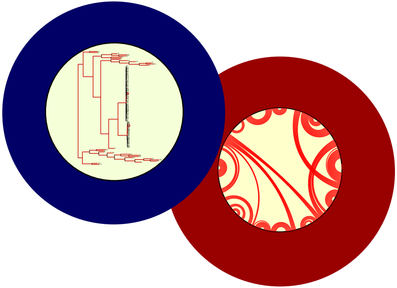
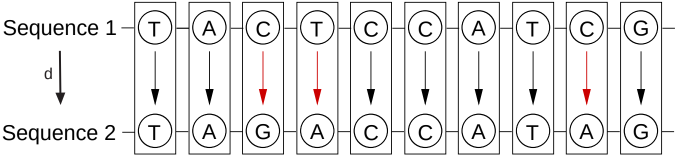
A phylogenetic view on RNA structure evolution

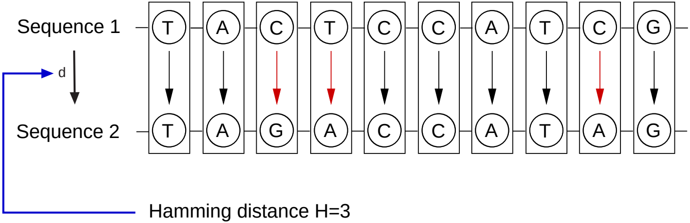


Tanja Gesell, Bioinformatics Institute, Heinrich-Heine University
Düsseldorf, Germany - February 2006, Bled, Slovenia,

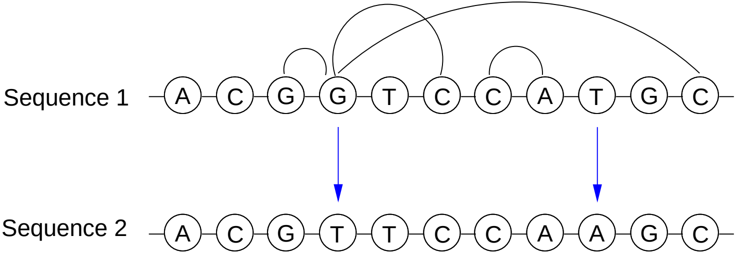
Modeling sequence evolution



Modeling sequence evolution

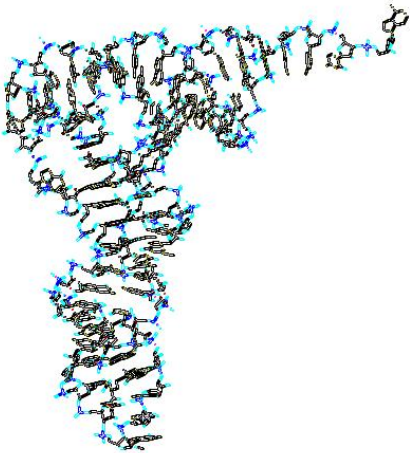


Modeling sequence evolution

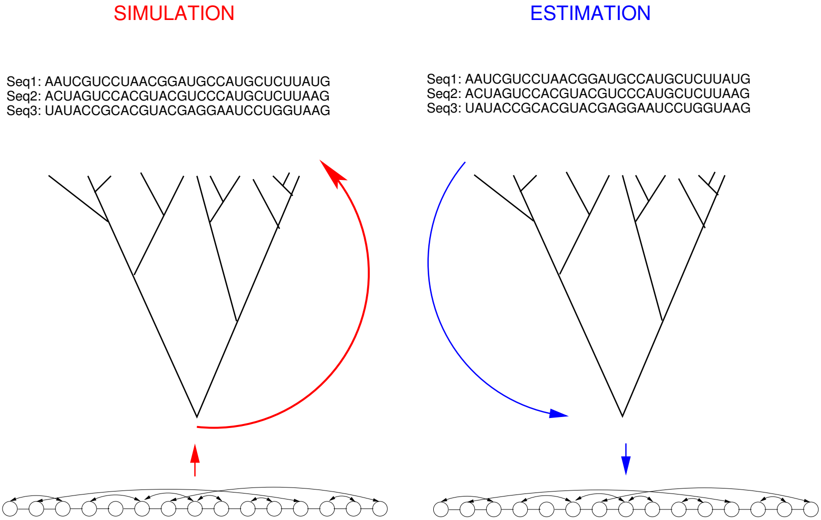


each sequence site does not evolve independently of the others

Example for site-specific interactions



- 2D-structure
- 3D-structure
- of RNAs, e.g. tRNA, mRNA ...
- of proteins
- CpG
- codon positions
- . . .



