BIOINFORMATICS

Vol. 00 no. 00 2007
Pages 1–2

Phylocomposer and Phylodirector: Analysis and Visualization of Transducer Indel Models

Ian Holmes

Department of Bioengineering, University of California, Berkeley CA, USA.

Summary: Finite-state string transducers are probabilistic tools similar to Hidden Markov Models that can be systematically extended to large numbers of sequences related by indel and substitution processes on phylogenetic trees. The number of states in such models grows exponentially with the number of nodes in the tree, with the consequence that even quite small trees can be difficult to analyze or visualize. Here, we present two tools, `phylocomposer` and `phylodirector`, for working with string transducers. The former tool implements previously described *composition* algorithms for extending transducers to arbitrary tree topologies, while the latter generates short animations for arbitrary input alignments and phylogenetic trees, illustrating the state path through the composed transducer.

Availability: `phylocomposer` and `phylodirector` are freely available at `http://biowiki.org/PhyloComposer` and `http://biowiki.org/PhyloDirector`

Contact: `ihh@berkeley.edu`

String transducers are finite-state automata with input and output tapes. They are closely related to Hidden Markov Models (HMMs) and indeed may be regarded as conditionally-normalized Pair HMMs. Given a string transducer and a phylogenetic guide tree, one can systematically derive a multi-sequence HMM that correctly handles nested indels and neighbor-dependent substitutions (Holmes, 2003).

The algorithm to do this, known as *transducer composition*, has considerable application throughout bioinformatics. These applications include "statistical alignment" (Holmes and Bruno, 2001), comparative genome annotation (Lunter *et al.*, 2006), phylogeny (Lunter *et al.*, 2005) and measurement of evolutionary rates (Holmes, 2005).

Figure 1 and Table 1 illustrate why the composition problem is nontrivial, prompting automation: the number of states in the composed machine increases exponentially with number of nodes in the tree. Even though the single-branch transducer (A) corresponds to the simplest possible indel model (Thorne *et al.*, 1991), the largest composed transducer (E) has 620 transitions.

These data were generated with `phylocomposer`. The program takes as input a file describing a phylogenetic tree, one or more transducers (represented as Moore machines) & a mapping from branches to transducers. The file format is based on Lisp S-expressions (Rivest, 1997). The output is a description of the composed phylo-transducer, including the algebraic structure of transition & emission probabilities. In addition, the program generates dotfiles suitable for input to GraphViz, which was used to produce Figure 1. It can also be used for dynamic programming alignment and inference. The program is written in C++ and requires the GNU `gcc` and `make` utilities.

The `phylocomposer` distribution includes the examples of Figure 1 along with more realistic transducers, incorporating geometric and mixture-of-geometric length distributions (i.e. affine & general convex gap penalties).

Applications of the HMMs generated by the program include multiple alignment (including progressive (Higgins and Sharp, 1989), multidimensional (Sankoff and Cedergren, 1983) and MCMC (Holmes and Bruno, 2001)); simultaneous phylogeny and alignment (Lunter *et al.*, 2005); inference of ancestral "protosequences"; measurement of evolutionary rates; and analysis of indel models.

Figure 2 illustrates the output of `phylodirector`. The program takes as input a Stockholm-format alignment, including a Newick-format phylogenetic tree. The output is an MPEG-format animation illustrating the state path through the composed phylo-transducer required to generate the tree, along with a still image for each column of the alignment. The program requires Perl, the GD graphics library (and GD.pm Perl module), and the Berkeley MPEG Encoder.

Animations produced by `phylodirector` are available at `biowiki.org/PhyloFilm`.

ACKNOWLEDGMENTS

IH was funded in part by NIH/NHGRI grant 1R01GM076705-01.

REFERENCES

Higgins, D. G. and Sharp, P. M. (1989). Fast and sensitive multiple sequence alignments on a microcomputer. *Computer Applications in the Biosciences*, **5**, 151–153.

Holmes, I. (2003). Using guide trees to construct multiple-sequence evolutionary HMMs. *Bioinformatics*, **19 Suppl. 1**, i147–157.

Holmes, I. (2005). Using evolutionary Expectation Maximization to estimate indel rates. *Bioinformatics*, **21**(10), 2294–2300.

Holmes, I. and Bruno, W. J. (2001). Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, **17**(9), 803–820.

Lunter, G., Miklos, I., Drummond, A., Jensen, J. L., and Hein, J. (2005). Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, **6**, 83.

Lunter, G., Ponting, C. P., and Hein, J. (2006). Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Computational Biology*, **2**(1).

Rivest, R. (1997). S-expressions. Internet Working Draft. http://theory.lcs.mit.edu/~rivest/sexp.txt.

Sankoff, D. and Cedergren, R. J. (1983). Simultaneous comparison of three or more sequences related by a tree. In D. Sankoff and J. B. Kruskal, editors, *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*, chapter 9, pages 253–264. Addison-Wesley, Reading, MA.

Thorne, J. L., Kishino, H., and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, **33**, 114–124.

