Inverting the Viterbi Algorithm: An Abstract Framework For Structure Design

Michael Schnall-Levin

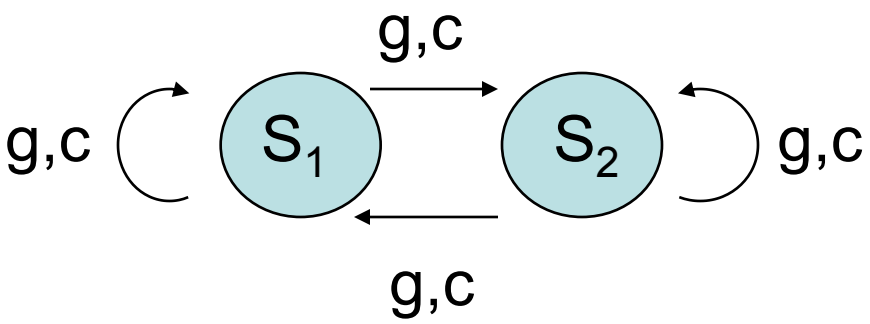
Massachusetts Institute of Technology

Joint work with:

Leonid Chindelevitch and Bonnie Berger

HMMs and SCFGs

Hidden Markov Model



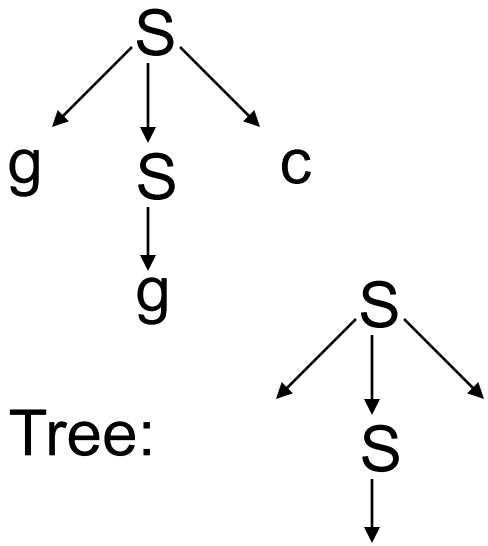
State-path: $S_1S_2S_2S_1$

Emission sequence: ggc

Stochastic Context-Free Grammar

S ⌫ cSg | gSc

S ⌫ cS | gS | c | g



Derivation Tree:

Emission sequence: ggc

HMMs and SCFGs: 3 Fundamental Problems

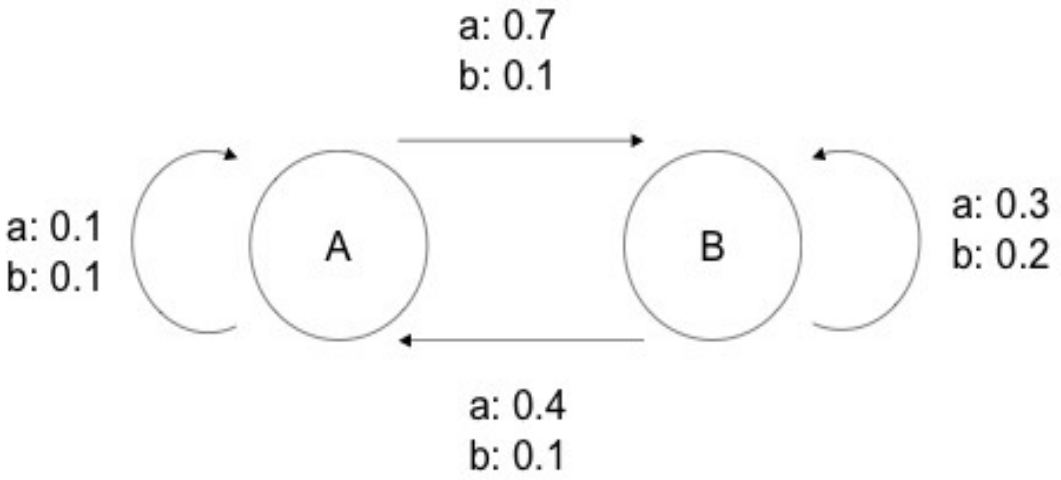
- **Decoding Problem**: Given model and sequence find most likely state-path
- **Evaluation Problem**: Given model find probability of sequence being emitted
- **Learning Problem**: Given set of sequences learn parameters of a model

A Novel Problem: The Design Problem

- **Design Problem:** Given model and state-path, find a sequence for which this state-path is optimal

