Principles of nucleic acid structure
Biophysical Chemistry 1, Fall 2010

Reading assignment: Chap. 3
Web assignment: http://w3dna.rutgers.edu
Nucleic acids: phosphates, sugars, bases

The phosphodiester linkage is directional.

The 3´-oxygen of nucleotide \( i \) is joined to the 5´-oxygen of nucleotide \( i+1 \).
The sugar-phosphate backbone

Main categories, E and T. If four atoms of the aldofuranose ring lie approximately in a plane, the conformation is described as envelope (E), due to the pictorial resemblance to an envelope, otherwise the conformation is described as twist (T) (Fig. 3.5).

The puckering of the ribose is conveniently described by the phase angle $P$, which is defined in terms of the internal ribose conformation angles $\nu_0$–$\nu_4$ (see Fig. 3.6):

$P$ must be in the interval $0^\circ$–$360^\circ$, so if $\nu_0 < 0$ then $P = P + 180^\circ$.

The various conformational nomenclatures of the ribose ring can conveniently be drawn along a circle as a function of $P$ (Fig. 3.7).

Adenine and guanine are the two most common purine bases, but inosine is found in some nucleic acid molecules. DNA has two types of pyrimidines, cytosine and thymine. In RNA, thymine is normally replaced by a similar base, uracil, which lacks the methyl group found in thymine (Fig. 3.8).

$P = \arctan \left( \frac{\sin(\nu_0)}{\sin(\nu_4)} \right)$.

Basics of Nucleic Acid Structure
There are two preferred ways of arranging the base in relation to the ribose, syn and anti (Figs. 3.9 and 3.10). In pyrimidine nucleotides only the anti conformation is found, because this avoids collision between the oxygen and the ribose. Purines can have both orientations, but the anti conformation is the most common of them.

Mammalian DNA is known to contain a methylated base, 5-methylcytosine (m5C). Bacterial DNA contains this one and two other methylated bases, namely,

**FIGURE 3.5** Various sugar ring puckering conformations. Those on the left are denoted S (for south); those on the right, N (for north). The C3’-endo conformation is seen at the top right, and the C2’-endo conformation at the top left. The notation of E and T conformations is also given.
Sugar puckering

FIGURE 3.7 Diagram showing the correlation between the phase angle $\phi$ and the ribose conformation. Conformational angles of $\phi$ are divided into two categories, north (N) and south (S).

FIGURE 3.8 The most common bases found in nucleic acids: the top row is purines; the bottom row pyrimidines. The atom-numbering scheme of purines and pyrimidines is given.
FIGURE 3.7
Diagram showing the correlation between the phase angle and the ribose conformation. Conformational angles of $P$ are divided into two categories, north (N) and south (S).

FIGURE 3.8
The most common bases found in nucleic acids: the top row is purines; the bottom row pyrimidines. The atom-numbering scheme of purines and pyrimidines is given.
The glycosidic torsion parameter

FIGURE 3.9

The anti and syn conformations of adenine monophosphate.

FIGURE 3.10

Diagram defining the torsion angle \( \chi \) around the N-glycosidic bond. The penta-gon illustrates the ribose unit, and the base is seen edge-on. The sequence of atoms chosen to define this angle is \( O4' - C1' - N9 - C4 \) for purine and \( O4' - C1' - N1 - C2 \) for pyrimidine derivatives. Thus when \( \chi = 0^\circ \) the \( O4' - C1' \) bond is eclipsed with the \( N9 - C4 \) bond for purine and the \( N1 - C2 \) bond for pyrimidine derivatives. The syn conformation is defined as \( \chi = \pm 90^\circ \) and anti as \( \chi = 180^\circ \pm 90^\circ \); thus, the syn conformational region is given by the upper half-circle, and the anti conformation by the lower half. The purine on the left is therefore in the syn conformation, and the pyrimidine on the right, in the anti conformation.
Watson-Crick base pairing

In the sugar-phosphate part the phosphate groups connect to the 3′ carbon of one deoxyribose moiety and the 5′ carbon of the next moiety, thereby linking successive deoxyriboses together. The two ends of a chain differ; the end where the 5′ carbon is not connected to another nucleotide is called the 5′ end. The other end is called the 3′ end. The two ends may have or lack free phosphate groups.

