Genetics and Molecular Biology, 23, 2, 411-415 (2000)

Sequence characterization of hypervariable regions in the soybean genome: leucine-rich repeats and simple sequence repeats

Everaldo G. de Barros[1], Scott Tingey[2] and J. Antoni Rafalski[2]

Abstract

The genetic basis of cultivated soybean is rather narrow. This observation has been confirmed by analysis of agronomic traits among different genotypes, and more recently by the use of molecular markers. During the construction of an RFLP soybean map (*Glycine soja* x *Glycine max*) the two progenitors were analyzed with over 2,000 probes, of which 25% were polymorphic. Among the probes that revealed polymorphisms, a small proportion, about 0.5%, hybridized to regions that were highly polymorphic. Here we report the sequencing and analysis of five of these probes. Three of the five contain segments that encode leucine-rich repeat (LRR) sequence homologous to known disease resistance genes in plants. Two other probes are relatively AT-rich and contain segments of $(A)_n/(T)_n$. DNA segments corresponding to one of the probes (A45-10) were amplified from nine soybean genotypes. Partial sequencing of these amplicons suggests that deletions and/or insertions are responsible for the extensive polymorphism observed. We propose that genes encoding LRR proteins and simple sequence repeat region prone to slippage are some of the most hypervariable regions of the soybean genome.

INTRODUCTION

Studying the genetic diversity among and within plant species is informative not only from an evolutionary point of view but it is also useful for breeding purposes. Soybean (*Glycine max* L. Merrill) is an autogamous species (2n = 40) with 1.81 x 10^9 pairs of nucleotides in its genome distributed in repetitive (60%) and non-repetitive sequences (40%) (Goldberg, 1978). In spite of the extensive variability of the species, several studies have demonstrated the low diversity of cultivated soybeans (Delannay *et al*., 1983; Hiromoto and Vello, 1986; Abdelnoor *et al*. 1995; Powell *et al*., 1996b). In the United States, 88% of the cultivars used in the Northern part of the country and 70% of those used in the South derive from only 10 ancestors (Delannay *et al*., 1983). In Brazil, 80% of the cultivars recommended for 1983/84 introduction were derived from nine ancestors (Hiromoto and Vello, 1986).

More recent molecular marker data suggest that the limited genetic diversity of cultivated soybean (*G. max*) is due not only to selection during the breeding process but also due to its domestication from *G. soja* (Morgante *et al*., 1994; Powell *et al*., 1996b).

Several genetic maps have been constructed for soybean, using populations derived from interspecific (Shoemaker and Olson, 1993) as well as from intraspecific crosses (Lark *et al*., 1993). We constructed an RFLP map for soybean (*G. soja* x *G. max*) with low copy number probes generated from a *Pst*I soybean genomic library (Rafalski and Tingey, 1993). Among more than 2,000 probes analyzed, only a few (0.5%) were extremely polymorphic when tested in different soybean lines. In this work, we characterized five of these probes in order to understand the genetic basis of the polymorphism revealed by them.

MATERIAL AND METHODS

Probe isolation and sequencing

The five selected RFLP probes (A1-10, A2-08, A45-10, A53-09, and A75-10) hybridized to multiple, polymorphic DNA fragments on a Southern blot of soybean genomic DNA (Rafalski and Tingey, 1993).

Each DNA, cloned in the vector pBluescript (Stratagene), was sequenced using dye primer chemistry (ABI 373A, Perkin-Elmer) using T3 and T7 primers. The sequencing was completed with dye terminator chemistry using custom-designed oligonucleotide primers. The sequences were deposited in the GenBank under accession numbers: AF 215727, AF 215728, AF 215729, AF 217488, and AF 217489. They were searched for open reading frames (ORFs) and compared to sequences contained in the GenBank release 109 using Blast (Altschul *et al*., 1997).

