A Zero-Knowledge Based Introduction to Biology

Jim Notwell
09 January 2013
Q: What is your genome?

A:
Q: What is your genome?

A: The sum of your hereditary information.
From DNA to Organism

You are composed of ~ 10 trillion cells
From DNA to Organism Cell
Proteins do most of the work in biology
Central Dogma of Biology

1. DNA Polymerase
   DNA → DNA

2. RNA Polymerase
   DNA → RNA

3. Ribosome
   RNA → Protein
DNA: “Blueprints” for a cell

- Genetic information encoded in long strings

- Deoxyribonucleic acid comes in four flavors: adenine, thymine, guanine, and cytosine
Phosphate-deoxyribose Backbone

to previous nucleotide

3'

5'

to base

to next nucleotide
Nucleobase Complementary Pairing

- **Purines**: Adenine (A) and Guanine (G)
- **Pyrimidines**: Thymine (T) and Cytosine (C)

[Diagram showing the structures of the nucleobases and their complementary pairs]
DNA Double Helix

- Hydrogen
- Oxygen
- Nitrogen
- Carbon
- Phosphorus

Pyrimidines
Purines
DNA Packaging
Q: What is your genome?

A: The sum of your hereditary information.
Q: What is your genome?

A: The sum of your hereditary information. Humans bundle two copies of the genome into 46 chromosomes in every cell.
Central Dogma of Biology

- **DNA Polymerase**
  - Replication (DNA -> DNA)
- **RNA Polymerase**
  - Transcription (DNA -> RNA)
- **Ribosome**
  - Translation (RNA -> Protein)
DNA vs RNA

to previous nucleotide

O\(\rightarrow\)PO\(\rightarrow\)C\(\rightarrow\)O\(\rightarrow\)H\(\rightarrow\)H\(\rightarrow\)H\(\rightarrow\)H\(\rightarrow\)H\(\rightarrow\)C\(\rightarrow\)C

5’

to base

to next nucleotide

3’

to previous ribonucleotide

O\(\rightarrow\)PO\(\rightarrow\)C\(\rightarrow\)O\(\rightarrow\)H\(\rightarrow\)H\(\rightarrow\)H\(\rightarrow\)H\(\rightarrow\)C\(\rightarrow\)C

5’

to base

to next ribonucleotide

3’
RNA Nucleobases

purines

Adenine (A)  Guanine (G)

Uracil (U)    Cytosine (C)

pyrimidines
Gene Transcription

3' GATTACA... 5'
3' CTAATGT... 5'
Gene Transcription
Gene Transcription

Strands are separated (DNA helicase)
Gene Transcription

An RNA copy of the 5’→3’ sequence is created from the 3’→5’ template
Gene Structure

5' UTILITY

promoter

introns

exons

3' UTR

coding

non-coding
Central Dogma of Biology

DNA Polymerase

DNA

RNA Polymerase

RNA

Ribosome

Protein

replication (DNA → DNA)

transcription (DNA → RNA)

translation (RNA → Protein)
From RNA to Protein

• Proteins are long strings of amino acids joined by peptide bonds

• Translation from RNA sequence to amino acid sequence performed by ribosomes

• 20 amino acids → 3 RNA letters required to specify a single amino acid
There are 20 standard amino acids
Proteins

from 5’

N-terminus

(start)

H

OH

R

C-terminus

(end)

3’ mRNA

to previous aa

H

O

C

C

H

to next aa

N-terminus (start)

C-terminus (end)

Proteins
The ribosome (a complex of protein and RNA) synthesizes a protein by reading the mRNA in triplets (codons). Each codon is translated to an amino acid.
<table>
<thead>
<tr>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUU</td>
<td>Phenylalanine (Phe)</td>
<td>UCU</td>
<td>Serine (Ser)</td>
</tr>
<tr>
<td>UUC</td>
<td>Phe</td>
<td>UCC</td>
<td>Ser</td>
</tr>
<tr>
<td>UUA</td>
<td>Leucine (Leu)</td>
<td>UCA</td>
<td>Ser</td>
</tr>
<tr>
<td>UUG</td>
<td>Leu</td>
<td>UCG</td>
<td>Ser</td>
</tr>
<tr>
<td>CUU</td>
<td>Leucine (Leu)</td>
<td>CCU</td>
<td>Proline (Pro)</td>
</tr>
<tr>
<td>CUC</td>
<td>Leu</td>
<td>CCC</td>
<td>Pro</td>
</tr>
<tr>
<td>CUA</td>
<td>Leu</td>
<td>CCA</td>
<td>Pro</td>
</tr>
<tr>
<td>CUG</td>
<td>Leu</td>
<td>CCG</td>
<td>Pro</td>
</tr>
<tr>
<td>AAU</td>
<td>Isoleucine (Ile)</td>
<td>ACU</td>
<td>Threonine (Thr)</td>
</tr>
<tr>
<td>AUC</td>
<td>Ile</td>
<td>ACC</td>
<td>Thr</td>
</tr>
<tr>
<td>AUA</td>
<td>Ile</td>
<td>ACA</td>
<td>Thr</td>
</tr>
<tr>
<td>AUG</td>
<td>Methionine (Met) or START</td>
<td>ACG</td>
<td>Thr</td>
</tr>
<tr>
<td>GUU</td>
<td>Valine (Val)</td>
<td>GCU</td>
<td>Alanine (Ala)</td>
</tr>
<tr>
<td>GUC</td>
<td>Val</td>
<td>GCC</td>
<td>Ala</td>
</tr>
<tr>
<td>GUA</td>
<td>Val</td>
<td>GCA</td>
<td>Ala</td>
</tr>
<tr>
<td>GUG</td>
<td>Val</td>
<td>GCG</td>
<td>Ala</td>
</tr>
</tbody>
</table>
Translation

