ICB Fall 2009

G4120: Introduction to Computational Biology

Oliver Jovanovic, Ph.D.
Columbia University
Department of
Microbiology & Immunology

Lecture 5
Introduction to
RNA Analysis
October 22, 2009

Analysis of RNA Sequences

Transcription

- **Promoters** (e.g. consensus, sequence matrices, neural nets)
- **Transcriptional Terminators** (e.g. hairpins, TransTerm)
- **Splice Sites** (e.g. consensus, GeneSplicer)

Translation

- **Ribosome Binding Sites** (e.g. consensus, sequence matrices)
- **Coding Region Prediction** (e.g. ORFs, %GC, hexamer frequency, uneven positional base frequency)

Secondary Structure

- **Hairpins** (e.g. DNA Strider, GCG StemLoop)
- **Folding** (e.g. MFold)

Transcriptional Initiators

Prokaryotic Promoters

- Prokaryotic sigma 70 (σ^{70}) promoters are characterized by a **-35** consensus sequence (**TTGACA**) and a **-10** consensus sequence (**TATAAT**) which are respectively located 35 bp and 10 bp upstream of the transcriptional start point, which is labeled +1 (no 0 exists in transcriptional nomenclature).
- The consensus sequences vary for promoters using other sigma factors, and they also vary somewhat from species to species.
- They can be detected by consensus or matrix searching using software such as SeqMatrix or the use of specialized software trained to recognize prokaryotic promoters, such as a neural network.

Eukaryotic Promoters

- Three kinds of promoters exist for each major eukaryotic RNA polymerase (I, II and III):
  - I) **Pol I** promoters are associated with rRNA genes, and have a **GC rich Upstream Control Element** at -170 to -110, and a **core promoter element** at -40 to +20.
  - II) **Pol II** promoters have a **TATA box** at -25, a **CAA initiator sequence** at +1, upstream **GC rich elements**, an upstream **CCAAT box**, and **enhancer elements** that can be kilobases away from the +1.
  - III) **Pol III** promoters consist of at least three types, two with a **pair of control elements** at +50 to +100, and the last type with **three upstream control elements.**
- Although the procedures are more involved, since their structure can be complex and varies from one class to another, they can be detected by consensus or matrix searching or the use of specialized software trained to recognize a particular kind of eukaryotic promoter, such as PROSCAN and Pol3Scan.

E. coli σ^{70} Promoter

Consensus

```
 -35                     -10
TTGACA.................TATAAT
```

Matrix

```
                                    -35 Region
                                    T  T  G  A  C  A
A   11 8  8  7  8  7  3  5  5  0  1  0  14 5  9  5
C   3  4  2  4  4  3  5  2  8  1  1  2  3  11 2  5
G   3  2  4  2  4  5  5  5  5  2  1  17 1  2  3  3
T   4  7  7  8  5  6  8  9  3  17 18 2  4  3  7  9

             Spacer Region
Length   9  10  11  12  13  14  15
         1   6  14   6   1   1   1

                   -10 Region
                   T  A  T  A  A  T
A   4  5  3  4  4  0  20 5 12 11  0  7  4  6
C   5  4  5  4  5  2  0  3  3  4  1  2  7  6
G   2  5  5  8  7  2  0  3  3  3  0  6  5  6
T   10 6  8  5  6  17 1  9  3  4  20 6  5  4
```

Transcriptional Terminators

Rho Independent Prokaryotic Terminators

- These are characterized by a hairpin loop structure followed by a string of approximately six **U**s in the RNA transcript. The hairpin loop is typically centered about 20 to 30 bases upstream of the last nucleotide in the transcript.
- They can be detected by searching for hairpin loops (using DNA Strider's Seek Hairpin function, or the GCG StemLoop program), then looking for a nearby string of **U**s, or by using specialized software such as TIGR TransTerm.

