GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Identification and isolation of full-length cDNA sequences by sequencing and analysis of expressed sequence tags from guarana (*Paullinia cupana*)

L.C. Figueirêdo[1], A.C. Faria-Campos[2], S. Astolfi-Filho[1] and J.L. Azevedo[3]

[1]Centro de Apoio Multidisciplinar, Universidade Federal do Amazonas, Manaus, AM, Brasil
[2]Laboratório de Universalização de Acesso à Internet, Departamento de Ciência da Computação, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
[3]Departamento de Genética, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, SP, Brasil

Corresponding author: L.C. Figueirêdo
E-mail: livio_cf@ufam.edu.br

Genet. Mol. Res. 10 (2): 1188-1199 (2011)
Received October 25, 2010
Accepted December 9, 2010
Published June 21, 2011
DOI 10.4238/vol10-2gmr1124

ABSTRACT. The current intense production of biological data, generated by sequencing techniques, has created an ever-growing volume of unanalyzed data. We reevaluated data produced by the guarana (*Paullinia cupana*) transcriptome sequencing project to identify cDNA clones with complete coding sequences (full-length clones) and complete sequences of genes of biotechnological interest, contributing to the knowledge of biological characteristics of this organism. We analyzed 15,490 ESTs of guarana in search of clones with complete coding regions. A total of 12,402 sequences were analyzed

using BLAST, and 4697 full-length clones were identified, responsible for the production of 2297 different proteins. Eighty-four clones were identified as full-length for N-methyltransferase and 18 were sequenced in both directions to obtain the complete genome sequence, and confirm the search made *in silico* for full-length clones. Phylogenetic analyses were made with the complete genome sequences of three clones, which showed only 0.017% dissimilarity; these are phylogenetically close to the caffeine synthase of *Theobroma cacao*. The search for full-length clones allowed the identification of numerous clones that had the complete coding region, demonstrating this to be an efficient and useful tool in the process of biological data mining. The sequencing of the complete coding region of identified full-length clones corroborated the data from the *in silico* search, strengthening its efficiency and utility.

Key words: Guarana; Transcriptome; Full-length cDNA; Caffeine synthase; Bioinformatics

INTRODUCTION

The high-throughput sequencing of genomes has revolutionized biology over the last ten years, generating a lot of new information that has changed dramatically our knowledge of hundreds of species, including us (Salzberg, 2007).

Research studies in molecular biology, mainly in genome sequencing, produce large amounts of data and a huge volume of information, and the techniques used in these studies, nowadays, are not capable of extracting all the possible knowledge about the study organisms; in this way, computer analysis techniques are used and they are indispensable to help researchers in the task of understanding these organisms (Luscombe et al., 2001).

The availability of sequenced genomes has increased the volume of information, making it more difficult to store and manipulate, requiring increasingly efficient computer resources and data analysis personnel with more training. The detailed analysis and correct interpretation of existing information have provided the right conditions for innovative research studies (Malone et al., 2006).

Unfortunately, for a gene stored in GenBank - NCBI (Benson et al., 2009), mainly if it is an incomplete and/or eukaryotic gene, there is a good chance that the amino acid sequence is wrong. And, depending on when the genome was sequenced and annotated, there is also a chance that its function is wrong (Salzberg, 2007).

It is against this backdrop that bioinformatics has emerged, a new field of knowledge, which studies and applies computer techniques and tools to organize and interpret information related to structures studied in molecular biology (Luscombe et al., 2001).

Among the models used in bioinformatics, the analysis of sequences is of great importance. The analysis of sequences is a subarea of bioinformatics that has grown a lot in recent years, as a response to the great increase of data generated from molecular biology. This trend will probably continue with new transformations that emerge from its integrative, quantitative and impressive nature (Fuchs, 2002). This growing proliferation of data from biological sequences made possible the development of many algorithms for the analysis and mining of

knowledge (Lu et al., 2008).

One of the most valuable resources of any species is a non-redundant group of mRNA transcripts from expressed genes that serve as a data resource and also as physical reagents, such as clones with full-length cDNA, which present the complete coding region and microarrays (Gilchrist et al., 2004).

