Semantics and Ambiguity of Stochastic RNA Family Models

25th TBI Winterseminar

Robert Giegerich [1] Christian Höner zu Siederdissen [2]

[1]Center of Biotechnology and Faculty of Technology
Bielefeld University

[2]Institute for Theoretical Chemistry
University of Vienna

February 16, 2010

Uses for Stochastic Models

- SCFGs: single RNA folding
- HMMs: protein families
- CMs: RNA families

an example CFG ($G1$): $S \rightarrow \varepsilon \mid aS \mid Sa \mid aSb \mid SS$

Syntactic vs. Semantic Ambiguity

Syntactic different parses on the same sequence produce different objects (wanted)

```
gcaagc  ((..))  (....)  .(..).  ......
        0.5     0.2     0.2     0.1
```

Semantic different parses on the same sequence produce the same object (unwanted)

```
gcaagc  ((..))  (....)  ((..))  ((..)) etc
        0.15     0.2     0.15    0.05
```

Systematic evaluation of semantic ambiguity requires a *canonical representation*

Dealing with Ambiguity

Compensation

- sum over all identical objects
- number of different objects increases exponentially
- infeasible for: RNA-folding, RNA family models
- feasible for: classified DP, RNA shapes

Grammar Design

- clean approach: fewer tricks and assumptions required
- harder to design unambiguous grammar?
- more complex grammar required? (cf. early folding algorithms)

Does Semantic Ambiguity Matter

- ignore ambiguity, assume:
  most likely parse = most likely object
- Dowell & Eddy, 2004: yes, it matters:
  simple grammars give wrong answer in 20% - 98% of the cases
- problem for family models? (Infernal & Rfam)

Ambiguity in Infernal?

- complex SCFG with:
- 7 types of non-terminals with O(model size) multiplicity
- up to 6 rules per non-terminal
- ambiguous prototype grammar
- no formal definition of prototype grammar and build process

a formal proof is non-trivial!

Automated Ambiguity Checking

- transform CM grammar G into equivalent ADP grammar G_{adp}
- partial evaluation of G_{adp} with canonical string mapping gives G_{csm}
- theorem[1]: G is semantically ambiguous iff G_{csm} is syntactically ambiguous
- apply ACLA[2] ambiguity checker on G_{csm}

[1] J. Reeder et al, Effective ambiguity checking in biosequence analysis, 2005

[2] C. Brabrand et al, Analyzing Ambiguity of Context-Free Grammars, 2007

First Results

- Infernal models are unambiguous
- as proven using an automated ambiguity checking pipeline
- . . . given the current Infernal *semantic* model

First Results

- Infernal models are unambiguous
- as proven using an automated ambiguity checking pipeline
- . . . given the current Infernal *semantic* model

is the Infernal semantic model:

- correct?
- a good choice?

Semantics of Family Models

sequence alignment recap:

```
ACAGGGG---CAC   ACA---GGGGCAC     ACA[GGGG]CAC
ACA----TTTCAC   ACATTT----CAC     ACA[TTT] CAC
```

three meaningful semantics can be defined for family models:

```
 Consensus **<<**>>

 Alignment **<<*--*>>    **<<*--*>>    **<<**>>
           __((....))    .._(....)_    ..((..))

     Trace *-*<<**>>    **<-<**>>
           ..__(..)_    .__.(..)_
           allowed      banned

  Strucural **<<*-*>>     consensus implicit only
             ((...))
```

(Non-)ambiguity of Infernal

- Infernal grammar unambiguous for alignments
- many alignments form one trace
- many traces form one structure
- Infernal is ambiguous for traces!

Non-ambiguous Trace Semantics for Family Models

```
remember:     ACA[GGGG]CAC
              ACA[TTT] CAC
```

$$
\begin{aligned}
A &\rightarrow \overset{-}{.}\ A \mid M \\
M &\rightarrow \varepsilon \mid \overset{*}{.}\ A \mid \underline{*}\ M \mid \\
&\quad \overset{<}{(}\ A\ \overset{>}{)}\ A \mid \overset{<}{.}\ A \geq M \mid \\
&\quad \leq M\ \overset{>}{.}\ A \mid \leq M \geq M
\end{aligned}
$$

- proved unambiguous using the ACLA ambiguity checker
- by virtue of construction, the above grammar generates unambiguous model grammars

Counting Alignments and Traces

Model length (size)	RF00163 45 (31)	RF01380 19 (12)
$\|x\| = 12$		
structures	8,958	2,048
traces	35×10^{9}	141,120,525
alignments	715×10^{12}	35,330,137,025
$\|x\| = 31$		
structures	n.a.	n.a.
traces	2×10^{21}	30,405,943,383,200
alignments	2×10^{27}	208,217,738,981,165,823

RF00163 consensus:

<<<<<<*******<<<<******>>>>***<<<<>>>>*>>>>>>

RF01380 consensus:

<<<<<<<****>>>*>>>>

Summary

- simplified grammar design process using ADP
- automated ambiguity checking: canonical representation, *cm2adp*, acla
- new trace grammar to better capture the biological process
- more accurate scoring for distant members of RNA families
- trace grammar smaller by a factor of 2 – 4
- ideally: 'Infernal 2.0' with same target language as *cm2adp*

Outlook

- *cm2adp*: designed for semantic enrichment
  with Robert Giegerich
  - multiple algebras & product operation for rich output
  - extract more information
- *cmcompare*: comparison of covariance models
  with Ivo Hofacker
  - improve design of existing models
  - detect similarities between models

- Robert Giegerich, Christian Höner zu Siederdissen
  Semantics and Ambiguity of Stochastic RNA Family Models
  IEEE/ACM Transactions on Computational Biology and Bioinformatics
- Christian Höner zu Siederdissen, Robert Giegerich
  Semantic Enrichment of Covariance Models
  in preparation
- Christian Höner zu Siederdissen, Ivo Hofacker
  Comparison of Covariance Models
  in preparation