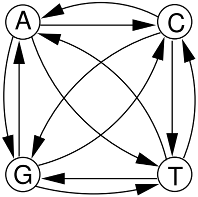
SIMULATION

Seq1: AAUCGUCCUAACGGAUGCCAUGCUCUUAUG
Seq2: ACUAGUCCACGUACGUCCCAUGCUCUUAAG
Seq3: UAUACCGCACGUACGAGGAAUCCUGGUAAG

ESTIMATION

Seq1: AAUCGUCCUAACGGAUGCCAUGCUCUUAUG
Seq2: ACUAGUCCACGUACGUCCCAUGCUCUUAAG
Seq3: UAUACCGCACGUACGAGGAAUCCUGGUAAG

Model-based approaches



stationary and time homogeneous Markov model
the probability that sequence x evolves to sequence y

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

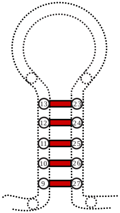
4x4 instantaneous rate matrix

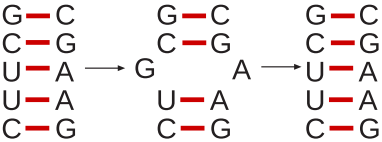
Generally Reversible (REV)

$$Q = \begin{pmatrix} -(a\pi_C + b\pi_G + c\pi_T) & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & -(a\pi_A + d\pi_G + e\pi_T) & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & -(b\pi_A + d\pi_C + f\pi_T) & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & -(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix}$$

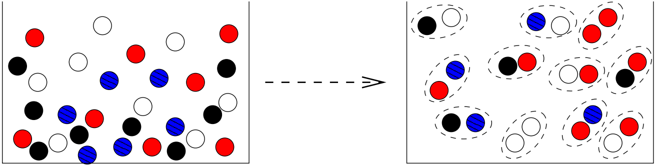
	A	C	G	T	A	C	G	T	A	C	G	T
	JC69-Modell				K80-Modell				HKY-Modell			
A	$*$	α	α	α	$*$	β	α	β	$*$	$\beta\pi_C$	$\alpha\pi_G$	$\beta\pi_T$
C	α	$*$	α	α	β	$*$	β	α	$\beta\pi_A$	$*$	$\beta\pi_G$	$\alpha\pi_T$
G	α	α	$*$	α	α	β	$*$	β	$\alpha\pi_A$	$\beta\pi_C$	$*$	$\beta\pi_T$
T	α	α	α	$*$	β	α	β	$*$	$\beta\pi_A$	$\alpha\pi_C$	$\beta\pi_G$	$*$
	TN93-Modell				F81-Modell				GTR-Modell			
A	$*$	$\beta\pi_C$	$\alpha_1\pi_G$	$\beta\pi_T$	$*$	π_C	π_G	π_T	$*$	$a\pi_C$	$b\pi_G$	$c\pi_T$
C	$\beta\pi_A$	$*$	$\beta\pi_G$	$\alpha_2\pi_T$	π_A	$*$	π_G	π_T	$a\pi_A$	$*$	$d\pi_G$	$e\pi_T$
G	$\alpha_1\pi_A$	$\beta\pi_C$	$*$	$\beta\pi_T$	π_A	π_C	$*$	π_T	$b\pi_A$	$d\pi_C$	$*$	$f\pi_T$
T	$\beta\pi_A$	$\alpha_2\pi_C$	$\beta\pi_G$	$*$	π_A	π_C	π_G	$*$	$c\pi_A$	$e\pi_C$	$f\pi_G$	$*$

Compensatory mutation





Nucleotides in stem regions evolve in strong correlation with their pairing counterpart.



Compensatory mutation

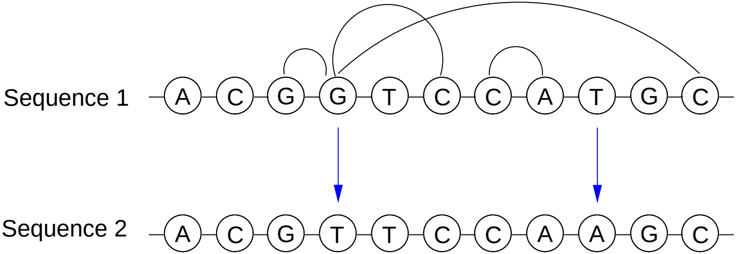
Schöniger and von Haeseler, 1994

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

How to simulate more complex interactions among nucleotide and other character based sequences?
A model, that represents a universal description of arbitrary complex dependencies among sites.

