BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 6(9)

Survey and characterization of NBS-LRR (*R*) genes in *Curcuma longa* transcriptome

Raj Kumar Joshi, Basudeba Kar, Sanghamitra Nayak*

Centre of Biotechnology, School of Pharmaceutical Sciences, Siksha O Anusandhan University, Bhubaneswar-751003, India; Sanghamitra Nayak - Email: sanghamitran@yahoo.com; Phone: 09437061976; *Corresponding author

Received July 07, 2011; Accepted July 12, 2011; Published July 19, 2011

Abstract:
Resistance genes are among the most important gene classes for plant breeding purposes being responsible for activation of plant defense mechanisms. Among them, the nucleotide binding site-leucine rich repeat (NBS-LRR) class *R*-genes are the most abundant and actively found in all types of plants. *Insilico* characterization of EST database resulted in the detection of 28 NBS types *R*-gene sequences in *Curcuma longa*. All the 28 sequences represented the NB-ARC domain, 21 of which were found to have highly conserved motif characteristics and categorized as regular NBS genes. The Open Reading Frames varied from 361 (CL.CON.3566) to 112 (CL.CON.1267) with an average of 279 amino acids. Most alignment occurred with monocots (67.8%) with emphasis on *Oryza sativa* and *Zingiber* sequences. All best alignments with dicots occurred with *Arabidopsis thaliana, Populus trichocarpa and Medicago sativa*. These detected NBS type *R*-genes from *Curcuma longa* can be used as a valuable resource for molecular marker development, molecular mapping of *R*-genes, and identification of resistance gene analogs and functional and evolutionary characterization of NBS–LRR-encoding resistance genes in asexually reproducing plants.

Keywords: *Curcuma longa*, expressed sequence tags, NBS-LRR, R-genes, TBLASTN

Background:
Pathogen attack has caused an estimated 12% loss of the global crop production in the last decade with up to 80% loses accounted from the tropical countries **[1]**. The most important group of genes that has been used by breeders for disease control is the plant resistance (*R*) genes. Resistance genes which are members of a very large multigene family are highly polymorphic and have diverse recognition specificities. As many as 70 different *R* genes showing resistance to major plant pathogens has been isolated, cloned, and characterized in different plants in the last 15 years **[2]**. These can be classified into five categories based on their predicted protein structure **[3, 4]**. Of the cloned plant disease resistance (R) genes, approximately 75% encode cytoplasmic receptor-like proteins characterized by an N-terminal nucleotide-binding site (NBS), leucine-rich repeat (LRR) domain and a leucine zipper (LZ), Toll interleukine 1-receptor (TIR) or a coiled-coil (CC) sequence **[5]**. The LRR region recognizes the pathogens, the TIR and CC regions are involved in signal transduction during many cell processes **[6]**, while the NBS usually signalizes for programmed cell death **[7]**. Many genes encode proteins of this class: *I2* **[8]** and *Sw5* **[9]** from tomato; *RPM1* **[10],** *RPS2* **[11]** and *RPS4* **[12]** from *Arabidopsis thaliana*; *Pib* **[13]**, *Pi-ta* **[14]** and *Xa1* **[15]** from *Oryza sativa* (rice); *Hero* **[16]**, *R1* **[17]** and *Rx2* **[18]** from potato, *L* **[19]**, and *P* **[20]** of flax, *N* **[21]** of tobacco etc. Infact, the whole-genome sequence analysis revealed that there are 150–175 NBS–LRR genes in the *Arabidopsis* genome **[22]** and approximately 600 NBS–LRR genes in the rice genome **[23]**. *Curcuma longa* L. (turmeric) of the family Zingiberaceae is one of the most important crop with great medicinal and economic significance. Turmeric rhizome is valued world over and has been in use from ancient time as a spice, food preservative, coloring agent, and in the traditional systems of medicine. India is the world's largest producer, and exporter of turmeric followed by China, Indonesia, Bangladesh and Thailand **[24]**. The International Trade Centre, Geneva, has estimated an annual growth rate of 10% in the world demand for turmeric. Continuous domestication of the preferred genotypes coupled with their exclusive vegetative nature seems to have eroded the genetic base of these crops and as a result, all of their cultivars available today are equally susceptible to major diseases such as rhizome rot caused by *Pythium aphanidermatum*, leaf blotch caused by *Taphrina maculans* and leaf spot caused by *Colletotrichum capsici*. Moreover, turmeric is completely sterile and is propagated exclusively by vegetative means using rhizome. In this context, characterization of resistance-related sequences may provide a lead towards retrieving resistance specificities suitable for the improvement of this crop. Recent advances in *Curcuma* genomic technologies have generated a large number of expressed sequence tags (ESTs) that have been made available in public database. As of July 2011, GenBank had released 12,593 EST sequences from *Curcuma longa*. This database can be used as a starting material for the characterization of NBS-LRR class *R* gene sequences in turmeric. Thus, our objective is to perform a data mining-based identification of plant NBS-LRR class *R*-genes in *Curcuma longa* EST database, by using well known *R*-genes sequences as template, comparing the identified sequences with known *R*-genes deposited in public DNA and protein databases.

