Distribution of motifs generated by a stochastic context-free grammar for RNA folding

Svetlana Poznanović

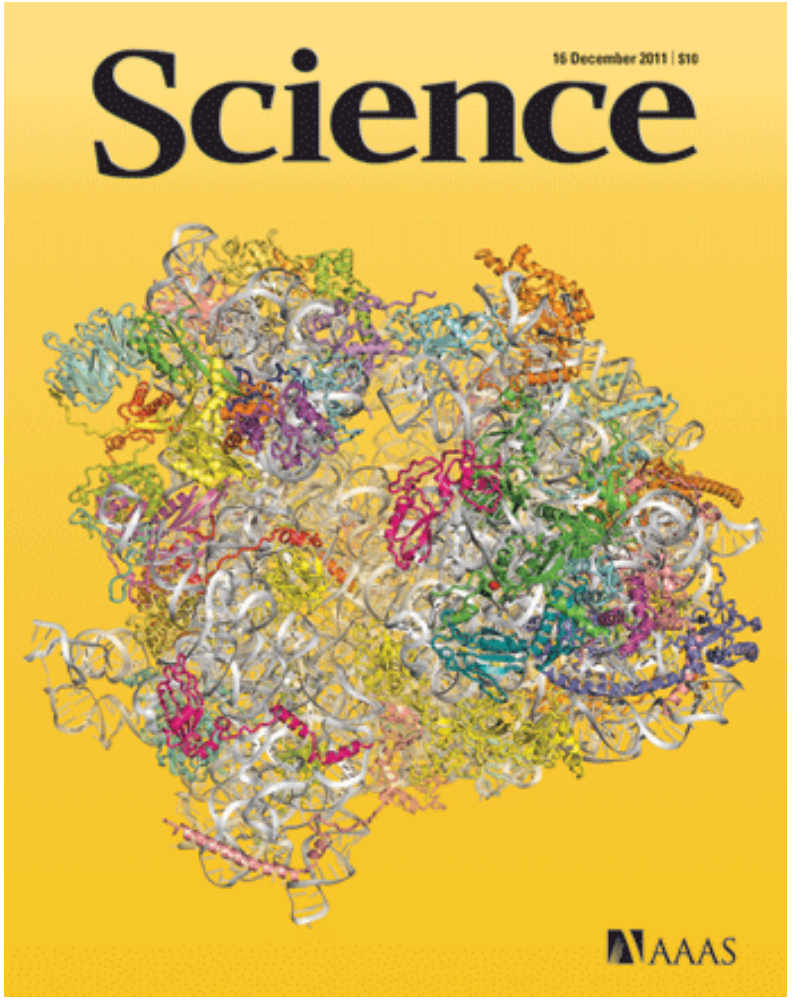
Department of Mathematical Sciences
Clemson University