Rho Dependent Prokaryotic Terminators

- These require the trans-acting Rho protein factor, and often have a hairpin loop structure similar to that of a *rho* independent terminator, but no string of **U**s, and are characterized by a *rut* (rho utilization) consensus sequence of roughly 85 nucleotides starting approximately 100 bases upstream of where transcription terminates.
- The *rut* site can be detected by consensus or matrix searching.

Eukaryotic Terminators

- A 20 to 200 nucleotide poly-A tail is added to the 3' end of mRNA approximately 20 bases downstream of a **AAUAAA polyadenylation signal** consensus sequence.
- The signal can be detected by consensus or matrix searching.

Eukaryotic RNA Splice Sites

- Splice sites typically consist of an upstream **AGGU donor site** consensus sequence and a downstream **AGG acceptor site** consensus sequence, from which an intron is spliced out to begin with **GU..** and end in **..AG**. In addition, a **branch site** consensus sequence **UAUAAC** is located 20 to 50 nucelotides upstream of the acceptor site.
- A fair amount of variation exists. Some donor sites end in **..GC**, and the exact residues conserved in a site can vary from species to species.
- At least one alternative splicing pathway exists which prefers different donor and acceptor sites (spliced to begin with **AU..** and end in **..AC**).
- Splice sites can be detected through consensus searching, matrix searching with software such as SeqMatrix, or the use of specialized software trained on a particular species such as GeneSplicer.

Translational Initiation of RNA

Prokaryotic Translation Initiation

- In prokaryotes, the **AGGAGG** Shine-Dalgarno consensus sequence is located 4 to 7 nucleotides 5' of the translation initiator **AUG** of most mRNAs.
- The sequence is complementary to a **CCUCCU** sequence at the 3' end of 16S rRNA.
- Other residues in the translation initiation region are also conserved, varying in detail from species to species.
- Prokaryotic translation initiators can be detected by consensus or matrix searching for the Shine-Dalgarno sequence, start codon, and other conserved residues using software such as SeqMatrix, or by the use of specialized software trained on prokaryotic translation initiation sequences such as the W101 Preceptron neural network algorithm.

Eukaryotic Translation Initiation

- In eukaryotes, translation typically initiates at the 5' **AUG** in the mRNA, and is not obviously complementary to rRNA, but also features other conserved residues.
- Initiators can be detected by consensus or matrix searching for the start codon and other conserved residues, or the use of specialized gene finding software trained on eukaryotic sequences.

Prokaryotic Ribosome Binding Sites

Shine-Dalgarno Consensus

AGGAGG....AUG

Plasmid RK2 Ribosome Binding Site Matrix

		A	G	G	A	G	G							C	A	A	U	G	A	A		
A	5	8	6	4	8	8	3	7	7	6	6	7	8	2	7	21	0	0	14	10	7	8
C	6	5	3	3	4	1	2	3	3	5	2	7	7	12	5	0	0	0	3	8	5	4
G	6	5	10	12	9	12	15	7	6	4	7	4	4	5	4	0	0	21	3	3	3	9
U	4	3	2	2	0	0	1	4	5	6	6	3	2	2	5	0	21	0	1	0	6	0