Neighbourhood system

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

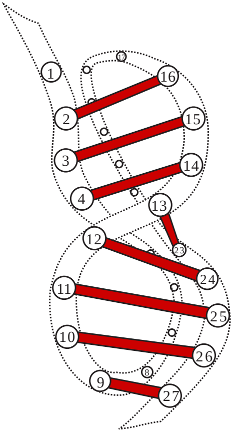


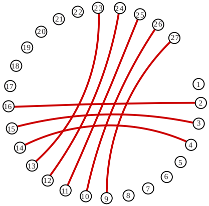
Neighbourhood system $\mathcal{N} = (N_k)_{k=1,2,\cdots,l}$:

1. $N_k \subset \{1, \ldots, l\}, k \notin N_k$ 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)



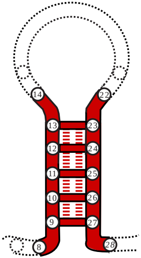


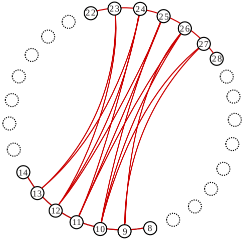
$N_1 = \{\}$
$N_2 = \{16\}$
$N_3 = \{15\}$
$N_4 = \{14\}$
$N_5 = \{\}$
$N_6 = \{\}$
$N_7 = \{\}$
$N_8 = \{\}$
$N_9 = \{27\}$

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

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

Example: $\mathcal{N}$(Stem including stacking)

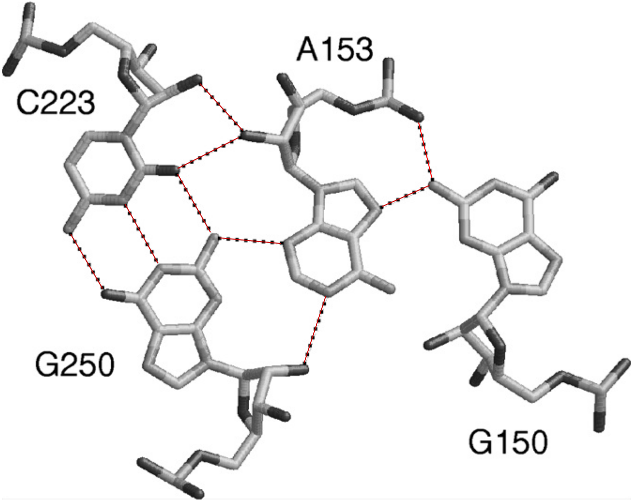




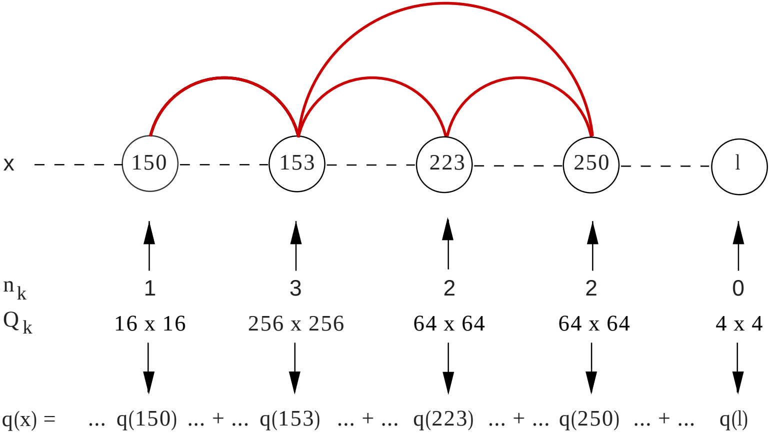
. . .
$N_8 = \{9\},$
$N_9 = \{8, 27, 10, 26\}$
$N_{10} = \{9, 27, 26, 25, 11\}$
$N_{11} = \{10, 26, 25, 24, 12\}$
$N_{12} = \{11, 25, 24, 23, 13\}$
$N_{13} = \{12, 23, 24, 14\}$
$N_{14} = \{13\}$
. . .

. . .
$N_{22} = \{23\}$
$N_{23} = \{22, 13, 12, 24\}$
$N_{24} = \{23, 13, 12, 11, 25\}$
$N_{25} = \{24, 11, 10, 12, 26\}$
$N_{26} = \{25, 9, 10, 11, 27\}$
$N_{27} = \{26, , 9, 10, 28\}$
$N_{28} = \{27\}$
. . .