- Design problem inverse to decoding problem

Design Problem: An Example



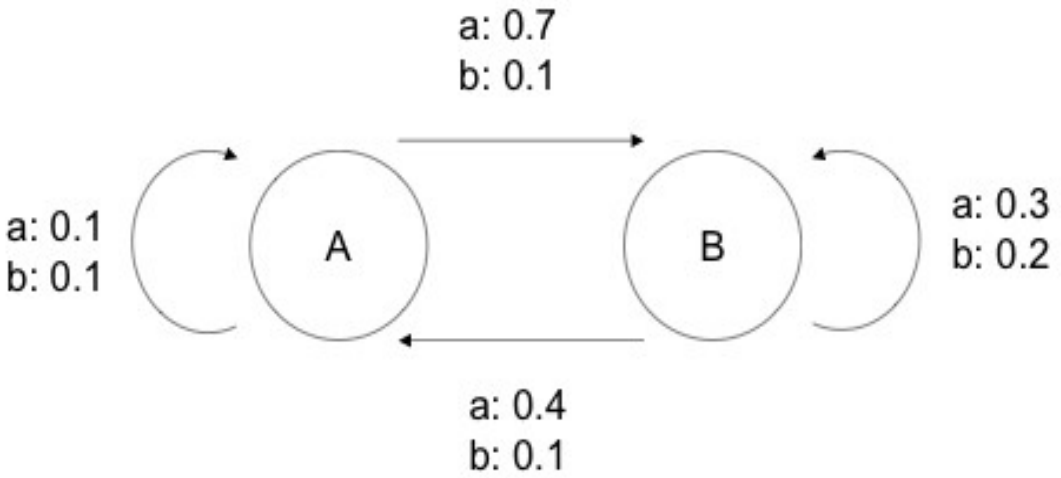
States: A, B

Emissions: a, b

Decoding Problem:
Input: abaa
Output: BABAB

Design Problem:
Input: BABAB
Output: abaa

Design Problem: Trivial Solutions Won't Work



States: A, B

Emissions: a, b

Design path B^n

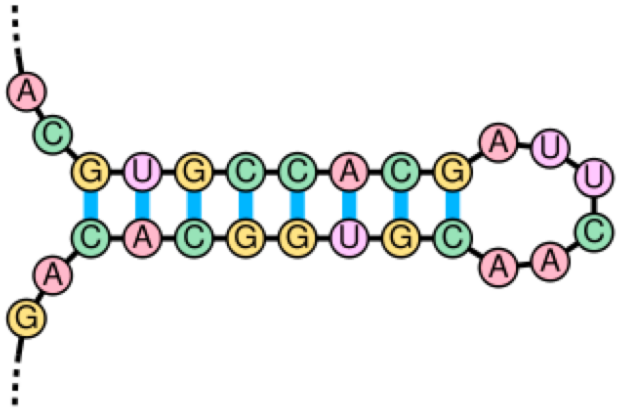
Unique answer b^{n-1}: least likely sequence

