BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 18 2012, pages 2388–2390
doi:10.1093/bioinformatics/bts457

Phylogenetics

Advance Access publication July 20, 2012

ANGES: reconstructing ANcestral GEnomeS maps

Bradley R. Jones[1], Ashok Rajaraman[1], Eric Tannier[2] and Cedric Chauve[1,*]

[1]Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada and [2]INRIA Rhône-Alpes, F-38334 Montbonnot, France; Université de Lyon, F-69000, Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France

Associate Editor: David Posada

ABSTRACT

Summary: ANGES is a suite of Python programs that allows reconstructing ancestral genome maps from the comparison of the organization of extant-related genomes. ANGES can reconstruct ancestral genome maps for multichromosomal linear genomes and unichromosomal circular genomes. It implements methods inspired from techniques developed to compute physical maps of extant genomes. Examples of cereal, amniote, yeast or bacteria ancestral genomes are provided, computed with ANGES.

Availability: ANGES is freely available for download at http://paleogenomics.irmacs.sfu.ca/ANGES/. Documentation and examples are available together with the package.

Contact: cedric.chauve@sfu.ca

Received on May 11, 2012; revised on July 14, 2012; accepted on July 16, 2012

1 INTRODUCTION

The reconstruction of the organization of ancestral genomes is a long-standing problem, motivated by several applications in comparative genomics (Pennisi, 2005). It has received a renewed interest during the past few years, due, among others, to the increasing number of sequenced and assembled genomes and major methodological advances. Two general approaches have been followed by recent methods. The global parsimony aims at computing ancestral gene orders that minimize the number of genome rearrangements; following the pioneering work of Sankoff *et al.*, it can now address hard problems of ancestral genome reconstruction (Zheng and Sankoff, 2012). The local parsimony approach that was pioneered in Ma *et al.* (2006) follows principles used in computing physical maps of extant genomes and was explored in several recent articles (Chauve and Tannier, 2008; Chauve *et al.*, 2010; Ma *et al.*, 2006; Muffato *et al.*, 2010).

In a series of recent articles, we followed the latter approach and described different methods for computing ancestral genome maps from the comparison of extant species. All methods are centered on variants of a combinatorial framework widely used to compute physical maps of extant genomes, the Consecutive-Ones Property (Chauve and Tannier, 2008; Chauve *et al.*, 2010; Gavranovic *et al.*, 2011). We implemented this framework for the reconstruction of animal, fungi and plant ancestral genomes and showed its validity on well-accepted ancestral genomes as well as its robustness to parameter change.

*To whom correspondence should be addressed.

In the present note, we describe ANGES, a suite of Python programs implementing and extending several of these methods, to reconstructing both eukaryotic and prokaryotic ancestral genome maps. It is currently the only available method able to handle a wide range of genetic or genomic data on any domain of life.

2 METHODOLOGICAL OVERVIEW

ANGES requires two kinds of data as input: a 'species tree', annotated to indicate an ancestral species (the ancestor), and a set of homologous 'markers' families.

The markers are genomic segments that are present in the current extant species. A family describes an ancestral marker, which is assumed to have been present in the ancestral genome, in single copy and with no overlap between different markers. They can be obtained through whole-genome alignments to detect families of genomic segments that evolved only through limited local rearrangements, as in Chauve and Tannier (2008), Gavranovic *et al.* (2011) and Ma *et al.* (2006) or using gene families as in Muffato *et al.* (2010) and Chauve *et al.* (2010). ANGES computes the organization of these markers into chromosomal segments of the ancestor, i.e. an ancestral genome map. ANGES proceeds in two stages:

(1) For each pair of species whose evolutionary path in the species tree contains the ancestor, ANGES detects genomic segments with a similar organization (in terms of markers) in both compared species. It then derives 'Ancestral Contiguous Sets' (ACS), which are sets of ancestral markers that should be contiguous in an ancestral genome. Each ACS is given a weight according to the pattern of occurrence of its support in the considered extant species.

(2) ANGES organizes the markers into linearly or circularly ordered ancestral chromosomal segments called 'Contiguous Ancestral regions' (Ma *et al.*, 2006), by extracting a subset of ACS that satisfies a variant of the Consecutive-Ones Property (the C1P).

This general approach can be implemented in several ways, and ANGES implements a wide range of variations.

ACS can be of two types. Adjacencies are sets of two markers that are believed to be contiguous in the ancestral genomes. Common intervals are sets of two or more markers that are also believed to be contiguous in the ancestral genomes. Common intervals are computed using efficient algorithms

(Bergeron *et al.*, 2008; Schmidt and Stoye, 2004). Markers missing in some extant genomes, due for example to gene loss, are a major issue, especially when considering a large number of extant genomes (Gavranovic *et al.*, 2011). To address this issue, ANGES implements several methods to post-process ACS in order to account for missing markers, which are described in detail in the ANGES manual. ANGES can, for example, handle whole-genome duplications, possibly followed by massive gene loss (Chauve *et al.*, 2010; Gavranovic *et al.*, 2011).

Ancestral genome maps are then computed from the set of ACS by selecting a subset of these ACS that satisfies the C1P (for multichromosomal linear genomes) or the circular C1P (for unichromosomal circular genomes). ANGES computes such subsets of ACS using a greedy heuristic defined in Ma *et al.* (2006), a branch-and-bound algorithm or a spectral seriation algorithm (Atkins *et al.*, 1998) (the latter method is a new feature that was not described in our previous articles). Ancestral genome maps are represented by the PQ-tree data structure [multichromosomal linear genomes, see Chauve and Tannier, 2008)] or the related PC-tree (circular genomes). The ability to compute circular ancestral genome is an important new feature of ANGES with respect to our previous works that considered only eukaryotic ancestral genomes. ANGES permits the use of C1P variants, such as the sandwich-C1P (Gavranovic *et al.*, 2011), by translating ACS into a correlation matrix that is used as input to a spectral seriation algorithm or the C1P with multiplicity that can include information about telomeric ACS (Chauve *et al.*, 2011).

