Open Problems

Problem 1. In the *Shortest Superstring Problem* (SSP), we are given a set $S = \{s_1, \ldots, s_n\}$ of string over a finite alphabet Σ and we have to find a shortest string s that contains all the s_i as substrings. This problem is NP-hard, so assuming $P \neq NP$, the best one can hope for, in terms of polynomial algorithms, is approximation. A natural greedy procedure is the following: Pick two strings s_i, s_j with largest overlap from S (breaking ties arbitrarily) and replace them with their merge. The algorithm stops when there is only one string left and, obviously, this string is a superstring of all the s_i.

Considering the set of $\{ab^k, b^k c, b^{k+1}\}$, we can see that the Greedy Algorithm cannot have an approximation factor better than 2. However, resolving the following conjecture due to Blum et al. [2] is a long standing and very important open problem.

Conjecture 1. *The Greedy Algorithm has approximation factor 2.*

For more information and some other problems related to it see for example [10, 5].

Problem 2. Given any two rooted binary X-trees T_1 and T_2, we consider the root of both T_1, T_2 as a vertex ρ at the end of a pedant edge incident to the original root. Furthermore, ρ is also considered as part of the label set of T_1, T_2. We give now the definition of an agreement forest as reported in [4]:

Definition 1. *An agreement forest F of two rooted binary phylogenetic X-trees T, T' is a collection of trees $\{T_\rho, T_1, \ldots, T_k\}$ such that for all $i \in \{\rho, 1, \ldots, k\}$ T_i is a rooted subtree with vertex set L_i such that:*

- *The label sets $L_\rho, L_1, \ldots, L_k$ partition $X + \rho$ and in particular $\rho \in L_\rho$*
- *For all $i \in \{\rho, 1, \ldots, k\}, T_i = T|L_i = T'|L_i$*
- *The trees in $\{T(L_i) : i \in \{\rho, 1, \ldots, k\}\}$ and $\{T(L_i) : i \in \{\rho, 1, \ldots, k\}\}$ are vertex disjoint subtrees of T and T', respectively.*

The maximum agreement forest of two phylogenetic trees T, T', is an agreement forest having the minimum number of components. Finding the maximum agreement forest for two trees is NP-hard [4] and various 3-approximation algorithms have been proposed (see for example [11, 3, 13]). Very recently in [12] the authors proposed a 2-approximation algorithm for the MAF problem.

Is it possible to find a simpler 2-approximation algorithm or even better to improve the factor 2?

Problem 3. In the problem of *reconciling* phylogenetic trees of hosts and their symbionts we are given a host tree H, a symbiont tree P, a mapping of the leaves of P to the leaves of H (that reflects current knowledge on which existing symbiont inhabit which hosts) and a cost vector $\underline{c} =< c_c, c_d, c_s, c_l >$ defining the cost of each of the four types of events that we can recover (*i.e.* cospeciation, duplication, host-switch and loss). We can then obtain a parsimonious solution by minimising the total cost of the mapping. Additionally, if timing information is not available or may be insufficiently reliable to be used with enough confidence, the reconciliation problem is NP-hard (see for example [8, 6, 14]).

In [7] it is defined the *k-bounded-reconciliaiton problem*: Given $H, P, \phi, \underline{c}$ find the reconciliation of minimum cost for which all host-switches are at a distance at most k (note that k is not part of the input). Recall that for a reconciliation γ, the distance of a host-switch is defined as follows: For an edge (p, p') in the symbiont tree P, for which p is associated to a host-switch, let $h = \gamma(p)$ and $h = \gamma(p')$. The *distance* of this host-switch is given as the distance in number of edges of the unique path between h and h'. If k is not part of the input the question remains whether the problem can be solved in polyonmial time.

The problem can be solved trivially in polynomial time for $k = 2$ and remains open for any ≥ 3.

References

[1] Allen, B. L., Steel, M.: Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of combinatorics*, 5(1), 1–15 (2001).

[2] Blum, A., Jiang, T., Li, M., Tromp, J., and Yannakakis, M.. Linear approximation of shortest superstrings. 328–336, 1991.

[3] Bordewich, M., McCartin, C., Semple, S.: A 3-approximation algorithm for the subtree distance between phylogenies, *Journal of Discrete Algorithms*, 6(3):458-471 (2008).

[4] Bordewich, M., Semple, S.: On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics* 8, 409-423 (2005).

[5] Bastien Cazaux, Eric Rivals: Approximation of Greedy Algorithms for Max-ATSP, Maximal Compression, Maximal Cycle Cover, and Shortest Cyclic Cover of Strings. *Stringology* 2014: 148-161.

[6] Conow C, Fielder D, Ovadia Y, Libeskind-Hadas R: Jane: a new tool for the cophylogeny reconstruction problem. *Algorithm. Mol. Biol.* 2010, 5:16.

[7] B. Donati, C. Baudet, B. Sinaimeri, P. Crescenzi and M-F. Sagot, EUCALYPT: Efficient tree reconciliation enumerator, *Journal of Algorithms for Molecular Biology* (AMB) 10(3), 2015.

[8] Doyon JP, Scornavacca C, Gorbunov KY, Szollosi GJ, Ranwez V, Berry V: An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In Proceedings of the 8th annual RECOMB Satellite Workshop on Comparative Genomics (*RECOMB-CG 2010*), Volume 6398 of Lecture Notes in Bioinformatics. 2011.

[9] Hein, J., Jiang, T., Wang, L., Zhang, K.: On the complexity of comparing evolutionary trees, *Discrete Applied Mathematics*, vol 71, Issues 13, (1996), 153–169.

[10] Haim Kaplan, Nira Shafrir, The greedy algorithm for shortest superstrings, *Information Processing Letters*, Volume 93, Issue 1, 2005, Pages 13-17,

[11] Estela Maris Rodrigues, Marie-France Sagot, Yoshiko Wakabayashi: The maximum agreement forest problem: Approximation algorithms and computational experiments. *Theor. Comput. Sci. 374(1-3): 91-110 (2007)*

[12] Frans Schalekamp, Anke van Zuylen, Suzanne van der Ster: A Duality Based 2-Approximation Algorithm for Maximum Agreement Forest. ICALP *2016: 70:1-70:14*

[13] Whidden, C., Zeh, N.: A unifying view on approximation and FPT of agreement forests. In: WABI 2009*. LNCS, vol. 5724, pp. 390401. Springer-Verlag (2009).*

[14] Tofigh A, Hallett M, Lagergren J: Simultaneous Identification of Duplications and Lateral Gene Transfers. IEEE/ACM Trans. on Comput. Biol. Bioinf. *2011, 8(2):517–535.*