Journal of Oil Palm Research Special Issue on Malaysia-MIT Biotechnology Partnership Programme: Volume 1-Oil Palm Tissue Culture (April 2008) p. 35-43

FUNCTIONAL ANNOTATION OF OIL PALM GENES USING AN AUTOMATED BIOINFORMATICS APPROACH

LAURA B WILLIS*; PHILIP A LESSARD*; JEFFERSON A PARKER*; XIAN M O'BRIEN* and ANTHONY J SINSKEY*

ABSTRACT

Recent advances in DNA sequencing technologies have led to a tremendous increase in the amount of sequence information available in public databases. To address the need for automated methods of assigning a putative function to each sequence, we have developed bioinformatics tools that can be run on a desktop computer and save significant time and effort. Elaeis guineensis *and* Elaeis oleifera *sequences were downloaded from PalmGenes and GenBank, and duplicate entries were eliminated by pairwise BLAST searches, resulting in a collection of unique oil palm sequences which we call the UniPalm dataset. We applied the CAPASA (Consensus Annotation by Phrase Anchored Sequence Alignment) software and automatically assigned functions to 5600 oil palm sequences in less than 8 hr. CAPASA mimics the human decision-making process by factoring in the degree of homology, taxonomic relationship and informational value when choosing a name. In addition, we applied COGsensus to place the UniPalm sequences into COG (Clusters of Orthologous Groups of genes) categories, and compared these results to a COGsensus analysis of the rice genome. COG classification is a homology-based method for distinguishing gene sets, particularly with regard to closely related genes found in different organisms. Our results indicate that the diversity of COG groups are well represented in the UniPalm set.*

Keywords: bioinformatics, expressed sequence tags, *Elaeis guineensis*, PalmGenes.

Date received: 31 July 2007; **Sent for revision:** 17 August 2007; **Received in final form:** 25 October 2007; **Accepted:** 7 November 2007.

INTRODUCTION

The oil palm, *Elaeis guineensis*, is the most productive oil producing plant under cultivation, with typical yields of 3.3 t of oil per hectare per year (Wahid *et al*., 2005). Palm oil is a rich nutritional source of vitamins, carotenoids, iron, and antioxidant activity (Sundram *et al*., 2003; Balasundram *et al.*, 2005). The palm fruit type grown in Malaysia is *tenera*, derived by crossing the thick-shelled *dura* with the thin-shelled *pisifera*. *Tenera* has a thick oil-rich mesocarp with a thin shelled covering the kernel. Continuous improvement of oil palm planting materials is being done by conventional breeding, supported by genetic engineering technologies and tissue culture propagation (Sambanthamurthi *et al*., 1996; Parveez *et al.*, 2000; Gorret *et al*., 2004; Tarmizi *et al*., 2004; Abdullah *et al.*, 2005).

Assessment of the genetic make-up of the plants derived from these programmes has been carried out by constructing genetic linkage maps using quantitative trait loci and genetic polymorphisms such as AFLPs (amplification fragment length polymorphisms) and RFLPs (restriction fragment length polymorphisms), and other molecular markers such as proteins and isoenzymes (Mayes *et al.*, 1996; Kularatne *et al*., 2000; Purba *et al*., 2000; Barcelos *et al*., 2002; Jaligot *et al*., 2004; Billotte *et al.*, 2005; Maizura *et al.*, 2006). Quantitative trait loci (QTLs) are linked to the phenotypic traits they follow and provide information about the transmission of

* Department of Biology,
Massachusetts Institute of Technology,
77 Massachusetts Avenue,
Cambridge, MA 02139,
USA.
E-mail: asinkey@mit.edu

traits from parents to progeny; researchers have identified the genetic regions correlated with data about height, oil yield and shell thickness, for example (Moretzsohn *et al*., 2000; Rance *et al*., 2001). However, as QTLs, AFLPs and RFLPs can represent a large region of a chromosome, more work must be done to narrow the region of interest down to a particular gene(s) associated with the desired traits.

To address this issue, researchers began an ambitious DNA sequencing programme to identify oil palm genes that are expressed during important development stages (*e.g*., fruit ripening and embryogenesis) (Cha and Shah, 2001; San and Shah, 2005). To improve their knowledge on the relevance of the genes discovered, researchers focused their efforts on cloning and sequencing expressed sequence tags (ESTs), increasing the likelihood that the genes being studied were physiologically significant (Abdullah *et al*., 1995; Singh and Cheah, 2000; Jouannic *et al*., 2005). Nonetheless, the immediate output of any large-scale sequencing project is a string of nucleotides, without inherent meaning. Learning what physiological role a single sequenced gene may have in oil palm is a challenge, but simultaneously assigning functions to the thousands of genetic sequences flowing from the cloning and sequencing projects would be impossible to do by hand.

