Contents

Chapter 1

A Biology Primer

1.1 The Central Dogma of Molecular Biology

The central dogma of molecular biology explains the flow of genetic information in the cell between between information-carrying biopolymers (DNA, RNA and protein). It states that the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible.

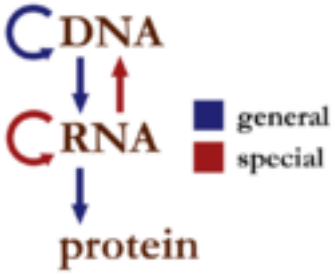

Figure 1.1: Information flows between DNA, RNA and protein. Source: Wikipedia

The genetic code of an organism is stored in DNA, which is converted into portable RNA messages in a process called transcription. These messages travel from the cell nucleus (where the DNA resides) to the ribosomes where they are used as template to make specific proteins in a process called translation. The central dogma states that the pattern of information that occurs most frequently in our cells is:

- From existing DNA to make new DNA (replication)
- From DNA to make new RNA (transcription)
- From RNA to make new proteins (translation).

Besides these, there are some notable possibilities. For instance, retroviruses are able to generate DNA from RNA via reverse-transcription, and some viruses use RNA to make protein. All this is shown in Figure 1.1. The generated proteins carry out most of the cellular functions such as metabolism, DNA regulation, and replication.

1.2 DNA

1.2.1 Function

The DNA molecule stores the genetic information of an organism. DNA contains regions called genes, which encode for the proteins that carry out most of the cellular function. Other regions of the DNA contain

regulatory elements, which partially influence the level of expression of each gene.

1.2.2 Structure

The DNA molecule consists of two strands that wind around to form a shape known as a double helix. Each strand has a backbone made of alternating sugar (deoxyribose) and phosphate groups. Attached to each sugar is one of the four bases: adenine, cytosine, guanine, and thymine, frequently represented using the letters A, C, G, and T respectively. The two strands are held together by bonds between the bases: A and T are connected by two hydrogen bonds, while C and G are connected by three bonds. This specificity in pairing means that one strand can be used as a template to generate the other strand.

The DNA strands also have directionality, which refers to the positions of the pentose ring where the phosphate backbone connects. This directionality convention comes from the fact that DNA and RNA polymerase synthesize in the 5' to 3' direction. The complementary pairing with directionality means that the DNA strands are anti-parallel. In other words the 5' end of one strand is adjacent to the 3' end of the other strand. As a result, DNA can be read both in the 3' to 5' direction and the 5' to 3' direction, and genes and other functional elements can be found in each direction (on either strand). By convention, DNA is written from 5' to 3'.

Base pairing between nucleotides of DNA constitutes its primary and secondary structure. In addition to DNA's secondary structure, there are several extra levels of structure that allow DNA to be tightly compacted and influence gene expression. The tertiary structure describes the twist in the DNA ladder that forms a helical shape. In the quaternary structure, DNA is tightly wound around small proteins called histones. These DNA-histone complexes are further wound into tighter structures seen in chromatin.

1.2.3 Replication

The structure of DNA with its weak hydrogen bonds between the bases in the center allows the strands to easily be separated for the purpose of DNA replication. In the replication of DNA, the two complementary strands are separated, and each of the strands are used as templates for the construction of a new strand. DNA polymerases attach to each of the strands at the origin of replication, reading each existing strand from the 3' to 5' direction and placing complementary bases such that the new strand grows in the 5' to 3' direction. Because the new strand must grow from 5' to 3', one strand (leading strand) can be copied continuously, while the other (lagging strand) grows in fragments that are later pasted together by DNA ligase. The end result is 2 double-stranded pieces of DNA, where each is composed of 1 old strand, and 1 new strand. For this reason, DNA replication is semi-conservative.