Example: Ribozyme domain



Basic idea: Different substitution matrix for each site



x 150 153 223 250 l

n_k 1 3 2 2 0

Q_k 16 x 16 256 x 256 64 x 64 64 x 64 4 x 4

q(x) = ... q(150) ... + ... q(153) ... + ... q(223) ... + ... q(250) ... + ... q(l)

Only one mutation is allowed at the current site

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

	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}	-	-	-	*

(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}	*

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

$$\begin{array}{cc} & \begin{array}{cccc} \mathbf{A}|A & \mathbf{C}|A & \mathbf{G}|A & \mathbf{U}|A \end{array} \\ \begin{array}{c} \mathbf{A}|A \\ \mathbf{C}|A \\ \mathbf{G}|A \\ \mathbf{U}|A \end{array} & \begin{pmatrix} * & \pi_{\mathrm{CA}} & \pi_{\mathrm{GA}} & \pi_{\mathrm{UA}} \\ \pi_{\mathrm{AA}} & * & \pi_{\mathrm{GA}} & \pi_{\mathrm{UA}} \\ \pi_{\mathrm{AA}} & \pi_{\mathrm{CA}} & * & \pi_{\mathrm{UA}} \\ \pi_{\mathrm{AA}} & \pi_{\mathrm{CA}} & \pi_{\mathrm{GA}} & * \end{pmatrix} \end{array}$$

$$\begin{array}{cc} & \begin{array}{cccc} \mathbf{A}|C & \mathbf{C}|C & \mathbf{G}|C & \mathbf{U}|C \end{array} \\ \begin{array}{c} \mathbf{A}|C \\ \mathbf{C}|C \\ \mathbf{G}|C \\ \mathbf{U}|C \end{array} & \begin{pmatrix} * & \pi_{\mathrm{CC}} & \pi_{\mathrm{GC}} & \pi_{\mathrm{UC}} \\ \pi_{\mathrm{AC}} & * & \pi_{\mathrm{GC}} & \pi_{\mathrm{UC}} \\ \pi_{\mathrm{AC}} & \pi_{\mathrm{CC}} & * & \pi_{\mathrm{UC}} \\ \pi_{\mathrm{AC}} & \pi_{\mathrm{CC}} & \pi_{\mathrm{GC}} & * \end{pmatrix} \end{array}$$

$$\begin{array}{cc} & \begin{array}{cccc} \mathbf{A}|G & \mathbf{C}|G & \mathbf{G}|G & \mathbf{U}|G \end{array} \\ \begin{array}{c} \mathbf{A}|G \\ \mathbf{C}|G \\ \mathbf{G}|G \\ \mathbf{U}|G \end{array} & \begin{pmatrix} * & \pi_{\mathrm{CG}} & \pi_{\mathrm{GG}} & \pi_{\mathrm{UG}} \\ \pi_{\mathrm{AG}} & * & \pi_{\mathrm{GG}} & \pi_{\mathrm{UG}} \\ \pi_{\mathrm{AG}} & \pi_{\mathrm{CG}} & * & \pi_{\mathrm{UG}} \\ \pi_{\mathrm{AG}} & \pi_{\mathrm{CG}} & \pi_{\mathrm{GG}} & * \end{pmatrix} \end{array}$$

$$\begin{array}{cc} & \begin{array}{cccc} \mathbf{A}|U & \mathbf{C}|U & \mathbf{G}|U & \mathbf{U}|U \end{array} \\ \begin{array}{c} \mathbf{A}|U \\ \mathbf{C}|U \\ \mathbf{G}|U \\ \mathbf{U}|U \end{array} & \begin{pmatrix} * & \pi_{\mathrm{CU}} & \pi_{\mathrm{GU}} & \pi_{\mathrm{UU}} \\ \pi_{\mathrm{AU}} & * & \pi_{\mathrm{GU}} & \pi_{\mathrm{UU}} \\ \pi_{\mathrm{AU}} & \pi_{\mathrm{CU}} & * & \pi_{\mathrm{UU}} \\ \pi_{\mathrm{AU}} & \pi_{\mathrm{CU}} & \pi_{\mathrm{GU}} & * \end{pmatrix} \end{array}$$

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

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

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

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

Normalisation:

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

The total instantaneous substitution rate for x:

$$q(\mathbf{x}) \quad = \quad \sum_{k=1}^{l} |\, Q_k(\mathbf{s}_k, \mathbf{s}_k)\,|$$

Relative mutability at site k:

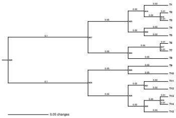
$$\mathbb{P}(k) \quad = \quad \frac{|\, Q_k(\mathbf{s}_k, \mathbf{s}_k)\,|}{q(\mathbf{x})}$$

Probability to replace x_k by y_0:

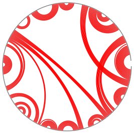
$$\mathbb{P}(x_k \rightarrow y_0) \quad = \quad \frac{Q_k(\mathbf{s}_k, \mathbf{y})}{|\, Q_k(\mathbf{s}_k, \mathbf{s}_k)\,|}$$

SISSI:

SImulating Sequence Evolution with Site-Specific Interactions
(Gesell and von Haeseler, Bioinformatics in press, Epub. 2005 Dec. 6)