DNA hybridization analysis

DNA samples from leaves of six different soybean genotypes (Bonus, PI 81.762, PI 416.937, N85-2176, PI 153.293 and PI 230.970) were extracted (Murray and Thompson, 1980) and digested with the restriction enzymes

[1]*Núcleo de Biotecnologia Aplicada à Agropecuaria, Universidade Federal de Viçosa, DBG/BIOAGRO, 36571-000 Viçosa, MG, Brasil. Send correspondence to E.G.B. Fax: +55-31-899-2864. E-mail: ebarros@mail.ufv.br*
[2]*DuPont Agricultural Biotechnology - Genomics, Delaware Technology Park, Suite 200, 1 Innovation Way, PO Box 6104, Newark, DE 19714-6104, USA.*

*Bam*HI, *Eco*R1, *Eco*RV, *Hin*dIII and *Pst*I (10 units per µg of DNA). The DNA fragments were separated on a 0.7% agarose gel (10 µg per lane), transferred to a nylon membrane (Hybond N^+, Amersham; Sambrook *et al*., 1989), and probed with plasmids A1-10, A2-08, A45-10, A53-09 or A75-10 labeled with ^{32}P- or GeneImages (Amersham) random primer labeling system. Labeling, washing and detection were according to the instructions contained in the GeneImages Kit (Amersham) or according to standard RFLP protocols (Rafalski *et al*., 1996).

PCR amplification of A45-10-related sequences

Primers 1 (5'-TTGACTGGTGATGTGCCTGT-3') and 2 (5'-CAGAATCCTATGCAAGCTCC-3') were used to amplify 25 ng of genomic DNA from soybean lines PI 230.970, PI 416.977, N85-2176, PI 51.293, PI 440.913, Bonus, Hardin, Noir1 and PI 81.762, using standard conditions, with annealing temperature of 55°C. Amplification products were purified (Qiagen) and cloned into pGEM-T using a commercial kit (Promega). Several individual transformants were picked for each soybean line, the insert size was verified by PCR, and plasmid DNA was sequenced using dye terminator chemistry (Perkin-Elmer/ABI) and T3/T7 and custom-designed primers.

RESULTS AND DISCUSSION

Five soybean clones isolated from a *Pst*I genomic library and used as RFLP probes detected multiple polymorphic regions in the soybean genome. We considered a probe hypervariable if it detected multiple polymorphisms among the six soybean lines tested. Most RFLP probes are either monomorphic (75%) or detect two allelic variants among the soybean lines tested (Powell *et al*., 1996a). Figure 1 shows the hybridization patterns obtained with a hypervariable probe A45-10.

It is well known that the soybean genome is relatively monomorphic due to the narrow genetic base of cultivated soybean and possibly due to the domestication process from *G. soja* (Morgante *et al*. 1994; Powell *et al*., 1996b). Therefore, identification of hypervariable regions is of considerable interest.

The five clones were sequenced and analyzed for sequence homology to known genes. As the genomic library from which the clones were isolated was a *Pst*I library, there was a high probability that they would map to transcriptionally active regions of the genome (Keim and Shoemaker, 1988). Two of the clones (A1-10 and A2-08) did not show significant homology to any sequences in GenBank. We also searched DuPont's collection of over 140,000 soybean ESTs and did not find significant sequence homology. One of these clones is AT rich (A2-08, 74% AT) and both clones contain regions of simple sequence repeats (SSRs). A1-10 contains a $(A)_{12}$ motif, and A2-08 contains three $(A/T)_8$ motifs and seven $(A/T)_7$ motifs. SSRs were found to be highly polymorphic (Powell *et al*., 1995; Powell *et al*., 1996b). We conclude that SSR-like sequences are likely to contribute to hypervariability of A1-10- and A2-08 homologous loci in soybean.

Analysis of clones A45-10, A53-09 and A75-10 revealed the presence of ORFs. Comparison of predicted amino acid sequences in all reading frames to sequence databases revealed homology to proteins coded by known disease resistance genes in plants (Table I).

