Plant Biotechnology 28, 351–360 (2011)
DOI: 10.5511/plantbiotechnology.11.0818a

Invited Review

Plant expressed sequence tags databases: practical uses and the improvement of their searches using network module analysis

Yoshiyuki Ogata[1,*], Hideyuki Suzuki[2]

[1] RIKEN Plant Science Center, Yokohama, Kanagawa 230-0045, Japan; [2]Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan
* E-mail: ogata@psc.riken.jp Tel & Fax: +81-45-503-9490

Received July 14, 2011; accepted August 18, 2011 (Edited by T. Demura)

Abstract Sequencing technology has been rapidly advancing. Giga-sequencers, which produce several gigabases of fragmented sequences per run, are attractive for decoding genomes and expressed sequence tags (ESTs). A variety of plant genomes and ESTs have been sequenced since the decoding of the genome of *Arabidopsis thaliana*, the model plant. ESTs are useful for functional analyses of genes and proteins and as biomarkers, which are used to identify particular tissues and conditions due to the specificity of their expression. Sequenced plant genomes and ESTs have been entered into public databases, where they are freely downloadable. Sequences representative of particular functions or structures have been collected from public databases to curate smaller databases useful for studying protein function. Here, we discuss the uses of the currently available plant EST datasets. We also demonstrate the use of network module analysis to perform more stable (or irrespective of the difference of performance in each analyzing PC) homology searches and to provide more information on molecular functions of plant ESTs and proteins.

Key words: Database, expressed sequence tag (EST), homology search, network module analysis.

Advances in sequencing techniques have promoted the expansion of DNA and RNA sequence datasets. In particular, the advent of next-generation sequencers (sometimes called "giga-sequencers"), which produce several gigabases of fragmented DNA or RNA sequence data per run, have succeeded in simplifying an international project for decoding a genome down to a laboratory task. Plant genome decoding projects include The *Arabidopsis* Genome Initiative (2000), an international project undertaken by 8 countries to decode the genome of *Arabidopsis thaliana*, and, more recently, the sequencing of the draft genome of *Jatropha curcas* by a research group working out several institutions (Sato et al. 2011). By adding datasets obtained using such high-throughput sequencers, genomes of many plants have been published; reviewed by Paterson et al. (2010). In addition, EST datasets have been published for hundreds of plant species, subspecies, and cultivars, before and in parallel with these genome decoding; reports on plant EST datasets of every family and genus are listed in Table 1 and Supplementary table, respectively, and reviewed (Batley et al. 2003; Fedorova et al. 2002; Hofte et al. 1993; Keith et al. 1993; Michalek et al. 2002; Newman et al. 1994; Park et al. 1993; Rounsley et al. 1996; Rudd 2003; Sasaki et al. 1994; Shoemaker et al. 2002; Uchimiya et al. 1992; Yamamoto and Sasaki 1997; Yuan et al. 2011).

EST data are useful for clarifying structural gene annotation, which can be applied in the functional genomics (Yonekura-Sakakibara and Saito 2009) and to make molecular markers (Kalia et al. 2011; Parida et al. 2009). Figure 1 shows a flowchart of the uses and user-friendly tools for plant EST datasets for plant scientists. Sequences stored in public databases such as NCBI (http://www.ncbi.nlm.nih.gov/), DDBJ (http://www.ddbj.nig.ac.jp/index-e.html), and EMBL-EBI (http://www.ebi.ac.uk/), have increased as shown in the top-left chart of Figure 1, obtained from DDBJ. These databases include EST datasets from all available species, including plants, animals, and microorganisms. PlantGDB (Duvick et al. 2008) contains plant ESTs, selected by plant sciences experts from such databases. Plant researchers can thus perform homology searches for query plant sequences using the selected datasets that include only plant ESTs, leading to a more stable (or irrespective of the difference of performance in each analyzing PC) search performance.

Plant EST datasets have also increased in number. To

Abbreviations: BLAST, Basic Local Alignment Search Tool; EST, expressed sequence tags; GBFF, GenBank flatfile; GFF, General Feature Format; GPFF, GenPept flatfime.

This article can be found at http://www.jspcmb.jp/
Published online September 25, 2011

