

Degenerate Primer Design via Clustering

Xintao Wei
*School of Computer Science,
Florida International
University, University Park,
Miami, FL 33199, USA.*
xwei001@cs.fiu.edu

David N. Kuhn
*Dept. of Biological Sciences,
Florida International
University, University Park,
Miami, FL 33199, USA.*
kuhnd@fiu.edu

Giri Narasimhan
*School of Computer Science,
Florida International
University, University Park,
Miami, FL 33199, USA.*
giri@cs.fiu.edu

&

*USDA-ARS, Subtropical
Horticulture Research Station,
13601 Old Cutler Road, Miami,
FL 33158, USA.*

Abstract

This paper describes a new strategy for designing degenerate primers for a given multiple alignment of amino acid sequences. Degenerate primers are useful for amplifying homologous genes. However, when a large collection of sequences is considered, no consensus region may exist in the multiple alignment, making it impossible to design a single pair of primers for the collection. In such cases, manual methods are used to find smaller groups from the input collection so that primers can be designed for individual groups. Our strategy proposes an automatic grouping of the input sequences by using clustering techniques. Conserved regions are then detected for each individual group. Conserved regions are scored using a BlockSimilarity score, a novel alignment scoring scheme that is appropriate for this application. Degenerate primers are then designed by reverse translating the conserved amino acid sequences to the corresponding nucleotide sequences. Our program, DePiCt, was written in BioPerl and was tested on the Toll-Interleukin Receptor (TIR) and the non-TIR family of plant resistance genes. Existing programs for degenerate primer design were unable to find primers for these data sets.

1. Introduction

A nucleotide sequence is called **degenerate** if one or more of its positions can be occupied by one of several possible nucleotides [1]. For example, AYGCNY is a sequence written down using IUPAC ambiguity codes

(see Table 2 for these codes), where Y stands for one of C or T, and N stands for A, C, G, or T. The *degeneracy* of a sequence is the number of different sequences that it represents. Thus, the degeneracy of AYGCNY is 16 (it has 2 Y's and one N).

One way to amplify a specific target sequence in a genome is to design a pair of primers (one forward and one reverse) that flank either end of the target sequence and to use the standard laboratory technique of Polymerase Chain Reaction (PCR). If some pair of primer sequences is strongly conserved in several genomes, then the same pair of primers can be used to amplify the target region from all the genomes. However, when the primer sequences are only weakly conserved, “degenerate” primers (i.e., primers whose sequences are degenerate) are needed. Degenerate primers are particularly useful in amplifying homologous genes from different organisms [2, 3]. In the “candidate gene approach”, known genes that affect similar processes in one organism could have their homologues amplified in other related organisms by the use of a well-designed degenerate primer pair [4]. Homologous genes display regions where they are highly conserved and also regions where they have evolved and are divergent. Primers can be found by searching in the highly conserved regions. In order to account for small mutations in the conserved regions, degenerate primers are used to match a large collection of similar sequences. Thus, degenerate primers can be used to isolate genes encoding proteins that belong to known protein families [3].

R-genes are a class of genes that code for proteins that impart plants with resistance to a variety of pathogens [5, 6]. In an effort to grow pathogen-resistant,

commercially important plants such as cacao, researchers at the USDA Subtropical Horticulture Research Station in Miami have been trying to design degenerate primers based on R-genes from available strains of a variety of plants [7, 8]. With more and more R-genes being identified, the corresponding protein families have been getting bigger and bigger [9, 10].

The first step in identifying primers for a set of homologous genes is to compute a sequence alignment of the genes or the protein sequences (or both). The next step is to identify at least two conserved regions in the alignment such that the target region to be amplified is contained in between them. Primers need to satisfy several properties. These include a feasible annealing temperature, an appropriate range for its GC-content, reasonably sticky ends that avoid degeneracy at the end of the primers, low degeneracy, and reasonable distance between two conserved regions [11-16]. Therefore the last step involves finding a primer pair that best satisfies all the constraints.