Prokaryotic Translation Initiation Regions

E. coli Translation Initiation Regions

	A	A	U	U	A	U	G	G	C	U	A
A	6	8	5	5	14	0	0	6	5	6	7
C	3	3	3	3	0	0	0	1	6	2	4
G	2	2	2	2	1	0	15	6	2	1	3
U	4	2	5	5	0	15	0	2	2	6	1

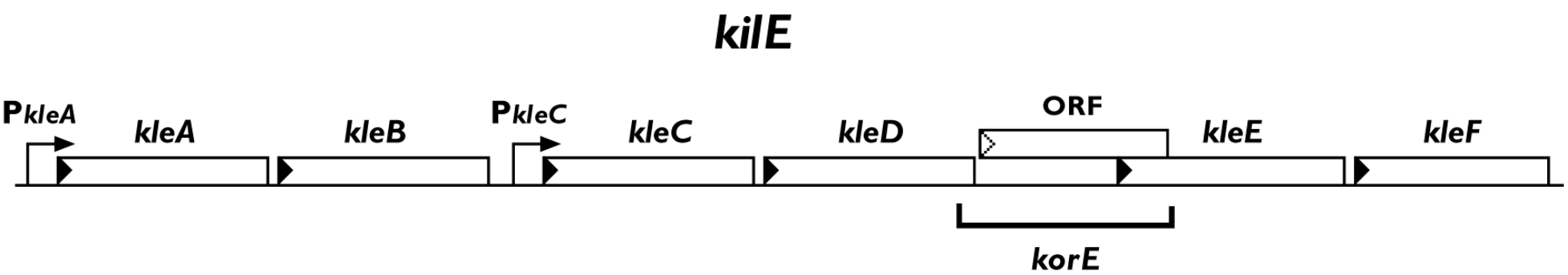
Plasmid RK2 Translation Initiation Regions

	C	A	C	A	A	U	G	A	A	A	G
A	7	8	2	7	21	0	0	14	10	7	8
C	7	7	12	5	0	0	0	3	8	5	4
G	4	4	5	4	0	0	21	3	3	3	9
U	3	2	2	5	0	21	0	1	0	6	0

Coding Region Prediction

- Due to the restraints imposed on them by having to code for the triplets of the genetic code, coding regions differ from non-coding regions in their nucleotide distributions.
- As a result, certain forms of computational analysis, including %G+C content, codon usage, hexamer frequency analysis, hidden Markov models and uneven positional base frequency can theoretically distinguish coding regions from non-coding regions.
- Coding sequences typically have a nonrandom distribution of bases, with an uneven distributions of bases in each codon position.
- This allows for more sophisticated prediction of coding regions and mRNAs than simply looking for an ORF, which may or may not actually be a coding region.
- It can also be useful when trying to identify a regulatory RNA or RNA with some function other than an mRNA.
- Various free software packages allow one to perform such analysis, including the Staden package, EMBOSS, and web based programs such as FramePlot, GeneHacker, GeneMark, GENSCAN and GRAIL.

RK2 *kilE* Locus



The linear map corresponds to a 2 kb region from 4,401 to 2,400 nt on the published RK2 plasmid sequence containing the operons of the kilE *locus. The minimal 201 bp* korE *region is marked with a bracket. The putative 57 aa ORF initially predicted to code for* korE *is marked with a box and a hollow triangle indicating its predicted direction of translation.*

RNA Secondary Structure Folding Prediction

- Individual hairpin loops are simple to predict (e.g. using DNA Strider Seek Hairpins) but may not accurately reflect the structure of an entire RNA.
- Accuracy of prediction is limited by accuracy of free energy calculations. Different tables of free energy values give different results.
- *In vitro* folding may be affected by the direction of synthesis, interaction of unpaired loops, interaction with other RNAs, interaction with other proteins, or environmental conditions.
- Modification of bases, such as which occurs with tRNAs, may affect folding.
- There may be multiple structures with similar or equally favorable free energy states.
- The structure prediction with the lowest free energy is not necessarily the functional conformation of the RNA.

