Bioinformatics: Basic Facts...

Bioinformatics in a Nutshell

Wikipedia[1] defines *bioinformatics* as follows:

> **Bioinformatics** is the application of information technology to the field of molecular biology.

The article continues:

> *Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data.*

While some of the problems addressed in the field of bioinformatics require years of education and advanced degrees in **both** Biology and Computer Science to study and solve, a *wide range* of bioinformatics-related tasks can be performed using relatively simple programs. This lab offers five such problems for you to solve.

Bioinformatics: the Data

Among the wide range of bioinformatics applications, we will concentrate on the problems associated with the field of *genomics*, i.e., the study of genomes of organisms. In particular, one of the key subject matters of *genomics* is the discovery of the DNA Sequences of various organisms.

DNA a.k.a. **deoxiribonucleic acid** is a molecule that contains the *genetic instructions* used in the development and functioning of all known living organisms[2].

DNA is an extremely large and complex molecule. It consists of a large number of simpler *components* called **nucleotides**. A nucleotide molecule consists of a three components, two of which, *the phosphate* and *the sugar* have the same

[1] http://en.wikipedia.org/wiki/Bioinformatics
[2] http://en.wikipedia.org/wiki/DNA

chemical structure for all nucleotides. The third component, **the base** is the one that distinguishes different nucleotides. There are four types of **nucleotide bases** present in DNA molecules:

- Adenine
- Cytosine
- Guanine
- Thymine

DNA Structure. One of the greatest scientific discoveries of the 20th century was the discovery of the structure of a DNA molecule. A DNA molecule is a **double helix** consisting of **two strands** of **nucleotides**[3]. One of the key properties of DNA is **base pairing**: the four nucleotides form specific pairings:

- If one strand has an Adenine base, then the corresponding position on the other strand is occupied by the Thymine base and vice versa.
- If one strand has a Cytosine, the the corresponding position on the other strand is occupied by Guanine and vice versa.

The **base pairing** property means that

> *A single DNA strand uniquely determines the full structure of a DNA molecule.*

The Adenine – Thymine and Cytosine – Guanine pairs of nucleotide bases are called base pairs. The two strands are called complementary to each other.

DNA Sequences. A **DNA Sequence**, is a representation of a DNA molecule as a string. In particular, a single DNA sequence represents one strand of a DNA molecule. The four nucleotides found in DNA are encoded with four letters according the following coding table:

Nucleotide base	Letter
Adenine	A
Cytosine	C
Guanine	G
Thymine	T

Thus, a sequence of Adenine, Adenine, Cytosine, Thymine, Guanine, Thymine will be encoded as a DNA sequence "AACTGT".

DNA Sequence Directionality and reverse complement. A strand on a DNA molecule has **orientation**. The two ends of a strand have different chemical structure, and this allows researchers to label them and to represent all DNA sequences as being headed in the same direction.

The two ends of a DNA strand are referred to as **5′** ("five prime") and **3′** ("three prime") – an allusion to the chemical structures found on each end.

[3] In addition to the double helix structure, a DNA molecule may have a non-trivial three-dimensional structure, but the problems we will be studying in this assignment do not take this into account.

