Job: Plant Functional Genomics--Grotewold
Chapter: Chapter 5
Pub Date: 7/1/2003
Template: MiMB/6x9/Template/Rev.02.03

Compositor: Nettype
Date: 3/15/2003
Revision: First Proof

5

Combined ESTs from Plant–Microbe Interactions: Using GC Counting to Determine the Species of Origin

Edgar Huitema, Trudy A. Torto, Allison Styer, and Sophien Kamoun

Summary

A diversity of microorganisms establishes intimate associations with plants. Whether pathogenic or symbiotic, such interactions are the result of complex recognition events between plants and microbes, leading to signaling cascades and regulation of countless genes involved in the interaction. A key step in unraveling the mysteries of plant–microbe interactions lies in defining the transcriptional changes that occur in both the host and the microbe during their association. The sum of the transcripts, from both host and microbe, which are produced during their association, has been defined as the interaction transcriptome. One approach to analyze interaction transcriptomes is to perform large-scale sequencing of cDNAs (expressed sequence tags or ESTs) obtained from infected plant tissue and representing a mixture of host and microbe sequences. In some cases, the two organisms have markedly different GC content, allowing most ESTs to be easily distinguished on this basis. In this chapter, we describe a GC counting method to determine the species of origin of ESTs obtained from interactions between plants and oomycetes or other high GC content microbes.

Key Words

plant–microbe interactions, *Phytophthora*, oomycetes, interaction transcriptome, EST annotation, GC content, GC counting

1. Introduction

A diversity of microorganisms establishes intimate associations with plants. Whether pathogenic or symbiotic, such interactions are the result of complex recognition events between plants and microbes, leading to signaling cascades and regulation of countless genes involved in the interaction. A key step in unraveling the mysteries of plant–microbe interactions lies in defining the genetic components involved and the transcriptional changes that are occur-

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: Erich Grotewold © Humana Press, Inc., Totowa, NJ

Job: Plant Functional Genomics--Grotewold
Chapter: Chapter 5
Pub Date: 7/1/2003
Template: MiMB/6x9/Template/Rev.02.03
Compositor: Nettype
Date: 3/15/2003
Revision: First Proof

ring in both the host and the microbe ***(1)***. The sum of the transcripts, from both host and pathogen, which are produced during their association, has been defined as the "interaction transcriptome" ***(1)***. Each interaction transcriptome has been hypothesized as being unique to a particular host–pathogen or host–symbiont association, and its characterization should help to define the complex mechanisms involved in establishing and maintaining their interaction ***(1)***.

The emergence of low-cost high-throughput DNA sequencing methods has allowed plant biology to enter the era of genomics. In particular, projects involving large-scale sequencing of cDNAs (expressed sequence tags or ESTs) are ongoing for a wide variety of plants and plant-associated microbes. Similarly, ESTs generated from mRNA isolated from plant tissue infected with microbial pathogens have emerged as useful data sets for dissecting interaction transcriptomes ***(1,2)***. For example, this approach has been used for two eukaryotic microbial pathogens, the oomycete *Phytophthora infestans*, which causes late blight on tomato and potato (E. Huitema and S. Kamoun, unpublished; B. Baker et al. NSF Potato Genomics Project, www.tigr.org/tdb/potato), and *Phytophthora sojae*, which causes root and stem rot on soybean ***(1,2)***. ESTs generated from cDNA libraries constructed from *Phytophthora*-infected plant tissue could be of either pathogen or host origin. Thus, the challenge is to distinguish between the plant and *Phytophthora* EST populations using sequence analyses. In this case, plant and *Phytophthora* ESTs have markedly different GC contents, allowing most ESTs to be easily distinguished on this basis. For example, the percentage of GC content was assessed for sequences from cDNA libraries derived solely from *P. sojae* and soybean ***(2)***. Both sets of sequences produced distinct slightly overlapping normal distribution curves, with the pathogen ESTs averaging 58% GC content, and the host ESTs averaging 46% GC content ***(2)***. A similar analysis of sequences from a *P. sojae*-infected soybean cDNA library revealed ESTs to be clustered around two peaks corresponding to 46 and 58% GC content, suggesting that about two-thirds of the ESTs from this library are likely to be from the pathogen ***(2)***. In this chapter, we provide step-by-step instructions on how to run the GC counting method to help distinguish between host and microbe sequences from ESTs from interactions between plants and oomycetes or other high GC content microbes.

2. Materials

2.1. Hardware and Operating System

A workstation running the Linux operating system. For example, we currently use a Pentium III personal computer (PC) running Red Hat Linux OS.