Table 1. A summary of plant EST datasets obtained from PlantGDB

Family name	Genus	Set	Sequence	CYP	GT	Representative*
Acanthaceae	3	3	2562	18	28	*Avicennia marina*
Acoraceae	1	1	9695	17	177	*Acorus americanus*
Actinidiaceae	2	9	163133	812	2223	*Actinidia deliciosa*
Aizoaceae	2	2	27386	57	209	*Mesembryanthemum crystallinum*
Alstroemeriaceae	1	1	2724	15	30	*Alstroemeria peruviana*
Amaranthaceae	11	21	38461	119	428	*Beta vulgaris*
Amaryllidaceae	4	5	28786	148	471	*Allium cepa*
Amborellaceae	1	1	26378	93	329	*Amborella trichopoda*
Anacardiaceae	2	2	1415	16	12	*Pistacia vera*
Apiaceae	4	6	2614	6	28	*Apium graveolens*
Apocynaceae	3	3	36334	273	613	*Catharanthus roseus*
Araceae	5	6	4500	28	21	*Zantedeschia aethiopica*
Araliaceae	3	6	17150	165	234	*Panax ginseng*
Araucariaceae	1	1	10	0	0	*Araucaria angustifolia*
Arecaceae	3	4	42630	145	361	*Elaeis guineensis*
Aristolochiaceae	3	3	27344	133	377	*Aristolochia fimbriata*
Asparagaceae	7	7	11614	43	109	*Asparagus officinalis*
Asteraceae	28	49	1081029	7341	14211	*Helianthus annuus*
Aulacomniaceae	1	1	439	0	6	*Aulacomnium turgidum*
Azollaceae	1	1	6	0	0	*Azolla caroliniana*
Berberidaceae	2	2	1137	27	11	*Sinopodophyllum hexandrum*
Betulaceae	1	2	5688	6	12	*Betula platyphylla*
Bixaceae	2	2	962	15	3	*Bixa orellana*
Boraginaceae	2	2	909	41	0	*Arnebia euchroma*
Botryococcaceae	1	1	85586	36	56	*Botryococcus braunii*
Brassicaceae	10	33	2898795	10377	21736	*Arabidopsis thaliana*
Bromeliaceae	2	2	5660	16	61	*Ananas comosus*
Cabombaceae	1	1	3097	13	39	*Cabomba aquatica*
Cactaceae	1	1	122	0	3	*Opuntia streptacantha*
Calycanthaceae	1	1	867	1	3	*Chimonanthus praecox*
Campanulaceae	1	1	870	8	3	*Codonopsis lanceolata*
Cannabaceae	2	2	29221	118	156	*Humulus lupulus*
Caricaceae	1	1	77393	278	901	*Carica papaya*
Caryocaraceae	1	1	958	0	0	*Caryocar brasiliense*
Caryophyllaceae	3	4	4547	11	26	*Silene latifolia*
Casuarinaceae	1	2	2081	19	17	*Casuarina glauca*
Celastraceae	1	1	51380	302	554	*Euonymus alatus*
Chlamydomonadaceae	2	6	218439	59	448	*Chlamydomonas reinhardtii*
Chlorellaceae	2	3	30147	46	209	*Chlorella variabilis*
Chlorodendraceae	1	1	1103	1	6	*Scherffelia dubia*
Cistaceae	1	2	2048	3	4	*Cistus creticus subsp. creticus*
Cleomaceae	3	3	36235	86	221	*Coccomyxa sp. C-169*
Clusiaceae	1	1	149	0	0	*Garcinia mangostana*
Colchicaceae	1	1	14	0	0	*Gloriosa superba*
Coleochaetaceae	1	3	9813	12	53	*Coleochaete scutata*
Combretaceae	1	1	9	0	0	*Terminalia arjuna*
Convolvulaceae	1	3	87095	674	1261	*Ipomoea nil*
Crassulaceae	2	2	350	0	6	*Kalanchoe x houghtonii*
Cucurbitaceae	4	11	151678	423	1152	*Cucumis melo*
Cupressaceae	4	5	66378	659	665	*Cryptomeria japonica*
Cycadaceae	1	1	21997	98	258	*Cycas rumphii*
Dennstaedtiaceae	1	1	424	1	2	*Pteridium aquilinum*
Desmidiaceae	1	1	25	0	0	*Micrasterias denticulata*
Dioscoreaceae	1	2	44165	193	537	*Dioscorea alata*
Ditrichaceae	1	1	1677	1	4	*Ceratodon purpureus*
Dunaliellaceae	1	2	4139	2	5	*Dunaliella salina*
Ebenaceae	1	1	9474	47	119	*Diospyros kaki*
Ericaceae	2	4	6560	28	117	*Vaccinium corymbosum*
Euphorbiaceae	6	9	257913	1377	2845	*Manihot esculenta*
Fabaceae	34	64	3178051	14592	23289	*Glycine max*
Fagaceae	4	11	194326	1047	2744	*Quercus robur*
Funariaceae	1	2	382587	1412	2343	*Physcomitrella patens subsp. patens*
Gentianaceae	1	1	647	34	9	*Eustoma exaltatum subsp. russellianum*
Geraniaceae	1	1	27	0	0	*Geranium dissectum*
Gesneriaceae	3	5	56	1	0	*Haberlea rhodopensis*

Table 1. (Continnue)

