Instructor: Sael Lee

CS549 Spring – Computational Biology

LECTURE 2: INFORMATION CONTENT IN BIOLOGY & DNA BINDING

Resources from:

1) Lecture Notes of Natasha Devroye devroye@ece.uic.edu http://www.ece.uic.edu/~devroye
2) F. Fabris "Shannon Information Theory and Molecular Biology" *JIM*, vol.12, n.1, february 2009, pp. 41-87.
3) T Cover & J Thomas "Elements of Information Theory 2nd ed." 2006

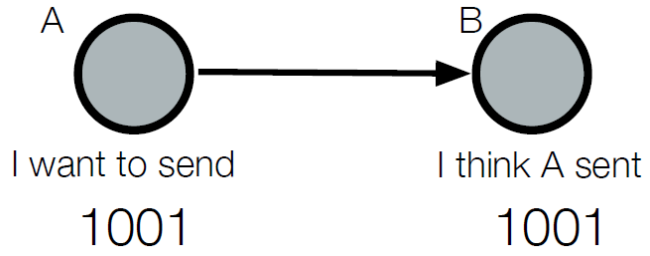
THE MATHEMATICS THEORY OF COMMUNICATION

Claude E. Shannon

Reprinted with corrections from *The Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656, July, October, 1948.

"The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point."

C.E. Shannon, 1948



A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

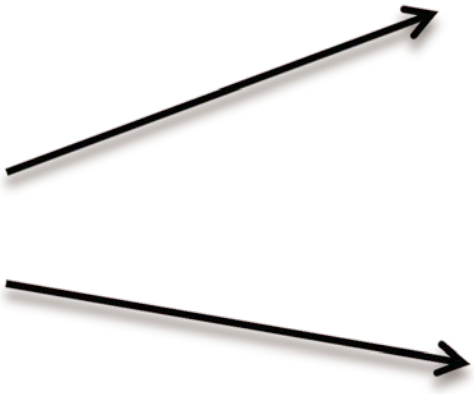
The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

Introduced a new field: Information Theory

SHANNON'S QUESTIONS

- *What is **information**?*
- *What is **communication**?*
- *How **fast** can we communicate?*
- *How much can we **compress** information?*

SHANNON'S FINDINGS



- Source Coding Problem:
  - Source = random variables
  - Ultimate **data compression** limit is the source's entropy H
- Channel Coding Problem:
  - Channel = conditional distributions
  - **Ultimate transmission rate** is the channel capacity C
- Relationship between input and output
  - Mutual Information
- Reliable communication possible ↔ H<C

GENERAL COMMUNICATION SYSTEM

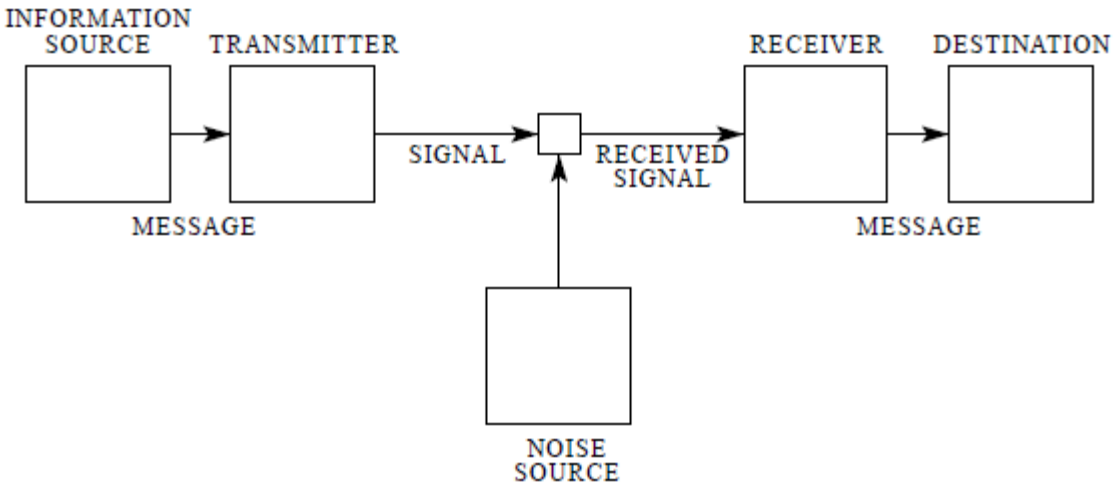
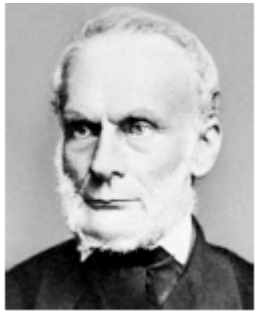


Fig. 1—Schematic diagram of a general communication system.

- ***Information source***: "produces a message or sequence of messages to be communicated to the receiving terminal"
- ***Transmitter:*** *"*operates on the message in some way to produce a signal suitable for transmission over the channel"
- ***Channel :*** *"*the medium used to transmit the signal from transmitter to receiver"
- ***Receiver:*** *"performs the inverse operation of that done by the transmitter reconstructing the message from the signal"*
- ***Destination:*** *"*person (or thing) for whom the message is intended"

ENTROPY: PHYSICS ORIGIN

* **Entropy**: Measure of disorder in a thermodynamic system.



