# Computational Linguistics and Computational Biology

Fernando Pereira

CIS, University of Pennsylvania

with

Aravind Joshi

Mark Liberman

*et al*

# Computational Linguistics

- Data-driven revolution:
  - Large text and speech datasets
  - Experimental, rigorous evaluation
  - Machine learning dominates
- Sequence modeling
  - Grammars for describing sequence structure
  - Learning grammars/parameters: classification, segmentation, parsing
- Information management
  - Entity and relation extraction
  - Integration

# Sequence Modeling

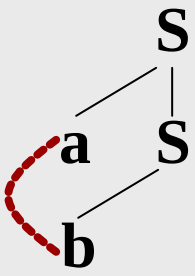
- Previous successes
  - Local sequence statistics: Markov models
  - Sequence structure: finite-state and context-free grammars
- Challenges:
  - More complex structures: folding
  - Long-range dependencies
  - Integrating multiple sources of evidence

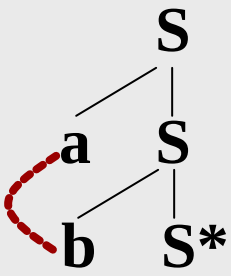
# Grammatical Techniques

- Modeling folded structures in RNA and proteins using tractable mildly context-sensitive formalisms
  - Pseudoknots
  - Doubly embedded pseudoknots
  - Complex beta-sheet folds
- Efficient exact computation of folding free energy $\Rightarrow$ probabilistic grammars

# Tree Adjoining Grammar

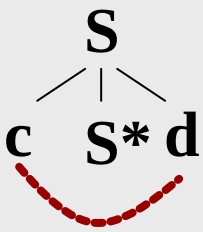
*Elementary trees:*



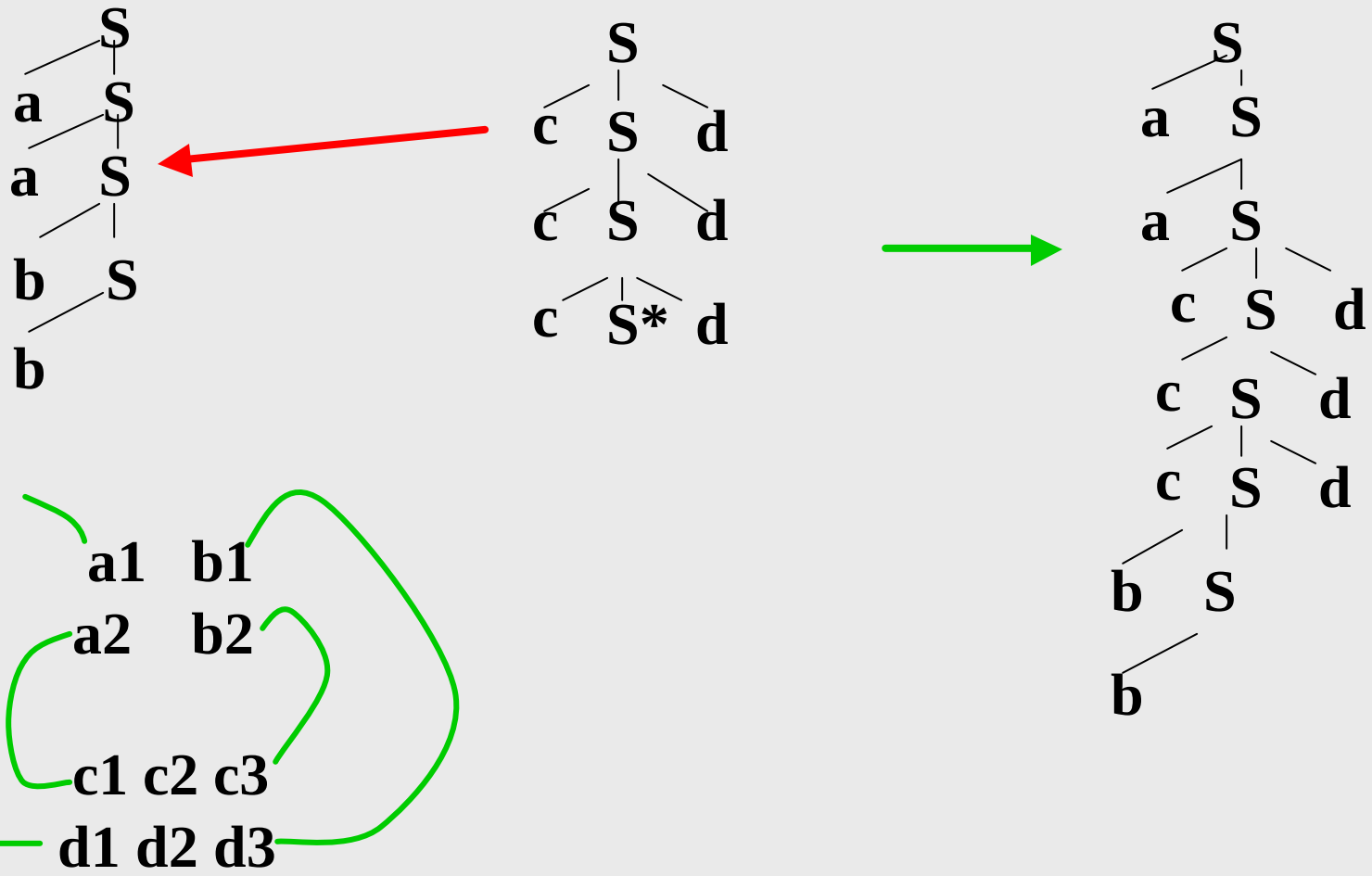

Assembly by

- substitution
- *adjoining* (inserting in)



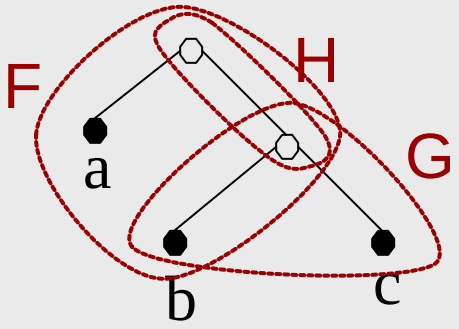

**matching base pairs**

# Pseudoknot


S
a S
a S
b S
b
S
c S d
c S d
c S* d
S
a S
a S
c S d
c S d
c S d
b S
b
a1 b1
a2 b2
c1 c2 c3
d1 d2 d3

# Grammars and Probabilities

- Learn the statistical distribution of segmentations/analyses/conformations
- Standard approach: probabilities of derivation steps
  - Problems: hard to model long-distance interactions, evidence combination
- New approach: conditional undirected graphical models/random fields




$$P(\text{analysis} \mid \text{sequence}) = \frac{\exp(V(F) + V(G) + V(H))}{Z(\text{sequence})}$$

# Proposed Applications

- Folded structures:
  - Verify structural predictions of grammatical models
  - Develop probabilistic parameterization for fold prediction
- Mining *Apicomplexa* genomes :
  - Better TIS recognition
  - Identify novel apicoplast-targeted proteins

# Better Information Extraction

- Create high-quality annotated corpus
- High-accuracy statistical parsing
- Shallow semantic analysis
- Database integration
- Proposed applications:
  - Enzyme inhibitors
  - Genotype/phenotype relations