The interpretation of growing information of raw sequences generated from modern techniques of genome sequencing faces multiple challenges, such as the analysis of genomic function and genome annotation. In fact, nearly 40% of plant genes code proteins with unknown functions. The functional characterization of these genes is one of the main challenges of modern biology. In this way, the availability of complete clones of cDNA can fill the gap created between the sequence information and biological knowledge. Clones with complete cDNA sequences facilitate the functional analysis of corresponding genes, allowing the manipulation of its expression in heterologous systems and the generation of a variety of labeled versions of native protein. Besides, the development of full-length cDNA sequences has the ability to enhance the quality of genome annotation (DeMarco and Verjovski-Almeida, 2008).

In recent years, the annotation of genes has been enormously increased by the integration of full-length cDNA sequences produced by the community. The importance of isolating full-length clones is based on "aggregated value", a characteristic not present in common expressed sequence tags (ESTs). Full-length cDNAs define the boundaries of transcriptional units and coding regions and enable identify the immediate upstream basal promoter and allow the characterization of sequence 5' and 3' of non-translated regions. Besides, they give a record of transcript diversity because of modifications in primary pre-mRNA transcripts, such as the use of an alternate promoter, alternative splicing, alternative polyadenylation, and RNA editing. On the other hand, cDNA libraries rich in full-length clones are a valuable tool for high throughput genome analysis (DeMarco and Verjovski-Almeida, 2008).

Full-length cDNAs are a very useful resource, not only for annotation and determination of initial regions of transcription, but also for functional analysis, mainly when analyzed under the context of genomic sequences (Nanjo et al., 2007).

Deciphered genomes are an abundant source of genes with biotechnological value and with unknown functions. Besides, modern methods in bioinformatics allow the inference of the probable metabolism of a living being from its gene sequences (*in silico* biochemistry). Computer simulation is emerging as a powerful tool to support research and industrial development. Considered an *in silico* analysis, this technique is located at the interface between theory and experiment, many times acting as a link between them. An enormous savings in resources, time and environmental damage have been accomplished with its use in many areas of human activity. In biology, medicine and pharmaceutics, this technique has been useful in comprehending biological processes, the action of drugs and diseases at atomic-molecular scales.

From this perspective, the constant study of data from transcriptome projects, genomics, metabolomics, etc., is of fundamental importance, so we can understand more about genes, metabolism, and even for new discoveries in biotechnology. And, since guarana (*Paullinia cupana* var. *sorbilis*), from the Amazon Region, is of economic, cultural and social importance, this species becomes an important candidate for new studies.

In this way, the objectives of the present study were to reevaluate data generated from the Guarana Transcriptome Sequencing Project, to identify full-length cDNA clones and to

increase knowledge of complete sequences of genes of biotechnology interest.

MATERIAL AND METHODS

Identification of full-length clones

The identification of full-length clones was done according to Faria-Campos et al. (2006), with some modifications. The number of codons calculated before the initial methionine was used for all ESTs and was obtained using a MySQL query to apply the search algorithm to the results of a BLAST search (tBLASTn) of protein sequences from the NR database against ESTs of *P. cupana*, since there are no sequences of proteins for *P. cupana* stored in public databases.

Calculation of number of codons before the initial methionine

$$((\text{sinit})/3) \rightarrow \text{qinit} = \text{number of codons before the initial methionine}$$

The initial positions of alignment of the orthologous protein and the selected ESTs were obtained from database using MySQL queries and used to calculate the number of codons before the initial methionine. The values obtained were grouped in three classes (values > 0, $= 0$ and < 0), and the number of events in each class was determined. Each event corresponds to a unique alignment for each pair of sequences.

Sequencing of selected cDNAs

Due to the importance of caffeine as a stimulant, two cDNAs related to its metabolic pathway, theobromine synthase and caffeine synthase, were selected for sequencing. Amino acid sequences of N-methyltransferase enzymes were selected, including the two stored in the UniProtKB database, due to its better characterization for a new identification of full-length clones, exclusively for these enzymes, but according to procedures previously described.

Sequencing

Through an SQL search, cDNAs from the guarana library with the best scores in alignments for the enzymes related to the metabolic pathway were selected for characterization.

