Development of a Tool for the Analysis of Plant Stress Proteins

Muhammad Ali[*], Rana Rehan Khalid, Muhammad Nawaz, Nauman Qamar

Department of Bioinformatics & Biotechnology
Government College University Faisalabad
Allama Iqbal Road – 38000, Pakistan
E-mails: Aly.binm@gmail.com, ray.binm@gmail.com, M.nawaz9@gmail.com, Rabia.nauman26@gmail.com

[*]*Corresponding author*

Received: June 06, 2011 **Accepted: December 27, 2011**

Published: January 31, 2012

Abstract: *The recent explosion of biological data and the accompaniment proliferation of distributed databases make it challenging for biologists and bioinformatists to discover the best and concise data resources for their needs, and the most efficient way to access and use them. For the biologist, running bioinformatics analyses involve a time-consuming management of data and tools. Users need support to organize their work, retrieve parameters and reproduce their analyses. They also need to be able to combine their analytic tools using a safe data flow software mechanism. Finally we have designed a system, Stress Gene catalog, to provide a flexible and usable web environment for defining and running bioinformatics analyses for the ease of researchers working in plants sciences. It embeds simple yet powerful data management features that allow the user to reproduce analyses and to combine tools using an adobe flex tool. Rice Stress gene catalog can also act as a front end to provide a unified view of already-existing rice stress gene families and their protein members along with their FASTA sequences. Users can analyze genomic and proteomic data by using the tools that has been integrated in the software (tools for alignments, multiple sequence comparison and to compare a novel sequence with those contained in nucleotide and protein databases).*

Keywords: *Stress gene catalog, Genomic and proteomic data, Manipulations, Programming logics, Action script.*

Introduction

A prominent goal of current plant genomics research is to establish an expandable platform for global analysis and classification of plant gene family space. A large fraction of genes in plant genomes are the product of novel gene creation and duplication processes that have occurred within plants over their 500-million-year history. Gene classifications that attempt to capture all of eukaryote diversity typically provide a poor representation of plant gene sets [16]. With many additional genome and transcriptome projects being initiated, and more than a dozen plant genomes scheduled for completion over the last two years, there is a need for flexible, gene family-focused tool that provide rich toolsets for analyses of plant genomes [14].

The implementation of large-scale data analysis initiatives, the volume of information in terms of biological data availability is overwhelming, as reflected by the hundreds of databases and web servers [2]. These resources have a great value to bioinformatician for proposing novel hypotheses and delineating further research.

Rice Stress Gene Catalog tool is a global classification of genes from rice plant genome. As the first completely sequenced crop genome, rice continues to be the best-annotated genome for monocots and offers a wealth of information on the function and structure of genes, polymorphisms and other functional elements anchored to the genome [5]. Other crop plants are in the process of being sequenced, but have not yet reached the level of completeness offered by rice [8]. This tool offers a unique view of objectively defined gene families that facilitates comparative analyses of rice plant genomes. For example, our tool allows one to identify all gene families in rice and quickly assess the range of protein members of those gene families in addition with FASTA sequence of each protein member. Families that have proliferated greatly in one genome compared to another's, or have remained stable in size can easily be identified [1]. Scaffold of gene families into which users can sort their genes of interest [4]. We have devised search and query tools that allow users to access this information. This catalog continually grows to add new gene families and their protein members, as well as to improve annotations and website access.

Materials and methods

This software is developed under the adobe flex 3 platform. The central database is established by SQLight, and is used only in the portal. We use MXML (www.adobe.com/) and Action Script 3.0 (www.actionscript.org/) as the major programming languages. MXML is used for mainly front end interface designing and Action Script is used in all programming logics. Database is also established in action script. Action Script is also used for communication with internet for using online resources in our software. We have already tested the system on computation node(s) with Linux, Windows XP and MAC platform.