Folding of RNA

Sequence $\longrightarrow$ Structure

GGGCGUAUGGCG
CGUAGUCGGUAG
CGCGCUCCCUUC
GCCUGGGAGACU
CCGGUGUUCCGG
ACACGUCCACCA

RNA structure prediction

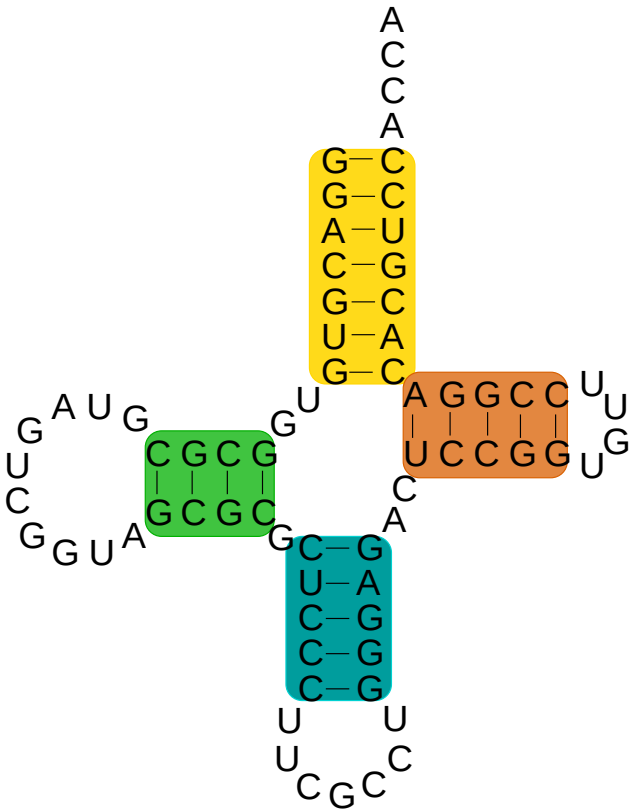


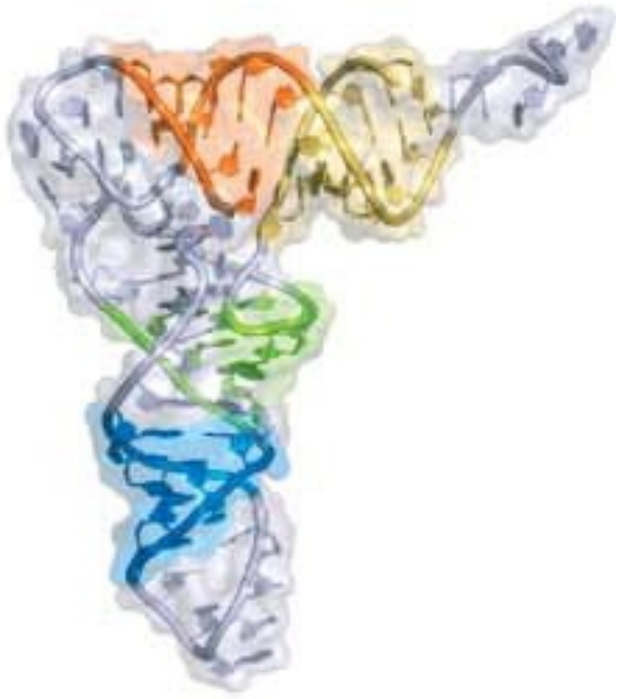
- Structure is often important for function
- Experimental determination of the structure is too often nontrivial

Intermediary structure: secondary structure

Sequence $\longrightarrow$ Secondary structure $\longrightarrow$ 3D structure

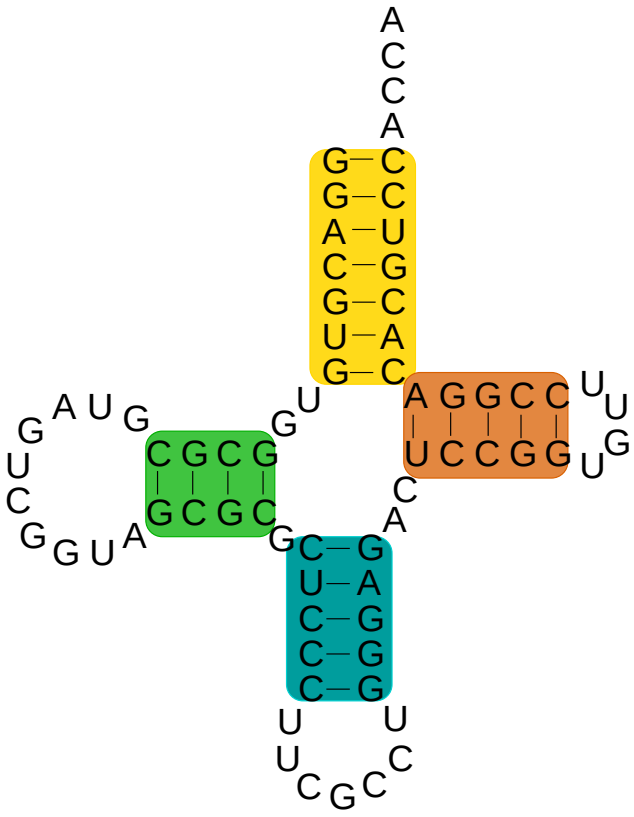
GGGCGUAUGGCG
CGUAGUCGGUAG
CGCGCUCCCUUC
GCCUGGGAGACU
CCGGUGUUCCGG
ACACGUCCACCA

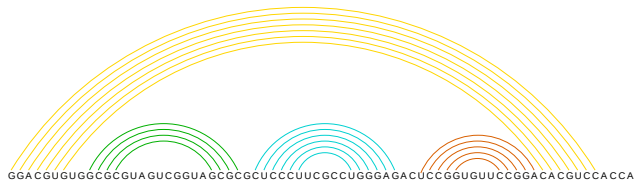




Secondary structure

- Prediction of 3D structure is difficult, but the secondary structure can reveal a lot of information.
- Secondary structure = Non-crossing matching