$Pr(b^{n-1} | B^n) = (0.2)^{n-1}$

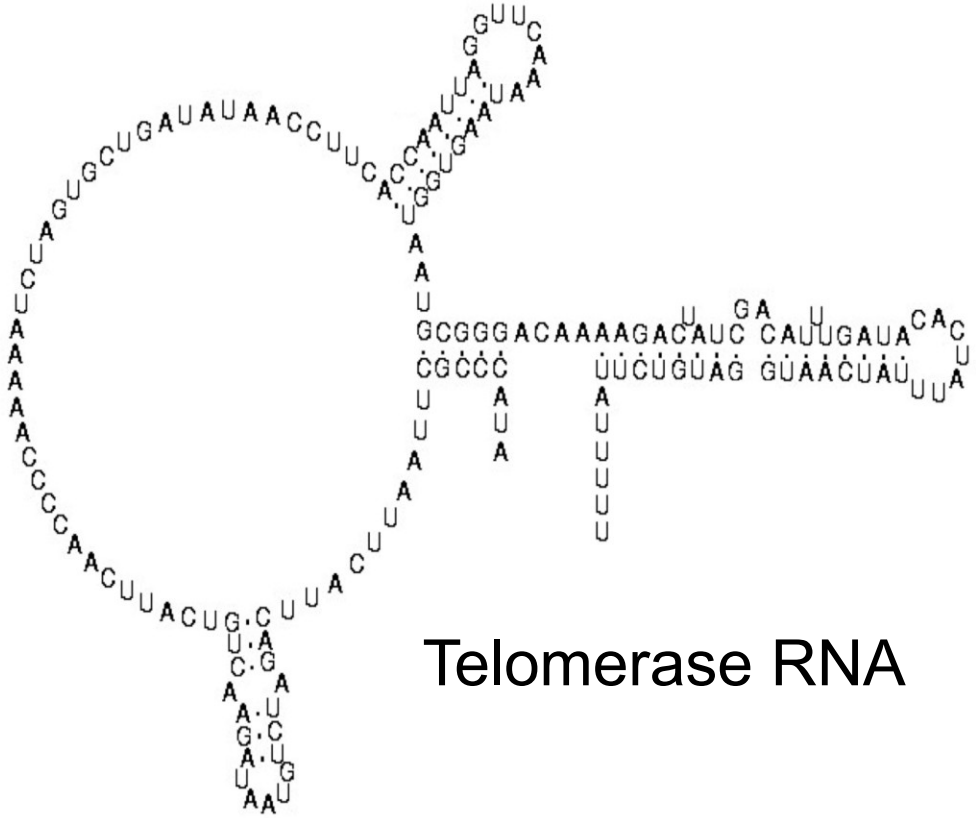
HMMs and SCFGs: Biology Applications

- Labeling DNA sequences
  - Gene prediction
  - Exon/Intron prediction
  - Protein binding sites
- Protein family classification
- Protein and RNA secondary structure prediction

RNA Secondary Structure



Hairpin Loop



Telomerase RNA

Pattern of base-pairs when chain folds on itself

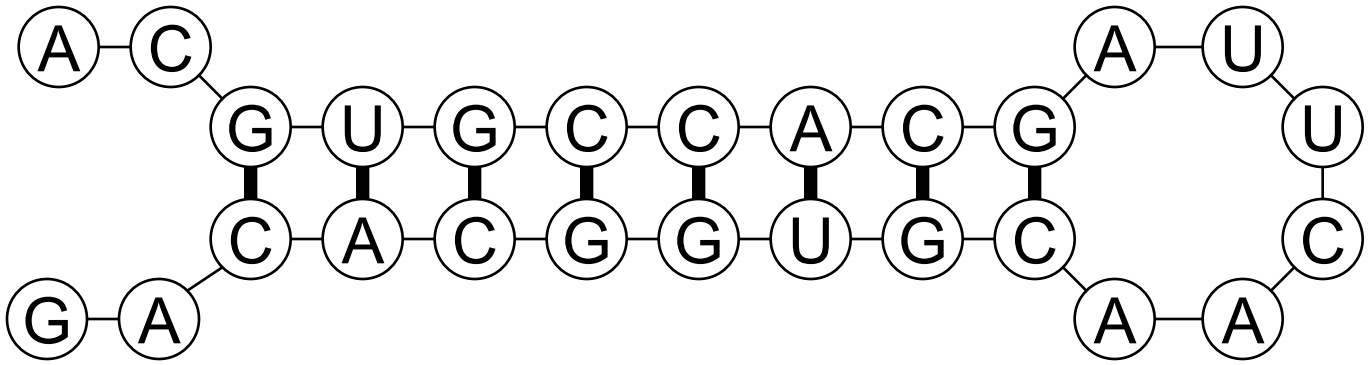
Canonical pairs: A-U, G-C and G-U wobble

Secondary Structure Prediction

- Input: Sequence of nucleotides
- Output: Base-pairing sequence will take

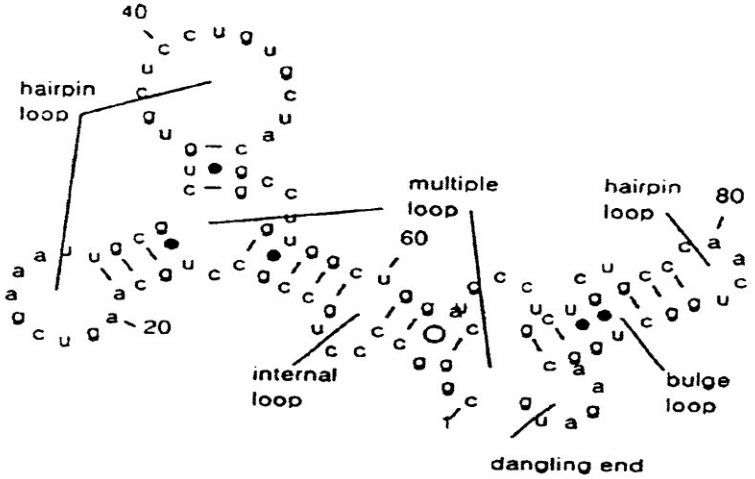
Input ACGUGCCACGAUUCAACGUGGCACAG

Output



Secondary Structure Prediction

- Minimum energy structure: Zuker's algorithm
  - Energy built from small components
  - $O(n^3)$ time, n length sequence

