



# Comparative Structure Determination

Complementarity has to be preserved for base pairing positions, so an aligned set of homologous sequences should exhibit compensating mutations.

Standard way of measuring compensation is by *mutual information*:

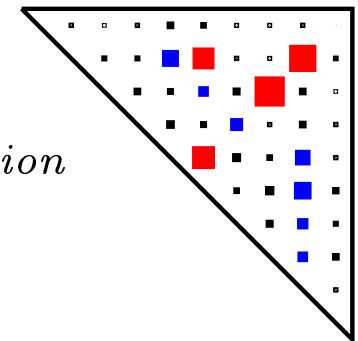
$$M_{ij} = \sum_{x,y} f_{x,i;y,j} \log_2 \frac{f_{x,i;y,j}}{f_{x,i} f_{y,j}}$$

where  $f_{x,i}$  is fraction of sequences having base  $x$  in alignment column  $i$  and  $f_{x,i;y,j}$  is fraction of sequences simultaneously having base  $x$  in column  $i$  and base  $y$  in column  $j$ .

Excellent if good alignment of many sequences is available.

<i>Phenylalanine (Phe)</i>	Acc-stem	D-stem	D-loop	D-stem	Ac-stem	Anticd-loop	Ac-stem	T-stem	T-loop	T-stem	Acc-stem	
	1	8 10		22	26 27	32	39	44	49	51	66	73
<i>Artibeus jamaicensis</i>	GTTGATG	TA GCTT	ATTAATA	AAGC A	AGGCA CTGAAAA	TGCCT	AGAT	GAGTC	GTA	TGACT	CATAAAG	A
<i>Balaenoptera musculus</i>	GTTAATG	TA GCTT	AAGCACTGATA	AAGC A	AGACA CTGAAAA	TGCTT	AGAT	GGGTT	AAATT	AACCG	CATTGGC	A
<i>Balaenoptera physalus</i>	GTTGATG	TA GCTT	AAAGCACTAGA	AAGC A	AGAGA CTGAAAA	TGCTT	AGAT	GGGTC	TAGCC	AACCG	CATTGAC	A
<i>Bos taurus</i>	GTTGATG	TA GCTT	AACCCA	AAGC A	AGGCA CTGAAAA	TGCCT	AGAT	GAGTC	TCCC	AACTC	CATAAAG	A
<i>Canis familiaris</i>	GTTAATG	TA GCTT	AATTAATA	AAGC A	AGGCA CTGAAAA	TGCCA	AGAT	GAGT	GGCAC	GACTC	CATAAAG	A
<i>Ceratotherium simum</i>	GTTAATG	TA GCTT	AACAACCTA	AAGC A	GGGCA TTGAAAA	TGCCC	AGAT	GAGCC	CACC	AGTCC	CATAAAG	A
<i>Dasyurus novemcinctus</i>	GTCAACG	TA GCTT	AAGTCTA	AAGC G	AGGCA CTGAAAA	TGCCT	AAAC	GAATC	CTAAT	GATTC	GGCAGAC	A
<i>Didelphis virginiana</i>	GTTAATG	TA GCTT	AATTTA	AAGC A	AAGCA CTGAAAA	TGCTT	AGAT	GGTIT	ATATGTT	AAACG	CATAAAG	A
<i>Equus asinus</i>	GTTAATG	TA GCTT	AATGATATCA	AAGC A	AGGCA CTGAAAA	TGCCT	AGAT	GAGTA	TCC	TACTC	CATAAAG	A
<i>Equus caballus</i>	GTTAATG	TA GCTT	AATAATATA	AAGC A	AGGCA CTGAAAA	TGCCT	AGAT	GAGTA	TTCT	TACTC	CATAAAG	A
<i>Erinaceus europaeus</i>	GTTAACG	TA GCTT	AAAATTA	AAGC A	AAGCA CTGAAAA	TGCTT	AGAT	GGGCC	TA	TACCC	GGTAAAG	A
<i>Felis catus</i>	GTTAATG	TA GCTT	AAACATATA	AAGC A	AGGCA CTGAAAA	TGCCT	AGAT	GAGTC	GCCA	GACTC	CATAAAG	A
<i>Gorilla gorilla</i>	GTTTATG	TA GCTT	ACCTGGCCA	AAGC A	ATACA CTGAAAA	TGTTT	CGAC	GGGCT	CACAT	CACCC	CATAAAG	A
<i>Halichoerus grypus</i>	GTTAATG	TA GCTT	AATAAACCA	AAGC A	AGGCA CTGAAAA	TGCCT	AGAT	GAGCC	ATAA	GGCTC	CATAAAG	A
<i>Hippopotamus amphibius</i>	GTTAAGG	TA GCTC	AAACACCCA	AAGC G	AGGCA CTGAAAA	TGCCT	AGAT	GGGCT	CACCC	AGCCG	CGTAAAG	A
<i>Homo sapiens</i>	GTTTATG	TA GCTT	ACCTCCTGA	AAGC A	ATACA CTGAAAA	TGTTT	AGAC	GGGCT	CAGAT	CACCC	CATAAAG	A
<i>Hylobates lar</i>	GTTTATG	TA GCTT	AAGTACCCA	AAGC A	AAAGA CTGAAAA	TGTGG	AGAC	GGCTC	ACC	CGCCG	CATAAAG	A
<i>Macropus robustus</i>	GTTAATG	TA GCTT	AATGCA	AAGC A	AAGCA CTGAAAA	TGCTT	AGAT	GGACT	TCAATA	AGTCC	CATAAAG	A
<i>Mus musculus</i>	GTTAATG	TA GCTT	AATAACA	AAGC A	AAGCA CTGAAAA	TGCTT	AGAT	GGATA	ATTG	TATCC	CATAAAG	A
<i>Myoxus glis</i>	GTTAATG	TA GCTT	ATAAT	AAGC A	AAGCA CTGAAAA	TGCTT	AGAT	AGGTA	GGTA	CACCC	CATAAAG	A
<i>Ornithorhynchus anatinus</i>	GCACCTGG	TA GCTT	AAACTCTTA	AAGC A	ATACA CTGAAAA	TGTTT	AGAT	GATTC	GTAECT	GAACC	GGAGCGG	A
<i>Oryzologus cuticulus</i>	GTTAATG	TA GCTT	AACAACA	AAGC A	AAGCA CTGAAAA	TGCTT	AGAT	GAGCC	TCCC	GGCTC	CATAAAG	A
<i>Ovis aries</i>	GTTAATG	TA GCTT	AAACTTA	AAGC A	AGGCA CTGAAAA	TGCCT	AGAT	GAGTC	FACT	GACTC	CATGAAG	A
<i>Pan paniscus</i>	GTTTATG	TA GCTT	ACGCCCTTA	AAGC A	ATACA CTGAAAA	TGTTT	CGAC	GGGTT	TATAT	CACCC	CATAAAG	A
<i>Pan troglodytes</i>	GTTTATG	TA GCTT	ACGCCCTGA	AAGC A	ATACA CTGAAAA	TGTTT	CGAC	GGGTT	TACAT	CACCC	CATAAAG	A
<i>Papio hamadryas</i>	GTTTATG	TA GCTT	AAACATACCCA	AAGC A	AGACA CTGAAAA	TGCCT	AGAT	GGGTT	GGCA	CGCCG	CATAAAG	A
<i>Phoca vitulina</i>	GTTAATG	TA GCTT	AATAAACCA	AAGC A	AGGCA CTGAAAA	TGCCT	AGAT	GAGCC	ACAA	GGCTC	CATAAAG	A
<i>Pongo pygmaeus</i>	GTTTATG	TA GCTT	ATTCCATCCA	AAGC A	ATACA CTGAAAA	TGCTT	CGAT	GGGCC	CACA	CGCCG	CATAAAG	A
<i>Rattus norvegicus</i>	GTTAATG	TA GCTT	ATAATA	AAGC A	AAGCA CTGAAAA	TGCTT	AGAT	GGATT	CAAA	AATCC	CATAAAG	A
<i>Rhinoceros unicornis</i>	GTTAATG	TA GCTT	AATGATTA	AAGC A	AGGCA TTGAAAA	TGCCT	AGAT	GAGAC	FACC	AACTC	CATAAAG	A
<i>Sus scrofa</i>	GTTAATG	TA GCTT	AAATTATCA	AAGC A	AGGCA CTGAAAA	TGCCT	AGAT	GGGCC	TC	ACAGC	CATAAAG	A

