Using Bioinformatics to Identify Promoters in Genome Sequences

Carlos M Hernandez-Garcia* and Gabriel Abud

Department of Horticulture and Crop Science, OARDC/The Ohio State University, 1680 Madison Ave., Wooster, OH 44691, USA

*E-mail: hernandez-garcia.1@buckeyemail.osu.edu

Introduction

With the continual release of plant genome sequences, the accumulation of genomics information has been exponentially increasing. In addition to managing all this information, one of the main challenges for users is to efficiently access these genome sequences to identify and extract specific sequences of interest. The aim of this tutorial is not to train potential users how to write programming scripts but to show them step-by-step how to run BioPerl scripts, which can be obtained from collaborators who specialize in bioinformatics.

In our laboratory, we frequently identify numerous gene families and validate their promoter sequences using different tools for validation of gene expression (http://www.oardc.ohio-state.edu/SURE/). Promoters possess great importance in both basic research and in the development of improved transgenic crops. The example described in this tutorial employs a highly-conserved domain usually found in plant ubiquitin proteins (http://en.wikipedia.org/wiki/Ubiquitin) to identify *ubiquitin* or *polyubiquitin* genes containing respective single or multiple repeats of this domain in a draft genome of tomato. As promoters from *polyubiquitin* genes drive strong constitutive gene expression in important crops such as soybean (Hernandez-Garcia et al. 2009), potato (Garbarino et al. 1995), tomato (Rollfinke et al. 1998), rice (Sivamani and Qu 2006) and maize (Christensen and Quail 1996), here we identify *polyubiquitin* genes of tomato and automatically extract their promoter sequences using a provided BioPerl script. With patience, dedication, and guidance you will soon be able to manipulate, modify, and even create BioPerl scripts for use in your own research.

Procedure

1. Software installation. In our case, we installed Cygwin, which is a UNIX emulator suitable for PC users with Windows running as the default operating system. Cygwin installation guide and software are publicly available at http://www.cygwin.com/. Then, install BioPerl available at http://www.bioperl.org/wiki/Main_Page. Since we are going to perform BLAST (Basic Local Alignment Search Tool) to identify gene sequences in our target genome sequence with high homologies to our query sequence, we need to install StandAlone BLAST (BLAST that can be run locally as a full executable and be used to run BLAST searches against private, local databases, or downloaded copies of the NCBI databases). A well-detailed manual to install StandAlone BLAST is freely available at http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/pc_setup.html
2. Download a draft genome sequence of tomato from the Sol Genomics Network at http://solgenomics.net/. Pass your cursor over ***Genomes & Sequences*** and click on ***Search/browse tomato genome*** link. Then click on ***scaffolds*** for the most updated release and save the compressed '****.gz***' file in your local disk. Extract your compressed file using any compressor software like the publicly open 7-Zip (http://www.7-zip.org/). For the tomato genome sequence downloaded on 08/05/2010, the name for the compressed database was ***S_lycopersicum_scaffolds.2.10.fa.gz*** and the decompressed file was ***SL2.10sc***.
3. Create a text file containing the query sequence in fasta format using a word editor. e.g using Notepad++ you can open a new blank file and paste your sequence of interest. Then convert the format to UNIX by clicking on the ***Format*** tab followed by a click on ***Convert to UNIX Format***. Save the file using ***All types (*.*)*** option. This will allow the following software to recognize this file (If you are not using cygwin, just change the extension of the file to .pl). The nucleotide query sequence used in this tutorial corresponds to a single ubiquitin domain found in *ubiquitin* and *polyubiquitin* genes of soybean (Hernandez-Garcia et al. 2010b). This domain is highly conserved among different plant species.
4. Place your decompressed database (***SL2.10sc***) file and the file containing the query sequence (named ***Gmubidomain***) in a new folder (named ***Solanum_Lycopersicum***), previously open in the root directory (in our case, C:\cygwin\home\Carlos***Solanun_Lycopersicum***). Note that Carlos is the username for the computer used in this analysis.
5. Format your genome database before use BLAST. Open Cygwin and locate in the ***Solanun_Lycopersicum*** folder (**Carlos@Carlos-PC ~/Solanum_Lycopersicum**) and type

the following command ***formatdb.exe –i SL2.10sc.fasta –p F***. ***-i*** indicates the input file to format into a searchable database. ***-p*** asks if the input data is protein sequence. ***F*** indicates "false", which therefore specifies a nucleotide database. For protein sequences use ***–p T***. If formatting was successfully performed now you are able to see the following three new index files (***SL2.10sc.fasta.nhr, SL2.10sc.fasta.nin, SL2.10sc.fasta.nsq***) along with a txt file named ***Formatdb*** indicating your success.