Each of the four bases in DNA has a unique set of hydrogen bond donors and acceptors that allows it to form base pairs with the other bases. In double-stranded DNA we have AT (adenine-thymine) base pairs with two hydrogen bonds and GC (guanine-cytidine) base pairs with three hydrogen bonds (Fig. 3.13). These interactions are called Watson–Crick base pairs to honor the scientists who first suggested that these base pairs are the basis of heredity.

In a double helix, the flat bases that base pair in the center are perpendicular to the helical axes. The two chains, which form a cylindrical spiral, run in opposite directions. The 5′ end of one chain is paired with the 3′ end of the other strand, and vice versa. This can be represented with an arrow for each chain, running from the 5′ end to the 3′ end. The arrows point in different directions — they are antiparallel. The orientation parameters are crucial during the copying of a single chain/strand.

In an ideal DNA double helix, there are approximately 10 base pairs per turn of DNA. Usually, the helical conformation is right-handed, that is, it twists to the right as do the threads on most screws. The sugar-phosphate backbone of the two antiparallel chains forms ridges on the edges of the helix. There are two grooves between the ridges formed by the ribose-phosphate backbone. The two grooves are of different widths and thus traditionally called the major and minor grooves. The major groove is wider and the bases are more accessible than in the minor groove. The exposed edges of the base pairs contain different hydrogen bonds.

FIGURE 3.13 /square6

The Watson–Crick base pairs. The sugar moieties are represented by R. Notice that the GC base pair on the left interacts via three hydrogen bonds, whereas the AT base pair on the right has only two. This makes the GC base pair and thus GC-rich DNA more stable than the AT base pair and AT-rich DNA.
A and B-form helices

A-DNA base pairs inclined with respect to helical axis and untwisted *cf.* B DNA.

A-DNA minor groove wider and more shallow, major groove narrower and more deep *cf.* B DNA.

Base pairs displaced from A-DNA helical axis.

A-DNA

11 bp/turn

+H$_2$O

$\rightarrow$

+salt and/or +alcohol

B-DNA

10 bp/turn
A and B-form helices

For each base pair, four different combinations of donors and acceptors exist at the exposed edges in the major and minor grooves. This creates a variety of patterns along the DNA double helix, available for highly specific DNA-protein interactions (see Sec 7.2).

Two factors are mainly responsible for the stability of the double helix structure: the base pairing between complementary strands and — especially — the stacking between adjacent base pairs. DNA molecules show considerable variation in conformation, dependent on the actual base sequence, the AT/GC content, and the presence of counter ions and stabilizing/destabilizing proteins. Since the AT base pair only contains two hydrogen bonds, it is easy to distort. AT-rich sequences are functionally important, and often serve as binding sites of DNA-binding proteins. In the crystal structure of the TATA box binding protein complex, the protein induces a significant bending of the DNA (see Sec. 7.5.2).

Double-stranded DNA can adopt several conformations, also called forms (Fig. 3.14). These conformations are determined by the activity of water and the nature of counterions. In aqueous solutions the DNA helix most often occurs in the B form, which has 10 bases per turn. One full turn measures 3.4 nm in the axial direction.
The overlap of successive base pairs depends on duplex form. Top-down "stacking diagrams" of dG₂·dC₂ and dA₂·dT₂ units in canonical A and B forms.
Concentrate on the basepair structures

Just help to neutralize the negative phosphate charge and thereby help to pack the molecule into a more dense structure, other proteins actively change the structure by causing local openings to form in the double helix. DNA helicases (Fig. 3.19) hydrolyze ATP when they are bound to single strands of DNA. The ATP hydrolysis changes the helicase structure and thereby enables it to perform mechanical work (see Secs. 6.1.1 and 7.5). Helicases propel themselves through the helix by separating the two strands and exposing the locally denatured DNA to other proteins, such as DNA polymerases. The situation is even more complex in eukaryotes, where the majority of DNA is located in the nucleus and packed into distinct units called chromosomes. A human chromosome

**FIGURE 3.17**

Getting DNA to bend

Global bend: 360°/75 bp left-handed superhelix

Combined B, A and B, C deformations tighten the bending of DNA:
The nucleosome core particle

The nucleosome core particle is composed of three α-helices connected by two loops. This structure is preserved in all eukaryotes. There are hundreds of hydrogen bonds, salt linkages and hydrophobic interactions formed between the histone octamer and the corresponding DNA helix. The DNA structure is not uniformly wound but several distortions and kinks can be seen.

