Knowledge System for Supporting High-throughput Protein Function Annotations. **Journal of Bioinformatics and Computational Biology**, Vol. 2, No. 4 (2004) 595-617
- o **Yu, G.X.**, Geist, Al, Ostrouchov, George, Samatova, Nagiza, "An SVM-based Algorithm for Identification of Photosynthesis-specific Genome Features", Proceedings of the IEEE Bioinformatics Conference, Stanford, CA, August 11-14, 2003.
- o B.-H. Park, G. Ostrouchov, **G.X. Yu**, A. Geist, A. Gorin, and N. Samatova (2003). Inference of Protein-Protein Interactions by Unlikely Profile Pair. Proceedings of the Third IEEE International Conference on Data Mining, Melbourne, Florida, USA, November 19 - 22, 2003.
- o Maltsev, N., **Yu, G.X.**, Marland, E. and et al. 2002. Sentra, a database of signal transduction proteins. **Nucleic Acids Research**. 30(1): 349-350.

- Yu, GX, Braun, E. and RP Wise. 2001. *Rds* and *Rih* mediate hypersensitive cell death independent of gene-for-gene resistance to the oat crown rust pathogen, *Puccinia coronata* f. sp. *avenae*. **Molecular Plant Microbe Interactions** 14:1376-1383.
- Yu, G.X., and Wise, R.P. 2000. An anchored AFLP and retrotransposon-based map of diploid *Avena* Genome 43:736-749 (also in Graingene database http://wheat.pw.usda.gov/ggpages/Wise_2X_Avena/).
- Yu, G.X., Bush, A.L., and Wise, R.P. 1996. Comparative mapping of homoeologous group 1 regions and genes for resistance to obligate biotrophs in *Avena*, Hordeum, and Zea mays. **Genome**, 39: 155-164.

Project Report
- G.X Yu, Joe Gabbard, Anjan Purkayastha, Mark Hance, Eric E. Snyder, Michael Czar, Joao Setubal and Oswald R. Crasta. Definition and Creation of novel algorithm for Reference Gene set for PATRIC Organism Classes. Internal Report of Virginia Tech.

Public Server development
- **WIT3**, a new generation information system for genome annotation and metabolic reconstruction (http://www-wit.mcs.anl.gov/Gongxin/).
- **SVMMER**, a SVM-based protein function annotator for the functional differentiation of highly homologous proteins (http://www-wit.mcs.anl.gov/svmmer/).

Research Proposals
- "Micro-Electro-Mechanical Systems for Rapid Enzyme Characterization." This proposal addresses needs identified in the research area "Technologies for the high-throughput synthesis of proteins and their biophysical characterization" identified in the call for proposals from the DOE Office of Science Program Announcement to DOE National Laboratories LAB 03-05, Genomes to Life.

Research funds
- Gong-Xin Yu, Nagiza F. Samatova, Al Geist, Andrey Gorin, George Ostrouchov, Byung-Hoon Park. High-throughput Biological Data Analysis and Modeling Tools for Genomes To Life Facilities. LDRD of Oak Ridge National Laboratory
- Nagiza F. Samatova, Andrey Gorin, Frank Larimer, Dong Xu, Byung-Hoon Park, Gong-Xin Yu, Pan Chongle, and David Tabb. High-throughput Biological Data Analysis and Modeling Tools for Genomes To Life Facilities. LDRD of Oak Ridge National Laboratory