

TSD: ฐานข้อมูลของการจัดเก็บข้อมูลชีววิทยาของอ้อย และเครื่องมือที่เป็นคอมพิวเตอร์ซอฟต์แวร์ผ่านเว็บ

TSD: a database of sugarcane biological data collection and web-based computer software tools

**ปิยรัตน์ พลยะเรศ¹, พุษดี ศิริแสงตระกูล², ศิษฏาท ทองลิมา³, ชุตีพงศ์ อรรคแสง⁴
และ ณัฐปภัสร์ ตันติสุขวิชญ์^{1*}**

**Piyarat Ponyared¹, Pusadee Seresangtakul², Sissades Tongsim³, Chutipong Akkasaeng⁴
and Nathapat Tantisuwichwong^{1*}**

บทคัดย่อ: อ้อยเป็นพืชเศรษฐกิจที่สำคัญของไทย ที่มียังชีพของเราให้ความสนใจศึกษาชีววิทยา พันธุศาสตร์ และชีวโมเลกุลของอ้อย และได้ผลิตข้อมูลจำนวนมากจากสองแนวทาง คือ งานทดลองที่อาศัยคอมพิวเตอร์ และงานทดลองที่อาศัยสารเคมี ข้อมูลเหล่านี้ ได้แก่ เครื่องหมายดีเอ็นเอชนิด EST-derived SSR เครื่องหมายดีเอ็นเอชนิด EST-derived SNP เครื่องหมายโปรตีนจากการระบุด้วยเทคนิค 2D-gel และ LC-MS ข้อมูลที่มีความหลากหลายเหล่านี้จำเป็นต้องมีการบูรณาการเพื่อช่วยอธิบายถึงความสัมพันธ์ระหว่างกัน ดังนั้นเพื่อให้บรรลุถึงเป้าหมายดังกล่าว ที่มียังชีพของเราจึงได้สร้างระบบชีวสารสนเทศสำหรับฐานข้อมูลทางชีววิทยาของอ้อย ให้ชื่อว่า Thai sugarcane database (TSD) ซึ่งบทความนี้จะนำเสนอส่วนของ TSD ที่ใช้จัดเก็บเครื่องหมายดีเอ็นเอชนิด EST-derived SSR และเครื่องหมายดีเอ็นเอชนิด EST-derived SNP นอกจากนี้ยังมีการติดตั้งเครื่องมือที่เป็นคอมพิวเตอร์ซอฟต์แวร์ใช้งานผ่านทางเว็บ มาพร้อมกับ TSD เครื่องมือทางคอมพิวเตอร์ซอฟต์แวร์นี้จะทำให้ผู้ใช้เข้าสู่ TSD และเข้าแหล่งข้อมูลอื่นผ่านทางลิงค์ การติดตั้งเครื่องมือนี้ช่วยรับรองการบูรณาการของการใช้ประโยชน์ของ TSD กับฐานข้อมูลสาธารณะ เช่น GenBank ระบบชีวสารสนเทศที่ประกอบด้วยฐานข้อมูลและเครื่องมือคอมพิวเตอร์ที่ใช้งานผ่านทางเว็บ จึงมีคุณค่าต่อการวิเคราะห์ข้อมูลชีววิทยาจากการทดลองที่อาศัยสารเคมีของกลุ่มวิจัยที่มีการสะสมปริมาณงานเพิ่มมากขึ้น และเป็นแนวทางให้ทีมวิจัยกำหนดงานวิจัยในอนาคตทั้งด้านการทดลองด้วยคอมพิวเตอร์และการทดลองที่อาศัยสารเคมี และทำให้เกิดองค์รู้ในด้านชีววิทยาของอ้อยอย่างลึกซึ้ง