Mutual information  
in T domain

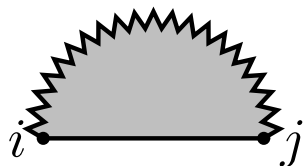


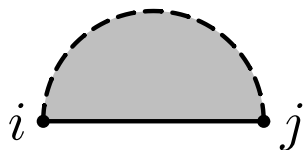
# Notation

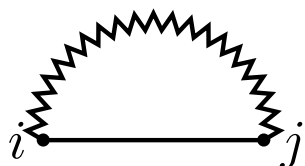
We will use a graphical notation to specify structural parts. Zigzagged lines indicates a definite base pairing, while dashed lines just delimits a region where the flanking bases may, or may not, be paired, but where no base inside the region forms a base pair with a base outside the region.

$i \text{---} j$  Base pairing of bases  $i$  and  $j$

$i \text{---} j$  Bases  $i$  and  $j$  not necessarily base paired

 (Optimal) structure on substring between bases  $i$  and  $j$  with bases  $i$  and  $j$  base paired

 (Optimal) structure on substring between bases  $i$  and  $j$

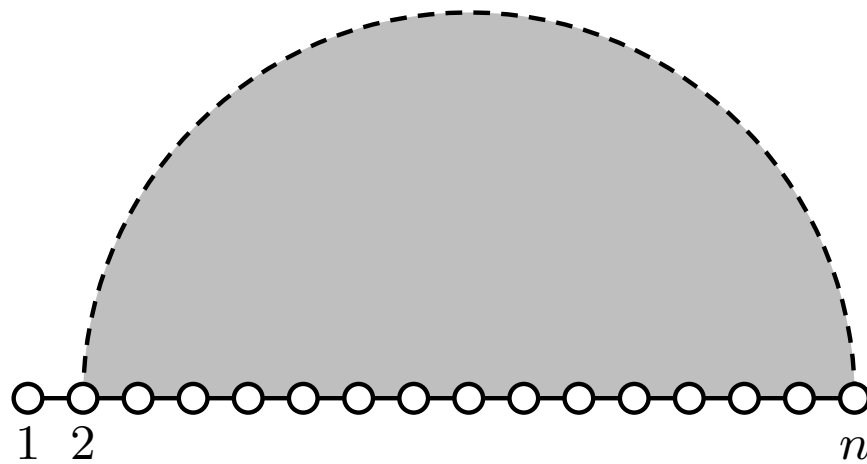
 Structure on substring between bases  $i$  and  $j$  only containing the base pair  $i \cdot j$

$i \text{---} j$  Empty structure on substring between bases  $i$  and  $j$

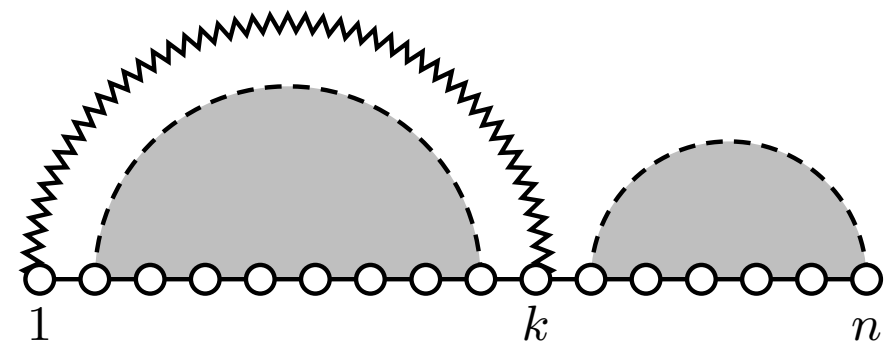
# Prediction by Maximum Number of Base Pairs

The hydrogen bonds formed by base pairs stabilise RNA structures – so a first attempt at structure prediction would be a structure with a maximum number of base pairs.

We might not know how such a structure looks, but for any structure either



First base is unpaired



First base pairs with some other base

So for a sequence  $s = s_1 s_2 \dots s_n$  the maximum number of base pairs is

- maximum number of base pairs for  $s' = s_2 \dots s_n$  or
- $1 + \max \#$  of b.p. of  $s' = s_2 \dots s_{k-1} + \max \#$  of b.p. for  $s'' = s_{k+1} \dots s_n$ , for some  $k$  such that  $s_1$  and  $s_k$  can form a base pair

# Recursive Algorithm

**Fact:** Two bases separated by less than three bases cannot form a base pair.

With at most four bases, the maximum number of base pairs is thus zero.

