International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.10 (2015), pp.67-74
http://dx.doi.org/10.14257/ijmue.2015.10.10.07

Insertion - Deletion as Informative Characters in DNA Barcoding

Goutam Sanyal, Asim Kumar Mahadani, Pradosh Mahadani and Partha Bhattacharjee

Department of Computer Science & Engineering,
National Institute of Technology,
Durgapur, West Bengal, India
Department of Computer Science & Engineering,
Bankura Unnayani Institute of Engineering, West Bengal, India
Department of Genetics, ICAR- National Research Center for Orchids, Pakyong, Sikkim, India
Department of Cybernetics, Central Mechanical Engineering Research Institute,Durgapur,India
nitgsanyal@gmail.com, asimmahadani@yahoo.com, pmahadani@gmail.com, partha_cmeri@yahoo.com

Abstract

DNA barcoding involve rapidly sequencing the standardized region from the genome for species identification and this homologous sequence information plays an important role to resolve the phylogenetic issues. However in case of complex groups, incorporation of insertion and deletion (Indel) informative sites in DNA barcode sequences are becoming more important due to low substitution rate. Indels are largely ignored in phylogenetics analysis and removed from sequence alignment assigning them a missing data. We review the current trends on mining the indels analysis, focusing especially on the topics of rapidly evolving indel containing loci and methods of indel treatment for phylogenetic relationship. Among the indel coding methods, Simple indel coding is easy to implement in indels contain sequences. But, this method does not utilized the all the available information and Complex Indel Coding rules are very difficult to translate into a clearly formulate algorithm for determining the values. But this coding method suffers from internal inconsistence when a long indel has a number of shorter subset indel and triangular in equality exists. However, SIDIER software package infer evolutionary relationships based on both the indels and substitutions. SIDIER is promising software for intra and inter specific calculation in DNA barcode studies as well as to infer phylogenetic relationships.

Keywords*: Indel, Gaps, DNA Barcoding, Simple Indel Coding, Modified Complex Indel Coding*

1. Introduction

After the development of DNA barcoding concept, the fields of genomics have revolutionized the way that we classify and identify species. In general, barcode is used to cataloging grocery store item similarly DNA barcode follow the same principle to recognize biological species. DNA barcoding involve rapidly sequencing the standardized region from the genome for species identification [1]. Since, assessing the whole genome is very intricate, the part of the genome that show significant sequence variation between the species is analyzed as a marker in molecular systematic and phylogenetics. In 2003, Paul DN Hebert and his coworkers recognized a 648bp region of *Cytochrome c oxidase subunit I* (COI) of mitochondrial DNA as the species level signature for animals [2]. Similarly, Consortium of Barcode of Life (CBOL) plant researcher proposed plastid genes

ISSN: 1975-0080 IJMUE

rbcL and *matK* either singly or in combination as the standard DNA barcode for plants [1]. But animal DNA barcode is more popular and successful in comparisons of plant barcode regions. Besides that, ITS of the nuclear ribosomal RNA standardized as the universal barcode marker for fungi and 16S rRNA sequences singly gain huge popularity for prokaryotic.

The Barcode of Life Data System (BOLD) is specialized repository for DNA barcode sequences and emerged as a global bio-identification system for species [3]. Neighbour-Joining (NJ) method is mainly used in DNA barcode based phylogenetic tree reconstruction. The DNA barcode sequence based studies passed by one decade with a huge success; many challenges are not yet answered. Among them, low divergence between closely related organisms got major attention in different studies. Bud mutation, cultivation and wide cross compatibility with in the plant groups are major future challenges to be resolve through DNA barcoding. From last few years, scientists' community gave a major attention for the utility of insertion and deletion (indels) in species identifications and phylogenetics. However, recently researchers started the investigation of the statistical properties of indels treatment. In this review paper, we provide a current trend of both the indels occurring regions and indel coding method for incorporating the indel informative sites in the phylogenetic analysis.