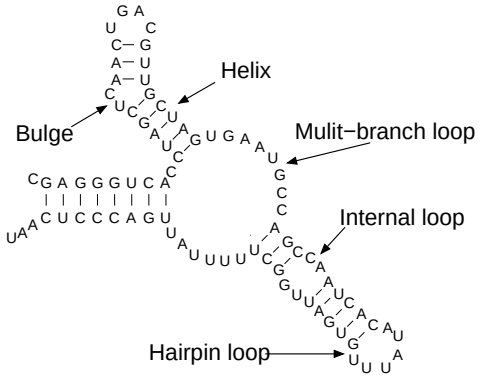




Secondary structure prediction methods

Physics-based methods

- Discrete optimization problem
- Model uses thousands of measured thermodynamic parameters



Knowledge-based methods

- Parameters obtained through Maximum Likelihood Estimation

Stochastic context-free grammar

Stochastic context-free grammar (SCFG)

$$G = (V, \alpha, S, R, P_R)$$

- V is a nonterminal alphabet
- α is a terminal alphabet
- $S \in V$ is a special start symbol
- R is a set of productions in which the left hand-side is a nonterminal symbol
- P_R is a probability distribution on productions:

$$\sum_{\lambda} P(X \to \lambda) = 1, \qquad \forall X \in V.$$

The Pfold SCFG

- Pfold – program for RNA secondary structure prediction which uses a SCFG (Knudsen, Hein, 1999, 2003)
- Nonterminal symbols $= \{S, L, F\}$
  Terminal symbols $= \{d, d', t\}$

$$
\begin{array}{llllll}
S \to LS & (p_1) & or & L & (q_1) \\
L \to dFd' & (p_2) & or & t & (q_2) \\
F \to dFd' & (p_3) & or & LS & (q_3)
\end{array}
$$

- $p_i, q_i > 0$ – probabilities
- $p_i + q_i = 1$

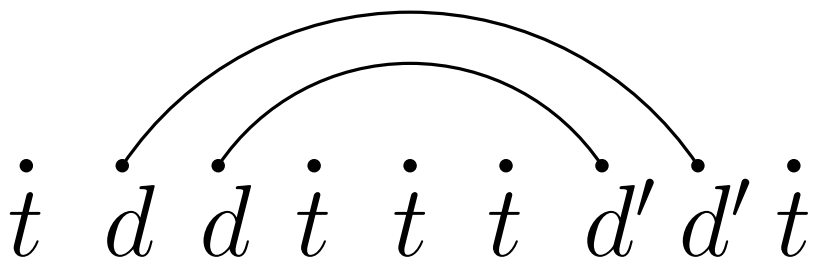
Left-to-right parse $\longleftrightarrow$ Secondary structure

S $\overset{p_1}{\Rightarrow}$ LS $\overset{q_2}{\Rightarrow}$ tS $\overset{p_1}{\Rightarrow}$ tLS $\overset{p_2}{\Rightarrow}$ tdFd'S $\overset{p_3}{\Rightarrow}$ tddFd'd'S $\overset{q_3}{\Rightarrow}$

tddLSd'd'S $\overset{q_2}{\Rightarrow}$ tddtS d'd'S $\overset{p_1}{\Rightarrow}$ tddtLSd'd'S $\overset{q_2}{\Rightarrow}$ tddttSd'd'S $\overset{q_1}{\Rightarrow}$

tddttLd'd'S $\overset{q_2}{\Rightarrow}$ tddtttd'd'S $\overset{q_1}{\Rightarrow}$ tddtttd'd'L $\overset{q_2}{\Rightarrow}$ tddtttd'd't