Otherwise, we can use the recursive breakdown of the previous overhead.

$$\text{maxbp}(s, i, j) = \begin{cases} 0 & \text{if } i > j - 4 \\ \max \left\{ \text{maxbp}(s, i+1, j), 1 + \max_{\substack{i+4 \leq k \leq j \\ i \cdot k}} \left\{ \text{maxbp}(s, i+1, k-1) + \text{maxbp}(s, k+1, j) \right\} \right\} & \text{otherwise} \end{cases}$$

maxbp(s, i, j)

**if**  $i > j - 4$  **then**

$m = 0$

**else**

$m = \text{maxbp}(s, i + 1, j)$

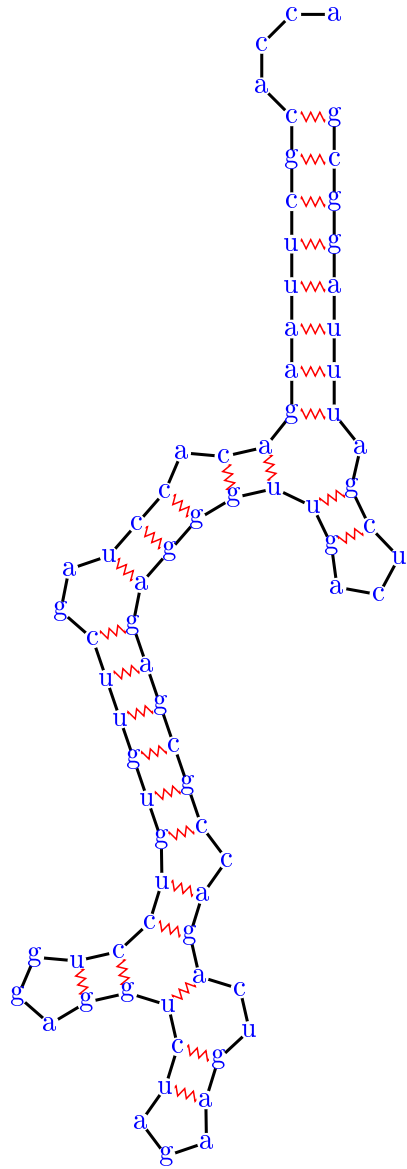
**for**  $k = i + 4$  **to**  $j$  **do**

**if**  $s_i$  and  $s_k$  can form a base pair **then**

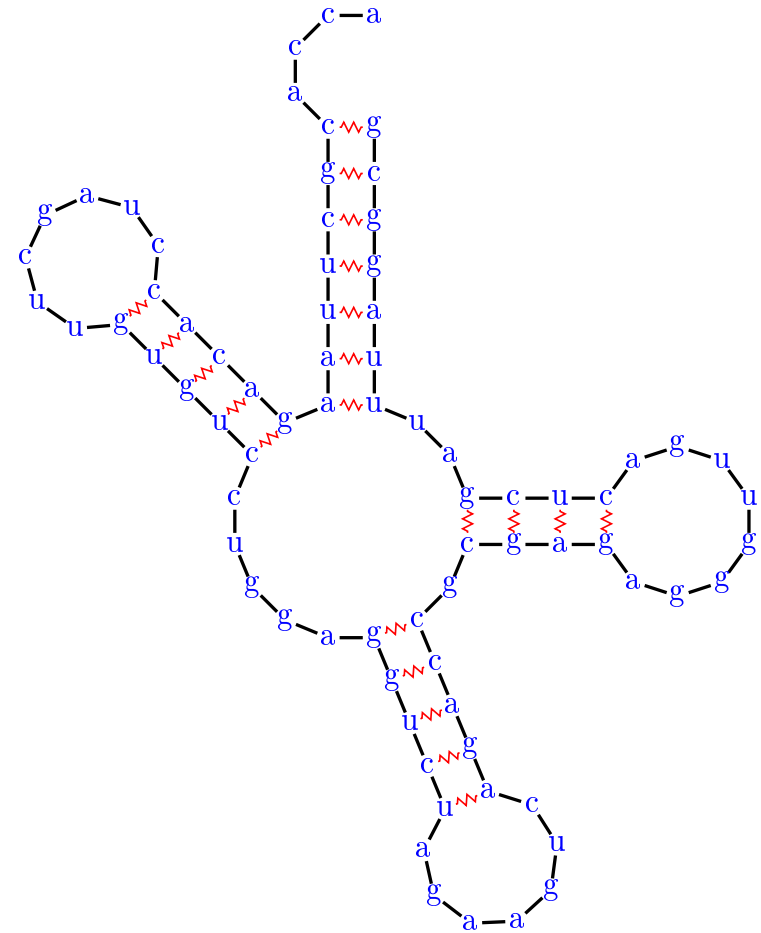
$m = \max\{m, 1 + \text{maxbp}(s, i + 1, k - 1) + \text{maxbp}(s, k + 1, j)\}$

**return**  $m$

# Algorithm Applied to Yeast tRNA<sup>Phe</sup>



Predicted Structure



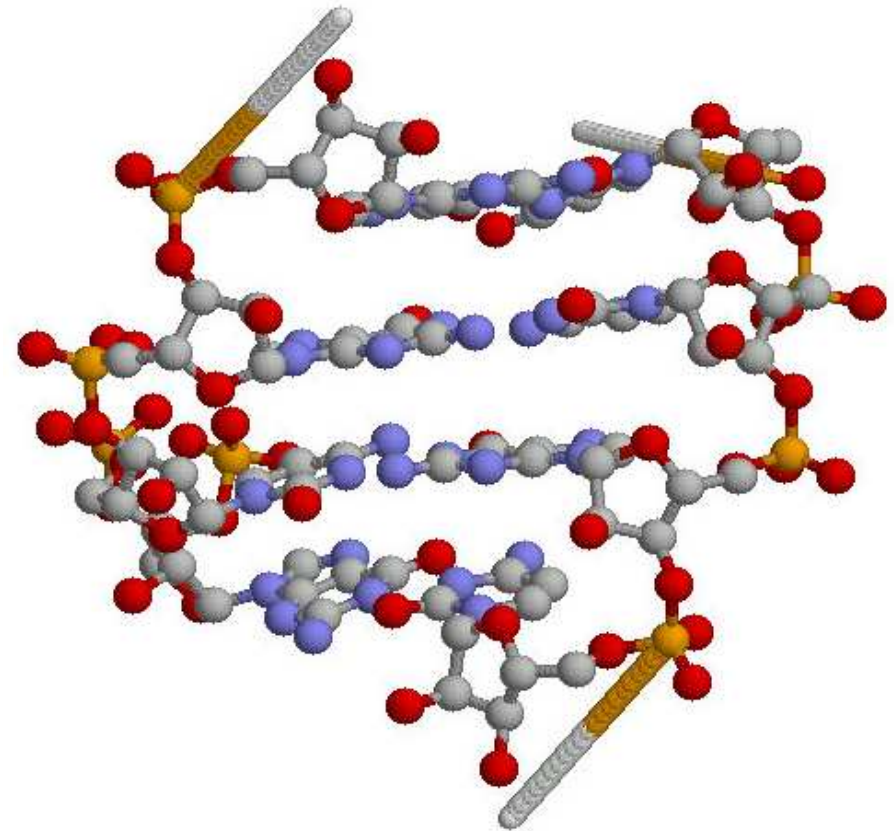
True Structure



# Base Pair Stacking

You might already have noticed that most of the base pairs for tRNA clumped together as consecutive base pairs.