Each of the core histones has a long N-terminal tail, which is subject to several modifications in the cell. These changes strongly influence the nucleosome and its higher structures and play important regulatory roles, for example during gene expression. The highly variable histone modifications may affect the interactions between the DNA and the histones as well as between different nucleosomes. In addition, the pattern of modifications may be recognized for the binding of other proteins. In particular, histone modifications can regulate the transcription of a certain part of the genome (Chap. 7 and App. E). The combination of histone modifications has been called the histone code.

It has been suggested that these modifications work as combinatorial signals that direct the binding of different molecules and subsequent events. When the DNA is replicated for a daughter cell the structure of the chromatin is also copied. Thus, the histone code is also copied. This phenomenon is called epigenetics.

The nucleosomes are packed into higher level chromatin structures. The first level is the 30-nm fiber, where an additional histone, H1, plays the important role.
DNA binding to the histone core proteins
Comparison to elastic rod models

Bending rigidity: $A = \frac{M(2\pi \nu n)^2}{LP_n^4}$; $a = A/(k_B T)$

Twisting rigidity: $C = \frac{I(2L\nu n)^2}{n^2}$

Stretching rigidity: $Y = \frac{ML(2\nu n)^2}{n^2}$

Lord Rayleigh, *The Theory of Sound*, 1894
Bending rigidity for linear duplex DNA

<table>
<thead>
<tr>
<th>n</th>
<th>GB frequencies</th>
<th>Analytical frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.114</td>
<td>0.116</td>
</tr>
<tr>
<td>2</td>
<td>0.294</td>
<td>0.295</td>
</tr>
<tr>
<td>3</td>
<td>0.522</td>
<td>0.532</td>
</tr>
</tbody>
</table>

d(GACT) 60 base pairs

\[ A = 2.44 \times 10^{-19} \text{erg.cm} \quad (2.26 \times 10^{-19} \text{erg.cm}) \quad a = 594 \, \text{Å} \quad (550 \, \text{Å}) \]
Stretching rigidity for linear duplex DNA

\[ Y = 1502 \, pN \quad (1000 - 1500 \, pN) \]

<table>
<thead>
<tr>
<th>n</th>
<th>GB frequencies</th>
<th>Analytical frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.664</td>
<td>0.619</td>
</tr>
<tr>
<td>2</td>
<td>1.289</td>
<td>1.237</td>
</tr>
<tr>
<td>3</td>
<td>1.807</td>
<td>1.846</td>
</tr>
</tbody>
</table>

\( n = 1 \) \hspace{1cm} \( n = 2 \) \hspace{1cm} \( n = 3 \)
Salt dependence of bending and stretching are not the same


Sequence dependence of bending rigidity

$d(GACT)$
$d(GC)$
$d(AT)$
$d(G)$
$d(C)$
$d(T)$
$d(CGG)$

Sequence (60 bp)
Now consider circular DNA

\[ \nu_n = f(\Omega, R, \Delta Tw, n, \rho) \]
- \( R \) is the circle radius
- \( \Delta Tw \) is the excess twist
- \( \Omega = C/A \)
- \( \rho \) is the mass density

Matsumoto, Tobias, Olson, *JCTC* 1, 117 (2005)
In-plane and out-of-plane modes for circular DNA

### In-plane bending motions

<table>
<thead>
<tr>
<th>n</th>
<th>GB frequencies</th>
<th>Analytical frequencies</th>
<th>GB frequencies</th>
<th>Analytical frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.165</td>
<td>0.172</td>
<td>0.121</td>
<td>0.166</td>
</tr>
<tr>
<td>3</td>
<td>0.394</td>
<td>0.452</td>
<td>0.366</td>
<td>0.443</td>
</tr>
<tr>
<td>4</td>
<td>0.672</td>
<td>0.724</td>
<td>0.697</td>
<td>0.704</td>
</tr>
</tbody>
</table>

