PathoGene™: a pathogen coding sequence discovery and analysis resource

Kar-wai Ng, Jonathan Lawson, and Harold R. Garner
The University of Texas Southwestern Medical Center, Dallas, TX, USA


PathoGene™ is a web-based resource that streamlines the process of predicting genes in microorganisms and designs PCR primers for amplification to facilitate sequence analysis and experimentation. PathoGene currently supports primer design for every complete microbial, viral, and fungal genome as cataloged in GenBank® by the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/). The resulting primers can then be subjected to a stand-alone Basic Local Alignment Search Tool (BLAST) system called PathoBLAST, in which the predicted PCR product and/or primers can be compared against the genome of interest or a similar genome to find related genes or estimate primer quality.

INTRODUCTION

Since the September 11th terrorist attacks and subsequent anthrax attacks, there has been a growing determination to prepare for future bioterror strikes or other public health crises, such as severe acute respiratory syndrome (SARS) or West Nile virus. On September 4, 2003, Health and Human Services Secretary Tommy G. Thompson announced grants to establish 10 Regional Centers of Excellence for Biodefense and Emerging Infectious Diseases Research (RCE).

The RCE program provides a coordinated and comprehensive mechanism to support the interdisciplinary research that will lead to new and improved therapies, vaccines, diagnostics, and other tools. We have created a resource, PathoGene™, which provides one capability important to researchers involved in this work and general microbiology. PathoGene is a web-accessible, up-to-date, quick, and validated tool for the identification of protein-coding DNA sequences (CDSs) and the design of computed reagents (primers, sequences, PCR products, etc.).

To study a microbial CDS, its sequence must first be found within the genome. Its flanking sequences, on which PCR primers are to be designed, must also be found. Although PCR primer selecting software tools are available, generation of the input files is often cumbersome if performed manually. This is made more difficult due to strict formatting requirements, potential user length constraints, and the number of CDSs to be studied. PathoGene was designed to assist microbiologists accomplish the above tasks in a fraction of the time. It is a web-based computer program (CGI) that integrates public databases and bioinformatics tools to completely automate PCR primer design process. At the time of writing, PathoGene supports primer design for 518,782 CDSs in 160 microbial organisms from 238 genomic annotations, 1063 viruses from 1518 annotations, and 30 fungi from 69 annotations. PathoGene, along with documentation, is available as a free service of the Region VI RCE Computational Biology Group at http://pathogene.swmed.edu. Its efficacy has been validated in the laboratory via PCR amplification of a test set of CDSs from gram-positive Bacillus anthracis and gram-negative Yersinia pestis.

MATERIALS AND METHODS

Computational Resources

The PathoGene source code was written in the Perl scripting language (v.5.8.0) and uses a standard CGI module. All web interfaces were created using standard HTML (v.4.01) and embedded Javascript. PathoGene and its pathogen database currently reside on a Linux (kernel v.2.4.19) web server. The database is continuously updated by a suite of local tools as new sequences, and annotations become available in GenBank® (1).

Validation Resources

PathoGene was tested and validated using 24 randomly chosen CDSs from B. anthracis (Ames strain) and Y. pestis. Y. pestis genomic DNA (gDNA) was isolated from a cultured clinical isolate. The gDNA of B. anthracis and Y. pestis was provided by C. Rick Lyons (University of New Mexico). The PCR primers were designed using the optimal conditions (2) of 800 bp product size, 23 bp primer size, 60°C primer melting temperature, and 50% GC content. The primers were obtained commercially from Qiagen (Valencia, CA, USA). Primer resuspension and redistribution were performed using a Genesis Robotic Sample Processor 200 (TECAN, Maennendorf, Switzerland) to a final concentration of 10 μM. PCR was carried out in the PTC-225 Peltier Thermal Cycler (MJ Research, Waltham, MA, USA) in 20-μL reactions (1x Taq buffer with MgCl2, 0.2 M dNTPs, 2 U Taq DNA polymerase, 0.4 μM each primer, 10–20 ng gDNA) using a standard cycling protocol (95°C for 5 min, then 30 cycles of 94°C for 45 s, 55°C for 45 s, and 72°C for 90 s, followed by 5 min at 72°C).

