8.3.1 Simple energy minimization

Maximizing the number of base pairs as described above does not lead to good structure predictions. Better predictions can be obtained by minimizing the following *energy function* for an RNA sequence x and the set P of base pairs

$$E(x,P)=\sum_{(i,j)\in P} e(x_i,x_j), \tag{8.1}$$

where $e(x_i,x_j)$ is the amount of *free energy* associated with the base pair (x_i,x_j).

Reasonable values for e at 37^oC are -3, -2 and -1 kcal/mol for base pairs $\mathtt{C}-\mathtt{G}$, $\mathtt{A}-\mathtt{U}$ and $\mathtt{G}-\mathtt{U}$, respectively.

Using this we generalize the Nussinov algorithm such that the free energy of a base pair is considered. In the algorithm we now use $e(x_i,x_j)$ rather than the simple $\delta(i,j)$ function. Since the free energy of a base pair is negative we search for the structures with overall minimal energy. Thus the recursion formula is

$$E(i,j)=\min\begin{cases} E(i+1,j), \\ E(i,j-1), \\ E(i+1,j-1)+e(x_i,x_j), \\ \min_{i<k<j}[E(i,k)+E(k+1,j)] \end{cases}$$

Unfortunately, this approach does not produce good structure predictions because it does not take into account that helical stacks of base pairs have a stabilizing effect, whereas loops have a destabilizing effect on the structure. A more sophisticated approach is required.

8.3.2 Free energy minimization and the Zuker algorithm

In Thermodynamics, the *Gibbs free energy* G describes the energetics of molecules in aqueous solution. The change ΔG of the free energy in a chemical process, such as nucleic acid folding, determines the direction of the process:

- $\Delta G = 0$ indicates equilibrium,
- $\Delta G > 0$ indicates an unfavorable process and
- $\Delta G < 0$ indicates a favorable process.

Hence, biomolecules in solution arrange themselves so as to minimize the free energy of the entire system (biomolecules + solvent).

> **Note:** *Today: article on Bioinformatics in Financial Times Germany, see:*
> `http://www.ftd.de/forschung/138783.html`

A sophisticated algorithm for folding single RNAs is the *Zuker* algorithm (due to M. Zuker), an energy minimization algorithm which assumes that the correct structure is the one with the lowest *Gibbs free energy* G.

RNA molecules fold by intramolecular base pairing and are stabilized by hydrogen bonds that result from the base pairing. In addition, the stacking of base pairs in a helix also stabilizes the molecule and decreases the free energy of the folded RNA. Loops and bulges destabilize the structure.

The free energy G of an RNA secondary structure is thus approximated as the sum of individual contributions from:

- *loops*,
- *stacked base pairs*, and

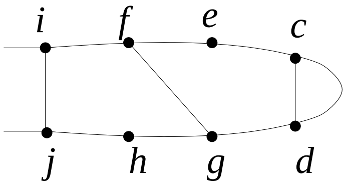
- *other* secondary structure elements.

Zuker's algorithm uses a dynamic programming approach that is based on the same four reduction steps as Nussinov's algorithm. We will see that an important difference is that the Zuker algorithm focuses on loops rather than base pairs. This provides a better fit to experimentally observed data.

8.3.3 The k-loop decomposition

Definition 8.3.1 *If (i, j) is a base pair in P and $i < h < j$, then we say that base h is* accessible *from (i, j) if there is no base pair $(i', j') \in P$ such that $i < i' < h < j' < j$.*

Definition 8.3.2 *The set of all bases accessible from a base pair $(i, j) \in P$ is called a* loop*. The* size *of the loop is the number of unpaired bases it contains.*



Definition 8.3.3 *The set s of all $k-1$ base pairs and k' unpaired bases that are accessible from (i, j) is called the k-*loop *closed by (i, j).*

The null *k-loop consists of all* free *base pairs and unpaired bases that are accessible from no base pair.*

Here is a formal definition of the secondary substructure elements that we introduced earlier:

