BMC Bioinformatics

BioMed Central

Methodology article

Open Access

GapCoder automates the use of indel characters in phylogenetic analysis

Nelson D Young*[1] and John Healy[2]

Address: [1]Department of Biological Sciences, Duquesne University, Pittsburgh, PA 15219, USA and [2]Biology Department, Trinity University, 715 Stadium Dr., San Antonio, TX 78212, USA

Email: Nelson D Young* - youngnd@duq.edu; John Healy - praehotec8@yahoo.com

* Corresponding author

Published: 19 February 2003

BMC Bioinformatics 2003, **4**:6

Received: 29 November 2002
Accepted: 19 February 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/6

Abstract

Background: Several ways of incorporating indels into phylogenetic analysis have been suggested. Simple indel coding has two strengths: (1) biological realism and (2) efficiency of analysis. In the method, each indel with different start and/or end positions is considered to be a separate character. The presence/absence of these indel characters is then added to the data set.

Algorithm: We have written a program, GapCoder to automate this procedure. The program can input PIR format aligned datasets, find the indels and add the indel-based characters. The output is a NEXUS format file, which includes a table showing what region each indel characters is based on. If regions are excluded from analysis, this table makes it easy to identify the corresponding indel characters for exclusion.

Discussion: Manual implementation of the simple indel coding method can be very time-consuming, especially in data sets where indels are numerous and/or overlapping. GapCoder automates this method and is therefore particularly useful during procedures where phylogenetic analyses need to be repeated many times, such as when different alignments are being explored or when various taxon or character sets are being explored. GapCoder is currently available for Windows from http://www.home.duq.edu/~youngnd/GapCoder.

Background

The position of insertion/deletion mutations (indels) in molecular data sets can be useful phylogenetic information [1–4], yet this information is rarely used, especially in large data sets with many indels. There are three main reasons for this. First, some workers believe that indels may be unreliable as characters [5]. However, numerous studies in which indel characters were compared with already established tree topologies have found that these indels are reliable in constructing phylogenies [6–11]. Second, it can be very time-consuming to determine character states based on gaps and enter this information into a data matrix by hand. Third, there is disagreement as to the best method of defining homologous character states for indels. Several different methods for incorporating indels into phylogenetic analyses have been used. We discuss five of the most useful of these methods.

The computer program MALIGN uses the first of these methods of including indels in sequence alignment and phylogenetic analysis of sequences [12]. In this method, gap characters are considered to be a fifth character state for bases in DNA, as in Eernisse and Kluge [1]. Therefore, adjacent gap characters are considered independently of their neighbors, although subsequent gap characters after the first may be weighted less heavily to reflect the possibility of longer indel regions [12,13]. Essentially, each individual gap position is considered as if it were a separate indel event. This is not very realistic. Insertion or deletion events often consist of multiple bases [14–16]. Since many gap characters do not arise independently of one another, counting each gap character as a separate event causes indel events to be considered multiple times in determining phylogenetic relationships. This overweights the indels and can distort phylogenies. Simmons and Ochoterena [16] also note a theoretical objection: because gaps are the product of the alignment procedure, and are not actually found in organisms or their sequences, sequences with gap characters do not have anything to compare with other sequences at the point where the gap occurs. For these reasons, gaps should not be considered as a fifth character state for nucleotide characters.

The second method, optimization alignment, is implemented in the program POY [17]. POY achieves a phylogenetic analysis, including indels as character state changes, without ever creating a multiple-sequence alignment. Allthough this avoids the major problems with MALIGN, it has a limitation. Indel changes may be weighted more heavily than substitutions, but the same weight is used for the determining the position of indels and phylogenetic analysis. For example, it is not possible to use a gap weight of 10 (an indel is equivalent to 10 substitutions), as is common in protein-coding regions, without also weighing that change 10 times as much as a substitution in phylogenetic analysis.

