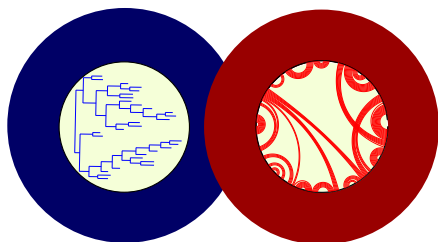


Isaac Newton Institute, 21st, December 2007

A Phylogenetic Definition of Structure

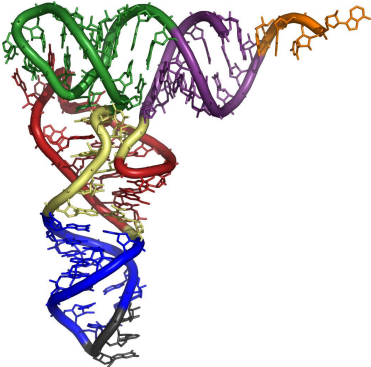
Accurate Background Models for Comparative Genomic Screens



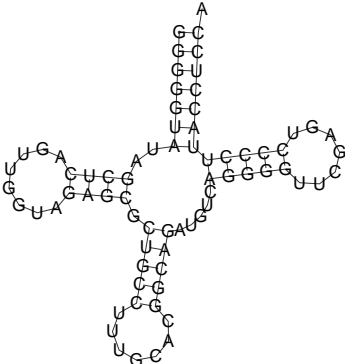
Tanja Gesell and Arndt von Haeseler
Center for Integrative Bioinformatics Vienna (CIBIV),
Max F. Perutz Laboratories (MFPL), Vienna, Austria

RNA Structure

Tertiary Structure



Secondary Structure

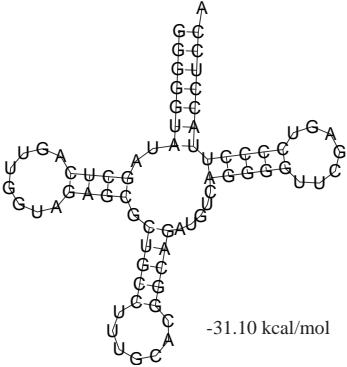
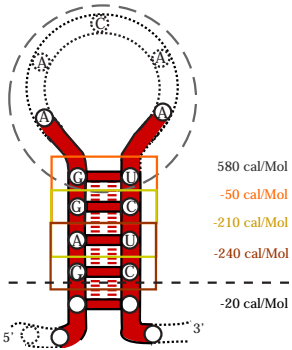


Primary Structure

GGGGUUAUAGCUCAGUUGGUAGAGCGCUGCCUUUGCACGGCAGAUGUCAGGGUUCGAGUCCCCUACCUCCA

RNA Structure

Minimum Free Energy Secondary Structure



The standard energy model expresses the free energy of a secondary structure \mathcal{S} as the sum of the energies of its components L :

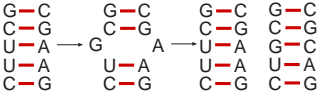
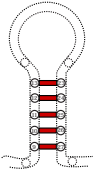
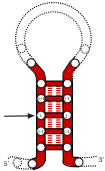
$$E(\mathcal{S}) = \sum_{L \in \mathcal{S}} E(L)$$

RNA Structure

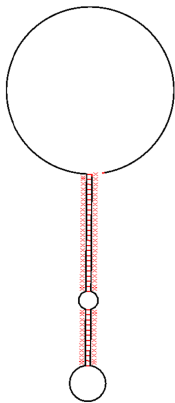
Consensus Structure

((((((...(((.....))))).((((.....)))))).....((((.....)))))))).
 GTTT**CC**GTAGTGTAGCGGTTATCACAT**TC**GCCTCACACGC**GAA**AGGTCCCCGGTTCGATCCCCGGGCG**GAA**ACA
 GTTT**CC**GTAGTGTAGTGGTTATCACGT**TC**GCCTAACACGC**GAA**AGGTCCCCGGTTCGAAACCGGGCG**GAA**ACA
 GTTT**TC**GTAGTGTAGTGGTTATCACGT**GT**GCTTCACACGC**ACA**AGGTCCCCGGTTCGAACCCGGGCG**A**AAACA

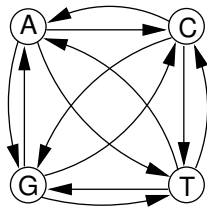
- 1) Energy contributions of the single sequences are averaged
- 2) Covariance information (e.g. compensatory mutations)