Definition 8.3.4 *1. A 1-loop is called a* hairpin *loop.*

2. Assume that there is precisely one base pair (i', j') accessible from (i, j). Then this 2-loop is called

(a) a stacked pair*, if $i' - i = 1$ and $j - j' = 1$,*

(b) a bulge loop*, if $i' - i > 1$ or $j - j' > 1$, but not both, and*

(c) an interior loop*, if both $i' - i > 1$ and $j - j' > 1$.*

3. A k-loop with $k \geq 3$ is called a multi-loop*.*

The following is a consequence of nestedness:

> **Fact:** The number of non-null k-loops of a structure equals the number of base pairs it contains.

Each k-loop s_i is assigned an *energy* $e(s_i)$ and the energy of a structure P with non-null k-loops is given by:

$$E(P) := \sum_{i=0}^{m} e(s_i). \qquad (8.2)$$

Note that the energy is a function of k-loops and *not* a function of base pairs.

8.4 Zuker's algorithm for folding RNA

We will now develop a more involved dynamic program that uses loop-dependent rules. It is due to M. Zuker and Stiegler[3]. We will use two matrices, W and V.

Let $x = (x_1, x_2, \dots, x_L)$ be a string over the alphabet $\Sigma = \{A, G, C, U\}$. For $i < j$, let $W(i, j)$ denote the minimum folding energy of all non-empty foldings of the subsequence $x_i, \dots, x_j$.

Additionally, let $V(i, j)$ denote the minimum folding energy of all non-empty foldings of the subsequence $x_i, \dots, x_j$, *containing the base pair* (i, j). The following obvious fact is crucial:

$$W(i, j) \leq V(i, j) \text{ for all } i, j.$$

The two matrices V and W are initialized as follows:

$$W(i, j) = V(i, j) = \infty \text{ for all } i, j \text{ with } j - 4 < i < j.$$

(We now enforce that two paired bases are at least 3 positions away from each other).

8.4.1 Loop-dependent energies

We define different energy functions for the different types of loops:

- Let $eh(i, j)$ be the energy of the hairpin loop closed by the base pair (i, j),
- let $es(i, j)$ be the energy of the stacked pair (i, j) and $(i + 1, j - 1)$,
- let $ebi(i, j, i', j')$ be the energy of the bulge or interior loop that is closed by (i, j), with (i', j') accessible from (i, j), and
- let a denote a constant energy term associated with a multi-loop (a more general function for this case will be discussed later).

Predicted free-energy values (kcal/mol at 37^oC) for base pair stacking:

			A/U	C/G	G/C	U/A	G/U	U/G
A/U	-0.9	-1.8	-2.3	-1.1	-1.1	-0.8		
C/G	-1.7	-2.9	-3.4	-2.3	-2.1	-1.4		
G/C	-2.1	-2.0	-2.9	-1.8	-1.9	-1.2		
U/A	-0.9	-1.7	-2.1	-0.9	-1.0	-0.5		
G/U	-0.5	-1.2	-1.4	-0.8	-0.4	-0.2		
U/G	-1.0	-1.9	-2.1	-1.1	-1.5	-0.4		

Predicted free-energy values (kcal/mol at 37^oC) for features of predicted RNA secondary structures, by size of loop:

	size	internal loop	bulge	hairpin
1	.	3.9	.	
2	4.1	3.1	.	
3	5.1	3.5	4.1	
4	4.9	4.2	4.9	
5	5.3	4.8	4.4	
10	6.3	5.5	5.3	
15	6.7	6.0	5.8	
20	7.0	6.3	6.1	
25	7.2	6.5	6.3	
30	7.4	6.7	6.5	

