BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 6 2003, pages 776–777
DOI: 10.1093/bioinformatics/btg090

AFLPinSilico, simulating AFLP fingerprints

Stephane Rombauts[1,*], Yves Van de Peer[1] and Pierre Rouzé[2]

[1]Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), K.L. Ledeganckstraat 35, Gent 9000, Belgium and [2]Laboratoire associé de l'INRA (France), K.L. Ledeganckstraat 35, Gent 9000, Belgium

Received on August 22, 2002; revised on October 29, 2002; accepted on December 13, 2002

ABSTRACT

Summary: A drawback of the Amplified Fragment Length Polymorphism (AFLP) fingerprinting method is the difficulty to correlate the different fragments with their DNA sequence. The AFLPinSilico application presented here simulates AFLP experiments run on either cDNA or genomic sequences, producing virtual fingerprints that allow high throughput identification of AFLP fragments. The program also enables biologists to manage experiments through simulations done beforehand, thereby reducing the number of experiments that have to be run. AFLPinSilico is available through the www or as a stand-alone version, through a command line executable (available upon request, for any platform running PERL).

Availability: For academic use http://www.psb.rug.ac.be/bioinformatics/AFLPinSilico.html.

Contact: strom@gengenp.rug.ac.be

Amplified fragment length polymorphism (AFLP), first described by Vos *et al.* (1995), is a fingerprint technique based on selective PCR amplification of DNA restriction fragments. AFLP consists of a double digestion of the DNA and ligation of the fragments to oligonucleotide adapters specific for each restriction enzyme. Those adapters, with nucleotides from the fragment ends, referred to as the selective nucleotides, serve as target sites for primer annealing and PCR amplification. As a consequence, only the subset of the DNA templates with the two different restriction sites will be amplified (Vos *et al.*, 1995). One drawback of this technique is the difficult characterization of the fragments, which need to be sequenced after DNA extraction from gel bands or capillary electrophoresis fractions. Nevertheless, AFLP is more and more performed in organisms for which complete sequences are available. In this case, the pattern and electrophoretic migration of AFLP fragments can be anticipated from the sequence, allowing identification of each fragment.

Here, we describe an application called AFLPinSilico that mimics AFLP experiments and produces virtual fingerprints that enable the identification of fragments based on their length and the choice of selective nucleotides. AFLPinSilico has already been successfully applied to reproduce cDNA-AFLP results and to test restriction enzymes as well as digestions of BAC sequences, contributing to the creation of high-resolution maps based on AFLP fingerprints (Peters *et al.*, 2001; http://www.arabidopsis.org/search/marker_search.html).

The algorithm searches for all the restriction sites, given as 'strings', and returns all fragments that would result from the digestions and selective amplification. The addition of the selective nucleotides is implemented by adding them to the 'string' that represents the restriction site, taking into account the strand, which is important for non-palindromic restriction sites. To run the AFLPinSilico program, the user needs to provide a file in FASTA-format with either genomic or full-length cDNA sequences and the sequence of both restriction sites of the chosen enzymes, using the extended UIPAC-code. The selective nucleotides, if relevant, should also be given. Optionally for higher accuracy, the length of the adapters can be taken into account. The output is the list of fragments ordered by their respective length, as it would appear with a real experiment (Fig. 1) The limitations of AFLPinSilico depend on the sequence availability and quality. For example, at this moment, modeling full genome AFLP mapping is feasible for *Arabidopsis* while genome-wide cDNA-AFLP fingerprinting is not since many cDNAs are lacking or not yet full length. Indeed, for cDNA-AFLP experiments 3′ UTR sequences are essential for the reliability of the output. As a consequence we can only expect reliable results on subsets of all the genes present in the genome, which in many cases might be sufficient. For analysis of large genome sequences, the main problem resides in fragments of similar size, which albeit different, do co-migrate. This lack of resolution is inherent to the AFLP technique itself. Therefore, in practice, smaller genomic sequences—about the size of a BAC—should be used to produce usable fingerprints and allow their safe modeling. Nevertheless, some unique fragments were found looking genome wide, and used as markers for the *Arabidopsis* genetic map (Peters *et al.*, 2001). When applying AFLP,

*To whom correspondence should be addressed.