Methodology:
Curcuma longa transcriptome database was searched for NBS-LRR *R*-gene homologues using Amino-acid sequences of known genes as query. Accession numbers of sequences used at NCBI (National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov) are shown in **Table 1 (see Supplementary material)**, together with sequences features and accession numbers. They are grouped according to the conserved domains previously

described. All turmeric sequences used during this work were obtained from *Curcuma longa* EST database. EST database of NCBI contains 12953 *Curcuma longa* express sequence tag data. We have mined 12593 EST sequences consisting of two tissue libraries of rhizomes 6870 (DY395309-DY388440) and leaves 5723 (DY388439-DY382717). The EST sequences were screened against the UniVec database from NCBI (ftp://ftp.ncbi.nih.gov/pub/ UniVec/) for detecting vector and adapter sequences by using the program Cross_Match. CAP3 program was used to assemble the EST sequence into contigs for creating a non-redundant dataset. The program TBLASTN **[25]** was used to perform reverse alignment on *Curcuma longa* contigs. The clusters frame of the TBLASTN alignment was used to predict the Open Reading Frames (ORFs) for each searched contig. For this purpose, the Expasy Translate Tool (bo.expasy.org/tools/dna.html) was used, which predicts the correct ORF for a DNA sequence in the corresponding amino acid FASTA sequence. The obtained ORFs were subsequently submitted to a Reverse Position Specific BLAST (RPS-BLAST) against Conserved Domain Database **[26]** aiming to identify patterns or motifs in predicted cluster products. Reciprocal alignments were conducted for ORFs by using the nr databank and stand-alone BLAST package from NCBI. Matched sequences were annotated for latter comparison.

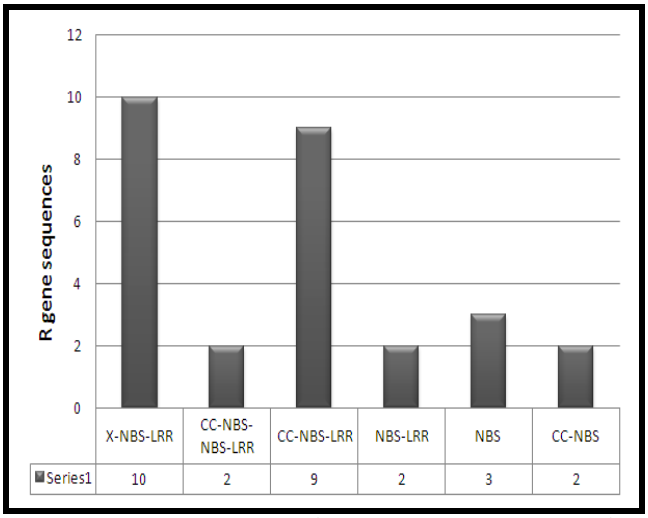

Figure 1: Graphical representation of the NBS-LRR *R*-genes retrieved from *Curcum longa* EST database.

