An Automated System for EST Functional Analysis and Its Application in Rice Genes

Hsiu-Ying Lu (呂秀英)[13], Cheng-Tao Chen (陳政道)[1], Yi-Chia Chiu (邱怡嘉)[2], Chun-Tang Lu (呂椿棠)[1], Meng-Li Wei (魏夢麗)[1]

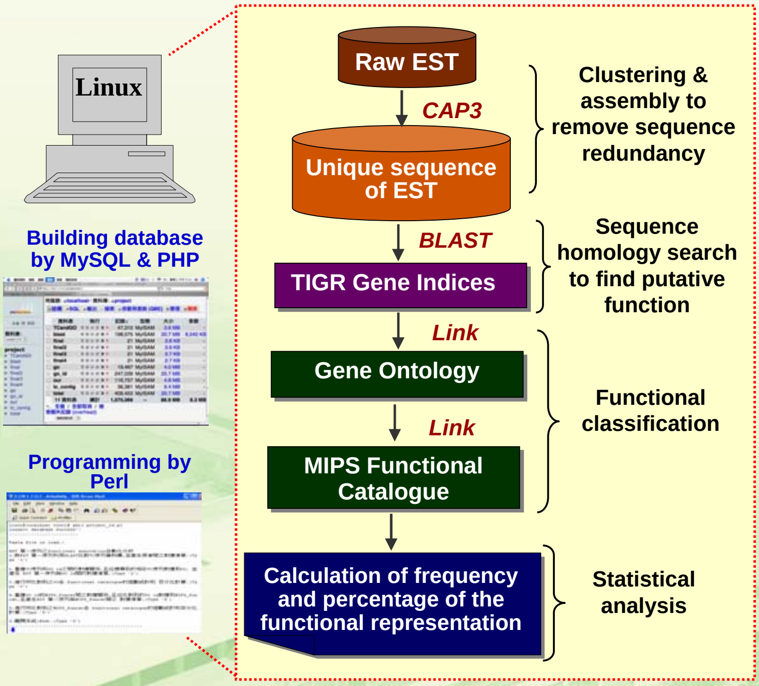
[1]Agricultural Research Institute, Taiwan, ROC.
[2]Graduate Institute of Biotechnology, Chaoyang University of Technology, Taiwan, ROC.
[3]Correspondence author (iying@wufeng.tari.gov.tw)

Motivation

Expressed sequence tag (EST), the partial fragment of cDNA sequence, represents the most extensive available survey of the transcribed portion of the genome. Rice (*Oryza sativa* L.) is the world's most important food crop; it is also a good model for studies of monocot plants. A vast number of rice EST sequences in the public databases provide an important resource for functional annotation of rice genome. However, ESTs are generated and deposited in the public domain, as redundant, un-annotated, single-pass reactions, with virtually no biological content. While efforts to assemble, organize and annotate raw EST sequence data have been developed employing diverse strategies, they typically read single input files, produce single output files and require extensive manual intervention by the user. Thus, there is a need for developing an integrated and fully automated program to perform the assembly, annotation, functional classification and statistical analysis of large quantities of EST sequences.

Material and Methods

- The complete procedure for EST functional analysis was established as (Figure 1): generation of a non-redundant data set (unique sequences) after clustering and assembling the raw ESTs (by CAP3 software), functional annotation of unique sequences (by BLAST similarity search to public annotated TC database in TIGR), functional classification (by linking to GO and MIPS catalog systems from tentative annotation of the selected TC sequences), and statistical analysis of gene functional representation.
- We firstly downloaded all the sequences of TC database, CAP3 and BLAST softwares, as well as GO and MIPS catalogs. Then, TC-sequence to function-catalog relations were established and a MySQL database was created. All programs of this system were written in Perl script and developed on the Linux platform. To store and manage sequence data and analytical results, PhpMyAdmin package was used to connect the user's web browser and MySQL database.
- Brown planthopper (*Nilaparvata lugens* (Stal.)) is the major pest in rice so is the rice blast disease (infected by *Magnaporthe grisea*) the most serious problem. The automated system was implemented in the rice ESTs induced by *N. lugens* (with the total amount as 188) and *M. grisea* (with the total amount as 84,705) from NCBI-dbEST database. We spent 3 months finishing the functional analysis of 188 ESTs induced by *N. lugens* using manual intervention. However, it's impossible to accomplish the entire process of the huge amount of ESTs induced by *M. grisea* within one year without automated system.