To address the need for automated methods of assigning a putative function to each sequence we have developed bioinformatics tools that can be run on a desktop computer and that can result in significant savings in time and effort. In this work, we report the use of our programme CAPASA (Consensus Annotation by Phrase Anchored Sequence Alignment) (Parker, 2004) for functional annotation of oil palm sequences; generation of the 'Unipalm' dataset of 7404 non-redundant oil palm sequences distilled from the PalmGenes and National Centre for Biotechnology Information (NCBI) databases; and our evaluation of how much of the total oil palm genome is represented in this dataset, based on our computational analysis of the COG functions (clusters of orthologous genes) associated with the ESTs in the dataset using the COGsensus software developed in our laboratory. The information generated with these tools adds value to the DNA databases cataloguing oil palm genes, and provides support for oil palm DNA microarray and genetic engineering studies.

EXPERIMENTAL

DNA Sequences (PalmGenes)

We wrote a custom PERL script (download_mpob.pl) and used it to obtain *Elaeis guineensis* DNA sequence information from the PalmGenes database of the Palm Oil Information Online Service of the Malaysian Palm Oil Board (PALMOILIS, http://palmoilis.mpob.gov.my). Information was downloaded serially to capture the MPOB Accession number, gene description/identity, the source clone identification and the DNA sequence of the gene. Each MPOB accession number was processed to extract information for the full range of genes available. Date accessed: 26 February 2005.

DNA Sequences (GenBank)

We wrote a custom PERL script (download_NCBI_palm) to collect DNA sequence from *E. guineensis* and the closely related oil palm *Elaeis oleifera* from the NCBI (NCBI, http://www.ncbi.nlm.nih.gov/) (Benson *et al*., 2007). The NCBI search toolbar was used to determine the full list of gene identifier numbers of sequences from these two species. These lists were used to obtain the NCBI accession number, gene description, organism name and DNA sequence of all the oil palm genes available in the NCBI nucleotide database. Date accessed: 27 February 2005 (*E. guineensis*) and 25 February 2005 (*E. oleifera*).

Sequence Unification (UniPalm)

The *E. guineensis* and *E. oleifera* nucleotide sequence information from the PalmGenes and Genbank databases was unified to a single set of non-redundant sequences using a custom PERL program: palm_orthologs.pl. This process involved several rounds of BLAST searches among the sequence sets to remove highly similar sequences within the sequence collection of a single species from either database, or between species from either database. Sequences were deemed highly similar if the product of the percentage aligned and percent identical residues was greater than 0.97. The sequence with the lower number of gaps and unknown sequence residues was maintained as the representative DNA sequence for the gene. The final list of DNA sequences from *E. guinensis* and *E. oleifera* from PalmGenes and Genbank, with identical or highly similar sequences removed, was renamed the UniPalm (Unified Oil Palm) sequence set.

Functional Annotation Prediction

Functional annotation predictions of the UniPalm DNA sequence collection were made using CAPASA program (Parker, 2004). Briefly, the CAPASA algorithm analyses the output of a BLASTx (DNA versus Protein) sequence similarity search of the query sequence (UniPalm DNA sequence) against the non-redundant protein database at NCBI (Gish and States, 1993). The significant similarity matches

of the protein database were analysed to measure the degree of alignment, similarity of the organism taxonomy within the results and consistency of the word usage in function descriptions across the set of similarity results. The functional description of the protein sequence with the highest combination of sequence similarity, taxonomy similarity and name consistency was transferred to the UniPalm sequence as a putative functional annotation.

Functional Domain Prediction

A second functional annotation prediction was performed using COGsensus (Parker *et al*., unpublished) to classify putative domains by the COG system (Tatusov *et al.,* 2000) within the UniPalm DNA sequences. Sequences of the UniPalm DNA sequence collection were searched against the set of all COG sequences using the BLASTx sequence similarity program. Since COGs are comprised of domains, or partial components of full genes, the COGs are aligned with subsets of the full length gene. BLAST similarity matches were grouped by position and degree of overlap, relative to the full length UniPalm gene. The functional identity was transferred from the COG group with highest average similarity and alignment coverage within each group.

RESULTS AND DISCUSSION

Generation of a Unified Oil Palm Dataset

To support the efforts of the oil palm DNA sequencing and DNA microarray projects of the Malaysia-MIT Biotechnology Partnership Programme, we applied bioinformatics tools to assign putative functions to oil palm DNA sequences. We began by collecting publicly available oil palm EST sequences from PalmGenes and GenBank. PalmGenes is a database devoted to oil palm sequences developed by the Malaysian Palm Oil Board (Cheah *et al.,* 2003). The GenBank database (Benson *et al.,* 2007), operated by the US NCBI, also contains many oil palm sequences.