### Out-of-plane bending motions

<table>
<thead>
<tr>
<th>n</th>
<th>GB frequencies</th>
<th>Analytical frequencies</th>
<th>GB frequencies</th>
<th>Analytical frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.217</td>
<td>0.231</td>
<td>0.211</td>
<td>0.211</td>
</tr>
<tr>
<td>3</td>
<td>0.525</td>
<td>0.545</td>
<td>0.496</td>
<td>0.520</td>
</tr>
<tr>
<td>4</td>
<td>0.793</td>
<td>0.851</td>
<td>0.864</td>
<td>0.802</td>
</tr>
</tbody>
</table>
Moving on to RNA

In DNA, the base pairs are centered over the helix axis. In an RNA double helix, the base pairs slide \( \sim 5 \text{ Å} \) away from the helix center. All these factors contribute to the tighter packing of the RNA double helix. The surface of an RNA helix is also quite different from the DNA double helix. The major groove of RNA is very narrow and deep, accentuated by the fact that RNA does not have the thymine methyl group, which resides in the major groove. In contrast, the minor groove is wide and shallow. For this reason, the major and minor grooves in RNA are more descriptively referred to as the deep groove and the shallow groove. The deep groove of RNA is a favored binding site of cations, water molecules and protein side chains.

### 3.4.2 Primary, Secondary and Tertiary Structure

We can classify the different structural levels of RNA in the same way as with proteins. The secondary structure is determined by base pairing. In addition to the normal Watson–Crick base pairs, mismatches such as GU base pairs are found. The secondary structure can form complicated patterns like the cruciform of tRNA. The secondary structure is formed by hydrogen bonds, with an energy of \( \sim -12 \text{kJ/mol} \). The formation of just a single base pair tends to be energetically unfavorable. The stacking of base pairs gives the largest stabilizing energy contribution of \( \sim -23 \text{kJ/mol} \). The stacking energies of base pairs are not symmetric, i.e. stacking of GC on AU is not the same as the stacking energy of AU on GC.

The most complex structural level is the tertiary one. The tertiary structure is formed by the spatial organization of the secondary structural elements, i.e. the stem-loop regions. A loop may form hydrogen bonds to another part of the backbone, resulting in a more compact helical conformation.

**FIGURE 3.23** The two types of sugar pucker most commonly found in nucleic acids. The C3′-endo pucker is prevalent in RNA and A-form DNA, whereas the C2′-endo pucker is characteristic of B-form DNA. It is seen that the C3′-endo pucker produces a significantly shorter phosphate-phosphate distance in the backbone, resulting in a more compact helical conformation.
RNA has more base-pairing possibilities

- Guanine (N7 and O6) and an acceptor and a donor on adenine (N7 and N6).
- Reverse Hoogsteen base pairing: two parallel chains, with the pyrimidine partner flipped and the ribose rings in the trans position.

**Figure 3.25**

- Left: Canonical Watson–Crick GC base pair (cis).
- Right: GC reverse Watson–Crick base pair (trans).

**Figure 3.26**

- AU Hoogsteen base pair.
- AU reverse Hoogsteen base pair.
- AU reverse Watson–Crick base pair. The blue dashed line shows the line of symmetry used to define the cis/trans conformation of the base pair. The AU Hoogsteen base pair is thus cis-H/WC, and the AU reverse Hoogsteen is trans H/WC.

**Figure 3.27**

- GU wobble base pair, one of the most common alternative base pairing patterns.
- GU reverse wobble, where the uracil group is simply flipped.

*Basics of Nucleic Acid Structure*
RNA has more base-pairing possibilities

RNA has more base-pairing possibilities

FIGURE 3.25 shows a standard Watson–Crick GC base pair (cis WC/WC) compared to the GC reverse base pair (trans WC/WC) that can be obtained between parallel backbone strands. Since parallel strands do not result from the usual stem-loop folding pattern of RNA, reverse Watson–Crick interactions are tertiary interactions.

Another fact, apparent in three-dimensional structures of RNA, is that the distance between backbones is different in the cis and trans configurations (Fig. 3.26).

