

Cheol-Goo Hur

Plant Genomics Lab, Genome Research Center, Korea Research Institute of Bioscience and Biotechnology(KRIBB)
hurlee@kribb.re.kr
<http://crop.kribb.re.kr/>

CG1222

Development of gene mining system

Grant total : \$80,000/year

Objectives

To support the researchers working with the crop related species, we aimed providing valuable information at a goal of the first step of the CFGC project by using and analyzing public plant EST sequences, microarray data, proteomic data, and protein interaction data. As a result, we developed the softwares for the microarray analysis and the peptide mass fingerprinting (PMF). We also constructed the systems for the predicted regulatory motifs of co-expressed genes from the results of microarray experiments, plant function catalogs, and EST analyses of crop-related species and nine representative plant species. In addition, we constructed the integrated systems on the basis of gene function catalog

linked with microarray and PMF data reciprocally. All of our results are freely available at <http://crop.kribb.re.kr> as a global integrated web interface.

Background

EST analyses Function catalogs

The gene analysis based on ESTs involves clustering and assembly of EST sequences, chromosomal mapping of consensus sequences obtained from clustering and assembly, and finally gene annotation process. For the EST clustering process, we used StackPACK software from SANBI to make virtual mRNA candidate with high coverage and high quality. To analyze these sequence data further, these process are needed as follow: homology search of consensus sequences using BLASTX against NCBI NR



database, protein function categorization according to MIPS (Munich Information) and Gene Ontology(GO) catalog, chromosomal mapping using sim4 tool and the upstream sequence analysis obtained from genomic mapping. We applied the public softwares such as Gibbs sampling, Multiple EM for Motif Elicitation (MEME), and TRANSFAC to our research for promoter analysis.

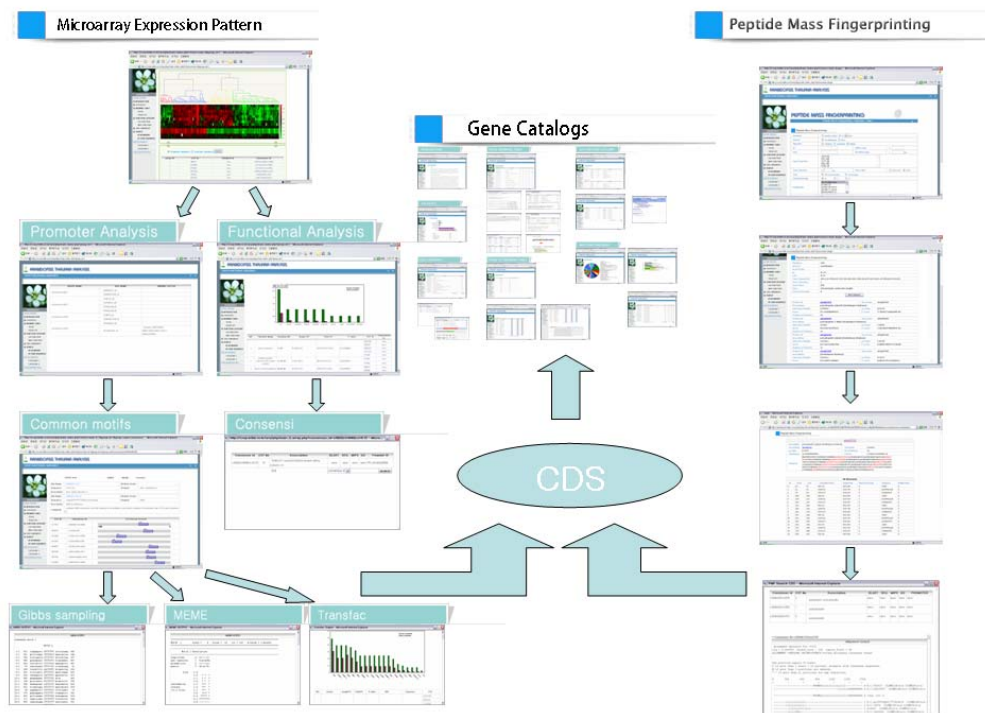
DNA chip analysis

In the case of clustering analysis of microarray study, validation is one of important data processing step. The notion is generally accepted that co-expressed genes are under the same transcriptional control and are probably categorized into same or similar functional group biologically or biochemically. Thus, the efforts to find common regulatory motifs in their promoter region, or functional grouping the genes in the same cluster are common process for the interpretation and validation of microarray data.