The most noticeable feature of all the putative amino acid sequences, especially the one derived from clone A45-10, is the presence of leucine-rich repeats (LRRs). The protein sequence resulting from conceptual translation of the ORF present in this clone was arranged in a pattern of imperfect LRRs (Figure 2; Kajava, 1998). These repeats correspond to protein structural elements that are thought to be associated with protein-protein interactions (Kajava, 1998). Proteins coded by plant disease resistance genes frequently contain LRR. They normally fall into two classes: those with extracytoplasmatic LRRs with the 24-amino acid consensus LxxLxxLxxLxLxxNxLxGxIPxx, and

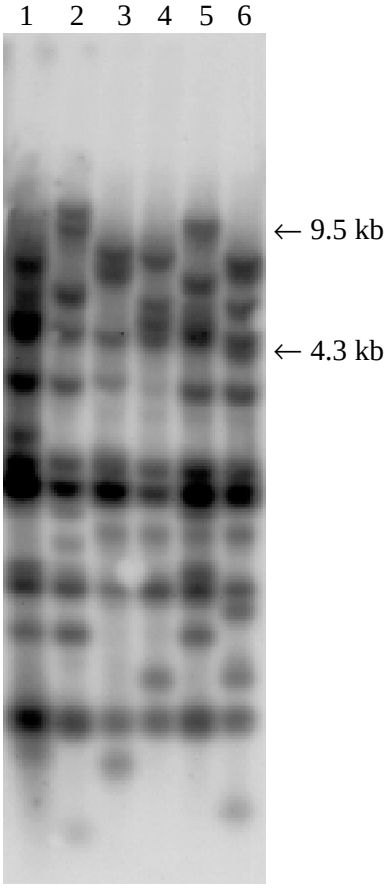

Figure 1 - Soybean genomic DNA hybridization analysis with RFLP probe A45-10. Lanes are as follows: *Hin*dIII digest of soybean genomic DNA from genotypes: 1, Bonus; 2, PI 81762; 3, PI 416937; 4, N85-2176; 5, PI 153293; 6, PI 230970.

Table I - Blast analyses of amino acid sequences deduced from soybean RFLP clones A45-10, A53-09, and A75-10.

Soybean clone	GenBank matches	GenBank entry No.	Species	P value
A45-10	AWJL218 protein	X81369	*Triticum aestivum*	$3e^{-75}$
	Cf - 2.2	U42445	*Lycopersicon pimpinellifolium*	$6e^{-64}$
	Cf - 2.1	U42444	*Lycopersicon pimpinellifolium*	$2e^{-64}$
	Cf - 9 protein precursor	U15936	*Lycopersicon pimpinellifolium*	$4e^{-59}$
A53-09	PG inhibiting protein	AF 020785	*Prunus armeniaca*	$1e^{-12}$
	LRR protein	X95269	*Lycopersicon esculentum*	$1e^{-12}$
	Cf - 9 protein precursor	U15936	*Lycopersicon pimpinellifolium*	$2e^{-12}$
	PG inhibiting protein	L26529	*Lycopersicon esculentum*	$9e^{-12}$
A75-10	TMV resistance protein N	U15605	*Nicotiana glutinosa*	$7e^{-35}$
	L6 protein	U27081	*Linum usitatissimum*	$1e^{-29}$
	Rust resistance protein M	U73916	*Linum usitatissimum*	$1e^{-26}$
	Downy mildew resistance protein RPP5	U97106	*Arabidopsis thaliana*	$8e^{-23}$

those with cytoplasmatic LRRs with the 23 or 24-amino acid consensus LxxLxxLxxLxLxx(N/C/T)x(x)LxxIPxx regions (Jones and Jones, 1997).

Disease resistance genes are expected to evolve much more rapidly than, for example, genes of central metabolism, because of the selection pressure from the pathogen (Michelmore and Meyers, 1998). In fact, several rice disease resistance-like genes do not cross-hybridize to maize genomic DNA (Tarchini, R., unpublished observations), while genes encoding metabolic enzymes from corn and rice cross-hybridize (Chen *et al.*, 1997). A45-10, A53-09 and A75-10 may in fact be disease resistance genes, explaining the high variability of RFLP patterns when they are used as hybridization probes.