Existing programs for the design of degenerate primers include GeneFisher [17], CODEHOP [18], and HYDEN [19]. GeneFisher uses a straightforward algorithm of starting with an amino acid sequence alignment and then back-translating conserved portions to obtain degenerate primers. CODEHOP designs “hybrid” degenerate primers that contain a short 3’ degenerate core region (about 11-12 bp) and a longer 5’ consensus clamp region (about 18-25 bp). It requires the input to contain a set of conserved amino acid blocks. CODEHOP then uses position-specific scoring matrices of aligned nucleotide

sequences to design primers with low degeneracy. Both GeneFisher and CODEHOP work well for small sets of sequences that have strong consensus blocks. HYDEN, the most recent of the three, designs primers for aligned DNA sequences. The algorithm represents a trade-off between low degeneracy and large coverage (number of matched sequences). Greedy “hill-climbing” is used at the end to improve the coverage.

In this paper, we focus on the second and third steps in the process of designing primers for a given set of aligned amino acid sequences. In other words, we focus on identifying conserved regions in the alignment and designing primers with low degeneracy and high coverage. We assume that a reliable multiple alignment is already provided to us for the entire set of input proteins. Often, it may only be possible to find primers with very high degeneracy, or it may be impossible to find conserved regions in an alignment. In such cases, it becomes necessary to manually divide the alignment into several groups, and detect degenerate primers for each group separately. Since primers are fairly inexpensive (less than \$3 for a primer of length 20), finding more than one primer pair is a feasible solution.

The first contribution of this paper is an algorithm to automate this manual process by using **clustering** techniques to group the sequences in the multiple alignment. The idea was to cluster sequences based on the presence of conserved regions for the members of the cluster. Clustering techniques require a concept of similarity and/or distance between members. The second

Table 1. Genetic code table

	T	C	A	G
T	TTT Phe (F)	TCT Ser (S)	TAT Tyr (Y)	TGT Cys (C)
	TTC	TCC	TAC	TGC
	TTA Leu (L)	TCA	TAA Ter	TGA Ter
	TTG	TCG	TAG Ter	TGG Trp (W)
C	CTT Leu (L)	CCT Pro (P)	CAT His (H)	CGT Arg (R)
	CTC	CCC	CAC	CGC
	CTA	CCA	CAA Gln (Q)	CGA
	CTG	CCG	CAG	CGG
A	ATT Ile (I)	ACT Thr (T)	AAT Asn (N)	AGT Ser (S)
	ATC	ACC	AAC	AGC
	ATA	ACA	AAA Lys (K)	AGA Arg (R)
	ATG Met (M)	ACG	AAG	AGG
G	GTT Val (V)	GCT Ala (A)	GAT Asp (D)	GGT Gly (G)
	GTC	GCC	GAC	GGC
	GTA	GCA	GAA Glu (E)	GGA
	GTG	GCG	GAG	GGG

Table 2. IUPAC ambiguity codes

Code	Description
M	AC
R	AG
W	AT
S	CG
Y	CT
K	GT
V	ACG
H	ACT
D	AGT
B	CGT
N	ACGT

contribution of this paper is the introduction of a novel measure called **BlockSimilarity** to measure the quality of an alignment between amino acid sequences, a measure that is best suited for designing degenerate primers. The *BlockSimilarity* score uses a similarity metric between amino acids, which is based on their coding in the *Genetic Code*. As explained later in detail, two amino acids are considered “similar” if they are identical, or if they are in the same row or column of the Genetic Code Table (see Table 1).

Once conserved amino acid blocks are found, degenerate primers are designed by reverse translating the conserved blocks to nucleotide sequences using the genetic code. Finally, if the degeneracy of the designed primer is too high, then we show how to use the

corresponding nucleotide sequence alignment, if available, to reduce its degeneracy.

Our program, which is dubbed DePiCt 1.0, is written in BioPerl. It was successfully tested on the TIR and non-TIR subfamilies of R-genes.

2. Algorithms and Implementation

2.1 Codons and Similar Amino Acids

Table 1 shows the genetic code. Table 2 shows the IUPAC ambiguity codes. Table 1 is reinterpreted in Table 3, which contains the nucleotide sequences (using the IUPAC ambiguity codes) for each of the 20 amino acids along with their degeneracy.