คำสำคัญ: Sugarcane, Expressed Sequence Tags (ESTs), Bioinformatics and Database

¹ ภาควิชาชีววิทยา คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น, 40002

Department of Biology, Faculty of Science, Khon Kaen University, 40002

² ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น, 40002

Department of Computer Science, Faculty of Science, Khon Kaen University, 40002

³ สถาบันจีโนม ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ กระทรวงวิทยาศาสตร์และเทคโนโลยี, 12120

Genome Institute, National Center for Genetic Engineering and Biotechnology, Ministry of Science and Technology of Thailand, 12120

⁴ ภาควิชาพืชศาสตร์และทรัพยากรการเกษตร คณะเกษตรศาสตร์ มหาวิทยาลัยขอนแก่น, 40002

Department of Plant Science and Agricultural Resources, Faculty of Agriculture, Khon Kaen University, 40002.

* Corresponding author: nathapat2012@gmail.com

ABSTRACT: Sugarcane is an economic important crop of Thailand. Physiology, genetics and molecular biology of sugarcane are of our interest. A wealth of data has been produced by both computational-based and chemical reagent-based experimental works such as EST-derived SSR markers, EST-derived SNP markers, protein markers identified by 2D-gel and LC-MS. These diverse data would be integrated to describe the meaningful relationships in biological processes of sugarcane. To achieve this goal, a bioinformatic system for sugarcane biology database has been established namely Thai sugarcane database (TSD). A section of TSD database collecting EST-derived SSR markers and EST-derived SNP markers is presented in this article. Another feature, web-based application is equipped with TSD database allowing querying of TSD database and several links. This feature guarantees a high integration with other public database such as GenBank. Bioinformatic system combining the database and web-based applications are therefore valuable tool for the analysis of the increasing quantities of biological data resulting from chemical reagent laboratory works of our group. It will also guide us to design future computational-based and chemical reagent-based experimental works and gain more in-depth knowledge of sugarcane biology.

Keywords: Sugarcane, Expressed Sequence Tags (ESTs), Bioinformatics and Database

Introduction

Sugarcane (*Saccharum* spp.) is the one of tropical and subtropical economic crops. It is cultivated in more than 24 million hectares in the world, producing up to 1.68 billion metric tons of crushable stems. There are approximately 1.3 million hectares widely grown in the middle, north and northeast of Thailand (Office of Agriculture Economics, 2011). It is generally used to produce sugar, accounting for almost two thirds of the world's production (Moore, 1995).

Commercial sugarcane cultivars are one of the most complex plant genomes. Yields of sugarcane crop have increased through traditional breeding strategies over the last century. However, the *Saccharum* hybrid cultivars have reduced in genetic variability, caused by recent speciation (Grivet and Arruda, 2002; Glaz, 2003). To overcome this problem, molecular biology tools are applied into sugarcane breeding programs and can contribute to the production of improved cultivars (Butterfield et al., 2001; Menossi et al., 2008; Ming et al., 2006)

Recently, large amounts of molecular biological data including genomics, transcriptomics, and proteomics of sugarcane are gradually re-

leased in public databases (Vettore et al., 2003).

To fully explore these massive data, bioinformatics have been developed for biological knowledge analyse and management (Goodman 2002; Wu et al., 2004). One of our interest is sugarcane transcriptomics, both full-length cDNA and EST. Over 200,000 sugarcane ESTs have been stored in Genbank database. Our group has retrieved those sugarcane ESTs and a computational pipeline has been constructed for sugarcane EST analysis (Ponyared et al., 2008; Ponyared and Tantisuwichwong, 2009; Ponyared et al., 2009). SSR and SNP markers have computationally been identified in the EST (Ponyared, 2009). The markers have been used in PCR to detect genetic variability among 15 sugarcane cultivars that commonly grow in Thailand (Tantisuwichwong et al., 2009). Sequences of the markers were read and revealed genetic polymorphism (Jariyajirawattana et al., 2010). The EST analysis pipeline and the EST-derived DNA markers are the valuable tools and resources concerning sugarcane molecular biology research. As increasing bioinformatic tools and molecular data, we set forth to establish Thai sugarcane database (TSD) for providing platform for sugarcane EST analysis pipeline and recording EST, EST-derived DNA markers,

marker sequences. The aim of this database to (1) develop an organized and integrated resource for sugarcane molecular data and (2) to develop computer tools for sequence retrieval and analysis. In this article, we describe the structure and content of the database and reveal the database access utility and tools.