Results and Discussion:

R-genes are quite abundant in higher plants but the most functionally defined R genes belong to a class that encode cytoplasmic receptor-like proteins characterized by an N-terminal nucleotide-binding site (NBS) and a leucine-rich repeat (LRR) domain. A set of 28 non-redundant NBS sequences were retrieved through TBLASTN alignment of 4035 *Curcuma longa* contig sequences. They have been annotated for one or more than one *R*-gene (data summarized in **Table 2 (see Supplementary material).** Earlier, five resistance gene analogues (RGAs) have been already isolated and characterized in *Curcuma longa* **[27]**. However, all the five RGAs isolated were of the CC-NBS-LRR class without exhibiting significant variations in the NBS type R-gene domain characterization**.** In contrast, it was expected that some similar genes grouped at the same class should cause some level of redundancy **[28]**. Contigs representing exclusive NBS type *R* genes with variability were (I) X-NBS-LRR: 10; CC-NBS-NBS-LRR: 2; CC-NBS-LRR: 9; NBS-LRR: 2; NBS: 3 and CC-NBS: 2 **(Figure 1)** In 21 out of 28 NBS genes, all the motifs characteristic of the NBS domain were conserved and categorized as regular NBS genes. The others were very different in their structures from the majority, or were simply truncated and categorized as non-regular NBS genes. Two non-regular NBS genes yielded higher P values when they were hit by TBLASTN in the NBS regions and had standard LRR regions while 3 genes had only some of the conserved NBS motifs. Two non-regular NBS genes encoded a coiled motif but were highly divergent in NBS region and lacked LRR regions. In the N-terminal region, 10 regular NBS genes contained some unknown motifs, which were symbolized as X. 11 regular NBS genes encoded the CC motif (CNL and CNNL) while the rest where without the CC motif (XNL). No genes were encoded with the TIR motifs. TIR motif is supposed to be absent in monocotyledonous plants **[4]**, being present in all dicotyledonous taxa actually studied. Sizes of *Curcum longa* contig aligned to NBS-LRR R-genes varied from 1256 (CL.CON.1529) to 452 nucleotides (CL.CON.1267). The prediction of contig coding regions revealed that ORFs were coded in both forward and reverse reading frames, with an average of 279 amino acids (aa) in length. ORF sizes varied from 361 (CL.CON.3566) to 112 amino acids (CL.CON.1267). The search for conserved domains (CD-Search) revealed conserved motifs in all the analyzed contig clusters. All the 28 contig *Curcuma longa* clusters represented the NB-ARC domain. In the LLR region, Pfam software detected 32 LRR motifs in the 28 NBS genes. This number is higher than the number of *Curcuma longa* contigs with NBS-LRR R genes, due to their occurrence in tandem repetitions. Sometimes these LRR sequences are imperfect and may be difficult to recognize with available *insilico* tools, so it is possible that a larger number may be identified manually. Two of the contig clusters CL.CON.1267 and CL.CON.3620 with a poorly developed NBS motif represented very short ORFs of 112 and 123 amino acids respectively. Considering the best matches to the 28 *Curcuma longa* NBS-LRR contigs identified, 9 were from plants of dicotyledenous families such as *Arabidopsis thaliana*, *Populus trichocarpa, Pyrus communis*, *Glycine max*, *Cajanus cajans* and *Medicago sativa*. From monocots, rice (*O. sativa*) sequences appeared as best matches (9 contig clusters) followed by *Zingiber officinale* (3 contig clusters). A comprehensive list of all the sequences that aligned with *Curcuma longa* NBS-LRR contig clusters are represented in **table 2 (see Supplementary material)**. The comparison of our results regarding the organization of detected *Curcuma longa* NBS-LRR genes was mainly with rice and ginger. It has been observed that most of the information regarding *R*-genes available in databases refers to herbaceous model and crop plants such as rice and *Arabidopsis*, may be because most identified and sequenced *R*-genes were a consequence of mapping approaches that have been abundantly performed in these plants. The larger number of sequences from *Oryza sativa* representing best alignments to *Curcuma* does not represent a higher similarity to this plant species, but it reflects the large number of sequences of this model plant deposited in GenBank. Barbosa-da-Silva *et al*., 2005 **[29]** has also found that *Eucalyptus* even being a woody plant exhibited maximum alignment of R-genes with herbaceous *Arabidopsis thaliana*. There can be other arguments as well such as (i) *Curcuma* belongs to the same family as ginger (Zingiberaceae) and (ii) both *Curcuma* and rice are monocots and exhibit similar levels of complexity. However, we cannot also rule out the fact that significant sequence similarity was also detected with dicot plants. This suggests that, *Curcuma longa* might be positioned at the transition point between dicots and monocots as far as resistance genes are concerned. However, detail characterization of the NBS–LRR gene in turmeric has to be made before making a valid conclusion on its evolutionary aspect. The number of NBS type *R*-genes identified here is quite low considering the total size of the EST database. However, there can be other types of *R*-genes in *Curcuma longa*, which were not targeted in this study. Moreover, the EST database has not been obtained under pathogen stress condition. This may suggest that the identified NBS sequences are expressed constitutively but also leads to the supposition that a higher number of *R*-genes may be present in *Curcuma* under other experimental conditions. Thus, the generation of additional ESTs especially under infection by pathogen, can make it possible to detect many new NBS genes from *Curcuma longa.*