To understand the molecular nature of the polymorphism, DNA segments homologous to A45-10 were isolated. To this end, oligonucleotide primers corresponding to A45-10 were designed and PCR was performed using DNA from nine soybean cultivars and PIs. Between one and three amplification products per cultivar were produced. Individual amplification products were cloned and sequenced (data not shown). Comparison of the DNA sequences revealed numerous insertions/deletions and single nucleotide changes among the different size clones from the same PCR. The interpretation of the result is complicated by the difficulty of assigning allelic relationships between multiple amplification products. The hybridization results (Figure 1) indicated that A45-10 was a member of a moderate-size family of related sequences, and several family members were represented among the amplification products. In the case of two *G. soja* accessions, PI 440.913B and PI 81.762, only one amplification product was identified from each accession. These were assumed to be allelic and compared. Several small deletions, 1-13 bp, explain the size differences between these PCR products. Many single nucleotide changes are also present. Such sequence variants could be the result of selection

```
LQVLNL---GANSLTGD-VPVT--LGTLSN
LVTLDL---SSNLLEGS-IKESNFVKLFT
LKELRL---SWTNLFLS-VNSGWAPPFQ
LEYVLL---SSFGIGPK-FPEW-LKRQSS
VKVLTM---SKAGIADL-VPSWFWIWTLQ
IEFLDL---SNNLLRGD-LSN---IFLN
SSVINL---SSNLFKGR-LPS---VSAN
VEVLNVA---NNSISGT-ISPF--LCGNPNATNK
LSVLDF---SNNVLSGD-LGHCW-VHWQA
LVHVNLG--SNN-LSGE-IPNS--MGYLSQ
LESLLL---DDNRFSGY-IPST--LQNCST
MKFIDM---GNNQLSDT-IPDW--MWEMQY
LMVLRLR--SNN-FNGS-IAQK--MCQLSS
LIVLDL---GNNSLSGS-IPNC--LDDMKTMAGEDDFFANPSSYSYGSDFSYNHYKET
LVLV-----PKK---DE-LEY---RDNLIL
VRMIDL---SSNKLSGA-IPSE--ISKLFA
LRFLNL---SRNHLSGE-IPND--MGKMKL
LESLDLSL--NN-ISGQ-IPQS--LSDL-SF
LSFLNL---SYHNLSGR-IPTST---QLQSF
-DELSYT--GNPLCGPPVTKNCTNKEWLRE
SASVGHGDGNFFGTSEFYIGMGVGFAAGFWGF
CSVVFFNRTWRLAYFHYLDHLRDLIYVMIVLKVRRLLGKL
```

Figure 2 - The amino acid sequence of clone A45-10 arranged in the pattern of leucine-rich repeats.

acting upon products of intragenic unequal crossing over, or replication errors caused by the repetitive nature of these sequences. This is in agreement with mechanisms proposed recently by Michelmore and Meyers (Meyers *et al*., 1998; Michelmore and Meyers, 1998).

To further validate putative identification of clones A45-10, A53-09 and A75-10 as fragments of actively transcribed genes, we searched the DuPont's soybean EST database for related sequences. Two cDNA clones isolated from developing soybean pods correspond closely to A53-09. cDNA clone sdp2c.pk007.p18 was 100% homologous to A53-09 from nucleotide 49 to 522. cDNA

clone sdp2c.pk007.a22 was 95% homologous to A53-09 from nucleotide 49 to 526. Both cDNA clones were homologous to *Arabidopsis* Cf-2.1-like protein, GenBank accession AC004238 (sdp2c.pk007.p18, pLog 14.4; sdp2c.pk007.a22, pLog 15.89), and to several tomato disease resistance genes, especially Hcr9-4C (accession AJ002235). This shows that A53-09 is an actively transcribed gene and a member of disease resistance-like gene family.

One EST, sls1c.pk010.j1, isolated from soybean infected with *Sclerotinia sclerotiorum* has 83% similarity at the nucleotide level to clone A75-10. This cDNA was similar to TMV resistance protein N (Table I) and shows that A75-10 is also likely to represent a disease resistance gene, although we do not have a direct evidence for its transcriptional activity.

No ESTs corresponding to A45-10 were identified in our collection. Nevertheless, highly significant homologies to LRR-containing disease resistance genes were identified throughout the length of the ORF (Table I).