1.3 Transcription

1.3.1 mRNA generation

Transcription is the process to produce RNA using a DNA template. The DNA is partially unwound to form a bubble, and RNA polymerase is recruited to the transcription start site (TSS) by regulatory protein complexes. RNA polymerase reads the DNA from the 3' to 5' direction and placing down complementary bases to form messenger RNA (mRNA). RNA uses the same nucleotides as DNA, except Uracil (U) is used instead of Thymine (T).

1.3.2 Post-transcriptional modifications

Messenger RNA (mRNA) in eukaryotes experience post-translational modifications, or processes that edit the mRNA strand further. Most notably, a process called splicing removes introns (intervening regions which don't code for protein), so that only the coding regions (the exons), remain. Different regions of the primary transcript may be spliced out and each can lead to a different protein product. This phenomenon is referred to as alternative splicing. In this way, an large number of protein products can be generated based on different splicing permutations. In addition to splicing, both ends of the mRNA molecule are processed. The

5’ end is capped with a modified guanine nucleotide. At the 3’ end, roughly 250 adenine residues are added to form a poly(A) tail.

Chapter 2

Epigenomics1

From Professor Chongzhi Zang's lecture slides "Regulatory DNA, Transcription factors, Sequence motifs". Scribed by Zhaoxia Ma.

2.1 Epigenome

Different cells with similar genome sequences have different genes expression. The epigenome can control gene activities to decide which genes are turned on or off.

> "The epigenome is a multitude of chemical compounds that can tell the genome what to do. The epigenome is made up of chemical compounds and proteins that can attach to DNA and direct such actions as turning genes on or off, controlling the production of proteins in particular cells."
>
> —from genome.gov

2.2 Epigenomic marks

	Chemical compounds	Proteins	Other molecules
DNA-associated	DNA methylation	Histones; DNA-binding proteins (transcription factors*)	RNA(e.g., R loops)
Chromatin-associated	Histone modifications: methylations, acetylations, …	Histone variants; Chromatin regulators; Histone modifying enzymes: writer, readers, erasers; Chromatin remodeling complexes	Non-coding RNAs

In addition to DNA-associated and chromatin-associated epigenomic marks, there are some other information served as epigenomic marks, such as nucleosome positioning, chromation accessibility and 3D genome orgenization, etc.

2.3 Epigenomics

2.3.1 Transcription factors

- The transcription factors (TF) is one of the most important group of proteins which can directly interact with DNA. In this case, the DNA region should be open/accessible to proteins.
- The transcription factors should have DNA-binding domains which are used to recognize specific DNA sequences and sites. They also have effector domains which can regulate TF activity, such as ligand binding domains, can mediate protein-protein interactions, such as BTB domain, and can have enzymatic activities, such as SET domain.
- There are some functional studies related to transcription factors, such as studying for the cell-type specific gene expression, binding DNA sequence motif, genome-wide binding sites, target genes, TF co-factors, etc.

2.3.2 Sequence motif

In general, a motif is a distinctive pattern that occurs repeatedly. In biomelecular studies, a sequence motif is a pattern common to a set of DNA, RNA, or protein sequences that share a common biological property, such as functioning as binding sites for a particular protein.

2.3.3 Sequence motif finding

For motif finding, the input data is a set of DNA sequences and the output is enriched sequence patterns (motifs). - Motif representation

Single-letter and ambiguity codes for nucleotides (Table)

Symbol	Meaning	Origin of designation
G	G	**G**uanine
A	A	**A**denine
T	T	**T**hymine
C	C	**C**ytosine
R	G or A	pu**R**ine
Y	T or C	p**Y**rimidine
M	A or C	a**M**ino
K	G or T	**K**eto
S	G or C	**S**trong interaction (3H bonds)
W	A or T	**W**eak interaction (2H bonds)
H	A or C or T	not-G, H follows G in the alphabet
B	G or T or C	not-A, B follows A
V	G or C or A	not-T (not-U), V follows U
D	G or A or T	not-C, D follows C
N	G or A or T or C	a**N**y

It is derived by IUPAC. one limitation is that they can not measure continue relation/difference because it is binary decision.

- Entropy

  Entropy is used to measure the orderliness.