6. Run a BLAST search for your query sequence against the previously downloaded and formatted genome database. Located inside the ***Solanun_Lycopersicum*** folder in cygwin, type the following command ***blastall.exe –p blastn –d SL2.10sc.fasta –i gmubidomain –o SL_Gmubi.out.*** The program to be used (blastn, in this case) is specified by ***-p***, ***-d*** specifies the database, ***-i*** indicates the query file and ***-o*** the desired name for the output file. The blast process usually takes a while, so you can grab a cup of fresh coffee and update your facebook profile in the meantime.
7. Check the output file using a text editor. You will be able to see all the hits your query sequence found in the searched database.
8. Write or obtain your BioPerl script (attached) to parse the output file generated during the BLAST procedure. Make sure that the '***use lib…***' lines in the provided script specify the directory where the BioPerl modules were installed. The current lines in this script specify the directory where the modules are installed by default, but check these two lines if you changed the directory or are having problems running the program. Because we work with promoters, we prefer the script give us the upstream regions of coding sequences for the *ubiquitin* genes of tomato with the highest similarities with our query sequence, as determined with BLAST. The biological interpretation, in our example, is to get the promoter sequences from the *ubiquitin* genes with at least one ubiquitin domain (similar to the query sequence). The script also allows us to specify a desired e-value and the number and the length of promoters.
9. Parse your output file (***SL_Gmubi***) using a BioPerl script. This procedure creates a file containing the sequences and descriptors for the genes with the highest similarities to the query sequence. Go to Command Prompt and locate at the ***Solanum_Lycopersicum*** folder, C:\cygwin\home\Carlos***Solanum_Lycopersicum>***
10. Type your command and specify your parameters. You can see the parameters that this script accepts by typing ***find_blast_matches.pl –help.*** The following example shows a parsing

procedure with specified parameters.

C:\cygmin\home\Carlos\Solanum_Lycopersicum>***find_blast_matches.pl SL_Gmubi.out SL2.10sc.fasta –p 2000 –n 100.*** This command will extract 100 promoters or less from the genome sequence based on the results in the output blast file and an e-value less than 0.01, which is specified by default in this script. ***-p 2000*** indicates promoters with 2000 bp length. You can also modify the e-value parameter by entering ***-e*** followed by the desired value in front of either ***-p*** or ***-n***.

11. One can easily create an output file containing the results from the parsing procedure. e.g typing C:\cygmin\home\Carlos\Solanum_Lycopersicum>***find_blast_matches.pl SL_Gmubi.out SL2.10sc.fasta –p 2000 –n 100 > c:\users\Carlos\Desktop\ubipromoters*** will create an output txt file on our desktop.
12. Check the created file for completion. In our example, we identified 31 promoter sequences in a draft genome sequence of tomato, corresponding to 31 *ubiquitin* genes containing at least one ubiquitin-domain with high similarity (e-value= 0.01) to a soybean ubiquitin domain used as a query sequence.
13. Manipulate your sequences according to your purposes. In our personal case, the purpose is promoter isolation and functional characterization using different tools for gene expression analysis. A video-article (Hernandez-Garcia et al. 2010a) showing the methods for genetic transformation and rapid validation of plant promoters is freely available at (http://www.jove.com/index/details.stp?id=1733). Eventually, the coding regions for specific genes are extracted and phylogenetic analysis is performed to predict functionality of promoters.