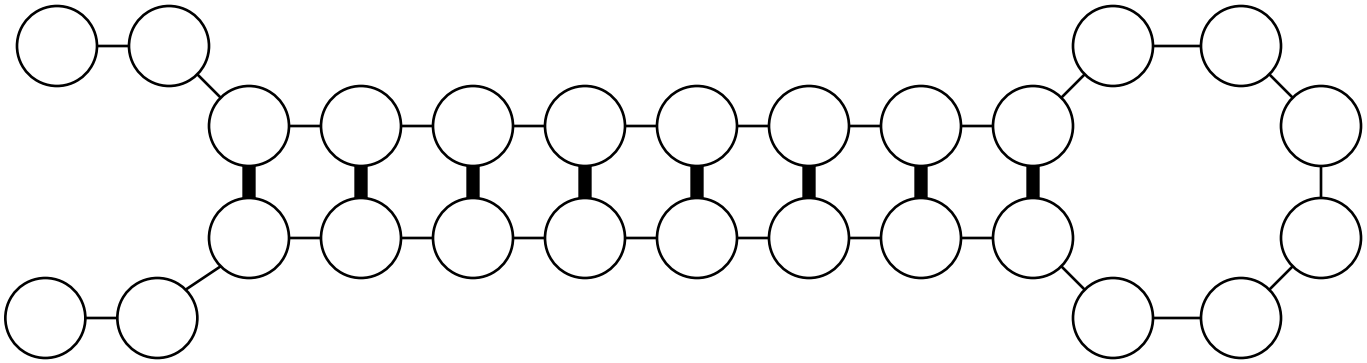


- Stochastic Context Free Grammars
  - Zuker's model can be put into SCFG
  - Minimum energy structure is best parse