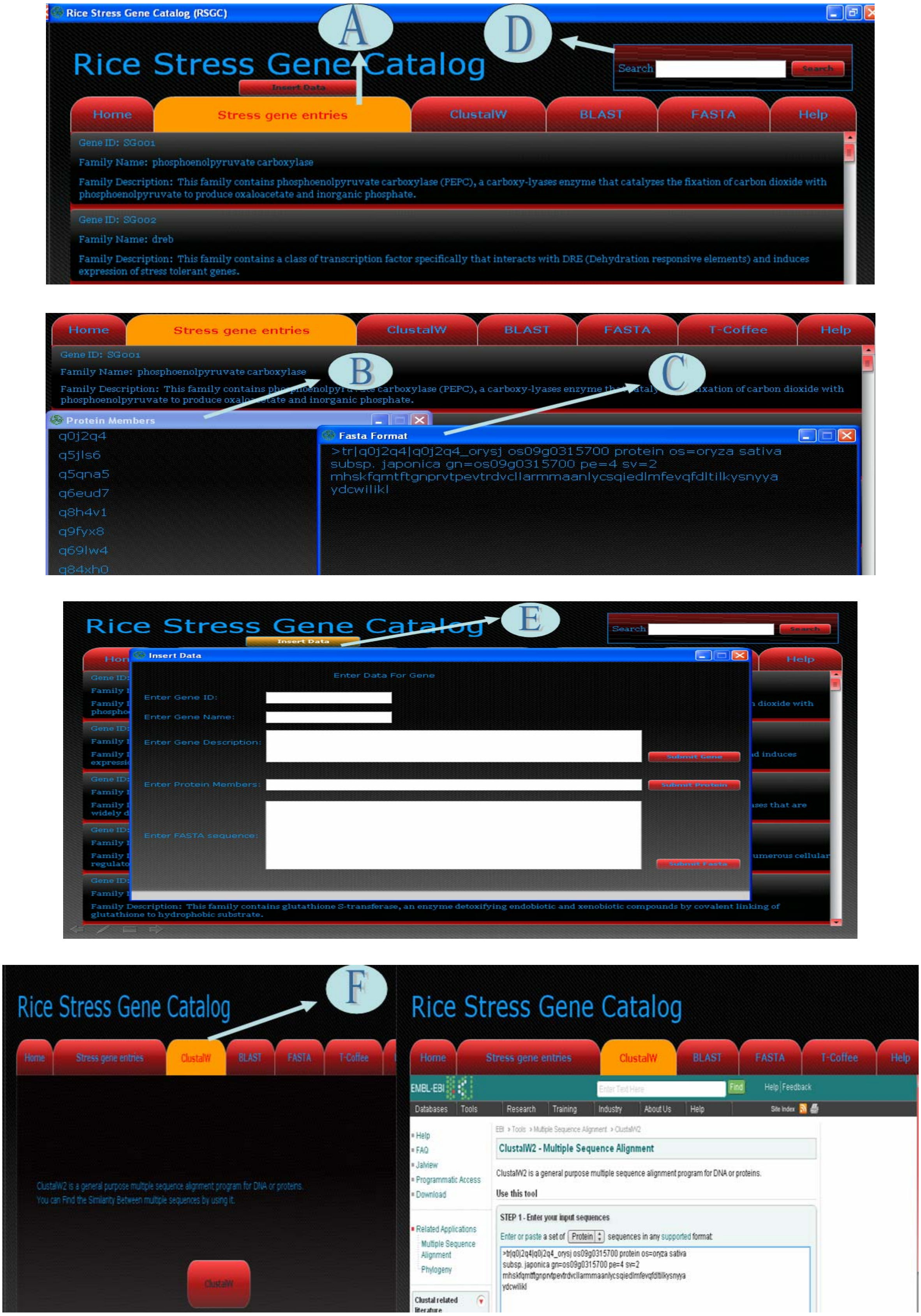
A number of services are already available through the Rice stress gene catalog [15]. To validate the feasibility, reliability, stabilization and compatibility of our system schema we built it in flex, we integrate many tools that provide the analysis of genomic and proteomic data, we facilitate the portal to find the homology between the sequences, profile searching, sequence comparison etc. For each function, some well-known packages are provided for users, e.g. Clustal W [13] and T-Coffee [11] for multiple sequence alignment, FASTA [10] and BLAST [7] for sequence comparison. These services could run independently.

Main interface

There are four main options for using the tool: browsing stress gene entries, search against tool by gene names or protein names, insert data and manipulations on FASTA sequence (Fig. 1).

Browsing stress gene entries

The tool can be used by browsing the entire set of stress protein families of rice that have been constructed (Fig. 1A) [3]. Users can select for browsing the tool from the main drop-down menu. A list of stress gene entries are shown on the front page, inside each entry are the entire protein members of that particular gene (Fig. 1B) and by clicking on any protein name users can have access to the FASTA sequence of that protein (Fig. 1C). As protein members of genes are named with their exact Uniprot IDs Details about each protein family, for example, the list of Uniprot IDs, protein names, Gene Ontology (GO) terms and key publications for each protein can be accessed through searching particular protein name against the Uniprot database [9].



Gene ID: SGoo1

Family Name: phosphoenolpyruvate carboxylase

Family Description: This family contains phosphoenolpyruvate carboxylase (PEPC), a carboxy-lyases enzyme that catalyzes the fixation of carbon dioxide with phosphoenolpyruvate to produce oxaloacetate and inorganic phosphate.

