PHYLOGENETICS: CHALLENGES AND CONJECTURES

These problems were posed during the 4-month program on Phylogenetics at the *Isaac Newton Institute for Mathematical Sciences*, September-December 2007, Cambridge UK.

List maintained by Mike Steel (m.steel@math.canterbury.ac.nz) This update: May 2010

1. Network and supertree challenges

[**NC1**] k–**level networks consistent with dense sets of rooted triples: SOLVED! - see below** Given a finite set X a *rooted triple* on X is a rooted binary phylogenetic tree on three leaves chosen from X. A set R of rooted triples on X is *dense* if, for every subset S of X of size 3, there is a tree in R that has leaf set S. A phylogenetic network N on X is *consistent with* a set R of rooted triples, if, for each $ab|c \in R$, N contains vertices $u \neq v$ and pairwise internally vertex-disjoint paths $u \to x, u \to y, v \to u$ and $v \to z$. Finally N is a k*–level network* if each biconnected component contains at most k recombination vertices.

Question Is there a polynomial-time algorithm that will determine, for any dense set of rooted triples R and integer $k \geq 1$, whether there is a k–level network N that is consistent with R?

Comments A polynomial-time algorithm is known for $k = 1$ (Jansson and Sung, 2006), and more recently for $k = 2$ (Iersel, Keijsper, Kelk and Stougie, 2007 *Constructing level-2 phylogenetic networks from triplets*, where this problem was posed). For full details see: arXiv:0707.2890v1 [q-bio.PE].

Update May 2010: A polynomial-time algorithm for any fixed k was provided by Thu-Hien To and Michel Habib, in their paper: Level-k Phylogenetic Networks Are Constructable from a Dense Triplet Set in Polynomial Time, in In *Combinatorial Pattern Matching*, LNCS 5577, pp, 275-288 (2009).
Steven Kelk and I recently found out that the problem is NP-hard when k is part of the input: http://arxiv.org/abs/1004.5332
The natural follow-up question is whether the problem is FPT when parametrized by the level k. Or, more generally, to find a 'practical' algorithm solving this problem.

[**NC2**] **Phylogenetic diversity for a 2-tree split system: SOLVED! (see below)**

Given a collection Σ of X–splits, a positive weighting function $w : \Sigma \to R^{>0}$, and a subset S of X, the *phylogenetic diversity* of S, $PD_{(\Sigma,w)}(S)$ is

$$PD_{(\Sigma,w)}(S) := \sum_{\{A|B \in \Sigma : A \cap S \neq \emptyset, B \cap S \neq \emptyset\}} w(A|B).$$

Question Is there a polynomial-time algorithm for the following problem: When a collection Σ of X–splits is the union of the sets of splits of two (unrooted) phylogenetic X–trees find a subset S of size k (variable) that maximizes $PD_{(\Sigma,w)}(S)$.

Comments The greedy algorithm provides an optimal solution in the case where the two trees agree (or more generally when one is a refinement of the other). In case Σ is a cyclic split system, a polynomial-time algorithm exists (Minh, Klaere and von Haeseler), however not all 2–tree split systems are cyclic. If Σ is the union of the sets of splits of 3 or more trees, the problem is NP-complete (Spillner, Nguyen and Moulton, 2007).

Update May 2010: Problem [NC2] has just been SOLVED! see Bordewich, M., Semple, C., and Spillner, A. (2009). 'Optimizing phylogenetic diversity across two trees', *Applied Mathematics Letters*, 22, 638-641.]

[**NC3**] **How many species must two trees agree on?**

Let $B(n)$ denote the set of unrooted binary phylogenetic trees with leaf set $\{1, 2, \ldots, n\}$. Define

$$f(n) = \min_{T,T' \in B(n)} \max\{|S| : S \subseteq X, T|S = T'|S\}.$$

It is fairly easy to show that $f(n)$ grows at a rate of between $\sqrt{\log(\log(n))}$ and $\log(n)$.

Question Determine the (asymptotic) rate of growth of $f(n)$.

