Patho-Genes.org: a website dedicated to gene sequences of potential bioterror bacteria and PCR primers used to amplify them

Julien Gardès,* Dipankar Bachar, Olivier Croce and Richard Christen
CNRS UMR 7138 Systématique Adaptation et Evolution, Université de Nice-Sophia Antipolis, Parc Valrose BP71, F06108 Nice cedex 02, France.

Summary
Pathogenic agents can be very hard to detect, and usually they do not cause illness for several hours or days. To improve the speed and the accuracy of detection tests and satisfy the needs of early diagnosis, molecular biology methods such as PCR are now used. However, selecting a proper target gene and designing good primers is often not easy. We present a dedicated website, http://patho-genes.org, where we provide every sequence, functional annotation, published primer and relevant article for every annotated gene of major pathogenic bacterial species listed as key agents to be used for a bioterrorism attack. Each published primer was analysed to determine its melting temperature, its specificity and its coverage (i.e. its sensitivity against every allele of its target gene). Data generated have been organized in the form of data sheet for each gene, which are available through multiple browser panels and query systems.

Introduction
The detection protocols of pathogens are traditionally based on cell cultures and biochemical tests (or use of antibodies), which are inexpensive and largely automated. However, these procedures often require a long time. The advent of the polymerase chain reaction (PCR) allowed the emergence of molecular diagnoses. Faster [7 min (Belgrader et al., 1999) or less (http://www.nhdiag.com/profile_one.shtml)] and often more accurate (allowing to characterize which alleles are present), molecular detections are progressively replacing cell cultures and biochemical tests (Christen, 2008; Fricke et al., 2009), allowing early diagnosis (Saravolatz et al., 2003) and therapy monitoring (Zaph, 2010).

Nevertheless, the development of molecular techniques requires selecting the proper target gene and using good primers. A valid primer should be specific of the target species and should cover every intraspecific genetic variation of the target gene. Its melting temperature (Tm) should be 55°C or above for an optimized amplification (Arun and Saurabha, 2003). In silico analyses theoretically allow verifying the specificity, the sensitivity and the thermodynamic conditions of a molecular protocol, before doing the real experiment.

Retrieving every sequence for a given gene is not always easy. Two different strategies are commonly used, retrieval by sequence similarity or by gene names (keywords). The most popular program to search sequences by similarity is BLAST (Altschul et al., 1990). However, selecting the threshold (expect value, e-value) to avoid noise (similar but different genes) may be quite difficult. This e-value depends upon the size of the database and the proper value to select is not known a priori. When retrieving sequences using keywords, the difficulty lies in non-annotated gene sequences, variations in naming, or wrong annotations. Finally, retrieval of PCR primers from the literature is extremely tedious because hundreds of articles have to be carefully read. As a result, primers used are often retrieved from a recent publication but with no real warranty of specificity and coverage.

We propose a web resource, http://patho-genes.org, which includes every genetic information required for developing or using molecular detection: every sequence, functional annotation, relevant article, published primer for each pathogenicity gene of major potential bioterror bacteria. For each published primer, an in silico analysis was performed to determine the specificity, coverage and melting temperature of each primer. Navigation panels and query systems allow retrieving these datasets.

Results and discussion
Patho-Genes.org: Content
Data provided by this website are freely available. One can choose a pathogenic bacterium from the homepage by selecting a species or by providing primers. Table 1 lists the main information available. A browsing system allows consulting the data sheet of a given gene (Fig. 1). Genes can be selected from an alphabetical list of gene or

Received 28 September, 2011; revised 24 April, 2012; accepted 15 May, 2012. *For correspondence. Email jgardes@gmail.com; Tel. (+33) 4 92 07 69 47; Fax (+33) 4 92 07 64 05.

