## DRAFT

# **Primer for Synthetic Biology**

(7/18/07) by Scott C. Mohr

"prim·er (prĭm'ər) n. A book that covers the basic elements of a subject." (The American Heritage College Dictionary)

#### Introduction

"Synthetic biology," in the modern sense, means using engineering principles to create functional systems based on the molecular machines and regulatory circuits of living organisms. However, it also includes going beyond them to develop radically new systems. At the very least, synthetic biology represents a merger of molecular biology, genetic engineering and computer science<sup>1</sup> – and given the scale of the components that it works with, it also qualifies as a form of nanotechnology. Students and others who wish to learn about or – better still – to *do* synthetic biology often approach this exciting field with only some of the background necessary to grasp its essentials. This threepart primer represents an effort to assist them in filling gaps in their knowledge. Part I, **"Molecular Biology for Novices,"** summarizes the key aspects of biochemistry and molecular biology that bear on synthetic biology. Part II, **"Engineering for Biologists,"** introduces some basic engineering principles and the most important tools used in synthetic biology. Part III, **"Ethics for Everyone,"** briefly outlines the key ethical issues facing synthetic biology today.

The document has been subdivided into numbered sections for convenient access to specific topics. The Outline below links to the corresponding sections. Within the sections key words and phrases are highlighted in boldface type; these should be helpful in locating more in-depth information in the References. A list of Acronyms and Abbreviations and a useful Glossary<sup>2</sup> are appended.

This primer has been written with a view to brevity and clarity. Thus, it is information-dense and should be read slowly (probably in short installments). Please send any and all criticisms or suggestions to the author (<u>mohr@bu.edu</u>).

<sup>&</sup>lt;sup>1</sup> One facetious quip defines synthetic biology as "genetic engineering on steroids." A recent critical review calls it "extreme genetic engineering."

<sup>&</sup>lt;sup>2</sup> Whenever appropriate, terms that appear boldface in the text are defined in the glossary.

# **A Primer for Synthetic Biologists – OUTLINE**

# PART I. Molecular Biology for Novices

- 1. <u>Cells<sup>3</sup></u>
- 2. <u>Small biomolecules and metabolism</u>
- 3. <u>Biological macromolecules</u>
- 4. <u>Protein functions</u>
- 5. <u>Amino acids</u>
- 6. <u>Peptide bonds</u>
- 7. <u>Protein primary (1°) structure</u>
- 8. Three-dimensional (3D) structure of proteins
- 9. Encoded information generates structure
- 10. Nucleic acid (RNA and DNA) primary structure
- 11. The double helix
- 12. <u>RNA the *ur*biomolecule</u>
- 13. Polysaccharides Structure and Information
- 14. <u>The central dogma</u>
- 15. Protein synthesis
- 16. <u>The genetic code</u>
- 17. <u>Regulation of small-molecule metabolism (the metabolome)</u>
- 18. Protein control of gene expression
- 19. <u>The *lac* operon</u>
- 20. <u>Regulation of gene expression by RNA</u>
- 21. Gene regulation by DNA packaging
- 22. Epigenetic control
- 23. <u>Summary</u>

<sup>&</sup>lt;sup>3</sup> Underlined items are completed.

# **PART II – Engineering for Biologists**

- 1. Engineering basics
- 2. Restriction enzymes
- 3. Plasmids
- 4. Recombinant DNA technology
- 5. Protein engineering
- 6. Genetic circuits
- 7. Metabolic engineering
- 8. Intercellular communication
- 9. Multicellular systems
- 10. Engineering biology
- 11. Abstraction hierarchy
- 12. Standardization

# **PART III – Ethics for Everyone**

- 1. The Meaning of Ethics
- 2. Origins of Ethics
- 3. Ethics and Law
- 4. Risk
- 5. Biosafety
- 6. Biosecurity
- 7. Who Decides?

# **POSTSCRIPT – The Future of Synthetic Biology**

# PART I Molecular Biology for Novices

"Biology is the highest form of applied chemistry." Richard E. Dickerson and Irving Geis (*Chemistry,Matter and the Universe,* W.A.Benjamin, Menlo Park, CA, 1976)

#### 1. Cells

Living organisms consist of microscopic compartments called **cells** that usually have largest dimensions in the range of 2-20 micrometers (about 1-10% of the diameter of a human hair).<sup>4</sup> Because cells are so small, ordinary biological samples typically contain enormous numbers of them. An adult human body, for example, contains approximately 100,000,000,000 (10<sup>14</sup> or one hundred trillion) cells and a liquid culture of *E. coli* bacteria near the end of logarithmic growth contains several billion cells per milliliter.

Almost all biochemical reactions occur *within* cells<sup>5</sup> and for the purposes of synthetic biology, most of the reactions we want to engineer are **intracellular**. However, two important types of process involve communication *between* cells and their external environment: (1) **uptake and excretion** of nutrients, waste products and some subcellular components<sup>6</sup> and (2) **detection of signals** that indicate the presence of nutrients, hormones, other cells, etc. Many synthetic biology projects involve engineering the components that carry out such interactions between cells and the outside world. Nevertheless, to manipulate these processes we must also work with the intracellular machinery.

Thus cells provide the framework or "**chassis**" for synthetic biology. At the present time most synthetic biology experiments involve simple **singlecelled organisms** like **the bacterium** *Escherichia coli* or **the yeast** *Saccharomyces cerevisiae*. Many researchers anticipate the development of drastically modified

<sup>&</sup>lt;sup>4</sup> Cells from eukaryotes ("higher" organisms) range from 10-100 μm, with animal cells normally smaller than plant cells. However, egg cells are notably larger (even visible to the naked eye), muscle cells can fuse into large multinuclear syncytia that extend for several centimeters, and – famously – some nerve cells in giraffes extend for the entire length of the neck (though their diameters are still in the micrometer range). There are also some extraordinarily large unicellular bacteria and protozoa. All these cases represent "outliers" – relatively rare exceptions to the general rule given above.

<sup>&</sup>lt;sup>5</sup> Exceptions are degradative processes catalyzed by secreted enzymes.

<sup>&</sup>lt;sup>6</sup> Including complex materials like DNA and viruses.

or even entirely **artificial cells** and the eventual expansion of synthetic biology to encompass **multicellular organisms**. The latter case involves **differentiation** into **multiple cell types** (about 200 kinds in the human body), greatly increasing system complexity! Figure 1.1 (next page) shows electron micrographs of some typical cells. Figure 1.1 Representative cells important in synthetic biology.

(a) Views of *E. coli* cells by scanning (left) and transmission (right) electron microscopy. Cell lengths are *ca.* **1-1.5**  $\mu$ m. The cells on the right have been negatively stained and show the nuclear zone with DNA.



(b) Views of *S. cerevisiae* cells by scanning (left) and transmission (right) electron microscopy. Largest cell dimensions are *ca.* **10**  $\mu$ **m**. The cell on the right has been negatively stained and shows the true eukaryotic nucleus as well as other intracellular compartments..



### 2. Small biomolecules and metabolism

When we take cells apart and examine their contents, these fall into two major categories: **small molecules** and "**macromolecules**" (big molecules). The most abundant constituent (about 70% by mass) is H<sub>2</sub>O (water), a very small molecule. Then come amino acids, sugars, fat molecules and salts like sodium chloride, potassium phosphate, etc. On the molecular weight scale where a hydrogen atom has a mass of 1 unit, the "small" biomolecules mostly have masses of a few hundred units or less. **Biological macromolecules**, on the other hand, have **molecular masses of 5000 units or more**.<sup>7</sup> The atomic mass unit is named a "dalton"<sup>8</sup> in honor of John Dalton (1766-1844), who first presented experimental evidence for the existence of atoms and devised a scale of their relative masses. When referring to proteins, DNA, and other macromolecules, it's convenient to use **kilodaltons (kDa = 10<sup>3</sup> daltons)** or **megadaltons (MDa = 10<sup>6</sup> daltons)** as units of mass.

Table 2.1 lists a representative subset of the several hundred smallmolecule "species"<sup>9</sup> typically found in living cells. Figure 2.1 shows space-filling models<sup>10</sup> for a few common metabolites on the list. Every cell contains its own characteristic set of small molecules. Strikingly, however, a large majority of these molecules are *common to all living cells*. This fact is an example of "**biochemical unity**," the remarkable similarity that organisms exhibit at the molecular level. **Metabolism** – a term familiar to scientists and non-scientists alike – constitutes **the complete set of biochemical reactions taking place inside a cell**.<sup>11</sup> Interconversions among the small molecules make up a large fraction of these reactions. In the spirit of the word genome (which means the complete set of genes) we now use the term "**metabolome**" to designate **all the small biomolecules** (metabolites) present in any given cell.<sup>12</sup> In synthetic biology the metabolome provides the source of raw materials (and energy).

<sup>&</sup>lt;sup>7</sup> This is an arbitrary cut-off; different authors may use slightly different values.

<sup>&</sup>lt;sup>9</sup> Chemists like to use the word species to indicate a precisely defined molecular entity. Normally this doesn't cause confusion with biologists' use of the same term (to mean a kind of organism), since the context normally makes clear which meaning is meant.

<sup>&</sup>lt;sup>10</sup> Space-filling models show the atoms with a size corresponding to their contact radii and a geometry corresponding to the atom positions in three dimensions.

<sup>&</sup>lt;sup>11</sup> In the case of multicellular organisms, "metabolism" refers to all the reactions taking place in *all* the cells (and some extracellular compartments like the stomach as well).

<sup>&</sup>lt;sup>12</sup> Note that "metabolome" as defined here omits macromolecules and the intermediates involved in their synthesis and breakdown. Although "metabolism" includes macromolecular metabolism, current usage tends to restrict "metabolome" to the small molecules.

Molecule		
	(Da)	
water	18	
sodium ion (a positively charged sodium atom)	23	
magnesium ion (a magnesium atom carrying two positive charges)	24	
chloride ion (a negatively charged chlorine atom)	35	
potassium ion (a positively charged potassium atom)	39	
carbon dioxide	44	
glycine (the smallest amino acid)	75	
lactate (the <i>in vivo</i> form of lactic acid)	89	
glycerol (a product of fat digestion)	92	
sulfate ion (a cluster of one sulfur atom with four attached oxygen	96	
atoms carrying two negative charges)		
dihydrogen phosphate ion (a cluster of one phosphorus atom with	97	
four attached oxygen atoms plus two attached		
hydrogen atoms with one negative charge)		
phenylalanine (an amino acid)	165	
ascorbate (vitamin C)	175	
glucose ("blood sugar")	180	
citrate (the <i>in vivo</i> form of citric acid)	189	
tryptophan (the largest amino acid)	204	
uridine-5'-monophosphate (a building block of RNA)	323	
deoxyadenosine-5'-monophosphate (a building block of DNA)	331	
lactose ("milk sugar")	342	
<pre>sucrose (plain old "sugar")</pre>	342	
cholesterol (a key cell-membrane component, precursor to all steroid	387	
hormones as well as bile salts)		
adenosine triphosphate (aka ATP, the cellular energy carrier)	504	
tripalmitin (a fat)	807	
cyanocobalamin (vitamin B <sub>12</sub> )	1355	

**Table 2.1** Some important small biomolecules and their molecular masses.

In a simple bacterial cell like *Escherichia coli*, for example, the metabolome numbers only a few hundred different kinds of molecule. Not counting water, the total number of small molecules per cell, however, amounts to roughly

250,000,000<sup>13</sup> (some species are present in many, many copies). Cells of higher organisms (eukaryotes) have approximately 1000 times the volume of bacterial cells, hence their metabolomes consist of **hundreds of billions of small molecules** – but the number of different types, though larger than that in bacteria, in most cases remains no more than a few thousand. The metabolism of all cells consists of a very large number of interlocking, enzyme-catalyzed reactions, simultaneously regulated so as to provide a **dynamic steady state**.<sup>14</sup> The exact concentration of any small biomolecule in the cell responds to the cell's needs at any given moment as a result of a **large network of regulatory interactions**. For example, if we stimulate an *E. coli* cell to start synthesizing a new protein, the reactions that lead to amino acids (the building blocks of proteins) will be "upregulated" (turned on) so as to provide the necessary materials.