- d, d' – paired nucleotides
- t – unpaired nucleotide

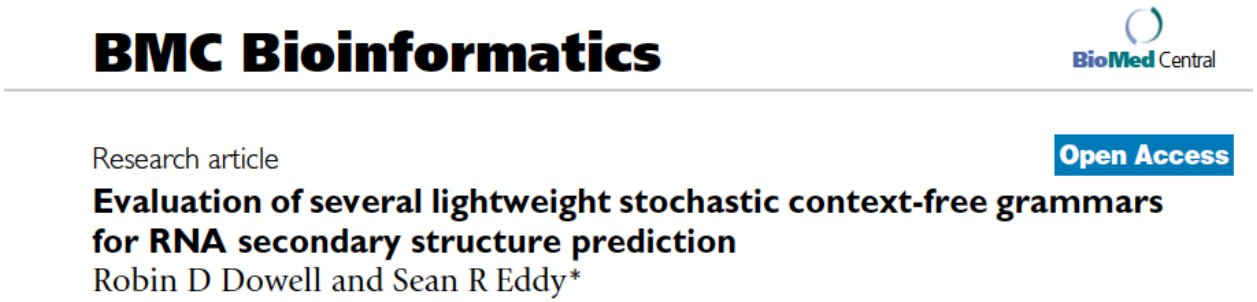


$$\begin{array}{cc} t & t \\ d & d' \\ d & d' \\ t & t \\ & t \end{array}$$

$$\begin{array}{llllll} S \to LS & (p_1) & \text{or} & L & (q_1) \\ L \to dFd' & (p_2) & \text{or} & t & (q_2) \\ F \to dFd' & (p_3) & \text{or} & LS & (q_3) \end{array}$$

Evaluation of the Pfold SCFG

"Apparently without considering alternative designs, Knudsen and Hein had already described the simple and effective G6 grammar, and extended it to analysis of input multiple sequence alignments in the program Pfold [...] Because our exploration has not been exhaustive, we can not say that G6 is the best possible simple grammar. However, after exploring various alternative SCFG designs, we confirm that the Knudsen/Hein grammar is an excellent, simple framework in which to develop some probabilistic RNA analysis methods."


BMC Bioinformatics

BioMed Central

Research article

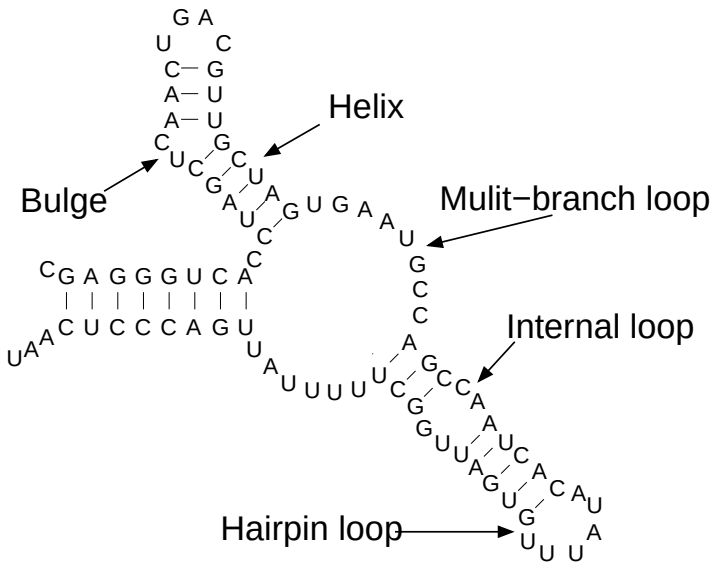
Open Access

Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction

Robin D Dowell and Sean R Eddy*

So ...

Q: How "sensitive" is the probability distribution defined by the SCFG under perturbation of the probability parameters?



Central limit theorems

- $\mathbb{X}_n$ – number of occurrences of a fixed biologically interesting motif in a secondary structure with n nucleotides
- Biologically interesting motifs: base pairs, helices, multi-branch loops, hairpin loops, internal loops, bulges

Theorem (Heitsch, P.)