Clausius: thermodynamics

1. The energy of the universe is constant.
2. The entropy of the universe tends to a maximum.

$$dS = \bar{d}Q/T$$

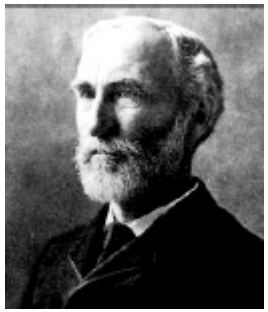
Where S is entropy, Q is the heat or internal energy and T the temperature.

Boltzman: statistical mechanics

quantifies entropy of an equilibrium thermodynamic system

$$S = K \log W$$

Where S is entropy, K is Boltzman constant, and W is number of microstates in the system



Gibbs: statistical mechanics

general entropy expression for a thermodynamic system

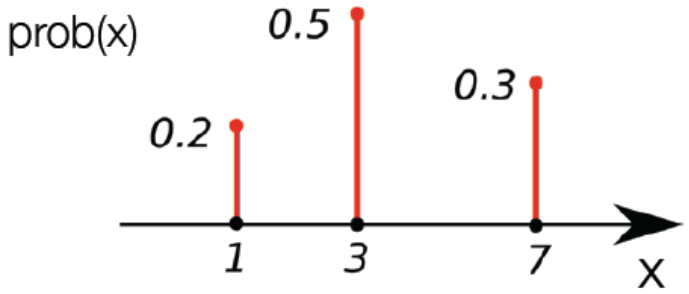
$$S = -\sum_j p_j \log p_j$$

where pj is the probability that the system is at microstate j.

* If pj = 1/W, then Gibbs' definition agrees with Boltzman's.

SHANNON'S ENTROPY

- Entropy is the measure of **average uncertainty** in the random variable
- Entropy is the **average number of bits** needed to describe the random variable
- Entropy is a lower bound on the **average length of the shortest description** of the random variable
- Entropy of a deterministic value is 0



What is the **entropy** of a random variable X with distribution p(x)?

$$H(X) = -\sum_{x} p(x) log_2(p(x))$$

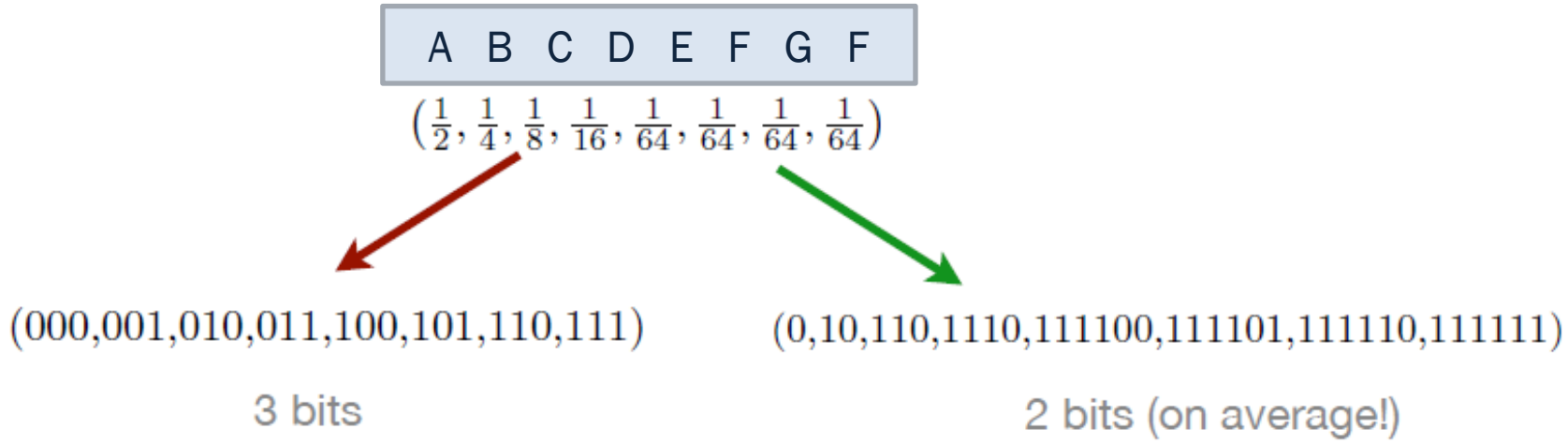
Entropy measured in bits

ENTROPY OF A NON-UNIFORM DISTRIBUTION

- Suppose X represents the outcome of a horse race with 8 horses, which win with probabilities $\left(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{16},\frac{1}{64},\frac{1}{64},\frac{1}{64},\frac{1}{64}\right)$

$$H(X) = -\frac{1}{2}log_2\left(\frac{1}{2}\right) - \frac{1}{4}log_2\left(\frac{1}{4}\right) - \frac{1}{8}log_2\left(\frac{1}{8}\right) - \frac{1}{16}log_2\left(\frac{1}{16}\right) - 4\frac{1}{64}log_2\left(\frac{1}{64}\right)$$
$$= \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{4}{16} + 4\frac{6}{64} = 2(\text{bits})$$

- 8 outcomes, 3 bits? But on average can represent with 2 bits.