The third method to be considered is the multistate gap region method [4,18–20]. In this method, areas of overlapping indels, gap regions, are coded as individual characters. Different indels within each region are considered to be different states for the corresponding multistate gap region characters [4]. Within the DNA sequences, gap characters are coded as missing data, and the gap region characters are then placed at the end of each sequence. This method is useful because it does code indels as separate characters and does consider contiguous gap characters as related. However, the number of character states for each gap region can be quite large. Since there are so many different possible states, these characters can be less informative regarding relationships than other methods.

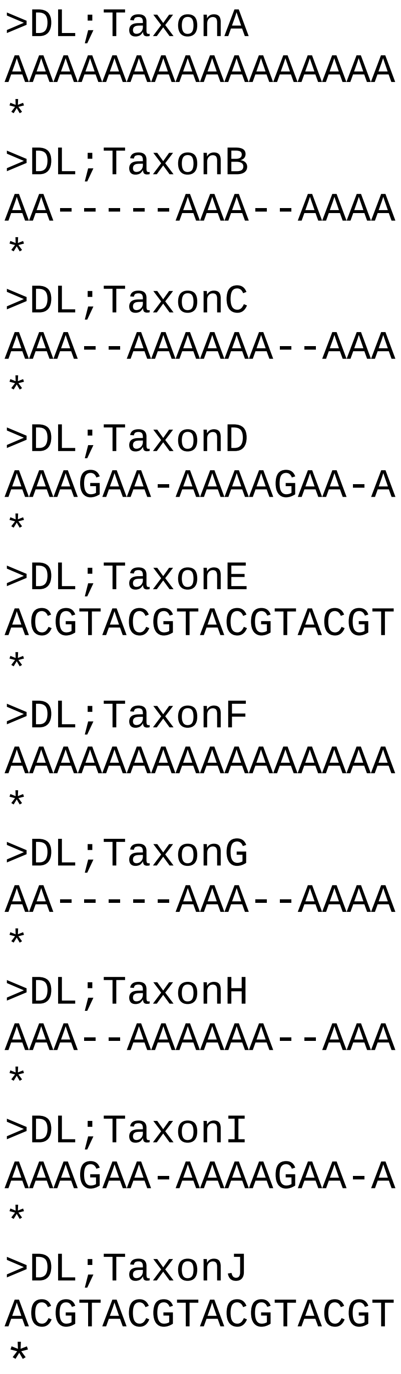
Simmons and Ochoterena have proposed a fourth method for coding indels [16]. This method is termed "simple indel coding". Similar to the third method, this process codes indels as separate characters in a data matrix, which is then considered along with the DNA base characters in phylogenetic analysis. Each indel with different start and/or end positions is considered to be a separate character, which all of the taxa under consideration either have or lack. If one of the indels completely overlaps an indel contained within another sequence, the sequences containing the longer indel are coded as being inapplicable for the shorter indel. This is done because it is impossible to determine whether or not the shorter indel is present in the sequences containing the longer one. Simple indel coding has the advantages of being conservative and easy to implement while still allowing indels to be highly informative in determining a correct phylogeny [16].

The final method for indel coding is also described by Simmons and Ochoterena [16]. This method is called complex indel coding. This method attempts to better account for the fact that indels are evolutionarily related to one another, and that an indel region may be modified through additional insertion/deletion events to yield a different indel region in another sequence. Complex indel coding, like simple indel coding, codes indels with different start and end positions as individual characters. However, overlapping indels may represent an evolutionary transition sequence [16]. Step matrices are constructed to accommodate this possibility. Complex indel coding utilizes more of the available information and never implies fewer steps than what is biologically realistic. However, this method generates some multi-state characters and step matrices and is thus more complicated to program. Also, the step matrices slow down phylogenetic programs. For a more thorough discussion of indels and their purpose in phylogenetic analysis, see reference [16].