+



+

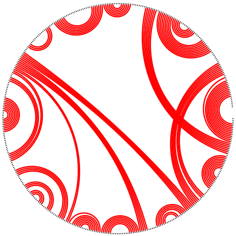
substitution model

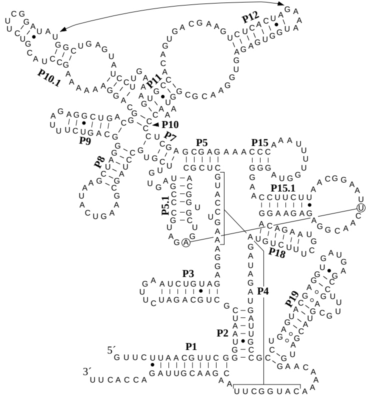
↓

```
 15 401
T1         AGACGGUCUGGUUGCGGGGGUGAUCACGACGAACGGUCGUGAUUGCCUUAGGCCGGUGGGCCUUGGUCAAGUCAGAUGAGCUC
T3         AGACGGUCUGGUUGCGGGGGUGAUUACGACGAACGGUCGUGAUUGCCUAAGGCCGGUGGGCCUUGGUCAAGUCGGAUGAGCUC
T2         AGACGGUCUGGUUGCGGGGGUGAUCACGACGAACGGUCGUGAUUGCCUAAGGCCGGUGGGCCUUGGUCAAGUCGGAUAAGCUC
T4         AGACGGUCUGGUUGCGGGGGUGAUCACGGCGAACGGUCGUGAUUGCCUACCGCAGGUGGGCCUAGGUCAAGUCGGAUGAGCUC
T5         AGACGGUCUGGUUGCGGGGGUGAUCACGACGAACGGUCGUGAUUGCCUAACGCAGGUGGGCCUAGGUCAAAUCGGACGAGCUC
T6         GGGCGGUCUGGUUAUGGGGGUGAUCACGGCGAACGGCCGUGAUGGCCUAAGGGAGGUUAGCCUGAGUUGAGUCGGAUUAGGUC
T7         GGGCGGUCUGGUUAUGGGGGUGAUCACGGCGAACGGCCGUGAUGGCCUAAGGGAGGUUGGCCUAAGUUGAGUCGGAUUAGGUC
T8         GGGCGGUCUGGUUAUGGGGGUCAUCACGGCGAACGGCCGUGAUGGCCUAAGGGAGGUUGGCCUAAGUUCAGUCGGAUUUGGUC
T9         CUAUGGUCUGGUUACGGGGGUGAUCAUGGCGGGCAGCCGUGAUUGCCGUGUGCAGGUGGGUUUAAGUUUAGUAGAAUUAGUGC
T10        CUAUGGUCUGGUUACGGGGGUGAUCAUGGCGGGCGCCCGUGAUCGCCGUGUGCAGGUGGGUCUAAUUUUAGUCGAAUUGGCGC
T11        CUAUGGCCUGGUUACGGGGGUGAUCAUGGUGGGCGGUCGUGAUUGCCGUGUGCAGGUGGGUCUAAGUUUAGGCGGAUUGGCGC
T12        CUAUGGUCUGGUUACGGGGGUGAUCAUGGUGGGCGGCCGUGAUUGCGGUGUGCAGAUGGGUCCAAGUUUAGGCGGAUUGGCGC
T13        CUAUGGUCUGGUUACGGGGGUGAUCACGGUGGGCGACCGUGAUUGCCCUGUGCAGGUGGGUCUAAGUUUAGGCGAAUUGGCGU
```

Example: *Bacillus subtilis*

RNase P database (Brown, 1999)



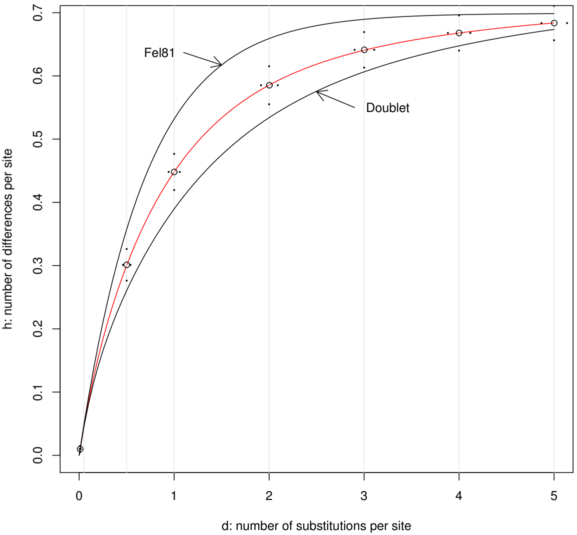


Sequence M13175, image created by Brown

Example: Counted frequencies from a RNase P sequence of *Bacillus subtilis* taken from the RNase P database:

	$f_{n_k=1}$				$f_{n_k=0}$
	second nucleotide in doublet				
first nucleotide	A	C	G	U	
A	0.000423	0.004228	0.012685	0.169133	0.422360
C	0.004228	0.000423	0.262156	0.000423	0.105590
G	0.012685	0.262156	0.000423	0.042283	0.236025
U	0.169133	0.000423	0.042283	0.016915	0.236025

Relationship between number of substitutions per site d and number of observed differences per site h:



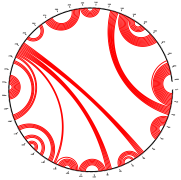
A pilot study of SISSI

Does phylogeny matter?