There are a number of duplicate oil palm sequences in both the PalmGenes and NCBI datasets. One reason for this is that PalmGenes includes not only the sequences generated at MPOB but also sequences collected from other databases, including GenBank, DNA Data Bank of Japan (Tateno *et al*., 2002), EMBL (Cochrane *et al*., 2006), Protein Information Resource (Barker and Wu, 2001), and SwissProt (Boeckmann *et al.,* 2003). Another important factor for the duplication is the nature of EST sequencing projects: a large number of ESTs are sequenced, often without enriching for the unique sequences, which can result in the overrepresentation of highly expressed genes such as the genes encoding ribosomal proteins.

In order to unify the palm EST sequences in one database, we wrote custom PERL scripts to download 5610 *Elaeis guineensis* sequences from PalmGenes and 2797 *Elaeis guineensis* sequences and 51 *Elaeis oleifera* sequences from GenBank. Then, to streamline the dataset and eliminate duplicate samples, we used local BLAST to conduct two rounds of homology searching with the *E. guineensis* or *E. oleifera* sequence collections from MPOB and NCBI.

The first round of searching was a self- *vs*. -self comparison to remove identical sequences from within a single sequence dataset. The identical sequences were removed (with preference to the lower numbered sequences), and a new BLAST database was created from this reduced set. The second round of searching was a self- *vs*. -self and self- *vs*. -other search to find highly similar sequences. The non-self best hits were scored on a value system for having fewer gaps, fewer ambiguous residue or more residues. The sequence with the most points between the pair was kept and the ortholog marked for deletion. If the ortholog and the query sequence had a combined homology (percent aligned residues x percent identical residues) 0.97 the ortholog was deleted. After the removal of 1054 duplicates, the PalmGenes set contained 4605 sequences, the *Elaeis guineensis* NCBI set 2752 samples, and the *Elaeis oleifera* set contained 47 sequences for a total of 7404 *Elaeis* sequences in the UniPalm dataset.

Automated Assignment of Names to the UniPalm Sequences

To add information to the Unipalm dataset, we carried out analyses to infer the functionality of a given gene sequence based initially on homology to known genes. The workhorse for this analysis was Batch BLAST, which is used for high throughput alignment to the NCBI sequence databases. The normal manual mode for BLAST analysis of DNA sequences (Altschul *et al.,* 1990; 1997) requires the researcher to carry out an alignment, collect the homologues and sort through this collection in the hopes of identifying an insightful relationship between the homologues that makes it possible to infer a function for the gene in question. Not surprisingly, this process is tedious and time-consuming. Furthermore, it is somewhat subjective. Although scientists attempt to make consistent decisions regarding the meaning and significance of the relationships discovered, it is impossible to ensure uniformity of the decision-making process; thus, the result may not be entirely objective. In fact, the common practice of transferring the 'best hit' in a BLAST search to assign a name to an unknown

sequence has been reported to contribute to the increase in erroneous information accumulating in sequence databases (Bork and Bairoch, 1996).

To overcome the limitations of throughput, objectivity and reproducibility of the method, we analysed Batch BLAST output information with the CAPASA software, developed in our laboratory (Parker, 2004). CAPASA automatically assigns a functional annotation to each sequence, taking into account the strength of sequence alignments between homologous genes, phylogenetic relationship between the query and subject sequences, and the information content of the gene names. A flow chart showing the operations carried out by the CAPASA software is shown in *Figure 1*. The functions assigned by CAPASA are robust. In tests of this tool comparing annotation of genes from fully sequenced genomes with the annotation provided by 'expert annotators' CAPASA returned functional assignments that agreed with the expert assignments in well over 90% of all cases (Parker and Sinskey, personal communication).

This computational tool operates thousands of times more quickly than human annotation. In this work, CAPASA assigned functions to sequences at a rate of more than 700 sequences per hour while running on a desktop computer. Example data from the CAPASA annotation of oil palm genes are shown in *Table 1*. The full CAPASA annotation of UniPalm is available as supplementary information at http://web.mit.edu/biology/sinskey/www/software/oilpalmannotation.html. The vast majority of the annotations assigned by CAPASA to the UniPalm set agree with the existing expert annotation. In many cases, CAPASA was able to add additional information. This demonstrates the utility of CAPASA for annotating new, unannotated sequences generated in EST sequencing projects.