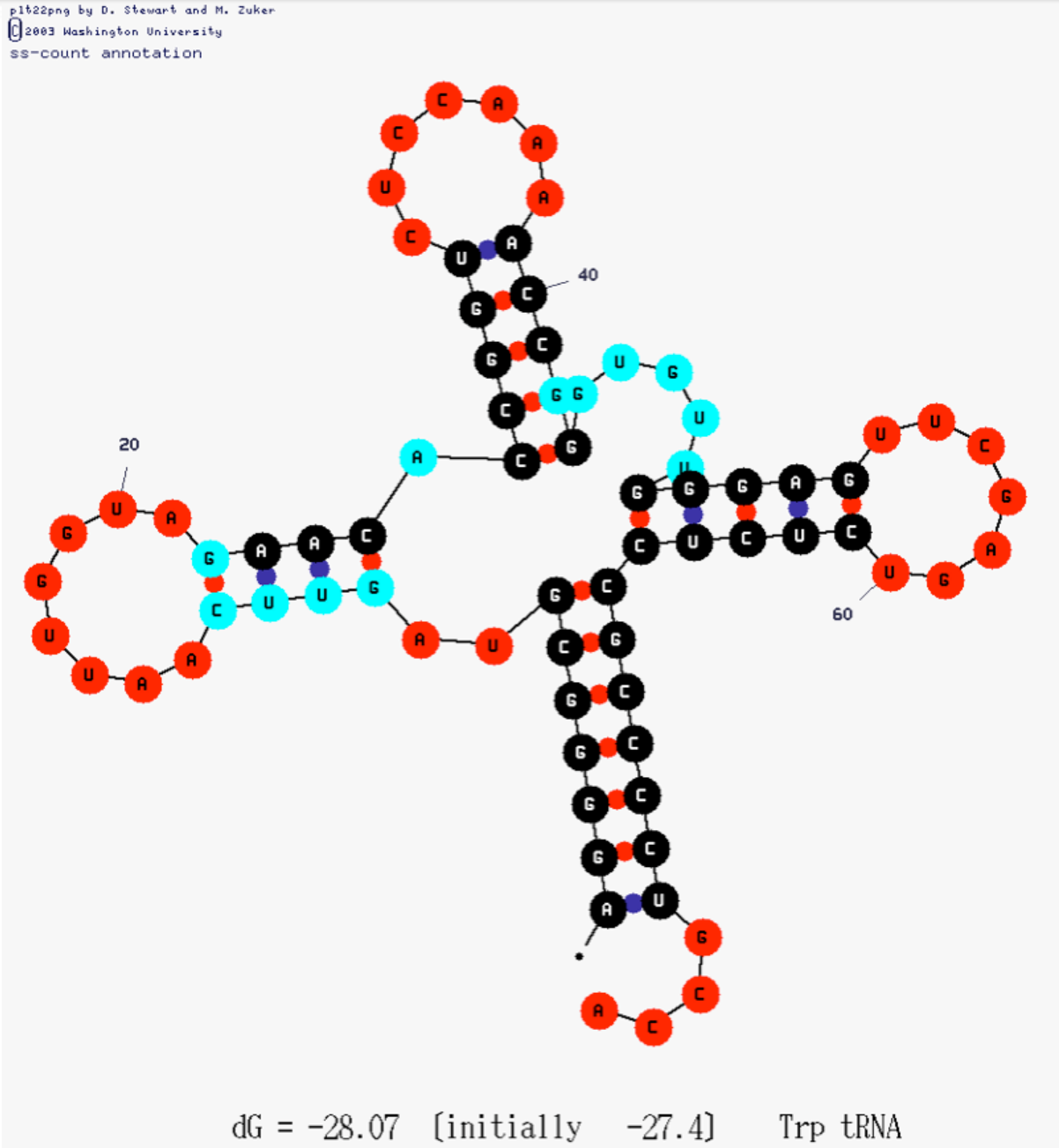
RNA Secondary Structure Prediction with Mfold

Mfold (**Michael Zucker, RPI**)

- Predicts optimal and suboptimal RNA secondary structures based on a table of free energy, looking for the structure with the lowest total free energy (ΔG) value.
- Tries to find the base pairings that vary the least in a series of optimal and suboptimal RNA secondary structures.
- Can also be used to find DNA secondary structures.
- Can run Mfold and display output with PlotFold as UNIX applications.
- Can run Mfold on the web at:

  http://frontend.bioinfo.rpi.edu/applications/mfold/cgi-bin/rna-form1.cgi

Tryptophan Transfer RNA Mfold Secondary Structure Prediction



Lecture 5
Introduction to RNA Analysis
October 22, 2009