Algorithm

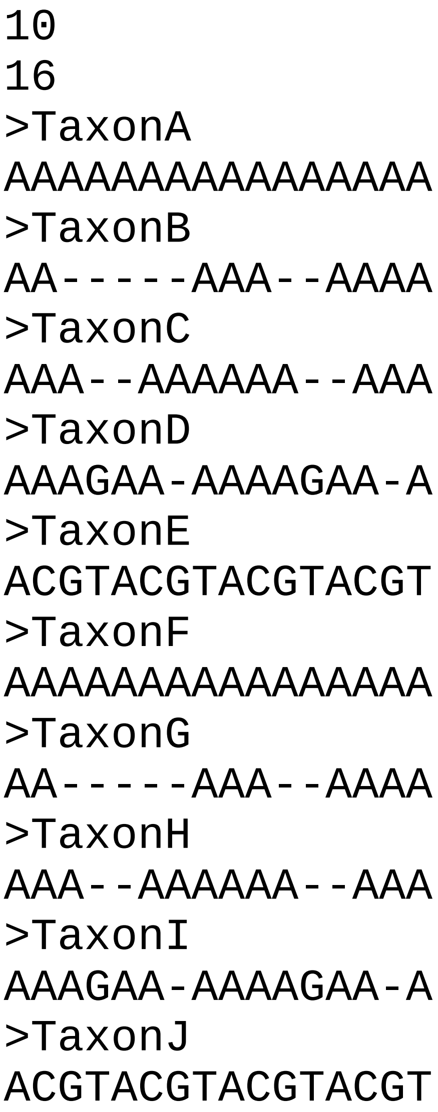
The GapCoder program

Simple indel coding [16] was chosen for implementation because it is a relatively simple algorithm. In addition, simple indel coding does not make as many assumptions as complex indel coding. As a result, GapCoder should be acceptable to a wide range of researchers with different views about the exact nature of indels. GapCoder considers homologous indels or gaps to be those with the same start and end positions in the nucleotide sequences. Indels are not homologous if they have differing lengths, because it would take additional mutations to transform one into another [16]. GapCoder takes a pre-aligned PIR-format or modified FASTA-format file as input, and examines it to gather information about the positions of the indel regions. Figures 1 and 2 illustrate the two valid input file types. The first of these file types, the PIR-format file, can be automatically generated by programs such as ClustalX. The second file type, the modified FASTA-format file, is shown in Figure 2. This file differs from the standard FASTA-format by the inclusion of the two numbers at the top of the file. The first number is the number of taxa contained in the file, and the second number is the number of bases in each of the taxa. The taxon names and sequences are placed below the numbers. The output from GapCoder is a NEXUS-format file. The new characters created



```
>DL;TaxonA
AAAAAAAAAAAAAAAA
*
>DL;TaxonB
AA-----AAA--AAAA
*
>DL;TaxonC
AAA--AAAAAA--AAA
*
>DL;TaxonD
AAAGAA-AAAAGAA-A
*
>DL;TaxonE
ACGTACGTACGTACGT
*
>DL;TaxonF
AAAAAAAAAAAAAAAA
*
>DL;TaxonG
AA-----AAA--AAAA
*
>DL;TaxonH
AAA--AAAAAA--AAA
*
>DL;TaxonI
AAAGAA-AAAAGAA-A
*
>DL;TaxonJ
ACGTACGTACGTACGT
*
```

Figure 1
Sample input file, PIR format



```
10
16
>TaxonA
AAAAAAAAAAAAAAAA
>TaxonB
AA-----AAA--AAAA
>TaxonC
AAA--AAAAAA--AAA
>TaxonD
AAAGAA-AAAAGAA-A
>TaxonE
ACGTACGTACGTACGT
>TaxonF
AAAAAAAAAAAAAAAA
>TaxonG
AA-----AAA--AAAA
>TaxonH
AAA--AAAAAA--AAA
>TaxonI
AAAGAA-AAAAGAA-A
>TaxonJ
ACGTACGTACGTACGT
```