Family name	Genus	Set	Sequence	CYP	GT	Representative*
Ginkgoaceae	1	1	21590	162	140	*Ginkgo biloba*
Gnetaceae	1	1	10724	29	93	*Gnetum gnemon*
Grimmiaceae	1	1	996	3	2	*Grimmia pilifera*
Grossulariaceae	1	2	8490	55	88	*Ribes nigrum*
Haematococcaceae	1	1	999	1	2	*Haematococcus pluvialis*
Hydrocharitaceae	1	1	70	1	1	*Hydrilla verticillata*
Hypericaceae	1	2	18	0	0	*Hypericum hookerianum*
Iridaceae	2	4	13512	173	106	*Crocus sativus*
Isoetaceae	1	1	338	4	0	*Isoetes lacustris*
Juglandaceae	2	5	19091	76	208	*Juglans hindsii x Juglans regia*
Klebsormidiaceae	1	1	4827	3	35	*Klebsormidium subtile*
Lamiaceae	8	12	55342	706	637	*Ocimum basilicum*
Lauraceae	1	1	16558	54	198	*Persea americana*
Liliaceae	2	6	5293	9	41	*Fritillaria cirrhosa*
Limnanthaceae	1	1	15331	10	29	*Limnanthes alba*
Linaceae	1	1	286852	921	1401	*Linum usitatissimum*
Linderniaceae	2	2	270	1	2	*Torenia fournieri*
Lycopodiaceae	1	1	3451	9	9	*Huperzia serrata*
Lythraceae	3	3	2119	3	20	*Cuphea paucipetala*
Magnoliaceae	1	1	24132	124	215	*Liriodendron tulipifera*
Malvaceae	8	15	552468	3493	7495	*Gossypium hirsutum*
Marchantiaceae	1	1	33692	107	144	*Marchantia polymorpha*
Marsileaceae	1	2	61	0	0	*Marsilea vestita*
Mesostigmataceae	1	1	15972	2	20	*Mesostigma viride*
Micractiniaceae	1	1	800	0	2	*Micractinium sp. HK002*
Moraceae	4	6	22132	98	415	*Musa ABB Group*
Musaceae	1	6	20841	63	189	*Musa acuminata AAA Group*
Myrtaceae	3	9	37491	166	643	*Eucalyptus gunnii*
Nelumbonaceae	1	1	2207	7	11	*Nelumbo nucifera*
Nyctaginaceae	1	1	8	0	0	*Mirabilis jalapa*
Nymphaeaceae	1	1	20589	40	224	*Nuphar advena*
Oleaceae	2	2	18102	92	157	*Fraxinus excelsior*
Onagraceae	1	1	3530	5	43	*Oenothera elata subsp. hookeri*
Orchidaceae	11	14	11688	81	131	*Phalaenopsis equestris*
Orobanchaceae	2	3	176369	1160	2200	*Striga hermonthica*
Osmundaceae	2	3	28381	24	352	*Ostreococcus 'lucimarinus CCE9901*
Paeoniaceae	1	1	2204	3	6	*Paeonia suffruticosa*
Pandanaceae	1	1	977	0	9	*Pandanus odoratissimus*
Papaveraceae	2	2	31814	185	239	*Papaver somniferum*
Pedaliaceae	1	1	3328	14	23	*Sesamum indicum*
Phrymaceae	1	3	279620	1885	3527	*Mimulus guttatus*
Phyllanthaceae	1	1	62	2	0	*Phyllanthus amarus*
Phytolaccaceae	1	1	451	3	2	*Phytolacca americana*
Pinaceae	4	21	1015269	5501	10426	*Pinus taeda*
Piperaceae	1	4	130	0	0	*Piper nigrum*
Plantaginaceae	4	4	25989	112	232	*Antirrhinum majus*
Platanaceae	1	1	7	0	0	*Platanus x acerifolia*
Plumbaginaceae	2	3	7314	13	15	*Limonium bicolor*
Poaceae	45	116	6676087	28920	72322	*Zea mays*
Podostemaceae	1	1	9679	36	50	*Polypleurum stylosum*
Polemoniaceae	1	1	5445	17	62	*Ipomopsis aggregata*
Polygonaceae	4	6	9568	23	63	*Polygonum sibiricum*
Polyphysaceae	1	1	4411	2	15	*Acetabularia acetabulum*
Posidoniaceae	1	1	3089	13	4	*Posidonia oceanica*
Pottiaceae	1	1	9991	32	67	*Syntrichia ruralis*
Primulaceae	3	3	2170	14	19	*Cyclamen persicum*
Proteaceae	1	1	24	1	0	*Gevuina avellana*
Pteridaceae	1	1	5125	16	37	*Ceratopteris richardii*
Pycnococcaceae	1	1	126	0	0	*Nephroselmis olivacea*
Ranunculaceae	4	4	92197	969	1782	*Aquilegia formosa x Aquilegia pubescens*
Rhizophoraceae	3	5	22562	84	100	*Bruguiera gymnorhiza*
Rosaceae	11	34	513044	2403	6392	*Malus x domestica*
Rubiaceae	6	12	265964	1725	3636	*Coffea arabica*
Rutaceae	4	30	567435	4855	9389	*Citrus sinensis*
Salicaceae	2	24	422902	2155	5045	*Populus trichocarpa*

Table 1. (Continnue)

Family name	Genus	Set	Sequence	CYP	GT	Representative*
Sapindaceae	2	2	14954	134	121	*Paullinia cupana var. sorbilis*
Saururaceae	1	1	15	0	0	*Saururus chinensis*
Scenedesmaceae	1	3	6630	4	15	*Scenedesmus obliquus*
Schisandraceae	1	1	233	1	0	*Illicium parviflorum*
Sciadopityaceae	1	1	11	0	0	*Sciadopitys verticillata*
Selaginellaceae	1	2	97503	807	1454	*Selaginella moellendorffii*
Selenastraceae	1	1	41	0	0	*Selenastrum capricornutum*
Simmondsiaceae	1	1	385	0	1	*Simmondsia chinensis*
Solanaceae	7	42	1316011	8259	16239	*Nicotiana tabacum*
Tamaricaceae	2	5	22731	46	80	*Tamarix hispida*
Taxaceae	1	1	161	21	0	*Taxus cuspidata*
Theaceae	1	3	13993	45	123	*Camellia sinensis*
Tropaeolaceae	1	1	10507	18	35	*Tropaeolum majus*
Typhaceae	1	2	126	0	1	*Typha angustifolia*
Ulmaceae	1	1	1277	40	6	*Ulmus americana*
Ulvaceae	1	3	2290	1	13	*Ulva linza*
Urticaceae	1	2	418	1	4	*Boehmeria nivea*
Velloziaceae	1	1	400	1	7	*Xerophyta humilis*
Violaceae	1	1	43	0	0	*Viola baoshanensis*
Vitaceae	2	17	526766	1788	4726	*Vitis vinifera*
Welwitschiaceae	1	1	10129	37	125	*Welwitschia mirabilis*
Woodsiaceae	1	1	10	0	0	*Athyrium distentifolium*
Xanthoceraceae	1	1	4	0	0	*Xanthoceras sorbifolium*
Zamiaceae	1	3	20677	57	95	*Zamia vazquezii*
Zingiberaceae	2	3	50779	392	773	*Zingiber officinale*
Zosteraceae	1	1	10659	33	57	Zostera marina
Zygnemataceae	1	1	7294	1	46	Spirogyra pratensis
na	6	7	9426	18	51	Micromonas sp. CCMP490

* In each family, a dataset with the maximal number of sequences is set as the representative.