Conclusion

Using bioinformatics tools, it was possible to detect and characterize NBS type *R*-genes from *Curcuma longa* transcriptome. Twenty eight (28) NBS type *R*-genes were detected with distinct NB-ARC domain, 21 of which were regular NBS genes. This *insilico* method of detecting NBS-LRR type *R* genes in *Curcuma longa* has been done for the first time in this study. The identified sequences will be valuable resources for the development of markers for molecular breeding and identification of RGAs (resistance gene analogs) in *Curcuma* and other related species. A few of the NBS type *R*-genes of *Curcuma* isolated in this study may also be used for fluorescent *insitu* hybridization (FISH) on *Eucalyptus* chromosomes, also helping in the comparison of different parental species and the respective hybrids. Further, these *insilico* detected NBS type *R*-genes will reveal furthers insights on the organization, function and evolution of the NBS–LRR-encoding resistance genes in asexually reproducing plants.

Acknowledgement:

The authors are grateful to Dr. Manoj Ranjan Nayak, President, Siksha O Anusandhan University for his encouragement and support.

References:

[1] www.crcpress.com/product/pest management

[2] Liu J *et al. J Genet Genomics*. 2007 **34**: 765 [PMID: 17884686]

[3] Song WY *et al. Plant Cell* 1997 **9**: 1279 [PMID: 9286106]
[4] Ellis J & Jones D. *Curr Opin Plant Biol*. 1998 **1**: 288 [PMID: 10066601]
[5] Hammond-Kosack KE & Jones JD. *Annual Rev of Plant Physiol Plant Mol Biol*. 1997 **48**: 575 [PMID: 15012275]
[6] Martin GB *et al. Annu Rev Plant Biol*. 2003 **54**: 23 [PMID: 14502984]
[7] van der Biezen EA & Jones JD. *Curr Biol*. 1998 **8**: R226 [PMID: 9545207]
[8] Ori N *et al. Plant Cell*. 1997 **9**: 521 [PMID: 9144960]
[9] Brommonschenkel SH *et al. Mol Plant Microbe Interact*. 2000 **13**: 1130 **[**PMID: 11043474]
[10] Grant MR *et al. Science* 1995 **269**: 843 [PMID: 7638602]
[11] Mindrinos M *et al. Cell* 1994 **78**: 1089 [PMID: 7923358]
[12] Gassmann W *et al. Plant J*. 1999 **20**: 265 [PMID: 10571887]
[13] Wang ZX *et al. Plant J*. 1999 **19**: 55 [PMID: 10417726]
[14] Bryan GT *et al. Plant Cell*. 2000 **12**: 2033 [PMID: 11090207]
[15] Yoshimura S *et al. Proc Natl Acad Sci U S A* 1998 **95**: 1663 [PMID: 9465073]
[16] Ernst K *et al. Plant J*. 2002 **31**: 127 [PMID: 12121443]
[17] Ballvora A *et al. Plant J*. 2002 **30**: 361 **[**PMID: 12000683]
[18] Bendahmane A *et al. Plant J*. 2000 **21**: 73 [PMID: 10652152]
[19] Lawrence GJ *et al. Plant Cell*. 1995 **7**: 1195 [PMID: 7549479]
[20] Dodds P *et al. Plant Cell*. 2001 **13**: 163 [PMID: 11158537]
[21] Whitham S *et al. Proc Natl Acad Sci U S A*. 1996 **93**: 8776 [PMID: 8710948]
[22] Meyers BC *et al. Plant Cell*. 2003 **15**: 809 [PMID: 12671079]
[23] Zhou T *et al. Mol Genet Genom*. 2004 **271**: 402 [PMID: 15014983]
[24] http://www.printsasia.com/book/Indian-Spices-Production-and-Utilization-H-P-Singh-K-Sivaraman-M-Tamil-Selvan
[25] Altschul SF *et al. Nucleic Acids Res*. 1997 **25**: 3389 [PMID: 9254694]
[26] Marchler-Bauer A *et al. Nucleic Acids Res*. 2002 **30**: 281 [PMID: 11752315]
[27] Joshi RK *et al*. *Genet Mol Res*. 2010 **9**: 1796 [PMID: 20830672]
[28] Meyers BC *et al. Plant J*. 1999 **20**: 317 [PMID: 10571892]
[29] Barbosa-da-Silva A *et al. Genet Mol Biol*. 2005 **28**: 562