Figure 2
Sample input file, modified FASTA format

```
#NEXUS
BEGIN DATA;
     DIMENSIONS NTAX=10 NCHAR=22;
     FORMAT MISSING=? DATATYPE=DNA GAP=- EQUATE="0=A 1=C";
     OPTIONS GAPMODE=MISSING;
MATRIX

                        [0000000000000000000000]
                        [0000000000000000000000]
                        [0000000001111111111222]
                        [1234567890123456789012]

TaxonA                   AAAAAAAAAAAAAAAA000000
TaxonB                   AA-----AAA--AAAA1--100
TaxonC                   AAA--AAAAAA--AAA010010
TaxonD                   AAAGAA-AAAAGAA-A001001
TaxonE                   ACGTACGTACGTACGT000000
TaxonF                   AAAAAAAAAAAAAAAA000000
TaxonG                   AA-----AAA--AAAA1--100
TaxonH                   AAA--AAAAAA--AAA010010
TaxonI                   AAAGAA-AAAAGAA-A001001
TaxonJ                   ACGTACGTACGTACGT000000
;
END;

[ Indel Character     Sequence Region ]
[ ---------------     --------------- ]
[                                     ]
[ 17             3-7         ]
[ 18             4-5         ]
[ 19             7-7         ]
[ 20             11-12         ]
[ 21             12-13         ]
[ 22             15-15         ]
```

Figure 3

Sample output file. Output files are in the NEXUS format and ready to be input into PAUP or other programs that use this format. The indel characters have been added to the matrix and a table of correspondences is appended in the form of a comment, showing each indel character and the position of the indel upon which it is based. The Equate command allows 0 and 1 to be used, while maintaining the data type as 'DNA'. This allows one to perform maximum likelihood and other analyses that require this data type, though if a model of DNA substitution is applied, it may be most appropriate to exclude the indel characters from the analysis. They probably don't evolve according to the same model as substitutions.

by the algorithm are placed at the end of the data. In addition, a table of correspondences between the indels and their codes is placed at the bottom of the file. If regions are excluded from an analysis, this table makes it easy to identify the corresponding indel characters for exclusion. An example output file corresponding to the input given in Fig. 1 or Fig. 2 is shown in Fig. 3. The indel characters coded by the program can be seen listed at the end. Each indel character can be in one of three states for each taxon: present, missing or inapplicable. The indel characters are coded with a '1' for present, '0' for missing, and '-' for inapplicable. When one or more indels are contained completely within a larger indel, all of the taxa that have the larger indel are coded with inapplicable ('-') characters for the smaller indels. For example, consider the first two indels listed in the correspondence table at the bottom of Figure 3. The table lists these indels as characters 17 and 18 at the ends of the sequences. The indel represented by character 17 occurs from characters 3–7 in the matrix. TaxonB and TaxonG have the indel in place of bases 3–7 and receive a '1' for character 17. Character 18 occurs from characters 4–5. TaxonC and TaxonH clearly have the indel and are scored as '1' for character 18. However, since the first indel completely covers the entire region of the second indel, it is unclear whether TaxonB or TaxonG could have had the first indel. Therefore, these taxa are given a '-' for character 18.

Discussion

GapCoder has the potential to be useful in phylogenetics, especially in non-protein-coding regions where indels can be as plentiful as substitutions. Whenever multiple phylogenetic analyses are performed, or greater resolution is required, GapCoder provides an efficient way to incorporate the phylogenetic information contained in the indels. For example, the output resulting from GapCoder may be used in exploratory analyses of optimal DNA sequence alignment. Such an analysis would likely include GapCoder as part of an objective method with four stages. In the first stage, several alignments would be created using a program such as ClustalX. GapCoder would then be used to code the indels into the data matrix. Next, a phylogenetic analysis of the data would be performed using software such as PAUP. Finally, the best alignment could be chosen using the desired optimality criterion. GapCoder is also useful when different character sets and/or taxon sets are being explored, such as when different combinations of outgroups are tried. This often requires re-aligning the data set for each taxon set; GapCoder allows the indel characters to be quickly added each time.

Authors' contributions

NY conceived of and oversaw the project, wrote a small portion of the code and participated in the testing. JH designed and wrote the program itself, and also did much of the testing. Both authors read and approved the final manuscript.