perform more stable homology searches and to provide more information on molecular functions of ESTs, downsizing while maintaining the precision of homology search and also constructing local modules, in which sequences are highly homologous and thus belong to a group with a common feature, are useful. The PSI-BLAST algorithm and its derivatives (Altschul et al. 1997; Lee et al. 2009; Li et al. 2011) focus on sequence-to-sequence hits between multiple sequences. To evaluate relationships between multiple elements (e.g., gene or metabolite), network module analysis is a useful approach (Saito et al. 2008). Network module analyses have been applied to plant gene co-expression, in which a plant gene is related to other genes based on similar expression profiles (Aoki et al. 2007; Ficklin and Feltus 2011; Huber et al. 2007; Marino-Ramirez et al. 2009; Ogata et al. 2010; Winden et al. 2011). This approach allows a co-expression module, which includes co-expressed gene to be assigned to a particular biological process. By identifying homologies between sequences, network module analysis can be used to create, a homology network in which a sequence (node) is connected to other sequences on the basis of high homology. To perform such analysis for a homology network, we used our algorithm (Ogata et al. 2009) according to the following processes: 1) performing BLAST for any pairs of sequences, 2) calculating association indices between pairs as described in "User-friendly tools for using plant EST and protein sequence databases", 3) depicting a network composed of sequences (nodes) and node-to-node links with high association indices, and 4) detecting local network modules with high *NC* values (Ogata et al. 2009). A module may include sequences representing both known and unknown molecular functions. Those encoding unknown functions can be assigned a function based on high homology to sequences in the module with known function. Moreover, due to high intra-modular homology, a single sequence included in a module can be substituted for the module for functional analysis of sequences. In the example shown in Figure 1, network module analysis can assign more sequences (5 vs. 3) and downsize the databases by one-fifth. User-friendly tools for functional analysis have been made available by applying these advantages of network module analysis.

We introduce the practical and potential uses of plant ESTs in the second section ("Uses of plant ESTs") of this report. Sequences of plant genomes and ESTs have been entered in public databases where they are freely available to anonymous users. These sequences are provided with their metadata, which are essential to extract their functional regions and to identify the sequences. In the third section ("Plant EST databases"), we discuss public databases available for storage of DNA

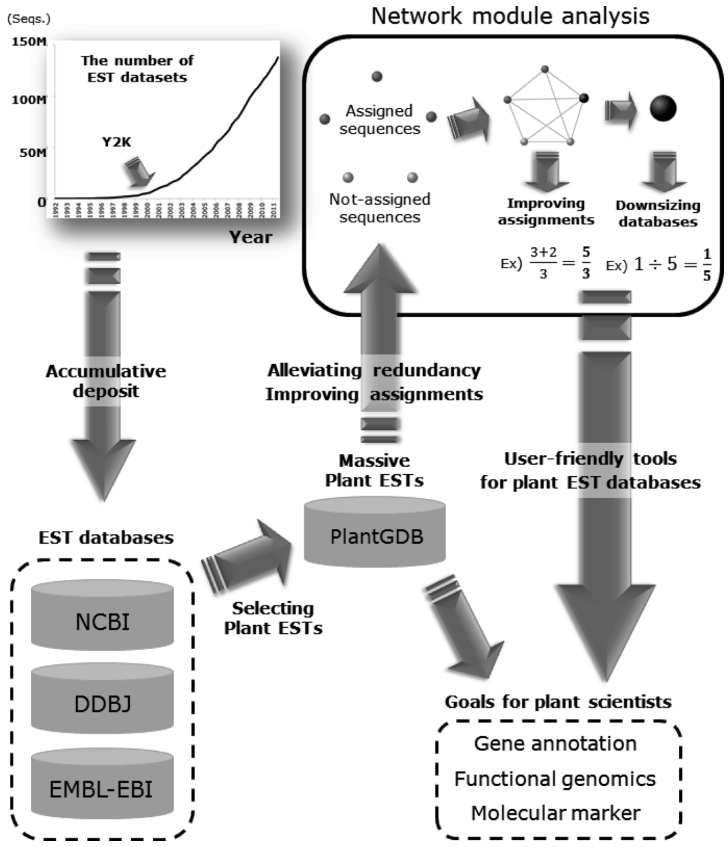

Figure 1. A flowchart showing the application of plant EST datasets and network module analysis of the datasets.

and RNA sequences from various plants. Studies of plant enzymes lead to the determination of enzyme functions and also the understanding of plant physiology and metabolism. In the fourth section ("Categorization of plant proteins"), we present websites that provide information on enzymes with specific functions. A group of sequences with a particular function tend to include various levels of redundancy; using one-to-one BLAST search, this redundancy may lead to extremely poor performance in homology searches, and failure to assign a function to a sequence. Network module analysis supports development of tools that are more user-friendly; they can reduce such redundancy leading to improved performance on the downsized databases. In the last section ("User-friendly tools for using plant EST and protein sequence databases"), we introduce the application of network module analysis to downsizing databases and improving assignment of function.

Uses of plant ESTs

Plant EST resources are used in the following 3 areas of research

(i) *Plant genome research*, to clarify the structural gene annotation including exon–intron boundaries, alternative splicing variants, and demarcation of untranslated regions in comparison with plant genomic sequence. Many plant genome scientists make use of plant EST sequence data. Sequences of full-length cDNA clones such as RIKEN *Arabidopsis* full-length cDNA (RAFL) clones (Seki et al. 2004) exist among a variety of EST clone sequences; plant EST data provide a new resource for plant full-length cDNA clone sequences.

(ii) *Plant functional protein research*, for biochemical and protein structural analyses of identified gene products. These gene products synthesize or modify basic metabolic structures, especially, functional analyses for biochemistry and protein structure of genes involved in plant natural products biosynthesis lead to the understanding of plant metabolic system (Yonekura-Sakakibara and Saito 2009). To isolate a cDNA clone of interest, at first we try to submit a key word or execute homology search in public databases such as NCBI, EMBL-EBI, and DDBJ. If an EST clone obtained from plant EST databases is a truncated clone, a full length cDNA clone provided by rapid amplification of cDNA ends (RACE) technology is necessary for functional analyses in a heterologous expression system.