Published 2012. This article is a U.S. Government work and is in the public domain in the USA.
Table 1. Major potential bioterror bacteria and statistics at Patho-Genes.org.

| Organism                      | Public Databases | Patho-Genes.org |
|-------------------------------|------------------|-----------------|
|                               | CDS             | CDS/genome GC% | Entries CDS | Published primers | Articles |
| Bacillus anthracis            | 28 835          | 5288           | 34.8        | 1190              | 204      | 3 934 |
| Burkholderia mallei           | 21 292          | 5213*          | 68.8*       | 1426              | 6 413    | 0     | 58    |
| Burkholderia pseudomallei     | 31 752          | 6152*          | 68.2*       | 2171              | 11 403   | 0     | 725   |
| Chlamydomphila psittaci       | 2 248           | 963            | 38.9        | 412               | 1 124    | 103   | 224   |
| Coxiella burnetti             | 11 590          | 1903           | 42.6        | 846               | 5 047    | 0     | 38    |
| Francisella tularensis        | 14 470          | 1698           | 32.2        | 886               | 8 135    | 117   | 1 965 |
| Mycobacterium tuberculosis    | 22 828          | 4046           | 65.6        | 1814              | 11 886   | 0     | 13 632|
| Rickettsia prowazekii         | 2 067           | 893            | 29.0        | 836               | 1 932    | 0     | 178   |
| Yersinia pestis               | 43 032          | 3863           | 46.9        | 1058              | 31 700   | 187   | 3 307 |

Entries of Patho-Genes.org correspond to annotated genes for each species. CDS: number of CDS sequences available in the EMBL database. CDS/genome: average of CDS per genome. Articles: number of relevant articles found. B. mallei and B. pseudomallei genomes are composed of two chromosomes. Asterisks indicate that the average of CDS/genome for these species is the sum of averages of CDS/chromosome (GC% is an average of over the two chromosomes).

The most sequenced or the most divergent ranked results are also useful to start a de novo primer design or to re-orient a strategy of detection. This can ease the selection of a target gene, for example by selecting most studied and conserved genes for a higher coverage. However, genes having several copies in the same genome can bias these rankings. Indeed, our method of collecting sequences by similarity gathers, in a single cluster, identical or nearly identical genes whose loci are different. It is typically the case of housekeeping genes from transposons and other selfish genetic elements (e.g. transposases), which are known to be present in several copies in bacterial genomes (Gilbert et al., 1990; Beuzon et al., 2004). Although ranks may be overvalued, this type of clustering was retained because these different loci correspond to copies of the same gene and the difference between sequences is limited to a few single nucleotide polymorphisms (SNPs).

From the data sheet of a gene (Fig. 1), users have access to several resources such as strains sequenced for the gene, links to the main biological databases (KEGG, GO, Interpro, PDB and Uniprot), links to the NCBI sequence viewer and relevant publications. It is also possible to download FASTA files containing only unique sequences, aligned, with published primers or every sequence for a gene. A text file, compatible with regular spreadsheet, is available with the coordinates and the strain name of each sequence. Each published PCR primer is displayed with Tms computed using five different methods, its position within aligned sequences, estimates of coverage and specificity, as well as links to publications in which they were found. Some genes had a majority of forward/reverse primers or only one published primer, although a minimum of two is required for PCR amplification. These results were seemingly caused by a design in non-coding regions, by the pres-
ence of an additional restriction site added to the primers leading to the failure of our automated retrieval process, or finally when a larger genomic fragment was amplified with primers located within two different genes (Gardès et al., 2012). However, users can combine two primers thanks to the alignment map of primers with their target gene, when several primers are available.

Patho-Genes.org: Tools

From the navigation panel, two query systems are available in Patho-Genes. First, entering a gene, a protein or an author name or selecting a species or a group of species allows searching for genes. Using the option ‘include alternate names’, it is possible to find every sequence of a gene using any alternate name used to describe a gene. To select only genes having published primers, users can check the case ‘published primer’. With this action and by not filling the sections gene, protein and author names, the list of genes having published primers for a species or for the whole database can be displayed too. Unlike current online retrieval system by keywords, our query system retrieves even non-annotated sequences of a gene because of the combination of similarity and keywords used to build this database. Second, querying by using a primer or a couple of primer’s sequences allows retrieving target genes, the amplicon sequences and the positions of primers in aligned sequences.