Materials and Methods

Data sources

Molecular biological data collected in the database were sugarcane ESTs, retrieved from GenBank (<http://www.ncbi.nlm.nih.gov/dbEST>), and data resulting from sugarcane EST analysis pipeline (Ponyared et al., 2008; Tantisuwichwong et al., 2009; Ponyared and Tantisuwichwong, 2009; Ponyared et al., 2009). These data were stored in text file (.txt, .seq), image file (.png, .pdf) and excel file format.

Database architecture

The Window 7 was used as server. The TSD database was constructed using open source technologies, Perl script, Python, MySQL database management and PHP-based web interfaces. Sugarcane ESTs, EST-derived DNA markers were imported into a MySQL database as schema shown in **Figure 1**. Scripting were performed either Perl script or Python using MySQL connectivity. Sugarcane EST analysis pipeline (Ponyared, 2009) was used to interact with the database. The web interface runs on Apache web server. The bioinformatics system is designed, based on a three-tier architecture model: application, middle and database layer. **Figure 1** depicts the overall architectural design of the system. On the application layer, PHP-based scripts were used as a common gateway interface (CGI) on the Apache web server to render the graphical web-interface.

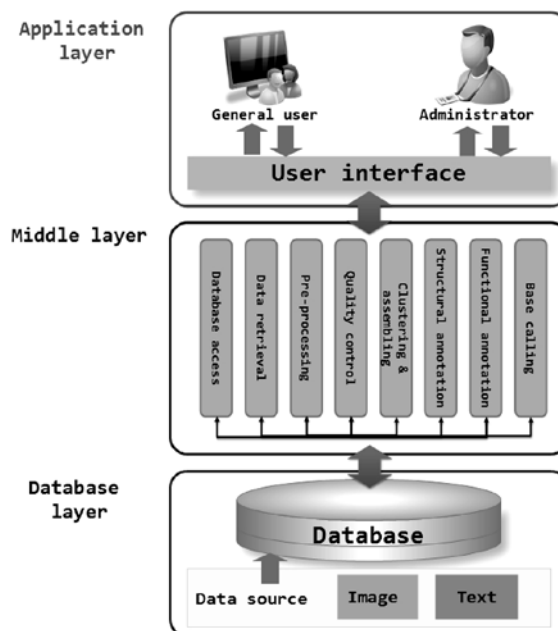


Figure 1 The overview of TSD system architecture

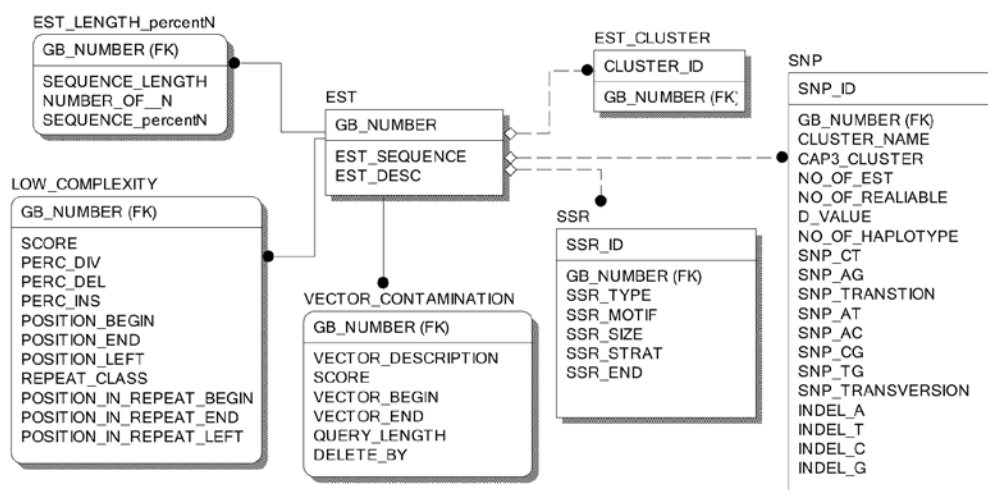


Figure 2 Schema of the relational model of TSD

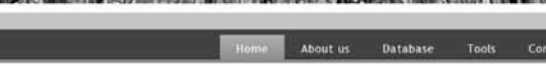
Database system design and development

The TSD database was constructed based on an entity-relational (ER) model (Figure 2). The TSD has 7 entities includes both strong and weak entities. The strong entities include EST, EST_cluster, SSR and SNP. The weak entities that related with strong entity include EST_Length_percentN, low_complexity and vector_contamination (Ponyared *et al.*, 2009).

Web-based application