Table 3 highlights the fact that several amino acids have very similar codons. It suggests, for example, that for the purpose of designing degenerate primers, Cysteine (C) and Tyrosine (Y) ought to be considered as “similar”, since they can be represented by the nucleotide sequences TGY and TAY, respectively, which differ only in the middle base. The “similarity” of Cysteine and Tyrosine, in this sense, is a consequence of their position in the genetic code table (Table 1), i.e., they lie in the same row. Note that Cysteine and Tyrosine are not considered similar in terms of their physic-chemical properties. Table 4 provides some examples of the nucleotide sequences for several “similar” amino acids along with their degeneracy. Not all “similar” amino acids are included in Table 4.

In summary, for the purpose of this algorithm, two amino acids are said to be “similar” if they lie along the same row or column in the Genetic Code Table (Table 1).

Table 3. Genetic code tables with IUPAC codes

<i>Amino Acid</i>	<i>IUPAC Ambiguity Code</i>	<i>Degen-eracy</i>	<i>Amino Acid</i>	<i>IUPAC Ambiguity Code</i>	<i>Degen-eracy</i>
Ala (A)	GCN	4	Met (M)	ATG	1
Cys (C)	TGY	2	Asn (N)	AAY	2
Asp (D)	GAY	2	Pro (P)	CCN	4
Glu (E)	GAR	2	Gln (Q)	CAR	2
Phe (F)	TTY	2	Arg (R)	CGN & AGR or MGR & CGY	6
Gly (G)	GGN	4	Ser (S)	TCN & AGY	6
His (H)	CAY	2	Thr (T)	CAN	4
Ile (I)	ATH	3	Val(V)	GTN	4
Leu (L)	CTN & TTR or YTR & CTY	6	Trp (W)	TGG	1
Lys (K)	AAR	2	Tyr (Y)	TAY	2

Table 4. Similar amino acids and their IUPAC codes

Amino Acid Sets	IUPAC Ambiguity Codes	Degeneracy
[HQ]	CAN	4
[HN]	MAY	4
[IM]	ATN	4
[IV]	ATH & GTN or RTH & GTG	7
[IT]	ATH & CAN or AYH & ACG	7
[IVM]	RTN	8
[TA]	RCN	8
[CWG]	KGB & GGA or TGB & GGN	7
[CW]	TGB	3
[YH]	YAY	4
[YN]	WAY	4
[YD]	KAY	4
[HD]	SAY	4
[ND]	RAY	4
[NK]	AAN	4
[DE]	GAN	4
[QK]	MAR	4
[QE]	SAR	4
[K.E]	RAR	4
[HQNK]	MAN	8
[HQDE]	SAN	8
[NKDE]	RAN	8
[VDE]	GWN	8
[VD]	GWY & GTR or GTN & GAY	6
[MT]	AYG & ACH or ATG & ACN	5
[MNK]	ATG & AAN or AWG & AAH	5
[MK]	ATG & AAR or AWG & AAA	3
[IN]	ATH & AAY or AWY & ATA	5
[CY]	TRY	4
[CF]	TKY	4
[FY]	TWY	4

2.2 BlockSimilarity

Given a set of aligned protein sequences, the *BlockSimilarity* score between sequences is a measure of the length of a conserved sequence, a necessary ingredient to find a pair of primers. Based on the notion of “similarity” of amino acids (see Section 2.1), a position in an alignment is said to be “conserved” if the amino acids occupied by that position in all the sequences are either identical or “similar”. To find primer sequences, we are

interested in blocks of such conserved positions. However, the blocks need to be sufficiently long (greater or equal to a prescribed threshold k) in order to find a primer sequence of desired length. Therefore, only sufficiently long blocks are “scored” by the algorithm. If it is of length less than k then it is not scored (i.e., given a score of 0). If the conserved block is of length $n \geq k$, then its score is computed as n . An example is shown in Figure 1.