The distance between backbone strands is shorter in the AU Hoogsteen pair than in the AU reverse Hoogsteen pair.

The term “wobble” base pairing was proposed by Francis Crick to account for the non-complementary GU base pairings observed in codon–anticodon interactions, manifested in the degeneracy of the genetic code (Chap. 8). Figure 3.27 shows the GU wobble base pair, which is one of the most common alternative base pairing patterns, and the GU reverse wobble, where the uracil group is simply flipped.

FIGURE 3.26 ■ Left: AU Hoogsteen base pair. Center: AU reverse Hoogsteen base pair. Right: AU reverse Watson–Crick base pair. The blue dashed line shows the line of symmetry used to define the cis/trans conformation of the base pair. The AU Hoogsteen base pair is thus cis-H/WC, and the AU reverse Hoogsteen is trans H/WC.
RNA has more base-pairing possibilities around the axis of the amine hydrogen bond. The GU wobble base pairing results in the loss of a hydrogen bond from the guanine, but the vacant amino group often forms hydrogen bonds to other bases nearby, perhaps in concert with the neighboring imino group. The GU wobble base pairings can be viewed as a canonical Watson–Crick pattern, with a shift of the pyrimidine partner.

The GU wobble pair has geometric properties that enable it to fit very well into a regular A-form helix, and therefore frequently substitute for regular Watson–Crick pairs. In certain chemical environments, the A base can become protonated at the N1 position. When this happens, the A base is able to form hydrogen bonds with a C base, forming an (A+):C base pair. The geometry of the (A+):C base pair is isosteric with the GU wobble pair.

3.4.4 Base Triplets: A Prominent Tertiary Structural Motif

The alternative base pairing patterns described in the previous section lead to a rich selection of multiple base interactions. Whenever a base pair is formed from two nucleotides, several hydrogen donors or acceptors are still available for alternate interactions to occur, either from the amino acid side chains of proteins, or from other nucleotides. The most important of these multi-base interactions are base triplets, important in maintaining the tertiary structure of RNA molecules. Several examples of triple base interactions will be presented in the following sections.

The possibility of triple base interactions were first realized in 1957 by Felsenfeld, who demonstrated that a (poly-A):(poly-U) duplex molecule could interact with a second poly-U strand to create a triple strand complex (Fig. 3.28). The additional poly-U strand interacts with the major groove of the duplex by forming Hoogsteen base pairs with the poly-A strand of the duplex. Later it was found that several other sequence combinations can lead to triple helix formation, as long as they contain two pyrimidines and one purine, for example C:G-C. Triple helix formation has been proposed to have a role in gene repair.
RNA uses chemically-modified bases thought to preserve a flanking cleavage site or perhaps to repress and to regulate translation of genes in the vicinity.

### 3.4.5 RNA Contains Modified Bases

Posttranscriptional modification of RNA results in an even greater diversity of modified bases than in DNA. This is especially true in functional RNAs such as tRNA and rRNA, where both ribose and base can be modified. The modified bases can profoundly change the chemical characteristics of the RNA molecule and can contribute to the stability of the molecule as well as partake in its external interactions and reactions.

A telling example of the importance of posttranslational modification is given by an *E. coli* isoleucine-specific tRNA^2\textsubscript{Ile}, specific for the codon AUA. This tRNA contains a modified base, lysidine, at the first position of the anticodon. The base lysidine is a cytidine that has been posttranslationally modified by the addition of the amino acid lysine to the C2 position (Fig. 3.32). If the lysidine residue is replaced by the native cytidine, a marked reduction of isoleucine incorporation is the result, and surprisingly, the appearance of methionine-accepting activity. How can this be? It turns out that in *E. coli*, the cognate isoleucyl tRNA ligase recognizes and charges tRNA\textsubscript{Ile} only if the lysidine residue is present in the anticodon loop. Furthermore, the anticodon sequence of tRNA^2\textsubscript{Ile} is CAT, which normally codes for methionine. If lysidine is absent, the tRNA will instead be recognized and mischarged by the CAT-recognizing methionyl tRNA ligase. So a single posttranslational modification is responsible for both the codon and amino acid specificity of this tRNA.