MUTUAL INFORMATION BETWEEN 2 RANDOM VARIABLES:

- Mutual Information I(X;Y) is the **reduction** in the uncertainty about X due to knowledge of Y
- if X, Y are independent I(X;Y) = 0
- if X=Y then I(X;Y) = H(X)
- I(X;Y) is non-negative

$$I(X;Y) = -\sum_x p(x,y) log_2 \left(\frac{p(x,y)}{p(x)p(y)}\right)$$

DATA COMPRESSION

- Given a source X with distribution p(x), what is the fundamental limit of compression of this source's information?

$$H(X) = -\sum_{x} p(x) log_2(p(x))$$

- Can we construct good codes to achieve this limit?

CHANNEL CAPACITY

- Information channel capacity

$$C = max_{p(x)} I(X;Y)$$
$$= max_{p(x)} \left(-\sum_x p(x,y) log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right) \right)$$

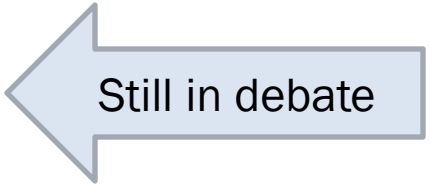
- Operational channel capacity
  - Highest rate (bits/channel use) that can communicate at reliably

- Channel coding theorem says

information capacity = operational capacity

BIOLOGICAL INFORMATION

- Double nature of "information"
  - Semantic (level of meaning): the functional story
    - What we dream of
  - Syntactic (level of physical data): sequence, structure, etc.
    - Shannon's perspective
- Source of information
  - DNA
  - Structure
  - Environment
- What is the information content of DNA?



THE DNA-TO-PROTEIN BIO-MOLECULAR CHANNEL

- Central Dogma of Molecular Biology states there is a flow of "biologic information" from DNA towards proteins:
- -> that the DNA carries information that, after transcription and translation, drives the synthesis of the proteins.