Update May 2010: The lower bound on $f(n)$ of $\sqrt{\log(\log(n))}$ has been increased to $\log(\log(n))$ in Szekely, L. and Steel, M. (2009). 'An improved bound on the Maximum Agreement Subtree problem', *Applied Mathematics Letters* 22: 1778-1780.

[**NC4**] **Computing the full closure of a general set of rooted triples**

Let R be a set of rooted triples (binary trees on three leaves). We say that R is *closed* if, for every compatible subset S of R, any rooted triple that is displayed by all trees that display S is contained in R. Define the *closure* of R to be the intersection of all closed sets containing R (this is well-defined).

Question Is there a polynomial-time algorithm for determining the closure of R?

Comments In case R is compatible, the answer is easily seen to be 'yes' by exploiting the BUILD algorithm of Aho et al. (1981). Thus the problem is of interest in case R is incompatible.

[**NC5**] **Groves of phylogenetic trees: SOLVED! (see below)** In the paper http://www.stat.wisc.edu/Department/techreports/tr1123.pdf the authors introduced the concept of a 'Grove' of phylogenetic trees, but they left open a fundamental conjecture as to whether maximal groves always partition a collection of taxon sets.

Question If two groves intersect, is their union a grove? (see the paper above for details).

Update May 2010: Mareike Fischer (Vienna) has shown that the answer to this question is 'no' by way of an explicit counterexample.

2. Parsimony challenges

[**PC1**] **Short sequences specifying an MP tree: SOLVED! (see below)**
Is the following conjecture true?

Conjecture There exists a constant $c > 0$ such that, for any fully-resolved phylogenetic tree $\mathcal{T}$, there exists a sequence of at most $\lfloor c \cdot \log(n(\mathcal{T}))\rfloor$ binary characters on X that has $\mathcal{T}$ as its unique maximum parsimony tree, where $n(\mathcal{T})$ is the number of leaves of $\mathcal{T}$.

Comments This conjecture was posed by M. Steel in 2005 and comes with a $100 prize. It would follow as a direct consequence of an affirmative solution to the next problem.

Update May 2010: Juanjuan Chai and Elizabeth A. Housworth (Indiana University) appear to have proved this conjecture in their paper currently under review: 'On the Number of Binary Characters Needed to Recover a Phylogeny Using Maximum Parsimony'.

[**PC2**] **Sequence length requirements for MP**
Consider sequences generated i.i.d. by the Poisson process (eg. symmetric 2-state, or Jukes-Cantor) on fully-resolved phylogenetic trees. Recall that the *branch length* for an edge e is the expected number of substitutions on that edge (given by $-\frac{1}{2}\log(1-2p(e))$ where $p(e)$ is the probability of a net substitution between the endpoints of e.

Question Is the following true: For any $\epsilon > 0$, and value $K \geq 1$ there exist constants a, b with $0 < a < b < \frac{1}{2}$ and $C > 0$ so that for any fully resolved phylogenetic tree (on any number n of leaves) provided that the branch lengths all lie between a and b and the ratio of any two branch lengths is at most K then MP will correctly return the underlying tree with probability at least $1-\epsilon$ whenever the sequence length is at least $C \cdot \log(n)/a^2$.

Comments This conjecture is (essentially) equivalent to one stated by Vic Albert in the introductory chapter of his recent book; accordingly (in addition to the bottle of wine for any of these problems), Vic has offered a bottle of Braastad Chteau de Triac (Réserve de la Famille) cognac for a proof of the full version of this conjecture. The special case $K = 1$ (i.e. all branch lengths equal) is perhaps easier, and if the claim is true in that case it would imply the validity of [PC1]. However, even in the special case $K = 1$ it is not known whether MP is statistically consistent when b is less than a value close to 0.1 (it is inconsistent on some trees when $p(e) = p > 1/8$).

[**PC3**] **MP trees for 2-tree data: SOLVED! – see below**
Suppose we take two fully-resolved phylogenetic X–trees, encode their splits by a sequence of binary characters, and concatenate these two sequences.

Question What can be said about the maximum parsimony tree(s) of the resulting data set?