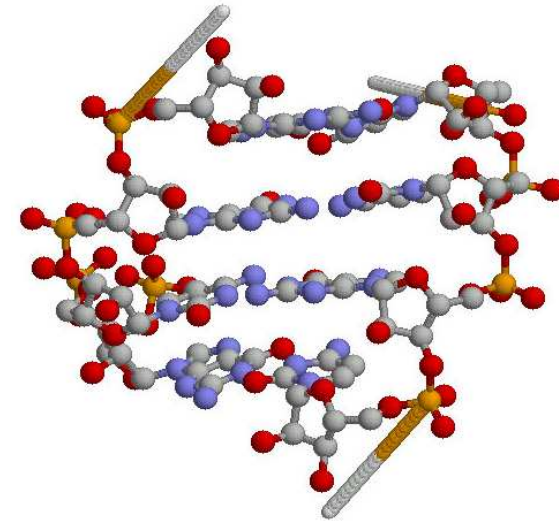
An isolated base pair usually *destabilises* a structure – several base pairs stacking next to each other are required for stability.



# Base Pair Stacking

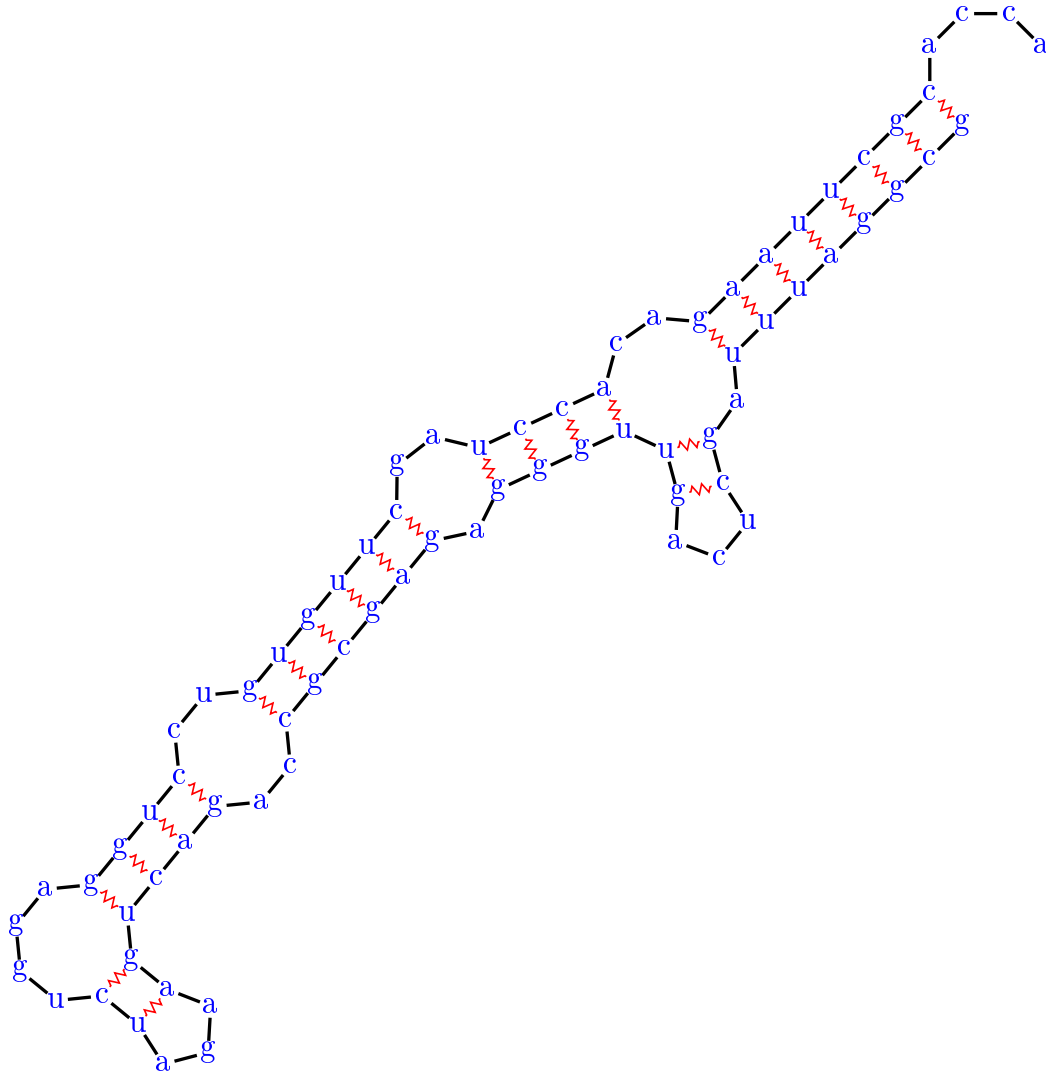
You might already have noticed that most of the base pairs for tRNA clumped together as consecutive base pairs.

An isolated base pair usually *destabilises* a structure – several base pairs stacking next to each other are required for stability.