After identification of the clones, they were recovered, grown, and plasmidial DNA was extracted for later sequencing. The sequencing was done using the M13 reverse promoters (5'-GGAAACAGCTATGACCATG-3') and direct (5'-GTTTTCCCAGTCACGAC-3') or T7 (5'-TAATACGACTCACTATAGGG-3') and SP6 (5'-ATTTAGGTGACACTATAG-3'), with DYEnamic™ ET Dye Terminators kit for MegaBACE™ (GE Healthcare), according to manufacturer instructions, with a MegaBACE™1000 automatic sequencer, in both directions (5' and 3') in four replicates. The files from the chromatogram were analyzed according to parameters previously described for base naming and trimming, removal of vectors and contaminant sequences, and grouping of contigs.

Sequence analysis through bioinformatics

The contigs were translated into amino acid sequences using the Transeq software (EMBOSS package, available at http://emboss.sourceforge.net), and globally aligned with the proteins theobromine synthase and caffeine synthase of several organisms using the Multialin software (available at http://multalin.toulouse.inra.fr/multalin/multalin.html). For a better characterization of global alignment, new alignments were done using the muscle algorithm (Edgar, 2004) with the SeaView software (Galtier et al., 1996; PBIL, 2009), and a phylogenetic tree was generated using the neighbor-joining method and a bootstrap of 1000 repetitions and by the method of maximum likelihood estimation with the PhyML algorithm from the Phylip package (Guindon and Gascuel, 2003; Felsenstein, 2008).

According to the alignment of the sequences, it is possible to corroborate the presence of the complete coding sequence when: 1) the initial methionine is in the expected position for the initiating codon ATG; 2) there is one stop codon compatible with the average length expected for that protein; 3) total alignment of the sequence or when, at least, the start and ending sequences are aligned with the homologous proteins.

RESULTS AND DISCUSSION

Identification of full-length clones

The identification of full-length cDNA clones takes into consideration the number of codons in the test sequence (EST - subject) before the initial methionine of the protein used in the calculation, and a clone must have a positive number of codons before the initial methionine selected.

From the 15,490 ESTs used in the study only 12,402 sequences could be used in the search for full-length clones, being aligned to proteins from the NR database under the established criteria. According to the analysis, an elevated number of events (46.15%) had a negative number of codons before the initial methionine, a result already expected for ESTs. Faria-Campos et al. (2006) demonstrated that these values can vary from 20 to 70% of the ESTs analyzed, depending on the source of organisms of the sequences used in the alignments.

However, the number of events with a positive or equal to zero number of codons before the initial methionine reached significant values, representing 37.87% of total ESTs used in the analysis, when compared to other studies using sequences of proteins from the organism itself analyzed, where 54% of the events had a positive number of codons before the initial methionine.

Of all the ESTs analyzed, 4697 presented positive or equal to zero number of codons before the initial methionine, thus identified as relative to full-length clones. These sequences represented 2297 different proteins in the NR database of NCBI.

Figure 1 represents a distribution of clones containing cDNAs with a positive number of codons before the initial methionine, where there is a bigger distribution between values from 0 and 14, representing 46.82% of the data. Or in other words, almost half the ESTs were identified as relative to possible full-length cDNA clones. Figure 2 presents a regression analysis demonstrating the distribution of all the values of a number of codons before the initial methionine related to the position of initial alignment with the protein, with the observance of

a prevalence of values close to the beginning of the protein sequence, with a tendency for the number of codons before the initial methionine to stay around 50.