2. Basic of Phylogenetics

In phylogenetic, molecular sequences are analyzed to estimate relatedness of gene or species. There are two major steps involve, first alignment of the homologous sequences and construct a tree based on alignment. A multiple sequence alignment (MSA) arranges a set of sequences in a scheme where positions believed to be homologous are written a common column. Rates or patterns of changes in sequences cannot be analyzed unless the sequences can be aligned. One of the most common reasons for generating alignments is that they are an essential prerequisite for phylogenetic analyses. Basically, phylogenetic methods are broadly categorized into distance-matrix methods, also familiar as clustering method (*e.g*. UPGMA, Neighbour-Joining), and discrete data methods, also known as tree searching methods (*e.g*. Maximum Parsimony (MP), Maximum Likelihood (ML), Bayesian method).

3. Insertion and Deletion (Indels)

Insertion and deletion are one of major factor in early divergence and are likely the most rapid and significant form of sequence variation in eukaryotic evolution. However, still there is very little information about evolutionary processes of indels and their impact speciation. When an insertion occurs, a string of nucleotides is received into a sequence. Deletions are the opposite genetic event of insertions, in which part of a sequence is removed. Indels are only observable when the alignment of sequences of unequal length during primary homology assessment. A gap (-) is a strategy used to deal with sequences that contain indels. Insertion–deletion polymorphisms are a type of biallelic short DNA length variation that has been subject to a growing interest in the phylogenetic as well as the mathematical biology field. In comparison to protein coding region, noncoding regions showed high frequency of Indels. Polymerase slippages during DNA replication are well known consequences for formation of Indels. However, in many cases we do not know the mechanism of indel formation. Li [12] classified indels based on size and mechanisms and divided into (i) short indels, which are mostly caused by errors of the cellular machinery during replication and(ii) long indels, which are caused primarily by unequal crossing over and transposition. Indels are largely ignored in phylogenetic reconstruction and gaps created by Indels are commonly removed from sequence alignments assigning them as missing data. However, several studies showed the importance of indels in phylogenetic for better resolution. So, incorporation of indel informative sites in phylogenetic analyses is becoming more important. Several studies reported different DNA region with high indel polymorphism for species-identification of plant (*trnH-psbA*, *trnL-trnF*, *matK*), fungi (ITS), bacteria (rRNA) and animals (rRNA, D-loop) *etc.*

3.1. MaturaseK (matK)

MaturaseK of chloroplast is immerged as a most conserved gene in plant kingdom and functioned as Group II intron splicing. About 1500 bp long, *matK* gene is found within the intron of trnK of chloroplast DNA and encodes maturase like protein. In major lineages of grass, the chloroplast *matK* showed indels, nucleotide substitution at the extreme 3' end of gene. Indels and nucleotide substitution at the 3' end of the *matK* provided valuable phylogenetic markers in Poaceae. *matK* sequences exhibited indels in multiple of 3 at 5´ end of family Apocynaceae. But, few numbers of indels in *matK* is one of the major limitations in protein coding sequences. Among the intergenic spacers of Chloroplast DNA, *trnH-psbA*, trnL-trnF are largely used as markers in plant population genetics and species‐level authentication. Length variation of these regions is due to the occurrence of indels and potential for higher gap rates than base pair substitutions. Indels also showed phylogenetic informative sites at inter and intraspecific level. *psbA-trnH* intergenic spacer region is highly variable among non-coding parts (~450 bp) in angiosperms and is easily recovered by using universal primer. Beside that Indels in *trnL-trnF* and *trnH-psbA* regions were promising approach for discriminate the closely related taxa and indel coding methods should be considered in DNA barcoding of closely associated plant species. However, insertions, deletions and simple sequence repeats were more common across the families when aligning the sequences.

3.2. Internal Transcribed Spacer (ITS)

Internal transcribed spacer (ITS) of the nuclear DNA is one of the most popular loci in fungi systematic and phylogenetic studies. About 400-800bp of ITS, makes it easier for sequencing and provide sufficient discrimination power among the species. In phylogenetic analysis, indels are reliable source of information and more conserved than base substitution. If indels are excluded from analyses that also affect the tree structure and bootstrap value. Thus, rapidly evolving indel-rich loci play an important factor of phylogenetic issue yet to solve.