Published 2012. This article is a U.S. Government work and is in the public domain in the USA, Microbial Biotechnology, 5, 594–598
Patho-Genes.org: Conclusions

Patho-Genes.org is a website containing every information useful for the development of molecular detection tests based on PCR and the study of genes in major potential bioterror bacteria. Contrary to Multilocus Sequence Typing (MLST) databases (e.g. http://www.mlst.net/ or http://pubmlst.org/) that focus on seven housekeeping genes per species by proposing every allele, couples of primers and MLST profiles, Patho-Genes.org targets every annotated gene of a species but does not provide any MLST profile. In particular, our platform offers an analysis of the quality of published primers. Indeed, in the case of *Vibrio cholerae*, we discovered that only one-third of published PCR primers are able to amplify every allele of their target gene (Gardès et al., 2012). This tool thus allows to rapidly check the main characteristics of primers (specificity, coverage and thermodynamic parameters). Presently, primers are available only for four species. Primers for the other species of Table 1 will be available soon.

Methods

Ethics Statement: this study did not involve any living being or biological sample

Protein coding DNA sequences belonging to each potential pathogenic bacterium were collected using the ACNUC database and its retrieval system (Gouy and Delmotte, 2008). For each species, the occurrence of gene names was computed to obtain a raw list of most frequent names. From this list, tBLASTx analyses were performed to cluster similar sequences. Every annotation of such similar sequences was collected and the gene was internally described using the most used gene and protein names. Relevance and consistency of each term were checked among species. For each gene, sequences were then ‘de-replicated’: sequences contained into a longer sequence or identical sequences were removed in order to obtain a set of unique sequences. This set was aligned with MUSCLE version 3.8.31 (Edgar, 2004).

Using species name and annotations of similar sequences, requests were done with Entrez at NCBI (PubMed), Jane (Schuemie and Kors, 2008) and eTBLAST (Errami et al., 2007) in order to retrieve a combined list of relevant PubMed IDentification numbers (PMID) for each gene. Some requests yielded up to hundreds of publications. Each article was downloaded in PDF format and relevant short nucleic acid sequences were extracted from each file using regular expressions. Oligomers found at least once in the set of unique sequences were considered as published primers. Because different methods to calculate a Tm can give different results, the basic (Mann et al., 2009) (bas), the salt adjusted (Howley et al., 1979) (Sal) and three nearest-neighbour (Breslauer et al., 1986; Sugimoto et al., 1996; SantaLucia, 1998) (Bre, San and Sug) methods were used to compute Tm of published primers via dnaMATE (Panjkovich et al., 2005).

An Apache2 web server hosts Patho-Genes. From the navigation panel, two query systems are available. The first one is written in python and allows searching for genes using keywords: gene, protein or author names. It is possible to specify if the search should take in account every annotation used in INSDC databases, or be limited to genes having published primers. The program checks the existence of submitted keywords and selects corresponding entries. The second query system is based on a C program developed by our team, which returns genes containing a primer or a set of primers. Unlike BLAST, this software allows the use of the IUPAC code for degenerate positions.

Future directions

Patho-Genes will be updated regularly to incorporate the last release of the public databases. In the future, other organisms, like pulmonary pathogens or new emerging pathogens, will be included. The main goal of this project is to provide an exhaustive and user-friendly web resources for DNA-based detection technologies of pathogenic bacteria.

Acknowledgements

Our study was supported by a PhD fellowship from the Direction Générale de l’Armement (DGA), Ministère de la Défense Française. We thank the two referees for constructive comments that allowed us to improve this manuscript and our website.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410.

Arun, A., and Saurabha, D. (2003) PCR primer design. In *PCR Primer: A Laboratory Manual*. Diefenbach, C.W., and Dveksler, G.S. (eds). New York, USA: Cold Spring Harbor Laboratory Press, pp. 61–64.