(iii) *Plant breeding research*, to make molecular markers such as simple sequence repeats (SSRs), for use in the examination of genetic relationships for plant breeding, mapping of useful genes, and construction of

linkage maps (Kalia et al. 2011). Because EST-derived markers come from transcribed regions of the genome, they are likely to be conserved across a broader taxonomic range than are other types of markers. EST-based SSR markers (EST-SSRs) can be rapidly and inexpensively developed from existing plant EST databases.

In order to more effectively use plant EST resources, user-friendly tools based on plant EST sequence data are necessary, especially for the purpose of individual researches. User-friendly plant EST databases allow access to data for use in studies of structural gene annotation, functional genomics, and genetic relationships for plant breeding, without requiring time-consuming procedures.

Plant EST databases

EST sequences are stored in public databases such as NCBI, EMBL-EBI, and DDBJ. These corresponding websites provide access to EST datasets, which include data from all types of organisms and environmental samples, in FASTA and metadata (GBFF, GPFF, and GFF) formats. The FASTA format includes metadata only for identifying individual sequences and is used for homology searches using BLAST. The metadata format includes the following metadata: locus name, sequence length, defined name, several accession numbers for other databases, sample source, taxonomy, journal reference, sequence features such as gene or protein names and information on functional regions, and nucleotide or amino acid sequence. To use the metadata for retrieving functional information, it is essential to trim functional regions (e.g., domains or motifs) or to access different databases. Although the datasets are useful for functional analysis, it is difficult to select plant EST datasets from among the datasets of various organisms. The PlantGDB website selects out plant ESTs from public databases and lists them by individual plant including species, subspecies, and cultivars. As of 2011, PlantGDB provides EST datasets for 848 plants in FASTA file format, which includes 22 933 800 sequences, 428 genera, and 157 families (Table 1). Supplementary table represents all dataets of plant ESTs obtained from PlantGDB. A direct link of each dataset to the publication site is included. Maximum members of datasets include the family Poaceae with datasets for 116 plants, and the genus *Citrus* with datasets for 27 plants. The *A. thaliana* EST dataset includes 1 529 700 sequences, which is approximately 20 times more than the number of cDNA sequences for the plant; 77 461 sequences were found in the file "ATcdna171", obtained from PlantGDB. The dataset with the most EST sequences is maize (2 019 105 sequences). The 2 large groups of enzymes, cytochrome P450 (CYP) and glycosyltransferase (GT), which are related to the enzymatic diversity of plant natural products, account for 0.3% and 0.5% of the whole genome-level dataset, respectively (Table 1). Plant EST datasets in PlantGDB are useful resources for functional analysis of enzymes and other proteins. However, sequence datasets are exponentially accumulating as shown in Figure 1; it is thus important to reduce sequence redundancy and to assign accumulating sequences to particular molecular functions in a high-throughput mode.

Categorization of plant proteins

To analyze protein function of a sequence of interest, several databases provide sequence data for specific functional groups. Protein sequence datasets are, in general, available at public databases such as RefSeq, published by NCBI. RefSeq provides 6 types of datasets: FASTA- and GBFF-formatted files of genomic DNA, RNA, and protein sequences. For plant researchers, RefSeq provides sequence datasets of plants (available at the FTP site; ftp://ftp.ncbi.nih.gov/refseq/release/plant/). As of May 2011, 396 895 protein sequences were available. The sequences are stored in this website, but not categorized into functional groups on the website; it is difficult for plant researchers to perform systematic functional analyses. By collecting sequences for specific protein functional groups, the Cazy website (Cantarel et al. 2009; http://www.cazy.org/) categorized glycoside hydrolases, GTs (the numbers of the sequences are described in the 6th column in Table 1), polysaccharide lyases, carbohydrate esterases, and carbohydrate-binding modules into 125, 92, 22, 16, and 64 functional families, respectively. CYPs (the numbers of the sequences are described in the 5th column in Table 1) were similarly categorized in the database maintained by Nelson (2009) and in the CYP450 Engineering Database (Sirim et al. 2009; http://www.cyped.uni-stuttgart.de/). The CYP450 Engineering Database categorizes the large CYP enzyme family, composed of 11 195 sequences, 8614 proteins, and 620 homologous families, into 249 superfamilies, named as "CYP1"to "CYP772". According to categorization by Nelson (2009), CYP1 to CYP9 exist only in animals, CYP71 to CYP99 and CYP701 to CYP772 exist only in plants, and CYP101 to CYP281 exist in bacteria. These sequences are useful resources for identifying a species or genus and for identifying *de novo* proteins or ESTs. These categorizations are curated by experts to evaluate homology groups and proteins pertinent to particular functions and they can prove useful for functional analysis of plant ESTs and protein sequences. On the other hand, it is difficult to curate any kind of EST and protein functions, and the application of functional analysis to various ESTs and proteins should be further improved. Additionally, for a more stable

