Molecular Phylogenetics and Evolution 101 (2016) 46–55

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

On simulated annealing phase transitions in phylogeny reconstruction

Maximilian A.R. Strobl [a,b,1], Daniel Barker [a,*]

[a] School of Biology, University of St Andrews, St Andrews, Fife KY16 9TH, UK
[b] School of Mathematics and Statistics, Mathematical Institute, North Haugh, St Andrews, Fife KY16 9SS, UK

ARTICLE INFO

Article history:
Received 15 December 2015
Revised 29 April 2016
Accepted 1 May 2016
Available online 2 May 2016

Keywords:
Phylogeny
Optimisation
Heuristic methods
Simulated annealing
Phase transition
Search landscape

ABSTRACT

Phylogeny reconstruction with global criteria is NP-complete or NP-hard, hence in general requires a heuristic search. We investigate the powerful, physically inspired, general-purpose heuristic simulated annealing, applied to phylogeny reconstruction. Simulated annealing mimics the physical process of annealing, where a liquid is gently cooled to form a crystal. During the search, periods of elevated specific heat occur, analogous to physical phase transitions. These simulated annealing phase transitions play a crucial role in the outcome of the search. Nevertheless, they have received comparably little attention, for phylogeny or other optimisation problems. We analyse simulated annealing phase transitions during searches for the optimal phylogenetic tree for 34 real-world multiple alignments. In the same way in which melting temperatures differ between materials, we observe distinct specific heat profiles for each input file. We propose this reflects differences in the search landscape and can serve as a measure for problem difficulty and for suitability of the algorithm's parameters. We discuss application in algorithmic optimisation and as a diagnostic to assess parameterisation before computationally costly, large phylogeny reconstructions are launched. Whilst the focus here lies on phylogeny reconstruction under maximum parsimony, it is plausible that our results are more widely applicable to optimisation procedures in science and industry.

1. Introduction

Global optimisation is an important step in modern phylogeny reconstruction. To identify the most plausible phylogenetic tree under a given criterion, one has to find the tree with optimal fit amongst all conceivable tree topologies. The maximum parsimony criterion (MP) requires optimisation of tree length (Fitch, 1971) and maximum likelihood (ML), as the name suggests, requires optimisation of a likelihood function (Felsenstein, 1981). However, MP is an NP-Complete problem (Foulds and Graham, 1982) and ML phylogeny reconstruction is NP-Hard (Roch, 2006). It follows that in order to be sure of obtaining the optimal tree, one would theoretically need to examine all feasible topologies; a number which grows factorially with the number of taxonomic units in the analysis (Felsenstein, 1978a). Even for studies of moderate size, this number exceeds the number of atoms in the universe and makes an exhaustive search infeasible. Despite reductions in complexity achieved by the branch-and-bound algorithm (Hendy and Penny, 1982) and revolutions in computing power, it is likely that run time for exact optimisation will remain astronomical.

As a workaround, heuristic optimisation algorithms are used. Instead of setting out to determine all globally optimal solutions, these methods use shortcuts to aim to find optima approximately. As such, heuristics allow one to obtain very good solutions in practical time scales when exact optimisation is too costly. This is also true in phylogeny reconstruction. Heuristic searches underlie the majority of today's MP and ML phylogeny reconstruction programs, for example PAUP, TNT, PhyML and RAxML (Swofford, 2003; Goloboff et al., 2008; Guindon et al., 2010; Stamatakis, 2014). However, in employing heuristic searches we fundamentally lose the guarantee of global optimality. Using a heuristic search it remains possible that even if the true phylogeny was hypothetically inferable from an alignment, one might not be able to retrieve it – simply because the algorithm did not happen to search a specific area of the search space. This problem is particularly pressing for analyses involving many taxonomic units.

The increasing ease and decreasing costs of DNA sequencing allow opportunities for very large phylogeny reconstructions. At the same time, phylogeny is being applied to even more areas of

[*] Corresponding author at: Institute of Evolutionary Biology, University of Edinburgh, Charlotte Auerbach Road, The Kings Buildings, Edinburgh EH9 3FL, UK.
E-mail address: Daniel.Barker@ed.ac.uk (D. Barker).

[1] Current address: Mathematical Institute, University of Oxford, Andrew Wiles Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford, Oxfordshire OX2 6GG, UK.

http://dx.doi.org/10.1016/j.ympev.2016.05.001
1055-7903/© 2016 The Authors. Published by Elsevier Inc.

the life sciences. Applications include, for example, ancestral state reconstruction (Lutzoni et al., 2001), assessing biodiversity (Flynn, 2011), predicting gene function (Eisen, 1998; Barker et al., 2007) and investigations of cancer and pathogen evolution, with implications for treatment (Gerlinger et al., 2012; Köser et al., 2012). Such analyses build on the solution of large and complex optimisation problems, making further development of heuristic searches in phylogeny increasingly urgent.