[3] Zuker & Stiegler 1981, Zuker 1989

8.4.2 The main recursion

Here is the complete main recursion. For all i, j with $1 \leq i < j \leq L$:

$$\begin{aligned}
W(i,j) &= \min \begin{cases} W(i+1,j) \\ W(i,j-1) \\ V(i,j) \\ \min_{i<k<j}\{W(i,k)+W(k+1,j)\}, \end{cases} \\
V(i,j) &= \min \begin{cases} eh(i,j) \\ es(i,j)+V(i+1,j-1) \\ VBI(i,j) \\ VM(i,j), \end{cases} \\
VBI(i,j) &= \min_{\substack{i<i'<j'<j \\ i'-i+j-j'>2}} \{ebi(i,j,i',j')+V(i',j')\}, \text{ and} \\
VM(i,j) &= \min_{i<k<j-1}\{W(i+1,k)+W(k+1,j-1)\}+a.
\end{aligned}$$

In the following we discuss each of the four cases in detail.

The first case considers the four possibilities in which (a) i is unpaired, (b) j is unpaired, (c) i and j are paired to each other and (d) i and j are possibly paired, but not to each other. In case (c) we reference the auxiliary matrix V.

$$W(i,j) = \min \begin{cases} W(i+1,j) & \text{(a)} \\ W(i,j-1) & \text{(b)} \\ V(i,j) & \text{(c)} \\ \min_{i<k<j}\{W(i,k)+W(k+1,j)\}. & \text{(d)} \end{cases} \tag{8.3}$$

The second case considers the different situations that arise when bases i and j are paired, closing (a) a hairpin loop, (b) a stacked pair, (c) a bulge or interior loop or (d) a multi-loop. The two latter cases are more complicated and are obtained from equations 8.5 and 8.6.

$$V(i,j) = \min \begin{cases} eh(i,j) & \text{(a)} \\ es(i,j)+V(i+1,j-1) & \text{(b)} \\ VBI(i,j) & \text{(c)} \\ VM(i,j). & \text{(d)} \end{cases} \tag{8.4}$$

The third case takes into account all possible ways to define a bulge or interior loop that involves a base pair (i', j') and is closed by (i, j). In each situation, we have a contribution from the bulge or interior loop and a contribution from the structure that is on the opposite side of (i', j').

$$VBI(i,j) = \min_{\substack{i<i'<j'<j \\ i'-i+j-j'>2}} \{ebi(i,j,i',j')+V(i',j')\}. \tag{8.5}$$

The fourth case considers the different ways to obtain a multi-loop from two smaller structures and adds a constant contribution of a to close the loop.

$$VM(i,j) = \min_{i<k<j-1}\{W(i+1,k)+W(k+1,j-1)\}+a. \tag{8.6}$$

8.4.3 Time analysis

The minimum folding energy E_{min} is given by $W(1,L)$.

There are $O(L^2)$ pairs (i,j) satisfying $1 \leq i < j \leq L$.

The computation of

1. W takes $O(L^3)$ steps,
2. V takes $O(L^2)$ steps,
3. VBI takes $O(L^4)$ steps, and
4. VM takes $O(L^3)$ steps,

and so the total run time is $O(L^4)$.

The most practical way to reduce the run time to $O(L^3)$ is to limit the size of a bulge or interior loop to some fixed number d, usually about 30. This is achieved by limiting the search in Equation 8.5 to $2 < i' - i + j - j' - 2 \leq d$.

8.4.4 Modification of multi-loop energy

In Equation 8.6 we used a constant energy function for multi-loops. More generally, we can use the following function