Peptide mass fingerprinting

Public protein sequence database such as SWISS-PROT is used practically for the protein identification from the result of Matrix-Assisted Laser Desorption Ionization-Time Of Flight (MALDI-TOF) data, which is one of popular proteomic studies. However, for the less of protein information for the specific plant species in these databases it is needed to construct the private protein database containing sufficient protein information for interpreting massive PMF results about each specific plant species. Thus we tried to make the protein database by translating enormous coding region sequences obtained from EST analysis and the PMF software working on these databases.

Therefore, in this study, we tried to make the individual systems about EST based data analysis, regulatory motif information from chromosomal mapping of ESTs, microarray data and PMF information from bench works at first and finally integrate these individual



information modules as a whole. That is,

we constructed the web-based integrated systems in which the results from gene analysis, microarray analysis, and proteomic analysis are reciprocally connected and complemented with each other.

Conclusion

We developed the integrated systems based on the gene catalogs from EST analysis results which are reciprocally connected with microarray analysis data and PMF data. Gene function catalogs help making customized cDNA microarray and the output of microarray experiments are directly analyzed on these system. In addition, the data from MALDI-TOF can be easily explained in our systems in an interactive mode with EST based information, microarray based information and vice versa.

Further Research Plan

In the second step, on the basis of the results from the first step, i. e. information from EST analysis, microarray analysis, the prediction systems for plant gene functions, proteomic analysis, and protein-protein interaction analysis, we'll perform the research about plant stress/resistance mechanisms massively by collecting, processing, and analyzing plant stress-related and resistance(R) genes and hopefully provide the specialized information such as R gene-specific motifs, type III secretion information, pathogene-related microarray analysis, protein-protein interaction and subcellular localization of plant stress-related genes.

Acknowledgments : *especially for collaborators*

Publications *CFGC acknowledged only in the order of SCI Journal to Non-SCI*

Lee S, SY Kim, E Chung, YH Joung, HS Pai, **Cheol-Hoo Hur**, D Choi, 2004. EST and microarray analyses of

pathogen-responsive genes in hot pepper(Capsicum annum L.) non-host resistance against soybean pustle pathogen(Xanthomonas axonopodis pv, glycines), Functional & Intergrative Genomics (Published online at Feb. 4th 2004)

Seong-Eui Hong, **Cheol-Goo Hur***, Sung Hoon Lee, Hae Young Chung, "Improved algorithms for the identification of yeast proteins and significant transcription factor and motif analysis", Proteomics, 2004. *Submitted.*

Tae-Hoon Chung, Sunyoung Park, **Cheol-Goo Hur**, Yangsuk Kim*, "Comparative assessment of differentially expressed gene identification algorithms in DNA microarray systems", Nucleic Acids Research, 2004. *Submitted.*

Tae-Hoon Chung, Sunyoung Park, **Cheol-Goo Hur***, "An open source cDNA microarray data analysis system with GUI: Quint", Genome Biology, 2003. *Submitted.*

Patents(program S/W)

Peptide Mass Fingerprinting Database Management program Using AMWISE and fBIND technique, Registration Number: 2004-01-12-835
Peptide Mass Fingerprinting program Using AMWISE and fBIND technique, Registration Number : 2004-01-12-836
cDNA Microarray data Classification tool, Registration Number : 2004-01-22-839
cDNA Microarray data Clustering tool, Registration Number : 2004-01-22-840

Technology transfer or licensing

Hur, **Cheol-Goo** et al. 2004. peptide mass spectrum analysis S/W and Database management program by AMWISE and fBIND technique, License to WithUSTech Company in Korea at \$25,000.