  Boltzmann entropy: $S_B = -k_B \sum_i p_i \ln(p_i)$

  Shannon entropy: $H(X) = -\sum_i P(x_i) \log_2 P(x_i)$

- Position weight matrix

  Here are some DNA sequences. The first line is the position:

$$\begin{array}{ccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ G & A & G & G & T & A & A & A & C \\ T & C & C & G & T & A & A & G & T \\ C & A & G & G & T & T & G & G & A \\ A & C & A & G & T & C & A & G & T \\ T & A & G & G & T & C & A & T & T \\ T & A & G & G & T & A & C & T & G \\ A & T & G & G & T & A & A & C & T \\ C & A & G & G & T & A & T & A & C \\ T & G & T & G & T & G & A & G & T \\ A & A & G & G & T & A & A & G & T \end{array}$$

  In each position, we count the number of A, C, G, and T, respectively. Then we get the corresponding position frequency matrix (PFM):

$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}$$

  Then we simply do a normalization. Each number is divided by the total number of sequences. The corresponding position probability matrix (PPM) is:

$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}$$

  When we talk about TF binding sites or motifs, we always see some sequence logo. The sequence logo consists of stacks of symbols, one stack is for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position (information content). The height of symbols within the stack indicates the relative frequency of nucleic acid at that position. For example, in this sequence logo, the height of each position is calculated as $R_i = \log_2(4) - H_i$, in which $H_i = -\sum_k M_{k,i} \times \log_2 M_{k,i}$

Figure 2.1: sequence logo

Given the PPM (M) and a background model b, we can calculate the position weight matrix (PWM). In the PWM, $M_{k,i} = \log_2(M_{k,i}/b_k)$, in which $b = (b_1, b_2, b_3, b_4) = (p_A, p_C, p_G, p_T)$. For nucleotides, $b_k = 0.25$. In general, b_k does not have to be equal for each symbol. For example, if the organisms we studied with a high GC-content, the b_k for C and G will be higher than that for A and T. Besides, in practice, in order for convenience for calculation, we will give a pseudo count (such as 0.0001) to 0 to avoid the logarithm of 0.

- Motif matching score

  Given the PPM and a background model b, we can also calculate the motif matching score using the likelihood ratio score. For example, the score for GAGGTAAAC $= \log_2 \frac{p_G \times p_A \times p_G \times p_G \times p_T \times p_A \times p_A \times p_A \times p_C}{b_G \times b_A \times b_G \times b_G \times b_T \times b_A \times b_A \times b_A \times b_C}$

2.3.4 De novo sequence motif finding

The goal of de novo sequence motif finding is to look for common sequence patterns enriched in the input data compared to a background genome. There are two kinds approaches to do de novo sequence motif finding: deterministic approach and probabilistic approach.

2.3.4.1 Deterministic approach

The deterministic approach is regular expression enumeration. The basic idea for this approach is to check over-representation for every w-mer by comparing observed w occurrence in data and expected w occurrence in data. The over-represented w is potential TF binding motif. The advantages of this approach are that it is exhaustive, can guatantee to find global optimum, and can find multiple motifs. For disadvanteges, one is that it is not as flexible with base substitutions and long list of similar good motifs, and the other is that it's limited with motif width.

2.3.4.2 Probabilistic approach

Different from deterministic approach which is pattern driven approach, the probabilistic approach is data driven approach. Expectation-Maximization (EM) approach and Gibbs Sampling are two probabilistic approaches. Here we talk about the EM approach.

The objects of this approach are as follows: seq is sequence data to search for motif; θ_0 is non-motif probability (genome backgrouhnd) parameter; θ is motif probability matrix parameter; π is motif site location. The problem of this approach is to estimate $P(\theta, \pi | seq, \theta_0)$. The approach is to alternately estimate one of π and θ each time by fixing the other, in which the two steps are called E-step and M-step respectively. Here is an example for this approach:

- E-step: given $theta_0$, seq and θ to estimate π, in which θ_0 and seq are known while θ is given a initial value. In alternative steps, θ is calculated by M-step.