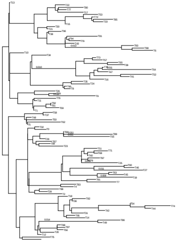
A phylogenetic view on some existing structure prediction methods.

Influence of the tree topology

Examples with 5 bifurcating trees, with the same topology, but different mean branch length.



$+$



$+$ substitution-model $\longrightarrow$

$T_{0.03}$, $T_{0.075}$, $T_{0.1}$, $T_{0.3}$, $T_{0.5}$

ConStruct

Construction of RNA consensus structures (Lück et al. 1999), (Wilm, A. & Steger, G. 2006, submitted)

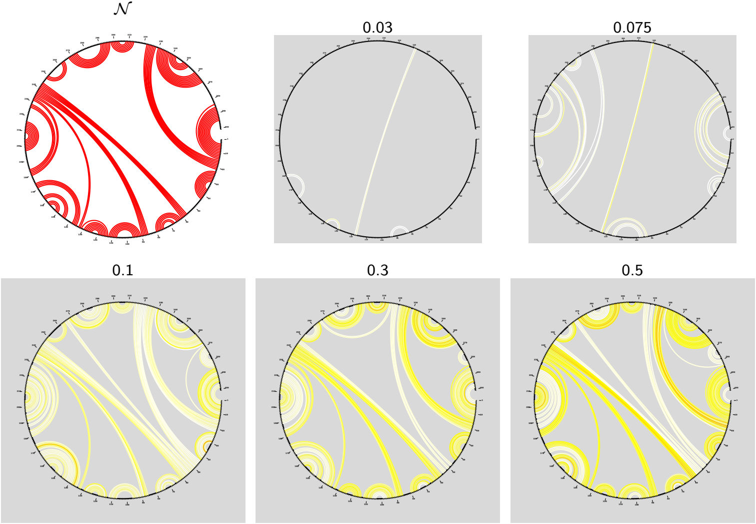
Combination of Sequence Alignment, Thermodynamics and Mutual Information Content.

Consensus Structure

- Thermodynamic Consensus Dotplot:
  Consensus Dotplot using RNAfold: Hofacker et al. (1994)
- Mutual Information Content: MIC:
  (Chiu & Kolodziejczak, 1991, Gutell et al. 1992)
- Prediction of Tertiary Interaction
  Maximum Weighted Matching: Tabaska et al. (1998)