An important step towards more efficient algorithms is a better understanding of the nature of heuristic searches. In the current paper we present an analysis of the search behaviour of the simulated annealing heuristic. Simulated annealing, available in the phylogeny reconstruction packages LVB, MetaPIGA, SAMPARS and RAxML (Barker, 2004; Helaers and Milinkovitch, 2010; Richer et al., 2013; Stamatakis, 2014), is inspired by the physical processes occurring during the crystallisation of a liquid by gentle cooling (Kirkpatrick et al., 1983; Cerny, 1985). Convergence to the final solution is controlled by a parameter mimicking temperature in a physical system, with certain ranges of this parameter, called phase transitions, being particularly important to the search (Kirkpatrick et al., 1983; Cai and Ma, 2010; Hasegawa, 2012). In the following paper, we investigate the nature and role of these phase transitions during simulated annealing searches under the MP criterion. We identify the phase transitions for 34 real world phylogeny problems. We find that properties of phase transitions are repeatable for the same analysis and input and vary across different inputs, but in all cases correspond to the onset of effective resolution of the tree structure. Subsequently, we discuss how knowledge of the phase transition can help advance our understanding of the functioning of simulated annealing. We hypothesise phase transitions can serve as a useful diagnostic for finding suitable parameterisations for the search and be a stepping stone for future algorithmic improvements. Whilst in the current study we focus on phylogeny reconstruction under MP, conceptual links between MP and ML (see ‘Maximum Parsimony’, below) and the general nature of the simulated annealing algorithm make it plausible that our results are relevant for other areas of phylogeny reconstruction and beyond.

1.1. Maximum parsimony

For evolution of discrete traits on a given tree topology, the minimum number of changes consistent with the observed characters is known as tree length. Maximum parsimony seeks to find the topology of lowest length for the data matrix at hand.

Fitch (1971) provides a rapid, dynamic programming method to calculate the length for a given tree topology.

Because of its simplicity the most parsimonious tree problem provides a good model for studying global optimisation in phylogeny reconstruction and was therefore the chosen focus of this study. Calculations are significantly quicker than for the ML case and one avoids complications arising from needing to select appropriate models and parameters for the data at hand, which might confound the interpretation. This allows one to gain a fundamental understanding of the optimisation algorithm which in the future can then be extended to more complex optimality criteria.

Nevertheless, our immediate results are of general interest to the phylogeny community. Although MP is usually regarded as a distinct method, the MP tree is also the ML tree, at least under a constrained ‘no common mechanism’ model (Tuffley and Steel, 1997; Steel and Penny, 2000; see also Cavalli-Sforza and Edwards, 1967, pp. 239–240). For the model implied by MP the number of parameters increases with the amount of data, leading to statistical inconsistency (Felsenstein, 1978b; Yang, 2006, pp. 198–204). Simulation studies suggest the biological accuracy of MP is lower than that of statistically consistent ML approaches (e.g. Huelsenbeck, 1995). But even where one desires statistical consistency, the MP tree may still be useful as an initial tree, for further refinement by ML whose evaluation function is more time-consuming to calculate. This approach combines the speed of calculation of tree length (MP) for the initial part of the search, with a consistent approach to finding the final result (ML). This is an option in, for example, PhyML (Guindon et al., 2010).

Tree length varies with the input data, as well as with the optimality of the tree for those data. For greater comparability across different input files we used the tree consistency index (CI), with a theoretical range of 0–1 (Kluge and Farris, 1969), which seeks to normalise tree length and improve comparability. CI provides a measure of the amount of homoplasy on the proposed tree. For a given data matrix, a higher consistency index indicates reduced homoplasy, hence a shorter tree. Tree CI is given by:

$$\mathrm{CI} = \frac{K}{\text{Tree length}}, \tag{1}$$

where K is the sum of the minimum number of changes for each column in the multiple alignment individually (without reference to any tree structure). As such, K is a theoretical minimum of the number of mutations that has to have occurred to produce the given alignment from a single ancestor sequence. To cast the problem as a minimisation, which is more typical, we seek trees of minimum homoplasy index (HI), where HI = 1 – CI (Swofford, 1993).

1.2. Simulated annealing

The simulated annealing algorithm was independently developed by Kirkpatrick et al. (1983) and Cerny (1985). It is inspired by the processes which occur during the cooling of physical systems and is a simple but powerful optimisation technique. If a liquid is cooled slowly, the atoms anneal to form a crystal structure that minimises their energy. Since the particles in the liquid are in continuous motion at each instance they go through a plethora of positions and arrangements, some of which will be energetically more favourable than others. When cooling is applied and the system is given sufficient time at each temperature, the distribution of states visited will shift towards – and finally become – the very low energy crystalline state (van Laarhoven and Aarts, 1988).

Simulated annealing mimics this process: in the same way that the physical system seeks the state of minimal energy, it seeks the solution of minimal cost. The algorithm will start with an initial, often randomly generated solution X and perturb it according to some neighbourhood function N, to propose an updated solution X'. Similarly to how particles will always adopt a state that is an improvement over the current one, the algorithm will always move to X' if it has lower cost. However, also if X' has higher cost, in accordance with the physical analogy it will be occasionally accepted with probability:

$$P_{acc} = \exp(-\Delta H/T), \tag{2}$$

where ΔH is the change in cost and T is a control parameter playing the role of temperature. The algorithm will perform a certain number of such moves to allow the system to equilibrate. Then the temperature T is decreased according to a decrement rule known as the cooling schedule. The cooling schedule is often chosen as a geometric law of the form: $T_{n+1} = \alpha\, T_n$, $0 < \alpha < 1$ (e.g. Barker, 2004; Kirkpatrick et al., 1983). If a certain number of temperature decrements fails to bring about further improvements, the system is considered frozen and the search is terminated. Despite its relative simplicity, simulated annealing yields high quality solutions for a wide range of optimisation problems (e.g. van Laarhoven and Aarts, 1988; Ingber, 1993; Lindorff-Larsen et al., 2005; Alavi and Gandomi, 2011; Kolish and Dahlmann, 2015), including