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