The information obtained from the execution of the sugarcane EST analysis pipeline is stored in the TSD. The TSD database provides a data warehouse useful for further investigations. All kinds of information collected in the database can support biologically interesting analyses both to check the quality of the experimental results and to define structural and functional features of the

data. For this purpose the database can be queried through SQL calls implemented in a suitable PHP-based interface. We provide a pre-defined web based query system to support non expert users (researchers) as shown in figure 3. Different views are possible. Database browser with web-based application allows users to formulate flexible queries considering three different aspects. Firstly, users can submit data in FASTA format or excel format into TSD. The submitted data are collected in specific folder and the format of submitted data will be rechecked by administrator. Secondly, users can retrieve the information of sugarcane stored in TSD. Thirdly, search filter designed by web-based application facilitates ease of information retrieval system by entering EST sequence number, vector name, type of SSR and SNP etc.



[Home](#) [About us](#) [Database](#) [Tools](#) [Contact Us](#)

Menu

- * Forum & Sign In
- * People
- * Publication
- * Downloads & Uploads
- * Help & Faq
- * Link

Welcome to Thai Sugarcane Database



Sugarcane is an economic important crop of Thailand. The crop is grown predominantly under rain fed conditions and may suffer from moisture stress during any stage of its 10 to 12-month crop cycle. Therefore, drought is an important stress that reduces the crop production and subsequent sugar yields. Physiology, genetics and molecular biology of sugarcane under water-deficit stress are of our interest. A wealth of data has been produced

Menu

- * Forum & Sign In
- * People
- * Publication
- * Downloads & Uploads
- * Help & Faq
- * Link

Downloads and Uploads

Download data

Select

[Upload data](#)