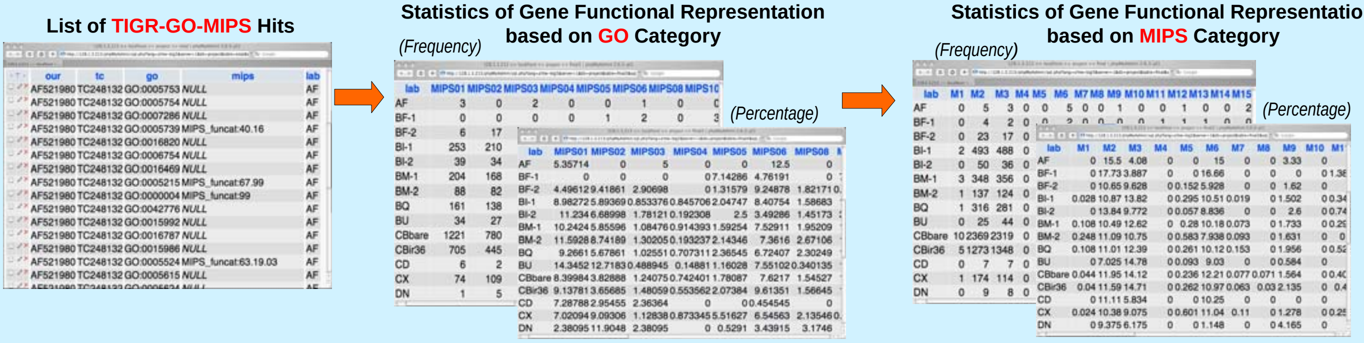
Linux
Building database by MySQL & PHP
Programming by Perl
Raw EST
CAP3
Unique sequence of EST
BLAST
TIGR Gene Indices
Link
Gene Ontology
Link
MIPS Functional Catalogue
Calculation of frequency and percentage of the functional representation
Clustering & assembly to remove sequence redundancy
Sequence homology search to find putative function
Functional classification
Statistical analysis

Figure 1. System Structure and Pipeline Strategy of automated system

Results and Discussion

- This automated system considerably reduced analysis time from 3 months to 2 hours for the 188 rice ESTs induced by *N. lugens*. It also allows us to finish the fully functional analysis of 84,705 rice ESTs induced by *M. grisea* within 7 days.
- The output results contain 5 kinds of records and can be saved to an EXCEL file (an example shown in Figure 2).
- The findings of rice EST annotation help clarify the functional mechanisms of rice genes resistant to pest and disease.

Figure 2. Output Results of Functional Analysis for Rice ESTs induced by *M. grisea*



List of TIGR-GO-MIPS Hits

our	tc	go	mips	lab
AF521980	TC248132	GO:0005753	*NULL*	AF
AF521980	TC248132	GO:0005754	*NULL*	AF
AF521980	TC248132	GO:0007286	*NULL*	AF
AF521980	TC248132	GO:0005739	MIPS_funcat:40.16	AF
AF521980	TC248132	GO:0016820	*NULL*	AF
AF521980	TC248132	GO:0006754	*NULL*	AF
AF521980	TC248132	GO:0016469	*NULL*	AF
AF521980	TC248132	GO:0005215	MIPS_funcat:67.99	AF
AF521980	TC248132	GO:0000004	MIPS_funcat:99	AF
AF521980	TC248132	GO:0042776	*NULL*	AF
AF521980	TC248132	GO:0015992	*NULL*	AF
AF521980	TC248132	GO:0016787	*NULL*	AF
AF521980	TC248132	GO:0015986	*NULL*	AF
AF521980	TC248132	GO:0005524	MIPS_funcat:63.19.03	AF
AF521980	TC248132	GO:0005615	*NULL*	AF

