Figure S5: Production rules of the non-structural component (a) and the structural component (b). Nomenclature: ‘|’ denotes a choice between different productions; x denotes single column emissions; xl and xr denotes the left and right part of pair emissions, respectively. A corresponding graphical overview of these grammar components are given in fig. S6.

(fg. S5b and fig. S6b), and a high-level component (not shown), which combines the two other components. In contrast, the background model consists of only the non-structural component.

The overall structure of the fRNA grammar is identical to the grammar used by RNA-decoder [11], but the structural and the non-structural components are different. The structural component uses the RNA-grammar proposed by [13]. Note that we write these grammars using the restricted set of production rules given above, this causes the structural grammar to be written with three more states than originally used by Knudsen and Hein. A minimum loop-size of three bases is enforced by the emission distribution of the stem pair state. No minimum stem size is enforced.

Phylogenetic models

Since the input of EvoFold is multiple sequence alignments, the emission distributions of its SCFGs are defined over sets of alignments columns. There are exceedingly many such alignment columns for even a moderate number of aligned sequences: \(16^n\) in the case of paired columns with \(n\) sequences in the alignment. Explicitly defining a distribution over these columns is therefore infeasible.

The use of phylogenetic models for studying molecular evolution and evolutionary relationships has a long history [14, 15], and they are now becoming