Comments Several questions arise. For example, can a maximum parsimony tree be found in polynomial time? how many MP trees might there be? Under what conditions are the two initial trees the unique to MP trees for the data? Can one establish good bounds on the parsimony score of the MP tree? If these questions prove to be too difficult it may be useful to consider the more restricted case where one tree is obtained from the other by one (or a small number of) tree rearrangement operation(s) such as SPR.

Update May 2010:The main questions in [PC3] have been been SOLVED by Vincent Moulton and Stefan Grünwald in their paper 'Maximum parsimony for tree mixtures', *IEEE/ACM Computational Biology and Bioinformatics*, 6(1), 2009, 97-102.

[**PC4**] **Optimal tree refinement for parsimony: SOLVED - see below!**
Suppose we take have a phylogenetic X–tree, $\mathcal{T}$, which has maximum vertex degree d, and a sequence $F = (f_1, \ldots, f_k)$ of characters on X (we may suppose for convenience that these are binary characters). An *optimal tree refinement of $\mathcal{T}$ with respect to parsimony* (OTR-P) is a phylogenetic X–tree $\mathcal{T}'$ that contains all the splits present in $\mathcal{T}$ (i.e. which 'refines $\mathcal{T}$') and which has minimal parsimony score on F amongst all such refining trees.

Question Is there an algorithm for finding an OTR-P tree for the pair $(\mathcal{T}, F)$ whose complexity, for each fixed $d > 3$, is polynomial in n, k (where $n = |X|$)? If so, might there even be an algorithm with run time complexity polynomial(n, k)$\times f(d)$ for some (non-polynomial) function d.

Comments We may assume, without loss of generality, that any OTR-P tree is binary. (Bonet et al. (J. Comp. Biol., 5(3): 393-407, 1998) introduced this problem, and has some weaker results).

Update May 2010: A polynomial time algorithm for this problem has now been described in: Wu, T., Moulton, V., and Steel, M. (2009). 'Refining phylogenetic trees given additional data: Algorithms based on parsimony', *IEEE/ACM Transactions in computational biology and bioinformatics* 6(1): 118-125.

[**PC5**] **Hereditary MP trees: SOLVED! – see below**
Suppose we take have a phylogenetic X–tree, $\mathcal{T}$, which is a maximum parsimony tree for a sequence $F = (f_1, \ldots, f_k)$ of characters on X (one may consider the special case where these are binary characters).

Question Is the following conjecture true? For each subset $k \in \{4, \ldots, |X| - 1\}$ there exists a subset S of X of size k so that $\mathcal{T}|S$ is an MP tree for $F|S$ (the sequence of characters restricted to the taxa in S).

Comments This problem was posed by Arndt von Haeseler. It is true in the special case where the characters are homoplasy-free on $\mathcal{T}$.

Update May 2010:Mareike Fischer (Vienna) has found a counterexample to this conjecture.

3. Stochastic challenges

[**SC1**] **Explicit analysis of the star-tree paradox: SOLVED! – see below**
Suppose we generate k sites i.i.d. under the 2-state symmetric model of site substitution on the 3-taxon star tree with all three branches of fixed length t_1. Consider a prior distribution on the three resolved trees and their branch lengths that assigns equal probability (namely $\frac{1}{3}$) to the three trees, and that assigns independent (but non-identical) exponential priors to the node heights (i.e. the branch lengths are considered to satisfy a clock). Let P_i be the posterior probability of resolved tree T_i $(i = 1, 2, 3)$ from the sequences generated by the star tree.

Question Find an analytical expression for the limiting joint density $f(P_1, P_2, P_3)$ or the marginal density $f(P_1)$ as $k \to \infty$.

Comments See Ziheng Yang, 'Fair-Balance Paradox, Star-tree Paradox, and Bayesian Phylogenetics', *Molecular Biology and Evolution* 2007 24(8):1639-1655, and the references therein.

Update May 2010: Ed Susko has made major progress on understanding this limiting distribution in his paper: Susko, E. (2008). 'On the Distributions of Bootstrap Support and Posterior Distributions for a Star Tree', *Systematic Biology*, 57:602–612.