Seq 1:	. .	i g e m l a a	. lv	. pb . s e p f	..	y g a l q t	.			
Seq 2:	. y	i g d m l a a	. fv	. v . . s e p f	..	f a a l h t	.			
Conserved seq:	. .	i g * m l a a	. *v s e p f	..	**a l * t	.			
Amino acids block:		i g * m l a a	*v	s e p f		**a l * t				
BlockSimilarity scores:		7	+	0	+	0	+	6	=	13

Figure 1. Example of a BlockSimilarity score computation. Here Seq1 and Seq2 are two multiply aligned protein sequences with “.” indicating a gap in the alignment. A conserved sequence is found based on the “similarity” of amino acids. Note that “*” in the conserved sequence implies that the amino acids occupying that position are “similar”. Then, for each block of conserved amino acids, the BlockSimilarity score is computed; the sum of BlockSimilarity scores is the score of the pair of protein sequences. Here MinPrimerLength was set to 18.

The value k is an integer value dictated by the minimum number of nucleotides that is required in the primer for the PCR reaction. It is therefore usually computed as $\text{MinPrimerLength} / 3$ (rounded to the next integer). Note that MinPrimerLength is the number of nucleotides. If a set of sequences has more than one conserved block, then the BlockSimilarity score is obtained by simply adding the scores for each conserved block. Higher scores are more desirable. However, note that since we need to design a pair of primers, a set of amino acid sequences must have at least two blocks with non-zero BlockSimilarity scores (or there must be a sufficiently long block).

2.3 Clustering

As mentioned before, the idea behind clustering is to group together sequences that have high BlockSimilarity scores so that a primer pair can be designed for each group. Clustering is a well-researched problem and numerous algorithms exist for the problem [20]. We use *Hierarchical clustering*, although any of the other known clustering techniques could have been used.

Hierarchical clustering is usually implemented as a bottom-up algorithm. It starts off with each sequence forming a singleton group. In each iteration, the two groups with the highest similarity are merged to form one group, if the resulting group corresponds to a “feasible” grouping. The process continues until the groups cannot be merged any more. The resulting groups are then reported as the clusters output by the algorithm. A grouping is said to be “feasible” if it has at least one block of length greater than or equal to the parameter

MinPrimerProductLength, or if it has at least two blocks of amino acids separated by a length in the range $[\text{MinPrimerProductLength}, \text{MaxPrimerProductLength}]$. MinPrimerProductLength and MaxPrimerProductLength are user-specified threshold values indicating the minimum and maximum length for a PCR product if a primer pair is successfully chosen from the conserved blocks in the sequences.

2.4 Designing Degenerate Primers

The degenerate primers are designed by reverse translating conserved blocks of residues using their corresponding genomic codons. Primers are usually required to be within a certain length range. Our algorithm requires that the MinPrimerLength and MaxPrimerLength be specified. The algorithm also requires that the maximum degeneracy of the primer be specified. The algorithm also requires the specification of the range of length for the PCR product. Note that complementary DNA sequences are used to compute the reverse primers.

2.5 Decreasing Degeneracy

In case the degeneracy is higher than the specified MaxDegeneracy value, then three choices exist. One possibility is that the degeneracy can be decreased by splitting primers. For example, a primer with the code “N” (referring to the set [ACGT]) can be split into two primers, one having “R” (set [AG]) instead of “N”, and the other having “Y” (set [CT]) instead of “N”. The two