A Phylogenetic Definition of Structure

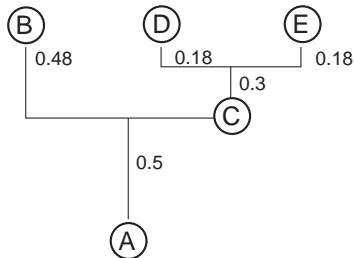


\mathcal{N} ighbourhood
System



$$\mathbf{P}(t) = \exp(\mathbf{Q}t)$$

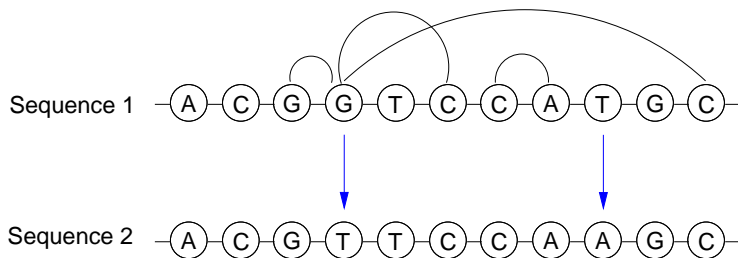
Model



Phylogenetic T ree

Neighbourhood system

$k = 1, \dots, l$ sites in a (nucleotide) sequence $\mathbf{x} = (x_1, \dots, x_l)$

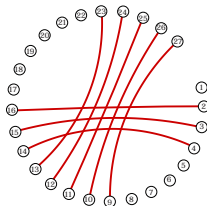
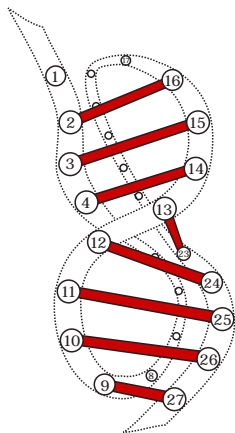


Neighbourhood system $\mathcal{N} = (N_k)_{k=1,2,\dots,l}$:

1. $N_k \subset \{1, \dots, l\}$, $k \notin N_k$ for each k
2. If $i \in N_k$ then $k \in N_i$ for each i, k .

n_k denotes the cardinality of N_k .

Example: \mathcal{N} (Pseudoknot)



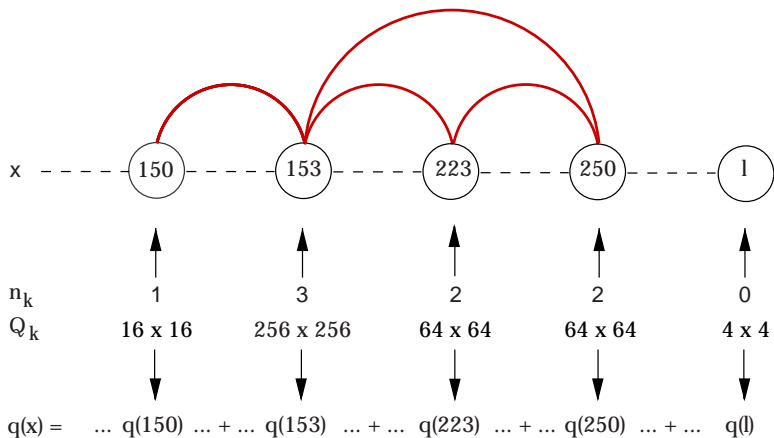
$N_1 = \{\}$
 $N_2 = \{16\}$
 $N_3 = \{15\}$
 $N_4 = \{14\}$
 $N_5 = \{\}$
 $N_6 = \{\}$
 $N_7 = \{\}$
 $N_8 = \{\}$
 $N_9 = \{27\}$

$N_{10} = \{26\}$
 $N_{11} = \{25\}$
 $N_{12} = \{24\}$
 $N_{13} = \{23\}$
 $N_{14} = \{4\}$
 $N_{15} = \{3\}$
 $N_{16} = \{2\}$
 $N_{17} = \{\}$
 $N_{18} = \{\}$

$N_{19} = \{\}$
 $N_{20} = \{\}$
 $N_{21} = \{\}$
 $N_{22} = \{\}$
 $N_{23} = \{13\}$
 $N_{24} = \{12\}$
 $N_{25} = \{11\}$
 $N_{26} = \{10\}$
 $N_{27} = \{9\}$

A model, that represents a universal description of arbitrary complex dependencies among sites.

Modeling Sequence Evolution with Arbitrary Dependencies



Basic idea: Different substitution matrix for each site

→ only one mutation is allowed at the current site.