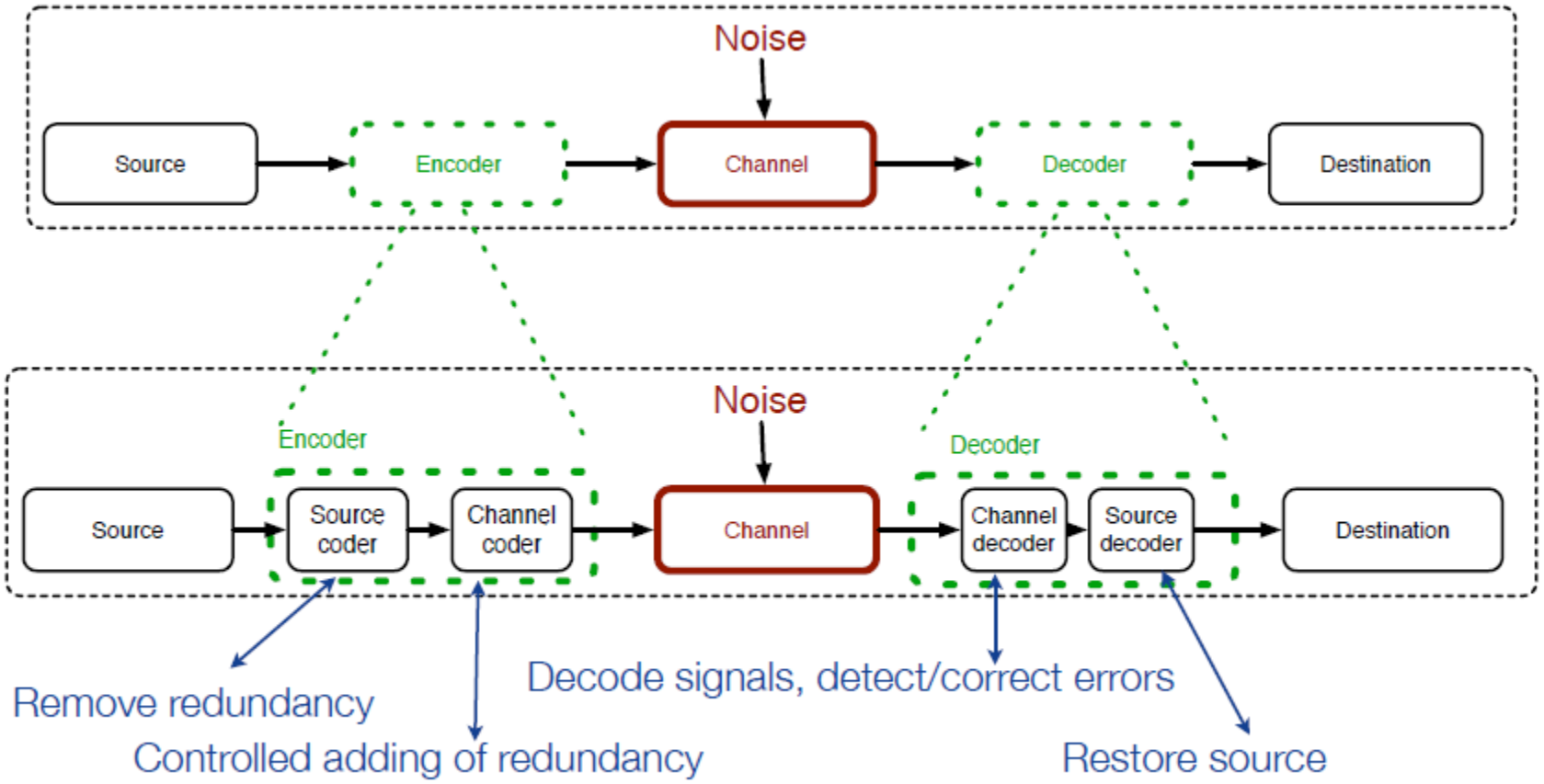
APPEALING METAPHOR

the flow of information that starts from DNA and reaches the proteins, in the biological communication system outlined by the Central Dogma, is analogous to the flow of information that starts from the sender and reaches the receiver (at the other side of the channel) in the communication system.

- DNA: interpreted as a sequence based on a 4-letters alphabet,
  - a sequence of nucleotides - Adenine, Thymine, Cytosine and Guanine (A, T,C,G),
- Protein: interpreted as a 20-letters alphabet sequence.
  - a sequence based on 20 amino acids (Metionine, Serine, Threonine etc.),

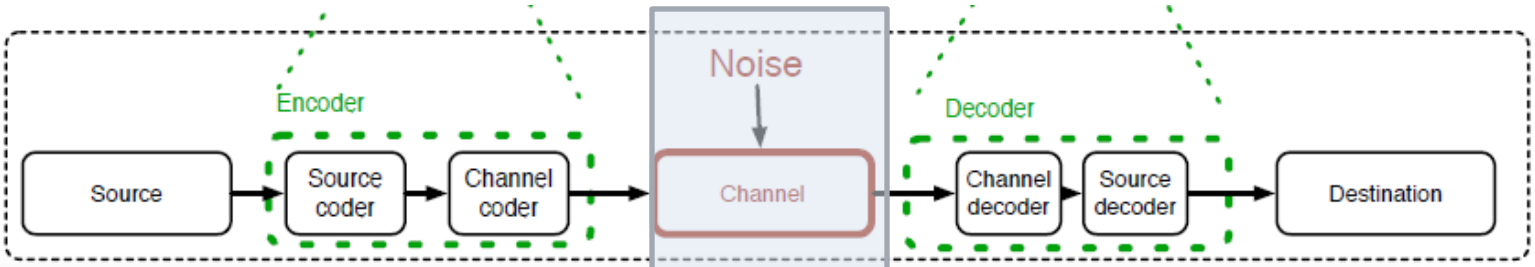
This approach seems to offer the opportunity of using Information Theory as a tool to build a model of biological information transmission and correction.