**Figure 2.1** Some representative small-molecule metabolites shown as space-filling models – approximately to the same scale. Top row: glycine, glucose, lactose; bottom row: adenosine triphosphate (ATP), water. Color code: hydrogen – white, carbon – black, oxygen – red, nitrogen – blue, phosphorus – orange.

<sup>&</sup>lt;sup>13</sup> This calculation assumes that an average metabolite has a molecular mass of 200 Da. That may be too high, thus the total number of molecules is a minimum estimate. Note that this number includes membrane lipid molecules as well as water-soluble small molecules.

<sup>&</sup>lt;sup>14</sup> Note that this is not the same as an equilibrium. The flow of energy through cells keeps molecular concentrations at *non*-equilibrium values.

## 3. Biological macromolecules

Turning to the macromolecular components of cells, we encounter a pleasant surprise – an over-arching simplicity! All cells of all organisms contain only three general kinds15 of true macromolecule: proteins, nucleic acids and polysaccharides. Furthermore, cells synthesize each type of biological macromolecule from a small set of building blocks, interconnecting them with standard linkages: **peptide bonds** in the case of proteins, **phosphodiester bonds** in the case of nucleic acids, and **glycosidic bonds** in the case of polysaccharides. These chemical terms have precise definitions, but for now we will simply use them as labels. Peptide bonds are "the linkages that hold amino acid residues<sup>16</sup> together in protein chains," phosphodiester bonds are "the linkages that hold nucleotide residues together in nucleic acid chains," and glycosidic bonds are "the linkages that hold sugar residues together in polysaccharide chains." Note that the building blocks of proteins and polysaccharides are familiar – amino acids and sugars. In the case of nucleic acids, the building blocks have a less familiar name (nucleotides), though it isn't difficult to learn! Table 3.1 summarizes the basic facts about biological macromolecules.

Type of	Building Blocks	Linkage	Examples
Macromolecule			
protein	amino acids	peptide bond	collagen, insulin,
	(twenty kinds)		hemoglobin
nucleic acid	nucleotides	phosphodiester	DNA, RNA
	(eight kinds)	bonds	
polysaccharide	sugars	glycosidic bonds	starch, cellulose,
	(ca. fifteen kinds)		blood-group anti-
			gens

**Table 3.1.** The biological macromolecules.

<sup>&</sup>lt;sup>15</sup> Lipids comprise a fourth broad category of biomolecule, but their molecular weights are less than 5000. Since they shun water and aggregate together in its presence, however, they often display macromolecular behavior (intense light scattering, ready sedimentation, exclusion from small pores, etc.). Thus in some sense they are "honorary" macromolecules. However, because they are random aggregates, they lack the sequence-specific informational character of the proteins and nucleic acids (and some polysaccharides).

<sup>&</sup>lt;sup>16</sup> See section 6 (below) for an explanation of the term "residue" that properly describes the connected monomer units of biopolymers.

### 4. Protein functions

In 1838, Gerardus Johannes Mulder (1803-1880) first applied the term "protein" to the **complex**, often insoluble, nitrogen-rich substances found in living tissues.<sup>17</sup> The term derives from the Greek word  $\pi \varrho \omega \tau \epsilon_{ioc}$  ("*proteios*" – of first importance), and could not be more apt. As we now know, proteins perform a tremendous variety of functions in all living organisms, ranging from powerful and highly selective **catalysis** by enzymes, through structural support by components such as collagen, to transduction of photons of light into electrochemical nerve impulses by sensor molecules like the rhodopsin in your retina that enables you to read these words... Table 4.1 lists some of the biological functions that have been discovered for proteins. The complete list of known functions would be longer, and it continues to grow. One supremely important biological function is *not* on this list, however – proteins do not encode genetic information.

Function	Examples	
1. catalysis	most enzymes	
2. membrane transport	channels & carriers	
3. structure	collagen, keratin, actin	
4. transport	hemoglobin, serum albumin,	
	transferrin	
5. storage	myoglobin, ferritin	
6. defense	antibodies, toxins, venoms	
7. gene regulation	transcription factors, repressors	
8. motion	actin, myosin, dynein, flagellin	
9. signaling	insulin, growth hormone, cytokines	
10. water properties	anti-freeze protein, lung surfactant	
11. enzyme regulation	kinases, serpins	
12. adhesion	cadherin, marine adhesives	
13. light refraction	lens crystallin	
14. light absorption/transduction	photosystems I & II, rhodopsin	
15. light emission	fluorescent proteins, luciferase	
16. chemotaxis	cheY, motA	
17. nutrition	ovalbumin, casein	

#### Table 4.1 Protein functions\*

\*Note that "moonlighting proteins" can perform more than one (unrelated!) function.

<sup>&</sup>lt;sup>17</sup> The eminent Swedish chemist Jöns Jakob Berzelius had suggested the name to Mulder.

After about two centuries of intensive effort, chemists and biologists today can **describe proteins in immense**, **exquisite detail**. In many cases they also have very good ideas about how these remarkable molecules perform some of the functions just listed. If you want to understand life in scientific terms you need to learn a lot about proteins, and if you want to engineer living systems to do your bidding, you need to put this knowledge to work. Scientists in the field of "protein engineering" have done this for more than 30 years. Fortunately this primer can present **a simplified picture of protein structure and function** that **will enable a basic understanding of what protein engineers and other synthetic biologists do** that needn't go in depth into the enormous complexity of our present knowledge. Figure 4.1 shows a map of proteins found in yeast cells. Since the sample includes *all* of the proteins from the cell, it constitutes a display of the organism's "proteome."

**Figure 4.1** Overview of the yeast proteome by two-dimensional gel electrophoresis. The first dimension separates the proteins according to their electric charges ("isoelectric focusing") in a tube. The gel tube is then placed at the side of a flat gel slab and a perpendicular separation done according to the proteins' molecular weights ("SDS-PAGE"). Abundant proteins stain as large, dark spots; less abundant ones are fainter (or invisible).



### 5. Amino acids

The first fact about proteins concerns their composition. They are built from amino acids – nitrogen-rich, water-soluble, small molecules found in all living cells. All protein amino acids have the same fundamental chemical structure: a pivotal central carbon atom termed the " $\alpha$  carbon"<sup>18</sup> with four atomic groups attached (see Figure 5.1). Three out of the four attached groups are the same in every case, giving all amino acids fundamental common characteristics. The fourth group (R) differs from amino acid to amino acid, giving each molecule its distinctive identity.

### GENERALIZED FORMULA FOR PROTEIN AMINO ACIDS



**Figure 5.1** Generalized formula for protein  $\alpha$ -amino acids. The prefix " $\alpha$ -" indicates the carbon atom to which the other groups are attached. All the variation between amino acids occurs in the side chain (R) groups which contain from one to 18 atoms. Note that at physiological pH (*ca.* 7.0) the amino group has a positive charge and the carboxylate group has a negative charge. Many older textbooks show these incorrectly as –NH<sub>2</sub> and –COOH, respectively. Since electrostatic (ionic) charges play an important role in the interactions between molecules this is a non-trivial mistake.

Over a span of 130 years, chemists isolated and identified twenty protein α amino acids, starting with asparagine in 1806 and ending with threonine in 1935. We now know the stunning chemical fact that **all proteins of all organisms are constructed from the exact same set of 20 amino acids**, despite the fact that hundreds of other amino acids exist – and in principle could be substituted for members of the standard set.<sup>19</sup> Nature has clearly chosen these

<sup>&</sup>lt;sup>18</sup> Read "alpha carbon." Note that this is different from the " $\alpha$ -helix" described below.

<sup>&</sup>lt;sup>19</sup> Two additional amino acids, selenomethionine and pyrroyllysine, occur in small amounts in the proteins of some organisms.

building blocks with care and cunning, considering that they enable the phenomenal range of biological properties proteins display (*cf.* Section 4). The set of 20 amino acids shared by the proteins of all organisms not only stands as one of the most powerful examples of biochemical unity, but also plays an essential role in enabling synthetic biology. **Since all organisms' proteins share the same set of building blocks, proteins from one organism can be readily synthesized in the cells of another.** Table 5.1 lists the 20 protein amino acids, their abbreviations, their solubility characteristics, and the year that they were discovered in proteins.

	Abbreviation		Polarity*	Year discovered
Name	3-letter	1-letter	@ pH7	in protein <sup>+</sup>
Asparagine	asn	Ν	<b>P</b> <sup>0</sup>	1806
Glycine	gly	G	$\mathbf{P}^0$	1820
Leucine	leu	L	Н	1820
Tyrosine	tyr	Y	$\mathbf{P}^{0}$	1849
Serine	ser	S	$\mathbf{P}^{0}$	1865
Glutamate	glu	Ε	P-	1866
Aspartate	asp	D	P-	1868
Glutamine	gln	Q	$\mathbf{P}^{0}$	1879
Phenylalanine	phe	F	Н	1881
Alanine	ala	Α	Н	1888
Lysine	lys	К	<b>P</b> +	1889
Arginine	arg	R	<b>P</b> +	1895
Histidine	his	Н	<b>P</b> +	1896
Cysteine	cys	С	$\mathbf{P}^0$	1899
Proline	pro	Р	$\mathbf{P}^0$	1901
Tryptophan	trp	W	Н	1901
Valine	val	V	Н	1901
Isoleucine	ile	Ι	Н	1903
Methionine	met	Μ	н	1922
Threonine	thr	Т	$\mathbf{P}^0$	1935

#### Table 5.1 Standard Protein Amino Acids

\* P<sup>+</sup> - polar, positively charged; P<sup>-</sup> - polar, negatively charged; P<sup>0</sup> – polar, neutral; H - hydrophobic

+ Source: J.S. Fruton & S. Simmonds, *Biochemistry*, 2<sup>nd</sup> ed. (Wiley, NY, 1958); see also H.B. Vickery & C.L.A. Schmidt, *Chem. Revs.* **9**, 169-318 (1931)

### 6. Peptide bonds

As mentioned in Section 3, the amino acids in proteins are connected together by peptide bonds. **No matter which two amino acids get connected, the chemical structure of the linkage is the same**. It's important to make the point that the connection of two amino acids by **a peptide bond is a condensation** – when the amino acid moieties become linked, they release a water molecule. This means that the product contains three fewer atoms than the starting material. Thus the protein chain consists of **amino acid** *residues*, sets of atoms, each missing two hydrogens and one oxygen compared to the free amino acids.<sup>20</sup> Breaking a peptide bond requires adding the water molecule back – in a process called **hydrolysis**. Analogous condensation and hydrolysis processes occur in the synthesis and breakdown of nucleic acids and polysaccharides.

A key point about protein structure needs to be stressed. Consider a chain with only two amino acid residues. It gets formed by linking the **carboxylate group** of one residue to the **amino group** of the next. Thus the first residue still has its amino group (centered on a nitrogen) free, but its carboxylate group has become part of the peptide bond. The opposite is true of the second residue – whose amino group has been used to make the peptide bond, but its carboxylate group is still free. Thus the two ends of the molecule are different as shown in Figure 6.1 (next page). Such a **dipeptide** has *directionality*<sup>21</sup> – the left-hand residue still has a free amino group (and is called the amino- or N-terminus), whereas the right-hand residue retains a free carboxylate group (and is called the carboxylate- or C-terminus).

The dipeptide shown in Figure 6.1 comes from connecting glycine and alanine and is called glycylalanine. Nowadays chemists and biologists prefer to abbreviate the amino acid residues in peptides and proteins with single, uppercase letters (*cf.* Table 5.1). G stands for glycine and A for alanine. Hence we represent the dipeptide of Figure 6.1 simply as GA, with the convention that **the first letter corresponds to the amino-terminal residue**. A third amino acid residue can be added to either end of the dipeptide by a second condensation

<sup>&</sup>lt;sup>20</sup> The "lasso" shown in Figure 6.1 emphasizes that point. Nowadays mass spectrometry is extensively used to analyze proteins and the loss of a water molecule for each peptide bond formed affects the masses calculated for protein chains and their fragments.

<sup>&</sup>lt;sup>21</sup> Many textbooks misleadingly refer to this directional property as "polarity." That usage can be easily confused with the common chemical meaning of polarity – possession of full or partial electrostatic charges that confer water solubility (and hydrocarbon insolubility) on a molecule.

reaction. Figure 6.2 shows the complete formula for the tripeptide made by attaching a serine residue to the *C-terminus* of GA. Since S stand for serine, the tripeptide can be designated as GAS. The alternative tripeptide would be SGA. The peptide bonds in these structures have strong hydrogen bond donor (N-H) and acceptor (C=O) groups that contribute to stabilizing the folded shapes of protein chains (*cf.* Section 9).

