Phylogenetics

Transducers: an emerging probabilistic framework for modeling indels on trees

Robert K. Bradley¹ and Ian Holmes²,*

¹Department of Physics and ²Department of Bioengineering, University of California, Berkeley CA, USA

Received on March 22, 2007; revised on August 1, 2007; accepted on August 3, 2007

Advance Access publication September 5, 2007

Associate Editor: Alex Bateman

Contact: ihh@berkeley.edu

1 INTRODUCTION

When it comes to dealing with indels, molecular evolution lags heuristic bioinformatics by decades. Sophisticated alignment algorithms have been widely known since the 1960s (and in bioinformatics since 1970), but we are still struggling to understand the corresponding phylogenetic models. Big ideas drive change: as we dream of reconstructing ancestral genotypes, it is ever clearer that indels cannot be ignored. We need to develop a robust understanding of probabilistic indel analysis and its relationship to alignment.

We believe that a suitable foundation for such analysis already exists, where evolutionary models meet automata theory: the framework of finite-state transducers. This framework links Hidden Markov Models (Brown et al., 1993; Churchill, 1992), sequence alignment algorithms (Gotoh, 1982; Miller and Myers, 1988; Needleman and Wunsch, 1970; Smith and Waterman, 1981), finite-state machines and Chomsky grammars (Durbin et al., 1998) and molecular phylogenetics (Miklós et al., 2004; Thorne et al., 1991). In this letter we outline this framework, also describing a preliminary analysis of one recent algorithm—Indelign—for reconstructing ancestral indel histories (Kim and Sinha, 2007).

Below, we briefly review the theory of transducers, concentrating not on the details of individual algorithms but rather on their unifying qualitative character. We show that Indelign, which reconstructs maximum-likelihood indel histories, is implicitly based on a transducer model. Thus, we can compare their unifying qualitative character. We show that Indelign, treating not on the details of individual algorithms but rather on annotation or parameter estimation. Transducers are natural models for indels on trees, just as continuous-time Markov chains are natural models for substitutions.

Although transducers have been known in the computer science literature since the 1950s (Mealy, 1955; Mohri et al., 2000), they have been applied in bioinformatics only lately (Holmes, 2003; Searls and Murphy, 1995). In fact, early probabilistic alignment algorithms share similarities to transducers (Bishop and Thompson, 1986), as (of course) do Pair HMMs (Holmes and Durbin, 1998). The breakthrough came

1.1 Theory of finite-state transducers

A transducer is a finite-state machine with an input tape (X), an output tape (Y), a symbol alphabet and a set of transition and (possibly) emission weights. It is therefore very similar to a Pair HMM (Hidden Markov model), which is also a two-tape finite-state machine with transition and emission weights (Durbin et al., 1998). As with a Pair HMM, each transducer state may be classified as Match, Insert, Delete, Start or End. In both cases, a path π through the machine corresponds to a pairwise sequence alignment with an associated likelihood, defined to be the product of transition and emission weights along the path. Pair HMMs and transducers have similar sets of algorithms for inference, including the Forward, Backward and Viterbi algorithms (Durbin et al., 1998).

The crucial difference is that the Pair HMM’s tapes are both considered to be outputs, whereas the transducer has one input and one output. The probabilistic interpretation is that the path probability for a Pair HMM is the joint likelihood P(π, X, Y), while for a transducer it is the conditional likelihood P(π, Y|X). Conceptually, a transducer represents the operation of a finite span of evolutionary time (ΔT), ‘evolving’ the input sequence into the output sequence by introducing substitutions and indels at random. We can represent this operation as X ΔT Y.