GENERIC COMMUNICATION BLOCK DIAGRAM

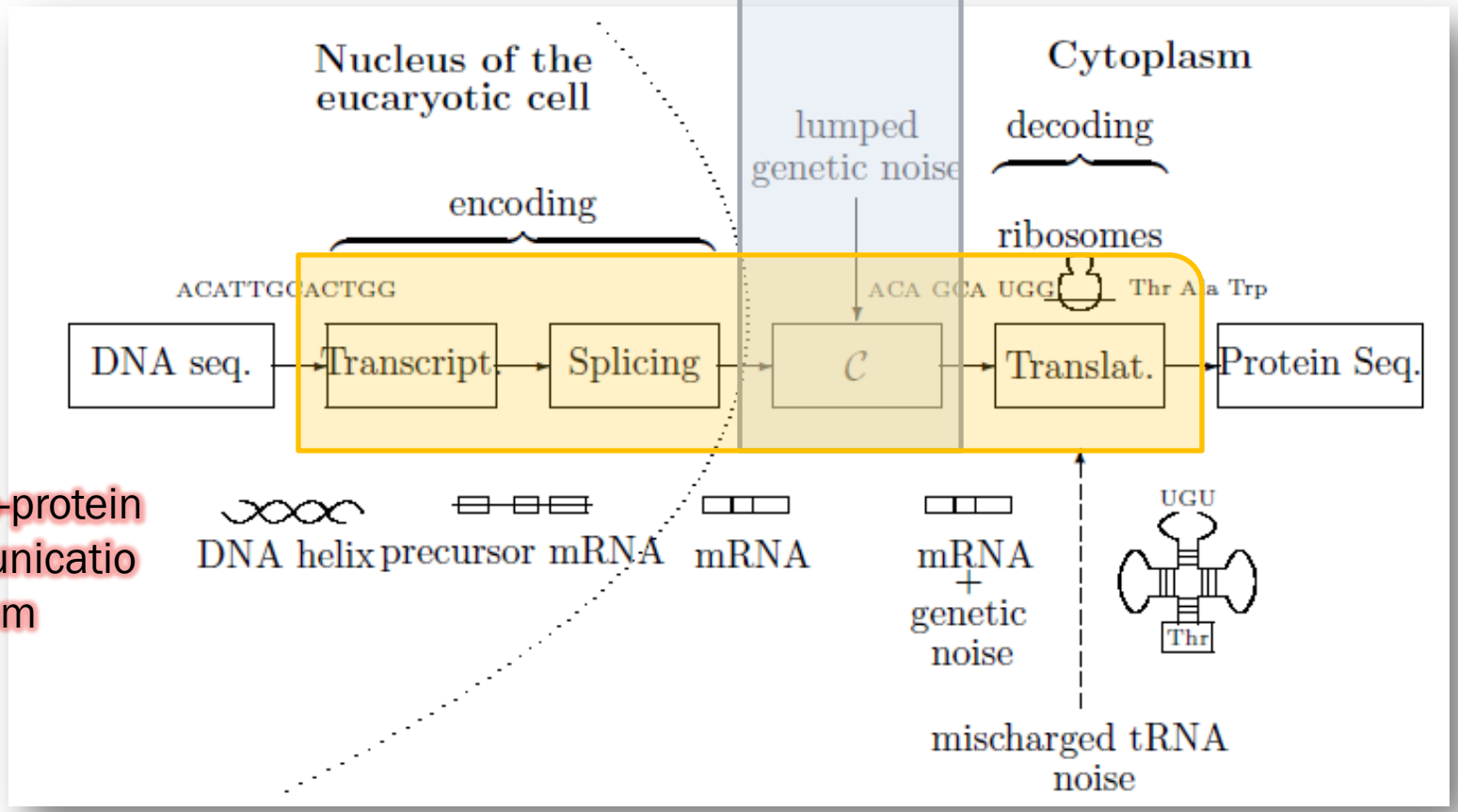


THE DNA-TO-PROTEIN BIO-MOLECULAR CHANNEL

Shannon unidirectional communication system



DNA-to-protein communicatio n system



DNA ERROR CORRECTION

biological mechanism of synthesizing a protein does have some mechanisms for DNA error correction

- Direct Chemical Reversal
  - Spontaneous addition of a methyl group (CH3-) to a C, that transform a C in a T after deamination
- Repaired by Enzymes (glycosylases)
  - remove the mismatched T restoring the correct C
- Excision Repair
  - damaged base or bases are removed and then replaced with the correct ones in a localized burst of DNA synthesis
- Etc.

Repair done in place

DNA backbone needs be broken

EFFECT OF THE GENETIC CODE REDUNDANCY

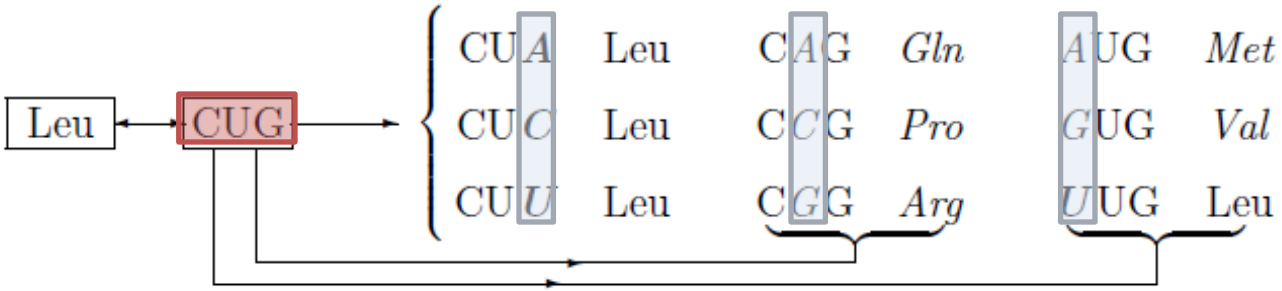


Table of the Genetic Code. The amino acids are characterized by the 1-letter and the 3-letters

Amino acids coding table				
Basic	Lysine	Lys	K	AAA AAG
	Arginine	Arg	R	CGU CGC CGA CGG AGA AGG
	Histidine	His	H	CAU CAC
Non Polar	Glycine	Gly	G	GGA GGC GGG GGU
	Alanine	Ala	A	GCA GCC GCG GCU
	Valine	Val	V	GUA GUC GUG GUU
	Leucine	Leu	L	CUU CUC CUA CUG UUA UUG
	Isoleucine	Ile	I	AUU AUC AUA
	Methionine	Met	M	AUG
	Phenylalanine	Phe	F	UUC UUU
	Tryptophan	Trp	W	UGG
	Proline	Pro	P	CCU CCC CCA CCG

Amino acid				DNA codons
Acidic	Aspartic Acid	Asp	D	GAU GAC
	Glutamic Acid	Glu	E	GAA GAG
Polar	Serine	Ser	S	UCU UCC UCA UCG AGU AGC
	Threonine	Thr	T	ACU ACC ACA ACG
	Cysteine	Cys	C	UGU UGC
	Asparagine	Asn	N	AAU AAC
	Glutamine	Gln	Q	CAA CAG
	Tyrosine	Tyr	Y	UAU UAC
	Terminator	Ter	end	UAA UAG UGA