Edited by P Kangueane
Citation: Joshi *et al.* Bioinformation 6(9): 360-363 (2011)

BIOINFORMATION

open access

Supplementary material:

Table 1: Classification and features of NBS-LRR R genes used as query against the *Curcuma longa* EST database. The genes are grouped in three classes (I: NBS+LRR; II: CC+NBS+LRR; III: TIR+NBS+LRR) with respective accession number in NCBI, source species, gene name and domain range.

R gene class	Accn no.	Source	Gene name	Sequence size (aa)	Domain range (aa)							
					LRR		NBS		CC/LZ/zf		TIR	
					Start	End	Start	End	Start	End	Start	End
NBS-LRR	ABB88855	*Oryza sativa*	*Pi9*	1032	603	755	172	462	-	-	-	-
	ABC94599	*Oryza sativa*	*Pi2*	1032	627	760	166	465	-	-	-	-
	BAA76281	*Oryza sativa*	*Pib*	1251	876	905	173 406	336 709	-	-	-	-
CC-NBS-LRR	NP172686	*Arabidopsis thaliana*	*RPS5*	889	540	636	140	444	3	110	-	-
	NP187360	*Arabidopsis thaliana*	*RPM1*	926	604	883	177	465	4	169	-	-
	AF118127	*Lycopersicon esculentum*	*I2*	1266	578	1231	154	457	5	142	-	-
	AAQ01784	*Triticum aestivum*	*Lr10*	921	534	888	198	502	17	197	-	-
	AAC05834	*Triticum aestivum*	*Cre3*	921	502	823	193	481	14	183	-	-
	AAK00132	*Oryza sativa*	*Pita*	928	533	891	211	504	13	193	-	-
	BAA25068	*Oryza sativa*	*Xa1*	1802	771	1773	283	593	17	189	-	-
	AAP81262	*Zea mays*	*Rp1*	1269	596	1228	148	457	13	157	-	-
TIR-NBS-LRR	AAC72977	*Arabidopsis thaliana*	*RPP1*	1189	668	1011	226	505	-	-	54	184
	AF440696	*Arabidopsis thaliana*	*RPP4*	1135	642	1043	185	441	-	-	15	145
	BAB11393	*Arabidopsis thaliana*	*RPS4*	1232	663	889	198	473	-	-	21	149
	U27081	*Linum usitatissimum*	*L6*	705	524	699	240	520	-	-	63	194
	AF093649	*Linum usitatissimum*	*L*	1294	607	1277	220	521	-	-	63	195
	T18548	*Linum usitatissimum*	*M*	1305	744	1288	235	534	-	-	78	210
	A54810	*Nicotiana glutinosa*	*N*	1144	597	908	172	447	-	-	14	147
	CAD29728	*Solanum tuberosum*	*HERO*	1283	-	-	504	811	-	-	54	184

Table 2: Blast results and sequence evaluation of *Curcuma* NBS-LRR genes, including data about the query: homologous sequence, NCBI gi/-number; features and evaluation results of *Curcuma* clusters related to R-genes: cluster size in nucleotides (n), ORF (Open Reading Frame) size in amino acids (aa) and e-value.