$$n_k = 1$$

	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
AA	*	-	-	-	π_{CA}	-	-	-	π_{GA}	-	-	-	π_{UA}	-	-	-
AC	-	*	-	-	-	π_{CC}	-	-	-	π_{GC}	-	-	-	π_{UC}	-	-
AG	-	-	*	-	-	-	π_{CG}	-	-	-	π_{GG}	-	-	-	π_{UG}	-
AU	-	-	-	*	-	-	-	π_{CU}	-	-	-	π_{GU}	-	-	-	π_{UU}
CA	π_{AA}	-	-	-	*	-	-	-	π_{GA}	-	-	-	π_{UA}	-	-	-
CC	-	π_{AC}	-	-	-	*	-	-	-	π_{GC}	-	-	-	π_{UC}	-	-
CG	-	-	π_{AG}	-	-	-	*	-	-	-	π_{GG}	-	-	-	π_{UG}	-
CU	-	-	-	π_{AU}	-	-	-	*	-	-	-	π_{GU}	-	-	-	π_{UU}
GA	π_{AA}	-	-	-	π_{CA}	-	-	-	*	-	-	-	π_{UA}	-	-	-
GC	-	π_{AC}	-	-	-	π_{CC}	-	-	-	*	-	-	-	π_{UC}	-	-
GG	-	-	π_{AG}	-	-	-	π_{CG}	-	-	-	*	-	-	-	π_{UG}	-
GU	-	-	-	π_{AU}	-	-	-	π_{CU}	-	-	-	*	-	-	-	π_{UU}
UA	π_{AA}	-	-	-	π_{CA}	-	-	-	π_{GA}	-	-	-	*	-	-	-
UC	-	π_{AC}	-	-	-	π_{CC}	-	-	-	π_{GC}	-	-	-	*	-	-
UG	-	-	π_{AG}	-	-	-	π_{CG}	-	-	-	π_{GG}	-	-	-	*	-
UU	-	-	-	π_{AU}	-	-	-	π_{CU}	-	-	-	π_{GU}	-	-	-	*

$$n_k = 1$$

(k, i)	AA CA GA UA	AC CC GC UC	AG CG GG UG	AU CU GU UU
AA	* π_{CA} π_{GA} π_{UA}	- - - -	- - - -	- - - -
CA	π_{AA} * π_{GA} π_{UA}	- - - -	- - - -	- - - -
GA	π_{AA} π_{CA} * π_{UA}	- - - -	- - - -	- - - -
UA	π_{AA} π_{CA} π_{GA} *	- - - -	- - - -	- - - -
AC	- - - -	* π_{CC} π_{GC} π_{UC}	- - - -	- - - -
CC	- - - -	π_{AC} * π_{GC} π_{UC}	- - - -	- - - -
GC	- - - -	π_{AC} π_{CC} * π_{UC}	- - - -	- - - -
UC	- - - -	π_{AC} π_{CC} π_{GC} *	- - - -	- - - -
AG	- - - -	- - - -	* π_{CG} π_{GG} π_{UG}	- - - -
CG	- - - -	- - - -	π_{AG} * π_{GG} π_{UG}	- - - -
GG	- - - -	- - - -	π_{AG} π_{CG} * π_{UG}	- - - -
UG	- - - -	- - - -	π_{AG} π_{CG} π_{GG} *	- - - -
AU	- - - -	- - - -	- - - -	* π_{CU} π_{GU} π_{UU}
CU	- - - -	- - - -	- - - -	π_{AU} * π_{GU} π_{UU}
GU	- - - -	- - - -	- - - -	π_{AU} π_{CU} * π_{UU}
UU	- - - -	- - - -	- - - -	π_{AU} π_{CU} π_{GU} *

$$n_k = 1$$

$$\begin{array}{c} \mathbf{A|A} \\ \mathbf{C|A} \\ \mathbf{G|A} \\ \mathbf{U|A} \end{array} \begin{pmatrix} * & \pi_{CA} & \pi_{GA} & \pi_{UA} \\ \pi_{AA} & * & \pi_{GA} & \pi_{UA} \\ \pi_{AA} & \pi_{CA} & * & \pi_{UA} \\ \pi_{AA} & \pi_{CA} & \pi_{GA} & * \end{pmatrix} \quad \begin{array}{c} \mathbf{A|C} \\ \mathbf{C|C} \\ \mathbf{G|C} \\ \mathbf{U|C} \end{array} \begin{pmatrix} * & \pi_{CC} & \pi_{GC} & \pi_{UC} \\ \pi_{AC} & * & \pi_{GC} & \pi_{UC} \\ \pi_{AC} & \pi_{CC} & * & \pi_{UC} \\ \pi_{AC} & \pi_{CC} & \pi_{GC} & * \end{pmatrix}$$

$$\begin{array}{c} \mathbf{A|G} \\ \mathbf{C|G} \\ \mathbf{G|G} \\ \mathbf{U|G} \end{array} \begin{pmatrix} * & \pi_{CG} & \pi_{GG} & \pi_{UG} \\ \pi_{AG} & * & \pi_{GG} & \pi_{UG} \\ \pi_{AG} & \pi_{CG} & * & \pi_{UG} \\ \pi_{AG} & \pi_{CG} & \pi_{GG} & * \end{pmatrix} \quad \begin{array}{c} \mathbf{A|U} \\ \mathbf{C|U} \\ \mathbf{G|U} \\ \mathbf{U|U} \end{array} \begin{pmatrix} * & \pi_{CU} & \pi_{GU} & \pi_{UU} \\ \pi_{AU} & * & \pi_{GU} & \pi_{UU} \\ \pi_{AU} & \pi_{CU} & * & \pi_{UU} \\ \pi_{AU} & \pi_{CU} & \pi_{GU} & * \end{pmatrix}$$