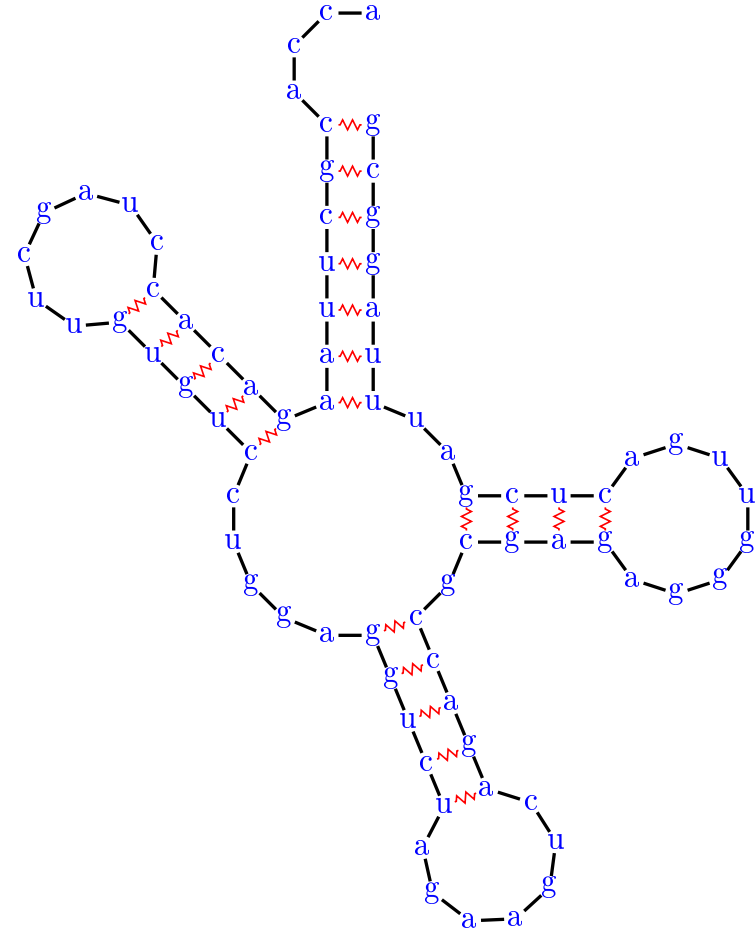


$$\begin{aligned}
 & \text{Diagram: A semi-circle with a dashed top edge and a solid bottom edge. The bottom edge is labeled with 'i' at the left end and 'j' at the right end.} \\
 & = \left\{ \begin{array}{l} 0 \text{ if } i > j - 4 \\ \max \left\{ \begin{array}{l} \text{Diagram: A semi-circle with a dashed top edge and a solid bottom edge. The bottom edge is labeled with 'i+1' at the left end and 'j' at the right end.} \\ \max_{i+4 \leq k \leq j} \left\{ \begin{array}{l} \text{Diagram: A semi-circle with a jagged top edge and a solid bottom edge. The bottom edge is labeled with 'i' at the left end and 'k' at the right end.} \\ \text{Diagram: A semi-circle with a dashed top edge and a solid bottom edge. The bottom edge is labeled with 'k+1' at the left end and 'j' at the right end.} \end{array} \right\} \end{array} \right\} \\
 \\
 & \text{Diagram: A semi-circle with a jagged top edge and a solid bottom edge. The bottom edge is labeled with 'i' at the left end and 'j' at the right end.} \\
 & = \left\{ \begin{array}{l} -\infty \text{ if } s_i \text{ and } s_j \text{ cannot form a base pair} \\ \max \left\{ \begin{array}{l} \text{Diagram: A semi-circle with a dashed top edge and a solid bottom edge. The bottom edge is labeled with 'i+1' at the left end and 'j-1' at the right end.} \\ \text{Diagram: A semi-circle with a jagged top edge and a solid bottom edge. The bottom edge is labeled with 'i+1' at the left end and 'j-1' at the right end.} \end{array} \right\} + 1 \end{array} \right\}
 \end{aligned}$$