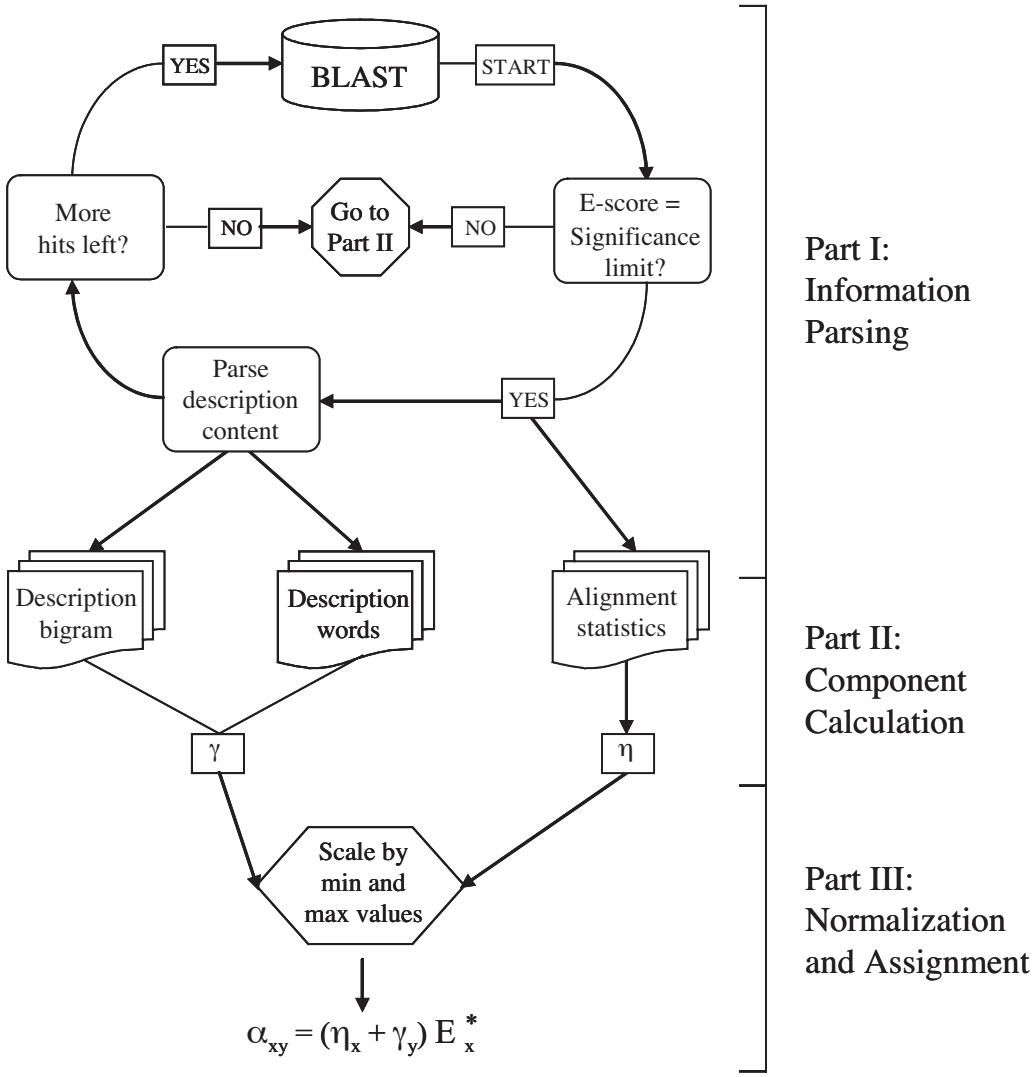

Figure 1. Schematic diagram of the steps in the CAPASA (Consensus Annotation by Phrase Anchored Sequence Alignment) annotation process. The annotation score is the sum of the individual scores for homology, taxonomic relationships and name value. The annotation associated with the database homolog showing the maximum CAPASA score is assigned as the annotation for the individual palm gene.

TABLE 1. EXAMPLE OF CONSENSUS ANNOTATION BY PHRASE ANCHORED SEQUENCE ALIGNMENT (CAPASA) AUTOMATED ANNOTATION COMPARED WITH PALMGENES EXPERT ANNOTATION

PalmGenes ID	PalmGenes Annotation	CAPASA Gene name	Annotation score	E-score
E0000001	At2g04520/T1O3.7	eukaryotic translation initiation factor 1A	1.9818	5E-23
E0000002	Elicitor inducible protein	unknown protein	1.5512	6E-08
E0000003	Xyloglucan endotransglycosylase	xyloglucan:xyloglucosyl transferase, putative/xyloglucan endotransglycosylase, putative/endo-xyloglucan transferase, putative	1.7254	3E-42
E0000004	Anti-viral protein	Similar to synechocystis anti-viral protein	1.914	2E-28
E0000005	ATP citrate lyase	ATP citrate lyase	1.6758	1E-15
E0000006	Ribosomal protein L32	ribosomal protein L32	1.7843	3E-25
E0000007	ORFX	expressed protein	1.7696	2E-35
E0000008	Putative epsin	epsin N-terminal homology (ENTH) domain-containing protein	1.6114	2E-21
E0000009	Homeodomain protein	homeodomain protein	1.7907	1E-24
E0000010	ADP-ribosylation factor	ADP-ribosylation factor 1	2.178	5E-70
E0000011	Methylenetetrahydrofolate reductase	5,10-methylenetetrahydrofolate reductase	1.6323	4E-12
E0000012	Beta-1, 3-glucananse	glucanase	2.0143	1E-33
E0000013	Putative protein	expressed protein	1.6151	1E-43
E0000014	Glutathione transferase (EC 2.5.1.18)	glutathione s-transferase	2.0511	4E-42
E0000016	Unknown protein	unknown protein	1.9033	2E-32
E0000017	Lactate dehydrogenase	lactate dehydrogenase-A	2.0044	7E-37
E0000019	Ribosomal protein L27a	ribosomal protein L27a	2.0435	6E-48
E0000020	Unknown protein	sterile alpha motif (SAM) domain-containing protein	1.8506	6E-18
E0000021	Peroxidase	peroxidase	1.9222	2E-23
E0000022	Hsp70 interacting protein/thioredoxin chimera	heat shock 70kD protein binding protein	1.4983	1E-29
E0000023	Aquaporin	aquaporin	2.0541	2E-66
E0000025	Peroxiredoxin	peroxiredoxin	1.8104	3E-64
E0000026	Aquaporin 2	aquaporin (plasma membrane intrinsic protein 1B)	1.9398	5E-59
E0000027	Hypothetical protein	Rad1-like protein	1.6629	6E-42
E0000028	26S proteasome regulatory subunit S2	putative 26S proteasome regulatory subunit S2	1.7238	2E-73

