The Duplication-loss Problem: Linear-time Algorithms for NNI Local Search

Cheng-Wei Luo[1,*], Meng-Han Li[1], Hsiao-Fei Liu[1], and Kun-Mao Chao[1,2,3]

[1] Department of Computer Science and Information Engineering,
National Taiwan University,
Taipei 106, Taiwan, ROC
b89902079@ntu.edu.tw

[2] Graduate Institute of Biomedical Electronics and Bioinformatics,
National Taiwan University,
Taipei 106, Taiwan, ROC

[3] Graduate Institute of Networking and Multimedia,
National Taiwan University,
Taipei 106, Taiwan, ROC

Received 10 September 2010; Revised 5 November 2010; Accepted 10 December 2010

Abstract. Given a set of gene trees, the DUPLICATION-LOSS problem is to infer a comparable species tree minimizing the number of gene duplications and losses (the mutation cost). This problem has been shown to be NP-hard. A standard heuristic performs the stepwise local searches on the tree space until a local minimum is reached. In this paper, we study the heuristic for the DUPLICATION-LOSS problem based on NNI local searches and propose a linear-time algorithm for the NNI LOCAL SEARCH problem under the mutation cost. Bansal et al. presented a near-linear time algorithm for the NNI LOCAL SEARCH problem under the duplication cost, and we also improve the result of Bansal et al. with a linear-time algorithm.

Keywords: Computational phylogenetics, gene duplication, local search, NNI, linear-time algorithm.

References

[1] J. E. Neigel and J. C. Avise, "Phylogenetic Relationship of Mitochondrial DNA under Various Demographic Models of Speciation," *Evolutionary Processes and Theory*, Academic Press, Orlando, FL, pp. 515-534, 1986.

[2] P. Pamilo and M. Nei, "Relationships between Gene Trees and Species Trees," *Molecular Biology and Evolution*, Vol. 5, No. 5, pp. 568-583, 1988.

[3] N. Takahata, "Gene Genealogy in Three Related Populations: Consistency Probability between Gene and Population Trees," *Genetics*, Vol. 122, No. 4, pp. 957-966, 1989.

[4] C. I. Wu, "Inferences of Species Phylogeny in Relation to Segregation of Ancient Polymorphisms," *Genetics*, Vol. 127, No. 2, pp. 429-435, 1991.

[5] M. Goodman, J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, G. Matsuda, "Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences," *Systematic Zoology*, Vol. 28, No. 2, pp. 132-163, 1979.

[6] M. S. Bansal and O. Eulenstein, "The Multiple Gene Duplication Problem Revisited," *Bioinformatics*, Vol. 24, No. 13, pp. i132-i138, 2008.

* Correspondence author

[7] P. Bonizzoni, G. D. Vedova, R. Dondi, "Reconciling a Gene Tree to a Species Tree under the Duplication Cost Model," *Theoretical Computer Science*, Vol. 347, No. 1-2, pp. 36-53, 2005.

[8] J. G. Burleigh, M. S. Bansal, A. Wehe, O. Eulenstein, "Locating Large-scale Gene Duplication Events through Reconciled Trees: Implications for Identifying Ancient Polyploidy Events in Plants," *Journal of Computational Biology*, Vol. 16, No. 8, pp. 1071-1083, 2009.

[9] K. Chen, D. Durand, M. Farach-Colton, "NOTUNG: A Program for Dating Gene Duplications and Optimizing Gene Family Trees," *Journal of Computational Biology*, Vol. 7, No. 3-4, pp. 429-447, 2000.

[10] P. G´orecki and J. Tiuryn, "On the Structure of Reconciliations," in *Proceedings of the 2nd RECOMB Comparative Genomics Satellite Workshop*, Bertinoro (Forli), Italy, pp. 42-54, 2004.

[11] R. Guig´o, I. Muchnik, T. F. Smith, "Reconstruction of Ancient Molecular Phylogeny," *Molecular Phylogenetics and Evolution*, Vol. 6, No. 2, pp. 189-213, 1996.

[12] C.W. Luo, M.C. Chen, Y.C. Chen, R.W.L. Yang, H.F. Liu, K.M. Chao, "Linear-time Algorithms for the Multiple Gene Duplication Problems," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 8, No. 1, pp. 260-265, 2011.

[13] B. Mirkin, I. Muchnik, T. F. Smith, "A Biologically Consistent Model for Comparing Molecular Phylogenies," *Journal of Computational Biology*, Vol. 2, No. 4, pp. 493-507, 1995.

[14] R. D. M. Page, "Maps between Trees and Cladistic Analysis of Historical Associations among Genes, Organisms, and Areas," *Systematic Biology*, Vol. 43, No.1, pp. 58-77, 1994.

[15] L. Zhang, "On a Mirkin-Muchnik-Smith Conjecture for Comparing Molecular Phylogenies," *Journal of Computational Biology*, Vol. 4, No. 2, pp. 177-187, 1997.

[16] B. Ma, M. Li, L. Zhang, "From Gene Trees to Species Trees," *SIAM Journal on Computing*, Vol. 30, No. 3, pp. 729-752, 2000.

[17] R. D. M. Page, "GeneTree: Comparing Gene and Species Phylogenies Using Reconciled Trees," *Bioinformatics*, Vol. 14, No. 9, pp. 819-820, 1998.

[18] B. L. Allen and M. Steel, "Subtree Transfer Operations and Their Induced Metrics on Evolutionary Trees," *Annals of Combinatorics*, Vol. 5, No. 1, pp. 1-15, 2001.

[19] M. S. Bansal, O. Eulenstein, A. Wehe, "The Gene-duplication Problem: Near-linear Time Algorithms for NNI-based Local Searches," *IEEE/ACM Trans. on Computational Biology and Bioinformatics*, Vol. 6, No. 2, pp. 221-231, 2009.

[20] M. Bordewich and C. Semple, "On the Computational Complexity of the Rooted Subtree Prune and Regraft Distance,"*Annals of Combinatorics*, Vol. 8, No. 4, pp. 409-423, 2004.

[21] M. S. Bansal, J. G. Burleigh, O. Eulenstein, "Efficient Genome-scale Phylogenetic Analysis under the Duplication-loss and Deep Coalescence Cost Models," *BMC Bioinformatics*, 11(Suppl 1):S42, 2010.

[22] M. S. Bansal, J. G. Burleigh, O. Eulenstein, A. Wehe, "Heuristics for the Gene-duplication Problem: A $\theta(n)$ Speed-up for the Local Search," in *Proceedings of the 11th Annual International Conference on Research in Computational Molecular Biology*, Oakland, CA, USA, pp. 238-252, 2007.

[23] D. L. Swofford, G. J. Olsen, P. J. Waddel, D. M. Hillis, *Phylogenetic Inference*, Molecular Systematics, Sinauer Associates, pp. 407-509, 1996.

[24] M. S. Bansal and O. Eulenstein, "An $\Omega(n2/\log n)$ Speed-up of TBR Heuristics for the Gene-duplication Problem," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 5, No. 4, pp. 514-524, 2008.

[25] D. Chen, O. Eulenstein, D. Fern´andez-Baca, J. G. Burleigh, "Improved Heuristics for Minimum-flip Supertree Construction," *Evolutionary Bioinformatics*, Vol. 2, pp. 347-356, 2006.

[26] M. A. Bender and M. Farach-Colton, "The LCA Problem Revisited," in *Proceedings of the 4th Latin American Symposium on Theoretical Informatics*, Punta del Este, Uruguay, pp. 88-94, 2000.