BIO-MOLECULAR CHANNEL

	Leu	Ser	Arg	Ala	Val	Pro	Thr	Gly	Ile	STOP	Tyr	His	Gln	Asn	Lys	Asp	Glu	Cys	Phe	Trp	Met
UUA	1-7α	α			α				α	2α									2α		
UUG	1-7α	α			α					α									2α	α	α
CUU	1-6α		α		α	α			α			α							α		
CUC	1-6α		α		α	α			α			α							α		
CUA	1-5α		α		α	α			α				α								
CUG	1-5α		α		α	α							α								α
UCU		1-6α		α		α	α				α							α	α		
UCC		1-6α		α		α	α				α							α	α		
UCA	α	1-6α		α		α	α			2α											
UCG	α	1-6α		α		α	α			α										α	
AGU		1-8α	3α				α	α	α					α				α			
AGC		1-8α	3α				α	α	α					α				α			
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
UUU	3α	α			α				α										α	1-8α	
UUC	3α	α			α				α		α								α	1-8α	
UGG	α	α	2α					α		2α	α								2α		1-9α
AUG	2α		α		α		α		3α						α						1-9α

Transition probability matrix of the Yockey bio-molecular channel (Yockey, 1974, 1992). = (x/y) is the probability of passing from nucleotide y to nucleotide x

APPEALING BUT HAS LIMITS

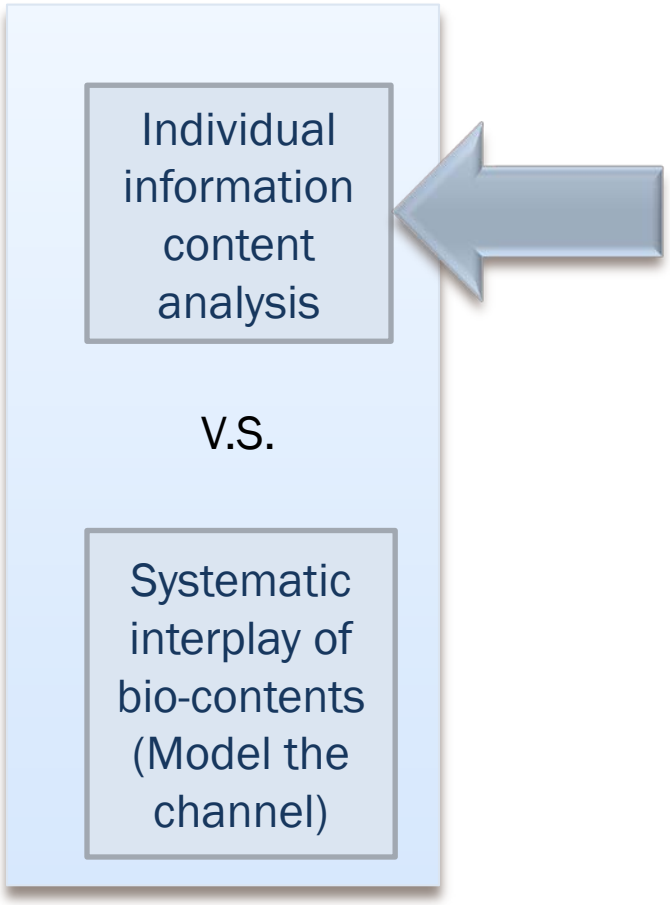
- Biology is much complex compared to general communication system.
  - Systematically complex: Feedback loops, granularity, multiple players
  - Model incomplete: Many biological relations yet to be learned

DNA / RNA / PROTEINS; GENE

Single "word" in genome

"A gene is a molecular unit of heredity of a living organism. It is widely accepted by the scientific community as a name given to some stretches of DNA and RNA that code for a polypeptide (protein) or for an RNA chain that has a function in the organism."
[http://en.wikipedia.org/wiki/Gene]

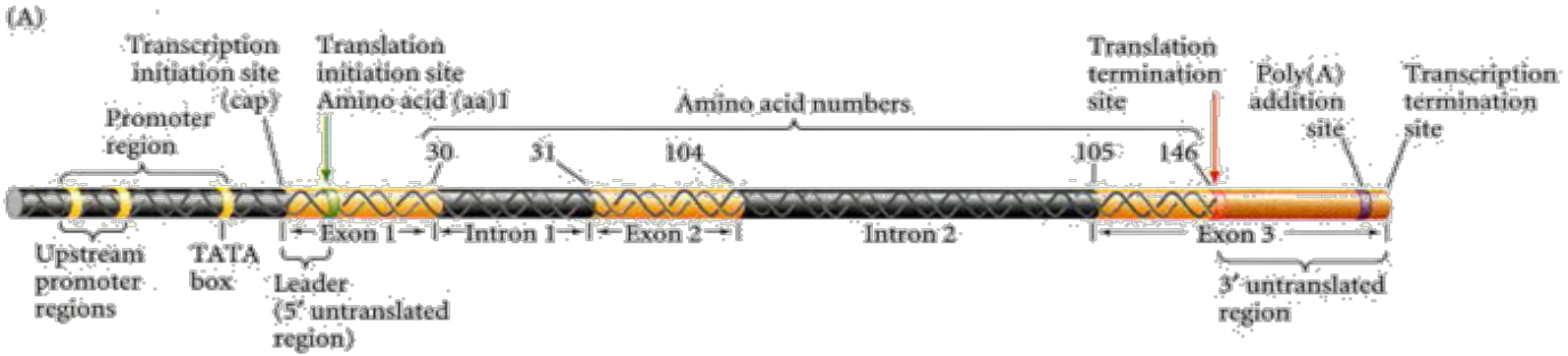
* The concept of genes preceded the knowledge of DNA. So, there is some controversies in linking genes to DNA.