Generally: $4^{n_k+1} \times 4^{n_k+1} \rightarrow 4^{n_k}$ Submatrices

$$\begin{array}{l} A|y_1, \dots, y_{n_k} \\ C|y_1, \dots, y_{n_k} \\ G|y_1, \dots, y_{n_k} \\ U|y_1, \dots, y_{n_k} \end{array} \left(\begin{array}{cccc} & * & \pi_{C|y_1, \dots, y_{n_k}} & \pi_{G|y_1, \dots, y_{n_k}} & \pi_{U|y_1, \dots, y_{n_k}} \\ \pi_{A|y_1, \dots, y_{n_k}} & & * & \pi_{G|y_1, \dots, y_{n_k}} & \pi_{U|y_1, \dots, y_{n_k}} \\ \pi_{A|y_1, \dots, y_{n_k}} & & \pi_{C|y_1, \dots, y_{n_k}} & * & \pi_{U|y_1, \dots, y_{n_k}} \\ \pi_{A|y_1, \dots, y_{n_k}} & & \pi_{C|y_1, \dots, y_{n_k}} & \pi_{G|y_1, \dots, y_{n_k}} & * \end{array} \right)$$

Generally: $4^{n_k+1} \times 4^{n_k+1} \rightarrow 4^{n_k}$ Submatrices

$$\begin{array}{l}
 A|y_1, \dots, y_{n_k} \\
 C|y_1, \dots, y_{n_k} \\
 G|y_1, \dots, y_{n_k} \\
 U|y_1, \dots, y_{n_k}
 \end{array}
 \left(
 \begin{array}{cccc}
 A|y_1, \dots, y_{n_k} & C|y_1, \dots, y_{n_k} & G|y_1, \dots, y_{n_k} & U|y_1, \dots, y_{n_k} \\
 * & \beta \pi_{C|y_1, \dots, y_{n_k}} & \alpha \pi_{G|y_1, \dots, y_{n_k}} & \beta \pi_{U|y_1, \dots, y_{n_k}} \\
 \beta \pi_{A|y_1, \dots, y_{n_k}} & * & \beta \pi_{G|y_1, \dots, y_{n_k}} & \alpha \pi_{U|y_1, \dots, y_{n_k}} \\
 \alpha \pi_{A|y_1, \dots, y_{n_k}} & \beta \pi_{C|y_1, \dots, y_{n_k}} & * & \beta \pi_{U|y_1, \dots, y_{n_k}} \\
 \beta \pi_{A|y_1, \dots, y_{n_k}} & \alpha \pi_{C|y_1, \dots, y_{n_k}} & \beta \pi_{G|y_1, \dots, y_{n_k}} & *
 \end{array}
 \right)$$

$$Q = \{Q_k | k = 1, \dots, l\}$$

$$Q_k^\gamma(\mathbf{s}_k, \mathbf{y}) = \begin{cases} \gamma(\mathbf{s}_k, \mathbf{y}) \cdot Q_k(\mathbf{s}_k, \mathbf{y}) & \text{if } H(\mathbf{s}_k, \mathbf{y}) = 1 \text{ and } x_k \neq y_0 \\ - \sum_{\substack{\mathbf{z} \in \mathcal{A}^{n_k+1} \\ \mathbf{z} \neq \mathbf{s}_k}} Q_k^\gamma(\mathbf{s}_k, \mathbf{z}) & \text{if } H(\mathbf{s}_k, \mathbf{y}) = 0 \\ 0 & \text{otherwise} \end{cases}$$

with $\mathbf{s}_k = (x_k, x_{i_1}, \dots, x_{i_{n_k}}) \in \mathcal{A}^{n_k+1}$, where $\{i_1, \dots, i_{n_k}\} = N_k$
 $\mathbf{y} = (y_0, y_1, \dots, y_{n_k}) \in \mathcal{A}^{n_k+1}$

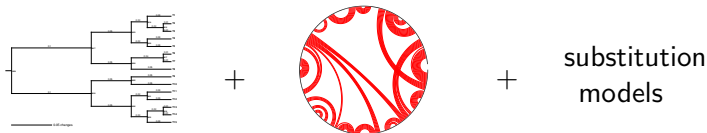
$\gamma(\mathbf{s}_k, \mathbf{y}) > 0$ neighbourhood constraints,
 e.g. as a function of energy values ΔG^0 .