Note: The full Consensus Annotation by Phrase Anchored Sequence Alignment (CAPASA) annotation of the UniPalm dataset can be accessed at http://web.mit.edu/biology/sinskey/www/software/oilpalmannotation.html.

Automated Assignment of Clusters of Orthologues Groups (COGs) to Unipalm Sequences

Another question arising from the analysis of oil palm genes is, "How much of the genome has been covered?" When examining thousands of sequenced ESTs, it is difficult to determine what portion of the genes that one might expect to find have actually been identified. One method for gauging the 'depth' or 'coverage' of this collection is to compare the number and types of genes it contains with the number and types of genes present in a well characterized reference genome. For example, if a related species is known to possess 300 genes involved in one aspect of cellular metabolism, and the oil palm collection includes 250 genes for the same aspect of cellular metabolism, then it can be inferred that coverage of the oil palm genes is rather good.

To carry out such an analysis of an oil palm EST database, we applied a novel computational tool, developed in our laboratory, to assign sequences to COGs. COG classification is a homology-based method for distinguishing gene sets, particularly with regard to closely related genes found in different organisms (Tatusov *et al*., 2000). We applied the COGsensus bioinformatics tool (Parker *et al*., unpublished) to automatically assign COG functional categories to each gene from rice (the most closely related species of monocot with a completed genomic sequence available at the time) (Goff *et al*., 2002; Yu *et al*., 2002) and the Unipalm dataset of oil palm sequences. By carrying out this analysis, it became possible to compare the numbers of genes each species had in each COG category.

The results from this analysis are summarized in *Figure 2*. This evaluation yielded a very encouraging assessment of the oil palm EST sequencing project - the genes in the UniPalm dataset, derived largely from EST sequencing projects, are primarily those that encoded recognizable functions; they represented a wide diversity of functions, and their coverage was surprisingly consistent with that for the rice genome. Of the 25 different classes of gene products recognized by COG classification, only the genes involved in transcription (COG class K), replication (COG class L), signal transduction (COG class T) and those that are poorly characterized (COG class R) or have no known function in other genomes (COG class S) appear to be underrepresented in the UniPalm dataset. While the UniPalm dataset represents only a small portion of the total oil palm genome, the diversity of COG groups are well represented. Additional cloning and sequencing projects are likely to yield genes of unknown function with increasing frequency. The complete COGsensus annotation of the UniPalm dataset is available as supplementary information at http://web.mit.edu/biology/sinskey/www/software/oilpalmannotation.html.