Map-based cloning, transposon tagging, and PCR amplification of conserved regions have been used to clone a great number of disease resistance genes and disease resistance gene analogs in the past few years (Martin *et al*., 1993; Kanazin *et al*., 1996; Yu *et al*., 1996; Liester *et al*., 1996, 1998). As we demonstrated here, sequences related to these genes explain some of the polymorphism present in the soybean genome. It is particularly striking that of the five most highly polymorphic probes studied here, three contain LRRs. Therefore, the use of disease resistance gene homologs and LRRs in particular as probes for genetic mapping and especially fingerprinting of accessions and cultivars may provide two significant benefits. These probes may reveal frequent polymorphisms, and these polymorphisms are likely to be related to agronomically relevant disease resistance phenotypes. Similarly, searching for LRR-containing disease resistance gene homologs, by degenerate PCR or other methods, provides an approach to the isolation of highly polymorphic mapping probes.

ACKNOWLEDGMENTS

We would like to thank Sylvia Stack and Maureen Dolan for DNA sequencing, Mike Hanafey for the development of EST data bases, and to Blake Meyers, Michele Morgante and Renato Tarchini for discussion of disease resistance genes and comments on the manuscript. Everaldo G. de Barros was the recipient of a fellowship from CAPES.

RESUMO

A base genética da soja cultivada é relativamente estreita. Essa observação foi confirmada por análises de características agronômicas entre diferentes genótipos e, mais recentemente, pelo uso de marcadores moleculares. Durante a construção de um mapa de RFLP da soja (*Glycine soja* x *Glycine max*), os dois progenitores foram analisados com mais de 2000 sondas, das quais 25% eram polimórficas. Entre as sondas que revelaram polimorfismos, uma pequena proporção, cerca de 0,5%, hibridizou com regiões que eram altamente polimórficas. Neste trabalho, são apresentados o seqüenciamento e análise de cinco dessas sondas. Três dessas sondas contêm segmentos que codificam repetições ricas em leucina que são homólogas a genes de resistência a doenças já conhecidos em plantas. As duas outras sondas são relativamente ricas em AT e contêm segmentos do tipo (A)n/(T)n. Segmentos de DNA correspondentes a uma das sondas (A45-10) foram amplificados a partir de nove genótipos de soja. Seqüenciamento parcial desses amplicons sugere que deleções e/ou inserções são responsáveis pelo extensivo polimorfismo observado. Nós propomos que os genes que codificam proteínas com repetições ricas em leucina e regiões de seqüências repetidas simples, que são passíveis do fenômeno de *slippage* (deslizamento), estão entre as regiões mais variáveis do genoma da soja.

REFERENCES

Abdelnoor, R.V., Barros, E.G. and **Moreira, M.A.** (1995). Determination of genetic diversity within Brazilian soybean germplasm using random amplified polymorphic DNA techniques and comparative analysis with pedigree data. *Rev. Bras. Genet. 18*: 265-273.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and **Lipman, D.J.** (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res. 25*: 3389-3402.

Chen, M., SanMiguel, P., De Oliveira, A.C., Woo, S.-S., Zhang, H., Wing, R.A. and **Bennetzen, J.L.** (1997). Microcolinearity in sh-homologous regions of the maize, rice and sorghum genomes. *Proc. Natl. Acad. Sci. USA 94*: 3431-3435.

Delannay, X., Rodgers, D.M. and **Palmer, R.G.** (1983). Relative genetic contribution among ancestral lines to North American soybean cultivars. *Crop Sci. 23*: 944-949.

Goldberg, R.D. (1978). DNA sequence organization in the soybean plant. *Biochem. Genet. 16*: 45-68.

Hiromoto, D.M. and **Vello, N.A.** (1986). The genetic base of Brazilian soybean (*Glycine max* (L.) Merrill) cultivars. *Rev. Bras. Genet. IX*: 295-306.

Jones, D.A. and **Jones, J.D.G.** (1997). The role of leucine-rich repeat proteins in plant defences. *Adv. Bot. Res. 24*: 89-167.

Kajava, A.V. (1998). Structural diversity of leucine-rich repeat proteins. *J. Mol. Biol. 277*: 519-527.

Kanazin, V., Marek, L.F. and **Shoemaker, R.C.** (1996). Resistance gene analogs are conserved and clustered in soybean. *Proc. Natl. Acad. Sci. USA 93*: 11746-11750.