Select

EST

EST

EST length 6141

EST low complexity

EST vector contamination

EST cluster

EST

submit

Database search: **EST**

Keyword
Search

GB_NUMBER	EST sequence	EST Description
CA184650	CTGACATTCGAGGCCACCGGACCAACCGGCGAGATCGGAGGGCGCTCTCGCCCTCGTCACTACCTGTC AGGGAGCACCAGCGCCGCTCGATGCGCGGCTGCCAGGGGTGACACGCCCGCAGCTCTCAGAGATGGAA CGATATCCATGAGAGAGGTGCTGATACASAGCAAGAGAGAGCTGGATCTGCTGAGGGAGAGAGATCCGTGT GTCTGTCTGTGTTCATGACGCGACAGCTTCCTGCTCTGTGATGATACATTAAGGCGTGAAGATCA CAGGGAGATTGGAGCGATGAAGCATACCTGCTGATTCCTGCTGATTGGATGGATCTTTGTTCGACAAAG CTAACATCATGCTCTCAAGAAGAGGATCTTACCACTGACAGCTGATGCTTCGCAATATTCGGGCGAGATGTG TGAGGGGCTGAGACATGACGAGCTTTGATCTCGATATGCCCATAGCATTTGAGCTGTGTATGTT CTTATAAGCTGGGCTAAAGGACAGCACTGTGGCACTTTAAATGGATTTCGAGATGCAAGGCGCTGCAAA AAGGAAGATCCCGTTCTGTCTGAGGCATACAGTTGACGAGATGGGCGCTGCAAGATGGCTGCGGCT TATCGGGGCGCTGATTGGAGCTGGCAGCATGCTTTTGTAGAGAGAGCAATTTGGCTTTAGATG CCCTTTTGTATTAAGTCAAAATTCACATTTGATGCTTGGGATTTGGAGATTTGACCTTCTGCT CACAGCATTTATGGGGCGCAGTGAGCATCTTTATGATGATTTGATTTG	-g335122625g(CA184650) 1(CA184650) SCLRTS173165F09 g 373 Saccharum officinense cDNA clone SCLRTS173165F09 5' mRNA sequence

Database search: **Vector Contamination**

Keyword
Search

GB_NUMBER	VECTOR_DESCRIPTION	VECTOR_BEGIN	VECTOR_END	QUERY_LENGTH	DELETE_BY	VECTOR_SCORE
CA216648		1	788	788		0
CA216649		1	718	718		0
CA216650		1	710	710		0
CA216653		1	777	777		0
CA216654		1	654	654		0
CA216655		1	667	667		0
CA216656		1	634	634		0

Figure 3 A screenshot of the EST Browser. ESTs were retrieved by sequence features collected in the input step. Some relevant data were visualized by web-based application tools.

Conclusions and Discussion

We designed the presented database system and user interface to perform an exhaustive analysis on EST datasets. Moreover, we implemented database to reduce execution time of the different steps required for a complete analysis by means of distributed processing. The database structure was design to collect and manage data from EST

analysis pipeline and integrate the relationships between EST raw data and the result of EST analysis pipeline. The TSD database system provided valuable tools for the utilization of the increasing quantities of data resulting from experiment works of our research group and provides the basis for the application of this bioinformatics system to perform our future works.

Acknowledgements

We thank Department of Biology and Computer Science, Faculty of Science, Khon Kaen University for computer facilities. Piyarat Ponyared is a recipient of studentships from The Research Professional Development Project under the Science Achievement Scholarship of Thailand (SAST).