For a generic choice of the grammar probabilities,

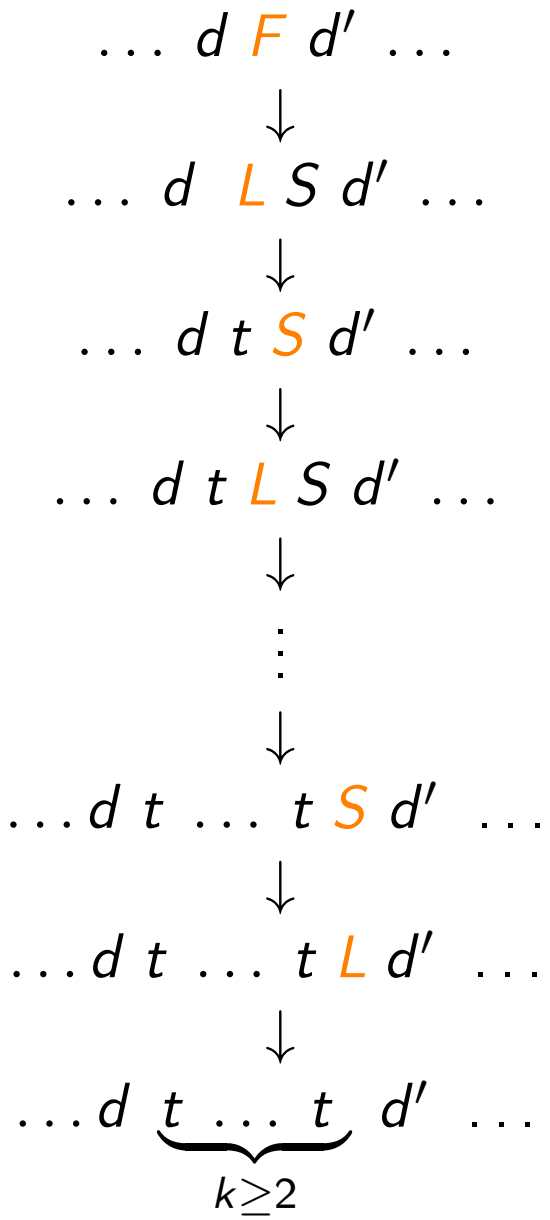
$$\mathbb{X}_n^* = \frac{\mathbb{X}_n - \mu n}{\sqrt{n\sigma^2}}$$

converge in distribution to a Gaussian variable, i.e.,

$$\lim_{n\to\infty} \mathbb{P}\left(\mathbb{X}_n^* < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{c^2}{2}}\, dc.$$

The constants μ and σ can be explicitly computed as functions of the probabilities.

Proof sketch: Hairpin loops



- Probability for this sequence of steps:

$$q_1 q_2^2 q_3 (p_1 q_2)^{k-2}.$$

- Hairpin loops created: 1
- Nucleotides created: k

$S \to LS$	(p_1)	or	L	(q_1)
$L \to dFd'$	(p_2)	or	t	(q_2)
$F \to dFd'$	(p_3)	or	LS	(q_3)

Proof sketch: From grammar to generating functions

$$\begin{cases} S \to L \text{ or } LS \\ L \to dFd' \text{ or } t \\ F \to dFd' \text{ or } LS \end{cases}$$

$$S(z,u) = \sum_{S \overset{*}{\Rightarrow} M} p(M) z^{\#\text{nucleotides}(M)} u^{\#\text{hairpin loops}(M)}$$

Proof sketch: From grammar to generating functions

$$\begin{cases} S \to L \text{ or } LS \\ L \to dFd' \text{ or } t \\ F \to dFd' \text{ or } LS \end{cases}$$

$$S(z,u) = \sum_{S \overset{*}{\Rightarrow} M} p(M) z^{\#\text{nucleotides}(M)} u^{\#\text{hairpin loops}(M)}$$

$$\begin{cases} S(z,u) = p_1 L(z,u) + q_1 L(z,u) S(z,u) \\ L(z,u) = p_2 z^2 F(z,u) + q_2 z \\ F(z,u) = p_3 z^2 F(z,u) + q_3 u \frac{q_1 q_2^2 z^2}{1 - p_1 q_2 z} + q_3 \left(L(z,u) S(z,u) - \frac{q_1 q_2^2 z^2}{1 - p_1 q_2 z} \right) \end{cases}$$

Proof sketch: Singularity analysis

$$S(z,u) = Q(z,u) - \frac{\sqrt{P(z,u)}}{2p_2q_3z^2(1-p_1q_2z)}$$

$Q(z,u)$ – rational function $P(z,u)$ – polynomial

Theorem (Flajolet, Sedgewick)

Let $G(z,u)$ be a function that is bivariate analytic at $(z,u) = (0,0)$ and has non-negative coefficients and let $\mathbb{X}_n$ be a random variable such that

$$\mathbb{P}(\mathbb{X}_n = k) = \frac{[z^n u^k]G(z,u)}{[z^n]G(z,1)}.$$

Under certain technical conditions, there exist constants μ and σ such that the normalized random variable

$$\mathbb{X}_n^* = \frac{\mathbb{X}_n - \mu n}{\sqrt{n\sigma^2}}$$

converges in distribution to a Gaussian variable with a speed of convergence $O(n^{-1/2})$.