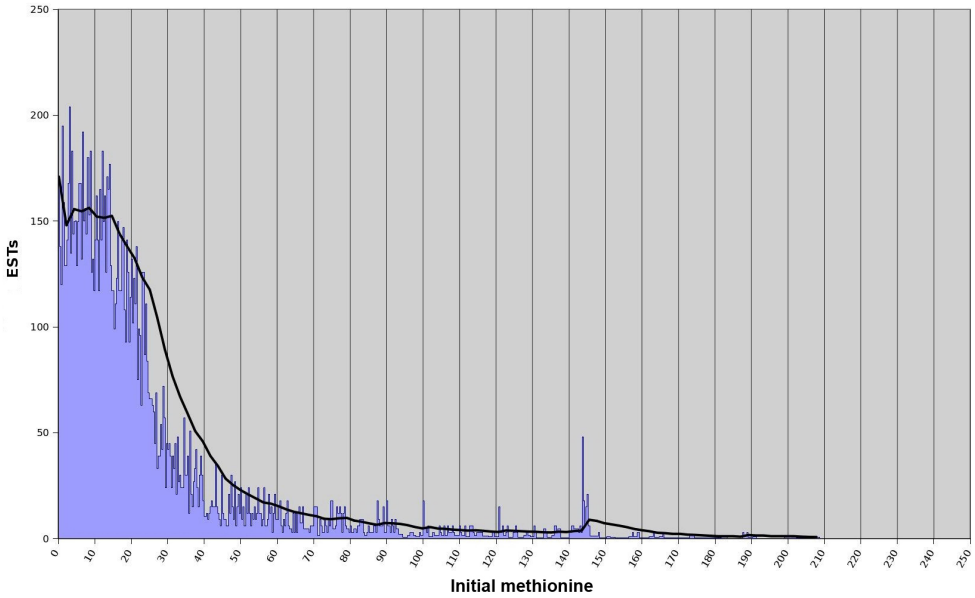

Figure 1. Number of expressed sequence tags (ESTs) distributed according to the number of codons before the initial methionine calculated as positives, obtained using BLAST, of proteins from the NR database against ESTs of guarana fruit.

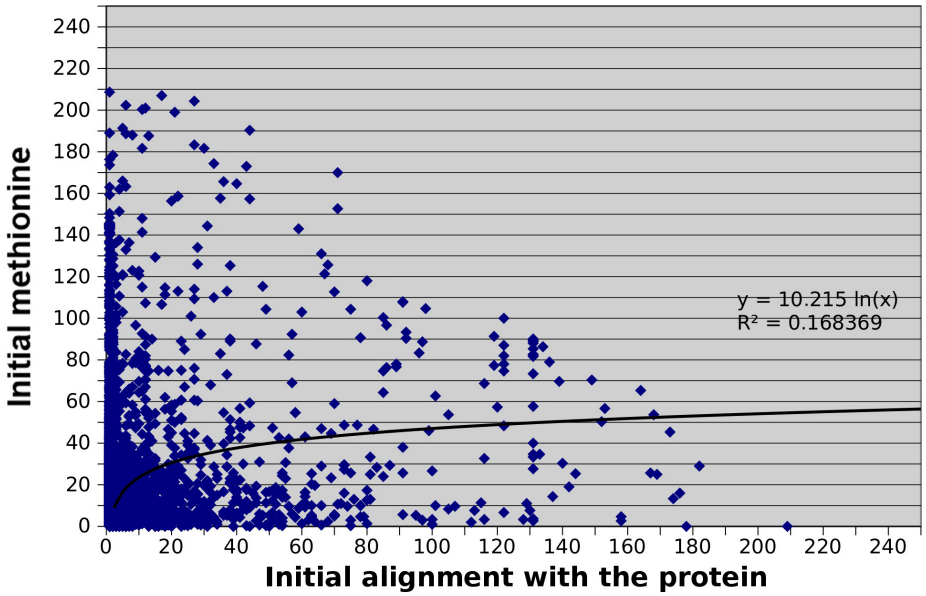

Figure 2. Distribution of values of positive codons, related to the alignment of the starting position of proteins from the NR database with ESTs from guarana fruit, obtained using BLAST.

For alignments where the initial position is smaller (closer to the beginning of the protein), the hit capacity varies less, regardless of the value of the positive number of codons before the initial methionine. However, when the beginning of the alignment is far from the beginning of the protein, the hit capacity is bigger when the value of the positive number of codons before the initial methionine is bigger.

Usually, attempts at genomic discovery through automatic predictions of genomic sequences and the use of ESTs in the identification of transcripts provide valuable information about genes present in a genome. However, these strategies present some limitations, since the predictions *ab initio* are far from accurate, and data from ESTs do not provide information about the complete coding region. cDNA clones containing the complete coding sequence constitute the best object of study to analyze a genomic transcript, providing information about the structure of the expressed gene and making easier posterior functional studies of these genes and their corresponding proteins.

Sequencing of the selected clones

The search for full-length cDNA clones using only N-methyltransferase enzymes resulted in the identification of 84 presumably full-length clones. From these, only 24 clones were selected for complete sequencing (Table 1). It is worth mentioning that, from these clones, only 18 had acceptable sequences after base calling to continue with the procedures of global alignment.