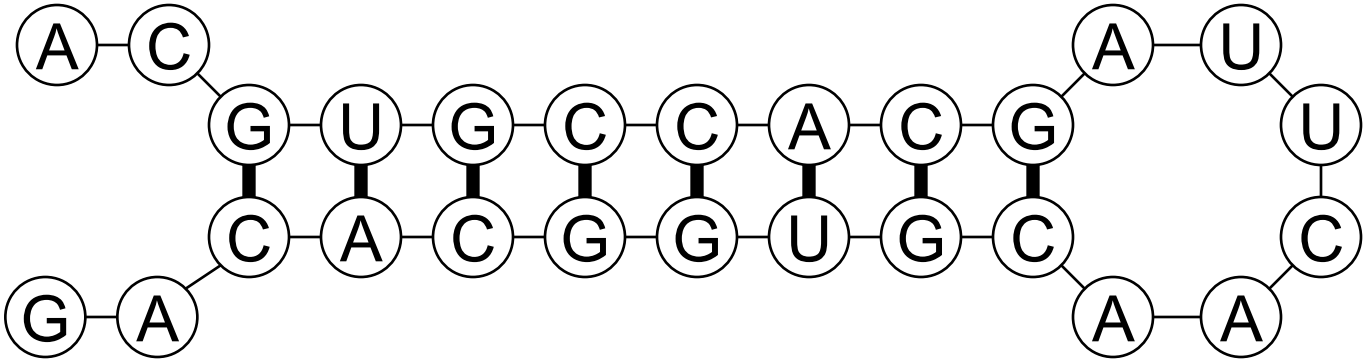
Secondary Structure Design

- Input: Desired base-pairing
- Output: Sequence that takes this base-pairing

Input

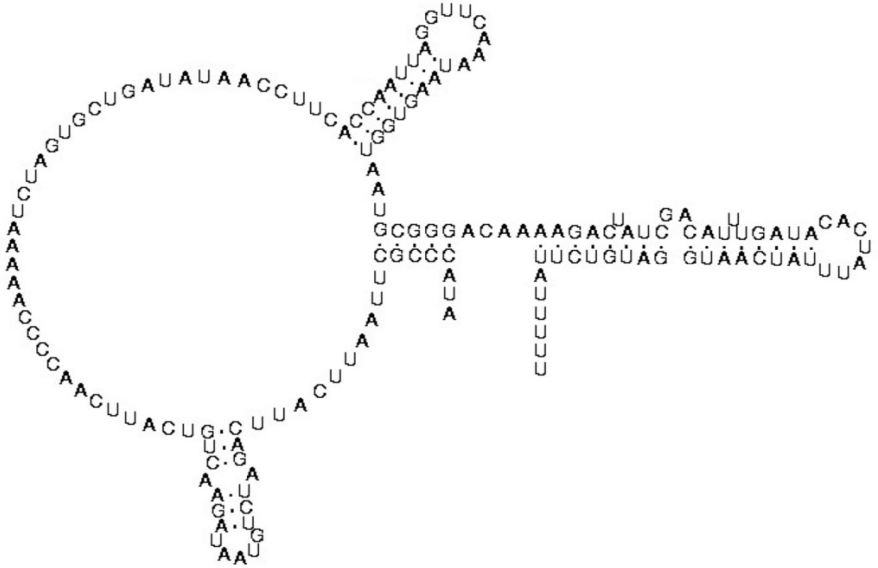


Output

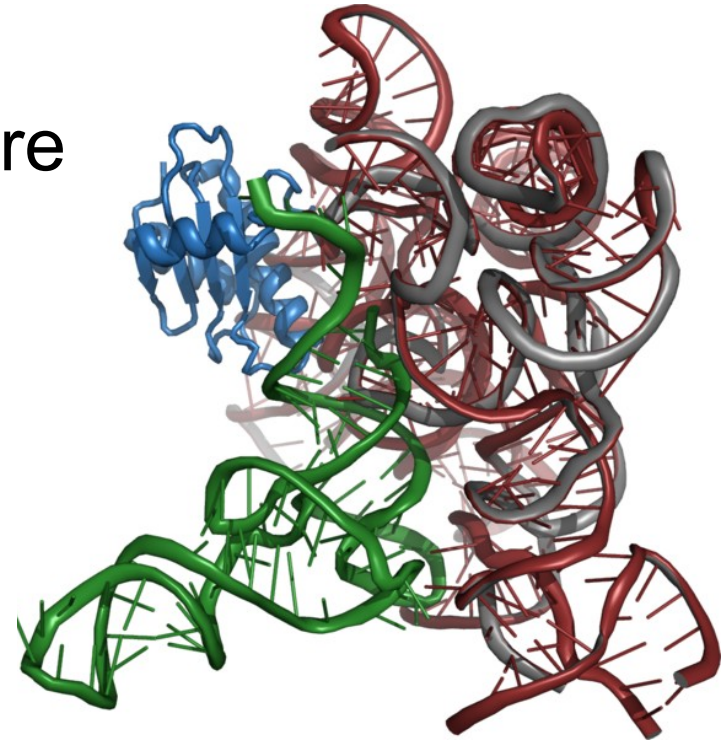


Secondary Structure Design Motivation

- RNA plays functional role
  - Ribozymes: RNA enzymes
  - Function depends on structure



Telomerase



RNase-P

Problem Correspondence

- Secondary structure prediction under SCFG model
  - Decoding problem: Given sequence find best structure
  - Design problem: Given structure find sequence so that structure is optimal

Secondary Structure Design

- Current methods for structure design all rely on heuristics to search for a sequence

- 3 software packages available:
  - RNAinverse (1994)
  - RNA-SSD (2004)
  - INFO-RNA (2006)

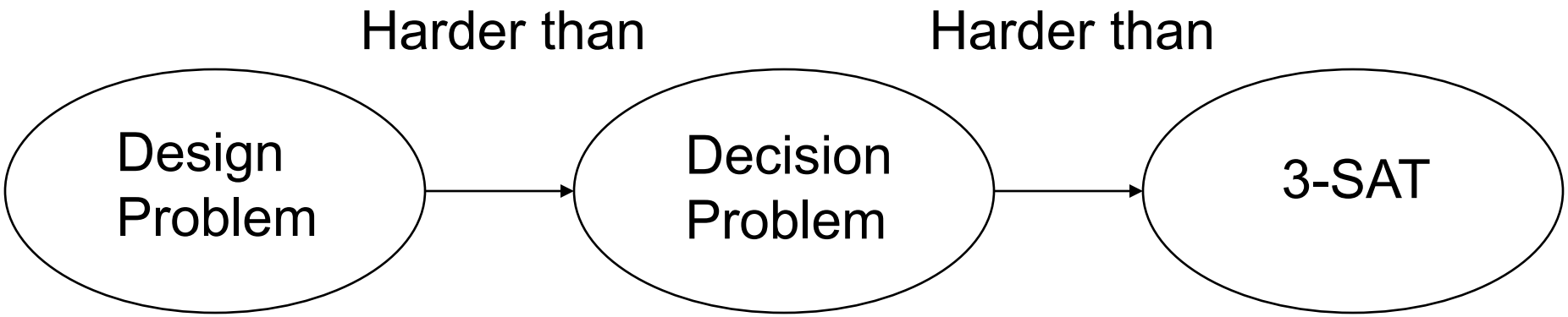
Problem Motivation

- How hard is design under models that make different assumptions (HMMs, SCFGs, …)
- Are heuristics best that can be done?
- Can a more general framework aid in developing algorithms for design in specific cases?