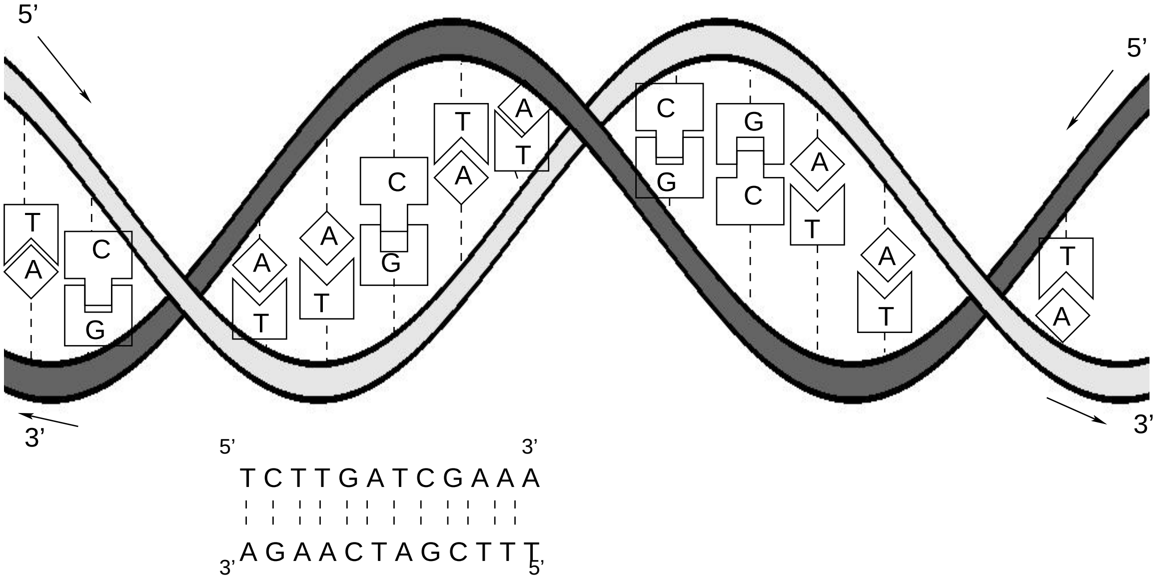

Figure 1: A double-helix DNA fragment and the DNA sequences representing it.

Direction of DNA sequences. All DNA Sequences are presented starting at the **5′** end and ending at the **3′** end.

DNA strands in a helix have opposite orientation. The two DNA strands have opposite orientation.

Figure 1 represents a DNA molecule fragment. The orientation of the light-gray DNA strand is left-to-right. The orientation of the dark-gray DNA strand is right-to-left. The the light-gray DNA strand contains the following sequence of nucleotides:

TCTTGATCGAAA

The dark-gray DNA strand contains the complementary sequence AGACTAGCTTT, **however**, when shown in this order, the sequence runs from the **3′** end to the **5′** end. To correctly represent this sequence, we must **reverse** it, i.e., read it from right to left. The resulting sequence of nucleotides,

TTTCGATCAAGA

has the correct orientation and is called the **reverse complement** of the sequence TCTTGATCGAAA.

reverse complements. In general given a DNA sequence $S = L_1 L_2 \dots L_N$, where L_i represents one of the four letters A,T,C,G, its **reverse complement**

is the DNA sequence that must occupy the corresponding bases on the other DNA strand, ordered from **5′** to **3′**.

Remember that we have the following complement relationship between the nucleotides:

$$\boxed{\begin{array}{l}\mathsf{compl}(\mathsf{A}) = \mathsf{T}\\ \mathsf{compl}(\mathsf{T}) = \mathsf{A}\\ \mathsf{compl}(\mathsf{C}) = \mathsf{G}\\ \mathsf{compl}(\mathsf{G}) = \mathsf{C}\end{array}}$$

To construct an reverse complement of a sequence $S = L_1 L_2 \ldots L_N$, we perform two steps:

1. **Step 1: complement.** Each letter L_i in the sequence is replaced with its complement $\mathsf{compl}(L_i)$. The resulting sequence is $S^C = \mathsf{compl}(L_1)\mathsf{compl}(L_2)\ldots\mathsf{compl}(L_N)$.
2. **Step 2: Reversion:** The sequence S^C is rewritten **from right to left** to form the **reverse complement** sequence

$$S^I = \mathsf{compl}(L_N)\mathsf{compl}(L_{N-1})\ldots\mathsf{compl}(L_2)\mathsf{compl}(L_1).$$

From Nucleotide Sequences to Amino Acids

Amino Acids. Amino acids are molecules that are used in building proteins. Amino acids are found in many other compounds in living organizms, and they play important roles in a variety of biochemical processes. More importantly (for us), amino acids form the next layer in the hierarchy of DNA encoding.

There are twenty amino acids. The table with their full names, three-letter abbreviations and (more importantly), single-letter codes is shown in Table 1.

Certain sequences of amino acids form **proteins**.

Genetic Code. In DNA molecules, **each triple of consecutive *nucleotides* encodes one *amino acid*** or serves as a special **marker**. The triple of nucleotides is called a codon.

The **translation** of the nucleotide triples (codones) into amino acid codes is called **the genetic code**.

With four nucleotides, there are **64 possible codons** that encode 20 amino acids and two special markers. A single amino acid can be encoded by multiple codons.