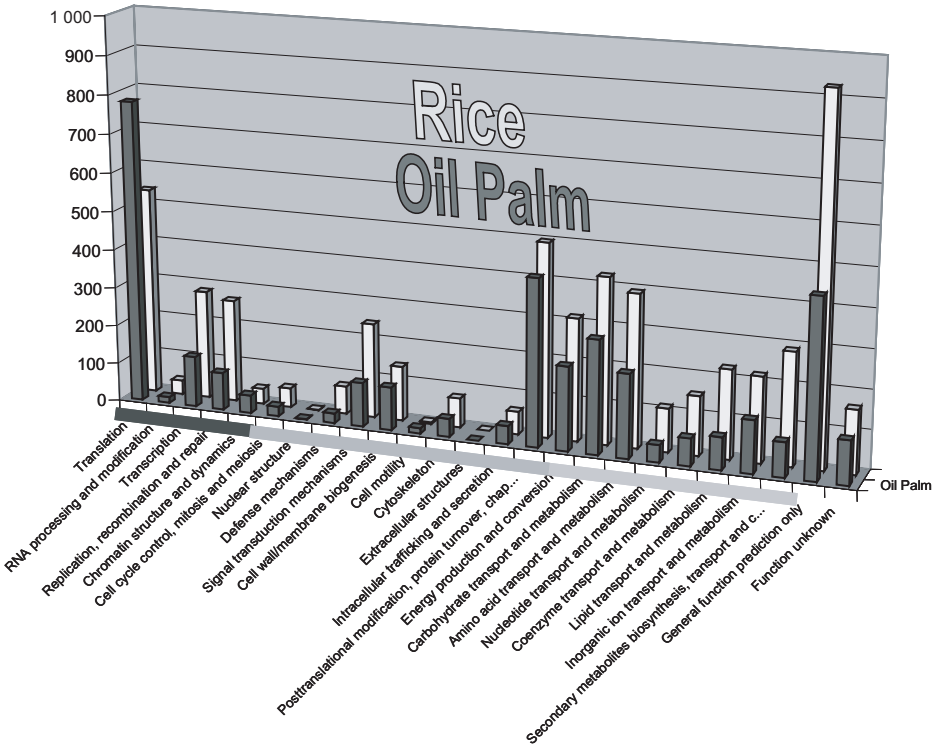

Figure 2. Expressed sequence tags (genes) from rice and oil palm were assigned to functional clusters of orthologous genes (COG) families using COGsensus software. COG families are indicated on the Y-axis, and the number of genes in each COG family plotted on the X-axis.

CONCLUSION

In this article, we described the compilation of the Unipalm dataset of unified oil palm sequences, the application of our CAPASA software to infer functional assignments for oil palm ESTs based on homology to known genes from other species, and the application of our COGsensus software to assign oil palm ESTs to COG groups. These analyses serve to improve the information content associated with the oil palm sequences. Our results indicate that the diversity of COG groups are well represented in the UniPalm set. The output from this research includes a DNA database of oil palm sequences that is suitable for tracking the identity of DNA microarray probes used to generate hybridization data.

ACKNOWLEDGEMENTS

This work was funded in part by the Ministry of Science, Technology and Innovation, Malaysia (MOSTI) and the Malaysia-MIT Biotechnology Partnership Programme. J.A.P. was funded by a Bioprocess Engineering Centre National Science Foundation Training Grant and grants to A.J.S. from DuPont and the Cambridge-MIT Institute.

REFERENCES

ABDULLAH, R; ZAINAL, A; HENG, W Y; LI, L C; BENG, Y C; PHING, L M; SIRAJUDDIN, S A; PING, W Y S; JOSEPH, J L and JUSOH, S A (2005). Immature embryo: a useful tool for oil palm (*Elaeis guineensis* Jacq.) genetic transformation studies. *Electronic J. Biotechnology, 8*: 24-34.

ABDULLAH, S N A; SHAH, F H and CHEAH, S C (1995). Construction of oil palm mesocarp cDNA library and the isolation of mesocarp-specific cDNA clones. *Asia-Pacific J. Molecular Biology and Biotechnology, 3*: 106-111.

ALTSCHUL, S F; GISH, W; MILLER, W; MYERS, E W and LIPMAN, D J (1990). Basic local alignment search tool. *J Mol Biol, 215*: 403-410.

ALTSCHUL, S F; MADDEN, T L; SCHAFFER, A A; ZHANG, J H; ZHANG, Z; MILLER, W and LIPMAN, D J (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research, 25*: 3389-3402.

BALASUNDRAM, N; AI, T Y; SAMBANTHAMURTHI, R; SUNDRAM, K and SAMMAN, S (2005). Antioxidant properties of palm fruit extracts. *Asia Pacific J. Clinical Nutrition, 14*: 319-324.

BARCELOS, E; AMBLARD, P; BERTHAUD, J and SEGUIN, M (2002). Genetic diversity and relationship in American and African oil palm as revealed by RFLP and AFLP molecular markers. *Pesquisa Agropecuaria Brasileira, 37*: 1105-1114.

BARKER, W C and WU, C H (2001). Protein information resource: a community resource for expert annotation of protein data. *Biophysical J., 80*: 33A-33A.

BENSON, D A; KARSCH-MIZRACHI, I; LIPMAN, D J; OSTELL, J and WHEELER, D L (2007). GenBank. *Nucleic Acids Res, 35*: D21-D25.

BILLOTTE, N; MARSEILLAC, N; RISTERUCCI, A M; ADON, B; BROTTIER, P; BAURENS, F C; SINGH, R; HERRAN, A; ASMADY, H; BILLOT, C; AMBLARD, P; DURAND-GASSELIN, T; COURTOIS, B; ASMONO, D; CHEAH, S C; ROHDE, W; RITTER, E and CHARRIER, A (2005). Microsatellite-based high density linkage map in oil palm (*Elaeis guineensis* Jacq.). *Theoretical and Applied Genetics, 110*: 754-765.

BOECKMANN, B; BAIROCH, A; APWEILER, R; BLATTER, M C; ESTREICHER, A; GASTEIGER, E; MARTIN, M J; MICHOUD, K; O'DONOVAN, C; PHAN, I; PILBOUT, S and SCHNEIDER, M (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research, 31*: 365-370.