$$e(\text{multi} - \text{loop}) = a + b \times k' + c \times k,$$

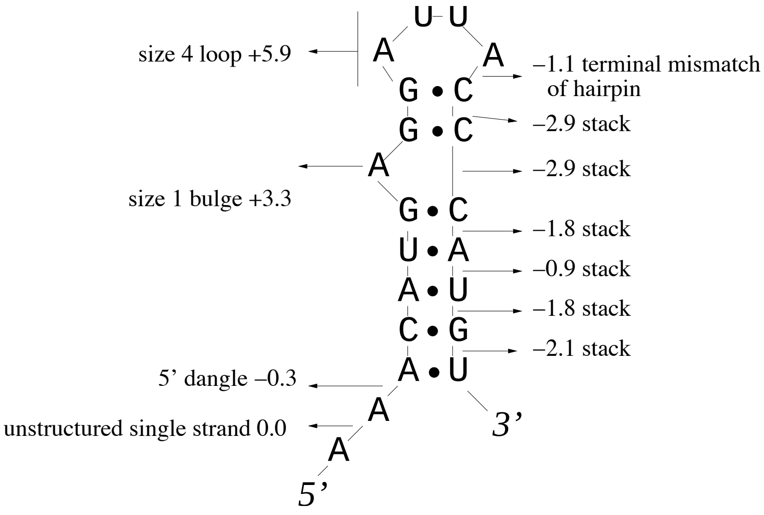
where a, b and c are constants and k' is the number of unpaired bases in the multi-loop.

This is a convenient function to use because, similar to the introduction of affine gap penalties in sequence alignment, a cubic order algorithm remains possible.

A number of additional modifications to the algorithm can be made to handle the stacking of single bases. These modifications lead to better predictions, but are beyond the scope of our lecture.

8.4.5 Example of energy calculation

Here is any example of the full energy calculation for an RNA stem loop (the wild type $R17$ coat protein binding site):



(from Durbin et al., 1999, p. 273).

Overall energy value: -4.6 kcal/mol

8.5 RNA folding via comparative analysis

An alternative to energy minimization techniques is a comparative approach. A guiding principle in molecular biology is that structure is more conserved than sequence. One example are tRNAs that are structurally conserved across anti-codon families and also across species.

When one base of a pair changes, we usually find that its partner also changes so as to conserve base pair complimentarity. This phenomenon is called a *compensatory* base change. How can we detect this?

The key idea is to identify compensatory (Watson-Crick) correlated positions in a multiple alignment.

For example:

seq1 G**C**CUUCGG**G**C
seq2 G**A**CUUCGG**U**C
seq3 G**G**CUUCGG**C**C

The two bold positions *covary* to maintain Watson-Crick complementarity.

Comparative methods require many diverse sequences and highly accurate multiple alignments to work well.

The amount of correlation of two positions can be computed as the *mutual information content* measure:

If you tell me the base at position i, how much do I learn about the base at position j?

8.5.1 Mutual information content

A method used to locate *covariant* positions in a multiple sequence alignment is based on the mutual information content of two columns.

First, for each column i of the alignment, the frequency $f_i(x)$ of each base $x \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}$ is calculated.

Second, the 16 joint frequencies $f_{ij}(x, y)$ of two nucleotides, x in column i and y in column j, are calculated.

For each pair of columns i, j we compute the ratio $\frac{f_{ij}(x,y)}{f_i(x) \times f_j(y)}$.

If the base frequencies of any two columns i and j are *independent* of each other, then the ratio of ≈ 1.

If these frequencies are *correlated*, then this ratio will be significantly greater than 1.

To calculate the *mutual information content* $H(i, j)$ in bits between the two columns i and j, the logarithm of this ratio is calculated and summed over all possible 16 base-pair combinations:

$$H_{ij} = \sum_{xy} f_{ij}(x, y) \log_2 \frac{f_{ij}(x, y)}{f_i(x) f_j(y)}. \tag{8.7}$$

For RNA sequences, we expect a value of 0 for complete randomness and/or complete conservation. We expect a maximum value of 2 bits when there is perfect correlation, because then $f_{ij}(x, y) = f_i(x) = f_j(y) = \frac{1}{4}$, and thus $\log_2 \frac{f_{ij}(x,y)}{f_i(x) f_j(y)} = \log_2 4 = 2$.

If either site is conserved, there is less mutual information: for example, if all bases at site i are `A`, then the mutual information is 0, even if site j is always `U`, because there is no covariance.