Expected number of motifs

$$\begin{aligned}
\mathbb{E}(\text{Base pairs}) &\sim \frac{\alpha}{\gamma} n \\
\mathbb{E}(\text{Helices}) &\sim \frac{\alpha\beta}{\gamma} n \\
\mathbb{E}(\text{Hairpins}) &\sim \frac{\alpha\beta(1+p_1 q_2 \rho_0)}{4\gamma} n \\
\mathbb{E}(\text{Right bulges}) &\sim \frac{\alpha^2\beta}{4\gamma} n \\
\mathbb{E}(\text{Internal loops}) &\sim \frac{p_1 q_2 \rho_0 \alpha\beta}{4\gamma} n \\
\mathbb{E}(\text{Multi-branch loops}) &\sim \frac{\alpha\beta}{4\gamma} n
\end{aligned}$$

$\alpha = 1 - p_1 q_2 \rho_0$, $\beta = 1 - p_3 \rho_0^2$, $\gamma = 3 - p_1 q_2 \rho_0 - p_3 \rho_0^2 - p_1 p_3 q_2 \rho_0^3$

ρ_0 is the smallest positive root of $(1 - p_1 q_2 z)^2 (1 - p_3 z^2) - 4 p_2 q_1 q_2 q_3 z^3$

Interesting relations

Corollary

If $p_i, q_i > 0$, *then*

(i) $\mathbb{E}(\mathbb{X}_n^{lb}) = \mathbb{E}(\mathbb{X}_n^{rb})$,

(ii) $\mathbb{E}(\mathbb{X}_n^{m}) = \frac{1}{4}\mathbb{E}(\mathbb{X}_n^{hel})(1 + o(n))$,

(iii) $\mathbb{E}(\mathbb{X}_n^{hp}) = (\mathbb{E}(\mathbb{X}_n^{i}) + \mathbb{E}(\mathbb{X}_n^{m}))(1 + o(n))$,

(iv) $\mathbb{E}(\mathbb{X}_n^{m}) = (\mathbb{E}(\mathbb{X}_n^{lb}) + \mathbb{E}(\mathbb{X}_n^{i}))(1 + o(n))$,

Prediction using CYK

	No. Sequences (Type)	Av. Length	St. Dev. Length
Set I	122 (5S)	121.17	3.1
Set II	37 (16S)	956.46	6.51
Set III	81 (16S)	1521.33	24.86
Set IV	50 (16S)	1787.1	20.09
Set V	34 (23S)	2912.85	23.08

The transmission probabilities are

$$p_1 = 0.868534, p_2 = 0.105397, p_3 = 0.787640,$$

$$q_1 = 0.131466, q_2 = 0.894603, q_3 = 0.212360$$

and the emission probabilities are

	A	U	G	C
A	0.001167	0.177977	0.001058	0.001806
U	0.177977	0.002793	0.049043	0.000763
G	0.001058	0.049043	0.000406	0.266974
C	0.001806	0.000763	0.266974	0.000391