# Base Pair Stacking Prediction for Yeast tRNA<sup>Phe</sup>

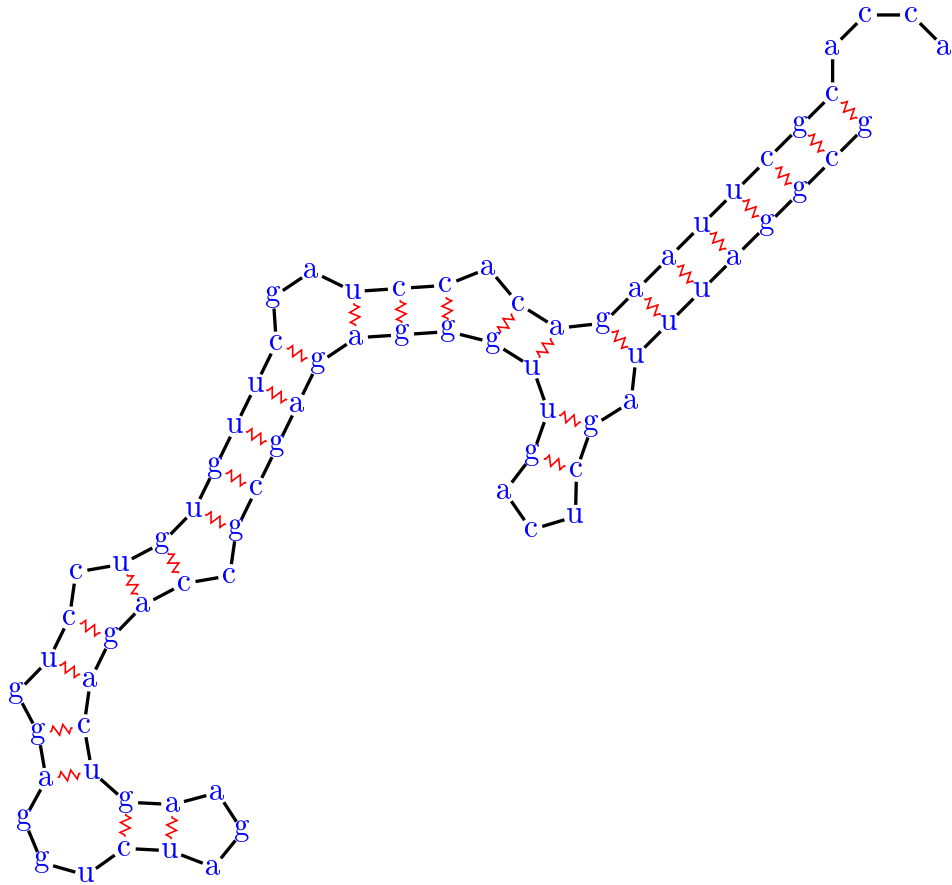


Predicted Structure

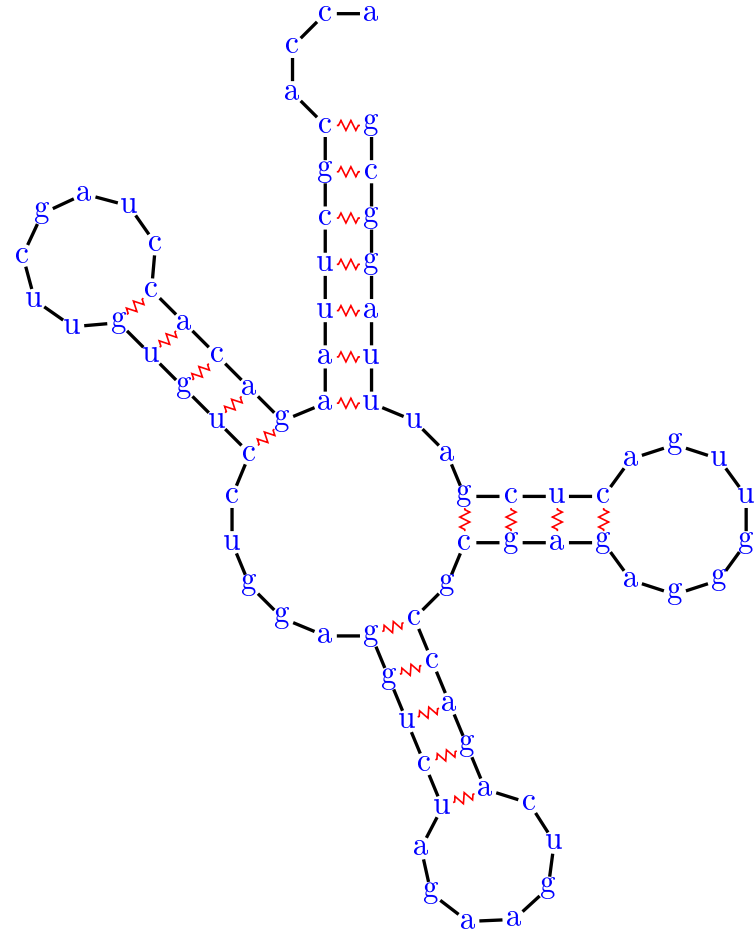


True Structure

# ... with Hydrogen Bond Counting

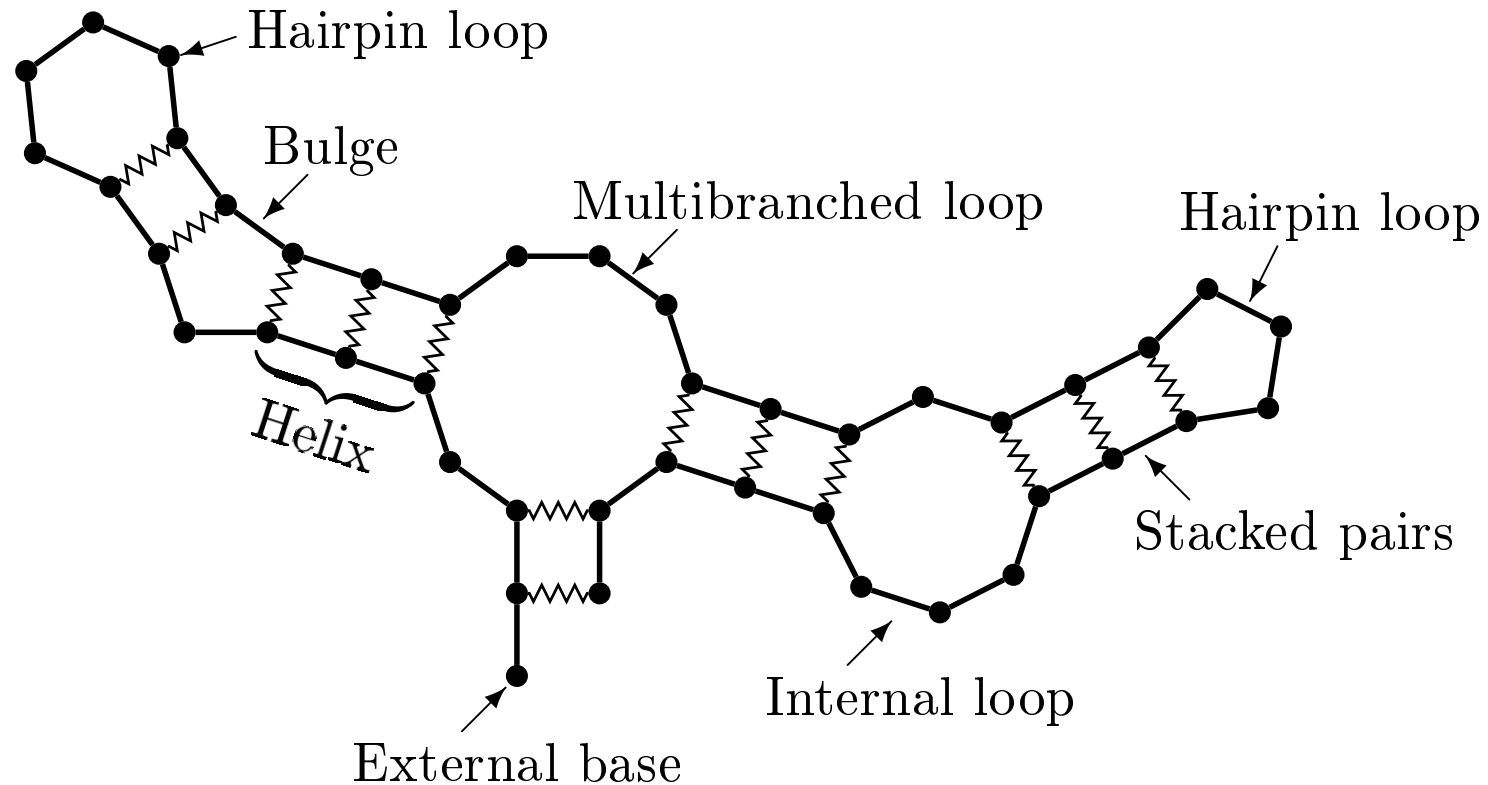


Predicted Structure



True Structure

# Loops



Over the last two and a half decades increasingly better data for the stability of various loops (measured in free energy) has been obtained.

$$V[i, j] = \max_{\substack{k \geq 0 \\ i < i_1 < j_1 < \dots < i_k < j_k < j}} \left\{ \text{score}(i \cdot j; i_1 \cdot j_1, \dots, i_k \cdot j_k) + \sum_{l=1}^k V[i_l, j_l] \right\}$$

# Energy Functions

$$eS \left( \begin{array}{c} \text{yellow} \quad \text{green} \\ \text{red} \quad \text{blue} \end{array} \right) = \text{tabulated for all combinations of base pairs}$$

$$eH \left( \begin{array}{c} \text{circle} \\ \text{red} \quad \text{blue} \end{array} \right) = \text{size} \left( \begin{array}{c} \text{circle} \end{array} \right) + \text{stacking} \left( \begin{array}{c} \text{black} \quad \text{black} \\ \text{red} \quad \text{blue} \end{array} \right)$$

$$eL \left( \begin{array}{c} \text{circle} \\ \text{yellow} \quad \text{green} \\ \text{red} \quad \text{blue} \end{array} \right) = \text{size} \left( \begin{array}{c} \text{blue bar} \quad \text{red bar} \end{array} \right) + \text{stacking} \left( \begin{array}{c} \text{black} \quad \text{black} \\ \text{red} \quad \text{blue} \end{array} \right) + \\ \text{stacking} \left( \begin{array}{c} \text{yellow} \quad \text{green} \\ \text{black} \quad \text{black} \end{array} \right) + \text{Ninio} \left( \begin{array}{c} \text{blue bar} \\ \text{red bar} \end{array} \right)$$

$$eM \left( \begin{array}{c} \text{circle} \\ \text{red} \quad \text{red} \\ \text{blue} \quad \text{blue} \end{array} \right) = a + b \cdot \# \left( \begin{array}{c} \text{blue} \quad \text{blue} \end{array} \right) + c \cdot \# \left( \begin{array}{c} \text{red} \end{array} \right) + \sum_{\begin{array}{c} \text{blue} \quad \text{blue} \end{array}} \text{stacking} \left( \begin{array}{c} \text{black} \quad \text{black} \\ \text{blue} \quad \text{blue} \end{array} \right)$$

# Mfold Recursions

We can now formulate recursions for computing the minimum free energy of any structure for a given sequence.