### FIGURE 3.32

Examples of modified bases in RNA. Modifications are marked in red. R stands for ribose.
RNA motifs: tetraloops

The RNA backbone has six degrees of freedom for each residue whereas the polypeptide backbone only has two. This extreme flexibility allows single-stranded RNA regions to adopt a wide range of conformations. Nevertheless, a couple of single-stranded RNA motifs are particularly common. The first is the S-turn, where an S shape is formed by two consecutive bends in the phosphate-sugar backbone and distinguished by inverted sugar puckers. The S-turn motif is found in the ribosomal loop E motif and the sarcin-ricin loop (see Sec. 8.3). The other important single-stranded RNA motif is the U-turn, which is a sharp bend in the backbone between the first and second nucleotides, followed by a distinctive stacking of the second and third nucleotides. Hydrogen bonds between the first and third residues often stabilize the motif. U-turns are typical of GNRA loops and are also found in the TΨC loop of tRNA.

3.4.7.1 The GNRA tetraloop

Tetraloops are a particularly common motif in RNA structures. An especially well-known case of this hairpin loop is the GNRAa loop motif, which closes the ribosomal loop E motif. A stand for puRine; N stands for aNy; Y stands for pYrimidine.

**FIGURE 3.36** Three-dimensional structures of various tetraloop folds. Left: GNRA loop from 5S rRNA (PDB: 1JJ2). The first G in the loop stabilizes the loop by hydrogen bonding to the fourth member. Middle: ANYA loop from MS2-RNA complex (PDB: 1DZS). Bases one and two form a stacking interaction, while bases of three and four of the loop are looped out and poised to interact with other species. Right: UNCG tetraloop from 16S rRNA (PDB: 1BYJ). The first U and the last G in the tetraloop interact via hydrogen bonds, while bases of one and two in the loop form a stacking interaction. The third base in the loop is available for interaction with other species.
An example of an internal loop that has gained a lot of attention is the loop E motif, first recognized in the 1980s when it was found that similar, conserved internal loops in eukaryotic 5S rRNA and in PSTV viral RNA shared a surprising disposition for cross-linking when exposed to UV light (Fig. 3.38). Later, the loop E motif was found in the sarcin-ricin loop of ribosomal 23S rRNA, which is involved in the binding of elongation factors EF-Tu and EF-G. In E. coli 5S rRNA, loop E is known to constitute a specific binding site for ribosomal protein L25 (Fig. 3.39). It was therefore apparent that this motif was an important site of activity and molecular recognition. Both 5S rRNA and the sarcin-ricin loop have been intensively studied by both high resolution NMR and crystallography. Loop E is an asymmetric internal loop characterized by a highly conserved stack of seven non-Watson–Crick base pairs. The hydrogen-bonding pattern is particularly well conserved. The first base pair is an AG sheared pair, followed by a UA trans Hoogsteen pair, a bulged G base, and finally a trans Hoogsteen AA base pair with locally parallel backbone conformation (Fig. 3.38). The AU reverse Hoogsteen base pair observed in loop E is the most abundant AU interaction in the ribosomal RNAs after the normal Watson–Crick base pair.
Even mRNA has structure

**FIGURE 3.50** Schematic representation of eukaryotic mRNA showing the 5’ cap, the coding region (red), and the 5’ and 3’ UTRs.
More on RNA secondary structure

Fortunately, a lot of interesting RNA is shorter than this, or can be thought of in terms of domains

- predicting secondary structure from sequence
- the inverse folding problem: finding sequences that are compatible with structure
- making the 2D $\rightarrow$ 3D transition
- force field simulations of nucleic acids
Nearest-neighbor energy function

Lu, Turner & Mathews, NAR 34, 4912 (2006)

Hairpin: 

\[
\begin{array}{cc}
\text{G} & \text{A} \\
\text{C-G} & \text{G-C} \\
\text{G-C} & \\
\end{array}
\]

\[
\Delta G_{\text{loop}} = \Delta G_{\text{loop}}^{(6)} + \Delta G\left(\frac{CG}{GA}\right) + \Delta G\left(\frac{G}{A}\right) = +2.9 \text{ kcal/mol}
\]

\[
\Delta G_{\text{stem}} = \Delta G\left(\frac{GG}{CC}\right) + \Delta G\left(\frac{GC}{CG}\right) = -6.3 \text{ kcal/mol}
\]

\[
\Delta G_{\text{hairpin}} = \Delta G_{\text{loop}} + \Delta G_{\text{stem}}
\]