Gene ID: SGoo2

Family Name: dreb

Family Description: This family contains a class of transcription factor specifically that interacts with DRE (Dehydration responsive elements) and induces expression of stress tolerant genes.

Protein Members

q0j2q4
q5jls6
q5qna5
q6eud7
q8h4v1
q9fyx8
q69lw4
q84xh0

Fasta Format

```
>tr|q0j2q4|q0j2q4_orysj os09g0315700 protein os=oryza sativa
subsp. japonica gn=os09g0315700 pe=4 sv=2
mhskfqmtftgnprvtpevtrdvcllarmmaanlycsqiedlmfevqfdltilkysnyya
ydcwilikl
```

Insert Data

Enter Data For Gene

Enter Gene ID:

Enter Gene Name:

Enter Gene Description:

Enter Protein Members:

Enter FASTA sequence:

Family Description: This family contains glutathione S-transferase, an enzyme detoxifying endobiotic and xenobiotic compounds by covalent linking of glutathione to hydrophobic substrate.

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins.
You can Find the Similarity Between multiple sequences by using it.

ClustalW2 - Multiple Sequence Alignment

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins.

Use this tool

STEP 1 - Enter your input sequences

Enter or paste a set of Protein sequences in any supported format:

```
>tr|q0j2q4|q0j2q4_orysj os09g0315700 protein os=oryza sativa
subsp. japonica gn=os09g0315700 pe=4 sv=2
mhskfqmtflgnprvtpevtrdvcllarmmaanlycsqiedlmfevqfdltilkysnyya
ydcwilikl
```

Fig. 1 An example of browsing stress gene entries (A), obtaining protein members (B), retrieving FASTA sequence (C), search option (D), insert data (E) and manipulation on FASTA sequence like Clustal W (F)

Search option

Users can search the database by keywords by searching Family name; the matched family which contains all protein members will be retrieved (Fig. 1D). For example, if users want to search about the DREB gene family, by putting the keyword 'DREB' and clicking search option only one desired entry will appear in stress gene entries window.

Insert data

One of the main features of this tool is that it can be modified by the user. By using insert data option user can insert new stress gene families (Fig. 1E) or can modify previously stored entries. All these stress gene entries have been added to this software through this insert data option. By clicking on Insert data button a new window will open where user can insert the new stress gene entry data.

Manipulation on FASTA sequence

For the ease of the user to perform different tasks on the FASTA sequences of protein like multiple sequence alignment, protein or nucleotide comparison and to compare a novel sequence with those contained in nucleotide and protein databases by aligning the novel sequence with previously characterized genes, links to different tools home page like BLAST, Clustal W, T-Coffee has been placed inside the tool in separate windows (Fig. 1F). So user can have access to perform various functions on FASTA sequence of protein inside a single interface.

Conclusions and future prospects

The rice stress gene catalog tool offers a unique and powerful view of rice plant genome. With many plant genome sequence projects in progress, Rice stress gene catalog will allow researchers to quickly determine and identify new gene families, errors in the initial annotations, rapidly identify the best gene models and increase the confidence in the limits and structure of existing gene families [6]. Rice stress gene catalog has been designed for ease of expansion and feature addition. As new genomes are sequenced, or large EST sets generated, Rice stress gene catalog will be continuously expanded to include these data [12]. As the number of sequenced genomes increases rapidly, the continued expansion of the Rice stress gene catalog tool will facilitate a multitude of gene-family studies and genome, particularly characterization of large gene families, genome-scale analysis of multiple gene families, homology-based annotation and subsets of genes with common domain architectures.

References

1. Bateman A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, R. Eddy Sean, S. Griffiths-Jones, K. L. Howe, M. Marshall, E. L. L. Sonnhammer (2002). The Pfam Protein Families Database, Nucleic Acids Research, 30, 276-280.
2. Batemen A. (2007) Editorial. Nucleic Acids Research, 35, D1-D2.
3. Brandt B. W., J. Heringa (2009). WebPRC: The Profile Comparer for Alignment-based Searching of Public Domain Databases, Nucleic Acids Research, 37, W48-W52.
4. Enright A. J., V. Kunin, C. A. Ouzounis (2003). Protein Families and Tribes in Genome Sequence Space, Nucleic Acids Research, 31, 4632-4638.
5. Guan Y. S., R. Serraj, S. H. Liu, J. L. Xu, J. Ali, W. S. Wang, E. Venus, L. H. Zhu, Z. K. Li (2010). Simultaneously Improving Yield under Drought Stress and Non-stress Conditions: A Case Study of Rice (Oryza sativa L.), Journal of Experimental Botany, 61, 4145-4156.