Table 2 shows the **genetic code**.

Start and Stop codons. Three codons, TAA, TAG and TGA do not encode any amino acids. Instead, these codons represent **the end of a protein molecule** encoded by a DNA sequence. They are called stop codons.

Additionally, the ATG codon encodes Methionine, the amino acid that serves as the **beginning of encodings of all proteins** within a DNA sequence. Whenever Methionine is found at the beginning of a protein, the ATG sequence encoding it is called a start codon.

Amino Acid	Three-letter code	One-letter code
Alanine	*Ala*	A
Arginine	*Arg*	R
Asparagine	*Asn*	N
Aspartic acid	*Asp*	D
Cysteine	*Cys*	C
Glutamic acid	*Glu*	E
Glutamine	*Gln*	Q
Glycine	*Gly*	G
Histidine	*His*	H
Isoleucine	*Ile*	I
Leucine	*Leu*	L
Lysine	*Lys*	K
Methionine	*Met*	M
Phenylalanine	*Phe*	F
Proline	*Pro*	P
Serine	*Ser*	S
Threonine	*Thr*	T
Tryptophan	*Trp*	W
Tyrosine	*Tyr*	Y
Valine	*Val*	V

Table 1: Amino Acids.

First Nucleotide	Second Nucleotide			
	T	C	A	G
T	TTT → *Phe* F	TCT → *Ser* S	TAT → *Tyr* Y	TGT → *Cys* C
	TTC → *Phe* F	TCC → *Ser* S	TAC → *Tyr* Y	TGC → *Cys* C
	TTA → *Leu* L	TCA → *Ser* S	TAA → **Stop**	TGA → **Stop**
	TTG → *Leu* L	TCG → *Ser* S	TAG → **Stop**	TGG → *Trp* W
C	CTT → *Leu* L	CCT → *Pro* P	CAT → *His* H	CGT → *Arg* R
	CTC → *Leu* L	CCC → *Pro* P	CAC → *His* H	CGC → *Arg* R
	CTA → *Leu* L	CCA → *Pro* P	CAA → *Gln* Q	CGA → *Arg* R
	CTG → *Leu* L	CCG → *Pro* P	CAG → *Gln* Q	CGG → *Arg* R
A	ATT → *Ile* I	ACT → *Thr* T	AAT → *Asn* N	AGT → *Ser* S
	ATC → *Ile* I	ACC → *Thr* T	AAC → *Asn* N	ACC → *Ser* S
	ATA → *Ile* I	ACA → *Thr* T	AAA → *Lys* K	AGA → *Arg* R
	ATG → *Met*/**Start** M	ACG → *Thr* T	AAG → *Lys* K	AGG → *Arg* R
G	GTT → *Val* V	GCT → *Ala* A	GAT → *Asp* D	GGT → *Gly* G
	GTC → *Val* V	GCC → *Ala* A	GAC → *Asp* D	GGC → *Gly* G
	GTA → *Val* V	GCA → *Ala* A	GAA → *Glu* E	GGA → *Gly* G
	GTG → *Val* V	GCG → *Ala* A	GAG → *Glu* E	GGG → *Gly* G

Table 2: Genetic Code.

Translation from Nucleotide sequences to Amino Acid sequences. Using the **genetic code** table, any DNA sequence written in the alphabet of nucleotides can be translated into a sequence of amino acids. To do this,

1. Start at the **5′** end of the DNA sequence.
2. For each triple of nucleotides, find, using the **genetic code table**, the matching amino acid (or start/stop codon).
3. Write out the amino acids and/or start/stop codons in a sequence following the **5′** to **3′** order.

Example. Consider the following DNA sequence:

TCTTGATCGAAA

We split it into triples of nucleotides:

TCT TGA TCG AAA

We then use the **genetic code** table, to substitute each triple with the matching amino acid:

TCT	TGA	TCG	AAA
S	**Stop**	S	K

Frames. Typically, DNA sequences represent portions of a DNA molecule. DNA molecules consist of millions of individual nucleotides, but current *DNA sequencing equipment* is capable of sequencing (i.e., discovering from a sample) only small "chunks" at a time. Given a DNA sequence in a nucleotide alphabet, there are three possibilities for translating it into a sequence of amino acids. These possibilities are called **frames**.