Similar calculations for \(\Delta H\), \(\Delta S\), and \(T_{\text{melt}}\)
A slightly more complex structure:

\[
\begin{align*}
\Delta G^\circ &= -1.7 - 3.4 - 2.4 - 0.2 - 2.1 - 3.3 - 2.5 + 5.4 \\
&= -10.2 \text{ kcal/mol}
\end{align*}
\]
A very nice folding program (Windows only) is RNAstructure: http://rna.urmc.rochester.edu/rnastructure.html

1. sequence editing, secondary structure prediction
2. dynalign, for multiple sequence alignment
3. partition function analysis, gives much improved confidence levels for the predictions

Prediction of secondary structure as a web service: http://www.bioinfo.rpi.edu/~zukerm/ (classical approach) or at http://sfold.wadsworth.org/ (statistical sampling of the Boltzmann distribution)

RNAmotif: scanning genomes for secondary structure motifs

desc
h5 (minlen=4, maxlen=6) ss (seq = "^uucg$"

returns:

<table>
<thead>
<tr>
<th>RNA</th>
<th>Score</th>
<th>Start</th>
<th>End</th>
<th>Seq1</th>
<th>Seq2</th>
<th>Seq3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AEURR16S</td>
<td>0.000</td>
<td>0</td>
<td>1035</td>
<td>12 tccc ttcg gggga</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGIRRDDX</td>
<td>0.000</td>
<td>0</td>
<td>180</td>
<td>14 ggagc ttcg gctcc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGIRRDDX</td>
<td>0.000</td>
<td>0</td>
<td>181</td>
<td>12 gagc ttcg gctc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AHNRRDX</td>
<td>0.000</td>
<td>0</td>
<td>181</td>
<td>12 gagc ttcg gctc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANNRRRO</td>
<td>0.000</td>
<td>0</td>
<td>39</td>
<td>12 cccc ttcg gggg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AOSRR18S</td>
<td>0.000</td>
<td>0</td>
<td>175</td>
<td>12 gccc ttcg gggc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APBRR16SD</td>
<td>0.000</td>
<td>1</td>
<td>1321</td>
<td>12 tacc ttcg ggta</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APSRRE</td>
<td>0.000</td>
<td>0</td>
<td>227</td>
<td>12 cacc ttcg ggtg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AQF16SRRN</td>
<td>0.000</td>
<td>1</td>
<td>39</td>
<td>12 agcg ttcg cgct</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AQF16SRRN</td>
<td>0.000</td>
<td>1</td>
<td>40</td>
<td>14 cagcg ttcg cgctg</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

.....
Motif for an artifically generated DNA enzyme

Hypothesis: this motif would self-destruct and never be found in real genomes

Looking for a analogue of an artificial DNA enzyme

```python
parms
    wc += gu;

descr
    h5( tag='h1', len=5, mispair=1 )
    ss( len=1, seq="r" )
    h5( tag='h2', minlen=5, maxlen=8, seq="^y", mispair=1 )
    ss( minlen=4, maxlen=200 )
    h3( tag='h2', seq="r$" )
    ss( tag='ggct', len=15, mismatch=1, seq="^ggctagcnacaacr:"
    h3( tag='h1' )

Results:

Arabidopsis thaliana chr II sect 146/255:
AC004747    1.000 1 15537 171 ttgtt a tccgc ttt...(135)...tgggtgga
ggctaTccacaaacaa ggtgg
```
How can we generate descriptors?

- **search with strict 2° struct descriptor**
- **choose seq with low T.E. and look for conserved nt**
- **include conserved nt in descriptor, search with looser 2° struct constraints**

This general procedure can often be used to find conserved nucleotides in families of structures, and to search for specific types of RNA in genomic databases.
Example: the tRNA motif

(a) Secondary structure of tRNAPhe from yeast. (b) Schematic representation of the three-dimensional folding of the tRNA molecule, using the same color scheme as in (a).
The common features of the secondary structure of all the cytoplasmic tRNAs are as follows (Jack et al., 1976; Kim, 1979):

1. There are 7 bp in the amino acid (AA) stem, 3 or 4 bp in the dihydrouracil (D) stem, 5 bp in the anti-codon (AC) stem, and 5 bp in the TC (T) stem.