5' . . . A U U A U G G C C U G G A C U U U G A . . . 3'

[Diagram showing the nucleotide sequence with base pairing indicated]
Translation

5' ... AUUAUGGCCUGGACCUUGA ... 3'

UTR  Met  Ala  Trp  Thr  Stop Codon
Central Dogma of Biology

1. **DNA Polymerase** (replication: DNA → DNA)
2. **RNA Polymerase** (transcription: DNA → RNA)
3. **Ribosome** (translation: RNA → Protein)
Protein coding 1%

Other Stuff 99%
Protein coding 1%
Non-coding exons 2%
Introns/promoters/polyA sites 37%
Intergenic transcribed RNA 19%
Regulatory elements 9%

Non-coding RNAs

• RNAs transcribed from DNA but not translated into protein

• Structural ncRNAs: Conserved secondary structure

• Involved in gene regulation
microRNA

1 A protein called exportin-5 transports a hairpin primary microRNA (pri-miRNA) out of the nucleus.

2 An enzyme called Dicer (not shown) trims the pri-miRNA and removes the hairpin loop, leaving a double stranded microRNA duplex molecule.

3 In plant cells, the microRNA is usually perfectly complementary to its target mRNA molecule. The microRNA will bond with it and cause the mRNA to break down.

2½ Meanwhile, one of the strands joins a group of proteins, forming an microRNA-protein complex. The other strand, known as a passenger strand, is usually discarded. How this all happens is still not very well understood.

4 In animal cells, the microRNA nucleotides typically don't pair up with the mRNA nucleotides as well. Their base pairing often follows a pattern though.

5 The microRNA-protein complex's presence blocks translation as well as speeding up deadenylation (breakdown of the Poly-A tail), which causes the mRNA to be degraded sooner and translated less.
Protein coding: 1%
Non-coding exons: 2%
Introns/promoters/polyA sites: 37%
Intergenic transcribed RNA: 19%
Regulatory elements: 9%

Subsets of the DNA sequence determine the identity and function of different cells
Gene Expression Regulation

• When should each gene be expressed?

• Why? Every cell has **same DNA** but each cell expresses **different proteins**.

• Signal transduction: One signal converted to another: cascade has “master regulators” turning on many proteins, which in turn each turn on many proteins
Central Dogma of Biology

- DNA Polymerase
  - Replication (DNA -> DNA)

- RNA Polymerase
  - Transcription (DNA -> RNA)

- Ribosome
  - Translation (RNA -> Protein)

- Protein
Transcription Regulation

- Transcription factors link to binding sites
- Complex of transcription factors forms
- Complex assists or inhibits formation of the RNA polymerase machinery
Gene Transcription

5' - GATTACA... - 3'
3' - CTAATGT... - 5'
Transcription Factor Binding Sites

• Short, degenerate DNA sequences recognized by particular transcription factors

• For complex organisms, cooperative binding of multiple transcription factors required to initiate transcription
Transcription Regulation

TF A Binding Site

Gene B

Transcription Factor A
understand how different permutations of the same regulatory elements alter gene expression. An understanding of how the combinatorial organization of a promoter encodes regulatory information first requires an overview of the proteins that constitute the transcriptional machinery.

### The Eukaryotic Transcriptional Machinery

Factors involved in the accurate transcription of eukaryotic protein-coding genes by RNA polymerase II can be classified into three groups: general (or basic) transcription factors (GTFs), promoter-specific activator proteins (activators), and coactivators (Figure 2).

GTFs are necessary and can be sufficient for accurate transcription initiation in vitro (reviewed in 141). Such factors include RNA polymerase II itself and a variety of auxiliary components, including TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. In addition to these “classic” GTFs, it is apparent that in vivo transcription also requires Mediator, a highly conserved, large multisubunit complex that was originally identified in yeast (reviewed in 38, 119).

GTFs assemble on the core promoter in an ordered fashion to form a transcription preinitiation complex (PIC), which directs RNA polymerase II to the transcription start site (TSS). The first step in PIC assembly is binding of TFIID, a multisubunit complex consisting of TATA-box-binding protein (TBP) and a set of tightly bound TBP-associated factors (TAFs). Transcription then proceeds through a series of steps, including promoter melting, clearance, and escape, before a fully functional RNA polymerase II elongation complex is formed. The current model of transcription regulation views this as a cycle, in which complete PIC assembly is stimulated only once. After RNA polymerase II escapes from the promoter, a scaffold structure, composed of TFIID, TFIIE, TFIIH, and Mediator, remains on the core promoter.