1. **Frame 1.** First codon starts at the first nucleotide.
2. **Frame 2.** First nucleotide is **ignored**; first codon starts at the second nucleotide.
3. **Frame 3.** First and second nucleotides are ignored; first codon start at the third nucleotide.

Intuitively, the frames can be explained as follows. Given a DNA sequence, we do not know whether it starts at the beginning of an amino acid encoding, or in the middle of it. There are three cases, and in order to translate a DNA fragment into a sequence of amino acids, we need to consider all possibilities. The three **frames** indicate where the first full amino acid of the fragment starts: at the beginning of the fragment, at the second nucleotide or at the third nucleodtide.

Example. Consider the following DNA sequence:

TCTTAATCGAATCGAT

This sequence can be split into codons in three different ways:

Frame 1: TCT TAA TCG AAT CGA T
Frame 2: T CTT AAT CGA ATC GAT
Frame 3: TC TTA ATC GAA TCG AT

In translating the DNA sequence in three frames, any nucleotides that are not part of a codon are ignored. The remaining codons are translated into amino acids using the **genetic code table**:

Frame 1:	TCT	TAA	TCG	AAT	CGA	T	
	S	**Stop**	S	N	R		
Frame 2:	T	CTT	AAT	CGA	ATC	GAT	
		L	N	R	I	D	
Frame 3:	TC	TTA	ATC	GAA	TCG	AT	
		L	I	E	S		

So, the same DNA sequence, may give rise to three different sequneces of amino acids: "S**Stop**SNR", "LNRID" and "LIES"[4], depending on which frame is actually correct.

Three more frames. DNA molecule has two strands. Given a DNA sequence that comes from one of the strands, it is possible that either this sequence, **or the corresponding sequence from the other strand** participates in describing a protein. Therefore, given a DNA sequence, we may need to convert

[4]This was not intentional.

both it, and **its reverse complement** to the amino acid sequences. Each of the two sequences (direct and the reverse complement) can be converted using three frames. Therefore, the total number of frames, i.e., plausible amino acid sequences represented by a DNA fragment is six: three for the fragment itself, and three for its **reverse complement**.

Example(continued). Consider again, the DNA sequence

TCTTAATCGAATCGAT

The example above shows the three frames of translation of this fragment into amino acid sequences. We now complete the full list of translations, by using the reverse complement of this fragment.

The reverse complement of the fragment is

Sequence:	TCTTAATCGAATCGAT
complement sequence:	AGAATTAGCTTAGCTA
reverse complement:	ATCGATTCGATTAAGA

The three frames for the reverse complement are:

Frame 4: ATC GAT TCG ATT AAG A
Frame 5: A TCG ATT CGA TTA AGA
Frame 6: AT CGA TTC GAT TAA GA

The translation to the sequences of amino acids is:

Frame 4:	ATC	GAT	TCG	ATT	AAG	A
	I	D	S	I	K	
Frame 5:	A	TCG	ATT	CGA	TTA	AGA
		S	I	R	L	R
Frame 6:	AT	CGA	TTC	GAT	TAA	GA
		R	F	D	Stop	

Combining the results of the two examples, here are the six amino acid sequences that may be encoded by the given DNA fragment:

S<Stop>SNR
LNRID
LIES
IDSIK
SIRLR
RFD<Stop>

Structure of DNA

Living organisms are subdivided into two groups: *prokaryotes* and *eukaryotes*.

Prokaryotes: a group of organisms that lack a cell nucelus. Most of prokaryotes are single-cell organisms. Prokaryotes form two taxonomic domains, *bacteria* and *archaea*.

Eukaryotes: a group of organisms whose cells envelop the genetic material (DNA) in a nucleus and that contain outher complex structures enclosed within membranes. Eukaryotesd form a domain of life by themselves. This domain includes three (or four – depending on who you ask) kingdoms: plants, fungi and animals (the fourth one is *Protista* - a kingdom (?) of simple eukaryotic life forms like amoebas).

Chromosome. An organized structure of DNA found in cells. Chromosomes contain *genes*, *regulatory elements* and other nucleotide sequences.