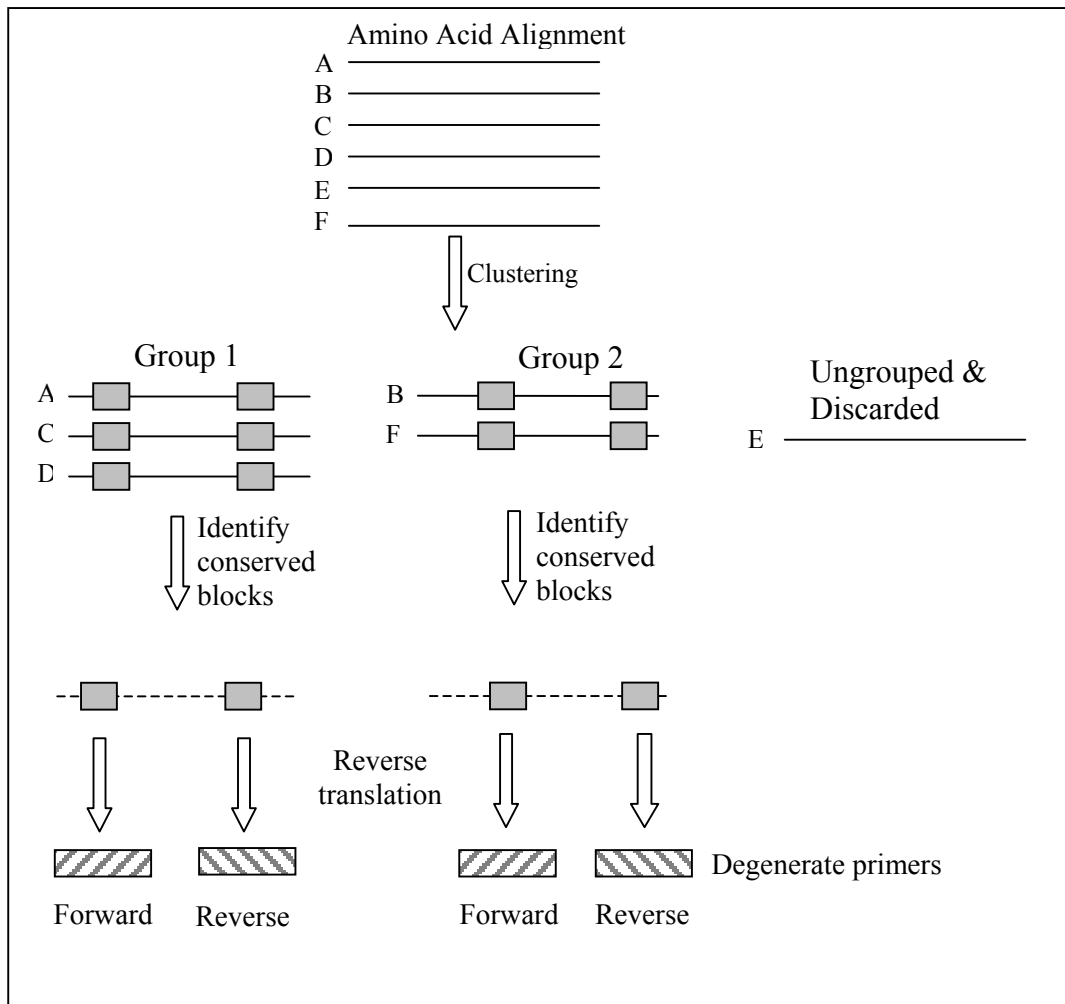


Figure 2. Algorithm of DePiCt

resulting primers now have only half the degeneracy of the original primer.

A second possibility can be used only if a nucleotide alignment is available, in which case one could check if specific location in the alignment is favored by a smaller set of bases than what the degeneracy would imply. For instance, using the nucleotide alignment, a primer with “N” could be replaced by “W” (set [AT]), if all the sequences in the alignment have either “A” or “T” occupying that location. Such a replacement would decrease the degeneracy.

It is also known that different organisms favor different codons. If the codon biases were known, then primers with lower degeneracy could be designed by using the more favored codons. Note that this would work

only if the primers fall within coding regions, and if the reading frames are known.

2.6 Algorithm Implementation

The algorithm DePiCt was implemented in BioPerl. The input to DePiCt is a multiple alignment of the target portion of the protein sequences, the length range for the length of the primer, the maximum degeneracy of the primer, and the minimum and maximum length of the PCR product. DePiCt clusters the sequences in the alignment and constructs degenerate primers for each group. A schematic diagram explaining the algorithm is shown in Figure 2.