**Figure 6.1** Peptide bond formation. The linkage in this case forms when the amino acid glycine (G) contributes its carboxylate group and alanine (A) its amino group to form the dipeptide glycylalanine (GA). Note that a water molecule (H<sub>2</sub>O) gets lassoed out in the process. The two ends of the peptide are different. On the left there is a free amino group (H<sub>3</sub>N<sup>+</sup>–), making the glycine residue the "amino" end (or N-terminus), and on the right a free carboxylate group (–COO<sup>-</sup>), making the alanine residue the "carboxylate" end (or C-terminus). The alternative arrangement, alanylglycine (AG) is a different molecule. Can you write its chemical formula?



**Figure 6.2** Chemical structure of a tripeptide. Here glycine is the N-terminal amino acid residue and serine is the C-terminal amino acid residue. Note the differences in the side chains (marked in green). This is *only one* of  $20 \times 20 \times 20 = 8000$  (!) different possible tripeptides that can be constructed from the standard set of 20 protein amino acids.



## 7. Protein primary (1°) structure

A medium-sized protein chain contains several hundred amino acid residues in a genetically determined sequence called its primary (1°) structure. Each residue contributes three atoms (-N $-C_{\alpha}-C$ -) to the backbone, so a protein with 300 residues has an unbranched chain 900 atoms long. In addition, the  $\alpha$ -carbon of every amino acid carries a specific, chemically distinct side chain (cf. Figure 6.2), and with 20 different amino acids to choose from at each position, it's clear that an astronomical number of different protein chains can be made. Twenty possibilities at the first position combine with twenty at the second to give 400 possible dipeptide sequences. Moving to the third position, this number gets multiplied by 20 more possibilities to give 8000 different tripeptides (cf. Fig. 6.2), and so on for successive positions all the way to the end of the chain. Obviously the total number (*n*) of different, possible 300-residue-long protein sequences can be calculated by multiplying 20 by itself 300 times (*i.e.*,  $n = 20^{300}$ ). This amounts to 2 x 10<sup>390</sup> (2 followed by 390 zeroes!). Such a number is *more* than astronomical – according to Einstein's theory of relativity there are only about 10<sup>80</sup> atoms in the entire universe!

Nowadays it's common to refer to all possible protein amino acid sequences as "**sequence space**," and this, of course, includes proteins both larger and smaller than the 300-residue chain just discussed. The smallest proteins are 50-60 residues long, and the largest one thus far described, a muscle protein appropriately named "titin," tops out at roughly 38,000 (!) residues (molecular mass 4.2 MDa). We could, of course, try to make a rigorous calculation of the size of protein sequence space, but this would be pointless given that the subset of that space corresponding to a 300-residue chain is already unimaginably huge. This fact has led to the realization that even over the entire span of evolutionary time (*ca.* 3.8 billion years) **nature has almost certainly not sampled all possible protein sequences**. Synthetic biologists thus have options to create and test *utterly* novel proteins.

### 8. Three-dimensional (3D) structure of proteins

Table 4.1 lists an amazing variety of biological functions carried out by proteins. However, though the amino acid sequences of different proteins characterize them, the sequences alone cannot explain their functions. Proteins behave rather differently from small molecules - to the dismay of early chemists whose techniques for purifying and studying small molecules generally failed when applied to proteins. If you subject a protein to comparatively mild, nonphysiological conditions such as temperatures above 60-70 °C, pH's below 4-5<sup>22</sup> or above 8-9, high concentrations of organic solvents like alcohol or solutes like urea or guanidinium salts, or strong shear forces, it usually loses its biological function(s). This phenomenon, called **protein denaturation** makes biochemistry a difficult art since the macromolecules biochemists study are often very delicate and denature easily.<sup>23</sup> Furthermore, with rare exceptions, denaturation is irreversible.<sup>24</sup> Cooking egg whites affords an everyday example. The so-called egg white is a very concentrated solution of many proteins that emerges from the raw egg as a viscous, transparent, almost colorless solution. Upon heating (or very rapid stirring - "beating") the proteins in the solution denature and become insoluble, forming the familiar, semi-solid white gel (see Figure 8.1). Cooling the denatured material (or stopping beating it) does not reverse the process.

An important fact about protein denaturation is that despite the complete loss of function and the dramatic change in physical properties of the protein, **its sequence remains unchanged.** Thus the biological function depends upon weak, non-covalent forces that stabilize a particular three-dimensional fold of the amino acid chain. **Destroy the fold and you destroy the function.** 

Even though proteins are enormous, complex and very delicate molecules, biochemists, beginning in the 1860s, discovered that many of them could be coaxed into forming crystals (just as smaller inorganic and organic compounds do). This key observation had two significant consequences: (1) it showed that proteins must have a *highly ordered structure*, and (2) it opened the door to the

<sup>22</sup> pH defines a logarithmic acidity scale. Pure, neutral water has a pH of 7.0. Lower values correspond to acidic solutions, while higher values correspond to basic solutions. A change of one unit on the scale corresponds to a 10-fold change in acidity (H<sup>+</sup> ion concentration), so that a pH 4 solution is 1000 times more acidic than a pH 7 solution. Living cells usually operate in the pH 5.5-8.5 range, *i.e.*, pH 7.0 ± 1.5, generally referred to as "physiological pH."

<sup>&</sup>lt;sup>23</sup> Of course, there are exceptions. Some proteins are tough and can withstand very harsh treatment, though often they are also small and comparatively uninteresting....

<sup>&</sup>lt;sup>24</sup> Only a few small proteins (most famously, bovine pancreatic ribonuclease A, M<sub>r</sub> = 13.7 kDa) have been completely unfolded, then refolded without the aid of chaperone proteins.

successful *determination of that structure* right down to the level of individual atoms **by X-ray crystallography**. Knowledge of the three-dimensional structure of proteins has proved inestimably valuable in understanding how they work and how they might be redesigned to work differently. The field of **protein engineering**, now more than three decades old, provides a major foundation for synthetic biologists. We are greatly indebted to the pioneers of protein crystallography<sup>25</sup> such as Max Perutz (1914-2002) and John Kendrew (1917-1997), who struggled for more than twenty years before successfully determining the first high-resolution protein structures in the late 1950s. Today (in mid-2007) **more than 44,578 structures of biological macromolecules are available** from the Protein Data Bank, the central repository for crystallographic results in biology.



**Figure 8.1** An example of denatured proteins (left) and a protein crystal (right). The myoglobin crystal (seen under the microscope) comes from a sample grown in space (where zero-gravity conditions favor more perfect crystallization).

<sup>&</sup>lt;sup>25</sup> The principles and techniques developed in the study of protein crystallography also apply to the other biological macromolecules.

#### 9. Encoded information generates structure.

Recognition that each protein has a unique primary structure and specific function(s) emerged in the 1950s, by which time biochemists had purified many proteins, and chemists had successfully developed techniques to determine their amino acid sequences. Using the single-letter abbreviations, an entire protein sequence, even a very long one, can be written out as a string of symbols. Like words in a book, these strings convey meaning to an informed observer (see Figure 9.1), *i.e.*, **proteins are informational macromolecules.** Given such sequence data, a chemist (or a computer!) can quickly write out the complete covalent molecular bonding pattern of the protein chain. Such a pattern applies equally well to the native as well as the denatured forms of the protein. How does the native, three-dimensional state of the protein arise from this primary structure? The obvious and simplest hypothesis is that somehow the chemical structure directs the folding of the chain into its native functional form, *i.e.*, **the information in the sequence determines the folding** – and thence the all-important biological function.

**Figure 9.1** Amino acid sequence of sperm whale myoglobin – the first protein to have its structure determined at atomic resolution. The information contained in this sequence determines the 3D fold of the native protein. *Cf.* Figures 8.1 and 9.3.

 $V-L-S-E-G-E-W-Q-L-V-L-H-V-W-A- \\ -K-V-E-A-D-V-A-G-H-G-Q-D-I-L-I- \\ -R-L-F-K-S-H-P-E-T-L-E-K-F-D-R- \\ -F-K-H-L-K-T-E-A-E-M-K-A-S-E-D- \\ -L-K-K-H-G-V-T-V-L-T-A-L-G-A-I- \\ -L-K-K-K-G-H-H-E-A-E-L-K-P-L-A- \\ -Q-S-H-A-T-K-H-K-I-P-I-K-Y-L-E- \\ -F-I-S-E-A-I-I-H-V-L-H-S-R-H-P- \\ -G-D-F-G-A-D-A-Q-G-A-M-N-K-A-L- \\ -E-L-F-R-K-D-I-A-A-K-Y-K-E-L-G- \\ -Y-Q-G \\ \\ \end{array}$ 

As mentioned above, denatured proteins almost never spontaneously renature. However, careful experiments in the 1960s showed that high yields of active, folded structures could be obtained with small, relatively stable proteins. Subsequently a variety of helping proteins or "**chaperones**" were discovered that have the ability to facilitate proper folding in more difficult cases by **preventing or reversing** *mis***folding**. Present evidence strongly supports the view that the sequence determines the folding – with the caveat that *in vivo*, chaperones often assist this process.

Support for the sequence-dependence of protein folding also derives from theoretical studies of protein-chain conformations and related experimental work on synthetic polymers of amino acids, *e.g.*, poly-alanine, poly-lysine, etc. As early as the 1940s, Linus Pauling (1901-1994) pointed out **the importance of non-covalent hydrogen bonding** in maintaining the fragile structures of proteins and other large biomolecules. Hydrogen bonding occurs when a hydrogen atom is attached to an oxygen or nitrogen atom, and as a result has a partial positive charge.<sup>26</sup> Such charges are attracted to concentrated lobes of negative charge ("unshared electron pairs") on other oxygen or nitrogen atoms. This attraction – only 5-10% as strong as the covalent bonds that hold atoms together in organic molecules – helps to stabilize a particular three-dimensional protein structure or "conformation." Individual H-bonds break easily at physiological temperatures, but like the Lilliputian ropes that immobilized Gulliver,<sup>27</sup> the sum of hundreds of them suffices to lock the flexible protein chain into its native, functional state.

The atomic groupings of **peptide bonds are particularly effective at creating hydrogen bonds**, and Pauling recognized that regular, repeating patterns of these bonds can generate stable local structures along the protein chain. The two **most important types of such "secondary (2°) structure,"** called the **\alpha-helix and \beta-sheet** occur very extensively in proteins. Figure 9.2 illustrates these geometrical arrangements of the protein chain.

While the  $\alpha$ -helix and  $\beta$ -sheet structures "emphasize the positive" by enabling H-bonds to form, another factor plays a key role in dictating the native,

<sup>&</sup>lt;sup>26</sup> The "electronegative" (electron-loving) character of oxygen and nitrogen means that when bonded to another atom like hydrogen, they distort the electron distribution around the other atom, leaving it with a partial positive charge.

<sup>&</sup>lt;sup>27</sup> The hero of *Gulliver's Travels* (Jonathan Swift, 1726) gets shipwrecked in Lilliput, a country where the people are only centimeters tall. While he lies on the ground sleeping, they use dozens of their miniature ropes to tie him down so that he cannot harm them. He can break the single ropes, but the whole collection suffices to imprison him.

functional folds of protein chains. This factor, often called the "**steric effect**"<sup>28</sup> since it stems from the fact that **two atoms cannot occupy the same space at the same time**, was the focus of the work of G.N. Ramachandran (1922-2001). He pointed out that at each amino acid residue the chain can readily turn around only two of the three backbone bonds,<sup>29</sup> and depending upon the rotation angles of these bonds it assumes different paths through space. By estimating the repulsive energy associated with each pair of values for these angles, Ramachandran plotted a two-dimensional **contour map** that had **hills of high energy and valleys of low energy** corresponding to minimal steric clashes. These valleys turned out to coincide with the geometry Pauling had predicted for the  $\alpha$ -helix and  $\beta$ -sheet! Thus, "eliminating the negative" for such arrangements reinforces "accentuating the positive," favorable energy contributions of H-bonding between backbone atoms.