6. Jagadish S. V. K., R. Muthurajan, R. Oane, T. R. Wheeler, S. Heuer, J. Bennett, P. Q. Craufurd (2010). Physiological and Proteomic Approaches to Address Heat Tolerance during Anthesis in Rice (Oryza sativa L.), Journal of Experimental Botany, 61, 143-156.
7. Li Y., N. Chia, M. Lauria, R. Bundschuh (2011). A Performance Enhanced PSI-BLAST based on Hybrid Alignment, Bioinformatics, 27, 31-37.
8. Liang C., P. Jaiswal, C. Hebbard, S. Avraham, E. S. Buckler, T. Casstevens, B. Hurwitz, S. McCouch, J. Ni, A. Pujar, D. Ravenscroft, L. Ren, W. Spooner, I. Tecle, J. Thomason, C.-W. Tung, X. Wei, I. Yap, K. Youens-Clark, D. Ware, L. Stein (2008). Gramene: A Growing Plant Comparative Genomics Resource, Nucleic Acids Research, 36, D947-D953.
9. Mi H., B. Lazareva-Ulitsky, R. Loo, A. Kejariwal, J. Vandergriff, S. Rabkin, N. Guo, A. Muruganujan, O. Doremieux, M. J. Campbell, H. Kitano, P. D. Thomas (2005). The PANTHER Database of Protein Families, Subfamilies, Functions and Pathways, Nucleic Acids Research, 33, D284-D288.
10. Miller P. L., P. M. Nadkarni, N. M. Carriero (1991). Parallel Computation and FASTA: Confronting the Problem of Parallel Database Search for a Fast Sequence Comparison Algorithm, Computer Applications in the Biosciences: CABIOS, 7, 71-78.
11. Poirot O., E. O'Toole, C. Notredame (2003). Tcoffee@igs: A Web Server for Computing, Evaluating and Combining Multiple Sequence Alignments, Nucleic Acids Research, 31(13), 3503-3506.
12. Reyes B. G., M. Morsy, J. Gibbons, T. S. N. Varma, W. Antoine, J. M. McGrath, R. Halgren, M. Redus (2003). A Snapshot of the Low Temperature Stress Transcriptome of Developing Rice Seedlings (Oryza sativa L.) Via ESTs from Subtracted cDNA Library, TAG Theoretical and Applied Genetics, 107, 1071-1082.
13. Thompson J. D., D. G. Higgins, T. J. Gibson (1994). CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-specific Gap Penalties and Weight Matrix Choice, Nucleic Acids Research, 22, 4673-4680.
14. Wall P. K., J. Leebens-Mack, K. F. Müller, D. Field, N. S. Altman, C. W. dePamphilis (2008). PlantTribes: A Gene and Gene Family Resource for Comparative Genomics in Plants, Nucleic Acids Research, 36, D970-D976.
15. Wanchana S., S. Thongjuea, V. J. Ulat, M. Anacleto, R. Mauleon, M. Conte, M. Rouard, M. Ruiz, N. Krishnamurthy, K. Sjolander, T. van Hintum, R. M. Bruskiewich (2008). The Generation Challenge Programme Comparative Plant Stress-responsive Gene Catalogue, Nucleic Acids Research, 36, D943-D946.
16. Zhou L., Y. Liu, Z. Liu, D. Kong, M. Duan, L. Luo (2010). Genome-wide Identification and Analysis of Drought-responsive microRNAs in Oryza sativa, Journal of Experimental Botany, 61, 4157-4168.

Mr. Muhammad Ali
E-mail: Aly.binm@gmail.com

Mr. Muhammad Ali, Persevering and Active Research Fellow, pursuing Master in Bioinformatics at Government College University, Faisalabad, Pakistan. Research Interests: Development of software useful for bioinformatists, protein structure prediction and Docking.

Mr. Rana Rehan Khalid
E-mail: Ray.binm@gmail.com

Mr. Rana Rehan Khalid is a Master student at Government College University, Faisalabad, Pakistan. This young researcher has interest in Image processing, Homology Modelling and software development.

Mr. Muhammad Nawaz
E-mail: M.nawaz9@gmail.com

Mr. Muhammad Nawaz is Assistant Professor at Government College University, Faisalabad, Pakistan. Professional teaching experience: about 5 year teaching and research experience. Research areas: Protein expression profiling, Functional Genomics and Molecular biology.

Mr. Nauman Qamar
E-mail: Rabia.nauman26@gmail.com

Mr. Nauman Qamar belongs to the field of Computer Science. He is running an educational institute named EDUCATORS in Rawalpindi. This young researcher has interest in DBMS and software engineering.