Mean branch length

Consensus Structures based on Mutual Information Content



$\mathcal{N}$

0.03

0.075

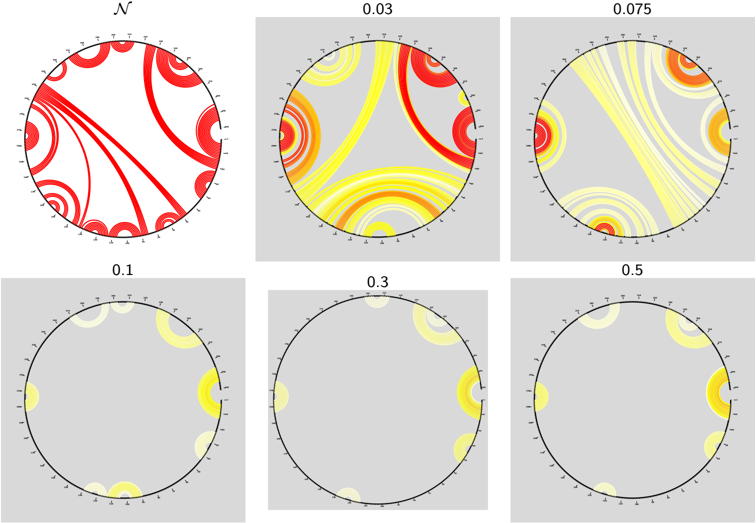
0.1

0.3

0.5

Mean branch length

Thermodynamic Consensus Structures



$\mathcal{N}$

0.03

0.075

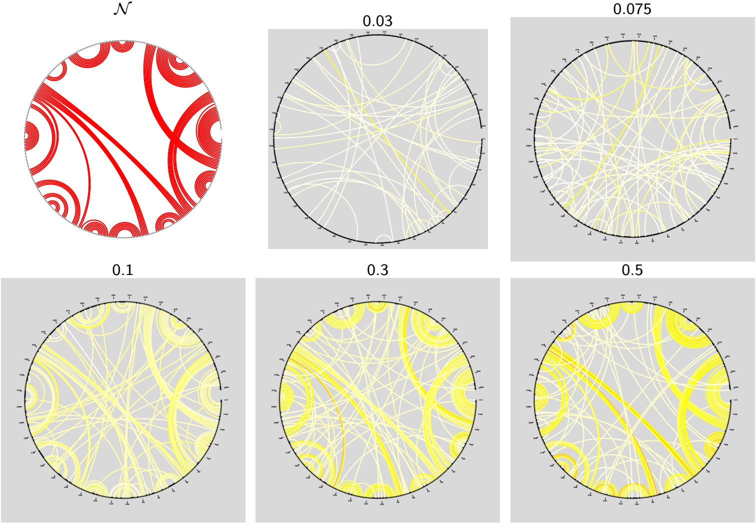
0.1

0.3

0.5

Prediction of tertiary interactions

Using Maximum Weighted Matching: Tabaska et al. (1998)



$\mathcal{N}$

0.03

0.075

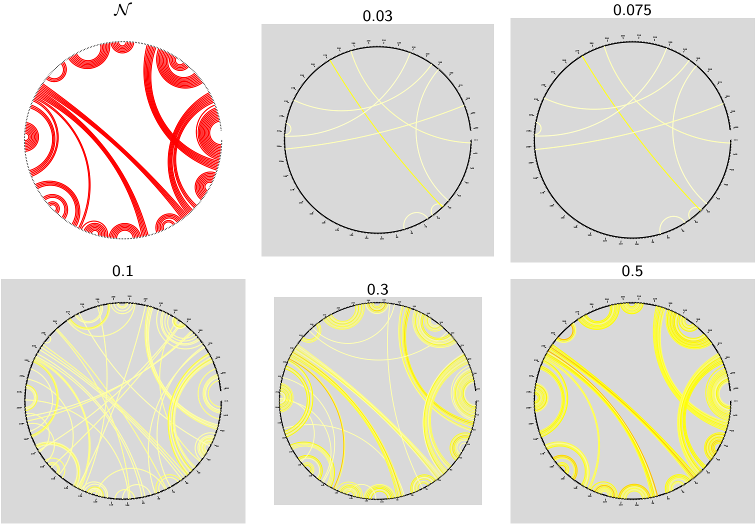
0.1

0.3

0.5

Prediction of tertiary interactions

Using Maximum Weighted Matching and a threshold



$\mathcal{N}$

0.03

0.075

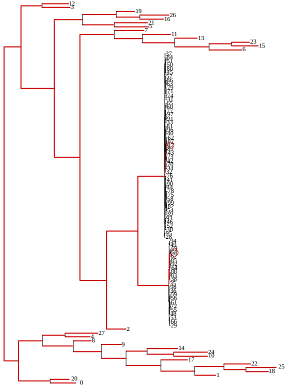
0.1

0.3

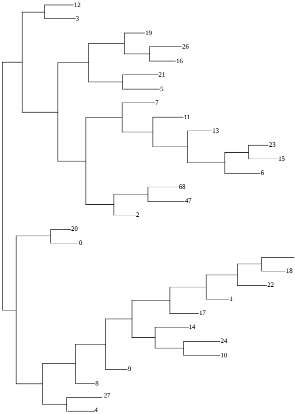
0.5

Tree topology

Fulltree: 100 sequences



Subtree: 30 sequences

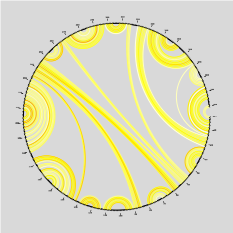


Is maximisation of evolutionary divergence useful for structure predictions?

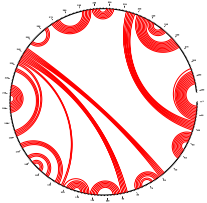
Evolve along the fulltree and the subtree

Consensus structures based on Mutual Information Content

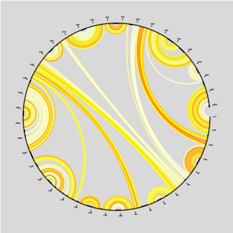
Fulltree



$\mathcal{N}$



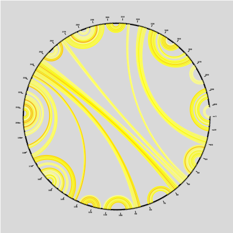
Subtree



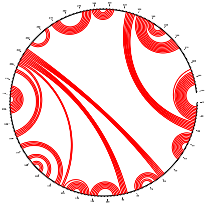
Evolve along the fulltree and the subtree, threshold