Keim, P. and **Shoemaker, R.C.** (1988). *Construction of a Random Recombinant DNA Library that is Primarily Single Copy Sequence*. Soybean Genetics Newsletter, No. 15. Department of Agronomy, Iowa State University and the United States Department of Agriculture, Agricultural Research Service, Ames, Iowa. pp. 147-148.

Lark, K.G., Weisemann, J.M., Matthews, B.F., Palmer, R., Chase, K. and **Macalma, T.** (1993). A genetic map of soybean (*Glycine max* L.) using a intraspecific cross of two cultivars: 'Minsoy' and 'Noir 1'. *Theor. Appl. Genet. 86*: 901-906.

Liester, D., Ballvora, A., Salamini, F. and **Gebhardt, C.** (1996). A PCR-based approach for isolating pathogen resistance genes from potato with potential for wide application in plants. *Nat. Genet. 14*: 421-429.

Liester, D., Kurth, J., Laurie, D.A., Yano, M., Sasaki, T., Devos, K., Graner, A. and **Schulze-Lefert, P.** (1998). Rapid organization of resistance gene homologues in cereal genomes. *Proc. Natl. Acad. Sci. USA 95*: 370-375.

Martin, G.B., Brommonschenkel, S.H., Chunwongse, J., Frary, A., Ganal, M.W., Spivey, R., Wu, T., Earle, E.D. and **Tanksley, S.D.** (1993). Map-based cloning of a protein kinase gene conferring disease resistance in tomato. *Science 262*: 1432-1436.

Meyers, B.C., Shen, K.A., Rohani, P., Gaut, B.S. and **Michelmore, R.W.**

(1998). Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection. *Plant Cell 11*: 1833-1846.

Michelmore, R.W. and **Meyers, B.C.** (1998). Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome 8*: 1-18.

Morgante, M., Rafalski, A., Biddle, P., Tingey, S. and **Olivieri, A.M.** (1994). Genetic mapping and variability of seven soybean simple sequence repeat loci. *Genome 37*: 763-769.

Murray, M.G. and **Thompson, W.F.** (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res. 8*: 4321-4325.

Powell, W., Morgante, M., Andre, C., McNicol, J.W., Machray, G.C., Doyle, J.J., Tingey, S.V. and **Rafalski, J.A.** (1995). Hypervariable microsatellites provide a general source of polymorphic DNA markers for the chloroplast genome. *Curr. Biol. 5*: 1023-1029.

Powell, W., Morgante, M., Andre, C., Hanafey, M., Vogel, J., Tingey, S. and **Rafalski, A.** (1996a). The comparison of RFLP, RAPD, AFLP and ssr (microsatellite) markers for germplasm analysis. *Mol. Breed. 2*: 225-238.

Powell, W., Morgante, M., Doyle, J.J., McNicol, W., Tingey, S.V. and **Rafalski, A.J.** (1996b). Genepool variation in genus *Glycine* subgenus soja revealed by polymorphic nuclear and chloroplast microsatellites. *Genetics 144*: 793-803.

Rafalski, J.A. and **Tingey, S.V.** (1993). RFLP map of soybean (*Glycine max*) 2N = 40. In: *Genetic Maps* (O'Brien, S.J., ed.). Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp. 149-156.

Rafalski, J.A., Vogel, J.M., Morgante, M., Powell, W., Andre, C. and **Tingey, S.V.** (1996). Generating and using DNA markers in plants. In: *Nonmammalian Genome Analysis. A Practical Guide* (Birren, B. and Lai, E., eds.). Academic Press, San Diego, pp. 75-134.

Sambrook, J., Fritsch, E.F. and **Maniatis, T.** (1989). *Molecular Cloning. A Laboratory Manual*. 2nd edn. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Shoemaker, R.C. and **Olson, T.** (1993). Molecular linkage map of soybean (*Glycine max* L. Merr.). In: *Genetic Maps* (O'Brien, S.J., ed.). Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp. 131-136.

Yu, Y.G., Buss, G.R. and **Saghai-Maroof, M.A.** (1996). Isolation of a superfamily of candidate disease-resistance genes in soybean based on a conserved nucleotide-binding site. *Proc. Natl. Acad. Sci. USA 93*: 11751-11756.

(Received January 6, 2000)