Theoretical Results

- Novel problem on HMMs and SCFGs
- Proof that problem is NP-hard on HMMs
  - NP-hard on SCFGs by extension
- Branch-and-bound algorithm in HMMs
  - Search strategy
  - Theoretical bounds for special cases
- ILP formulation

NP-hardness Proof: Overview

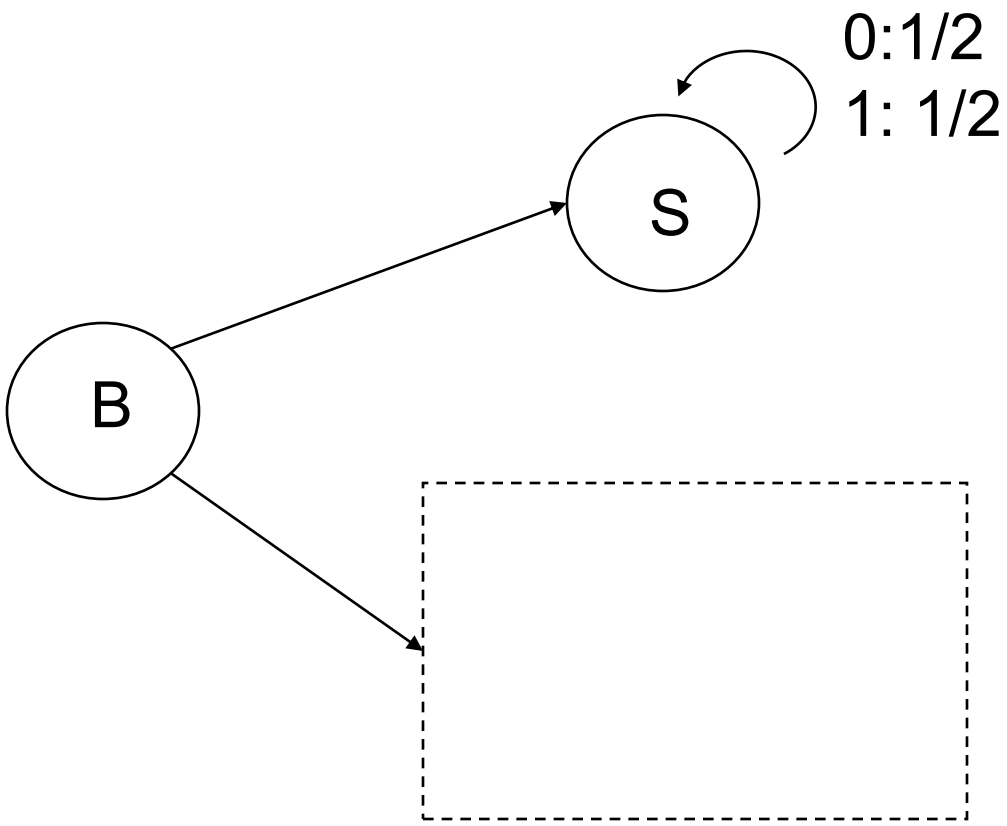


Since 3-SAT is NP-hard, Design Problem is NP-hard

NP-Hardness Proof

- Design problem:
  - Given state-path find sequence so that state-path is optimal
- Decision problem:
  - Given state-path is there a sequence so that state-path is optimal
- 3-SAT
  - Conjunction of m clauses over n variables
  - Eg $(x_1 \vee x_2 \vee \sim x_4) \wedge (\sim x_1 \vee \sim x_3 \vee x_4)$