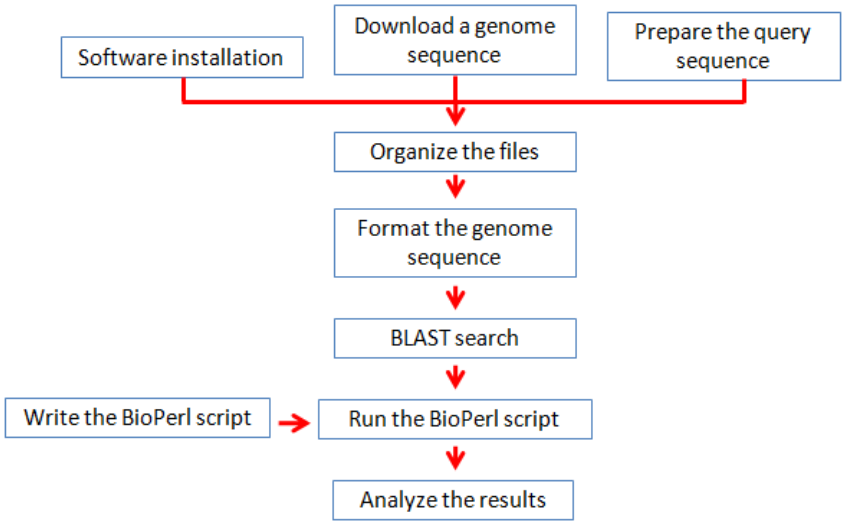

Fig. 1 Simplified diagram depicting the main steps of this tutorial

Acknowledgements

The authors would like to thank Drs David Francis and Heather Merk for reviewing this tutorial and providing helpful suggestions. Input from classmates enrolled in HCS806 course (Summer, 2010) is also highly appreciated. The work on promoter analysis mentioned above has being developed in the laboratory of Dr. John J. Finer (Department of Horticulture and Crop Science, The Ohio State University/OARDC). CMHG is funded by a Graduate Associateship from the Department of Horticulture and Crop Science, The Ohio State University, and partial support from CONACYT, Mexico.

References Cited

- Christensen A. H. and P. H. Quail. 1996. Ubiquitin promoter-based vectors for high levels of selectable and/or screenable marker genes in monocotyledonous plants. Transgenic Research 5:213–218.
- Garbarino, J. E., T. Oosumi, and W. R. Belknap. 1995. Isolation of a polyubiquitin promoter and its expression in transgenic potato plants. Plant Physiology 109:1371–1378.
- Hernandez-Garcia, C. M., R. A. Bouchard, P. J. Rushton, M. L. Jones, X. Chen, M. P. Timko and J. J. Finer. 2010b. High level transgenic expression of soybean (Glycine max) GmERF and Gmubi gene promoters isolated by a novel promoter analysis pipeline. BMC Plant Biology 10:237.
- Hernandez-Garcia, C. M., J. M. Chiera, and J. J. Finer. 2010a. Robotics and dynamic image analysis for studies of gene expression in plant tissues. Journal of Visualized

Experiments 39. (Available online at: www.jove.com/index/details.stp) (verified 10 Dec 2010).
- Hernandez-Garcia, C. M., A. P. Martinelli, R. A. Bouchard, and J. J. Finer. 2009. A soybean (Glycine max) polyubiquitin promoter gives strong constitutive expression in transgenic soybean. Plant Cell Reports 28:837-849.
- Rollfinke, I. K., M. V. Silber, and U. M. Pfitzner. 1998. Characterization and expression of a heptaubiquitin gene from tomato. Gene 211:267–276.
- Sivamani, E. and R. Qu. 2006. Expression enhancement of a rice polyubiquitin gene promoter. Plant Molecular Biology 60:225–239.

External Links

- BioPerl [Online]. Available at: www.bioperl.org/wiki/Main_Page (verified 10 Dec 2010).
- Cygwin [Online]. Red Hat. Available at: www.cygwin.com/ (verified 10 Dec 2010).
- National Center for Biotechnology Information. BLAST+ Setup: Procedures for Windows PC [Online]. U. S. National Library of Medicine, National Institutes of Health. Available at: www.ncbi.nlm.nih.gov/staff/tao/URLAPI/pc_setup.html (verified 10 Dec 2010).
- Praline Multiple Sequence Alignment [Online]. The Centre for Integrative Bioinformatics VU. Available at: www.ibi.vu.nl/programs/pralinewww/ (verified 10 Dec 2010).
- Sol Genomics Network [Online]. Available at: solgenomics.net/ (verified 10 Dec 2010).
- 7-Zip [Online]. Igor Pavlov. Available at: www.7-zip.org/ (verified 10 Dec 2010).