1 ID of fragment	2. Fragment borders	3. Fragment sequence	4. Start of Fragment	5. End of Fragment	6. Length of Fragment
18	SacI-MseI	GAGCTCCACATC(	36268	36770	503
17	MseI-SacI	TTAAAGTAGTTGTT	35801	36273	473
10	SacI-MseI	GAGCTCTGTGTCT	17218	17685	468
21	MseI-SacI	TTAAGACTTTGTTG	62406	62847	442
40	SacI-MseI	GAGCTCCATCACT	86964	87388	425
7	MseI-SacI	TTAAAGACGGAGG	16721	17094	374
11	MseI-SacI	TTAAACTTCTTTTT	22110	22453	344
35	MseI-SacI	TTAAAGGCTTAGC	82399	82680	282
31	MseI-SacI	TTAATAAGCATCT(	79406	79674	269
26	SacI-MseI	GAGCTCGGTTTTC	67245	67506	262
20	SacI-MseI	GAGCTCGAGTCG/	52290	52547	258
19	MseI-SacI	TTAATTGCAATCTT	52040	52295	256
28	SacI-MseI	GAGCTCAAACCA/	75067	75301	235
30	SacI-MseI	GAGCTCGTTTTTA	75320	75547	228
25	MseI-SacI	TTAAACTAGCACG	67033	67250	218
38	SacI-MseI	GAGCTCCTCCTAT	84131	84318	188
6	SacI-MseI	GAGCTCGCATTAG	14454	14635	182
5	MseI-SacI	TTAAAAACAAGTT(	14278	14459	181
37	MseI-SacI	TTAATAAGTTGTCT	83960	84136	177
33	MseI-SacI	TTAAAACCGTGAC	81906	82079	174
34	SacI-MseI	GAGCTCCGATGG	82074	82245	172
27	MseI-SacI	TTAAAGATCCTTTG	74917	75072	156
12	SacI-MseI	GAGCTCGATGATC	22448	22595	148
3	MseI-SacI	TTAATAGTATTCTA	12724	12866	143
2	SacI-MseI	GAGCTCTTCCATA	2906	3013	108
39	MseI-SacI	TTAAGAAATCTTC/	86867	86969	103
41	MseI-SacI	TTAAGCCGTCGTC	87723	87816	94
9	MseI-SacI	TTAAAGATGTTTAC	17139	17223	85

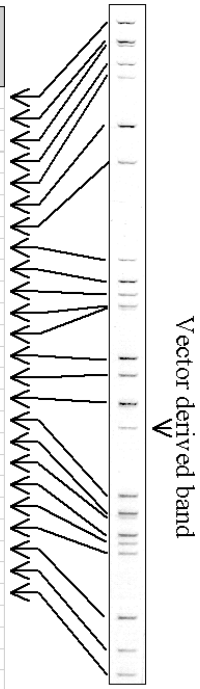

Fig. 1. Experimental and virtual AFLP data This figure shows on the right a typical AFLP gel electrophoresis pattern for a genomic DNA sequence (here the *Arabidopsis* BAC clone F26P21 (AL031804) after SacI and MseI double digestion, no selective nucleotides used) and on the left the AFLPinSilico output for this BAC. Every gel band has been extracted and sequenced to check if the experimental pattern was correctly modeled with the characteristic fragment length and sequence. This has been performed on different gels allowing us to conclude that unequivocal identification is possible for sequences up to the size of a BAC. The AFLPinSilico output is giving for each fragment: an ID (1) corresponding to the order of the fragment along the sequence, the enzyme restriction sites at the borders, with their relative order (2), the sequence (3), the respective position of the 5′ (4) and 3′ (5) borders relative to the DNA under analysis (here, the BAC clone) and the size of the fragment, in number of base pairs.

time-consuming and costly trials need to be performed. Simulation is therefore a suitable alternative to reduce the number of trials as different enzyme combinations can be tested beforehand. The number of primer combinations increases exponentially with the number of selective nucleotides, increasing the number of gels to be run (Bachem *et al.*, 1996). AFLPinSilico simulation might allow the choice of the most informative combinations. AFLPinSilico also allows to follow bands over different experiments or over a time-course. The combination of the semi-quantitative detection of transcripts by AFLP together with *in silico* simulation transforms the fingerprinting method into a worthy alternative for microarrays, finding even more efficiently the low abundant transcripts (Breyne and Zabeau, 2001). In conclusion, AFLPinSilico performs AFLP simulations with different combinations of restriction enzymes or selective nucleotides, on organisms or data sets for which DNA sequences are available. It allows identification of the fragments and the genes they are tagging, sparing the collection and sequencing task. Besides data analysis, such simulations can be used by experimentalists to efficiently plan and design AFLP based experiments, saving time and expensive trials. This is especially important when designing cDNA-AFLP experiments where the goal is to have, for all the genes of interest, one specific tag.

REFERENCES

Bachem,C.W., van der Hoeven,R.S., de Bruijn,S.M., Vreugdenhil,D., Zabeau,M. and Visser,R.G. (1996) Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. *Plant. J.*, **9**, 745–753.

Breyne,P. and Zabeau,M. (2001) Genome-wide expression analysis of plant cell cycle modulated genes. *Curr. Opin. Plant. Biol.*, **4**, 136–142.

Peters,J.L., Constandt,H., Neyt,P., Cnops,G., Zethof,J., Zabeau,M. and Gerats,T. (2001) A physical amplified fragment-length polymorphism map of *Arabidopsis*. *Plant. Physiol.*, **127**, 1579–1589.

Vos,P., Hogers,R., Bleeker,M., Reijans,M., Vandelee,T., Hornes,M., Frijters,A., Pot,J., Peleman,J., Kuiper,M. and Zabeau,M. (1995) AFLP: A new technique for DNA-fingerprinting. *Nucleic Acids Res.*, **23**, 4407–4414.