The feature of transducers that makes them so useful for comparative sequence analysis is the existence of algorithms for composing them in series or in parallel (Holmes, 2003; Mohri et al., 2000), where a series composition represents the consecutive operation of two transducers (X ΔT Y ΔT Z) and a parallel composition represents a bifurcation in a phylogenetic tree (X ΔT Y and X ΔT Z). By placing a transducer on each branch of a phylogenetic tree, we can automate the construction of systematic scoring schemes and algorithms for alignment, annotation or parameter estimation. Transducers are natural models for indels on trees, just as continuous-time Markov chains are natural models for substitutions.
in the field of statistical alignment (a term coined by Jotun Hein),
which attempts to unite bioinformatics and molecular evolution
via explicit birth-death models for indels and other events.
In pioneering work, the TKF91 model of Thorne et al. (1991)
was used to derive alignment algorithms with linear gap
penalties; these were then extended to multiple sequences on a
tree (Hein, 2001), recognized as examples of HMM algorithms
(Holmes and Bruno, 2001) and formulated using transducers
(Holmes, 2003).

Although the linear gap penalty of TKF91 is occasionally
quoted as a drawback of statistical alignment, this is a
misconception of the role of TKF91. Several transducers with
affine gap penalties have been derived from evolutionary
models (Knudsen and Miyamoto, 2003; Miklós et al., 2004).
TKF91’s role can be seen as a well-studied and canonical (albeit
simple) example, which can be used to illustrate nearly all the
relevant kinds of algorithm, such as HMM state pruning
(Lunter et al., 2003), Expectation Maximization (Holmes,
2005b) or alignment (Lunter et al., 2004).

Transducers provide a convenient bridge between rigorous
phylogenetic analysis of indel processes and the rich lore of
finite-state machine design. Many empirically observed char-
acteristics of genome evolution can be integrated with
transducers: they provide a systematic framework for analyzing
mutation rates, including variations in GC content, fluctuating
local conservation, methylation rate and codon substitution
patterns (Kosiol et al., 2007), and for modeling phenomena
involving indels, including probability distributions over exon
and intron length, stop codon avoidance and conservation of
codon reading frame (Kellis et al., 2003).

Further, transducers are not limited to models where the
indel and substitution processes are independent. Extensions
beyond HMM-like models allow transducers to, in principle,
model microsatellite expansion/contraction, transposon inser-
tion/deletion, local micro-duplications and micro-inversions,
and various other mutation processes that would otherwise
difficult to analyze mathematically.

Formal extensions to string transducers allow them to model
RNA and gene structure. Related machines, called tree trans-
ducers by linguists, are analogous to Pair Stochastic Context-
Free Grammars and are used to analyze RNA sequences
(Bradley and Holmes, 2007; Holmes, 2005a; Sakakibara, 2003).

Further discussion of transducers, including links to anima-
tions, may be found on our wiki at the following URL: http://
biowiki.org/StringTransducers

2 INDEL HISTORIES

We now turn to the evolutionary model for indels described in
Kim and Sinha (2007) and the associated algorithms. It can
immediately be seen that the Indelign model is a transducer:
conditionally normalizing the probabilities of Indelign’s Pair
HMM gives just the probability \( P(\pi, Y|X) \) associated with the
evolution \( X^{\otimes 2} Y \).

Indelign’s ANNOTATE algorithm returns the maximum-
likelihood indel history given a multiple alignment of \( n \) observed
sequences. It operates on ‘blocks’, defined as the spans of
maximal ungapped stretches of observed sequence, and
computes the maximum-likelihood indel history of sets of
consecutive dependent blocks by labeling each block as gap or
non-gap for all ancestral sequences. If there are \( k \) such
conditionally dependent blocks, then each node has a labeling in \( \{\ast, \ast\}^k \). A dynamic-programming (DP) version of
the algorithm, which iterates over combined labelings of sets of
three nodes (two siblings and their parent), has a worst-case time
complexity of \( O(N^2k^3) \), where \( N \) is the number of nodes in the
phylogenetic tree. Note that \( k \) is theoretically bounded only by
the alignment length.