Table 5. Groups output by DePiCt for the TIR subfamily

Group #	1	2	3	4	5	6
GenBank IDs of Proteins in Group	af098963	np_189178	np_198509	af211528	af175395	ab006706
	np_190053	ab025639	ab016877	a54810	af322632	np_197270
	np_190034	np_196686	t18548	stu9719	af175388	t08196
	af098964	np_198650	np_193686	stu9720	af327903	
	aal86339	fl2p19	np_192938	stu300266	ac022492	
	np_190049	np_179024	np_198969	ac000348		
	af098962	aal38864	np_197338	np_174037		
	np_190026	aam15274		np_174038		
	np_187072	np_199264		af316405		
	ac073178	np_193422		lus310164		
		np_199457		Lus310157		

Table 6. Degenerate primers for the 6 groups of sequences from the TIR subfamily.

The standard primers used by Aarts *et al.* is given in the row labeled STD.

Group#	Primers	Degeneracy	Primer Location	Amino acid sequence
1	GGNCCNGGRWSTCGKATTATDATCAC	768	530-538	GPGSRIIT
	GAYGCNYTTCARATHTTYTG	192	566-572	EA[FL]QIFC
	DRTYTCYCKNCCVARYTG	1152	746-741	Q[FL]GRE[TI]
2	ATNRTHGGNATHTGGGGN	1152	403-408	[MI][IV]GIWG
	RTCNGGDAKYTCYTYYAR	768	937-932	LKE[LI]PD
3	GGNATHGGNAARACNAC	384	411-416	GIGKTT
	AARCANGCHATRTSVARRAA	1152	664-658	FL[HD]IACF
4	GGNAARACNACNMTHGC	768	413-418	GKTT[IL]A
	RCANGCDATRTCYARRAA	384	663-658	FLDIAC
	RCANGCDATRCRAGRAA	192		
5	SARAAAYTAYGCNTCNTCN	1024	245-250	[EQ]NYASS
	SARAAAYTAYGCNAGYTCN	512		
	TCYCTNGTNGTDATDATDA	576	541-535	[VI]IITTRD
TCNCGNGTNGTDATDATDA	864			
6	TTYAGNGGNGARGAYGTN	512	187-192	FRGEDV
	GCDATNGTNGTYTTNCC	384	418-413	GKTTIA
STD	GGTATGGGTGGTGTGGTAARACNACN	32	408-416	GMGGVVGKTT
	CCWRATGGTRATCGTRATTTYCATGATCC	32	609-600	GLPLALKVLG

3. Results

Resistance gene homologues (RGH) are plentiful in plants. Most resistance genes contain a nucleotide binding site (NBS) motif and a leucine-rich repeat (LRR) that share significant identity of amino acid sequence with known and mapped resistance genes from other plants [5, 6]. The NBS region has some highly conserved portions from which degenerate primers have been designed [2, 3]. These primers have proved useful in amplifying resistance gene homologues from a variety of plants whose genetic structure is unknown or only poorly known, such as cacao and coffee [21]. More and more

RGH sequences have been cloned from a wide variety of plant species; the set of known genes has become large enough to make the design of primers a non-trivial problem.

Based on the motifs in the NBS region, R-genes are divided into two subfamilies: TIR (Toll-Interleukin Receptor-like regions) and non-TIR [9, 10]. We tested our program with a multiple alignment of 47 protein sequences from the TIR family and another alignment of 49 sequences from the non-TIR family. CODEHOP and Genefisher were unable to find primers on these data sets. The results from running DePiCt are shown in Tables 5-8.

Table 7. Groups output by DePiCt for input nonTIR49. Two of the 49 sequences were not covered.

Group #	1	2	3	4	5	6	7
GenBank IDs of Proteins in Group	af414176	ac025416	ac092553	af181730	af098970	ac084404	t06219
	af414179	np_172692	bab63659	af181728	af098971	bab44079	af017751
	af342991	np_172686	bab63657	af181729	np_188065	af158634	
	aam03018	np_172685	bab07969	af180355	np_188064	af118127	
	aak95831	bab90113	aal99361	af368301		t06403	
	af414177	np_172693	af414172			af306502	
	af414175	np_192816					
	af414171	np_176314					
	af414174	aak96709					
	af107294	ac004255					
		np_176451					
		np_176524					
		np_201107					
		np_196159					

Table 8. Degenerate primers for the 7 groups of sequences from the non-TIR subfamily.
The standard primers used by Shen *et al.* is given in the row labeled STD.