**Figure 9.2** The two principal forms of protein 2° structure:  $\alpha$ -helix (left) and  $\beta$ -sheet (right). Hydrogen bonds are indicated with dashed lines. Color code: carbon atoms – black/gray, nitrogen atoms – blue, oxygen atoms – red, hydrogen atoms – white, side chains – lavender.



<sup>&</sup>lt;sup>28</sup> "Steric," like stereo, refers to space.

<sup>&</sup>lt;sup>29</sup> The third bond (the peptide bond itself) strongly resists rotation because of "partial doublebond character" – a feature frequently encountered in organic molecules

This description of the  $\alpha$ -helix and  $\beta$ -sheet has stressed their relation to the backbone features of the protein chain, and these by definition have no information content (since all residues have the same backbone atoms). However, the **side chains of the amino acids** that *do* **convey information** also participate in steric interactions (and to a lesser extent H-bonding). A further important property of the side chains is their degree of affinity for water. Hydrophilic amino acid side chains favor interactions with water, while the hydrophobic side chains shun water and prefer to interact with one another (or with other hydrophobic molecules such as membrane lipids). About half of the protein amino acids fall into each of these two categories.<sup>30</sup> The net result of the variation in side-chain properties is that the tendency to form a particular structure like an  $\alpha$ -helix or a  $\beta$ -sheet (or to form no ordered structure at all) depends upon the sequence. Thus the information content of the sequence expresses itself in the stability of a particular fold: the sequence determines the folding via non-covalent, weak chemical forces. Such forces not only favor the presence of  $\alpha$ -helix and  $\beta$ -sheet segments of secondary structure at certain positions along the chain, but also have additional interactions that contribute to the collapse of the elements of secondary structure into the complete fold of the entire chain – termed the tertiary (3°) structure. See Figure 9.3.

In some instances the formation of tertiary structure completes the emergence of a native, functioning protein. Since such proteins contain only a single polypeptide chain, they are said to be **monomeric.** However, many proteins, probably a majority, assemble into aggregates of monomer units by matching up complementary "sticky" surfaces that allow formation of weak non-covalent interactions such as hydrogen bonds, hydrophobic interactions and attractions between oppositely charged ions. This type of **multi-subunit higher order is called quaternary (4°) structure**. Two identical self-complementary chains can stick together to form a **homodimer** (symbolized  $\alpha_2$ ). Alternatively two different chains can form a **heterodimer** (symbolized  $\alpha\beta$ ).

Extension of multimer formation to larger and ever-more-complex aggregates is very common. Hemoglobin, for example, is a tetramer with two copies each of two different kinds of chain,  $\alpha_2\beta_2$ . In *E. coli*, the enzyme pyruvate dehydrogenase, a central player in aerobic energy metabolism, has a quaternary structure designated  $\alpha_{24}\beta_{24}\gamma_{12}$  with a total molecular mass of 5.3 million daltons! The enzyme that makes mRNA **in yeast cells (RNA polymerase II) has 12** *different* **protein chains and a molecular mass of more than half a million daltons** – a complexity consistent with the importance of the task it performs.

<sup>&</sup>lt;sup>30</sup> In some cases the side chains have both hydrophobic and hydrophilic parts.

Just as the amino acid sequence encodes the information to direct folding of a protein chain into its native tertiary structure, the sequences of the chains that make up the multimeric proteins in addition contain the **information necessary to display the proper complementary features at the subunit interfaces** that cause the monomers to stick together correctly.

**Figure 9.3** Protein 3° structure – myoglobin in ball-and-stick representation (top) and as a cartoon ribbon diagram (bottom). Note the  $\alpha$ -helices (myoglobin has no  $\beta$ -sheet structure).





## 10. Nucleic acid (RNA and DNA) primary structure

Work on the chemical details of nucleic acids began later and proceeded more slowly than that on proteins. By the early 20<sup>th</sup> century, however, their **key building blocks** had all been identified. Called **nucleotides**, each consists of three parts: (1) a residue from a nitrogen-rich, flat, "aromatic"<sup>31</sup> compound called a **base**, (2) a five-carbon **sugar** moiety – either **ribose or deoxyribose** – and (3) a **phosphate group**. These three parts of the nucleotide are always connected in an analogous way (see Fig. 10.1 on the following page), so that all nucleotides have major features in common. The **distinction between RNA and DNA** consists of the **absence of one oxygen atom in the ribose sugar** part of the *deoxy*ribonucleotides. The presence of that oxygen in the sugar of the ribonucleotides may seem like a minor change, but it has profound effects on the structure and reactivity of the two different types of nucleic acid.

Table 10.1 summarizes the basic building blocks of RNA and DNA. The thymine base found in DNA is simply a modified form of the uracil in RNA and the modification has no effect on the base-pairing properties to be discussed shortly.

Nucleotide	Base	Sugar	Abbreviation
adenosine-5'-monophosphate	adenine (A)	ribose	AMP
guanosine-5'-monophosphate	guanine (G)	// //	GMP
cytidine-5'-monophosphate	cytosine (C)	// //	СМР
uridine-5'-monophosphate	uracil (U)	<i></i>	UMP
deoxyadenosine-5'-monophosphate	adenine (A)	deoxyribose	dAMP
deoxyguanosine-5'-monophosphate	guanine (G)	<i></i>	dGMP
deoxycytidine-5'-monophosphate	cytosine (C)	// //	dCMP
deoxythymidine-5'-monophosphate	thymine (T)	11 11	dTMP

**Table 10.1.** Nucleotides: the building blocks of nucleic acids (RNA & DNA).

Analogous to the peptide bonds between amino acids that allow the construction of a long chain by successive condensation reactions, nucleotides can be linked by condensation **to form phosphodiester bonds**. Also analogous to the protein chains, **nucleic acid chains have directionality**. In this case the labels chosen to designate the two distinct ends of the chain are 5' ("five prime")

<sup>&</sup>lt;sup>31</sup> "Aromatic" compounds derive their name from fragrant natural products, but now the term is used by chemists to designate a class of organic molecule that has an especially stable form of bonding in structures with atoms in flat rings (often hexagons).



(a) A *ribo*nucleotide (5'-AMP)



(**b**) A *deoxyribo*nucleotide (5'-dTMP)



(c) Generalized nucleotide structure

**Figure 10.1** Chemical structures of representative nucleotides. The simplified diagram at the bottom (c) summarizes the arrangement of *all* nucleotides in DNA and RNA. The middle formula (**b**) shows the atom-numbering scheme, including the origin of the "primed" atom numbers in the sugar (which must be distinguishable from the atom numbers of the base). Note the connection between atom N1 of the thymine (T) base and atom C1' of the deoxyribose sugar. This is the glycosidic bond. Bonds of this general type link the bases to the sugars in all nucleotides. Shading indicates the flat, aromatic character of the bases in (a) and (b).

and 3' ("three prime"), derived from the atom numbers of the sugar carbons to which the phosphate groups are attached. Thus starting at one end of the chain and walking along the backbone, if the first carbon atom you encounter is a 5' C, you have started at the 5' end of the chain (and *vice versa* for the 3' end). See Figure 10.2 on the next page.

RNA and DNA chains look very similar and can grow to great lengths. The **backbones consist of alternating sugar and phosphate fragments** to which



**Figure 10.2** Chemical structure of a tetranucleotide (4 nucleotide residues connected by 3 phosphodiester bonds. This short oligomer illustrates the principles of nucleic acid primary structure. *Note the directionality.* The glycosidic bonds (in red) are set angles that position the nucleobases in the *anti* conformation (pointed away from the backbone). This arrangement puts the hydrogen-bonding donor and acceptor groups facing outward where they can interact with the complementary bases to form the base pairs of the double helix (see Section 11).

**bases are attached at carbon 1' of the sugar** (ribose or deoxyribose). Each successive sugar group can carry any of the four bases, just as each  $\alpha$ -carbon atom in a protein chain can carry any of the twenty amino acid side chains. Thus **nucleic acids, like proteins, are informational macromolecules**. With only four choices of base at each position, they have significantly less potential variety for a given number of monomer units than do proteins.<sup>32</sup> Note that since each phosphodiester linkage has a single negative electric charge, the **intact RNA or DNA chains have as many electrostatic charges as nucleotides**. This fact means

<sup>&</sup>lt;sup>32</sup> A chain 300 nucleotide residues long can be constructed in only (!)  $4^{300}$  (*i.e.*,  $10^{91}$ ) ways. Contrast that with the 300-residue protein example given in section 7 (which can have any one of 2 x  $10^{390}$  different sequences).

that separation by electrophoresis works extremely well for nucleic acid molecules.

Generally speaking, native DNA molecules are much longer than RNA **molecules**. This reflects nature's selection of RNA – whose backbone breaks more easily than that of DNA – for tasks that require shorter fragments and/or rapid turnover. Among stable RNAs in cells, the largest<sup>33</sup> only reaches a few thousand nucleotide residues (nts) in length. By contrast, the DNA in each eukaryotic chromosome occurs as a single, gargantuan double-stranded piece millions of nucleotide residues long. In humans, for instance, the DNA of chromosome 1 (our largest chromosome) has more than 250,000,000 nucleotide residues in each of its two complementary chains.<sup>34</sup> The unicellular organisms popular with synthetic biologists have much smaller DNAs. The entire DNA of an E. coli K-12 cell, for example, is a single, double-stranded circular molecule with (only!) 4,650,000 nts in each strand. Yeast cells contain 16 chromosomes<sup>35</sup> with a total single-strand DNA length of 12,156,678 nts. The largest yeast chromosomal DNA molecule has 1,531,918 nts in each strand.<sup>36</sup> Thanks to the DNA-sequencing projects, we can write out these DNA lengths exactly, down to the last nucleotide residue. However, most discussions are better served by rounding the numbers off. Thus we normally say that the E. coli K-12 chromosome is 4.65 Mb ("megabases"<sup>37</sup>) long and the largest yeast chromosome 1.53 Mb.

- <sup>36</sup> The specific numbers given for yeast DNA molecules come from the standard organism
- described in the Saccharomyces Genome Database (SGD), http://www.yeastgenome.org/.

<sup>&</sup>lt;sup>33</sup> The RNA of the large ribosomal subunit – which is about 3600 nucleotides long in bacteria and 5000 nucleotides long in eukaryotes.

<sup>&</sup>lt;sup>34</sup> And thus carries a total of *half a billion negative ionic charges!* 

<sup>&</sup>lt;sup>35</sup> Not counting the small mitochondrial DNA (85,779 nt long).

<sup>&</sup>lt;sup>37</sup> Note that common parlance uses "bases" (or sometimes "base pairs") to express the sizes of DNA molecules. While this isn't exactly correct (since the chemical monomer units are nucleotides, not bases) it's convenient – and stresses the importance of the bases that embody the sequence information.

### 11. The double helix

Perhaps the greatest achievement of 20th-century biology was the realization that, at the chemical level, the structure and composition of living cells derives from a genetic program encoded in DNA. Discovered by Friedrich Miescher (1844-1895) in 1869, and called "nuclein" because of its location in the cell nucleus, DNA was not convincingly demonstrated to be the genetic material until the mid-1940s.<sup>38</sup> Before that time most chemists thought that it had a comparatively low molecular weight (perhaps a few thousand daltons) and a very boring structure – or perhaps no regular structure at all. It certainly did not seem to have the characteristics required to encode large amounts of information and biologists generally assumed – erroneously – that genes were proteins. By the 1930s, however, careful experiments had begun to show that the **molecular** weights of DNA molecules extend into the megadalton (MDa) range, and by 1950, analysis of their nucleotide residue compositions showed an intriguing regularity – the number of dAMP residues in a DNA sample always equals the number of dTMP residues and similarly, the number of dGMP residues equals the number of dCMP residues. (Or, put more simply, A = T and G = C.) The awareness of DNA's genetic significance put a premium on understanding its higher-order molecular structure in order to uncover the mechanism of storage, replication and transfer of information.

