“National research program for genomic medicine: advanced bioinformatics core: genomic statistics for complex diseases”
PI: Chun-Houh Chen (Academia Sinica)
Role: Co-principal investigator
National Science Council of Taiwan: 05/01/2008-04/30/2011

The objective of this Genomic Statistics (GS) component project is to provide effective and integrative statistics related bioinformatics services to the NRPGM disease research projects and other core facilities, and to expand the services and research accomplishments to international biomedical researchers. Over the past four years, collective efforts have been contributed to establish the “Gene Expression Study Design and Analysis Suite” (GESDAS) for microarray experiments, and extend it to “Gene-Environment Analysis Refining System” (GEARS) to facilitate more comprehensive statistical analysis for phenotypes and genotypes data. However the recent service requests made by the NRPGM core facilities and other collaborative laboratories indicated that the aforementioned efforts made by the GS Unit in fulfilling the service needs is not satisfactory. Tiling array, microRNA microarray, CHIP-on-chip array are top items among the current service request list in addition to services the GS Unit already provided. It is necessary to make efforts in preparing ourselves for providing high quality services related to these new generation of biomedical experiments in the next project period. We are also considering to open a statistics/bioinformatics consultation office for providing routine services to NRPGM members and other biomedical researchers.

Because of the complex disease nature for most of the NRPGM disease research projects and other core facilities, the whole spectrum of data to be collected in these projects involves very complicated data structure which can not be fully explored by currently available statistical or visualization procedures. We plan to introduce and develop most advanced visualization and statistical methods for quality checking, missing pattern diagnosing, and integration of these complex data sets.

Development of new statistical methods and software packages for genome-wide association studies to investigate the gene-environment interactions on phenotypes of complex diseases are among the most important objectives of this component project. Methods integrating endophenotype concept and latent class models will be developed. We will also develop multiple-marker screening algorithms using mixture regression models with visualization methods for detecting interactions within and between (endo)genotype / (endo)phenotype / (endo)environtype levels data.

Modern cancer class prediction (and disease subtyping) methods all involve 3 major strategic steps: resampling, variable selection, and prediction. We plan to present a matrix visualization based method to integrate and explore the association structure embedded in these 3 steps. In addition to the development of numerical methods for feature selection, we will also implement matrix visualization procedures for exploring the grouping and interaction structure embedded in several resampling related profile matrices.

It is crucial to develop simplified models to gain deep insights for large and complex biologic networks in system biology for the post-genomic studies. We plan to develop advanced statistical methods to analyze network structure related to human diseases. The integration of Boolean and Bayesian networks will be investigated as well.

Through the tight collaborations over the past four years, the GS Unit has established very strong relationships with all the other three component projects (FB, CB, IT) with deeper understanding and closer working experiences. The GS team plan to (1) develop an interactive statistical testing procedure for generating user-specific hypotheses for testing significance of pathways interactions with the Functional bioinformatics (FB) component by Dr. Ueng-Cheng Yang; (2) Collaborate with Comparative bioinformatics (CB) component by Dr. Chuan-Hsiung Chang using visualization and statistical methods for analyzing comparative genomics and infectious diseases data; and (3) Collaborate with Information Technology (IT) component by Dr. Chunnan Hsu in developing novel machine learning algorithms.

Through the accomplishments of all the aforementioned statistical consultation, data analysis, methodological research, and software development, the GS team should be able to assist the ABC Program Project in providing satisfactory statistical services to the NRPGM disease research projects and other core facilities