Job: Plant Functional Genomics--Grotewold
Chapter: Chapter 5
Pub Date: 7/1/2003
Template: MiMB/6x9/Template/Rev.02.03
Compositor: Nettype
Date: 3/15/2003
Revision: First Proof

2.2. Software

The GC counting program *GC* can be downloaded from (http://www.oardc.ohio-state.edu/phytophthora/gc.htm). The program was written in C++ and was only tested on the Linux platform.

Microsoft® Excel® or a similar spreadsheet program running on a Linux, PC, or Mac® platform.

2.3. Data Sets

Processed ESTs in a FASTA format ***(3)*** (see also [http://www.oardc.ohio-state.edu/phytophthora/gc.htm] for a sample input file). It is essential to remove vector sequences and to trim low quality sequences prior to processing.

3. Methods

3.1. Running GC to Count the Frequency of GCs

1. Download or transfer the program *GC* and the input file containing the ESTs to the appropriate directory in your Linux workstation (*see* **Note 1**).
2. Start the program by typing: gc.
3. At this point, you will be prompted to type the input file name and then the output file name.
4. The output file is a comma-formatted file that can be exported into Excel or a similar spreadsheet program.

3.2. Importing GC Output into Microsoft Excel

1. Open or import the output file with Microsoft Excel. The Text Import Wizard window will pop-up.
2. Select original data type: delimited.
3. Click Next.
4. Select delimiters: comma and deselect tab.
5. Click Next.
6. In data preview, assign column A to text format and the other columns to general format.
7. Click Finish.
8. The GC frequency data is now imported into the spreadsheet.

3.3. Description of Output

There are eight columns in the output file:

1. Column A: sequence ID.
2. Column B: GC content for frame 1 (based on the first base of the EST).
3. Column C: GC content for frame 2.
4. Column D: GC content for frame 3.
5. Column E: GC content for entire sequence.
6. Column F: Ratio of GC content frame 1/GC content entire sequence.

7. Column G: Ratio of GC content frame 2/GC content entire sequence.
8. Column H: Ratio of GC content frame 3/GC content entire sequence.

3.4. Identifying High GC Sequences

The table can be sorted in descending order based on column E to help identify high GC sequences:

1. Select columns A–H.
2. Select Data:Sort and sort based on column E and descending order.
3. Identify high GC sequences by scrolling down the file.
4. For oomycete–plant ESTs, we estimate that sequences with a GC content higher or equal to 53% have a 98% probability to be of pathogen origin (*see* **Note 2**).

3.5. Quality Check

A quality check can be performed by searching the high GC sequences against species-specific databases using the BLASTN algorithm ***(4)*** (*see* **Note 3**).

4. Notes

1. Ideally, the sequences should be generated using a robust base-calling program, such as phred ***(5,6)***. It is essential to trim the ESTs for low quality sequences. Some EST data sets may have an overrepresentation of long stretches of As or Ts due to the polyadenylation signals in the mRNA. In such cases, these stretches need to be removed.
2. This estimate is based on the observation that for tomato, less than 2% of the ESTs have a GC content higher or equal to 53%.
3. The clear differences in GC content between plant and oomycete cDNA sequences may not occur in other pathosystems. The GC content of cDNAs from the examined organisms need to be determined in order to establish a reliable threshold for discrimination. In cases in which there are no clear difference in GC content, the hexamer counting method described by Hraber et al. ***(7)*** could be a valuable alternative.

Acknowledgments

Supported by the OARDC Research Enhancement Grant Program and Syngenta Biotechnology, Inc. Salaries and research support were provided by State and Federal Funds appropriated to the Ohio Agricultural Research and Development Center, the Ohio State University.

References

1. Birch, P. R. J. and Kamoun, S. (2000) Studying interaction transcriptomes: coordinated analyses of gene expression during plant-microorganism interactions, in

New Technologies for Life Sciences: A Trends Guide (Wood, R., ed.), Elsevier Science, New York, pp. 77–82.

2. Qutob, D., Hraber, P. T., Sobral, B. W., and Gijzen, M. (2000) Comparative analysis of expressed sequences in *Phytophthora sojae*. *Plant Physiol*. **123,** 243–254.
3. Pearson, W. R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*. **183,** 63–98.
4. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **17,** 3389–3402.
5. Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*. **8,** 175–185.
6. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*. **8,** 186–194.
7. Hraber, P. T. and Weller, J. W. (2001) On the species of origin: diagnosing the source of symbiotic transcripts. *Genome Biol*. **2,** RESEARCH0037.

Uncorrected Proof Copy