Predicted protein domain	*Curcuma longa* contig	Homologous sequence	NCBI gi/nr	Size (n)	ORF (aa)	E.value
X-NBS-LRR	CL.CON.251	*Arabidopsis thaliana RPM1* protein	15231371	1103	336	7e-30
	CL.CON.1003	*Oryza sativa* Indica group NBS-LRR disease resistance protein, *Pi9*	82659480	998	312	2e-09
	CL.CON.1361	*Avena damascena* rga resistance gene for putative resistance protein, Clone DAM II-6	49640073	972	308	4e-05
	CL.CON.1681	*Avena vaviloviana* rga gene for putative resistance protein	49640105	857	287	4e-09
	CL.CON.1947	*Avena sativa* cultivar SunII NBS-LRR type disease resistance protein *O1*	3411224	1036	339	3e-27
	CL.CON.2236	NBS-LRR type R protein *NBS4-Pi, Oryza sativa*	86361429	937	298	2e-29
	CL.CON.2521	*Saccharum* hybrid cultivar Q117 RGA-Q3 resistance protein	56694164	881	273	4e-16
	CL.CON.3560	*Oryza sativa* indica group, *Xa1* protein	2943742	779	236	2e-08
	CL.CON.3602	*Populus trichocarpa* NBS-LRR resistance protein	224069218	981	302	0.052
	CL.CON.4029	*Pyrus communis* putative NBS-LRR disease resistance protein	40644865	1012	329	0.005
CC-NBS-NBS-LRR	CL.CON.370	Pib resistance protein, *Oryza sativa*	37777009	1132	356	2e-21
	CL.CON.3566	Pib resistance protein, *Oryza sativa*	37777009	1162	361	2e-17
CC-NBS-LRR	CL.CON.255	*Zingiber officinale* CC-NBS-LRR disease resistance protein like gene, Zop68	58918807	962	312	
	CL.CON.364	*Zingiber officinale* CC-NBS-LRR disease resistance protein like gene, Zop1010	58918311	967	317	1e-20
	CL.CON.832	*Zingiber officinale* CC-NBS-LRR disease resistance protein like gene, Zop103	58918197	886	289	1e-23
	CL.CON.1258	*Ipomoea batatas* SPRGA-2 NBS-LRR protein rsp2	82541821	1012	341	1e-21
	CL.CON.3339	*Ipomoea batatas* SPRGA-3 NBS-LRR protein rsp3	82541823	923	312	2e-11
	CL.CON.3440	*Arabidopsis thaliana RPS*2 protein	549979	998	309	2e-05
	CL.CON.3569	Rust resistance protein *Rp1, Zea mays*	32423732	876	258	3e-11
	CL.CON3614	*Oryza sativa* japonica group, *Pita* protein	12642090	926	293	1e-26
	CL.CON.3846	*Oryza sativa* japonica group, *Pita* protein	12642090	873	264	1e-28
NBS-LRR	CL.CON.837	*Pib* resistance protein, *Oryza sativa*	37777009	1231	342	3e-05
	CL.CON.1529	*Pib* resistance protein, *Oryza sativa*	37777009	1256	347	3e-05
NBS	CL.CON.574	*Glycine max* resistance protein *KNBS-4*	13111696	569	143	0.001
	CL.CON.1353	*Cajanus cajan* clone PP4 unknown gene	7107261	583	158	0.002
	CL.CON.2247	*Cajanus cajan* clone PP4 unknown gene	7107261	571	156	0.002
CC-NBS	CL.CON.1267	*Medicago sativa* resistance gene analog protein	8118174	452	112	1e-23
	CL.CON.3620	*Medicago sativa* resistance gene analog protein	8118177	483	123	1e-32