BORK, P and BAIROCH, A (1996). Go hunting in sequence databases but watch out for the traps. *Trends in Genetics, 12*: 425-427.

CHA, T S and SHAH, F H (2001). Kernel-specific cDNA clones encoding three different isoforms of seed storage protein glutelin from oil palm *Elaeis guineensis*. *Plant Science, 160*: 913-923.

CHEAH, S C; LOW, L E T; RAJINDER, S; FATIMAH, H A Z and MARDHIAH, M Z (2003). PalmGenes. *MPOB Information Series No. 186*: 1-2.

COCHRANE, G; ALDEBERT, P; ALTHORPE, N; ANDERSSON, M; BAKER, W; BALDWIN, A; BATES, K; BHATTACHARYYA, S; BROWNE, P; VAN DEN BROEK, A; CASTRO, M; DUGGAN, K; EBERHARDT, R; FARUQUE, N; GAMBLE, J; KANZ, C; KULIKOVA, T; LEE, C; LEINONEN, R; LIN, Q; LOMBARD, V; LOPEZ, R; MCHALE, M; MCWILLIAM, H; MUKHERJEE, G; NARDONE, F; PASTOR, M P G; SOBHANY, S; STOEHR, P; TZOUVARA, K; VAUGHAN, R; WU, D; ZHU, W M and APWEILER, R (2006). EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Research, 34*: D10-D15.

GISH, W and STATES, D J (1993). Identification of protein coding regions by database similarity search. *Nat Genet, 3*: 266-272.

GOFF, S A; RICKE, D; LAN, T H; PRESTING, G; WANG, R L; DUNN, M; GLAZEBROOK, J; SESSIONS, A; OELLER, P; VARMA, H; HADLEY, D; HUTCHINSON, D; MARTIN, C; KATAGIRI, F; LANGE, B M; MOUGHAMER, T; XIA, Y; BUDWORTH, P; ZHONG, J P; MIGUEL, T; PASZKOWSKI, U; ZHANG, S P; COLBERT, M; SUN, W L; CHEN, L L; COOPER, B; PARK, S; WOOD, T C; MAO, L; QUAIL, P; WING, R; DEAN, R; YU, Y S; ZHARKIKH, A; SHEN, R; SAHASRABUDHE, S; THOMAS, A; CANNINGS, R; GUTIN, A; PRUSS, D; REID, J; TAVTIGIAN, S; MITCHELL, J; ELDREDGE, G; SCHOLL, T; MILLER, R M; BHATNAGAR, S; ADEY, N; RUBANO, T., TUSNEEM, N; ROBINSON, R; FELDHAUS, J; MACALMA, T; OLIPHANT, A and BRIGGS, S (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp japonica). *Science, 296*: 92-100.

GORRET, N; BIN ROSLI, S K; OPPENHEIM, S F; WILLIS, L B; LESSARD, P A; RHA, C and SINSKEY, A J (2004). Bioreactor culture of oil palm (*Elaeis guineensis*) and effects of nitrogen source, inoculum size, and conditioned medium on biomass production. *J. Biotechnol., 108*: 253-263.

JALIGOT, E; BEULE, T; BAURENS, F C; BILLOTTE, N and RIVAL, A (2004). Search for methylation-sensitive amplification polymorphisms associated with the 'mantled' variant phenotype in oil palm (*Elaeis guineensis* Jacq.). *Genome, 47*: 224-228.

JOUANNIC, S; ARGOUT, X; LECHAUVE, F; FIZAMES, C; BORGEL, A; MORCILLO, F; ABERLENC-BERTOSSI, F; DUVAL, Y and TREGEAR, J (2005). Analysis of expressed sequence tags from oil palm (*Elaeis guineensis*). *Febs Letters, 579*: 2709-2714.

KULARATNE, R S; SHAH, F and RAJANAIDU, N (2000). Estimation of genetic diversity in some African germplasm collection of oil palm (*Elaeis guineensis* Jacq.) as detected by AFLP markers. *Asia-Pacific J. Molecular Biology and Biotechnology, 8*: 27-36.

MAIZURA, I; RAJANAIDU, N; ZAKRI, A H and CHEAH, S C (2006). Assessment of genetic diversity in oil palm (*Elaeis guineensis* Jacq.) using Restriction Fragment Length Polymorphism (RFLP). *Genetic Resources and Crop Evolution, 53*: 187-195.

MAYES, S; JAMES, C M; HORNER, S F; JACK, P L and CORLEY, R H V (1996). The application of restriction fragment length polymorphism for the genetic fingerprinting of oil palm (*E. guineensis* Jacq). *Molecular Breeding, 2*: 175-180.