Illustration of Reduction



Emits all strings ω

$Prob(S,\omega) = (1/2)^n$

Emits non-satisfying assignments to 3-SAT problem

$Prob(\pi,\omega) = (1/2)^{n-3}$

Path BS^{n+1} designable $\Leftrightarrow$3-SAT problem is satisfiable

Second Component

Emits non-satisfying assignments

=

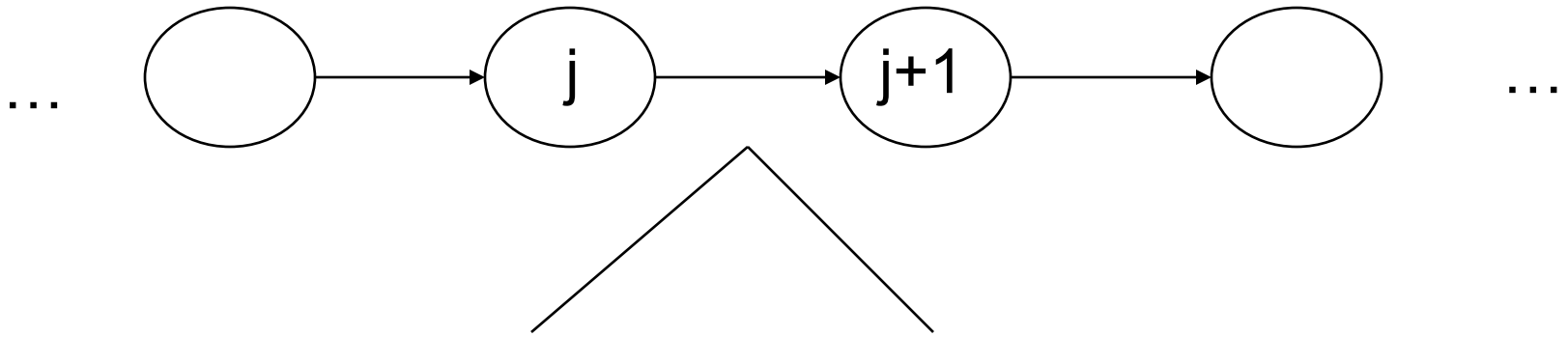
Emits strings don't satisfy clause 1

.
.
.

Emits strings don't satisfy clause m

Second Component

Emits strings don't satisfy clause k with prob = $(1/2)^{n-3}$



Emits 0 prob1, if x_j in clause k

Emits 1 prob 1, if $\sim x_j$ in clause k

Emits 0,1 prob 1/2, otherwise

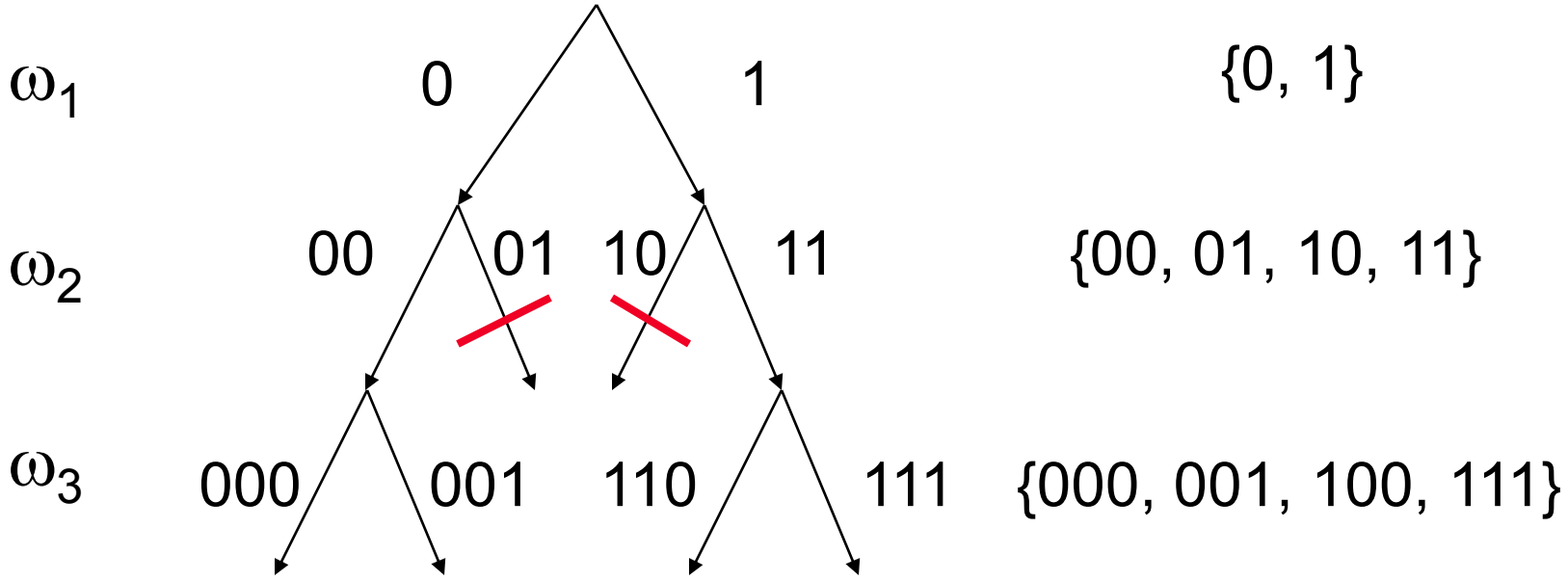
Branch-and-Bound Algorithm

- Search strategy of space of sequences
- Rules eliminate whole branches of search tree
- Step through positions in sequence and eliminate choices iteratively

Branch-and-Bound Algorithm

- At stage i have list of possible $\omega = \omega_1 \ldots \omega_i$
- Elimination Rule 1:
  - Does ω fail to satisfy Viterbi constraints?
- Elimination Rule 2:
  - Is there an ω' better than ω?