Using good DNA preparations, a gifted X-ray crystallographer, Rosalind Franklin (1920-1958), stretched fibers from concentrated DNA solutions and showed that they produced high-quality **X-ray diffraction patterns**. This meant that the DNA in these fibers had an **ordered structure** – analogous to the situation described above for crystalline proteins. Since the fiber diffraction patterns yielded only a limited amount of data (compared to true crystals) they could not directly reveal the "structure of the gene." To overcome this difficulty, in **1953**, **Francis Crick (1916-2004) and James Watson (1928- )** used all available information concerning the chemical and physical properties of DNA to construct a hypothetical model of the native molecule – the now-famous double helix.

The key step in developing the Watson-Crick model for DNA structure was the recognition that the A = T and G = C relationships depended upon specific **"base-pairing" interactions** between the aromatic moieties of the

<sup>&</sup>lt;sup>38</sup> Though, remarkably, the 19<sup>th</sup> century German cell biologists correctly identified its function using histological staining methods; those methods were considered unreliable, however, and their work was forgotten for 60 years!

nucleotide residues of the DNA. These interactions are dominated by **patterns of hydrogen bonding** that allow H-bond donors (groups with a partially charged H atom bound to an O or N) to match up with H-bond acceptors (groups with concentrated lobes of negative charge on O or N atoms), *cf.* Section 9 above. Figure 11.1 illustrates the correct matches. The incorrect ones (A-C and G-T) miserably fail to stabilize an interaction. In order to stabilize a higher-order (2°) structure to the maximum extent, all the bases must form pairs.



Figure 11.1 Watson-Crick base pairs shown attracting nucleotide residues from two different DNA strands to each other. Note the complementarity: In the G:C pair, at the upper left (in the red box), an Hbond acceptor group (=O) on the G base associates with an H-bond donor group (H-N) on the C base. In the middle position G has a donor (N-H) and C has an acceptor (N). At the lower right, G has another donor N-H and C has an acceptor (O=C) group. So, from top to bottom these groups encode the pattern "ADD" on the G base and "DAA" on the C (where D = donor and A = acceptor). The corresponding analysis for the A:T pair gives "**DA**" on the A base and "AD" on the T base. What happens when you try to switch the C and T bases and form complementary pairs? (Analyze this by attempting to match D and A groups.)

Various lines of evidence convinced Watson and Crick that the DNA structure involved **double-stranded molecules**. That, together with the basepairing almost defined the ultimate model. Additional constraints came from Franklin's X-ray data (peeked at by Watson without Franklin's knowledge in an episode that has spawned endless debates about ethics in science). Crick had previously done the fundamental work on the general theory of X-ray diffraction by helical molecules, and as soon as he learned the main features of Franklin's data, he could define the overall geometry of the secondary structure of DNA. It had to be a **right-handed**, **base-paired**, **double helix with approximately 10 base pairs per turn**. The flat aromatic bases "stack" on top of one another with their planes almost perpendicular to the helix axis. (See Figures 11.2 and 11.3.) Stacking interactions stabilize the structure.



**Figure 11.2** Base-paired DNA strands spontaneously wrap up into an *antiparallel* double helix. The figure on the right is copied from the original publication by Watson and Crick [*Nature* **171**, 737 (1953)]

One other feature of the proposed DNA structure must be understood to interpret correctly the many processes of DNA manipulation involved in synthetic biology: the **two strands run in opposite directions** (they are said to be **"anti-parallel"**). Thus at the end of a piece of helical DNA one strand (call it the "Watson") will have a 5′ group and the other (the "Crick") a 3′ group. Depending on how the DNA has been constructed and handled, either one or both of these groups may be a phosphate ( $-O-PO_{3^{-1}}$ ). If the phosphate is missing, the end consists of a "hydroxyl" group (symbolized as –OH).



**Figure 11.3** Space-filling models of three types of DNA double helix. The biologically important structure is called B-DNA. It predominates under physiological conditions. If the water content of a solvent is decreased (by adding ethanol, for example), the helix unwinds a bit and broadens to form A-DNA. Under very special conditions (in the presence of high contentrations of certain types of ions) the helix reverses its twist to adopt a *left-handed* form called Z-DNA. Up to now the evidence for biological functions for either A-DNA or Z-DNA has not been convincing.

The sequences of the two strands of "double-helical" DNA are said to be "reverse complements" of one another. If the Watson strand has a sequence segment that runs 5'-ATGCCGTT-3', for example, the Crick strand will have the matching segment 5'-AACGGCAT-3' opposite to it – and in real space, running in the other direction. Note the mental gymnastics necessary to match these segments when they are written in the conventional  $5' \rightarrow 3'$  direction.<sup>39</sup>

The Watson-Crick double helix constitutes the **secondary (2°) structure of DNA**. Such a structure is stable at physiological temperature (37 °C) and below, but if the temperature increases, the native **DNA double helix unwinds and denatures**, allowing the two strands to separate. If a solution containing the separate strands at a high temperature (95 °C, say), cools slowly, the complementary strands find each other and renature to form the original helix. Such "melting" and "reannealing" processes are important in manipulating

<sup>&</sup>lt;sup>39</sup> It takes practice to become accustomed to this, but the advantage of consistently writing DNA sequences in the conventional direction far outweighs the effort needed to manipulate their reverse complements to check the complementarity.

**DNA in the laboratory**, most notably in the **polymerase chain reaction (PCR)**. The temperature at which half of a sample of DNA molecules has denatured is referred to as the **T**<sub>m</sub> (**melting temperature**) for that sample. It **depends upon the salt concentration** (the higher the salt concentration the higher the T<sub>m</sub>). Similarly, increasing the **percentage of GC base pairs** in DNA increases T<sub>m</sub>.

The essential features of the 1953 Watson-Crick model (now called **"B-form" DNA** – see Figure 11.3) are still accepted today for most DNA sequences *in vivo*, though fine structural details of specific pieces have had to be adjusted on the basis of extensive X-ray diffraction and other physical studies. As it turns out, the **precise details of the double-helix geometry vary slightly** with the **base sequence** and the **molecular environment**. Such subtleties have been shown to play roles in the functions of genes, but their full implications are not yet clear. For the purpose of engineering genetic systems, however, we can probably overlook them, at least initially.

In their elegant 1953 paper announcing the double-helix model for DNA structure, Watson and Crick commented that "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." This surely ranks among the classic understatements of all time. The complementarity of the two DNA strands means that their **information content is redundant**. The sequence of one strand, together with the base-pairing rules, automatically determines the sequence of the second strand. Thus **duplication of the genetic information merely requires unwinding the double helix and synthesizing two new strands using the original strands as templates**. And that's exactly how nature does it – with the aid of a now well-characterized macromolecular machine called **DNA polymerase**. In synthetic biology (and in simpler genetic engineering procedures) this enzyme plays an important role as a tool to replicate and amplify DNA.

We thus reach the conclusion that the inherent physical properties of DNA chains – resulting from the same kinds of weak, non-covalent interactions that stabilize protein 3° structures – confer upon them the capacity for selfreplication, one of life's defining characteristics. Once a double-helical DNA molecule exists, it has the capacity for endless duplication.

### 12. RNA – the urbiomolecule

Although RNA chains, like DNA chains, consist of long sugar-phosphate backbones with bases dangling off at the side, they cannot arrange themselves into a DNA-like, B-form helix, even when the interacting molecules have complementary sequences. The **2'-OH group of the ribose sugar blocks formation of a stable B-form geometry**. Instead, although the bases still interact according to the base-pairing rules (A pairs with U and G pairs with C), the RNA helix is **less tightly wound** (*ca.* 11 base pairs per turn rather than 10) and the planes of the bases tilt away from the perpendicular to the helix axis. (See Figure 11.3.) We call this an **"A-form"** structure or **"double-helical RNA."**<sup>40</sup>

In sharp contrast to DNA, normally only a fraction of the nucleotides in an RNA molecule become part of a double-helical structure. A single strand of RNA can have complementary sequences at different locations, and if these align in an anti-parallel fashion, they will create short pieces of A-form RNA helix. Figure 12.1 shows how this might occur for an RNA chain 46 residues long so as to generate a "hairpin." Note that at least four or five unpaired bases in a loop are required to allow a stable double-helical hairpin structure to form in RNA at physiological temperature. Such helices constitute most components of **RNA secondary (2°) structure**. Often they are quite irregular, with some nucleotides sticking out of the side of the helix, forming a "bulge." Another irregularity often encountered in RNA helices is pairing between G and U residues (as well as G-C pairing). Figure 12.2 illustrates an actual RNA fragment with these types of irregular feature as determined by x-ray crystallography.

**Figure 12.1** An RNA "hairpin." This structure acts to terminate gene transcription in *E. coli* and is available as part number BBa\_B0011 from the Registry of Standard Biological Parts. [Note that the automated folding prediction program used to make this illustration (Mfold) has used DNA bases, thus T's are shown where the actual molecule would have U's.]



<sup>&</sup>lt;sup>40</sup> As noted in Section 11, DNA also can adopt the A-form geometry when it becomes dehydrated (by addition of alcohol to a DNA solution, for example), but this situation does not occur *in vivo* where the water concentration always remains high – except in dormant forms like seeds and spores.


**Figure 12.2** A segment of the large ribosomal subunit (23S) RNA showing how several hairpins can fold to generate a tertiary structure. This part of the rRNA binds to protein L11. (Taken from <u>http://www.scripps.edu/mb/millar/research/hairpin.htm</u>.)

**Transfer RNA** (tRNA) molecules are among the most thoroughly studied RNAs. They play **key roles in protein synthesis** (see below), and in addition to the formation of short hairpin helices (shown in the cloverleaf diagram of Figure 12.3a), they fold up into a more elaborate **tertiary (3°) structure** that includes hydrogen bonds involving atoms other than those that participate in the standard Watson-Crick base pairs (see Fig. 12.3b). These structures are analogous to the protein (3°) structures described earlier (see Section 8).



**Figure 12.3** Transfer RNA secondary (a) and tertiary (b) structure. The 3' end is on the right in (b). The amino acid corresponding to the anticodon (at the bottom) gets attached to the terminal AMP residue via an ester bond to the ribose 2' or 3' –OH group.

At this point, **no pure RNA 4° structures have been identified**, though in the complex aggregate of RNA and proteins that makes up the large ribosomal

subunit, interactions occur between different RNA molecules, and the interactions between ribosome subunits during translation (see below) mostly involve RNA-RNA contacts – essentially a form of RNA 4° structures.

RNAs have a diverse set of roles in all cells, some of which are listed in Table 12.1. Synthetic biology mainly involves the first four functions given.

Functional Role	Example
structural support	ribosomal RNAs (rRNAs)
group carrier	transfer RNAs (tRNAs)
information carrier	messenger RNA (mRNA), tRNA
terminator of RNA	GC-rich stem-loop followed by
synthesis	UUUU in <i>E. coli</i>
attenuator of mRNA	Trp-operon attenuation sequence
synthesis	in E. coli
catalyst	peptidyl transferase, RNase P
viral genome	influenza, HIV
RNA splicing	type I intron
mRNA regulator	interfering RNA (RNAi)
riboswitch	leader of guanylate operon mRNA
release ribosome	tmRNA
edit message	guide RNA
methylation guide	RdDM

Table 12.1RNA functions.

Note that although RNA acts as the genome for a wide range of viruses, it *never* does so for cells. This may be due to an ancient selective process in which cellular RNA genomes were eliminated from this role because **RNA is intrinsically less chemically stable than DNA**. Along with blocking formation of B-form helices, the 2'-OH group of ribose facilitates the cleavage of RNA chains. This makes RNA vulnerable to hydrolysis (breaking apart by water) and helps to explain why, in modern cells, it serves the role of temporary information carrier (messenger RNA, see below).