The methods implemented by ANGES are predictive and in most cases cannot be assessed against true results, as ancestral genomes at the evolutionary scale of tens or hundred millions of years are unavailable. Some specific ancestors, regarded as references by the genomics and cytogenetics communities, have been used to assess the accuracy of ancestral genome reconstruction methods, such as the boreoeutherian or saccharomyces ancestors [see (Chauve and Tannier, 2008; Chauve *et al.*, 2010)]. Comparison with simulations (Ma *et al.*, 2006), robustness studies (Ouangraoua *et al.*, 2011) or statistical support of local features (Ma *et al.*, 2006) has also been reported. Probably more useful, the number of discarded ACS to satisfy the C1P is a good indicator of the conflicting nature of the conserved synteny signal that defines ACS. For all the ancestors reported using the methods implemented in ANGES, we could observe a very low conflicting signal with well under 5% of ACS discarded.

3 IMPLEMENTATION

The general principle of our implementation is that every computational step is implemented into a single Python script, which takes as input a set of text files and computes, as output, a set of text files. Hence, the Python scripts that compose ANGES can be organized into a complete ancestral genome computation pipeline or can be used as stand-alone scripts. ANGES contains efficient implementations for generic tasks in comparative genomics such as computing common intervals and checking several variants of the C1P on binary matrices. In some cases, it is, as far as we know, the first available implementation of recent algorithms (Bergeron *et al.*, 2008; Chauve *et al.*, 2011; McConnell, 2004).

To use the ANGES scripts into an ancestral genome map computation pipeline, ANGES contains a 'master script' that reads the input files (markers and species tree, as well as possible optional files) and a 'parameters file' that records all user-defined choices that can be made regarding data processing. A simple graphical interface is provided to generate, read and modify parameters files.

Running times can vary depending on the nature of data and the ancestral map computation methods. The examples provided with the distribution have all completed within a few minutes on a bi-processor desktop computer. Larger datasets we analyzed, especially with non-unique markers or using the branch-and-bound method, required up to 2 days of computation.

4 FUTURE WORK

In the next release, we will include alternative methods to compute ancestral genome maps from ACS, and in particular an extension of the greedy heuristic for the sandwich-C1P framework. We will also include methods to compute ancestral genome structures that do not rely on the notion of contiguity but on the less rigid notion of conserved synteny, as described in Ouangraoua *et al.* (2011). And we are working on allowing multichromosomal genomes with circular and linear chromosomes, to be able to handle in a single framework all known genome forms in the living world. We are also integrating the possibility to provide gene families trees as input.

Funding: C.C. was funded by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (249834-2011). B.R.J. was funded by a Simon Fraser University Undergraduate Summer Research Award. A.R. was funded by a Simon Fraser University Graduate Fellowship. E.T. was funded by the Agence Nationale pour la Recherche grant ANR-10-BINF-01-01 Ancestrome.

Conflict of Interest: none declared.

REFERENCES

Atkins,J.E. *et al.* (1998) A spectral algorithm for seriation and the consecutive ones problem. *SIAM J. Comput.*, **28**, 297–310.

Bergeron,A. *et al.* (2008) Computing common intervals of K permutations, with applications to modular decomposition of graphs. *SIAM J. Discrete Math*, **22**, 1022–1039.

Chauve,C. *et al.* (2010) Yeast ancestral genome reconstructions: the possibilities of computational methods II. *J. Comput. Biol.*, **17**, 1097–1112.

Chauve,C. *et al.* (2011) Tractability results for the consecutive-ones property with multiplicity. In Giancarlo,R. and Manzini,G. (eds), *Combinatorial Pattern Matching–22nd Annual Symposium, CPM 2011, Lecture Notes Comput Sci.* Vol. 6661, Springer, Berlin, pp. 90–103.

Chauve,C. and Tannier,E. (2008) A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS Comput. Biol.*, **4**, e1000234.

Gavranovic,H. *et al.* (2010) Mapping ancestral genomes with massive gene loss: a matrix sandwich problem. *Bioinformatics*, **27**, i257–i265.

Ma,J. *et al.* (2006) Reconstructing contiguous regions of an ancestral genome. *Genome Res.*, **16**, 1557–1565.

McConnell,R. (2004) A certifying algorithm for the consecutive ones property. In Munro,J.I. (ed.), *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2004*. SIAM, Philadelphia, pp. 768–777.

Muffato,M. *et al.* (2010) Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*, **26**, 1119–1121.

Ouangraoua,A. *et al.* (2011) Reconstructing the architecture of the ancestral amniote genome. *Bioinformatics*, **27**, 2664–2671.

Pennisi,E. (2005) Extinct genome under construction. *Science*, **308**, 1401–1402.

Schmidt,T. and Stoye,J. (2004) Quadratic time algorithms for finding common intervals in two and more sequences. In Sahinalp,S.C., Mutukrishnan,S. and Dogrusöz,U. (eds), *Combinatorial Pattern Matching, 15th Annual Symposium, CPM 2004, Lecture Notes Comput Sci.* Vol. 3109, Springer, Berlin, pp. 347–358.

Zheng,C. and Sankoff,D. (2012) Gene order in Rosid phylogeny, inferred from pairwise syntenies among extant genomes. *BMC Bioinformatics*, Suppl 10, S9.