Table 1. cDNA clones of guarana fruit selected for complete sequencing-based transcriptome analysis.

Number	Prefix	Location*	Primer	Plate	Position
01	grn	IA	F01	0179g	E09
02**	grn	IA	F01	0170g	G09
03	grn	AM	F01	0128g	H03
04	grn	IA	F01	0018g	A03
05**	grn	EB	F01	0009s	D09
06	grn	AM	F01	0088g	F08
07	grn	PP	F01	0157g	A03
08	grn	AM	F01	0091g	D02
09**	grn	IA	F01	0167g	C04
10	grn	AM	F01	0187g	E12
11	grn	IA	F01	0171g	E05
12	grn	AM	F01	0092g	A09
13	grn	AM	F01	0186g	F01
14**	grn	AM	F01	0009g	D09
15	grn	PP	F01	0159g	D11
16	grn	AM	F01	0028g	H05
17	grn	MA	F01	0140g	C02
18**	grn	EB	F01	0021s	D11
19	grn	PS	F01	0227g	C11
20	grn	AM	F01	0227g	C11
21	grn	AM	F01	0200g	C04
22	grn	RR	F01	0200g	C04
23	grn	AM	F01	0086g	G07
24	grn	MA	F01	0147g	C01

*Acronym of labs participating in "Rede da Amazônia Legal de Pesquisas Genômicas - REALGENE": AM = UFAM; EB = EMBRAPA; IA = INPA; MA = UFMA; PP = UFPA (Belém); PS = UFPA (Santarém); RR = UFRR.
**Sequences were discarded after base calling.

Complete coding sequence of N-methyltransferase

Due to the sequencing capacity of MegaBACE™1000, which has produced sequences with sizes smaller than 700 bp in most of the runs, only three clones had their cDNA coding sequences completely sequenced: clones grn06, grn21 and grn22 (Figure 3). Since the coding sequences for theobromine synthase and caffeine synthase have nearly 1200 bp, even when sequencing genes in both ways, direct and reverse, frequently it was not possible to generate contigs with the sequences obtained, since there was no overlapping of the extremities, a fact that made the process even more difficult, causing it to be repeated four times.