Before the 1970s, biochemists dogmatically asserted that "all enzymes are proteins," *i.e.*, all biological catalysis depends upon protein enzymes. However, in a dramatic turnabout, Thomas Cech (1947-) and Sidney Altman (1939-) showed that **some reactions have "ribozyme" catalysts** – RNA molecules that specifically bind to substrates in cells and catalyze key processes. Since then, the

study of RNA catalysts has burgeoned, not least of all because RNA molecules can be sent through a process of **"directed evolution"** *in vitro* **to select ever better functional variants** from among vast numbers<sup>41</sup> of similar sequences. The most dramatic discovery in the ribozyme field so far has come from the structural analysis of ribosomes (the intracellular machines that synthesize proteins (see below). In the year 2000, Thomas Steitz (1940- ) and coworkers demonstrated that the actual catalyst for peptide bond formation in all cells is the largest piece of ribosomal RNA (rRNA). Previous to that, rRNA had been assumed to have only a structural role. Now the catch phrase has become **"the ribosome is a ribozyme!"** 

As will become evident (cf. sections 14-16 below), **RNA participates in** virtually every aspect of information transfer in all cells of all organisms. Furthermore, it has the capability to act as a genome (in viruses) and as a catalyst (in ribozymes). Speculations about the possible origin(s) and early evolution of life have focused on these facts and led to the proposal that the first "living" systems may have depended exclusively on RNA, with DNA emerging as a more stable genetic material at some later stage. Consistent with this idea, the biosynthesis of *deoxy*nucleotides (the building blocks needed to make DNA) occurs by conversion from the corresponding *ribo*nucleotides that are made earlier in the common nucleotide biosynthetic pathway. Researchers nowadays continue to find other roles of RNA in modern cells (cf. Table 12.1) that suggest an amazing ability to carry out all the essential functions of living systems except for the compartmentalization provided by cell membranes. The idea that life began with RNA-based systems has been dubbed "the RNA world **hypothesis.**" This view now dominates the thinking of most evolutionary biologists, but it's too soon to conclude that RNA truly was the *ur*biomolecule.<sup>42</sup>

Synthetic biologists need to understand the functions and behavior of RNA, not only because of its importance in the biological systems they seek to manipulate, but also because it affords (together with proteins) a **macromolecular material that can be engineered** to fill important roles in the systems we seek to create. Not least of all, the directed-evolution techniques pioneered by Jack Szostak (1952-) for creating RNA "aptamers" that have high specificity for binding and interacting with other molecules offer a powerful method for screening enormous numbers of RNAs to select those most suited to a particular design target. An alternative approach based solely on computation

<sup>&</sup>lt;sup>41</sup> A typical value is 10<sup>15</sup> (1,000,000,000,000,000).

<sup>&</sup>lt;sup>42</sup> In German the prefix "*ur*-" denotes the most ancient member of any historical lineage, and nowadays is also occasionally used with that meaning in English as well.

has been successfully used by Ronald Breaker (1966-) to create riboswitches (See Section 18).

Another issue for synthetic biology concerns the secondary structures RNA molecules can form. These modulate the processes of transcription and translation; hence in many cases **RNA sequences that can form 2° structures play important roles in engineering gene expression (especially terminators).** Conversely, inadvertent creation of RNA secondary structural elements by coupling genetic units together as part of a synthetic scheme may lead to inhibition of the desired functions. Efforts are underway to devise methods to detect this type of potential problem and find engineering solutions for it – such as **"insulators" that block unwanted RNA-RNA interactions between designed structural elements**.

# 13. Polysaccharides – structure *and* information

Polysaccharides, the third category of biological macromolecule, have long suffered a reputation for dullness. **Cellulose**, the **most abundant organic material on earth**, plays a structural role in plants. Its molecules occur as very long, unbranched polymers of glucose residues connected by the same repeating linkage (a  $\beta 1 \rightarrow 4$  glycosidic bond) over and over and over again. **Starch**, the **major energy source** in the diets of most humans, has both linear and branched molecules, again with repeating linkages ( $\alpha 1 \rightarrow 4$  and  $\alpha 1 \rightarrow 6$  bonds). Other polysaccharides such as **chitin**, the **basis for arthropod exoskeletons**, have similarly simple features. The animal body produces some more complex polymers that have roles in providing lubrication to joints and modulating the viscosity of fluids such as mucous. Many of these **glycosaminoglycans** (GAGs) have medical importance. While there are some interesting aspects to their chemistry and biology, they are not informational and only concern a subset of synthetic biologists at the present time.

By contrast, the pieces of polysaccharide attached to many proteins and some membrane lipids in eukaryotic organisms, including humans, have major consequences for synthetic biology. Called *oligosaccharides* (because they have a relatively small number of sugar residues) these structures are **attached to proteins destined for secretion** or function as **components of the plasma membrane** of the cell. They serve as barcode-like labels that provide information about the cell on whose surface they reside or the protein to which they are attached. Figure 13.1 illustrates the manner in which these groups differ from one another – they vary in the **positions and geometries of the intermonomer linkages** and in the **identity of the monomers**. The familiar ABO blood group typing for humans reflects the variation of such oligosaccharides on the surfaces of red blood cells, and as everyone knows, these variations confer immunological specificity that must be carefully considered when transfusing blood.

As synthetic biologists move towards experiments with higher organisms they will need to take into account the possible **roles of oligosaccharides in defining the behaviors of the cells** (and secreted proteins) that they engineer.



Figure 13.1 Oligosaccharide patterns on a variety of animal and viral proteins.

### 14. The "Central Dogma"

How does genetic information, encoded in the nucleotide (or "base") sequence of DNA direct the complex processes that give rise to a living cell? Obviously the most important issue concerns how the genetic material (DNA) directs production of the major functional components of cells (proteins). This posed a particularly acute problem for early molecular biologists studying eukaryotes, where DNA resides in the nucleus and the proteins are made and mostly function in the cytosol (outside the nucleus),. A related issue concerns the manner in which some proteins are "induced" when the cell needs them and otherwise get produced only at low, "constitutive" levels.

By following the uptake of radioactivity into protein after exposing cells to short pulses of labeled amino acids, workers in the 1950s showed that **proteins are synthesized in ribosomes** – cellular factories, gigantic on the macromolecule scale ( $M_r = 2.52$  MDa in *E. coli* and 4.22 MDa in animals), but barely visible with a light microscope. Rapidly growing *E. coli* contain about 50,000 ribosomes per cell. Analysis of ribosomes demonstrates that they consist of about 50% protein and 50% RNA by mass, implicating RNA in protein synthesis. Key discoveries made by Francois Jacob and Jacques Monod proved that the type of protein produced depends on a small *fraction* of the ribosomes (rRNA) is quite stable.

The **sequences of the short-lived RNA match genes in the DNA**, explaining how the information from the genes can reach the ribosome. A crucial enzyme called **RNA polymerase**<sup>43</sup> unwinds the double helix of the gene to be expressed and, with U substituting for T, **synthesizes an RNA copy** of one of the two DNA strands (called the **"coding strand"**). This copy then migrates to a ribosome where it directs the synthesis of the protein corresponding to the copied gene. Jacob and Monod termed such RNA molecules **"messenger RNA"** (mRNA).

We thus have a straightforward connection between the **sequence information in DNA** and the production of functional proteins. Note that this is a one-way street. **Proteins** *cannot* **provide the information** needed to produce

<sup>&</sup>lt;sup>43</sup> In all cells this enzyme is a large, multi-subunit protein (*Cf.* Section 9).

their genes. The simple scheme for genetic information flow, first spelled out in a 1958 paper by Francis Crick, is:

#### $DNA \rightarrow RNA \rightarrow protein$

This can be elaborated by saying that a deoxyribonucleotide (gene) sequence directs biosynthesis (by RNA polymerase) of a ribonucleotide (mRNA) sequence that in turn programs ribosomes to synthesize an amino acid (protein) sequence. Inherent in this scheme is the key to synthetic biology – **change the DNA sequence and you change the protein**.

Because the "languages" of DNA and RNA are the same (nucleotide sequences capable of recognition via Watson-Crick base-pairing) the RNA polymerase step is referred to as **transcription**. Proteins, consisting of amino acid residues, have a different language, thus ribosome-catalyzed protein synthesis is termed **translation**. Note that structural and regulatory RNA molecules are transcribed, but *not* translated.<sup>44</sup>

Section 11 introduced DNA replication, so we can elaborate the possible biological information transfers by adding that to the above schematic:

$$\bigcirc DNA \rightarrow RNA \rightarrow protein$$

This diagram summarizes what Francis Crick somewhat whimsically called the "Central Dogma" of molecular biology in his 1958 paper. Since that time two additional paths have been discovered that are indicated by smaller arrows in Figure 14.1: (1) synthesis of DNA strands copied from an RNA template by the enzyme **reverse transcriptase**, and (2) copying of RNA from an RNA template by **RNA replicase**. The first of these has important applications in biotechnology and synthetic biology. The bulk of sequence information transfers in biology occur by the paths given in the original simple version of the Central Dogma.

<sup>&</sup>lt;sup>44</sup> tmRNA is an interesting exception to this rule.



**Figure 14.1** The "Central Dogma" of molecular biology. Large red arrows denote the major paths of sequence information flow; smaller blue arrows show specialized cases. The *absence* of an arrow connecting protein to RNA has major significance.

# 15. Protein synthesis

**Translation of mRNA** occurs in<sup>45</sup> the ribosome by an elaborate process whose **basic features are conserved across all known biological species**. The ribosomes of archaea, bacteria and eukaryotes have similar 3D structures and carry out the same biochemical reactions in an essentially identical manner.<sup>46</sup> Table 15.1 summarizes the composition of bacterial and mammalian ribosomes. On the molecular scale these are truly **mammoth objects** — and have a corresponding complexity. Notably – in contrast to many other large biological structures like viruses or multi-subunit enzymes – they **utterly lack symmetry**!

PROKARYOTES* 70S	EUKARYOTES# 80S
Small subunit (SSU/30S)	Small subunit (SSU/40S)
<b>16S rRNA</b> (1542 nt)	<b>18S rRNA</b> (1874 nt)
<b>21 proteins</b> (avg. M <sub>r</sub> ~ 16,000 Da)	30 proteins
850,000 Da	
Large subunit (LSU/50S)	Large subunit (LSU/60S)
<b>5S rRNA</b> (120 nt)	<b>5S rRNA</b> (~120 nt)
<b>23S rRNA</b> (2904 nt)	<b>5.8S rRNA</b> (158 nt)
<b>31 proteins</b> (avg. M <sub>r</sub> ~ 15,000 Da)	<b>28S rRNA</b> (4718 nt)
1,450,000 Da	40 proteins
* = 11	

Table 15.1 Ribosome components.

\*E.coli

Sequence data for both the proteins and the rRNAs show that **remarkable similarities exist among the components of all ribosomes of all known living cells**, and we must conclude that **they derive from a common ancestral ribosome** that emerged at the earliest stages of evolution. In the year 2000, crystallographers first succeeded in determining the X-ray crystal structures of both ribosomal subunits at high resolution. Additional structural information has been published since then<sup>47</sup> and the structures have provided a gold mine of

*<sup>&</sup>lt;sup>#</sup> R. norvegicus* 

<sup>&</sup>lt;sup>45</sup> Using the word "in"(not "on!") in this situation better suits our present understanding of the nature of the ribosome.

<sup>&</sup>lt;sup>46</sup> Ribosomes from very divergent organisms have some clear functional differences. Eukaryotic ribosomes in particular employ more components to accomplish the same tasks, and mitochondrial ribosomes substitute protein in some cases for RNA to carry out their functions. Nevertheless, the main mechanism of protein synthesis has been conserved for 3.5 billion years!

<sup>&</sup>lt;sup>47</sup> Including very informative studies done by cryo-electron-microscopy.

opportunities to elucidate the detailed step-wise mechanism of protein synthesis. For the purposes of this primer, however, most of this detail will be omitted. (See **REFERENCES** for access to review articles at several levels of sophistication.) The summary of protein synthesis presented here, however, is consistent with these details.

Simply stated, the ribosome must attach itself to an mRNA molecule, and in response to the encoded signals, construct the corresponding protein chain. In addition to **selecting the correct amino acid for each position**, it **forms peptide bonds** and eventually **stops and releases the new protein**. Even **before synthesis has finished, the growing chain begins to fold** into a native, threedimensional structure (*cf.* Section 9), usually aided by chaperones.

The key challenge to understanding how this all-important cellular process occurs is to explain how the sequence of nucleotides in mRNA allows the ribosome to recognize which specific amino acid (out of the standard set of 20) to insert at each position. One early hypothesis was that the bases in the mRNA somehow interact directly with the side chains of the amino acids. Despite considerable effort, however, early molecular biologists failed to demonstrate such an interaction. In a stroke of genius, Francis Crick solved this problem in 1956-58 with the "**adaptor hypothesis**." He reasoned that a short sequence of ribonucleotides would be required to decode the mRNA message using basepairing for recognition. **Amino acids could be specifically attached to relatively small pieces of RNA** by "activating enzymes" that would recognize both the amino acid and the RNA adapter. The **RNA portions of these conjugates would then associate with appropriate segments of the mRNA through base-pairing**.