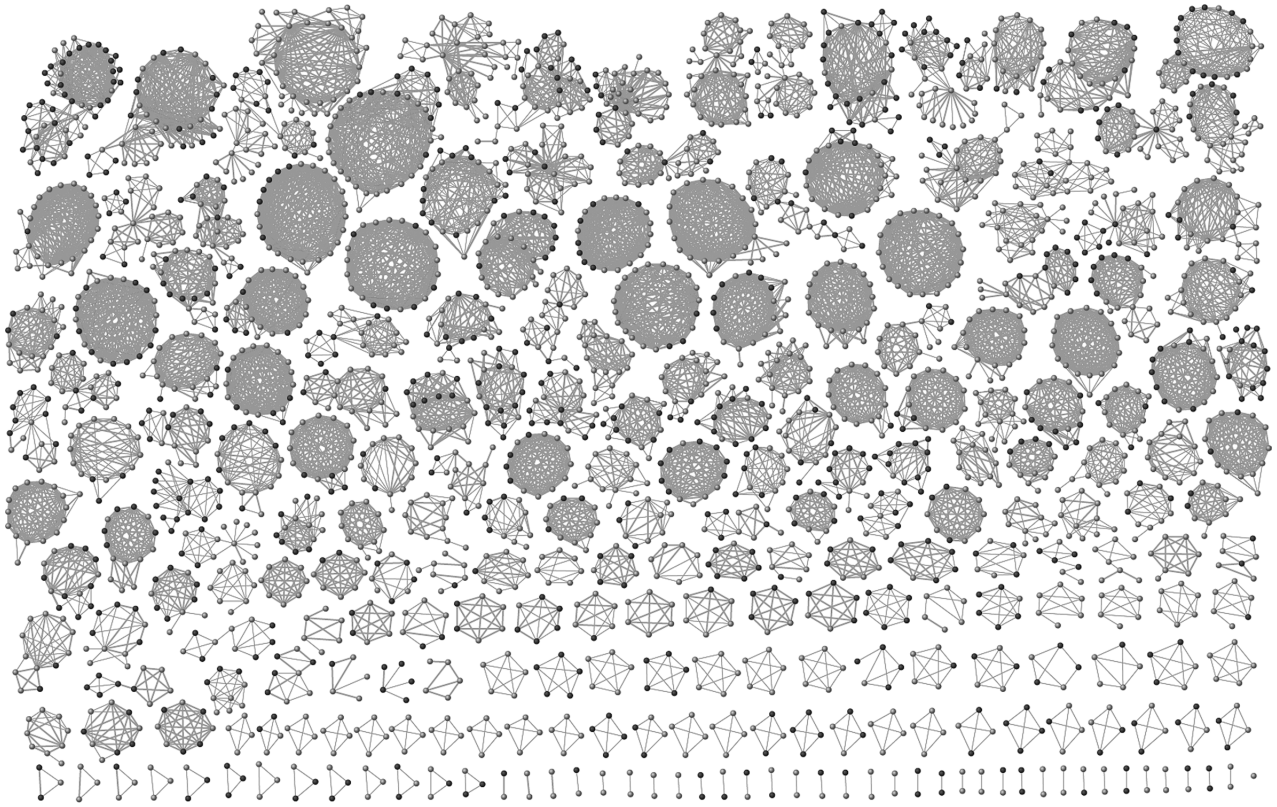

Figure 2. A homology network of plant cytochrome P450s (CYPs). In the network, individual nodes represent amino acid sequences of enzymes (circles) and sequence-to-sequence links are connected on the basis of association indices between sequences, which were calculated using "bit scores" of BLAST searches as follows: an index of sequence A to sequence B was calculated as the bit score of sequence A to sequence B divided by the bit score of sequence B to itself. The network includes 217 local modules with multiple sequences and 494 singletons with no link to other sequences (all singletons are expressed as a single node in the network). A tightly connected module indicates that module members (sequences) are highly homologous. If sequences assigned (dark-colored circles) and not assigned (light-colored circles) to a particular taxonomic level coexist in a module, the unassigned sequences can be assigned to the taxonomic level of the other sequences in that module. Of 3167 sequences in the CYP network, the number of assigned sequences changed from 871 (27.5%) to 2442 (77.1%) using network module analysis.

performance, it is useful to assemble stored plant EST and protein sequences into groups that are highly homologous. High homology indicates that a sequence in a homology group can be representative of the group.

User-friendly tools for using plant EST and protein sequence databases

The downsizing of plant EST and protein sequence databases and the assignment of sequences to particular functional categories are useful approaches to make such databases more user-friendly. Although the redundant sequences are useful for precise identification of species, they may cause homology searches to perform poorly. Downsizing a sequence database while maintaining high precision can circumvent this problem. Furthermore, assignment of plant EST and protein sequences to particular functions provides more information about the molecular functions of sequences of interest.

To downsize plant EST datasets and improve the assignment of sequences to particular functions, we applied network module analysis to obtain homology modules. These modules are composed of highly homologous sequences and can be used to downsize a database by selecting representative sequences. They can also assign functions to unassigned sequences using categories for sequences with known function in the same module (Figure 1); i.e., in Figures 2, 3, when a local network module includes sequences with known function (darker nodes) and with unknown function (lighter nodes), the unknown sequences can be assigned to the function assigned in the known sequences on the basis of their high homologies.

For this approach, we obtained sequences of 2 large families of enzymes: CYPs (3167 sequences) and GTs (5430 sequences). This data was obtained in GPFF file format from the RefSeq database domain information. Of these sequences, 871 in the CYP family and 888 in the GT family were assigned to functions according to Nelson (Nelson 2009) and Cazy, respectively. Within each dataset, we performed a BLAST search. The dataset was used as both query data and the database. Association indices between sequences were calculated, ranging from 0 to 1, based on indices representing the