MORETZSOHN, M C; NUNES, C D M; FERREIRA, M E and GRATTAPAGLIA, D (2000). RAPD linkage mapping of the shell thickness locus in oil palm (*Elaeis guineensis* Jacq.). *Theoretical and Applied Genetics, 100*: 63-70.

PARKER, J A (2004). Aromatic hydrocarbon metabolism by *Rhodococcus* sp. I24: computational, biochemical and transcriptional analysis. Ph.D. thesis. Department of Biology. Cambridge, MA, Massachusetts Institute of Technology.

PARKER, J A; LESSARD, P A and SINSKEY, A J (unpublished). COGsensus, Improved determination of COG function through combined BLAST and RPS-BLAST search.

PARVEEZ, G K A; MASRI, M M; ZAINAL, A; MAJID, N A;YUNUS, A M M; FADILAH, H H; RASID, O and CHEAH, S C (2000). Transgenic oil palm: production and projection. *Biochemical Society Transactions, 28*: 969-972.

PURBA, A R; NOYER, J L; BAUDOUIN, L; PERRIER, X; HAMON, S and LAGODA, P J L (2000). A new aspect of genetic diversity of Indonesian oil palm (*Elaeis guineensis* Jacq.) revealed by isoenzyme and AFLP markers and its consequences for breeding. *Theoretical and Applied Genetics, 101*: 956-961.

RANCE, K A; MAYES, S; PRICE, Z; JACK, P L and CORLEY, R H V (2001). Quantitative trait loci for yield components in oil palm (*Elaeis guineensis* Jacq.). *Theoretical and Applied Genetics, 103*: 1302-1310.

SAMBANTHAMURTHI, R; PARMAN, S H and NOOR, M R M (1996). Oil palm (*Elaeis guineensis*) protoplasts: Isolation, culture and microcallus formation. *Plant Cell Tissue and Organ Culture, 46*: 35-41.

SAN, C T and SHAH, F H (2005). Differential gene expression and characterization of tissue-specific cDNA clones in oil palm using mRNA differential display. *Molecular Biology Reports, 32*: 227-235.

SINGH, R and CHEAH, S C (2000). Differential gene expression during flowering in the oil palm (*Elaeis guineensis*). *Plant Cell Reports, 19*: 804-809.

SUNDRAM, K; SAMBANTHAMURTHI, R and TAN, Y A (2003). Palm fruit chemistry and nutrition. *Asia Pacific J. Clinical Nutrition, 12*: 355-362.

TARMIZI, A H; NORJIHAN, M A and ZAITON, R (2004). Multiplication of oil palm suspension culture in a bench-top (2-litre) bioreactor. *J. Oil Palm Research Vol., 16*: 44.

TATENO, Y; IMANISHI, T; MIYAZAKI, S; FUKAMI-KOBAYASHI, K; SAITOU, N; SUGAWARA, H and GOJOBORI, T (2002). DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res., 30*: 27-30.

TATUSOV, R L; GALPERIN, M Y; NATALE, D A and KOONIN, E V (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res., 28*: 33-36.

WAHID, M B; ABDULLAH, S N A and HENSON, I E (2005). Oil palm - achievements and potential. *Plant Production Science, 8*: 288-297.

YU, J; HU, S N; WANG, J; WONG, G K S; LI, S G; LIU, B; DENG, Y J; DAI, L; ZHOU, Y; ZHANG, X Q; CAO, M L; LIU, J; SUN, J D; TANG, J B; CHEN, Y J; HUANG, X B; LIN, W; YE, C; TONG, W; CONG, L J; GENG, J N; HAN, Y J; LI, L; LI, W; HU, G Q; HUANG, X G; LI, W J; LI, J; LIU, Z W; LI, L; LIU, J P; QI, Q H; LIU, J S; LI, L; LI, T; WANG, X G; LU, H; WU, T T; ZHU, M; NI, P X; HAN, H; DONG, W; REN, X Y; FENG, X L; CUI, P; LI, X R; WANG, H; XU, X; ZHAI, W X; XU, Z; ZHANG, J S; HE, S J; ZHANG, J G; XU, J C; ZHANG, K L; ZHENG, X W; DONG, J H; ZENG, W Y; TAO, L; YE, J; TAN, J; REN, X D; CHEN, X W; HE, J; LIU, D F; TIAN, W; TIAN, C G; XIA, H G; BAO, Q Y; LI, G; GAO, H; CAO, T; WANG, J; ZHAO, W M; LI, P; CHEN, W; WANG, X D; ZHANG, Y; HU, J F; WANG, J; LIU, S; YANG, J; ZHANG, G Y; XIONG, Y Q; LI, Z J; MAO, L; ZHOU, C S; ZHU, Z; CHEN, R S; HAO, B L; ZHENG, W M; CHEN, S Y; GUO, W; LI, G J; LIU, S Q; TAO, M; WANG, J; ZHU, L H; YUAN, L P and YANG, H M (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp indica). *Science, 296*: 79-92.