  Given an example, $\theta_0 : p_{0\mathrm{A}} = 0.3, p_{0\mathrm{C}} = 0.2, p_{0\mathrm{G}} = 0.2, p_{0\mathrm{T}} = 0.3$.

seq :

T	T	G	A	C	G	A	C	T	G	C	A	C	G	T	
T	T	G	A	C											LR_1
	T	G	A	C	G										LR_2
		G	A	C	G	A									LR_3
			A	C	G	A	C								LR_4
				C	G	A	C	T							LR_5
					G	A	C	T	G						LR_6
						A	C	T	G	C					LR_7
							C	T	G	C	A				LR_8
								.	.	.					

θ:

pos	A	C	G	T
1	0.7	0.1	0.01	0.2
2	0.01	0.01	0.8	0.1
3	0.32	0.02	0.3	0.18
4	0.03	0.42	0.1	0.47
5	0.2	0.5	0.1	0.2

Then, for LR_1,

$$\begin{aligned} P(\text{TTGAC}|\theta_0) &= p_{0\text{T}} \times p_{0\text{T}} \times p_{0\text{G}} \times p_{0\text{A}} \times p_{0\text{C}} \\ &= 0.3 \times 0.3 \times 0.2 \times 0.3 \times 0.2 \\ &= 1.08 \times 10^{-3} \end{aligned}$$

$$\begin{aligned} P(\text{TTGAC}|\theta) &= P(\text{T in pos1}) \times P(\text{T in pos2}) \times P(\text{G in pos3}) \times P(\text{A in pos4}) \times P(\text{C in pos5}) \\ &= 0.2 \times 0.1 \times 0.3 \times 0.03 \times 0.5 \\ &= 9 \times 10^{-5} \end{aligned}$$

Therefore, the likelihood ratio of the first motif π_1 is $LR_1 = \frac{P(\text{TTGAC}|\theta)}{P(\text{TTGAC}|\theta_0)} = \frac{9\times10^{-5}}{1.08\times10^{-3}}$. Then we can calculate LR_2, LR_3, LR_4, etc.

- M-step: given $theta_0$, seq, and π to estimate θ, in which θ_0 and seq are known while π with its likelihood ratio LR is calculated by E-step.

  Given an example, seq = TTGACGACTGCACGT, π and its likelihood ratio LR are:

π	LR
TTGAC	0.8
TGACG	0.2
GACGA	0.6
ACGAC	0.5
CGACT	0.3
GACTG	0.7
ACTGC	0.4
CTGCA	0.1
TGCAC	0.9
...	...

Then we can update θ by estimate the probability of A, C, G, T in any of the 5 positions (5 is the length of the motif):

$$P(\text{T in pos1}) = \frac{0.8 + 0.2 + 0.9 + ...}{0.8 + 0.2 + 0.6 + 0.5 + 0.3 + 0.7 + 0.4 + 0.1 + 0.9 + ...}$$

$$P(\text{T in pos2}) = \frac{0.8 + 0.1 + ...}{0.8 + 0.2 + 0.6 + 0.5 + 0.3 + 0.7 + 0.4 + 0.1 + 0.9 + ...}$$

$$P(\text{G in pos2}) = \frac{0.2 + 0.3 + 0.9 + ...}{0.8 + 0.2 + 0.6 + 0.5 + 0.3 + 0.7 + 0.4 + 0.1 + 0.9 + ...}$$

$$P(\text{C in pos5}) = \frac{0.8 + 0.5 + 0.4 + 0.9 + ...}{0.8 + 0.2 + 0.6 + 0.5 + 0.3 + 0.7 + 0.4 + 0.1 + 0.9 + ...}$$

... ...

After we get the updated θ from the M-step, we can re calculate the E-step. Iterate the E-step and M-step until θ does not improve. Then we can find the most frequent k-mers by calculate the likelihood ratio of each π.