Different species have different number of chromosomes present. The full set of chromosomes constitutes full, or major portion of the organism's *genome*.

Prokaryotes usually have *one circular chromosome*. *Eukaryotes* have multiple *linear chromosomes*.

Gene. Molecular unit of heredity in a living organism. Genes are encoded in the DNA stored in the cells of the organism.

Genes in *Prokaryotes* are a single "chunk" of the DNA.

Genes in *Eukaryotes* consist of multiple chunks, called **exons** separated by *non-coding DNA* fragments called *introns*. Most of the genes have multiple *exons*.

Regulatory elements. Segments of DNA, where *regulatory proteins* (i.e., proteins that bind to a DNA molecule) bind (attach themselves) preferentially. *Regulatory elements* are usually found short distance "upstream" of the gene, with which they are associated.

On-line Resources

A wide range of genomic databases and other resources is available for bioinformatics researchers and biologists. Some of the resources commonly used for a variety of tasks are:

1. **NCBI.** http://www.ncbi.nlm.nih.gov/. The genome database of the National Center for Biotechnology Information.

2. . **UCSC Genome Browser.** http://genome.ucsc.edu/. University of California at Santa Cruz on-line genome database complete with a tool for visualizing genomes.

Simple Problems

Most of the problems in this section occur naturally, when other, more complex, problems associated with working with DNA strings need to be solved.

The Reverse Complement

Take as input a DNA string in the nucleotides alphabet and output the **reverse complement** of this string.

Example. Consider the following input string:

```
ATTCCATGG
```

The output shall be.

```
CCATGGAAT
```

Conversion of Nucleotide Sequences into Amino Acid Sequences

Apply the **genetic code** to convert a DNA sequence in a nucleotide alphabet into **all possible** amino acid sequences it represents. Produce six strings in the amino acid alphabet, each string representing a translation of the input DNA sequence into an amino acid sequence on *one of the six possible frames.*

Example. If the input is

```
TCTTAATCGAATCGAT
```

then the output shall be:

```
S#SNR
LNRID
LIES
IDSIK
SIRLR
RFD#
```

(see excamples above for the translation).

Finding Palindromes

In linguistics, a **palindrome** is a string (sentence of text) that spells the same when read from left-to-right and from right-to-left. "Meaningful" palindromes exist in essentially every human language. Some examples of English palindromes are:

```
eye
Anna
tenet
rotator
redivider
madam
madam I'm Adam
never odd or even
```

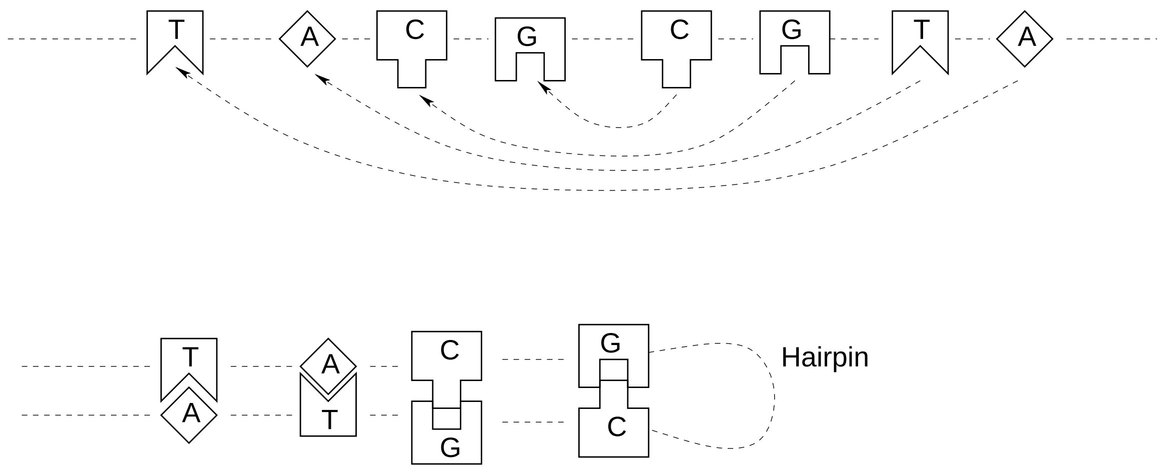

Figure 2: Palindromes in DNA sequences form *hairpins.*

```
murder for a jar of red rum
rats live on no evil star
God saw I was dog
Doc, note: I dissent. A fast never prevents a fatness. I diet on cod.
```

Palidromes in genomics. In genomics, a version of palindromes plays an important role in determining the 3D structure of a DNA molecule.