3.3. D-Loop of Mitochondrial DNA

Animal mitochondrial DNA (mtDNA) sequences are most accepted in the evolutionary study due to uniparental inheritance and high copy number *etc.* There is also a ~1122 bp 'control' region or displacement loop (D-loop) that does not encode any protein and contains the replication origin of one strand. Mitochondrial control region of the olive ridley turtle contains a seven base indel that is only specific to a single population [5].

4. Indel Coding Methods and Tools

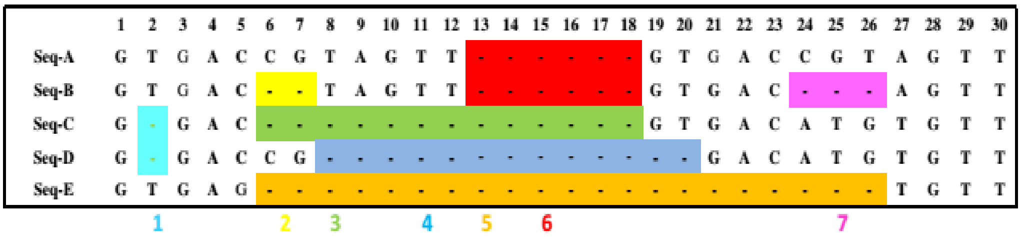
Insertion and deletion (indels) attract increasing interest because they play an important role in genomic evolution. Hardly a few studies incorporated the indel information in phylogenetic analysis and tried to treat different indel as separate binary characters or fifth state characters. Simmons and Ochoterena (2000) proposed simple indel coding (SIC) and the complex indel coding (CIC) procedures to treat indels with six rules[7]. Muller [6] developed a Modified Complex Indel coding (MCIC) based on state transformation cost. Simple indel coding method coded all indels characters by using Gap Coder and FastGap and IndelCoder menu of SeqState program deals with MCIC.

Simple indel coding is conservative and easy to apply in indels contain sequences. After sequence alignment, for each gap a number is assigned and each gap position is also reported. In the distance matrix, 1 represents presence of each gap, and 0 represent the absence of each gap (Figure 1). As per SIC, all gaps with unlike 5'and /or 3' termini are differentiated as presence/ absence characters. If one gap of a sequence completely overlaps a gap of distinct sequence, then small gaps are a subset of a longer one. Gaps showing both the 5' and the 3' termini within a larger gap in another sequence (*e.g.*, Gap 2 in sequences C and E) are represented as missing data (Figure 1). So, there are three types of the gap characters i. gap absent, ii. gap present and iii. Inapplicable. Here both the sequences are compared to show differences. If the identical positions in both the sequences are showing same values are represented by 0 and different values by 1 and the positions showing missing data in any of two sequences are not used. By summing the differences the final distance is obtained. But, this method is does not utilized the all the available information and CIC rules are very difficult to translate into a clearly formulate algorithm [8].

Modified Complex Indel Coding (MCIC) method primarily derived from the Complex indel coding. In MCIC method, first aligned sequences are numbered and converted into binary pattern. All the gaps are coded as 0 and A, T, G or C is coded as 1 (Figure 2, Step 1). There after combined the identical consecutive positions with ones as binary 1 and zeros as binary 0 (Figure2, Step 3) and removed the identical places with zero from both the sequences. Then identical places with one in both the sequences are also removed. Among the remaining positions; those showing either 0-1 or 1-0 in consecutive locations are merged. Now the columns are counted to calculate the distance between the two sequences (see Figure2, Step 6). But the coding method suffers from internal inconsistence when a long indel has a number of shorter subset indel and triangular in equality still exists. Ogden and Rosenberg [9]used gap information during tree construction under the maximum parsimony principle and concluded that all the three gap coding methods perform equally well in topological accuracy. However, maximum likelihood method based analysis (*e.g.* RAxML, PAUP, *etc.*) is statistically inconsistent when sequences evolved with indels. There is an urgent need of statistical- based methods that indels and substitution with statistical evidence and can run on large datasets[10]. Muñoz-Pajares [11] first time developed a software package, Substitution and Indel Distances to Infer Evolutionary Relationships (SIDIER), on R language. SIDIER combines both the gap distance and substitutions distance to infer evolutionary relationships. This combine distance can be used across a wider range of phylogenetic problems and also useful for barcode gap calculation. SIDIER is promising software for intra and inter specific calculation in DNA barcode studies as well as to infer phylogenetic relationships. We suggested that the indel-rich loci may be valuable for phylogenetic but careful attention should be require for selecting alignment algorithm and indel coding methods.