[**SC2**] **How much do two random trees have in common?**
Suppose we independently generate two fully-resolved phylogenetic X–trees $\mathcal{T}_1, \mathcal{T}_2$ under the uniform distribution (all such trees equally probably) or perhaps the Yule distribution. Consider the size of the largest subset S of X of X on which $\mathcal{T}_1$ and $\mathcal{T}_2$ agree (i.e. $\mathcal{T}_1|S = \mathcal{T}_2|S$), and the mean of this random variable $f(n) = \mathbb{E}[|S|]$.

Question Establish asymptotic lower bounds of the form $f(n) \geq cn^{\beta}$ for some $\beta > 0$.

Comments It is known that for the uniform distribution, $f(n)$ is $O(\sqrt{n})$ and so for the uniform distribution $\beta \leq \frac{1}{2}$. Simulations suggest that a value of β close to $\frac{1}{2}$ might be correct. For details see Bryant, McKenzie and Steel (2003). [http://www.math.canterbury.ac.nz/ m.steel/research/max-agree.pdf]

[**SC3**] **Admissibility of phylogenetic methods**
Consider sequences generated i.i.d. by the Poisson process (eg. symmetric 2-state, or Jukes-Cantor) on a phylogenetic tree $\mathcal{T}$. Given a method M for reconstructing phylogenetic trees from sequences let $P(M(X) = \mathcal{T}|\mathcal{T}, \lambda)$ denote the probability that M correctly returns the tree $\mathcal{T}$ when X is generated by the Markov process on $\mathcal{T}$ with branch lengths λ.

Following decision-theoretic terminology, we say that a method M is *inadmissible* if there exists another method M' for which

$$P(M'(X) = \mathcal{T}|\mathcal{T}, \lambda) \geq P(M(X) = \mathcal{T}|\mathcal{T}, \lambda)$$

for all fully resolved phylogenetic trees T and choice of (strictly positive, but finite) branch lengths λ, and for at least one such pair $(\mathcal{T}, \lambda)$ we have strict inequality.

Questions Is Maximum Parsimony is inadmissible? Is Maximum Likelihood is inadmissible?

Comments Note that ML is known to be admissible in case there is a fixed known value of λ for each tree, and ML is performed subject to this constraint.

[**SC4**] **Identifyability of the GTR+Gamma+I model: SOLVED - see below**

Recently Allman, Ane and Rhodes (August 2007) established that the GTR+Gamma model of site susbstitution has the property that the underlying tree (and associated parameters) can always be uniquely reconstructed from sufficiently long sequences (and in the process pointed out that a previous claimed proof by Rogers was incorrect). However this 'identifiability property' for the popular extension of the model – the GTR+Gamma+I model – is still open.

Question Determine whether or not the GTR+Gamma+I model satisfies the identifiability property.

Comments See arxiv:0709.0531v1.pdf. The identifiability question can also be asked for the closely related model in which the (continuous) Gamma distribution is replaced by a discretized Gamma distribution (plus invariable sites).

Update May 2010: The identifiability question has apparently been settled in this paper, Juanjuan Chai and Elizabeth A. Housworth. 2010. 'On Rogers's Proof of Identifiability for the GTR + Gamma + I Model', which is currently in revision at *Systematic Biology*.

[**SC5**] **Phylogenetic Invariants**

Under the general Markov model of site substitution on an n-taxon phylogenetic tree, the expected pattern frequencies satisfy certain polynomial equations known as *invariants*. Invariants for more restrictive models (e.g. the 2-state symmetric) have proved useful for solving constrained optimization formulations of the ML problem, and for finding examples of joint distributions on a mixture of two trees that agree with a joint distribution on a single tree.

- *Question* Find all phylogenetic invariants for the general Markov model of DNA site substitution on a 3-taxon tree.
- *Comments* If invariants for the 3-taxon tree were known completely, then invariants for any n-taxon tree could be explicitly constructed from these. Thus, the determination of phylogenetic invariants for the general Markov model for an n-taxon tree would be concluded. For a brief overview and information on a prize for its solution, see http://www.dms.uaf.edu/∼eallman/salmonPrize.pdf

Update May 2010: Scmuel Friedland gives a set-theoretic characterization (polynomials whose zero set is the right variety, i.e. a subset of 'all'). The link to the paper on the arXiv is: http://front.math.ucdavis.edu/1003.1968