Belgrader, P., Benett, W., Hadley, D., Richards, J., Stratton, P., Mariella, R., and Milanovich, F. (1999) PCR detection of bacteria in seven minutes. *Science* 284: 449–450.

Beuzon, C.R., Chessa, D., and Casadesús, J. (2004) IS200: an old and still bacterial transposon. *Int Microbiol* 7: 3–12.

Breslauer, K.J., Frank, R., Böcker, H., and Marky, L.A. (1986) Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci USA* 83: 3746–3750.

Christen, R. (2008) Identifications of pathogens – a bioinformatic point of view. *Curr Opin Biotechnol* 19: 266–273.

Published 2012. This article is a U.S. Government work and is in the public domain in the USA, *Microbial Biotechnology*, 5, 594–598.
Croce, O., Chevenet, F., and Christen, R. (2008) OligoHeat-Map (OHM): an online tool to estimate and display hybridizations of oligonucleotides onto DNA sequences. *Nucleic Acids Res* **36**: W154–W156.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.

Errami, M., Wren, J.D., Hicks, J.M., and Garner, H.R. (2007) eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. *Nucleic Acids Res* **35**: W12–W15.

Fredslund, J., Schauer, L., Madsen, L.H., Sandal, N., and Stougaard, J. (2005) PriFi: using a multiple alignment of related sequences to find primers for amplification of homologs. *Nucleic Acids Res* **33**: W516–W520.

Fricke, W.F., Rasko, D.A., and Ravel, J. (2009) The role of genomics in the identification, prediction, and prevention of biological threats. *PLoS Biol* **7**: e1000217.

Gardès, J., Croce, O., and Christen, R. (2012) *In silico* analyses of primers used to detect the pathogenicity genes of *Vibrio cholerae* [WWW document]. URL https://www.jstage.jst.go.jp/article/jsme2/advpub/0/advpub_ME11317/_article.

Gibert, I., Barbé, J., and Casadesús, J. (1990) Distribution of insertion sequence IS200 in *Salmonella* and *Shigella*. *J Gen Microbiol* **136**: 2555–2560.

Gouy, M., and Delmotte, S. (2008) Remote access to ACNUC nucleotide and protein sequence databases at PBIL. *Biochimie* **90**: 555–562.

Howley, P.M., Israel, M.A., Law, M.F., and Martin, M.A. (1979) A rapid method for detecting and mapping homology between heterologous DNAs. Evaluation of polymavirus genomes. *J Biol Chem* **254**: 4876–4883.

Mann, T., Humbert, R., Dorschner, M., Stamatoyannopoulos, J., and Noble, W.S. (2009) A thermodynamic approach to PCR primer design. *Nucleic Acids Res* **37**: e95.

Panjkovich, A., Norambuena, T., and Melo, F. (2005) dnaMATE: a consensus melting temperature prediction server for short DNA sequences. *Nucleic Acids Res* **33**: W570–W572.

SantaLucia, J. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci USA* **95**: 1460–1465.

Saravolatz, L.D., Manzor, O., VanderVelde, N., Pawlak, J., and Belian, B. (2003) Broad-range bacterial polymerase chain reaction for early detection of bacterial meningitis. *Clin Infect Dis* **36**: 40–45.

Schuemie, M.J., and Kors, J.A. (2008) Jane: suggesting journals, finding experts. *Bioinformatics* **24**: 727–728.

Seal, R.L., Gordon, S.M., Lush, M.J., Wright, M.W., and Bruford, E.A. (2011) genenames.org: the HGNC resources in 2011. *Nucleic Acids Res* **39**: D514–D519.

Sugimoto, N., Nakano, S., Yoneyama, M., and Honda, K. (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res* **24**: 4501–4505.

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., and Leunissen, J.A.M. (2007) Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res* **35**: W71–W74.

Zaph, C. (2010) Which species are in your feces? *J Clin Invest* **120**: 4182–4185.

Published 2012. This article is a U.S. Government work and is in the public domain in the USA, *Microbial Biotechnology*, 5, 594–598