DNA BINDING

- Many biological functions are realized by protein binding to a area in DNA
  - Transcription
  - Translation event in protein synthesis
  - DNA repair
  - Gene silencing
  - Gene expression enhancing
  - Splicing
- One way to find the DNA binding event is by finding the sequential pattern in DNA binding events

ANATOMY OF THE (EUKARYOTIC) GENE



Promoters Exons Introns

- **Promoters** are the sites where RNA polymerase binds to the DNA to initiate transcription.
- **Enhancer** is a DNA sequence that can activate the utilization of a promoter, controlling the efficiency and rate of transcription from that particular promoter. Located geometrically close to the promoter and gene but may not be close in sequence.
- **Exons**—are intervening sequences
- **Introns**—that have nothing whatsoever to do with the amino acid sequence of the protein.

* Father Reading: Differential Gene Transcription http://www.ncbi.nlm.nih.gov/books/NBK10023/

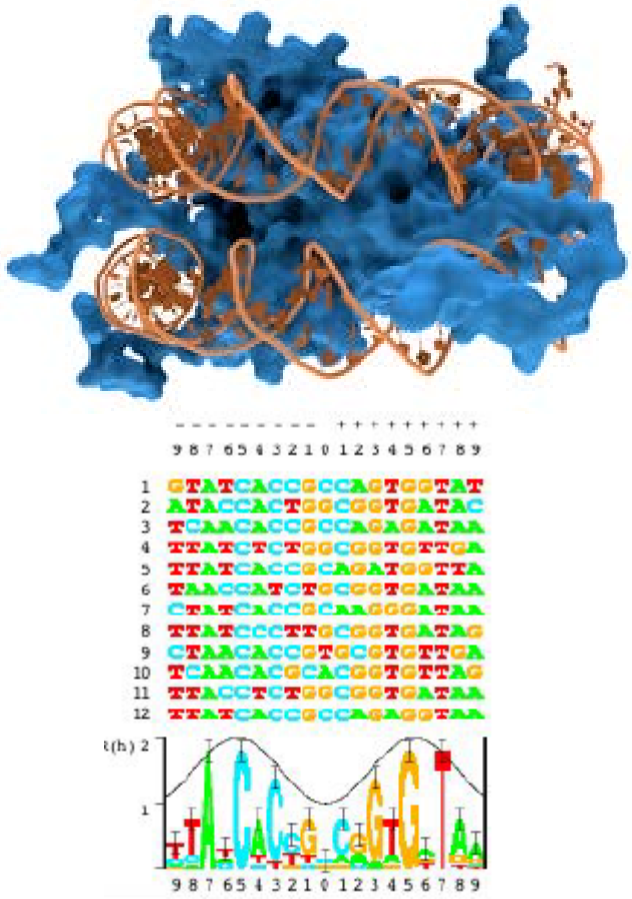
24

DNA-BINDING PROTEIN

http://en.wikipedia.org/wiki/DNA-binding_protein

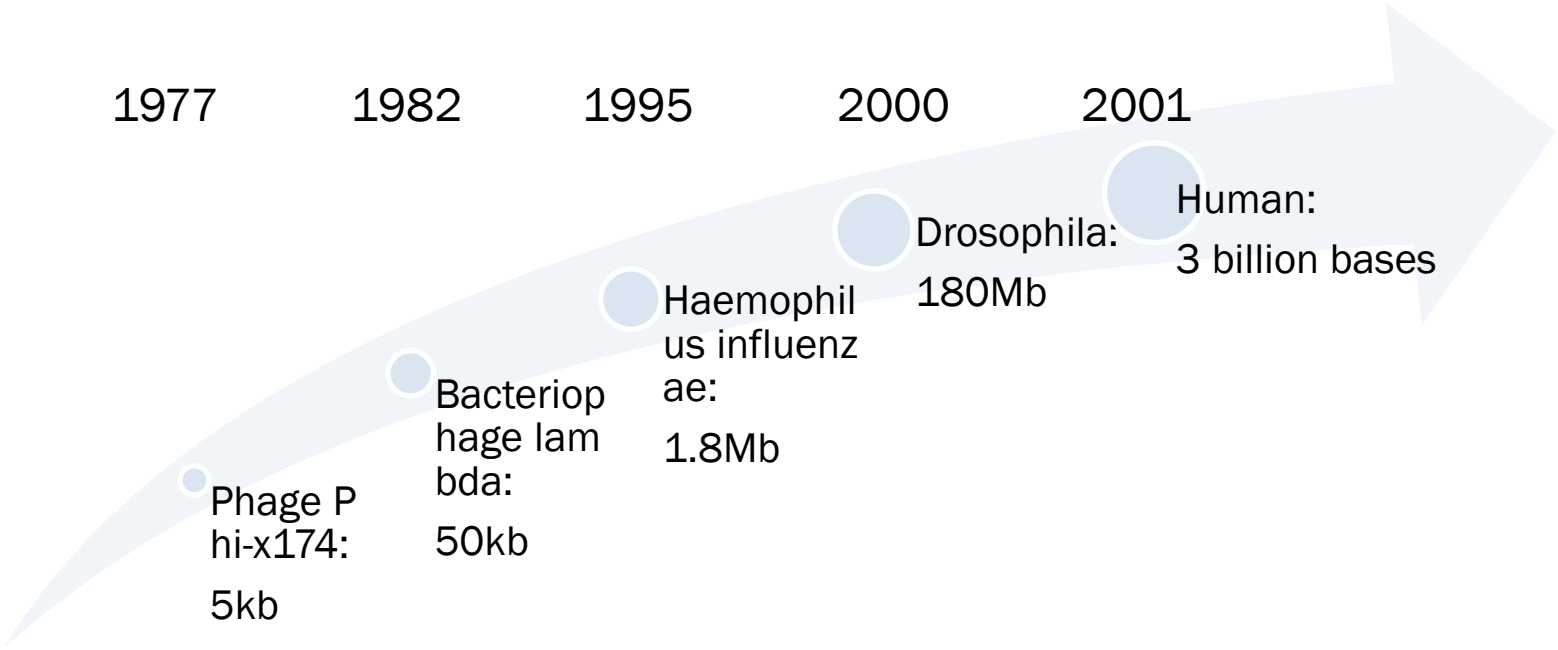