By 1958, biochemists had already begun to discover the key components required by the adaptor hypothesis. Activating enzymes driven by breakdown of adenosine triphosphate (ATP), an energy-rich molecule, were shown to **couple amino acids to** a class of **RNA molecules 70-90 nucleotide residues long**. Because these small RNAs transfer the activated amino acids to growing protein chains in the ribosome, they were eventually **called transfer RNAs (tRNA)**. Most cells<sup>48</sup> contain twenty different activating enzymes (now known as "**aminoacyl-tRNA synthetases**"), one for each of the 20 standard protein amino acids. They also contain at least one tRNA for each amino acid, and typically there are several distinct "isoaccepting" tRNA species for each amino acid.

<sup>&</sup>lt;sup>48</sup> A few minor exceptions occur, but they do not alter the fundamental pattern described here. Specific tRNAs always get "charged" with the correct amino acid.

Thus every cell contains a collection of 50-60 different tRNAs that normally are maintained in the "charged" or aminoacylated form by 20 aminoacyl-tRNA synthetases. When a protein is needed, **RNA polymerase** makes an mRNA copy of the coding strand of the appropriate gene, and this attaches to a ribosome. As seen from Table 15.1, ribosomes come in two pieces, the large subunit (LSU) and the small subunit (SSU).<sup>49</sup> When not making a protein chain these subunits remain separate. They come together in a process termed **translation initiation**. This begins with binding of mRNA to the SSU together with a special initiator tRNA and separate proteins called initiation factors. The SSU complex then associates with the LSU to form a primed 70S complex that then – with the help of additional protein factors – binds a second aminoacyl-tRNA. Formation of the first peptide bond (between amino acid no. 1 and amino acid no. 2) occurs next, after which the (now uncharged) tRNA from amino acid no. 1 exits the ribosome. Rearrangement of the complex of ribosome, mRNA and tRNA no. 2 (which now carries a dipeptide) assisted by elongation factors allows the third aminoacyl-tRNA to enter the active site of the ribosome and the second peptide bond is formed. **Successive rounds** of this process **continue to lengthen the nascent (newly made) protein chain**, which grows from its amino-terminus towards its carboxyl-terminus as the mRNA is "read" in the 5'-to-3' direction. Ultimately the mRNA presents a termination signal to the ribosome and protein "release factors" cleave the bond that links the protein chain to the last tRNA.

Folding begins even before the nascent protein chain emerges from the ribosome, and gets completed after the last amino acid has been added and the chain released. In some cases this completes formation of the native, biologically active protein. Other proteins – especially in eukaryotes – must undergo **post-translational modifications. Peptide bonds may be cleaved**, often with the loss of some pieces of the chain, **crosslinks may be formed**, and **additional chemical groups may be attached**. All of these changes come about as the result of the actions of enzymes in the cell, so if a synthetic biology project involves producing a protein that has post-translational modifications, the modifying enzymes must be on hand to create them. This becomes **a key issue in the biosynthesis of eukaryotic proteins in bacterial cells**.

Another complication in eukaryotic cells concerns **modification of the mRNA** before the ribosomes can effectively translate it. To summarize very simply, the **primary transcript of a protein-coding gene in a eukaryote must be** 

<sup>&</sup>lt;sup>49</sup> In *E. coli* (and other prokaryotes) these are known by the values of their sedimentation coefficients in the ultracentrifuge: 50S and 30S.

**capped, spliced and polyadenylated**. **Capping refers to chemical modification at the 5'-end** of the transcript. The modifications (not shown here) serve to protect the mRNA from premature degradation – and provide a distinguishing tag for ribosomes to recognize. **Polyadenylation occurs at the 3' end** of the mRNA and also provides some degree of protection from degradation and contributes to the efficient recycling of ribosomes. The 3'-untranslated region (3'UTR) of the initial transcript of most eukaryotic genes is longer than the finished message and encodes a **"cleavage and polyadenylation" site** where the primary transcript gets cut following which **a sequence of about 200 AMP residues (a polyA tail)** gets added.

**Splicing occurs in the middle of the primary transcript** where there are RNA sequences that do not get translated into the final protein product of the gene. Since they appear to be inserted into the message they were named **introns** to distinguish them from the sequences that are ultimately expressed (termed **exons**). Eukaryotic cell nuclei contain an **elaborate set of splicing factors** (about 150 proteins and five RNA molecules) that assemble into a large supramolecular complex called a **spliceosome**. The key reactions involved in splicing are thought to be catalyzed by the RNA components of the spliceosome – acting as sequence recognition elements and ribozymes. Thus it isn't surprising that **in some cases the introns are self-splicing**. This occurs mainly in lower organisms, including some prokaryotes.<sup>50</sup>

<sup>&</sup>lt;sup>50</sup> For a long time introns were thought to be exclusively found in eukaryotes. That's no longer true, but prokaryotes clearly have very few introns.

### 16. The genetic code

With the role of mRNA established and the adaptor hypothesis supported by experiments with tRNA and aminoacyl-tRNA synthetases, molecular biologists in the 1960s faced the challenge of determining the relationship between the nucleotide sequences of mRNAs and the amino acid sequences of the corresponding proteins. Put differently, **what is the code that allows the ribosome to interpret the information in mRNA as an amino acid sequence?** Since there are 20 amino acids and four nucleotides, each amino acid requires a minimum of three nucleotides to specify it uniquely.<sup>51</sup> In fact, the system has the capacity to encode more than 20 amino acids since there are  $4 \times 4 \times 4 = 64$ possible nucleotide triplets. *A priori* the code could be irregular, with some amino acids coded by three nucleotides and others by two, or it could be overlapping, or have still more complex features. Fortunately for us, the simplest situation prevails: **each amino acid corresponds to one or more nucleotide triplets (called "codons") that are adjacent and non-overlapping in the mRNA.** 

"Cracking" the genetic code – that is, assigning one of the 20 amino acids to each of the 64 possible triplets – began with the observation in 1961 by Marshall Nirenberg (1927- ) and Heinrich Matthei (1929- ) that in a cell-free protein-synthesizing system, addition of polyuridylate (polyU) resulted in the exclusive synthesis of polyphenylalanine. Thus **the codon UUU specifies phenylalanine**. Led by a tour-de-force of chemical synthesis of defined RNA molecules in the laboratory of Har Gobind Khorana (1922- ), **the complete genetic code was worked out by 1966**. Figure 16.1 displays the standard genetic code. Note that three codons (**UAA**, **UAG and UGA**) do not encode an amino acid, but rather **play the very important role of signaling the ribosome to stop**. Two amino acids (methionine and tryptophan) have only one codon, while the others have from two to six synonymous codons. **The sole methionine codon** (**AUG**) does double-duty because it **also functions as the initiation signal**.

It's important to recognize that the standard genetic code is only *near*universal. In eukaryotes the mitochondria have 2-5 substitutions in the code and in some organisms nuclear code changes also occur. The official repository for all gene sequences (GenBank) specifies 16 alternative codes in its taxonomy section: (<u>http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c</u>).

<sup>&</sup>lt;sup>51</sup> A one-to-one ratio of nucleotides to amino acids allows only 4 amino acids. A two-to-one ratio allows 16, which is still not enough. Thus the minimum codeword size is three nucleotides.

Table 16.1 The Standard ("Universal") Genetic Code

	Coondination						
		U	С	А	G		
	U	UUUC UUC UUA UUG } Leu	UCU UCC UCA UCG	UAU UAC UAA Stop UAG Stop	UGU UGC UGA UGG Trp	U C A G	
st letter	С	CUU CUC CUA CUG	CCU CCC CCA CCG	CAU CAC His CAA CAA GIn	CGU CGC CGA CGG	U C A G	Third
Firs	A	AUU AUC AUA AUG Met	ACU ACC ACA ACG	AAU AAC AAA AAA AAG	AGU AGC AGA AGG Arg	U C A G	letter
	G	GUU GUC GUA GUG	GCU GCC GCA GCG	GAU GAC GAA GAA GAG Glu	GGU GGC GGA GGG	U C A G	

# Second letter

As **synthetic biologists** retrieve genes from exotic organisms and install them in other cell types, they **need to consider** the possibility that the gene they are using has some **atypical codons**. That's mostly true, of course, when the gene comes from a mitochondrial genome.

Another complication related to the genetic code concerns "codon usage." Synonymous codons exist for all but two amino acids. Thus evolution can achieve the same final result in a protein sequence by using more than one mRNA sequence. In any particular cell, **the population of tRNAs** (most of which read more than one codon) **is optimized for the frequency with which particular codons are used** in that cell. A foreign gene may not be translated efficiently in such an environment. A final point needs to be made concerning the **"reading frame"** of the mRNA. Consider a simple 22-nt sequence: ...GUCCAUGACAUACAAUGCUAAC... This can be parsed into three different arrangements of nucleotide triplets (called reading frames):

Frame #1 5'...GUC CAU GCC AUA CAA UGC UAA C...3'
Frame #2 5'...GU CCA UGC CAU ACA AUG CUA AC...3'
Frame #3 5'...G UCC AUG CCA UAC AAU GCU AAC...3'

These can be translated (*cf.* Figure 16.1) as:

Peptide #1 N...val his ser ile gln cys **stop** c Peptide #2 N...pro cys his ser **met** leu... c Peptide #3 N...ser **met** pro tyr asn ala asn...c

Obviously these correspond to different chemical structures. So, **how does the ribosome know where to start reading?** The answer **in eukaryotes** is that it latches onto **the first available** methionine codon (**AUG**) that it encounters **after starting at the 5' end** of the message. Both frames #2 and #3 have an AUG codon in the segment given, so if there are no other AUG sequences between the 5' end of the RNA (which lies to the left) and the segment shown, the eukaryotic protein sequence would begin at the **met** residue shown in bold in frame #3. In **bacteria**, a special sequence of 4-6 nucleotides called a **ribosome binding site (RBS) lies about 10 nucleotides upstream of the AUG** codon where initiation begins. In either case the **initiator tRNA locks onto the AUG** and sets the reading frame – which the ribosome strictly maintains all the way to the end of the coding region of the message. Thus for any protein-coding gene, **only one of the three potential reading frames gets used**.

# 17. Regulation of the metabolome (small-molecule metabolism)

The metabolome, the collection of small-molecule metabolites in every cell (*cf.* Section 2), exists in a **dynamic steady state**.<sup>52</sup> The metabolites undergo rapid, enzyme-catalyzed interconversions, moving back and forth through a vast network of connected metabolic pathways. The **net flux** in these paths always **leads to states of lower free energy**. The question of what regulates this complex system lies at the heart of our understanding of cellular functions. It particularly concerns synthetic biologists, whose efforts to engineer cells inevitably influence (and are influenced by) metabolite concentrations.

Since the metabolites can only interconvert with the help of enzymes, understanding how their concentrations are regulated amounts to asking **"What controls enzyme activities in the cell?"** Obviously the **number of enzyme molecules** plays a major role – the next section on gene expression addresses that point. Once synthesized, however, the activity of enzymes depends upon (1) the **concentrations of substrate molecules**, (2) **non-covalent interactions with other cellular components** and (3) **covalent chemical modifications of the enzymes** (most of which are proteins).

To catalyze a metabolic reaction, an enzyme molecule must bind substrate(s) at some catalytic center on its surface (the "active site"). **Evolution has fine-tuned the binding affinities of enzymes** to match the corresponding concentrations of their substrates. **When the substrate concentration falls** below the range of a particular enzyme's binding affinity, the **rate of that metabolic step slows down**, and vice versa. Thus, if we express a foreign enzyme in a cell where the substrate concentration(s) is(are) significantly lower than in the cell of origin, it will not perform as expected.<sup>53</sup>

Another type of regulation involves **small-molecule binding to a special** "allosteric" site at a location on the enzyme surface different from the active site. When the small molecule – usually referred to as an "effector" or "ligand" – binds, it influences the enzyme's activity by inducing a subtle change in the three-dimensional structure of the protein (*cf.* Section 8). Such ligand binding may either increase or decrease the enzyme's catalytic rate, *i.e.*, the ligand can act as a positive or a negative effector. This type of regulation occurs very frequently in metabolism. Particularly important (and clear) examples involve

<sup>&</sup>lt;sup>52</sup> Not an equilibrium!