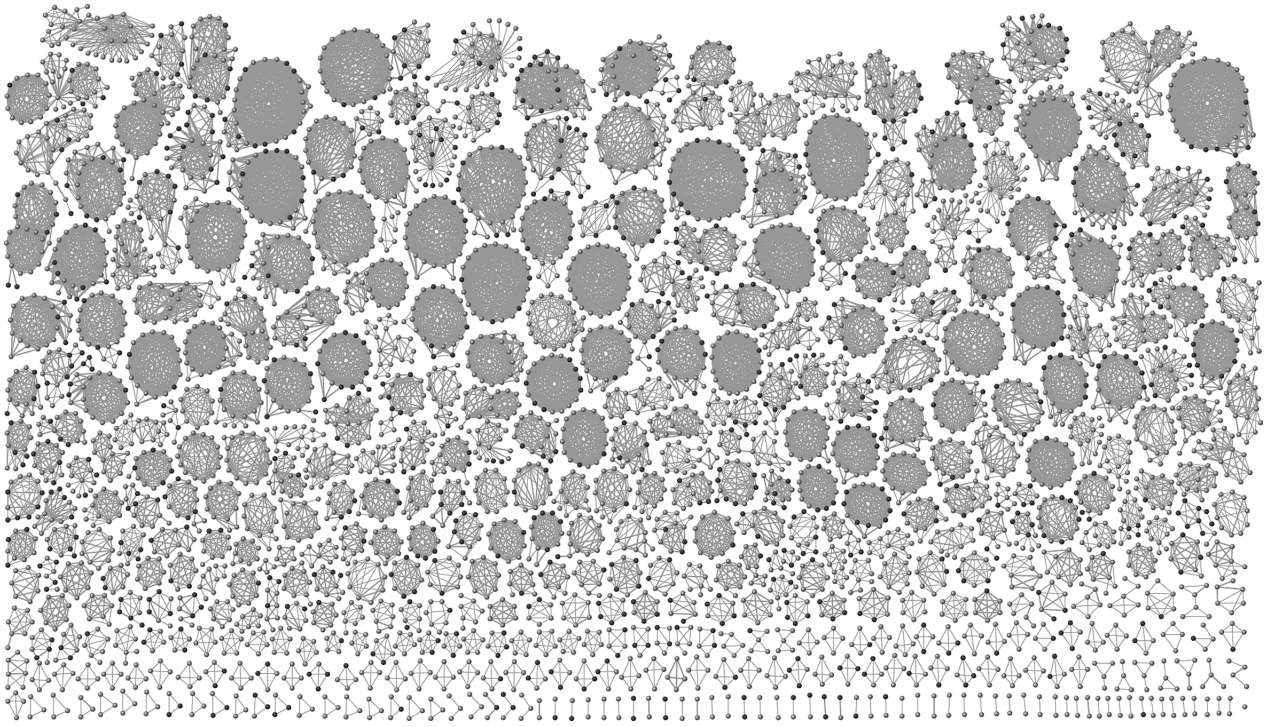

Figure 3. A homology network of plant glycosylhydrolases (GTs). This network was depicted similarly to Figure 2. The network includes 391 modules with multiple sequences and 318 singletons. Of 5430 sequences in the GT network, the number of assigned sequences changed from 888 (16.4%) to 3549 (65.4%).

degree of homology (called "bit score" in BLAST). For example, an index of sequence A to sequence B was calculated as the bit score of sequence A to sequence B divided by the bit score of sequence B to itself.

Network modules of CYPs composed of highly homologous sequences (Figure 2) were obtained by applying the network module analysis (Ogata et al. 2009) to a dataset of association indices between sequences. This network contained 217 modules and 494 singletons, which had no links to other sequences; they were not clearly homologous to any other sequences. All of these modules included few sequences that could be assigned to superfamilies (light-colored circles in Figure 2; (Nelson 2009) and few sequences that were categorized as superfamily members (dark-colored circles in Figure 2). Through analysis of these modules, we were able to categorize the previously unassigned sequences as members of the superfamily. Of 3167 sequences in the CYP network, the number of assigned sequences changed from 871 (27.5%) to 2442 (77.1%) using network module analysis.

We similarly identified a network of GTs that included 455 modules, and 655 singletons. We assigned the GT sequences to the categories established by Cazy (Figure 3). Each module can be downsized to a single representative sequence while maintaining high precision in functional analysis BLAST searches due to their high homologies. Of 5430 sequences in the GT network, the number of assigned sequences changed from 888 (16.4%) to 3549 (65.4%).

A database downsizing with high precision of homology search and improving the assignment of molecular function is applicable to any type of sequence or sequence family. Network module analysis will thus contribute to more stable performance and acquisition of more information via molecular function in BLAST searches. For providing these sequences for functional analysis, we developed a web tool called "E-class" (http://database.riken.jp/ecomics/eclass/), included in the ECOMICS suite. This version of E-class provides databases of small subunit of ribosomal RNA, micro-organic carbohydrate-binding module, and plant cytochrome P450 and GT as the two main protein families important for plant metabolism. We checked the precision of assignment to particular functions in both full-sized and modularized databases and acquired the result showing 99% or higher precision. Additionally, the database keeps information on memberships of the modules and thus comprehensiveness in homology search. We are ready for publishing furthermore databases modularized using our network module analyses in E-class.

Acknowledgements

We thank Daisuke Shibata and Jun Kikuchi for discussing the development of our algorithm for network module analysis and the construction of our web tool to provide downsized sequence

databases. We also would like to express our thanks to Yusuke Morioka for designing and managing the web tool. This work was supported in part by a grant from the New Energy and Industrial Technology Development Organization (NEDO). This research was also partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research on Innovative Areas, 20200062, 2008-2010 (to H.S) and 23108528, 2011 (to H.S).

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* 25: 3389–3402

Aoki K, Ogata Y, Shibata D (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol* 48: 381–390

Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol* 132: 84–91

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucl Acids Res* 37: D233-238

Duvick J, Fu A, Muppirala U, Sabharwal M, Wilkerson MD, Lawrence CJ, Lushbough C, Brendel V (2008) PlantGDB: a resource for comparative plant genomics. *Nucl Acids Res* 36: D959-965

Fedorova M, van de Mortel J, Matsumoto PA, Cho J, Town CD, VandenBosch KA, Gantt JS, Vance CP (2002) Genome-wide identification of nodule-specific transcripts in the model legume Medicago truncatula. *Plant Physiol* 130: 519–537

Ficklin SP, Feltus FA (2011) Gene Coexpression Network Alignment and Conservation of Gene Modules between Two Grass Species: Maize and Rice. *Plant Physiol* 156: 1244–1256

Hofte H, Desprez T, Amselem J, Chiapello H, Rouze P, Caboche M, Moisan A, Jourjon MF, Charpenteau JL, Berthomieu P, et al. (1993) An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from Arabidopsis thaliana. *Plant J* 4: 1051–1061