Table 1. List of Software Used in different Studies for Indel Coding Method

Software Name	Method used
GapCoder	SIC
FastGap	SIC
SeqState	MCIC
PAUP*	Missing data/fifth character
SPInDel	Diverse statistical methods
SIDIER	SIC/MCIC/fifthcharacters + substitution



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Seq-A	G	T	G	A	C	C	G	T	A	G	T	T	-	-	-	-	-	-	G	T	G	A	C	C	G	T	A	G	T	T
Seq-B	G	T	G	A	C	-	-	T	A	G	T	T	-	-	-	-	-	-	G	T	G	A	C	-	-	-	A	G	T	T
Seq-C	G	-	G	A	C	-	-	-	-	-	-	-	-	-	-	-	-	-	G	T	G	A	C	A	T	G	T	G	T	T
Seq-D	G	-	G	A	C	C	G	-	-	-	-	-	-	-	-	-	-	-	-	-	G	A	C	A	T	G	T	G	T	T
Seq-E	G	T	G	A	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	G	T	T

(a)

	Gap1	Gap2	Gap3	Gap4	Gap5	Gap6	Gap7
Gap position in alignment	1-3	5-8	5-19	7-21	5-27	12-19	23-27
Seq-A	0	0	0	0	0	1	0
Seq-B	0	1	0	0	0	1	1
Seq-C	1	-	1	0	0	-	0
Seq-D	1	0	0	1	0	-	0
Seq-E	0	-	-	-	1	-	-

(b)

	A	B	C	D	E
Seq-A	0				
Seq-B	2	0			
Seq-C	2	3	0		
Seq-D	2	4	2	0	
Seq-E	1	1	2	2	0

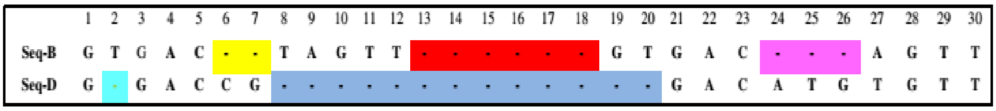
(c)

Seq-B	0	1	0	0	0	1	1
Seq-D	1	0	0	1	0	-	0
△	1	1	0	1	0	-	1

(d)