This enumeration of indel histories over blocks can be
contrasted with the state-enumeration approach typical of
transducer DP algorithms. Alignment to a composed transdu-
cer can be expressed as a one-dimensional DP problem over an
alphabet of strings in \( \{A, C, G, T, \ast\}^5 \), where each character in
the string corresponds to the residue or gap at a particular node
in the phylogenetic tree (with \( N \) nodes). The composed
transducer has \( O(a^N) \) states and hence at most \( O(a^{3N}) \)
transitions, where \( a \) is the number of states of a single
transducer. By analogy with standard path-inference to an
HMM, we can see that this state-enumeration approach has
time complexity \( O(N^3a^3) \), where \( L \) is the length of the multiple
alignment. For the simplest transducers \( a \approx 3 \), though it may be
possible to reduce this by redundant-state elimination.
In contrast to the above analysis of Indelign’s ANNOTATE
algorithm, this upper bound on the complexity is not input-
dependent. The term ‘phylo-HMM’ has been coined to describe
such phylogenetically-structured HMMs, particularly when the
multiple alignment is supplied as an external constraint.

The \( O(Na^{3N}) \) time complexity makes exact DP to a
composed transducer impractical for datasets of many
sequences, but Markov Chain Monte Carlo (MCMC)
approaches offer a principled alternative. Given a multiple
alignment of observed leaf sequences, we can sample exactly
from the posterior distribution over indel histories by starting
with some initial estimate of the indel history and modifying it
by successive local MCMC ‘moves’, e.g. the branch- and node-
sampling moves described in Holmes and Bruno (2001). These
moves involve sampling over only local transducer compositions
around the neighborhood of a single branch or node, allowing
us to avoid the computational cost of inference on the full phylo-
machine. Branch sampling has time complexity \( O(L^2a^2) \) and
node sampling \( O(La^3) \); much better than the \( O(Na^{3N}) \) cost of
exact inference.

The relative efficiency of Kim and Sinha’s enumerative
algorithm depends strongly on the dataset used. In Figures 1
and 2 we plot the distributions of \( k \)-values for two genomic
datasets, the 12 newly sequenced Drosophila genomes
(Drosophila Comparative Genome Sequencing and Analysis
Consortium, 2007) and data from the ENCODE project
(Margulies et al., 2007). Alignments were created with
MAVID (Bray and Pachter, 2004). Most blocks of the
Drosophila alignments belong to relatively short sequences of
conditionally dependent blocks and so are amenable to analysis
with Indelign, but the tail of the distribution stretches to
\( k \)-values of greater than \( 10^5 \). Multiple alignments of the highly
diverged genomes of the ENCODE project are dominated by
very high \( k \)-values of order \( 10^5 \). Indelign’s inference algorithm
grows exponentially in complexity with \( k \), making it likely
impractical for analysis of much of this data without further heuristics or constraints.

Dynamic programming to a composed transducer, on the other hand, can handle such datasets with a complexity that is (in the worst case) exponential in the number of tree nodes, but linear in alignment length. It is even possible to achieve sub-linear memory complexity with respect to alignment length, using recursive approaches (Hirschberg, 1975; Tarnas and Hughey, 1998). Other resource-saving techniques include sparse DP algorithms such as Treeterbi (Keibler et al., 2007). Such time- and space-saving approaches may make analysis of even extremely long genomic sequences increasingly feasible.

Kim and Sinha note that, in practice, the actual complexity of Indelign is often significantly reduced by the restrictions on evolutionary histories that they impose, namely that (i) nucleotides cannot be deleted and then re-inserted at the same position and (ii) indel event boundaries coincide with observed gap boundaries. Both of these assumptions significantly constrain the available paths through a phylogenetically composed transducer, and so should benefit any transducer-based method. Assumption (i) is often taken as standard in the statistical alignment literature (Thorne et al., 1991) and is implicit in the rules for transducer composition (Holmes, 2003). Assumption (ii) can be expressed as a restriction on the transitions that the transducer can use at each particular alignment column.

A strength of Indelign’s approach is the ease with which arbitrary distributions over indel lengths can be modeled. HMMs and transducers, in contrast, most naturally model geometric distributions. Extra states can be introduced to give arbitrary length distributions (this is the procedure Kim and Sinha use when describing the Pair HMM of their model) but much of the expressive power so conferred, such as long tails, can be compactly approximated by a transducer with a mixture of geometrics [see Do et al. (2005) for the PROBCONS program]. This is a long-understood design principle of bioinformatics state machines (Miller and Myers, 1988).