$Q_k(\mathbf{s}_k, \mathbf{y})$ rates given a chosen nullmodel

Normalisation:

$$d_k = - \sum_{\mathbf{z} \in \mathcal{A}^{n_k+1}} \pi_k(\mathbf{z}) \cdot Q_k(\mathbf{z}, \mathbf{z}) = 1.$$

(Gesell, T. and von Haeseler, A. (2006) ,*Bioinformatics*, 22, 716-722)

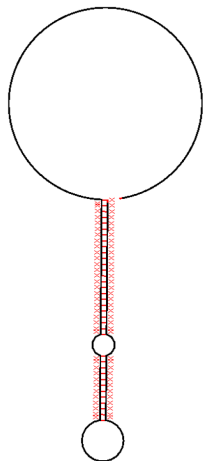


15 401

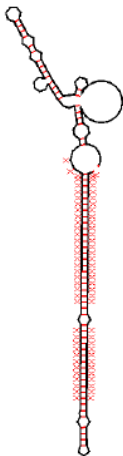
```

T1      AGACGGUCUGGUUUGCGGGGGUGAUCACGACGAACGGUCGUGAUUGCCUUAAGGCCGGUGGGCCUUGGUC AAGUCAGAUGAGCUC
T3      AGACGGUCUGGUUUGCGGGGGUGAUUACGACGAACGGUCGUGAUUGCCUUAAGGCCGGUGGGCCUUGGUC AAGUCGGAUGAGCUC
T2      AGACGGUCUGGUUUGCGGGGGUGAUCACGACGAACGGUCGUGAUUGCCUUAAGGCCGGUGGGCCUUGGUC AAGUCGGAUAAGCUC
T4      AGACGGUCUGGUUUGCGGGGGUGAUCACGGCGAACGGUCGUGAUUGCCUUAACCGCAGGUGGGCCUAGGUC AAGUCGGAUGAGCUC
T5      AGACGGUCUGGUUUGCGGGGGUGAUCACGACGAACGGUCGUGAUUGCCUUAACCGCAGGUGGGCCUAGGUC AAAUCGGACGAGCUC
T6      GGGCGGUCUGGUUAUGGGGGUGAUCACGGCGAACGGCCGUGAUGGCCUUAAGGAGGUUAGCCUGAGUUGAGUC GGAUUAGGUC
T7      GGGCGGUCUGGUUAUGGGGGUGAUCACGGCGAACGGCCGUGAUGGCCUUAAGGAGGUUUGGCCUUAAGUUGAGUC GGAUUAGGUC
T8      GGGCGGUCUGGUUAUGGGGGUGAUCACGGCGAACGGCCGUGAUGGCCUUAAGGAGGUUUGGCCUUAAGUUCAGUC GGAUUUGGUC
T9      CUAUGGUCUGGUUACGGGGGGUGAUCUUGGCGGGCAGCCGUGAUUGCCUGUGCAGGUGGGUUUAAGUUUAGUAGAAUUGAGUC
T10     CUAUGGUCUGGUUACGGGGGGUGAUCUUGGCGGGCCCGCCGUGAUCGCCUGUGCAGGUGGGUCUAAUUUUAGUCGAAUUGGGCG
T11     CUAUGGCCUGGUUACGGGGGGUGAUCUUGGUGGGCGGUCGUGAUUCCGUGUGCAGGUGGGUCUAAAGUUUAGGCGGAUUGGGCG
T12     CUAUGGUCUGGUUACGGGGGGUGAUCUUGGUGGGCGGCCGUGAUUCCGUGUGCAGAUUGGGUCAAGUUUAGGCGGAUUGGGCG
T13     CUAUGGUCUGGUUACGGGGGGUGAUCACGGUGGGCGACCCGUGAUUCCGUGUGCAGGUGGGUCUAAAGUUUAGGCGAAUUGGGCG
T14     CUAUGGUCUGGUUACGGGGGGUGAUCUUGGUGGGCGGCCGUGAUUCCGUGUGCAGGUGGGUCUAAAGUUUAGGCGAAUUGGGCG
T15     CUAUGGUCUGGUUACGGGGGGUGAUCUUGGUGAGCGGCCAUGAUUCCGUGUGCAGUUGGGUCUAAAGUUUAGGCGAAUUGGGCG
  