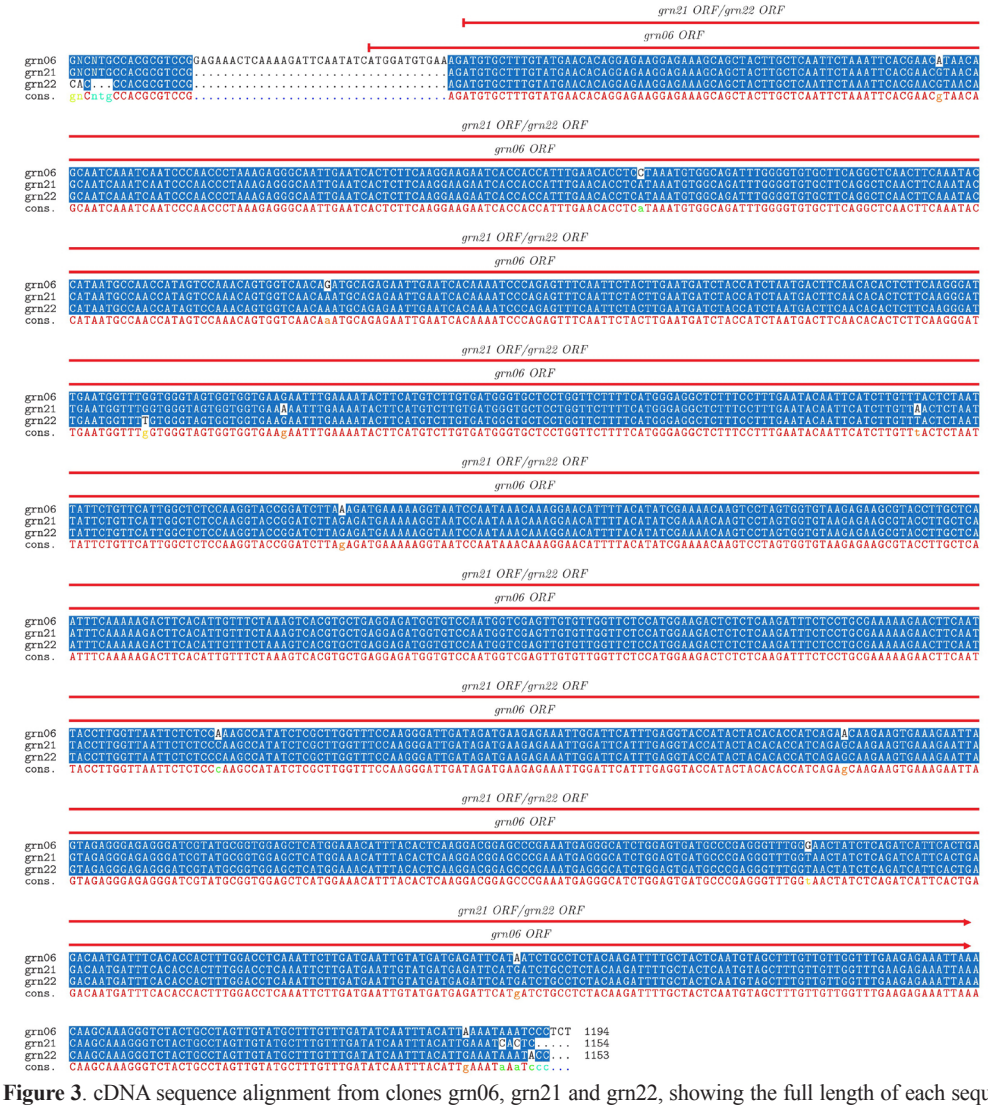

Figure 3. cDNA sequence alignment from clones grn06, grn21 and grn22, showing the full length of each sequence. Nucleotide identity among sequences is highlighted in blue; arrows in red point to possible open reading frames (ORF).

Although the sequences were incomplete, it was possible to confirm that the clones had the complete cDNA coding sequences, because they presented global alignments according to expectations, that is, the end sequences of their *in silico* translated proteins (aminoterminal and carboxyterminal) matching the end sequences of reference proteins theobromine synthase and caffeine synthase (Figure 4).

Figure 4. Similarity profile among translated sequences of 18 full-length cDNAs and 25 sequences from theobromine synthase and caffeine synthase proteins.

There is a great similarity among the three predicted sequences of amino acids of N-methyltransferase: there is only 0.017% of dissimilarity (Figure 5). Clones grn21 and grn22 present almost equal predicted sequences of amino acids, with the exception of two positions: position 115, which has a leucine (L) in clone grn21, and a phenylalanine (F) in clone grn22, and position 122, which has a lysine (K) in clone grn21, and a glutamic acid (E) in clone grn22. However, clone grn06 has polymorphism in positions 1 to 8, 27, 55, 82, 151, 166, 244, and 341, when compared to clones grn21 and grn22, indicating a bigger differentiation of this protein as compared to the other two.

Figure 5. Divergences among the predicted amino acid sequences of N-methyltransferases completely sequenced grn06, grn21 and grn22. Divergent amino acids are highlighted in red.

A phylogenetic tree based on the genetic distance among the predicted guarana N-methyltransferases, 13 theobromine synthases and 10 caffeine synthases is presented in Figure 6. The three proteins of guarana form an individual group, where grn21 and grn22 seem to be closer in relation to grn06. This group is phylogenetically closer to caffeine synthase of *T. cacao* (Q2HXL8 - EMBL:BAE79730), a result also seen in other studies of N-methyltransferase of guarana (Ângelo et al., 2008; Freitas, 2009). However, there was no group formed exclusively by caffeine synthase or theobromine synthase. These two enzymes seem to share the same phylogenetic profile, which was also observed in several studies (Uefuji et al., 2003; Yoneyama et al., 2006; McCarthy and McCarthy, 2007; Ângelo et al., 2008; Freitas, 2009; Ishida et al., 2009).

The search for full-length cDNA clones allowed the identification of a significant number of clones with the complete coding region, demonstrating it to be a very efficient and useful tool in the process of biological data mining. The sequencing of all the coding region of full-length identified clones corroborated data from the *in silico* search, strengthening its efficiency and utility.