> A DNA palindrome is a sequence of characters in the nucleotide alphabet $\{A, C, G, T\}$, such that **it is equal to its reverse complement.**

Example. Consider a sequence

AGTACT

It's complement is TCATGA. It's reverse complement is AGTACT, i.e., the original string.

The biological significance of the palindromes in DNA sequences is illustrated in Figures 2 and 3. Essentially, because if a palindrom is present in a DNA sequence on some strand, the DNA strand may become *entangled* in this place, as the nucleotides in the palindrome sequence may get "paired" with their complements in the palindrome sequence, rather than with the nucleotides on the second strand. Such entanglements, called hairpins are common in DNA sequences, and biochemists want to be able to find where they occur in the DNA.

Problem. Develop an algorithm takes as input a DNA sequence string in nucleotide alphabet, and reports all distinct palindroms found in it.

Multiple Sequence representation

A common occurrence in genomics is the situation when a biologist needs to look at a large number of DNA sequences encoding the same DNA region in different species (or different individuals of the same species). While more important and difficult problem of proper *alignment* of DNA sequences is one of

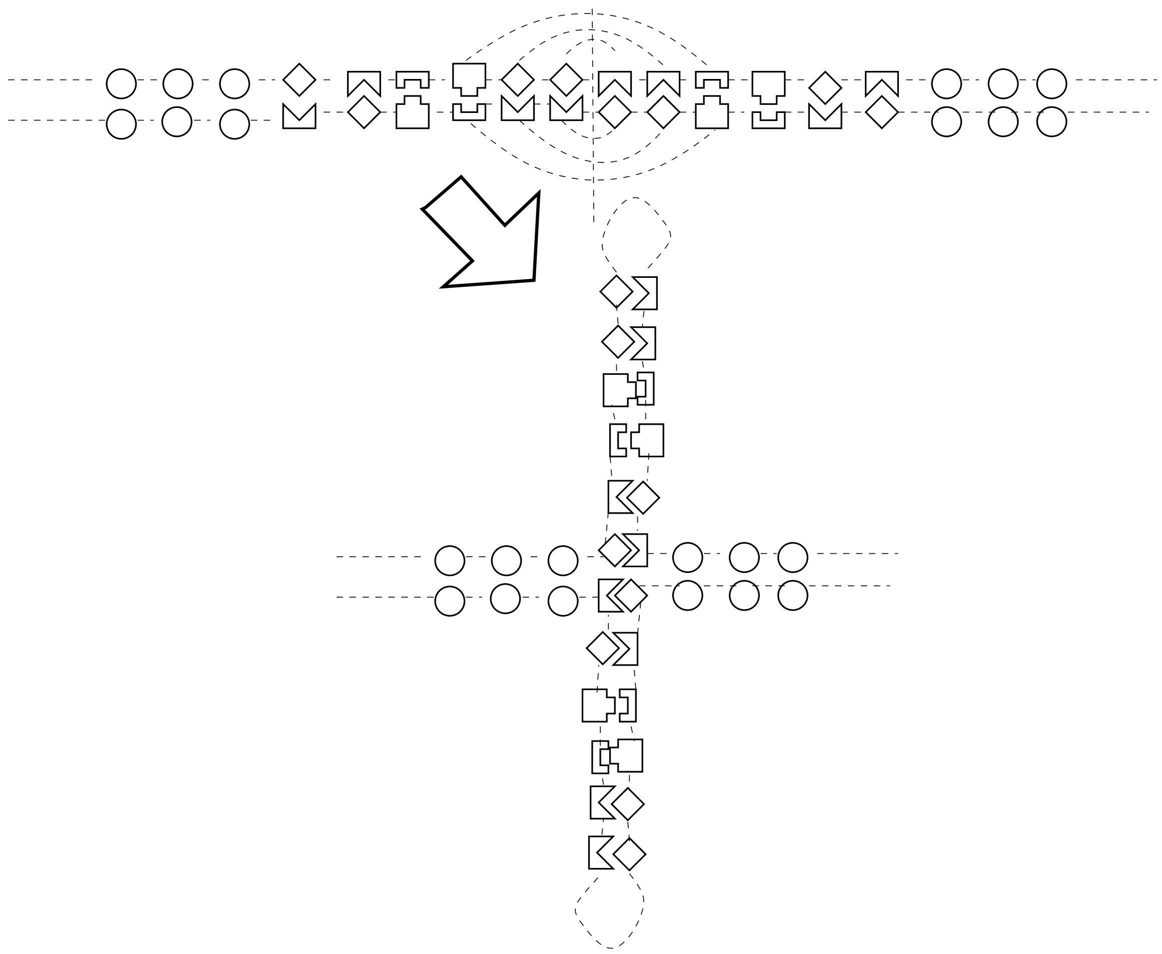

Figure 3: Palindromes in DNA sequences form *hairpins.* (part 2)

the most important problems we will be discussing throughout the quarter, here we consider a more simple task. We are given n already aligned sequences and are asked to

- Characterize the level of agreement of the sequences at each position.
- Determine the *consensus sequence* representing these DNA sequences.

We characterize the sequences, by creating a *histogram* of frequency counts of different nucleotide letters at each position. The create a *consensus* string by selecting, as a representative for each position, the nucleotide that appears in the plurality of strings.

For example, consider the following input.

```
ATATC
ATTTC
AATTC
ATGTC
ATTTC
ATCTC
AAAAC
AAGAC
ATATC
AATTC
```

