Computational analysis of large-scale genome expression data

Michael Zhang

Abstract

With the advent of DNA microarray and "chip" technologies, gene expression in an organism can be monitored on a genomic scale, allowing the transcription levels of many genes to be measured simultaneously. Functional interpretation of massive expression data and linking such data to DNA sequences have become the new challenges to bioinformatics. I will us yeast cell cycle expression data analysis as an example to demonstrate how special database and computational methods may be used for extracting functional information. I will also briefly describe a novel clustering algorithm which has been applied to the cell cycle data.

Curriculum Vitae

Michael Zhang

Research Interests

The long-term goal of research in our lab is to use mathematical and statistical methods to identify functional elements in eucaryotic genomes, especially the genes and their control and regulatory elements. A genome is the program book of a life, genome research will lead to eventual decoding of the entire genetic language of life and its grammar. Driven by the Human Genome Projetc, our current interest is on two related problems: gene-finding and gene expression analysis. Since most of eukaryotic genes are split by intervening sequences (called introns), after transcription of a gene into a precursor mRNA, the introns have to be spliced out and the remaining fragments (called exons) have to be joined together as a mature mRNA before it can be translated into protein. Therefore, the key of gene-finding is to identify these exons. Constitutive coding exons are relatively easy to identify, the greatest challenge lies in the identification of end exons and alternatively spliced exons. Since this requires the study of many important control and regulatory elements for gene expression. This link between gene structure and function at the genomic level requires high-throughput functional studies. Detecting cis regulatory elements and modeling gene expression networks are becoming new challenges in the functional genomics era. Working closely with bench-scientists, our investigation will undoubtedly contribute to the understanding of genome organization as well as their control and regulation mechanisms, which will in turn have a profound impact on biology and medicine.

Publications

Zhang, M.Q. (1997). Identification of Protein Coding Regions in the Human Genome Based on Quadratic Discriminant Analysis. Proc. Natl. Acad. Sci. USA, 94:565-568.

Chen, T., and Zhang, M.Q. (1998). POMBE: A Fission Yeast Gene-finding and Exon-intron Structure Prediction System. Yeast, 14:701-710.

Zhang, M.Q.,(1998). A Discrimination Study of Human Core-promoters, in Proceedings of Pacific Symposium on Biocomputing 1998. R.B. Altman et al. eds. pp240-251, World Scientific, Singapore.

Zhang, M.Q., (1998). Statistical Features of Human Exons and Their Flanking Regions. Hum. Mol. Genet., 7:919-932.

Zhang, M.Q., (1998). Identification of Human Gene Core-promoters In Silico. Genome Res., 8:319-326.

Zhang, M.Q., (1998). Identification of Protein Coding Regions in Arabidopsis thaliana Genome Based on Quadratic Discriminant Analysis. Mol. Plant. Biol., 37:803-806.
Liu, H.-X., Zhang, M.Q. and Krainer, A.R., (1998). Identification of Functional Exonic Splicing Enhancer Motifs Recognized by Individual SR Proteins. Gene & Dev. 12:1998-2012.
Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998). Comprehensive Identification of Cell Cycle Regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. Mol. Biol. Cell. 9:3273-3297.
Zhu, J. and Zhang, M.Q., (1999) SCPD: A Promoter Database of Yeast Saccharomyces cerevisiae, Bioinformatics, 15:607-611.
Zhang, M.Q., (1999) Promoter Analysis of Co-regulated Genes in the Yeast Genome. Computers and Chemistry, 23:233-250.
Ioshikhes, I., Trifonov, E.N. and Zhang, M.Q. (1999) Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. Proc. Natl. Acad. Sci. USA, 96:2891-2895.
Tabaska, J.E. and Zhang, M.Q.,(1999) Detection of polyadenylation signals in human DNA sequences. Gene, 231:77-86.
Zhang, M.Q. (1999) Large scale gene expression data analysis: a new challenge to computational biologists, Genome Research, 9:681-688.
Machida M, Yamazaki S, Kunihiro S, Tanaka T, Kushida N, Jinnno K, Haikawa Y, Yamazaki J, Yamamoto S, Sekine M, Ogudhi A, Nagai Y, Sakai M, Aoki K, Ogura K, Kudoh Y, Kikuchi H, Zhang MQ and Yanagida M. (2000) A 38 kb segment containing cdc2 gene from the left arm of fission yeast chromosome II: sequence analysis and characterization of the genomic DNA and cDNAs encoded on the segment, Yeast, 16:71-80.
Mayeda A, Badolato J, Kobayashi R, Zhang MQ, Gardiner EM and Krainer AR. (1999) Purification and characterization of human RNPS1: a general activator of pre-mRNA splicing, J. EMBL 18:4560-4570.
Zhu J and Zhang MQ (1999). Cluster, function and promoter: analysis of yeast expression array. In Proceedings of Pacific Symposium on Biocomputing 2000. R.B. Altman et al. eds. 5:476-487.
Liu, H-X, Chew SL, Cartegni L, Zhang MQ and Krainer AR , (1999). Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. MCB, 20:1063-1071.
Getz G, Levine E, Domany E and Zhang MQ (2000). Super-paramagnetic clustering of yeast gene expression profiles. Physica A, 279:457-464.
Ioshikhes I and Zhang MQ. (2000) Large-scale human promoter mapping using CpG islands. discrimination. Nature Genetics., 26:61-63.
Davuluri R, Suzuki Y, Sugano S and Zhang MQ (2000). CART classification of human 5'UTR sequences. Genome Res., 10:1807-1816.

Stamm S, Zhu J, Nakai K, Stoilov P, Stoss O and Zhang MQ. A database and statistical analysis of aternative exons. DNA & Cell Biol. in press.
Ando S, Sarlis N, Krishnan J, Feng X, Refetoff S, Zhang MQ, Oldfield EH and Yen PM. Brief report: A berrant alternative-splicing of thyroid hormone receptor is a mechanism for thyroid hormone resistance in a thyrotropin-producing pituitary tumor. (summitted)
Liu H-X, Cartegni L, Zhang MQ and Krainer AR. Mechanism of exon skipping caused by a nonsense mutation in the BRCA1 gene. (summitted)
Zhang MQ (2000). Discriminant analysis and its application in DNA sequence motif recognition. Based on a lecture given at EBI Symposium: Genome Based Gene Structure Determination.(Jun 1-2,2000, Cambridge, UK) (To appear in Briefings in Bioinformatics)
Zhang MQ (2000). Computational methods for promoter recognition. Chapter 10 in Current Topics in Computational Biology (Jiang T, Xu Y and Zhang MQ eds). to be publsihed by MIT Press.
Tabaska JE, Davuluri R and Zhang MQ. A novel 3'-Terminal exon recognition algorithm. (Result was presented at CSHL Computational Biology Workshop 9/99 and manuscript is submitted).
Wu Q, Zhang T, Cheng J-F, Kim Y, Grimwood J, Schmulz J, Dickson M, Noonan JP, Zhang MQ, Myers RM and Maniatis T. Comparative DNA sequence analysis of mouse and human protocadherin gene clusters. (submitted)
Alexander EK, Kel-Margoulis OV, Farnham PJ, Wingender E and Zhang MQ. Computer-assisted identification of cell cycle-related genes - new targets for E2F transcription factors.(submitted)