Consensus structures based on Mutual Information Content

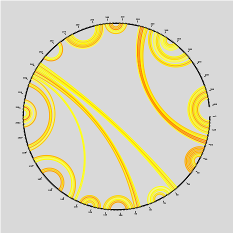
Fulltree



$\mathcal{N}$



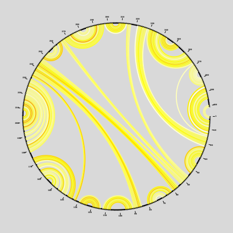
Subtree



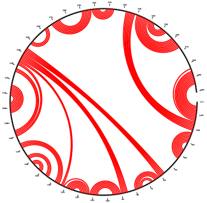
Reducing the alignment

Consensus structures based on Mutual Information Content

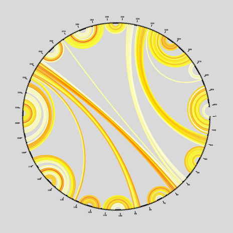
Fullalignment



$\mathcal{N}$

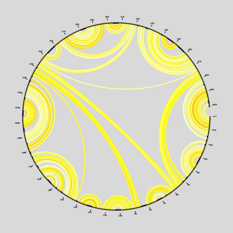


Subalignment

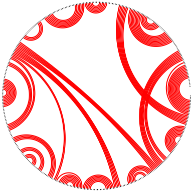


Reducing the alignment

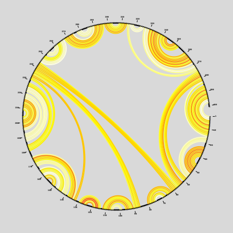
Consensus structures based on Mutual Information Content



Fullalignment



$\mathcal{N}$

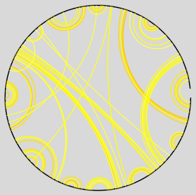


Subalignment

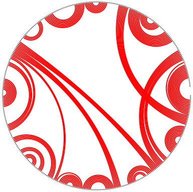
Reducing the alignment

Using Maximum Weighted Matching and a threshold

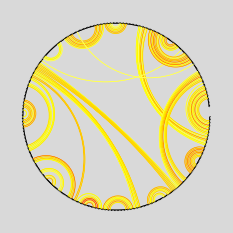
Fullalignment



$\mathcal{N}$



Subalignment



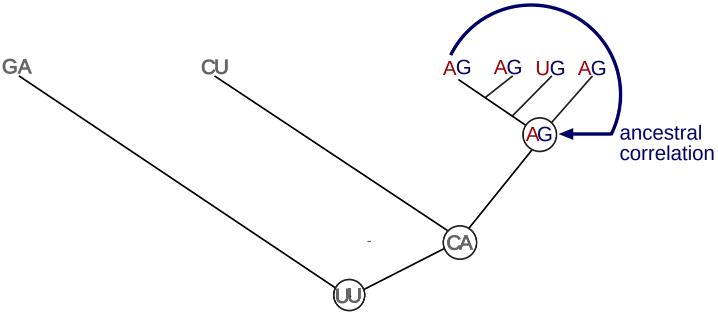
Influence of the tree topology

Mutual Information:

- Long branches → many true positives correlations
- Short branches → few true positives correlations
- Many false positives
- Comparative analysis ignores the phylogenetic information in the sequences, it tends to overestimate the amount of covariation between two positions.

Ancestral correlations

How long the branches need to be to avoid ancestral correlations?



Conclusion

- Including phylogenetic information in comparative analysis is potential useful
- Phylogeny can help to choose sequences for structure prediction methods:
    - Statistical study is necessary.
    - How looks the optimal tree for comparative structure prediction methods, with and without thermodynamics?
- Method for Reconstructing Dependencies with Phylogenetic Trees
    - Self-consistent method, where no threshold is needed.

Extension of SISSI

- SISSI is not limited to F81 types of rate matrices
  - E.g. inclusion of a transition-transversion parameter
  - Inclusion of codon position-specific heterogeneity
  - Studies with tertiary interactions
  - Inclusion of mixture models
- Inclusion of energy values
- Inclusion of indels

Acknowledgement

Arndt von Haeseler (Center for Integrative Bioinformatics Vienna)
Thomas Schlegel (Bioinformatics Institute, HHU Düsseldorf)
Minh Bui Quang (Center for Integrative Bioinformatics Vienna)
Steffen Kläre (Center for Integrative Bioinformatics Vienna)

Gerhard Steger (Institut für Physikalische Biologie, HHU Düsseldorf)
Andreas Wilm (Institut für Physikalische Biologie, HHU Düsseldorf)

Financial support from the DFG grant SFB-TR1 is gratefully acknowledged.

Special thanks to the big communities of Phylogenetics and RNA structures!