```

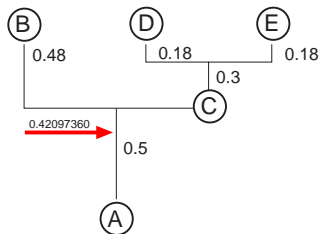

Phylogenetic RNAmovie



N



mfe



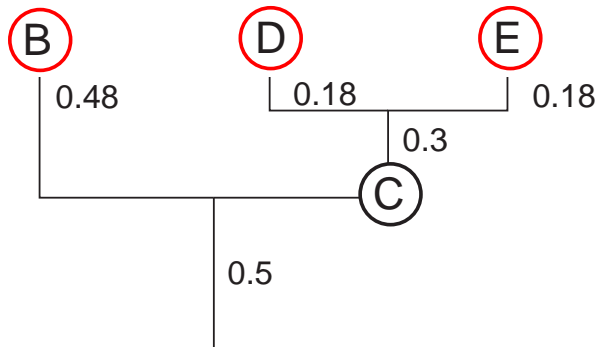
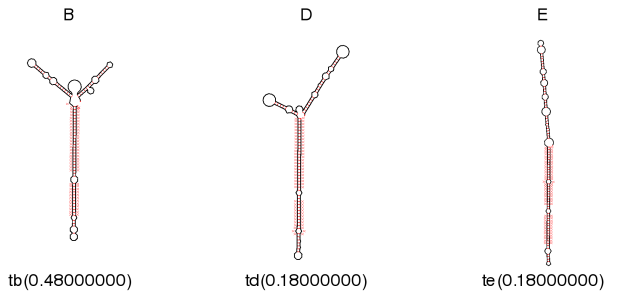
Movie

SISSI (Gesell, von Haeseler, 2006)

RNAfold (Hofacker et al., 1994)

RNAmovie (Evers, Giegerich, 1999)

Phylogenetic RNAmovie



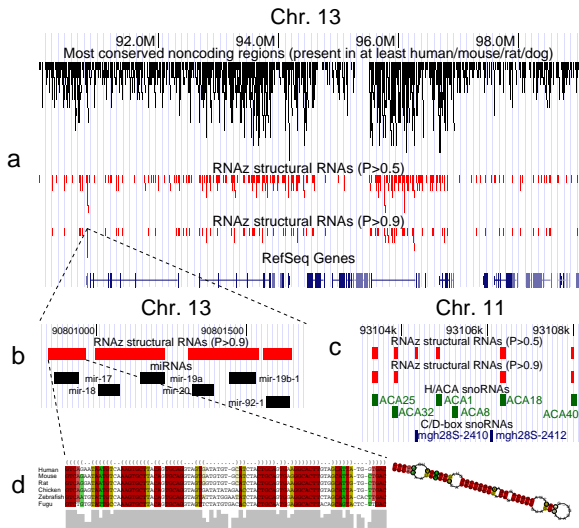
Applications of SISSI

- ▶ Energy landscape of RNA, Selection of RNA ...
- ▶ Performance of tree building method with dependencies
- ▶ Performance of structure prediction methods

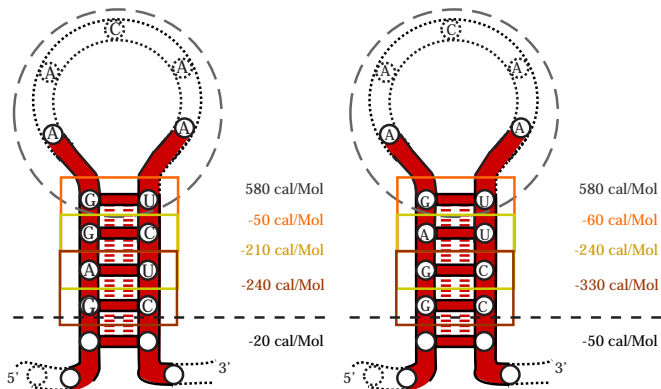
Applications of SSSI

- ▶ Energy landscape of RNA, Selection of RNA ...
- ▶ Performance of tree building method with dependencies
- ▶ Performance of structure prediction methods
- ▶ Accurate Background Models for Comparative Genomic Screens: **Applications to RNA Gene Prediction**
Joint work with Stefan Washietl

Maps of Structured ncRNAs in the Human Genome

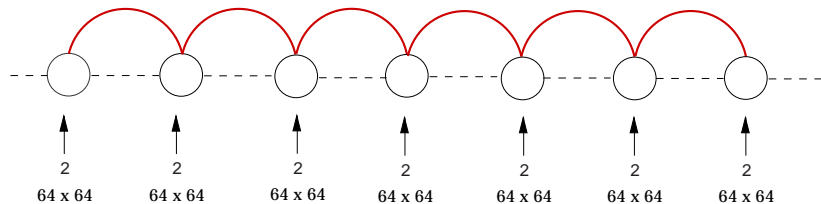


Dinucleotide Content Matters



How do we get dinucleotide controlled random alignments?