2. The base pairs in the stems are of the Watson–Crick type with an occasional G • U pair.

3. There are always two nucleotides between the AA and D stems, one between the D and AC stems and no nucleotide between the AA and T stems.

4. There are always 7 nucleotides in the AC and T loops, 7–10 nucleotides in the D loop, and 4–21 nucleotides in the variable arm (V arm).

5. Certain bases are conserved among all tRNAs except for the initiation tRNA: U between the AA and D stems; A at the beginning of the D loop; two G’s at the D loop; U in the AC loop; a GC base pair in the T stem; a TUC sequence and A in the T loop; and CCA at the 3’-end.

6. Some positions are semiconserved. For example, the base after the sequence TUC in the T loop is always a purine, either G or A.
1. Descriptor with no sequence requirement. GU’s are allowed, and no mispairs are allowed.

Result: 26 tRNAs not found, 5 false positives, all with higher Turner energies than true tRNAs.

2. Analyze sequence conservation for 50 hits with lowest Turner energies. Found 11 nucleotides that are 100% conserved.

3. Include conserved nucleotides in descriptor, but allow a mispair in each helix.

Result: 2 tRNAs not found, no false positives.
Generating an optimized tRNA descriptor for E. Coli

(lgth=7)
(lgth=3−4)
(lgth=5)
(lgth=8–11)
(lgth=5)
(lgth=4–22)
(lgth=7)

(GU basepairs not allowed; one mispair per helix; no sequence mismatches)

Performance for K12 and O157:H7: no false positives,
one missing tRNA for K-12,
one previously unidentified tRNA located
A more general bacterial tRNA descriptor

- GU pairs generally not allowed
- One mispair per helix is allowed
- One sequence mismatch is allowed
### Results using Optimized Descriptor

<table>
<thead>
<tr>
<th>Organism</th>
<th>total</th>
<th>false</th>
<th>false</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># tRNA</td>
<td>neg.</td>
<td>pos.</td>
</tr>
<tr>
<td>E.Coli K–12</td>
<td>86</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>E.Coli O157:H7</td>
<td>95</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>B. Subtilis</td>
<td>87</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>Aquifex aeolicus</td>
<td>43</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>H. influenzae</td>
<td>56</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Myc. pneumoniae</td>
<td>36</td>
<td>1</td>
<td>6</td>
</tr>
</tbody>
</table>

False positives can be further distinguished from true tRNAs using Turner energies.

(We show below how to eliminate false positives.)
Analysis using nearest-neighbor energies

Turner Energies

true tRNA  false pos.  true tRNA  false pos.

Bacillus subtilis
E. coli (K−12)
E. coli (O157:H7)
Haemophilus influenzae
Aquifex aeolicus
Mycoplasma pneumoniae
Moving to eukaryotes

<table>
<thead>
<tr>
<th>organism</th>
<th>tRNAs</th>
<th>tRNAs with introns</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. pombe</em></td>
<td>153</td>
<td>39</td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>273</td>
<td>59</td>
</tr>
<tr>
<td><em>arabidopsis</em></td>
<td>620</td>
<td>83</td>
</tr>
<tr>
<td><em>drosophila</em></td>
<td>284</td>
<td>15</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>584</td>
<td>34</td>
</tr>
<tr>
<td>human</td>
<td>496</td>
<td>28</td>
</tr>
</tbody>
</table>

*S. cerevisiae:*
- tRNAscan-SE found 275 tRNAs, 59 with introns
- we modified our descriptor to allow an 8-60 base insert after the sixth position of the anticodon loop
- *rnamotif* then found all but 3 of the tRNA’s identified by tRNAscan-SE
- *rnamotif* also found 1301 false positives, but all had high Turner energies
“Solving” the inverse folding problem (yeast):

tRNA folding energies for *S. cerevisiae*

![Graph showing tRNA folding energies for S. cerevisiae](graph.png)