All fragments have matching first and fifth nucleotides. In 80% of fragments the fourth nucleotide is T, while in the remaining 20%, it is A. The second nucleotide is evenly split between T and A, while the third nucleotide shows the least amount of consistency with the T: 40%, A: 30%, G: 20% and C: 10% distribution.

The Dot Plot

One of the most common problems studied by genome scientists is that of matching different DNA sequence fragments. The fragments may have come from the same DNA molecule (e.g., the actual DNA sequencing problem is the problem of combining multiple short DNA fragments into a longer one by arranging them in an appropriate order) or from DNA molecules of different spcecies (e.g., when trying to find how similar the DNA of two different species is).

In both cases, a construct called a **dot plot** can be utilized by the genome scientists to obtain a quick visualization of the similarity between two DNA fragments.

Dot Plot. Given two DNA sequences $S_1 = l_1 \ldots l_N$ and $S_2 = t_1 \ldots t_M$, where l_i and t_j come from **either** the nucleodite alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ **or** the amino acid alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{D}, \mathsf{E}, \mathsf{F}, \mathsf{G}, \mathsf{H}, \mathsf{I}, \mathsf{K}, \mathsf{L}, \mathsf{M}, \mathsf{N}, \mathsf{P}, \mathsf{Q}, \mathsf{R}, \mathsf{S}, \mathsf{T}, \mathsf{V}, \mathsf{Y}, \mathsf{W}\}$, the **simple dot plot** of S_1 vs. S_2 is a two-dimensional table $D = \{d_{ij}\}$, $i = 1 \ldots N$, $j = 1 \ldots M$, where

$d_{ij} = 1$ if $l_i = t_j$;
$d_{ij} = 0$ otherwise.

Example. Consider two DNA sequences $S_1 = \mathsf{TCAAGA}$ and $S_2 = \mathsf{TCGATT}$. Their **simple dot plot** is

	T	C	G	A	T	T
T	1	0	0	0	1	1
C	0	1	0	0	0	0
A	0	0	0	1	0	0
A	0	0	0	1	0	0
G	0	0	1	0	0	0
A	0	0	0	1	0	0

or

	T	C	G	A	T	T
T	1				1	1
C		1				
A				1		
A				1		
G			1			
A				1		

A general notion of a **dot plot** involves setting $d_{ij} = 1$ if **sequences around** l_i **and** t_j **are similar**. Examples or such *similarity are*:

- $l_{i-1}l_il_{i+1} = t_{j-1}t_jt_{j+1}$;
- $l_{i-2}l_{i-1}l_il_{i+1}l_{i+1} = t_{j-2}t_{j-1}t_jt_{j+1}t_{j+2}$;