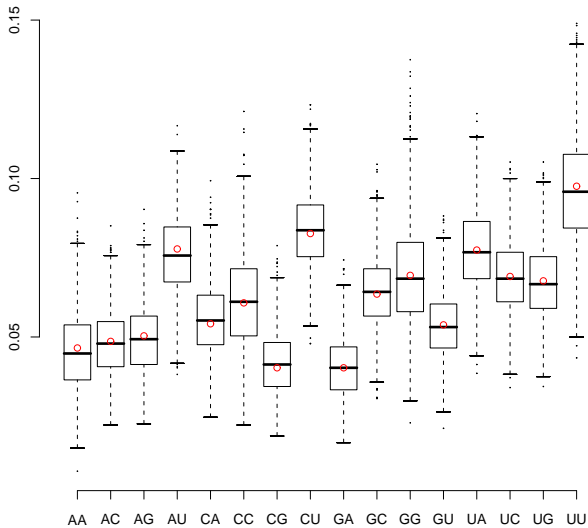
Neighbourhood System and Model for Dinucleotid Content



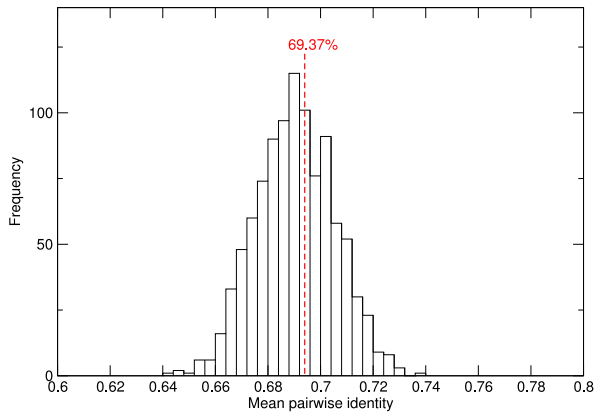
16 submatrices of the type:

$$\begin{array}{l}
 A|_{X_{k-1}, X_{k+1}} \\
 C|_{X_{k-1}, X_{k+1}} \\
 G|_{X_{k-1}, X_{k+1}} \\
 U|_{X_{k-1}, X_{k+1}}
 \end{array}
 \left(
 \begin{array}{cccc}
 A|_{X_{k-1}, X_{k+1}} & C|_{X_{k-1}, X_{k+1}} & G|_{X_{k-1}, X_{k+1}} & U|_{X_{k-1}, X_{k+1}} \\
 * & \beta\pi C|_{X_{k-1}, X_{k+1}} & \alpha\pi G|_{X_{k-1}, X_{k+1}} & \beta\pi U|_{X_{k-1}, X_{k+1}} \\
 \beta\pi A|_{X_{k-1}, X_{k+1}} & * & \beta\pi G|_{X_{k-1}, X_{k+1}} & \alpha\pi U|_{X_{k-1}, X_{k+1}} \\
 \alpha\pi A|_{X_{k-1}, X_{k+1}} & \beta\pi C|_{X_{k-1}, X_{k+1}} & * & \beta\pi U|_{X_{k-1}, X_{k+1}} \\
 \beta\pi A|_{X_{k-1}, X_{k+1}} & \alpha\pi C|_{X_{k-1}, X_{k+1}} & \beta\pi G|_{X_{k-1}, X_{k+1}} & *
 \end{array}
 \right)$$