Statistics of Gene Functional Representation based on GO Category

(Frequency)

lab	MIPS01	MIPS02	MIPS03	MIPS04	MIPS05	MIPS06	MIPS08	MIPS10
AF	3	0	2	0	0	1	0	0
BF-1	0	0	0	0	1	2	0	3
BF-2	6	17						
BI-1	253	210						
BI-2	39	34						
BM-1	204	168						
BM-2	88	82						
BQ	161	138						
BU	34	27						
CBbare	1221	780						
CBir36	705	445						
CD	6	2						
CX	74	109						
DN	1	5						

(Percentage)

lab	MIPS01	MIPS02	MIPS03	MIPS04	MIPS05	MIPS06	MIPS08
AF	5.35714	0	5	0	0	12.5	0
BF-1	0	0	0	0	7.14286	4.76191	0
BF-2	4.49612	9.41861	2.90698	0	1.31579	9.24878	1.82171
BI-1	8.98272	5.89369	0.853376	0.845706	2.04747	8.40754	1.58683
BI-2	11.234	6.68998	1.78121	0.192308	2.5	3.49286	1.45173
BM-1	10.2424	5.85596	1.08476	0.914393	1.59254	7.52911	1.95209
BM-2	11.5928	8.74189	1.30205	0.193237	2.14346	7.3616	2.67106
BQ	9.2661	5.67861	1.02551	0.707311	2.36545	6.72407	2.30249
BU	14.3452	12.7183	0.488945	0.14881	1.16028	7.55102	0.340135
CBbare	8.39984	3.82888	1.24075	0.742401	1.78087	7.6217	1.54527
CBir36	9.13781	3.65685	1.48059	0.553562	2.07384	9.61351	1.56645
CD	7.28788	2.95455	2.36364	0	0	0.454545	0
CX	7.02094	9.09306	1.12838	0.873345	5.51627	6.54563	2.13546
DN	2.38095	11.9048	2.38095	0	0.5291	3.43915	3.1746

Statistics of Gene Functional Representation based on MIPS Category

(Frequency)

lab	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15
AF	0	5	3	0	0	5	0	0	1	0	0	1	0	0	2
BF-1	0	4	2	0	0	2	0	0	0	0	1	1	1	0	0
BF-2	0	23	17	0											
BI-1	2	493	488	0											
BI-2	0	50	36	0											
BM-1	3	348	356	0											
BM-2	1	137	124	0											
BQ	1	316	281	0											
BU	0	25	44	0											
CBbare	10	2369	2319	0											
CBir36	5	1273	1348	0											
CD	0	7	7	0											
CX	1	174	114	0											
DN	0	9	8	0											

(Percentage)

lab	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
AF	0	15.5	4.08	0	0	15	0	0	3.33	0
BF-1	0	17.73	3.887	0	0	16.66	0	0	0	0
BF-2	0	10.65	9.628	0	0.152	5.928	0	0	1.62	0
BI-1	0.028	10.87	13.82	0	0.295	10.51	0.019	0	1.502	0
BI-2	0	13.84	9.772	0	0.057	8.836	0	0	2.6	0
BM-1	0.108	10.49	12.62	0	0.28	10.18	0.073	0	1.733	0
BM-2	0.248	11.09	10.75	0	0.583	7.938	0.093	0	1.631	0
BQ	0.108	11.01	12.39	0	0.261	10.12	0.153	0	1.956	0
BU	0	7.025	14.78	0	0.093	9.03	0	0	0.584	0
CBbare	0.044	11.95	14.12	0	0.236	12.21	0.077	0.071	1.564	0
CBir36	0.04	11.59	14.71	0	0.262	10.97	0.063	0.03	2.135	0
CD	0	11.11	5.834	0	0	10.25	0	0	0	0
CX	0.024	10.38	9.075	0	0.601	11.04	0.11	0	1.278	0
DN	0	9.375	6.175	0	0	1.148	0	0	4.165	0

Acknowledgments

This work was supported by the National Science Council, Taiwan, ROC with a grant no. NSC92-2313-B-055-006, 93-2313-B-055-002.

行政院農業委員會農業試驗所
Agricultural Research Institute, COA