Additional material

Additional File 1

GapCoder is currently available for the Windows platform. Instructions for use can be found by visiting http://www.home.duq.edu/~youngnd/GapCoder. The executable file Gapcoder.exe may be obtained by clicking on the link below or visiting the website. The source code is available on request.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-6-S1.zip]

References

1. Eernisse DJ and Kluge AG **Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology.** *Mol Biol Evol* 1993, **10:**1170-1195
2. Vogler AP and DeSalle R **Evolution and phylogenetic information content of the ITS-1 region in the tiger beetle *Cicindela dorsalis*.** *Mol Biol Evol* 1994, **11:**393-405
3. Simmons AM and Mayden RL **Phylogenetic relationships of the creek chubs and the spine-fins: An enigmatic group of North American cyprinid fishes (Actinopterygii: Cyprinidae).** *Cladistics* 1997, **13:**187-206
4. Freudenstein JV and Chase MW **Analysis of mitochondrial *nad1*b-c intron sequences in Orchidaceae: Utility and coding of length-change characters.** *Syst Bot* 2001, **26:**643-657
5. Golenberg EM, Clegg MT, Durbin ML, Doebley J and Ma DP **Evolution of a non-coding region of the chloroplast genome.** *Mol Phylogenet Evol* 1993, **2:**52-64
6. Lloyd DG and Calder VL **Multi-residue gaps, a class of molecular characters with exceptional reliability for phylogenetic analyses.** *J Evol Biol* 1991, **4:**9-21
7. Van Ham RCHJ, Hart H, Mes THM and Sandbrink JM **Molecular evolution of noncoding regions of the chloroplast genome in the Crassulaceae and related species.** *Curr Genet* 1994, **25:**558-566
8. Johnson LA and Soltis DE **Phylogenetic inference in Saxifragaceae sensu stricto and *Gilia* (Polemoniaceae) using *mat* K sequences.** *Ann Mo Bot Gard* 1995, **82:**149-175
9. Baldwin BG and Markos S **Phylogenetic utility of the external transcribed spacer (ETS) of 18S–26S rDNA: Congruence of ETS and ITS trees of *Calycadenia* (Compositae).** *Mol Phylogenet Evol* 1998, **10:**449-463
10. Prather LA and Jansen RK **Phylogeny of *Cobaea* (Polemoniaceae) based on sequence data from the ITS region of nuclear ribosomal DNA.** *Syst Bot* 1998, **23:**57-72
11. Simmons MP, Ochoterena H and Carr TG **Incorporation, relative homoplasy, and effect of gap characters in sequence-based phylogenetic analysis.** *Syst Biol* 2001, **50:**454-462
12. Wheeler WC and Gladstein DS **MALIGN: A multiple sequence alignment program.** *J Hered* 1994, **85:**417-418
13. Giribet G and Wheeler WC **On gaps.** *Mol Phylogenet Evol* 1999, **13:**132-143
14. Pascarella S and Argos P **Analysis of insertions/deletions in protein structures.** *J Mol Biol* 1992, **224:**461-471
15. Gu X and Li W-H **The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment.** *J Mol Evol* 1995, **40:**464-473
16. Simmons MP and Ochoterena H **Gaps as characters in sequence-based phylogenetic analyses.** *Syst Biol* 2000, **49:**369-381
17. Wheeler WC **Optimization alignment:the end of multiple alignment in phylogenetics?** *Cladistics* 1996, **12:**1-9

18. Baum DA, Sytsma KJ and Hoch PC **A phylogenetic analysis of *Epilobium* (Onagraceae) based on nuclear ribosomal DNA sequences.** *Syst Bot* 1994, **19:**363-388
19. Young ND, Steiner KE and dePamphilis CW **The evolution of parasitism in Scrophulariaceae/Orobanchaceae: plastid gene sequences refute an evolutionary transition series.** *Ann Missouri Bot Gard* 1999, **86:**876-893
20. Lutzoni F, Wagner P, Reeb V and Zoller S **Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology.** *Syst Biol* 2000, **49:**628-651
21. Swofford DL **PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4.** *Sinauer Associates, Sunderland, Massachussetts* 1998,