In silico analysis of complete bacterial genomes: PCR, AFLP-PCR, and endonuclease restriction

Joseba Bikandi*, Rosario San Millán, Aitor Rementeria, and Javier Garaizar

Department of Immunology, Microbiology and Parasitology, University of the Basque Country, Paseo de la Universidad, 7, 01006 Vitoria-Gasteiz, Spain.
ABSTRACT

Summary: We have developed a website, www.in-silico.com, which runs a software program that performs three basic tasks in completely sequenced bacterial genomes by in silico analysis: PCR amplification, AFLP-PCR, and endonuclease restriction. For PCR, after selection of the genome and introduction of primers, fragment size, DNA sequence and corresponding ORF identity of the resulting PCR product is computed. Plasmids of sequenced species may be included in the analysis. Theoretical AFLP-PCR analyzes similar parameters, and includes a suggestion tool providing a list of commercial restriction enzyme pairs yielding up to 50 amplicons in the selected genome. Endonuclease restriction analysis of complete genomes and plasmids calculates the number of restriction sites for endonucleases in a given genome. If the number of fragments is 50 or fewer, PFGE image and restriction maps are illustrated. Other tools that have been included in this site are ORF search by name and DNA to protein translation as well as restriction digestion of user-defined DNA sequences.

Availability: This is a new molecular biology resource freely available over the Internet at http://www.in-silico.com

Contact: oipbibij@lg.ehu.es

* To whom correspondence should be addressed.
INTRODUCTION

As the number of public completely sequenced bacterial genomes increases, specific analysis software packages are needed to extract useful information from the vast amount of sequence data. We have developed the software necessary to simulate three techniques widely used in molecular biology: PCR amplification, Amplified Fragment Length Polymorphism (AFLP-PCR), and endonuclease restriction of complete genomes (focused on typing by Pulsed Field Gel Electrophoresis – PFGE—).

The website was developed with open-source software (PHP running on an Apache server with Linux operating system). Over 100 completely sequenced bacterial genomes are included at www.in-silico.com, and they are regularly updated. Sequences and other related files were obtained from NCBI, and the list of endonucleases was from the Restriction Enzyme Database (Roberts et al, 2003). Cleavage by endonucleases in our scripts was cross-referenced against NEBcutter (Vincze et al., 2003; http://tools.neb.com/NEBcutter).

In silico analysis by PCR amplification

The aim of in silico PCR is to calculate theoretical PCR results by using up-to-date sequenced bacterial genomes, a technique which allows amplification of specific DNA sequences. After choosing the bacterial genome, primers and maximum length of amplicons must be specified by the user. Degenerate nucleotides (N) may be included in the primer sequence, while the maximum length of amplicons is 10,000 bp. Plasmids may be included in the experiment.

The program will show the starting position of each amplicon, its length, and a simulation of the electrophoretic mobility. Further information from each amplicon may be obtained following the corresponding links: DNA sequences, list of ORFs that are included in the amplicon, and a link to
the NCBI site, which displays a map of the chromosome around the amplicon. To elevate the number of potential binding sites the program can allow one mismatch.

Results obtained at www.in-silico.com for in silico PCR were compared to those obtained by "virtual PCR" 2.0 version (vPCR; Lexa et al., 2001). The latter program uses BLAST (Altschul et al, 1990) to search for amplicons, while in silico PCR calculates amplicons within genomes stored in our server. When long primers were used in the experiment (18-22 nucleotides), theoretical results obtained by vPCR and in silico PCR matched. On the other hand, when short primers were used in the experiment (10-12 nucleotides), results did not match, as predicted by the authors in their website for short primers or overrepresented sequences.

**In silico AFLP-PCR and suggestion tool**

AFLP-PCR is a widely used fingerprinting method originally described by Vos et al.(1995). Briefly, after double digestion of genome, fragments are ligated to restriction half-site specific adaptors, and selective amplification of some of the fragments is performed. Finally, electrophoretic separation of amplicons is performed, followed by visualization of banding pattern. The aim of in silico AFLP-PCR is to simulate a theoretical experiment by using the same principles described by those authors. After choosing a bacterial genome, endonucleases and selective nucleotides (optional) are chosen, and partial sequences of adaptors and primers together with a list of resulting amplicons will be computed. Availability of information about the amplicons is similar to the one described for PCR amplification above, including a graphical simulation of the electrophoretic mobility pattern. Comparison of our script to AFLPinSilico (Rombauts et al., 2003; http://www.psb.rug.ac.be/bioinformatics/AFLPinSilico.html;) and to RemComb commercial software (Keygene, The Netherlands) show similar results.

An AFLP-PCR suggestion tool is included to provide for each genome a combination of endonucleases that theoretically produces a small number of fragments without using selective
nucleotides. This eliminates the need of cumbersome trials to find ample combinations of restriction enzymes prior to performing microbial typing experiments. A script was used to find all combinations of two endonucleases yielding a maximum of 50 fragments shorter than 3,000 bp for each genome included in www.in-silico.com, and results are searchable based on the number of resulting products. After selecting a product range, the program will suggest restriction enzyme pairs suitable for the experiment with the chosen microorganism. Electrophoretic band pattern and resulting amplicons are shown. Some combinations of endonucleases may generate potential errors in the calculations due to overlapping recognition sites. These are marked with flags.

**In silico endonuclease restriction and PFGE**

Pulsed Field Gel Electrophoresis typing is a technique widely used for strain delineation. Briefly, genome is endonuclease restricted and the fragments are separated by PFGE for visualization. This typing method requires previous selection of restriction enzymes producing a limited number of cleavages within genome. The in silico program uses as template DNA the sequence of selected genome and calculates the theoretical number of fragments produced by each endonuclease. When the number of amplicons are 50 or fewer, PFGE is graphically simulated. As in the previous technique, potential errors are flagged.

**Other tools**

Three additional general tools have also been included in this site:

a) ORF search by name within a specific genome. The resulting page shows the list of ORFs matching the query string, and the results are linked to NCBI for additional information.

DNA sequence and length. This service also allows introduction of a user-defined genetic code.

c) Endonuclease restriction of a user-defined input string of DNA by all commercially available enzymes. The list of enzymes was obtained from The Restriction Enzymes database (Roberts et al., 2003), and this service was checked against NetCutter (Vincze et al., 2003) until no errors were detected. This restriction tool was used as a reference to perform restriction digest calculations at in silico AFLP-PCR and PFGE tools.
Acknowledgements

We thank Steffen Porwollik and Michael McClelland for critical reading of this manuscript. A grant from University of the Basque Country, Spain, No. 9/UPV 00093.125.13542/2001 supported this work.
References