A	0.364097
U	0.273013
G	0.211881
C	0.151009

Comparison with native secondary structures

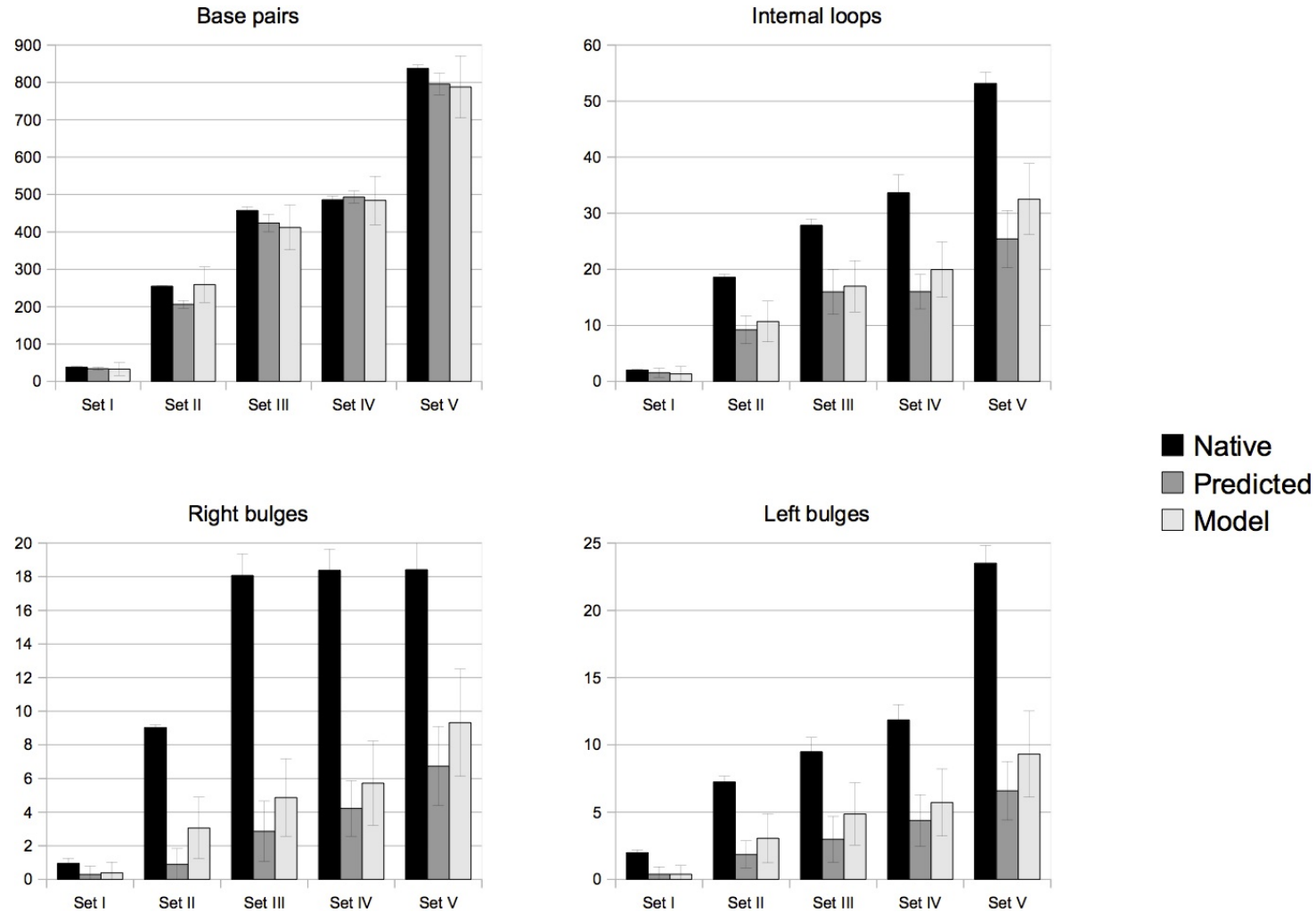


Figure : Motifs where there is agreement between the model expectations and the observed distribution in the most probable predicted structures.