Huber W, Carey VJ, Long L, Falcon S, and Gentleman R (2007) Graphs in molecular biology. *BMC Bioinformatics* 8: S6, S8

Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK (2011) Microsatellite markers: an overview of the recent progress in plants. *Euphytica* 177: 309–334

Keith CS, Hoang DO, Barrett BM, Feigelman B, Nelson MC, Thai H, Baysdorfer C (1993) Partial sequence analysis of 130 randomly selected maize cDNA clones. *Plant Physiol* 101: 329–332

Lee MM, Chan MK, Bundschuh R (2009) SIB-BLAST: a web server for improved delineation of true and false positives in PSI-BLAST searches. *Nucl Acids Res* 37: W53–56

Li Y, Chia N, Lauria M, Bundschuh R (2011) A performance enhanced PSI-BLAST based on hybrid alignment. *Bioinformatics* 27: 31–37

Marino-Ramirez L, Tharakaraman K, Bodenreider O, Spouge J, Landsman D (2009) Identification of cis-regulatory elements in gene co-expression networks using A-GLAM. *Meth Mol Biol* 541: 1–22

Michalek W, Weschke W, Pleissner KP, Graner A (2002) EST analysis in barley defines a unigene set comprising 4,000 genes. *Theor Appl Genet* 104: 97–103

Nelson DR (2009) The cytochrome p450 homepage. *Human Genomics* 4: 59–65

Newman T, de Bruijn FJ, Green P, Keegstra K, Kende H, McIntosh L, Ohlrogge J, Raikhel N, Somerville S, Thomashow M, et al. (1994) Genes galore: a summary of methods for accessing results from large-scale partial sequencing of anonymous Arabidopsis cDNA clones. *Plant Physiol* 106: 1241–1255

Ogata Y, Sakurai N, Suzuki H, Aoki K, Saito K, Shibata D (2009) The prediction of local modular structures in a co-expression network based on gene expression datasets. *Genome Inform* 23: 117–127

Ogata Y, Suzuki H, Sakurai N, Shibata D (2010) CoP: a database for characterizing co-expressed gene modules with biological information in plants. *Bioinformatics* 26: 1267–1268

Parida SK, Kalia SK, Sunita K, Gaikwad K, Sharma TR, Srivastava PS, Singh NK, Mohapatra T (2009) Unigene drived microsatellite markers for the cereal genomes. *Theor Appl Genet* 112: 808–817

Park YS, Kwak JM, Kwon OY, Kim YS, Lee DS, Cho MJ, Lee HH, Nam HG (1993) Generation of expressed sequence tags of random root cDNA clones of Brassica napus by single-run partial sequencing. *Plant Physiol* 103: 359–370

Paterson AH, Freeling M, Tang H, Wang X (2010) Insights from the comparison of plant genome sequences. *Annu Rev Plant Biol* 61: 349–372

Rounsley SD, Glodek A, Sutton G, Adams MD, Somerville CR, Venter JC, Kerlavage AR (1996) The construction of Arabidopsis expressed sequence tag assemblies. A new resource to facilitate gene identification. *Plant Physiol* 112: 1177–1183

Rudd S (2003) Expressed sequence tags: alternative or complement to whole genome sequences? *Trends Plant Sci* 8: 321–329

Saito K, Hirai MY, Yonekura-Sakakibara K (2008) Decoding genes with coexpression networks and metabolomics—'majority report by precogs'. *Trends Plant Sci* 13: 36–43

Sasaki T, Song J, Koga-Ban Y, Matsui E, Fang F, Higo H, Nagasaki H, Hori M, Miya M, Murayama-Kayano E, et al. (1994) Toward cataloguing all rice genes: large-scale sequencing of randomly chosen rice cDNAs from a callus cDNA library. *Plant J* 6: 615–624

Sato S, Hirakawa H, Isobe S, Fukai E, Watanabe A, Kato M, Kawashima K, Minami C, Muraki A, Nakazaki N, et al. (2011) Sequence analysis of the genome of an oil-bearing tree, Jatropha curcas L. *DNA Res* 18: 65–76

Seki M, Satou M, Sakurai T, Akiyama K, Iida K, Ishida J, Nakajima M, Enju A, Narusaka M, Fujita M, et al. (2004) RIKEN Arabidopsis full-length (RAFL) cDNA and its applications for expression profiling under abiotic stress conditions. *J Exp Bot* 55: 213–223

Shoemaker R, Keim P, Vodkin L, Retzel E, Clifton SW, Waterston R, Smoller D, Coryell V, Khanna A, Erpelding J, et al. (2002) A compilation of soybean ESTs: generation and analysis. *Genome* 45: 329–338

Sirim D, Wagner F, Lisitsa A, Pleiss J (2009) The cytochrome P450 engineering database: Integration of biochemical properties. *BMC Biochem* 10: 27

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408: 796–815

Uchimiya H, Kidou S, Shimazaki T, Takamatsu S, Hashimoto H, Nishi R, Aotsuka S, Matsubayashi Y, Kidou N, Umeda M, et al. (1992) Random sequencing of cDNA libraries reveals a variety of expressed genes in cultured cells of rice (*Oryza sativa* L.) *Plant J* 2: 1005–1009

Winden KD, Karsten SL, Bragin A, Kudo LC, Gehman L, Ruidera J, Geschwind DH, Engel J Jr (2011) A systems level, functional genomics analysis of chronic epilepsy. *PLoS One* 6: e20763

Yamamoto K, Sasaki T (1997) Large-scale EST sequencing in rice. *Plant Mol Biol* 35: 135–144

Yonekura-Sakakibara K, Saito K (2009) Functional genomics for plant natural product biosynthesis. *Nat Prod Rep* 26: 1466–1487

Yuan D, Tu L, Zhang X (2011) Generation, Annotation and Analysis of First Large-Scale Expressed Sequence Tags from Developing Fiber of Gossypium barbadense L. *PLoS One* 6: e22758