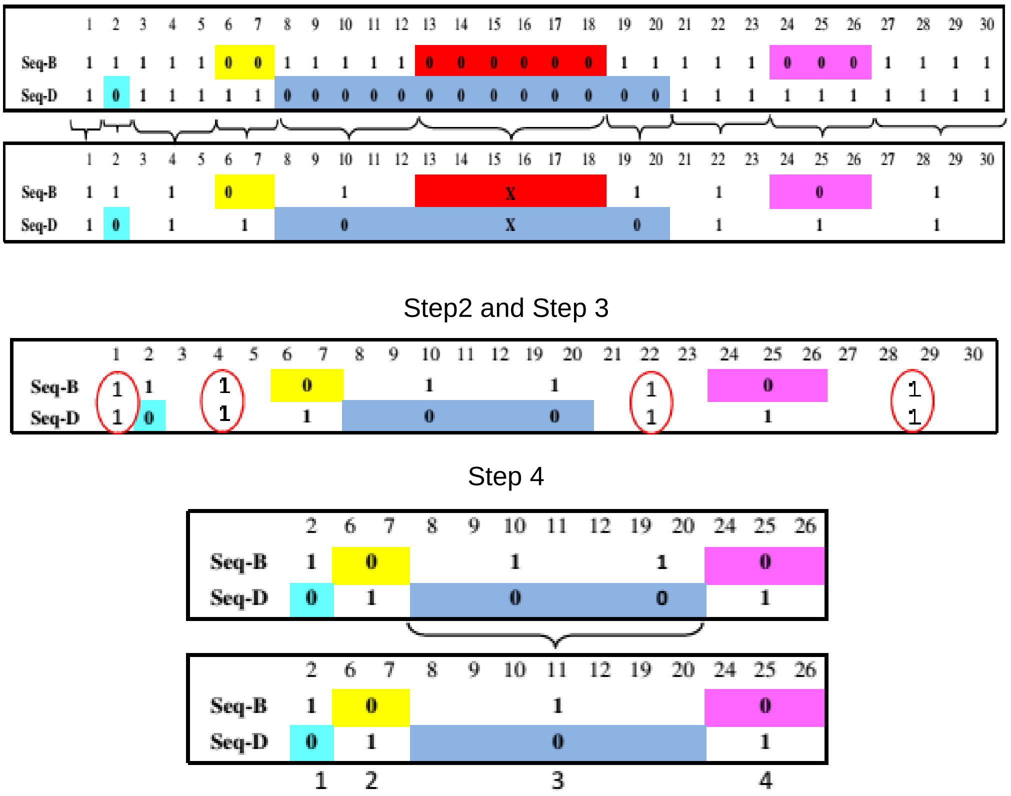
Figure 1. Schematic Representation of Simple Indel Coding Method. A. Multiple Sequence Alignment of Five Sequences. Different Colour indicated different Indel Position in the Alignment. B. Showing Position and the Presence as "1" or Absence as "0" of each Gap in the Aligned Sequences. C. Showing Indel Distance Matrix, Seq-B and Seq-D. D. Calculation of Indel Distances

🟦	represent Gap-1(1-3 position)
🟨	represent Gap-2(5-8 position)
🟦	represent Gap-3(5-19 position)
🟧	represent Gap-4(7-21 position)
🟥	represent Gap-5(12-19 position)
🟪	represent Gap-6(5-27 position)
🟩	represent Gap-7(23-27 position)



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Seq-B	G	T	G	A	C	-	-	T	A	G	T	T	-	-	-	-	-	-	G	T	G	A	C	-	-	-	A	G	T	T
Seq-D	G	-	G	A	C	C	G	-	-	-	-	-	-	-	-	-	-	-	-	-	G	A	C	A	T	G	T	G	T	T

Step 1



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Seq-B	1	1	1	1	1	0	0	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1	0	0	0	1	1	1	1
Seq-D	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Seq-B	1	1	1			0		1					X						1		1			0			1			
Seq-D	1	0	1			1		0					X						0		1			1			1			

Step2 and Step 3

	1	2	3	4	5	6	7	8	9	10	11	12	19	20	21	22	23	24	25	26	27	28	29	30
Seq-B	1	1	1			0		1					1		1			0			1			
Seq-D	1	0	1			1		0					0		1			1			1			

Step 4

	2	6	7	8	9	10	11	12	19	20	24	25	26
Seq-B	1	0		1					1		0		
Seq-D	0	1		0					0		1		

	2	6	7	8	9	10	11	12	19	20	24	25	26
Seq-B	1	0		1							0		
Seq-D	0	1		0							1		
	1	2		3							4		

Step 5 and Step 6

Figure 2. Schematic Representation of Modified Complex Indel Coding (MCIC) Methods. Step 1: For Example We Take the Seq-B and Seq-D from Figure 1. Step2:Translated all Characters as 1 and Gaps as 0.Step 3: Combined the Consecutive Ones (1) and Zeros (0) and Removed the Places that Contain Zeros in Both the Sequences. X Represents Removal of the places, with 0 from both the sequences. Step 4: Ignored the places where Both the Sequences Contain 1. Step 5: Among the Remaining Positions, those Showing either 0-1 or 1-0 are Merged. Step 6: The Number of Indel Events is Estimated by Counting the Number of Positions. So Here the Distance between B and D is 4.

