BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 21 2007, pages 2957–2958
doi:10.1093/bioinformatics/btm468

Phylogenetics

CREx: inferring genomic rearrangements based on common intervals

Matthias Bernt[1,*], Daniel Merkle[1], Kai Ramsch[1], Guido Fritzsch[2,3], Marleen Perseke[4], Detlef Bernhard[4], Martin Schlegel[4], Peter F. Stadler[2,3,5,6,7] and Martin Middendorf[1]

[1]Parallel Computing and Complex Systems Group, [2]Bioinformatics Group, Department of Computer Science, [3]Interdisciplinary Center for Bioinformatics, [4]Institute of Biology II, University of Leipzig, [5]Fraunhofer Institute IZI, Leipzig, Germany, [6]Department of Theoretical Chemistry, University of Vienna, Austria and [7]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Received on June 27, 2007; revised and accepted on September 7, 2007

Advance Access publication September 25, 2007

Associate Editor: Keith Crandall

ABSTRACT

Summary: We present the web-based program CREx for heuristically determining pairwise rearrangement events in unichromosomal genomes. CREx considers transpositions, reverse transpositions, reversals and tandem-duplication-random-loss (TDRL) events. It supports the user in finding parsimonious rearrangement scenarios given a phylogenetic hypothesis. CREx is based on common intervals, which reflect genes that appear consecutively in several of the input gene orders.

Availability: CREx is freely available at http://pacosy.informatik.uni-leipzig.de/crex

Contact: merkle@informatik.uni-leipzig.de

1 INTRODUCTION

Genomic rearrangement operations present a powerful approach to determine the phylogenetic relationship of species. In this context, unichromosomal genomes are usually encoded as signed permutations, in which each element represents a gene and the sign determines the orientation of the gene. Inversions (mostly referred to as 'reversals' in the bioinformatics literature) are by far the best-studied gene order rearrangement operations. Given a set of gene orders, software packages such as GRAPPA (Moret *et al.*, 2005) and amGRP (Bernt *et al.*, 2007a) try to find (i) a topology of a binary tree with the gene orders as leaf nodes and (ii) gene orders for the inner nodes of that tree, such that the overall reversal distance of all edges is minimal.

Phylogenetic reconstruction methods based on more general sets of rearrangement operations are rare, presumably because for some of the operations of interest the problem of computing an exact distance is unsolved or computationally hard. On the other hand, there is ample evidence in the biological literature that—particularly in mitochondrial genomes—rearrangements are the consequence of multiple mechanisms (Boore 1999). Biological findings thus strongly influenced methods for phylogenetic reconstruction methods; it is well known, for instance, that gene groups are often preserved during evolution. Algorithmically, this property is reflected by so-called common intervals. Strong common interval trees, which are close relatives of PQ-trees (Booth and Lueker, 1976), are a data structure that is particularly well suited for inferring rearrangement scenarios (Bérard *et al.*, 2007). Algorithms for gene order rearrangements that preserve common intervals are also discussed by Bernt *et al.*, (2007b). Strong common intervals, on the other hand, were already used to reconstruct rearrangement scenarios with the focus on reversals and transpositions (Parida 2006). CREx is based on these approaches and extends them to a larger set of operations.

2 METHODS

Formally, a common interval is a subset of genes that appear consecutively in two (or more) input gene orders (Bérard *et al.*, 2007). Two intervals I and J are said to commute if either $I \subset J$, or $I \supset J$, or $I \cap J = \emptyset$ holds. A common interval is a *strong interval* if it commutes with every common interval. A *strong interval tree* (SIT) for two gene orders, finally, is a rooted tree that has exactly one leaf for each gene and exactly one inner node for each strong common interval. The edges of the tree are defined by the inclusion order of the set of strong intervals. There exist two types of inner nodes. An inner node is called (increasing or decreasing) linear if its child nodes are in a left-to-right or right-to-left order, otherwise the child nodes are not ordered and the inner node is called prime.

The basic principle for computing heuristic gene order rearrangement scenarios with CREx is to detect patterns in the SITs that reflect the corresponding genome rearrangement operations. For example, reversals are very simple to detect because this operation is reflected as a sign difference of parent node and child nodes in the SIT. Prime nodes are good indicators for TDRL events. Both transpositions and reverse transpositions also lead to recognizable patterns in the SIT (for details see the webpage of CREx). CREx uses a stepwise approach to suggesting a genome rearrangement scenario: first it identifies transpositions and reverse transpositions, then reversals are identified based on the sign differences between connected nodes in the SIT. In these first two steps, CREx only operates on linear nodes. In the third step, the prime nodes are analyzed to identify combinations of reversal and TDRL operations (including transpositions), which can explain the corresponding prime nodes.

*To whom correspondence should be addressed.

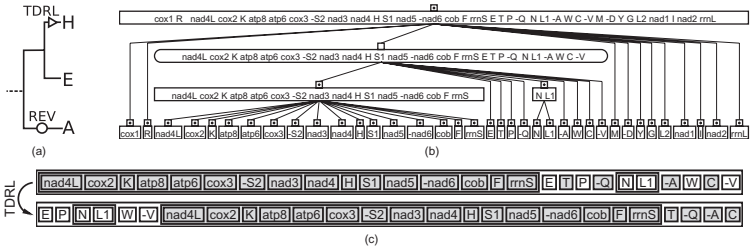

Fig. 1. (**a**) Mitochondrial gene order rearrangement scenario inferred by CREx for the given phylogeny of Asteroidea (A) Echinoidea (E) and Holothuroidea (H) showing one reversal (REV) and one TDRL; (**b**) SIT for E → H and (**c**) TDRL of E → H as suggested by CREx; (b) and (c) are exported from CREx.