$$\begin{array}{c}
 \text{Diagram: A semi-circle with a dashed green top and a solid green bottom. The bottom line is labeled with '1' at the left end and 'i' at the right end.} \\
 wx(1, i) \\
 = \min \left\{ \text{Diagram: A semi-circle with a dashed green top and a solid green bottom. The bottom line is labeled with '1' at the left end and 'i' at the right end.}, \min_j \left\{ \text{Diagram: A semi-circle with a dashed green top and a solid blue bottom. The bottom line is labeled with '1' at the left end and 'i' at the right end.} \right\} \right\}
 \end{array}$$

$$\begin{array}{c}
 \text{Diagram: A semi-circle with a jagged blue top and a solid blue bottom. The bottom line is labeled with 'i' at the left end and 'j' at the right end.} \\
 vx(i, j) \\
 = \min \left\{ \text{Diagram: A circle with a jagged black top and a solid black bottom. The bottom line is labeled with 'i' at the left end and 'j' at the right end.}, \min_{k,l} \left\{ \text{Diagram: A semi-circle with a jagged black top and a solid blue bottom. The bottom line is labeled with 'i' at the left end and 'j' at the right end. There are points 'k' and 'l' on the bottom line.} \right\}, \min_k \left\{ \text{Diagram: A semi-circle with a jagged black top and a solid red bottom. The bottom line is labeled with 'i' at the left end and 'j' at the right end. There are points 'k' and 'l' on the bottom line.} \right\} \right\}
 \end{array}$$



$$\begin{array}{c}
 \text{Diagram: A semi-circle with a dashed red top and a solid red bottom. The bottom line is labeled with 'i' at the left end and 'j' at the right end.} \\
 wx_I(i, j) \\
 = \min \left\{ \text{Diagram: A semi-circle with a dashed red top and a solid red bottom. The bottom line is labeled with 'i' at the left end and 'j' at the right end.}, \text{Diagram: A semi-circle with a dashed red top and a solid red bottom. The bottom line is labeled with 'i' at the left end and 'j' at the right end.}, \text{Diagram: A semi-circle with a jagged blue top and a solid blue bottom. The bottom line is labeled with 'i' at the left end and 'j' at the right end.}, \min_k \left\{ \text{Diagram: A semi-circle with a dashed red top and a solid blue bottom. The bottom line is labeled with 'i' at the left end and 'j' at the right end. There is a point 'k' on the bottom line.} \right\} \right\}
 \end{array}$$

Current energy parameters allows predictions that on average find between 56% and 70% of the base pairs in known structures.

# The Mfold Server

Linear RNA folding at 5%, window = 2, max folds = 50  
18 A's, 18 C's, 23 G's, 17 U/T's and 0 N's.

10	20	30	40	50
GCGGAUUUAG CUCAGUUGGG AGAGCGCCAG				
ACUGAAGAUC UGGAGGUCCU				
60	70	80		

 - *Output* - 

The *energy dot plot* for tma. ([Definition](#))  
File formats: [Text](#), [PostScript](#), [png](#) [jpg](#)

Computed Structures: New *RNAML Syntax*. ([File Formats](#))  
The computed foldings contain 36 base pairs out of 38 (95%) in the *energy dot plot*.

Extra files: [h-num](#) values; [p-num](#) values; [log file](#) for main computations.

Download all foldings:      zipped file:       compressed tar file:       PostScript

View ss-count information: ([Definition](#)) ([ss-count file](#)) ss value = 0.92 ± 0.73

Averaging window       Magnification       Base to magnify about       Plot format      

View Individual Structures:

[Click Here for New Structure Viewing Options](#)

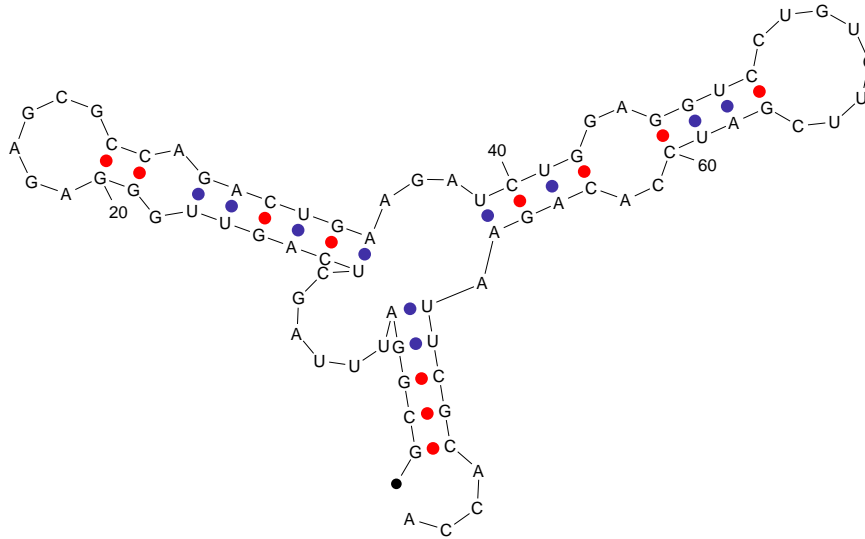
Structure 1: Initial dG = -22.9 kcal/mole, ([Thermodynamic Details](#)).  
Different file formats: [PostScript](#), [png](#), [jpg](#), [new .ct file](#), [RnaVic ct](#), [Mac ct](#), [GCG connect](#), [XRNA ss](#).

Structure 2: Initial dG = -21.9 kcal/mole, ([Thermodynamic Details](#)).