Acknowledgements

We acknowledge DBT-RA fellowships to the author (PM).

References

[1] Group C. P. W., "A DNA barcode for land plants", Proceedings of the National Academy of Sciences of the United States of America, vol. 106, no. 31, **(2009)**, pp. 12794-12797.

[2] Hebert P. D., Ratnasingham S., deWaard J. R., "Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species," Proceedings Biological sciences / The Royal Society 270 (Suppl 1):S, **(2003)**, pp. 96-99.

[3] Ratnasingham S. and Hebert P. D. N., "BOLD: The Barcode of Life Data System", (www.barcodinglife.org). Molecular Ecology Notes, no. 7, **(2007)**, pp. 355-364.

[4] Schocha C. L., Seifertb K. A., Huhndorfc S., Robertd V., Spougea J. L., André Levesqueb C., Chenb W. and Consortiuma F. B., "Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi", Proceedings of the National Academy of Sciences of the United States of America ,vol. 1009, no. 16, **(2012)**, pp. 6241-6246.

[5] Shanker K., Ramadevi J., Choudhury B. C., Singh L. and Aggarwal R. K., "Phylogeography of olive ridley turtles (Lepidochelys olivacea) on the east coast of India: implications for conservation theory", Molecular Ecology, vol. 13, no. 7, **(2004)**, pp. 1899-1909.

[6] Muller K., "Incorporating information from length-mutational events into phylogenetic analysis", Molecular Phylogenetics and Evolution, vol. 38, no. 3, **(2006)**, pp. 667-676.

[7] Simmons M. P. and Ochoterena H., "Gaps as characters in sequence-based phylogenetic analyses", Systematic Biology, vol. 49, no. 2, **(2000)**, pp. 369-381.

[8] Muller K., "SeqState: primer design and sequence statistics for phylogenetic DNA datasets", Applied Bioinformatics, vol. 4, no. 1, **(2005)**, pp. 65-69.

[9] Ogden T. H. and Rosenberg M.S., "How should gaps be treated in parsimony? A comparison of approaches using simulation", Molecular Phylogenetics and Evolution, vol. 42, no. 3, **(2007)**, pp. 817-826.

[10] Warnow T., "Standard maximum likelihood analyses of alignments with gaps can be statistically inconsistent", PLoS Currents 4:RRN1308, **(2012)**.

[11] Muñoz-Pajares A. J., "SIDIER: Substitution and indel distances to infer evolutionary relationships", Methods in Ecology and Evolution, vol. 4, **(2013)**, pp. 1195-1200.

[12] Li W. H., "Molecular Evolution", MA. Sinauer, Sunderland, **(1995)**.

Authors

Gautam Sanyal, (M. Tech, Ph.D.) National Institute of Technology, Durgapur, possesses more than 30 years' experience in teaching and research. He has published more than 50 research papers in International and National Journals / Conferences.

Asim Kr Mahadani, M. Tech (computer Science from Jadavpur University, Kolkata), Assistant Professor, Department of Computer Science, Bankura Unnayani Institute of Engineering. He is perusing his Ph. D in Computer Science from National Institute of Technology, Durgapur.

Pradosh Mahadani, PhD (Biotechnology), M.Sc. (Bioinformatics), research interest focuses on the areas of Bioinformatics, D NA barcoding, Molecular Biology. Currently working as Research Associate in ICAR-National Research Center for Orchids, Sikkim.

Partha Bhattacharjee, Senior Principal Scientist, CSIR-CMERI, He worked in various originations like Operation Research Group, India, C-DOT, Delhi. He published so many research papers in various Journals and Conferences.