Comparison with native secondary structures

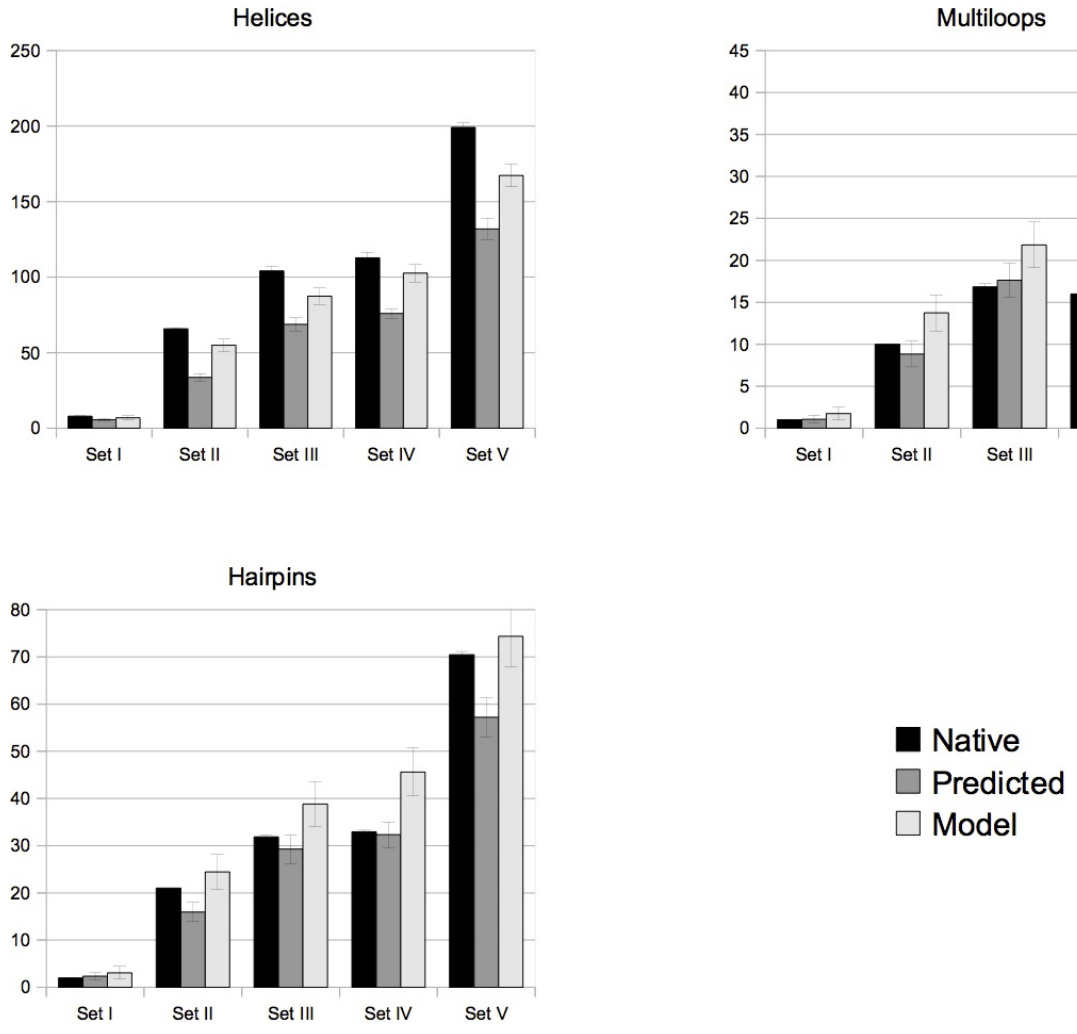


Figure : Motifs for which there is a larger discrepancy between the predicted structures and the model.

Prediction using CYK

		Ratios of Averages				
		$\frac{\text{RB}}{\text{LB}}$	$\frac{\text{Hel}}{\text{ML}}$	$\frac{\text{IL+ML}}{\text{HP}}$	$\frac{\text{IL+LB}}{\text{ML}}$	$\frac{\text{IL+RB}}{\text{ML}}$
Set I	Native	0.48	7.96	1.51	4.01	2.97
	Predicted	0.77	5.27	1.12	1.80	1.72
Set II	Native	1.25	6.58	1.36	2.58	2.76
	Predicted	0.50	3.80	1.13	1.24	1.14
Set III	Native	2.50	6.18	1.40	2.22	2.72
	Predicted	0.96	3.90	1.15	1.07	1.07
Set IV	Native	1.93	7.06	1.51	2.85	3.25
	Predicted	0.96	3.99	1.09	1.07	1.06
Set V	Native	0.78	5.92	1.23	2.28	2.13
	Predicted	1.02	3.67	1.07	0.89	0.89
Model		1	4	1	1	1

Ongoing work

- What is the effect of changing the emission probabilities?
- (with Heitsch and Greenwood) Why does this grammar give such nice ratios?

Example:
(slightly modified KH99 - strong performer (Anderson et al.))

$S \rightarrow LS|.|(F)$
$L \rightarrow .|(F)$
$F \rightarrow LS|(F)$

For this grammar, the ratio multiloops/helices is ≥ 4.

Thank you.