Compute the mutual information content:

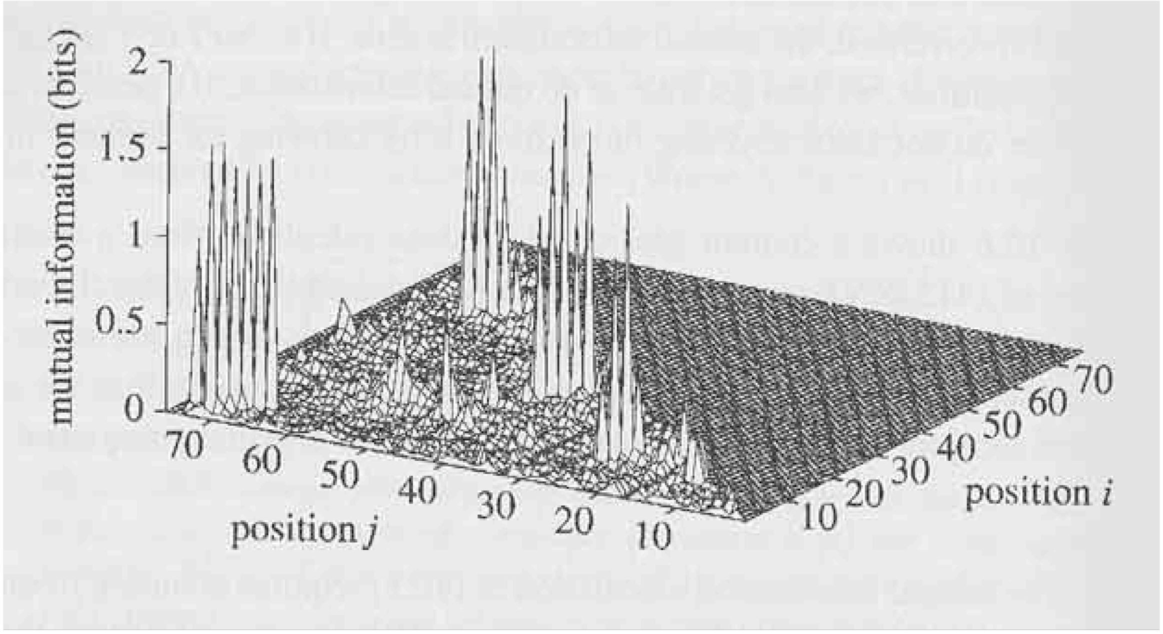
```
1  2  3  4  5  6

C  G  C  G  A  U
C  G  G  C  C  G
C  G  C  G  G  C
C  G  G  C  U  A
```

Compute: $H_{12} =$ ________
$H_{34} =$ ________
$H_{56} =$ ________

This comparative approach requires an accurate multiple alignment to get good structures. However, we need accurate structures to get a good alignment.

Example of a mutual information content plot of a tRNA:



(Source Durbin et al.,1999).

8.6 Important web sites with RNA folding servers

- Vienna RNA Secondary Structure Prediction:

  The web interface to the RNAfold program can be found at:
  `http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi`

- Zuker's mfold Server:

  The web interface to the `mfold` program can be found at:
  `http://www.bioinfo.rpi.edu/applications/mfold/` or
  `http://bioweb.pasteur.fr/seqanal/interfaces/mfold-simple.html`

- Sean Eddy's tRNAScan-SE:

  Allows to search specifically for tRNA genes in genomic sequences:

  `http://www.genetics.wustl.edu/eddy/tRNAscan-SE/`

- Nussinov:

  Web server at:
  `http://ludwig-sun2.unil.ch/~bsondere/nussinov/form.html`

- Important RNA Databases:

  – RFam: The Rfam database of RNA alignments and CMs
  (`http://www.sanger.ac.uk/Software/Rfam/`)

  – NonCode - database of non-coding RNAs
  (`http://noncode.bioinfo.org.cn/`)

  – RNAdb - mammalian non-coding RNA database
  (`http://research.imb.uq.edu.au/rnadb/`)

  – many more links at
  `http://www.imb-jena.de/RNA.html`