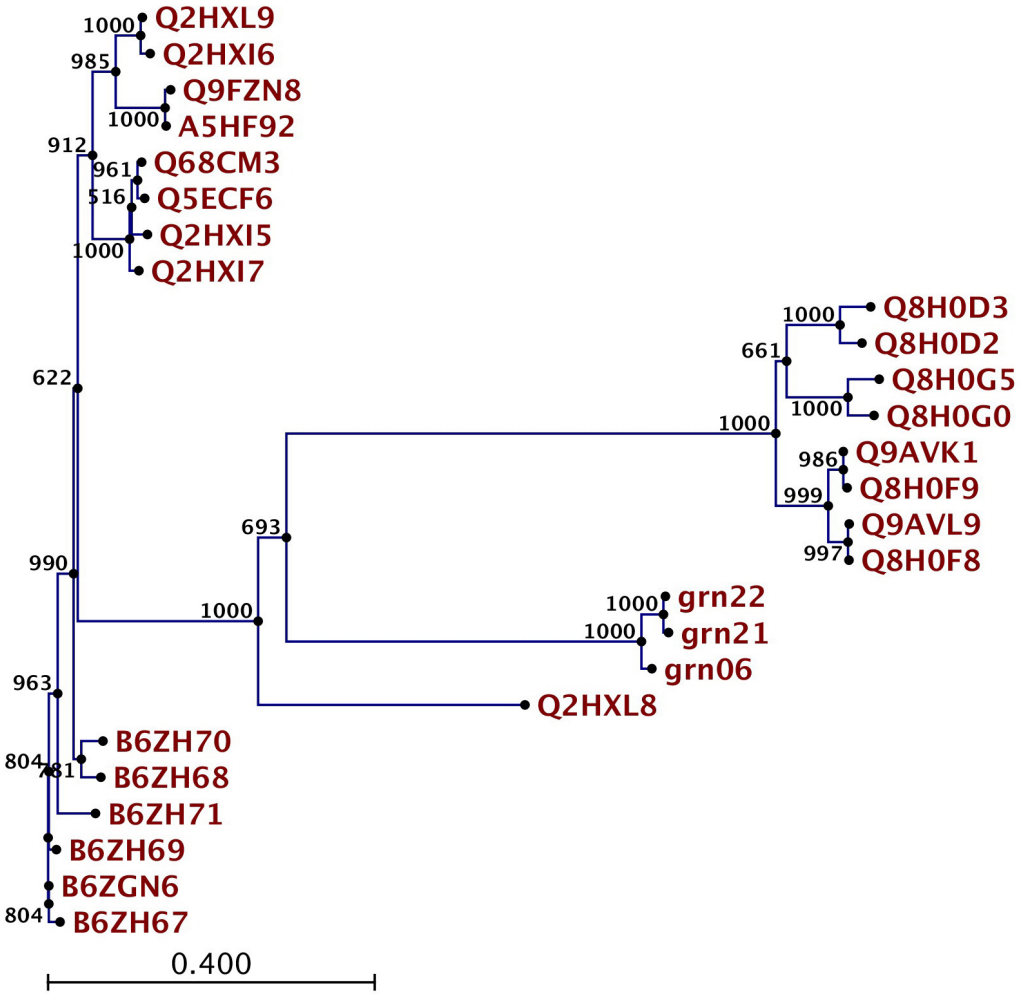

Figure 6. Phylogenetic tree of predicted guarana N-methyransferases, 13 theobromine synthases (B6ZGN6, B6ZH67, B6ZH68, B6ZH69, B6ZH70, B6ZH71, Q2HXI5, Q2HXI6, Q2HXI7, Q2HXL9, Q8H0G0, Q8H0G5, Q9AVK1) and 10 caffeine synthases (A5HF92, Q2HXL8, Q5ECF6, Q68CM3, Q8H0D2, Q8H0D3, Q8H0F8, Q8H0F9, Q9AVL9, Q9FZN8).

ACKNOWLEDGMENTS

We thank the Biotechnology Division of the Multi-Disciplinary Support Center and Bioinformatics Laboratory of Universidade Federal do Amazonas, for the use of its facilities; E.N. Assunção by sequencing technical support; the National Council for Scientific and Technological Development (CNPq) for providing grants to L.C. Figueirêdo, and to the Multi-Institutional Program of Post-Graduation in Biotechnology of Universidade Federal do Amazonas for supporting this project.