Branch and Bound Algorithm



Stage 1:
No eliminations

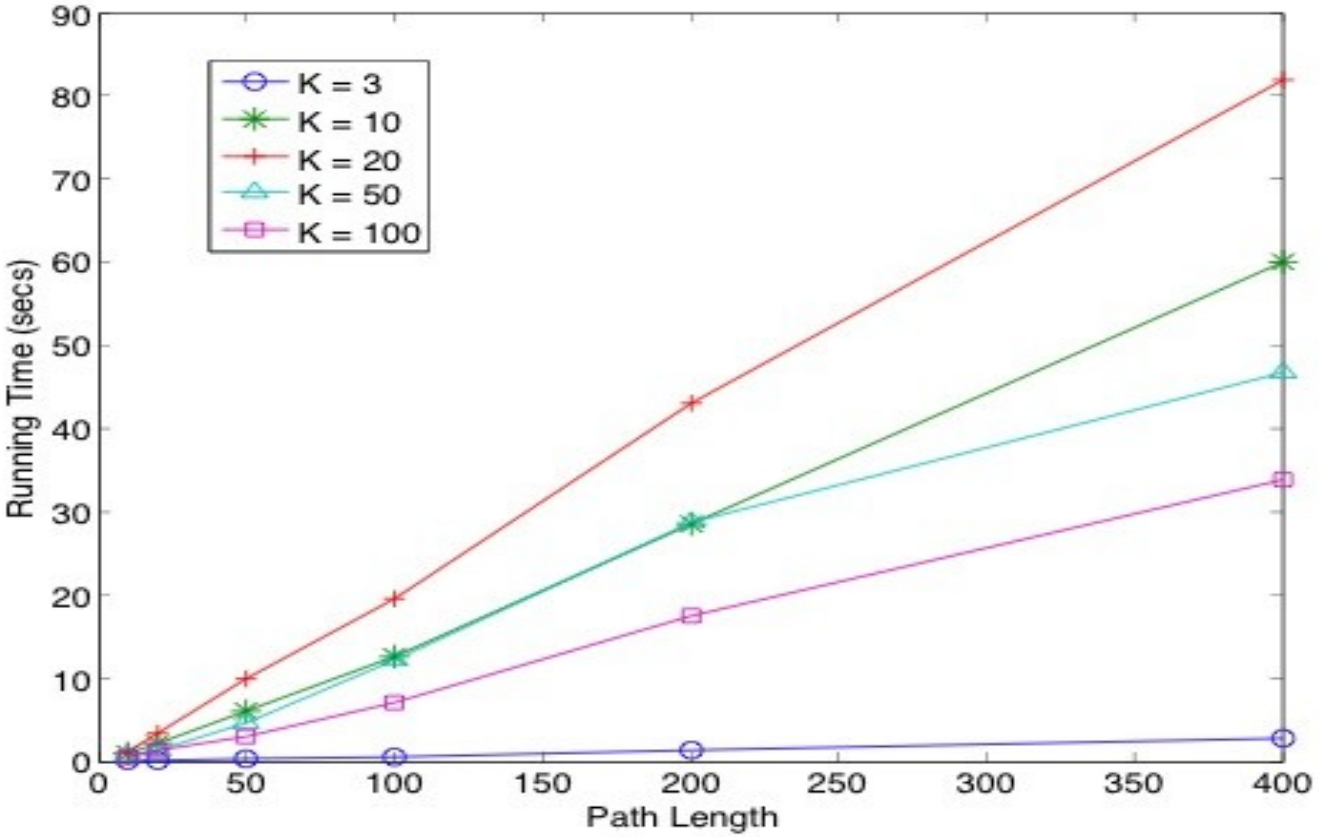
Stage 2:
01 doesn't satisfy Viterbi constraints
00 better than 10

Running Time Bound

- Branch-and-bound algorithm works on all inputs
- Under assumptions get worst-case running time:
  - All log-probabilities > -B
  - Log-probabilities known to precision δ
- Running time $O((2B/\delta)^{k-2}nk^2|\Sigma|)$
  - Fixed parameter tractable

n: length of sequence k: number of states Σ: emission alphabet

Simulations: Running Time



Running times versus path length and number of states on randomly generated HMMs (run on 3.0GHz Intel PC)

Conclusions

- Introduced novel problem on HMMs and SCFGs
- Theoretical Results:
  - Proof problem is NP-hard
  - Initial algorithmic results
- Biological implications:
  - RNA secondary structure design problem is algorithmically hard

Acknowledgements

Hertz Foundation
NDSEG
NSERC

Michael Collins
Mathieu Blanchette
Michael Baym
Jerome Waldispuhl
Andreas Schulz