References

- Butterfield, M. K., A. D'Hont, and N. Berding. 2001. The sugarcane genome: a synthesis of current understanding, and lessons for breeding and biotechnology. In *Proceedings of the South African Sugar Technologists Association (SASTA '01)* 75:1-5. Durban, South Africa, July-August 2001.
- Glaz, B. 2003. Integrated crop management for sustainable sugarcane production: recent advances. *Int. Sugar J.* 105:175-186.
- Goodman, N. 2002. Biological data becomes computer literate: new advances in bioinformatics. *Curr. Opin. Biotechnol.* 13:68-71.
- Grivet, L., and P. Arruda. 2002. Sugarcane genomics: depicting the complex genome of an important tropical crop. *Curr. Opin. Plant Biol.* 5(2):122-127.
- Jariyajirawattana, D., P. Ponyared, C. Akkasaeng, S. Roytrakul, T. Sansayavichai, W. Pongragdee, and N. Tantisuwichwong. 2010. Characterization and validation of molecular markers developed from expressed genes of sugarcane. GenBank, the sequences will be publicly available in October, 21, 2013.
- Menossi, M., M. C. Silva-Filho, M. Vincentz, M-A. Van-Sluys, and G. M. Souza. 2008. Sugarcane functional genomics: gene discovery for agronomic trait development. *Int. J. Plant Genomics*: 1-11.
- Ming, R., P. H. Moore, K.K. Wu, A. D'hont, J. C. Glaszmann, T. L. Tew, T. E. Mirkov, J. da Silva, J. Jifon, M. Rai, R. J. Schnell, S. M. Brumbley, P. Lakshmanan, J. C. Comstock, and A. H. Paterson. 2006. Sugarcane improvement through breeding and biotechnology. *Plant Breeding Reviews* 27:15-118.
- Moore, P.H. 1995. Temporal and spatial regulation of sucrose accumulation in the sugarcane stem. *Aust. J. Plant Physiol.* 22(4):661-679.
- Office of Agricultural Economics. 2011. Agricultural statistics of Thailand 2011. Ministry of Agriculture and Cooperatives, Bangkok, Thailand.
- Ponyared, P. 2009. Data mining of expressed sequence tags (ESTs) database for development of SSR and SNP markers. M. Sc. Thesis. Khon Kaen University, Khon Kaen.
- Ponyared, P., D. Jariyajirawattana, C. Akkasaeng, and N. Tantisuwichwong. 2008. Data mining of Expressed sequence tags (ESTs) for development of SSR markers in sugarcane. P. 17-18. In: *The 2nd Botanical Conference of Thailand* 26-28 March 2008. Department of Biology, Faculty of Science, Khon Kaen University.
- Ponyared, P. and N. Tantisuwichwong. 2009. Determination of nucleotide quality and redundancy within ESTs derived from sugarcane mature stalk. P. 161-171. In: *The 2nd on Graduate Research* 6 September 2009. Chandrakasem Rajabhat University, Bangkok.
- Ponyared, P., T. Remsungnen, N. Arch-int, C. Akkasaeng, and N. Tantisuwichwong. 2009. Computational analysis of sugarcane ESTs for high-quality clusters and SSR mining. *Thai. J. Genet.* 2(2): 131-144.
- Tantisuwichwong, N., P. Ponyared, D. Jariyajirawattana, T. Chavalit, T. Remsungnen, T. Sansayavichai, and C. Akkasaeng. 2009. Evaluation and validation of SSR marker identified by computer software. In: *The 16th National Genetics Conference* 25-27 March 2009 Department of Science and Technology, Kasetsart University, Pathum Tani. p. 27-31.
- Vettore, A. L., F. R. da Silva, E. L. Kemper, G. M. Souza, A. M. da Silva, M.I.T. Ferro, F. Henrique-Silva, E. A. Giglioti, M. V. F. Lemos, L. L. Coutinho, M. P. Nobrega, H. Carrer, S. C. Franca, B. J. M. M. H. S. Goldman, S. L. Gomes, L. R. Nunes, L. E. A. Camargo, W. J. Siqueira, M.-A. V. Sluys, O. H. Thiemann, E. E. Kuramae, R. V. Santelli, C. L. Marino, M. L. P. N. Targon, J. A. Ferro, H. C. S. Silveira, D. C. Marini, E. G. M. Lemos, C. B. Monteiro-Vitorello, J. H. M. Tambor, D. M. Carraro, P. G. Roberto, V. G. Martins, G. H. Goldman, R. C. de Oliveira, D. Truffi, C. A. Colombo, M. Rossi, P. G. de Araujo, S. A. Sculaccio, A. Angella, M. M. A. Lima, d. R. J. V. E, F. Siviero, V. E. Coscrato, M. A. Machado, L. Grivet, S. M. Z. D. Mauro, F. G. Nobrega, C. F. M. Menck, M. D. V. Braga, G. P. Telles, F. A. A. Cara G. Pedrosa, J. Meidanis, and P. Arruda. 2003. Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Res.* 13: 2725-2735.
- Wu, C.H., H. Huang, A. Nikolskaya, Z. Hu, and W.C.Barker. 2004. The iProClass integrated database for protein functional analysis. *Comput. Biol. Chem.* 28(1):87-96.