REFERENCES

Ângelo PCS, Nunes-Silva CG, Brígido MM, Azevedo JSN, et al. (2008). Guarana (*Paullinia cupana* var. *sorbilis*), an anciently consumed stimulant from the Amazon rain forest: the seeded-fruit transcriptome. *Plant Cell Rep.* 27: 117-124.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. (2009). GenBank. *Nucleic Acids Res.* 37: D26-D31.

DeMarco R and Verjovski-Almeida S (2008). Expressed Sequence Tags (ESTs) and Gene Discovery: *Schistosoma mansoni*. In: Bioinformatics in Tropical Disease Research. Available at [http://ncbi.nlm.nih.gov/bookshelf/br.fugi?book=bioinfo]. Accessed May 20, 2010.

Edgar RC (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5: 113.

Faria-Campos AC, Moratelli FS, Mendes IK, Ortolani PL, et al. (2006). Production of full-length cDNA sequences by sequencing and analysis of expressed sequence tags from *Schistosoma mansoni*. *Mem. Inst. Oswaldo Cruz* 101 (Suppl 1): 161-165.

Felsenstein J (2008). Phylip (Phylogeny Inference Package) Version 3.68. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle. Available at [http://evolution.genetics.washington.edu/phylip.html]. Accessed December 18, 2009.

Freitas DV (2009). Caracterização Genética e Molecular do Guaranazeiro (*Paullinia cupana* "*sorbilis*"). Doctoral thesis, Universidade Federal do Amazonas, Manaus.

Fuchs R (2002). From sequence to biology: the impact on bioinformatics. *Bioinformatics* 18: 505-506.

Galtier N, Gouy M and Gautier C (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.* 12: 543-548.

Gilchrist MJ, Zorn AM, Voigt J, Smith JC, et al. (2004). Defining a large set of full-length clones from a *Xenopus tropicalis* EST project. *Dev. Biol.* 271: 498-516.

Guindon S and Gascuel O (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52: 696-704.

Ishida M, Kitao N, Mizuno K, Tanikawa N, et al. (2009). Occurrence of theobromine synthase genes in purine alkaloid-free species of *Camellia* plants. *Planta* 229: 559-568.

Lu G, Zhang S and Fang X (2008). An improved string composition method for sequence comparison *BMC Bioinformatics* 9 (Suppl 6): S15.

Luscombe NM, Greenbaum D and Gerstein M (2001). What is Bioinformatics? An Introduction and Overview. Yearbook of Medical Informatics, 83-99. Available at [http://www.ebi.ac.uk/luscombe/publications.html]. Accessed June 22, 2009.

Malone G, Zimmer PD, Meneghello GE, Binneck E, et al. (2006). Gene prospection in cDNA libraries. *Rev. Bras. Agrocienc.* 12: 07-13.

McCarthy AA and McCarthy JG (2007). The structure of two N-methyltransferases from the caffeine biosynthetic pathway. *Plant Physiol.* 144: 879-889.

Nanjo T, Sakurai T, Totoki Y, Toyoda A, et al. (2007). Functional annotation of 19,841 *Populus nigra* full-length enriched cDNA clones. *BMC Genomics* 8: 448.

Pôle Bioinformatique Lyonnais (PBIL) (2009). SeaView Version 4.0. Available at [http://pbil.univ-lyon1.fr/software/seaview.html]. Accessed August 10, 2009.

Salzberg SL (2007). Genome re-annotation: a wiki solution? *Genome Biol.* 8: 102.

Uefuji H, Ogita S, Yamaguchi Y, Koizumi N, et al. (2003). Molecular cloning and functional characterization of three distinct N-methyltransferases involved in the caffeine biosynthetic pathway in coffee plants. *Plant Physiol.* 132: 372-380.

Yoneyama N, Morimoto H, Ye CX, Ashihara H, et al. (2006). Substrate specificity of N-methyltransferase involved in purine alkaloids synthesis is dependent upon one amino acid residue of the enzyme. *Mol. Genet. Genomics* 275: 125-135.