Sequence Analysis and Algorithm

PathoGene takes an organism selected by the user from a drop-down list, a user-supplied FASTA-formatted sequence, or an uploaded FASTA-formatted file as input. Users may also specify a particular CDS to design primers and choose the method in which CDSs are to be found. The organism input portion of the PathoGene user interface is depicted in Figure 1. Once an organism is chosen, PathoGene can find CDSs either by parsing the complete genomic annotation (Annotation Method) or by executing Glimmer (v.2.13) (3,4) to predict CDSs given options defined by the user and the genomic sequence prepared by our script (Glimmer Method).
If a FASTA-formatted sequence or file is submitted, PathoGene will automatically execute the Glimmer Method to predict CDS locations. However, it should also be noted that if a small sequence (approximately 2 to 3 kb) is provided, a sufficient training model may not be generated to allow Glimmer to find the remaining genes properly. If a fungal species is chosen, PathoGene will automatically default to the Annotation Method because Glimmer is not compatible with eukaryotic genomes.

Once the location of the target CDS(s) is known, PathoGene will extract its sequence and the user-defined flanking and buffer sequences from the genome. Primer design will be avoided within the CDS’s adjacent user-defined buffer regions to facilitate sequencing. Primer3 (v.0.9) (5) is then used to design PCR primers within the flanking sequences of the CDS. Primer picking parameters, such as product and primer sizes, GC content, and melting temperatures, can be adjusted by the user. In addition, real-time PCR primers can also be made, but the product size is restricted between 80–200 bp. Generation of internal primers (for in-frame cloning) is also possible but not recommended, as the optimal primer picking parameters are severely relaxed. Results are then reported to the user. The above process is generalized in Figure 2.

In the event that a CDS is large, it can be broken into 500-bp segments. Primers are designed for each segment, which can then be amplified individually. The products can be recombined into a single CDS by molecular methods (e.g., overlap PCR) or computationally via sequence analysis. By default, the separated products will have a maximum overlap of 250 bp, but this can vary depending on the chosen flanking and buffer sequence sizes. PathoGene also provides the option to use Basic Local Alignment Search Tool (BLAST) (6–10) with the PCR product and/or primers against the target genome or another user-defined microorganism using the PathoBLAST resource (http://rce.swmed.edu/genomes/). This can determine the uniqueness of the predicted PCR product and the theoretical specificity of the primers. PathoBLAST is an independently developed, stand-alone BLAST tool that shares the same microorganism library as PathoGene.

The promoter region, defined as the closest, upstream, same-strand, noncoding region between the target CDS and the next upstream CDS with a maximum size of 2 kb, can be found and have primers designed for it. PathoGene can also find the 2-kb upstream region of each CDS. This information can be valuable to researchers and can easily be introduced into promoter motif algorithms and tools (11,12).

Primer DB is a complementary database to the PathoGene resource. It was designed to serve as a repository of validated primers generated by PathoGene. With this tool, researchers can submit and retrieve primers and the parameters under which they were made.

Output Format

PathoGene generates a separate results page once primary processing (prior to Primer3 execution) is complete. This page will display general information for the selected organism, including its accession number, the total number of CDSs found, and the size of the genome. More specific information about the target CDS will then be provided, such as its name, locus tag, PathoGene number (position that the CDS appears in the annotation or Glimmer prediction), and its start and stop bases. If the CDS is found on the negative strand, this will also be noted. PathoGene will then output a series of sequences: (i) the sequence of the CDS; (ii) the promoter region and/or 2-kb upstream sequences if requested; and (iii) the sequence of the CDS along with flanking and buffer regions attached. Upon completion of Primer3, a hyperlink to the generated primer sets will be created. Hyperlinks are also created to PathoBLAST if the user chooses to BLAST the product and/or primers.

Validation Method

To confirm PathoGene as a rapid automated tool for primer design, we performed a comparison between Patho-
Gene and manual CDS processing. We processed 24 randomly chosen CDSs from *B. anthracis* and *Y. pestis*, and those larger than 500 bp were split into 500-bp segments. This yielded 30 segments for each organism, and primers were individually generated for each segment. Using PathoGene, the Annotation and Glimmer Methods completely processed a CDS in an average of 6 and 45 s, respectively. Manual processing was done using Microsoft® Development Environment (v.6.0), Glimmer, Primer3, the genomic annotation, the accompanying FASTA-formatted sequence, and find, cut, and paste operations. Per CDS, it took an average of 19 and 27 min to emulate the Annotation and Glimmer Methods, respectively. Manual processing utilized the same primer picking parameters for Primer3 as PathoGene. Although PathoGene and manual processing yielded identical results, the significant decrease in processing time and the standardization of the approach, which is not prone to user errors, demonstrates one value of the PathoGene resource.