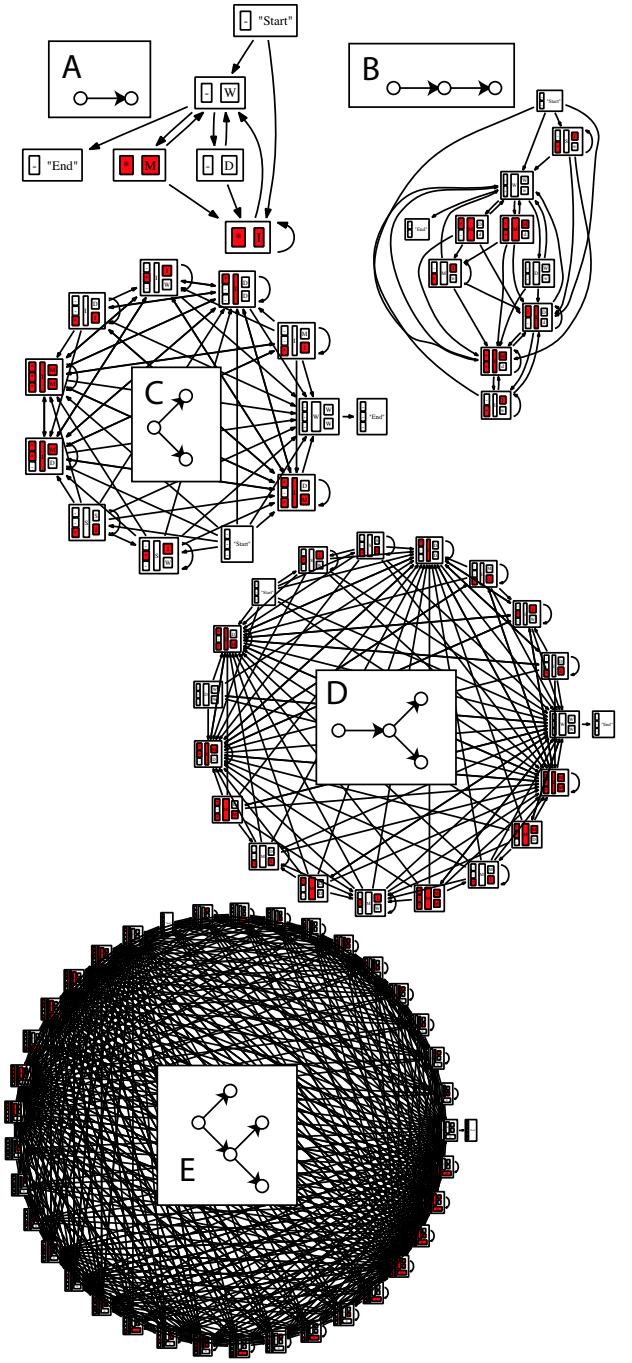

Fig. 1. Composed phylo-transducers for the TKF91 model (Thorne *et al.*, 1991), generated using `phylocomposer` and GraphViz. The basic transducer (A), which is self-similar when composed in series (B), can be used to infer common ancestors (C) or sample over alignments (D) or phylogenies (E).

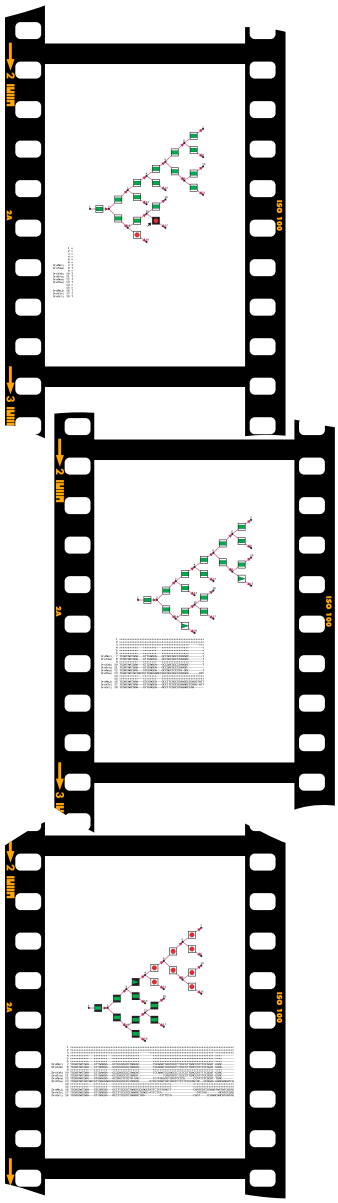

Fig. 2. Three frames from a `phylodirector` movie, showing alignments of twelve *Drosophila* species to Flybase gene CG31973.

Tree	Branches	States	Transitions
A	1	6	11
B	2 (series)	11	39
C	2 (parallel)	12	69
D	3	20	127
E	4	38	620

Table 1. State spaces of composed TKF91 transducers. Refer to Figure 1 for tree topologies.