Proteins that are composed of DNA-binding domains and thus have a specific or general affinity for either single or double stranded DNA.

- Types of Binding
  - **Sequence-specific DNA-binding**
    - generally interact with the major groove of DNA
  - Non-specific DNA-protein interactions
  - DNA-binding proteins that specifically bind single-stranded DNA



PROGRESS IN GENOME SEQUENCING

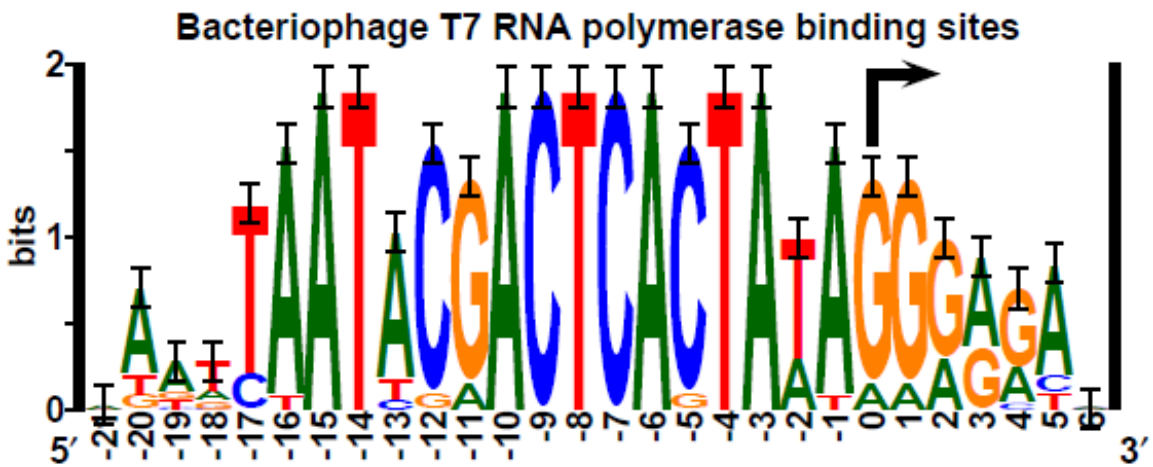
The **scale** of sequencing projects has been achieved, largely through automation:



Interest now revolves around **faster/cheaper/accurate** rather than larger. <- We get more data

SEQUENCE LOGO

- Sequence logo is a graphical representation of the sequence **conservation** of nucleotides (in a strand of DNA/RNA) or amino acids (in protein sequences)



Schneider & Stephens Nucl. Acids Res. 18: 6097-6100 1990

1	ttattaatacaactcactataaggagag
2	aaatcaatacgactcactatagagggac
3	cggttaatacgactcactataggagaac
4	gaagtaatacgactcagtatagggacaa
5	taattaattgaactcactaaagggagac
6	cgcttaatacgactcactaaaggagaca

6 of 17 sites

Sequence Alignment

READINGS FOR NEXT CLASS

- Chapter 2 of Elements of Information Theory.