TDRLs are of particular interest for phylogenetic analysis because the distance measure between gene orders that is based on the minimum number of TDRLs is not symmetric (Chaudhuri *et al*., 2006). In particular, in many cases a rearrangement can be explained by a single TDRL in one direction, while reversing the rearrangement would require more than a single operation. It follows that TDRLs imply the evolutionary direction of the rearrangement in many cases and hence allow the re-construction of the ancestral state from a comparison of two gene orders without considering an outgroup. In contrast, reversals, transpositions and reverse transpositions are inherently symmetric, hence ancestral states cannot be reconstructed without outgroup information.

3 CREX

CREx is a web-based application for analyzing gene orders based on the application server Zope. The algorithms handling the common interval data structures and for computing scenarios were implemented in C++. They were integrated into Zope via Python modules. For drawing publication ready downloadable versions of the SITs ReportLab was used.

After uploading the gene order data in FASTA format to CREx, a matrix is computed and displayed, with elements colored according to an evolutionary distance measure, such that gene orders with a small evolutionary distance can be easily identified. Therefore, the breakpoint distance, the reversal distance or the number of common intervals is utilized. Columns, rows and individual elements of the matrix can be selected; for the selected elements the SITs are computed and displayed as a tree or as a family diagram (Bergeron and Stoye, 2006). As briefly described in Section 2, the structure of the SIT can be used to infer genomic rearrangement operations connecting a pair of input gene orders. CREx uses its heuristic approach to suggest a rearrangement scenario based on common intervals. CREx allows the user to select individual operations of the scenario to highlight the affected common intervals.

4 CASE STUDY AND DISCUSSION

While testing CREx we found that, to the best of our knowledge, all reported cases of TDRL events in mitochondrial gene orders in the literature were correctly identified (Arndt and Smith 1998; Inoue *et al*., 2003). We furthermore used CREx to analyze the genomic rearrangement history of the complete mitochondrial genomes of echinoderms (including the newly sequenced mitogenomes of the crinoid *Antedon mediterranea* and the ophiuroid *Ophiura albida*), see Perseke (*et al*., 2007).

For the sake of space, we can discuss here only a small example to illustrate the functionality of CREx. We analyze the mitochondrial gene order rearrangement history of Asteroidea (A), Echinoidea (E), and Holothuroidea (H) (within these groups all sequenced species have the same gene order). CREx identifies three TDRL events for the evolution H → E, and one TDRL event, which was analyzed in detail by Arndt and Smith (1998), for E → H. This implies that the ancestral state conforms E. For A ⟷ E a single reversal is needed. As CREx suggests the same reversal and the same TDRL for A ⟷ H (not shown here), we can immediately infer the genome rearrangement events for the phylogeny in Figure 1a. A more detailed analysis including all published echinoderm mitochondrial genomes and different phylogenetic hypotheses is presented in Perseke *et al*. (2007). The SIT for E → H is given in Figure 1b, the corresponding TDRL is depicted in Figure 1c.

ACKNOWLEDGEMENT

This work was supported by the DFG grant STA 850/3-1 and SCHL 229/14-1 as part of the SPP-1174 (Metazoan Deep Phylogeny).

Conflict of Interest: none declared.

REFERENCES

Arndt,A and Smith,M.J (1998) Mitochondrial gene rearrangement in the sea cucumber genus cucumaria. *Mol. Biol. Evol.*, **15**, 9–16.

Bérard,S *et al*. (2007) Perfect sorting by reversals is not always difficult. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **4**, 4–16.

Bergeron,A and Stoye,J (2006) On the similarity of sets of permutations and its applications to genome comparison. *J. Comput. Biol.*, **13**, 1345–1354.

Bernt,M *et al*. (2007a) Using median sets for inferring phylogenetic trees. *Bioinformatics*, **23**, e129–e135.

Bernt,M *et al*. (2007b) A fast and exact algorithm for the perfect reversal median problem. In *Bioinformatics Research and Applications, Proc. of ISBRA 2007, number 4463 in LNBI*. Springer, Berlin/Heidelberg, pp. 305–316.

Boore,J.L (1999) Survey and summary animal mitochondrial genomes. *Nucleic Acids Res.*, **27**, 1767–1780.

Booth,K.S and Lueker,G.S (1976) Testing for the consecutive ones property, interval graphs, and planarity using pq-tree algorithms. *J. Comput. Syst. Sci.*, **13**, 335–379.

Chaudhuri,K *et al*. (2006) On the tandem duplication-random loss model of genome rearrangement. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*. ACM Press, New York, NY, pp. 564–570.

Inoue,J.G *et al*. (2003) Evolution of the deep-sea gulper eel mitochondrial genomes: large-scale gene rearrangements originated within the eels. *Mol. Biol. Evol.*, **20**, 1917–1924.

Moret,B *et al*. (2005) Reconstructing phylogenies from gene-content and gene-order data. *Mathematics of Evolution and Phylogenic*, Chapter 12, pp. 321–352.

Parida,L (2006) Using pq structures for genomic rearrangement phylogeny. *J. Comput. Biol.*, **13**, 1685–1700.

Perseke,M *et al*. (2007) *Evolution of mitochondrial gene orders in echinoderms* (submitted).