Group#	Primers	Degeneracy	Primer Location	Amino acid seq
1	GTNYTNGAYGAYGTNTGGTT	512	452-458	VLDDVWF
	WGGRTYYARYTTMTSRTA	512	603-598	Y[DEQ]KLDP
2	GGYATGGSNNGDGTGGNAARAC	1152	1086-1093	GM[GA]GVGKT
	BRTCCAHardGCCATBKC	648	1405-1400	[EA]MALW[IT]
3	MTHGTYTKGAYGAYGKTGG	384	451-457	[IL][VIM]LDDVW
	TYSARRTAMCKHARRTG	768	835-830	HLR[FY]L[DN]
4	GAYGTNTGGGARGARATHGAY	192	1173-1179	DVWEEID
	YTCYTCNGTYTCNCKRTG	512	1280-1275	HRETEE
5	CAHYTBAARMGATGYTTYGCY	576	609-615	[HQ]LKRCFA
	TTGAGWSAWGKYARHWGHCC	1152	1096-1090	G[QL]L[PT][SC]LK
6	GTYYTNGAYGAYRINTGG	512	452-457	VLDD[VI]W
	TTRGGRWATAHDSHRCA	864	617-612	C[SA][VIL][YF]PK
7	TTYGGNMNGWYRAYGA	1024	166-171	FGR[VD][DN]D
	CCANACRTCRTCNARNAC	1024	1170-1165	VLDDVW
STD	GGTATGGGTGGTGTGGTAARACNACN	32	357-365	GMGGVVGKTT
	CCWRATGGTRATCGTRATTTYCATGATCC	32	565-556	GLPLALKVLG

The MaxDegeneracy was set at 1200, the MinPrimerLength was 18, and the MinPrimerProduct was 300.

For the TIR subfamily, the program output 6 groups (Table 5). The primer pairs for each group are shown in Table 6. The degeneracy, the amino acid sequences and their locations (in the protein sequence alignment) are

shown in the last three columns. The last two rows of Tables 6 and 8 correspond to the (“hand-crafted” for TIR and non-TIR) standard primers used in previous research [2, 3]. For the 49 protein sequences from the non-TIR subfamily, DePiCt output 7 groups as shown in Table 7. Note that two of the 49 sequences were not covered by the algorithm since they did not cluster with any other

sequences. The primers designed for each group of the non-TIR subfamily are shown in Table 8.

4. Discussion and Conclusions

In this paper, we improve an algorithm for designing degenerate primers. Given a multiple alignment, clustering is used to group the sequences and the degenerate primers are designed for each group. A novel distance function called the *BlockSimilarity* measure, is proposed for quantifying the similarity between aligned sequences and for deciding which sequences should be clustered. A final cleanup step that uses nucleotide sequence alignment (if available) is employed to decrease the degeneracy. The algorithm was implemented in BioPerl and tested on TIR and non-TIR, which are multiple alignments of two sets of protein sequences from R genes. The BioPerl source code for DePiCt can be obtained upon request by email. Further information can be found at: <http://www.cs.fiu.edu/~giri/bioinf/DePiCt/>

5. Acknowledgements

The authors gratefully acknowledge the help of Kalai Mathee for critical comments on a draft of the paper and for several useful discussions.