To validate the generated primers, PCR was performed using the three best-fit primer pairs designed by Primer3 for the 30 segments in each organism. *B. anthracis* primers had a 98.9% success rate. All primers gave a single product of appropriate size except for one pair. *Y. pestis* primers had a 96.67% success rate. Three primer pairs failed to give any product, however they were all designed to amplify the viral transposase gene *tnp*. There are numerous viral transposases, and their presence and location are known to vary between strains of *Y. pestis*. The strain used in the PCR is a clinical isolate, thus its genomic sequence may vary from the annotated CO92 strain used to design the primers. Because all three primer pairs failed to generate a product, it is highly probable that the gene *tnp* is either not present in the clinical isolate or in a different genomic text within the isolate. All other primers gave a single product of the appropriate size. Therefore, the success rate may be higher than what is reported. The fact that the PathoGene primers, designed using a sequenced pathogen, were able to successfully amplify homologous

---

**Figure 2.** Flowchart describing PathoGene processing.

Vol. 37, No. 2 (2004)
genes from a clinical isolate lends confidence to the utility of the program for practical applications.

RESULTS AND DISCUSSION

PathoGene can facilitate microorganism investigations by automating most of the steps prior to PCR amplification. It has been shown to give identical results to manual processing but is on average 100 times faster. Furthermore, PathoGene reduces the potential of human error. We have also shown that the program is capable of rapidly selecting PCR primers that generate single products from either the target strain or similar nonsequenced strain. This can facilitate the development of PCR-based diagnostics or genotyping tools. Furthermore, the program can be used to design primer sets for the generation of PCR-based products for microarrays regardless of its annotation status. This will accelerate the study of pathogens for which detailed annotations have not been completed. While other studies and computer programs (13,14) have been published relating to these computational tasks for microarray design, none offer the flexibility of PathoGene in terms of annotation status or the ability to BLAST the results for quality and specificity determination. Also, many do not have the capability of genome-wide primer design nor are the primer design parameters optimized for pathogen genomes. The PathoGene output can be applied to almost any PCR-based study of microorganisms.

Future upgrades include the addition of a motif-finding function for the promoter region sequences. Also, a precomputed primer database for every CDS in every organism under default parameters is in the early stages of development. This database, PathoGeneDB, will offer users instantaneous results for primer queries. Although the running time of PathoGene is negligible, the database will help reduce computational costs when a large volume of simultaneous queries are received or when computationally expensive requests are made, such as primer generation for every CDS in an organism. Additional gene identification programs will also be added to provide the user with a variety of options.

We believe PathoGene is a rapid and versatile tool that can facilitate primer design and accelerate the study of microbial genomes. The immense size of PathoGene’s organism library and diverse functionality provide researchers with a wealth of capabilities. The ability to examine individual and proprietary sequences can aid annotation and genomic projects. Through the synergy of these benefits, we believe this resource will facilitate work leading to new advances in biodefense.

ACKNOWLEDGMENTS

PathoGene and PathoBLAST are available on the web at http://pathogene.swmed.edu/ and http://rec.smed.edu/geneomes/, respectively. This work was supported by a grant from the National Institute of Allergy and Infectious Diseases (NIAID) to H.R.G. through the Western Regional Center of Excellence for Biodefense and Emerging Infectious Disease Research, National Institutes of Health (NIH) grant no. U54 AI057156. The authors wish to thank David Trusty, Steve Crozier, Mark Burkart, Robert Powell, and Alexandra Smith for valuable comments and contributions. C. Rick Lyons for providing genomic DNA, and the Center for Biomedical Inventions for assistance with validation.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing interests.

REFERENCES


Received 1 March 2004; accepted 16 April 2004.

Address correspondence to Harold R. Garner, Eugene McDermott Center for Human Growth and Development, The University of Texas Southwestern Medical Center, 6000 Harry Hines Boulevard, Dallas, TX, 75390-8591, USA. e-mail: harold Garner@utsouthwestern.edu