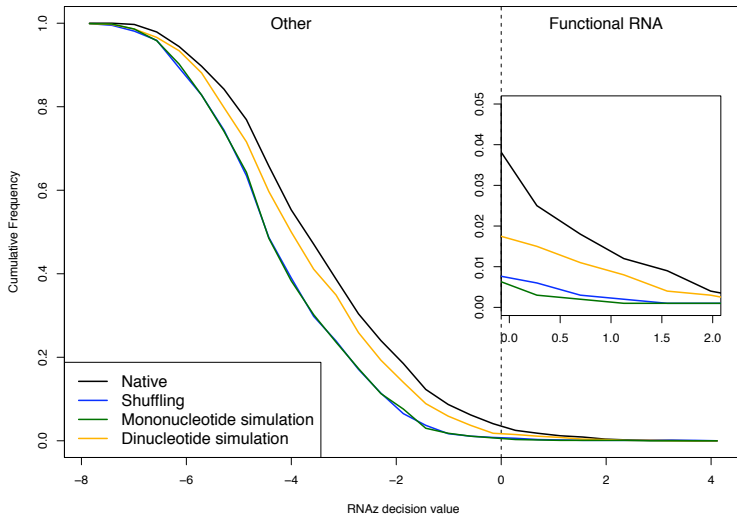
Results: Dinucleotide Content



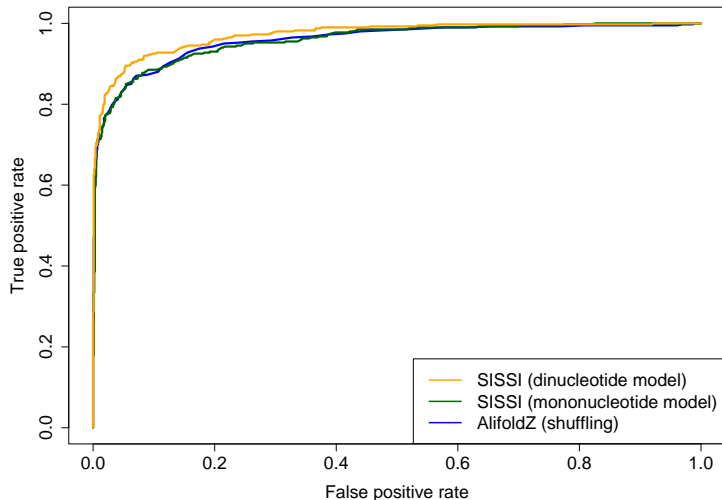
Results: Mean Pairwise Identity



Estimating False Positives for RNAz Gene Prediction



SISSlZ: Dinucleotide Based RNA Gene Finder

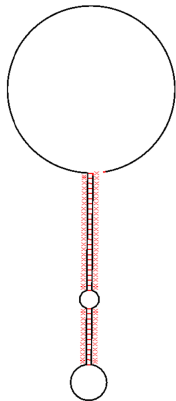


Summary

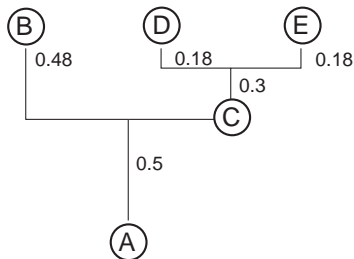
- ▶ We have presented a general framework and a model for arbitrary dependencies
- ▶ RNA gene predictions are biased by the genomic dinucleotide content.
- ▶ SSSI can be used to generate accurate null models giving more realistic estimation of false positives.
- ▶ SSSIz is introduced as a dinucleotide based gene finder (miRNA target search, transcription factor binding sites, ...)
- ▶ A phylogenetic definition of structure

Future Directions in Phylogenetic Methods and Models

A Phylogenetic Definition of Structure



MODEL

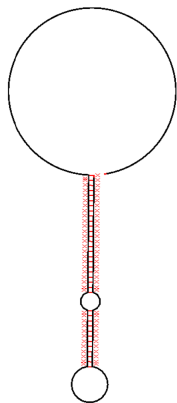


\mathcal{N}

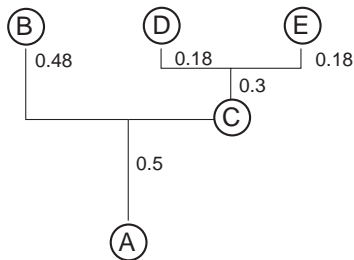
Defining lineage specific neighbourhood systems

Future Directions in Phylogenetic Methods and Models

A Phylogenetic Definition of Structure



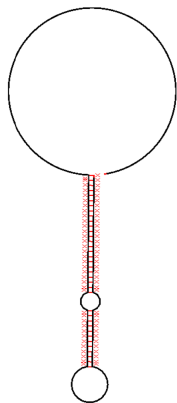
⇒ MODEL



\mathcal{N}

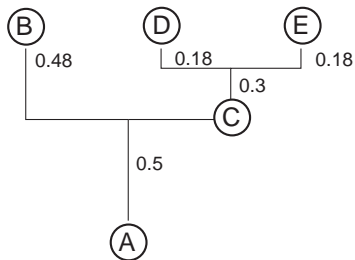
Future Directions in Phylogenetic Methods and Models

A Phylogenetic Definition of Structure

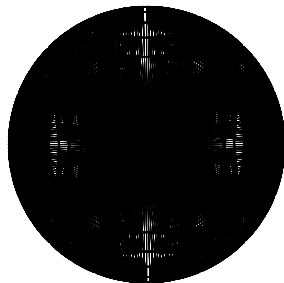


\mathcal{N}

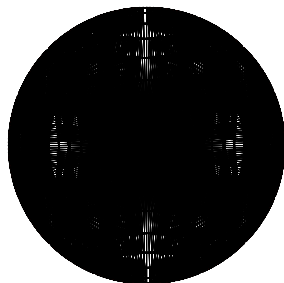
⇒ MODEL ⇐



...and Further Animations on the Sequence Space



...and Further Animations on the Sequence Space



28/29 April 2008: Discussion Meeting organised by Nick Goldman and Ziheng Yang at the Royal Society in London.

Acknowledgements



Thanks to:

Arndt von Haeseler
(CIBIV, MFPL, Vienna)

Ivo Hofacker & Stefan Washietl
(Theoretical Chemistry, Vienna)

Goldman Group at the EBI
(Carolin Kosiol & Simon Whelan)

Special thanks to Oswald Wiener
(Kunstakademie Düsseldorf)

Acknowledgements



Thanks to:

Arndt von Haeseler
(CIBIV, MFPL, Vienna)

Ivo Hofacker & Stefan Washietl
(Theoretical Chemistry, Vienna)

Goldman Group at the EBI
(Carolin Kosiol & Simon Whelan)

Special thanks to Oswald Wiener
(Kunstakademie Düsseldorf)

Thank you for your attention!