Available at <http://mfold.bioinfo.rpi.edu/cgi-bin/rna-form1.cgi>

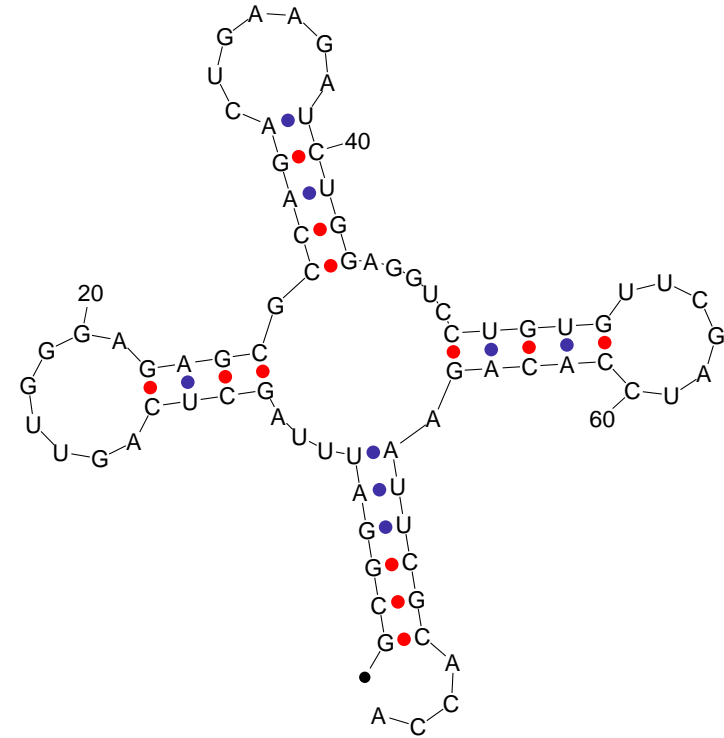
# Mfold Predictions

plt22ps by D. Stewart and M. Zuker  
© 2002 Washington University



dG = -20.84 [initially -22.9] trna

plt22ps by D. Stewart and M. Zuker  
© 2002 Washington University

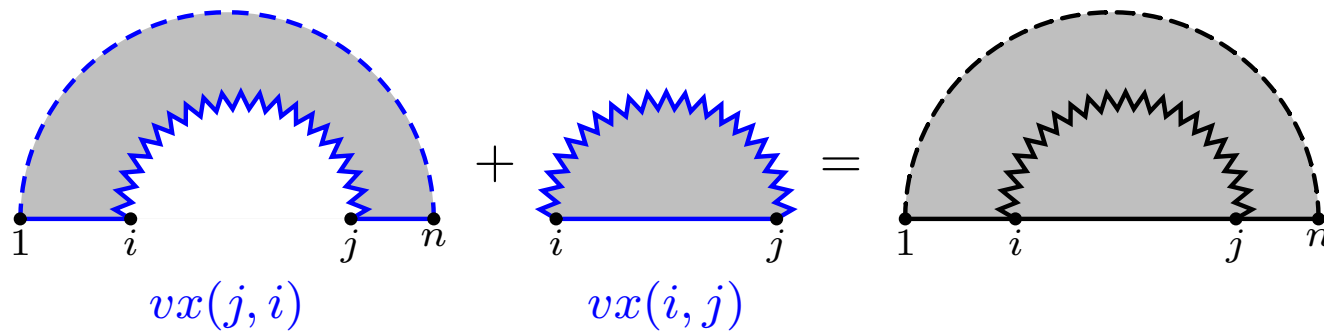


dG = -21.74 [initially -21.9] trna

Yes, it is still that yeast tRNA<sup>Phe</sup>...

# Suboptimal Structures

Just as we can compute the minimum energy of any structure between positions  $i$  and  $j$  and containing  $i \cdot j$ , we can compute the minimum energy of any structure *excluding* the positions between  $i$  and  $j$  and containing  $i \cdot j$ .



**Note:** Everything can still be handled in space  $O(n^2)$  and time  $O(n^3)$ .

# Boltzmann Distributions

For a system that can be in different states, the probability of it being in a particular state with energy  $E$  is  $e^{-E/k \cdot T}$

**Example:** Assume two magnets that can each be in one of two alignments

$$\begin{array}{cccc} \uparrow \uparrow & \uparrow \downarrow & \downarrow \uparrow & \downarrow \downarrow \\ E/k \cdot T = 0 & E/k \cdot T = \ln(2) & E/k \cdot T = \ln(2) & E/k \cdot T = 0 \\ Pr = \frac{1}{3} & Pr = \frac{1}{6} & Pr = \frac{1}{6} & Pr = \frac{1}{3} \\ & \underbrace{\hspace{10em}} & & \\ & Pr = \frac{1}{3} & & \end{array}$$

If we cannot distinguish between the magnets the two states with opposite alignments merge to one superstate

# The Full Partition Function

Our energy model defines a probability distribution on secondary structures

The partition function is the sum of Boltzmann terms from all structures

$$\begin{array}{c}
 \text{Diagram: A semi-circle with a dashed green top and a solid green bottom between points 1 and i.} \\
 wx(1, i) \\
 = \text{Diagram: A semi-circle with a dashed green top and a solid green bottom between points 1 and i, ending with a horizontal line to point i.} + \sum_j \text{Diagram: A semi-circle with a dashed green top and a solid blue bottom between points 1 and j, ending with a horizontal line to point i.}
 \end{array}$$

$$\begin{array}{c}
 \text{Diagram: A semi-circle with a jagged blue top and a solid blue bottom between points i and j.} \\
 vx(i, j) \\
 = \text{Diagram: A circle with a jagged black top and a solid black bottom between points i and j.} + \sum_{k,l} \text{Diagram: A semi-circle with a jagged black top and a solid blue bottom between points i and j, with a jagged blue top between points i and k, and a solid black bottom between points k and l.} + \text{Diagram: A semi-circle with a jagged black top and a solid red bottom between points i and j, with a dashed red top between points i and j.}
 \end{array}$$

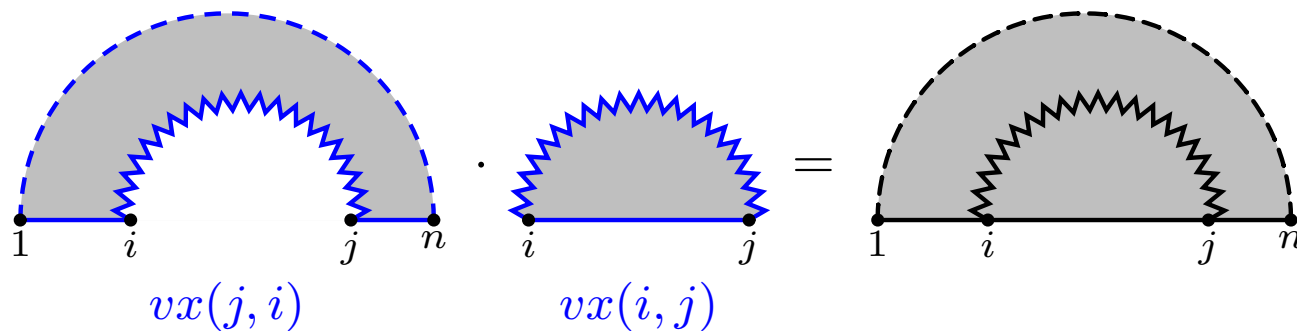
$$\begin{array}{c}
 \text{Diagram: A semi-circle with a dashed yellow top and a solid yellow bottom between points i and j.} \\
 wx_1(i, j) \\
 = \sum_k \text{Diagram: A semi-circle with a jagged blue top and a solid blue bottom between points i and j, with a solid black bottom between points i and k.} + \text{Diagram: A semi-circle with a dashed yellow top and a solid yellow bottom between points i and j, ending with a horizontal line to point j.}
 \end{array}$$

$$\begin{array}{c}
 \text{Diagram: A semi-circle with a dashed red top and a solid red bottom between points i and j.} \\
 wx_2(i, j) \\
 = \text{Diagram: A semi-circle with a dashed red top and a solid red bottom between points i and j, ending with a horizontal line to point j.} + \sum_k \text{Diagram: A semi-circle with a dashed yellow top and a solid blue bottom between points i and j, with a solid yellow bottom between points i and k, and a jagged blue top between points k and j.} + \sum_k \text{Diagram: A semi-circle with a dashed red top and a solid blue bottom between points i and j, with a solid red bottom between points i and k, and a jagged blue top between points k and j.}
 \end{array}$$

# Base Pair Boltzmann Probability

The probability of observing a base pair is the probability of observing a structure containing that base pair

Just as we can compute the full partition function for structures between positions  $i$  and  $j$  and containing  $i \cdot j$ , we can compute the full partition function for structures *excluding* the positions between  $i$  and  $j$  and containing  $i \cdot j$ .



Hence,  $Pr(i \cdot j) = vx(i, j) \cdot vx(j, i) / wx(1, n)$

**Note:** Everything can still be handled in space  $O(n^2)$  and time  $O(n^3)$ .