Addendum: During the review phase for this article, Diallo et al. (2007) published results using a phylo-HMM extremely similar to the one we have proposed in this section. In place of exact MCMC, they introduce a principled approximation that limits complexity by discarding low-valued cells from the DP matrix.

3 VERSATILE MACHINES

As we have shown, transducers provide a consistent language for many different flavors of alignment, including multiple alignment (Hein, 2001; Holmes, 2003) and post-alignment inference (Diallo et al., 2007; Kim and Sinha, 2007). The theory can frame questions of computational complexity in such models.

The range of possible algorithms extends beyond maximum-likelihood inference of ancestral indel history. One can sum over histories using the Forward-Backward algorithm (Durbin et al., 1998; Holmes, 2003), or sample histories from the posterior distribution using various flavors of MCMC (Holmes and Bruno, 2001; Lunter et al., 2004). Despite several assertions in the literature that MCMC or statistical alignment are unlikely to be practical for genomes, there is no reason to anticipate that this should be so. It is possible to construct transducer-based MCMC algorithms using similar resources to pairwise alignment (Holmes, 2003; Holmes and Bruno, 2001). While unconstrained pairwise alignment of genomic-scale sequences is impractical, several methods that impose constraints to reduce memory usage can be applied to MCMC (Bray and Pachter, 2004; Metzler et al., 2001; Myers and Miller, 1988).

One can readily estimate evolutionary rates and other parameters for transducer models. Measurement of evolutionary rates may reveal natural selection and other interesting signatures of evolution (Holmes, 2005b; Lunter et al., 2006). This can be achieved either by maximum-likelihood techniques such as Expectation Maximization (Durbin et al., 1998; Holmes and Rubin, 2002) or by MCMC (Lunter et al., 2005; Metzler et al., 2001). Bayesian methods, such as the use of priors, can easily be introduced (Brown et al., 1993).

Many bioinformatics analyses that use multiple alignments may benefit from reformulation in terms of transducers. Examples include phylogeny, where current techniques for sampling trees can be extended to co-sample alignments (Lunter et al., 2005); homology profiling, where HMMs that incorporate evolution have enhanced performance (Qian and Goldstein, 2004); comparative genome annotation using phylogenetic grammars (Klosterman et al., 2006); the detection of protein-coding genes via indels that preserve reading frame (Kellis et al., 2003) and the reconstruction of ancestral genomes (Ma et al., 2006). Transducers can also be used to model local context-dependent mutations, such as simultaneous substitutions at adjacent nucleotides (Averof et al., 2000), other context-dependent substitutions such as CpG effects (Lunter and Hein, 2004; Siepel and Haussler, 2004), or expansion/contraction of microsatellites. Tree transducers (Fülöp and
Vogler, 1998) can be used to model the evolution of structured features such as non-coding RNA genes (Holmes, 2005a) or protein-coding genes (Carmel et al., 2005). It may even be possible to model more context-dependent mutations, such as local duplications, inversions or transpositions, using models related to transducers.

3.1 Transducer software and algorithms

Several software tools for working with transducers are in common circulation, some of them unpublished. Tools for statistical alignment, phylogeny and/or parameter estimation include Fitch (Handel, Holmes and Bruno, 2001); Phylogeny Cafe (Miklos et al., 2007); BEAST (Drummond and Rambaut, 2003); BALI-Phy (Suchard and Redelings, 2006); MCMCALGN (Fleissner et al., 2005; Metzler et al., 2001) MCALIGN (Wang et al., 2006), FRANK (Loftynoja and Goldman, 2005) and Indelign (Kim and Sinha, 2007). Our lab provides several transducer-related tools and resources, including short illustrative animations (biowiki.org/PhyloFilm).

ACKNOWLEDGEMENTS

I.H. was funded in part by NIH/NHGRI grant 1R01GM076705-01. RB was supported under an NSF Graduate Research Fellowship.

Conflicts of Interest: none declared.

REFERENCES

Lunter, G. et al. (2005) Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinformatics, 6, 83.