6. References

1. Kwok, S., S.Y. Chang, J.J. Sninsky, and A. Wang, *A guide to the design and use of mismatched and degenerate primers*. PCR Methods Appl, 1994. **3**(4): p. S39-47.
2. Aarts, M.G., B. te Lintel Hekkert, E.B. Holub, J.L. Beynon, W.J. Stiekema, and A. Pereira, *Identification of R-gene homologous DNA fragments genetically linked to disease resistance loci in Arabidopsis thaliana*. Mol Plant Microbe Interact, 1998. **11**(4): p. 251-8.
3. Shen, K.A., B.C. Meyers, M.N. Islam-Faridi, D.B. Chin, D.M. Stelly, and R.W. Michelmore, *Resistance gene candidates identified by PCR with degenerate oligonucleotide primers map to clusters of resistance genes in lettuce*. Mol Plant Microbe Interact, 1998. **11**(8): p. 815-23.
4. Deng, C. and T.M. Davis, *Molecular identification of the yellow fruit color (c) locus in diploid strawberry: a candidate gene approach*. Theoretical and Applied Genetics, 2001. **103**(2/3): p. 316-322.
5. Hulbert, S.H., C.A. Webb, S.M. Smith, and Q. Sun, *Resistance gene complexes: evolution and utilization*. Annu Rev Phytopathol, 2001. **39**: p. 285-312.
6. Hammond-Kosack, K.E. and J.D.G. Jones, *Plant Disease Resistance Genes*. Annual Review of Plant Physiology and Plant Molecular Biology, 1997. **48**: p. 575-607.
7. Borrone, J.W., R.J. Schnell, and D.N. Kuhn, *Design of degenerate WRKY primers and potential usefulness of WRKY genes as molecular markers*. 2002.
8. Kuhn, D.N., M. Heath, R.J. Wisser, A. Meerow, J.S. Brown, L.R. J., and S.R. J., *Resistance gene homologues in Theobroma cacao as useful genetic markers*. 2002.
9. Cannon, S.B., H. Zhu, A.M. Baumgarten, R. Spangler, G. May, D.R. Cook, and N.D. Young, *Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS-LRR resistance gene subfamilies*. J Mol Evol, 2002. **54**(4): p. 548-62.
10. Meyers, B.C., A.W. Dickerman, R.W. Michelmore, S. Sivaramakrishnan, B.W. Sobral, and N.D. Young, *Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily*. Plant J, 1999. **20**(3): p. 317-32.
11. Kampke, T., M. Kieninger, and M. Mecklenburg, *Efficient primer design algorithms*. Bioinformatics, 2001. **17**(3): p. 214-25.
12. Hou, J., *Design of endonuclease restriction sites into primers for PCR cloning*. Bioinformatics, 2002. **18**(12): p. 1690-1.
13. Raddatz, G., M. Dehio, T.F. Meyer, and C. Dehio, *PrimeArray: genome-scale primer design for DNA-microarray construction*. Bioinformatics, 2001. **17**(1): p. 98-9.
14. Li, L.C. and R. Dahiya, *MethPrimer: designing primers for methylation PCRs*. Bioinformatics, 2002. **18**(11): p. 1427-31.
15. Podowski, R.M. and E.L. Sonnhammer, *MEDUSA: large scale automatic selection and visual assessment of PCR primer pairs*. Bioinformatics, 2001. **17**(7): p. 656-7.
16. Fernandes, R.J. and S.S. Skiena, *Microarray synthesis through multiple-use PCR primer design*. Bioinformatics, 2002. **18 Suppl 1**: p. S128-35.
17. Giegerich, R., F. Meyer, and C. Schleiermacher, *GeneFisher-software support for the detection of postulated genes*. Proc Int Conf Intell Syst Mol Biol, 1996. **4**: p. 68-77.
18. Rose, T.M., E.R. Schultz, J.G. Henikoff, S. Pietrokovski, C.M. McCallum, and S. Henikoff, *Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences*. Nucleic Acids Res, 1998. **26**(7): p. 1628-35.
19. Linhart, C. and R. Shamir, *The degenerate primer design problem*. Bioinformatics, 2002. **18 Suppl 1**: p. S172-81.
20. Han, J. and M. Kamber, *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems, ed. J. Gray. 2000: Morgan Kaufmann. 550.
21. Noir, S., M.C. Combes, F. Anthony, and P. Lashermes, *Origin, diversity and evolution of NBS-type disease-resistance gene homologues in coffee trees (Coffea L.)*. Mol Genet Genomics, 2001. **265**(4): p. 654-62.