[Figure 1: Schematic of a typical gene regulatory region. The promoter, which is composed of a core promoter and proximal promoter elements, typically spans less than 1 kb pairs. Distal (upstream) regulatory elements, which can include enhancers, silencers, insulators, and locus control regions, can be located up to 1 Mb pairs from the promoter. These distal elements may contact the core promoter or proximal promoter through a mechanism that involves looping out the intervening DNA.]

**GTF**: a factor that assembles on the core promoter to form a preinitiation complex and is required for transcription of all (or almost all) genes.

**Coactivators**: adaptor proteins that typically lack intrinsic sequence-specific DNA binding but provide a link between activators and the general transcriptional machinery.

**PIC**: preinitiation complex.

**TSS**: transcription start site.
The eukaryotic transcriptional machinery. Factors involved in eukaryotic transcription by RNA polymerase II can be classified into three groups: general transcription factors (GTFs), activators, and coactivators. GTFs, which include RNA polymerase II itself and TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, assemble on the core promoter in an ordered fashion to form a preinitiation complex (PIC), which directs RNA polymerase II to the transcription start site (TSS). Transcriptional activity is greatly stimulated by activators, which bind to upstream regulatory elements and work, at least in part, by stimulating PIC formation through a mechanism thought to involve direct interactions with one or more components of the transcriptional machinery. Activators consist of a DNA-binding domain (DBD) and a separable activation domain (AD) that is required for the activator to stimulate transcription. The direct targets of activators are largely unknown.

The DNA-binding sites for activators (also called transcription factor-binding sites (TFBSs)) are generally small, in the range of 6–12 bp, although binding specificity is usually dictated by no more than 4–6 positions within the site. The TFBSs for a specific activator are typically degenerate, and are therefore described by a consensus sequence in which certain positions are relatively constrained and others are more variable. Many activators form heterodimers and/or homodimers, and thus their binding sites are generally composed of two half-sites. Notably, the precise subunit composition of an activator can also dictate its binding specificity and regulatory action.

Although an activator can bind to a wide variety of sequence variants that conform to the consensus, in certain instances the precise sequence of a TFBS can impact the regulatory output. For example, TFBS sequence variations can affect activator binding strength (reviewed in 30), which may be biologically important in situations such as in early development, in which activators are distributed in a concentration gradient (84, 144). TFBS sequence variations may also direct a preference for certain dimerization partners over others (37, 124, 142). Finally, the particular sequence of a TFBS can affect the structure of a bound activator in a way that alters its activity (69, 104, 108, 154, 163). The best-studied examples are nuclear hormone receptors, a large class of ligand-dependent activators. Various studies have shown that the relative orientation of the half-sites, as well as the spacing between them, play a major role in directing the regulatory action of the bound nuclear hormone receptor dimer (37).
Protein coding: 1%
Non-coding exons: 2%
Introns/promoters/polyA sites: 37%
Intergenic transcribed RNA: 19%
Regulatory elements: 9%

Q: What if the transcription/translation machinery makes mistakes?

Q: What is the effect in coding regions?
Evolution = Mutation + Selection
Structural Abnormalities

Normal

Duplication

Deletion

Inversion

Insertion

Reciprocal Translocation
Single Nucleotide Changes

**Normal**

DNA: A A A A T A C G T G C A
mRNA: U U U U A U G C A C G U
Protein: Phe Tyr Ala Arg

**Missense Mutation**

DNA: A A A A T A C C T G C A
mRNA: U U U U A U G G A C G U
Protein: Phe Tyr Gly Arg

**Silent Mutation**

DNA: A A G A T A C G T G C A
mRNA: U U C U A U G C A C G U
Protein: Phe Tyr Ala Arg

**Nonsense Mutation**

DNA: A A A A T T C G T G C A
mRNA: U U U U A A G C A C G U
Protein: Phe Tyr STOP
Single Nucleotide Changes

**Normal**

DNA: 
```
AAAAATACGTGCA
UUUUAUGCACGU
```

mRNA: 
```
AAAAATACGTGCA
UUUUAUGCACGU
```

Protein: 
```
Phe  Tyr  Ala  Arg
```

**Frameshift (Deletion)**

DNA: 
```
AAAAACCTGCA
UUUUUGACGU
```

mRNA: 
```
AAAAACCTGCA
UUUUUGACGU
```

Protein: 
```
Phe  Leu  His  Val
```

**Frameshift (Insertion)**

DNA: 
```
AAATATACGTC
UUUAUAUGCACG
```

mRNA: 
```
AAATATACGTC
UUUAUAUGCACG
```

Protein: 
```
Phe  Ile  Cys  Thr
```

T
Evolution = Mutation + Selection
Evolution = Mutation + Selection
Summary

Evolution = Mutation + Selection
• All hereditary information encoded in double-stranded DNA

• Each cell in an organism has same DNA

• DNA $\rightarrow$ RNA $\rightarrow$ protein

• Proteins have many diverse roles in cell

• Gene regulation diversifies protein products within different cells
Further Reading

• See website: cs173.stanford.edu