<sup>&</sup>lt;sup>53</sup> In that case, it might be possible – though not easy – to re-engineer the enzyme's binding site to increase the affinity.

the inhibitory **binding of the end product** of a biosynthetic pathway **to the enzyme that catalyzes the first unique step** along the path that makes it.

A related kind of regulation involves **binding of another macromolecule** to an enzyme in such a way as to **alter its catalytic rate** – **or even the reaction that it catalyzes**. The synthesis of lactose in mammary gland provides a dramatic example. Before parturition (giving birth) an enzyme called **galactosyl transferase** attaches galactosyl groups<sup>54</sup> to glycoproteins. Even though the mammary cells have glucose available, the enzyme does not perform the closely related reaction of attaching galactosyl groups to glucose – which would produce lactose (aka milk sugar). Immediately after birth, however, the cells start to produce a small protein named  $\alpha$ -lactalbumin. This second protein binds tightly to galactosyl transferase, preventing it from reacting with glycoproteins, and causing it to make lactose instead. The tightly associated proteins constitute a new enzyme, **lactose synthase** – to which all mammals are indebted for most of the carbohydrate nutrition they enjoy during the first months of their lives!

Both allosteric regulation and binding of modulator subunits to enzymes leave their **covalent chemical structure** unchanged. Another, **more drastic kind of regulation involves altering that structure**. One very common type of alteration consists of cutting peptide bonds. **Digestive enzymes** are often synthesized in **inactive precursor forms** (so that they don't damage the cells in which they are made), then when the active forms are needed, **activating enzymes cut a small number of peptide bonds** and transform the precursors into powerful degradative catalysts (this activation occurs outside the cells that make the precursor). An elaborate cascade of similar processes controls **blood clotting**. Cleavage of short **"signal peptides"** that serve to localize proteins to the places they function in the cell also amounts to a regulatory process.

A second kind of covalent alteration consists of attaching chemical groups to enzymes. Protein phosphorylation (attachment of a –PO<sub>3</sub><sup>2-</sup> group)<sup>55</sup> is the most important example. Though especially common in eukaryotes, it also occurs in bacteria and archaea. The enzymes that phosphorylate proteins (using ATP as the source of the phosphoryl group) are called protein kinases. The human genome encodes 518 protein kinases,<sup>56</sup> some of which have already turned out to be critical targets for anticancer drugs. Current estimates are that at least one third of all mammalian cellular proteins get phosphorylated. The

<sup>&</sup>lt;sup>54</sup> Galactosyl groups contain the simple sugar galactose minus one of its –OH groups.

<sup>&</sup>lt;sup>55</sup> to the –OH groups on the side chains of serine, threonine or tyrosine.

<sup>&</sup>lt;sup>56</sup> Called the "kinome!"

effects of phosphorylation can be positive (increasing enzyme activity) or negative, and quite significantly, **protein phosphatases can remove the phosphoryl groups** and reverse them.

A number of **other chemical groups are attached post-translationally** (after synthesis of the protein chain) to proteins and in some cases regulate enzyme activity. Some important examples include acetylation, adenylation, ADP-ribosylation, carboxylation, glycosylation, hydroxylation, methylation, prenylation and sulfation. Not all cells can perform all of these modifications, so **synthetic biologists need to know when such protein refinements are necessary** to the function of one of their components, in order to insure that their chassis can provide them.

The previous paragraphs mostly refer to intracellular regulatory processes. It's important to realize that **extracellular signals** play significant roles in controlling intracellular processes. In multicellular organisms **hormones** have this function – and the responses to them include many of the processes described above, especially protein phosphorylation. The **quorum-sensing system in bacteria** responds to extracellular signals and leads to collective behavior of large numbers of cells.

## 18. Protein control of gene expression

Virtually all **cells have more genes than they are using** at any given time. That's abundantly true of multicellular organisms where different cell types need different gene products—especially different proteins—to display their specialized characteristics. Even bacteria, however, do not express all their protein-coding genes all the time. Furthermore, genes that are being expressed more often than not have varying levels of expression, depending upon the cell's needs at different times. This section describes protein-mediated **regulation of protein-coding gene expression** mainly in bacteria where it is best understood and of the **greatest current interest for synthetic biology**. The next section (**19**) describes the prototypical system (the "lac operon") that molecular biologists first studied in detail to analyze gene expression.

In the light of the Central Dogma (*cf.* Section 14), it's clear that **expression of protein-coding genes could be controlled either at the level of transcription** (**RNA synthesis**) **or translation (protein synthesis**). Not surprisingly – since it conserves more cellular resources – transcription control dominates. However, all cells exert control over expression of many genes at the level of translation as well.

Bacteria and bacteriophage<sup>57</sup> (bacterial viruses) have simpler regulatory systems than eukaryotes, though all cells have many features in common. In bacteria, transcription involves binding of the RNA polymerase (*cf.* Section 14) to DNA at a site centered about 20 nucleotides "upstream"<sup>58</sup> of the point where transcription begins. This RNA polymerase binding site is called a **"promoter,"** and the exact DNA sequence in that region strongly affects the ability of the polymerase to synthesize a message. **"Strong" promoters** lead to efficient binding and initiation of mRNA synthesis. In the absence of a repressor (see below) a strong promoter causes **formation of many mRNA molecules** that in turn get translated into **high levels of the corresponding proteins**. Other, weaker promoters lead to lower levels of gene expression. Thus, when designing a genetic system, one choice that a synthetic biologist must make is **the type of promoter to incorporate for each mRNA start site**.

<sup>&</sup>lt;sup>57</sup> "Bacteriophage" literally means "bacteria eater" and is often shortened simply to "phage." Note that both words are *collective nouns* (like deer or fish) and require no "s" for the plural. <sup>58</sup> On the 5′ side (by convention always placed on the left in written sequences or schematic diagrams).

A second mechanism for controlling gene expression involves **regulatory proteins called activators or repressors** that bind to DNA near the transcription start site and either stimulate or block RNA polymerase. A well-studied example is the **tryptophan repressor** that helps control the transcription of a set of five genes that encode enzymes required to make the amino acid tryptophan. These genes reside adjacent to one another on the *E. coli* chromosome,<sup>59</sup> and are transcribed into a single long piece<sup>60</sup> of mRNA. The **trp repressor protein binds** tightly to a specific DNA **"operator"** sequence **adjacent to the transcription start site** and **blocks RNA polymerase** so that it cannot initiate mRNA synthesis (See Figure 18.1). This is the *only site* in the entire *E. coli* chromosome where this repressor binds. It can only do so, however, if it has bound tryptophan first. When the intracellular **concentration of tryptophan drops below an optimum level** the repressor loses its bound tryptophan, changes conformation and dissociates from the DNA. **RNA polymerase then transcribes the genes.** 



Figure 18.1 Crystal structure of trp repressor/operator complex.

 <sup>&</sup>lt;sup>59</sup> Such a cluster of related genes is called an "operon." These arrangements are common in bacteria and archaea, but relatively rare in eukaryotes.
 <sup>60</sup> 7kB (7,000 nts)

**Catabolite-gene activator protein (CAP)**,<sup>61</sup> the **best-known activator** protein in bacteria, helps regulate a large number of genes. When glucose abounds, bacteria prefer it to all other sources of carbon and energy, but when intracellular glucose levels fall, they must switch on sets of genes encoding proteins that allow uptake and metabolism of other food/fuel molecules. (The lac operon discussed in the next section is one such set.) CAP binds to sites adjacent to the promoter regions upstream of the genes it activates, but to do so, it **must first bind cyclic AMP** (cAMP), a special nucleotide signaling molecule produced from ATP by the enzyme adenylate cyclase. High levels of glucose inhibit adenylate cyclase – and also stimulate systems that expel cAMP from the cell. Only **when glucose is in short supply do the cAMP levels increase, causing CAP to switch on its target genes.** 



**Figure 18.2** *E. coli* catabolite-gene activator protein bound to its DNA recognition site. The crystal structure, shows that CAP binding bends DNA. Different colors distinguish the two identical subunits of the protein dimer.

**Eukaryotic transcriptional regulator proteins** function similarly to their prokaryotic counterparts, though they come in a **wider variety of structural types** and tend to act as **components of multimeric complexes**. The eukaryotic promoter region has a binding site for a set of protein "general transcription factors" that act to localize RNA polymerase at the promoter. The key player for many promoters is the **TATA-binding protein (TBP)**, named for the dominant DNA sequence in its binding site. As Figure 18.3 shows, this "anchor" protein creates a very noticeable distortion of the DNA structure, helping to mark the site for other proteins to "feel." Additional regulatory proteins assemble **upstream of the promoter**, and in some cases they bind at quite distant locations (called "enhancers") and force the DNA to loop around when they interact with promoter bound proteins. (See Figure 18.4.)

Given the complexity of eukaryotic gene regulation as illustrated by

<sup>&</sup>lt;sup>61</sup> Also known as CRP (cyclic-AMP receptor protein).



**Figure 18.3** TATA-binding protein (TBP) attached to its binding site in the eukaryotic promoter. Note the severe bend it induces in DNA.



**Figure 18.4** Cartoon of eukaryotic RNA polymerase II at its promoter, poised to commence mRNA synthesis. The components are approximately drawn to scale, though the DNA loop may be thousands of base-pairs long. The mediator complex consists of about 20 proteins, evolutionarily conserved from yeast to humans. The chromatin remodeller complex and the HAT (histone acetyl transferase) assist in releasing the gene to be transcribed from its histone packaging. (See Section 21 for details on DNA packaging.)

Figure 18.4, it's not surprising that the term "**transcription factor**," commonly used to describe the proteins involved has a less-than-precise definition. For our purpose, we can define it as "**a protein that binds to specific regulatory sites in genomic DNA and modulates gene expression**." According to current estimates, **the human genome encodes more than 1,000 transcription factors**. Many of these respond to small-molecule ligand binding – by hormones, for example, providing a direct link between the endocrine system and gene expression. Figure 18.5 illustrates how transcription factors may combine in a specific instance to regulate an important gene (that encodes the enzyme phospho*enol*pyruvate carboxykinase). The obvious **combinatorial complexity** here demonstrates the types of **challenges higher eukaryotic organisms pose for synthetic biology**.

**Figure 18.5** Transcription factor binding sites in the promoter and upstream region of the gene for the liver enzyme phospho*enol*pyruvate carboxykinase (PEPCK). The figure and caption are taken from Hanson & Reshef, *Annu. Rev. Biochem.* **66**, 581-611 (1997). The ligands that interact with the transcription factors shown here include retinoic acid, thyroid hormone, cortisol, insulin, cyclic-AMP, leukotriene, fatty acids and possibly prostaglandins. Numbers at the bottom indicate nucleotide distances from the transcription start site (given as 0).



Figure 1 Transcriptional regulatory elements of the PEPCK-C gene promoter. The positions of regulatory elements in the PEPCK-C gene promoter are represented by ovals, and the various proteins that regulate transcription of the PEPCK gene are shown relative to their binding sites on the promoter. The abbreviations used are as follows: CRE, cAMP regulatory element; TRE, thyroid hormone regulatory element; GRE, glucocorticoid regulatory element; TBP, TATA binding protein(s); IRE, insulin regulatory element; PPARRE, peroxisome proliferator-activated receptor regulatory element; AF-1, accessory factor-1; C/EBP, CAAT/enhancer binding protein; CREB, cAMP regulatory element binding protein; DBP, D-binding protein; NF-1, nuclear factor-1; HNF-1, hepatic nuclear factor-1; HNF-3, hepatic nuclear factor-3; GRU, glucocorticoid regulatory unit; GR, glucocorticoid receptor; RAR, retinoic acid receptor; RXR, retinoic X receptor; PPAR $\gamma$ 2, peroxisome proliferator-activated receptor  $\gamma$ 2. This figure is from Nizielski et al (144).

### 19. The lac operon

The genes needed to allow *E. coli* to switch to the **use of lactose as a source of carbon and energy** constitute the lactose (or *lac*) operon – probably the most intensively analyzed genetic system ever. Not only does it illustrate some of the key principles described in the previous section, but it also has provided **well-understood genes and control mechanisms** that genetic engineers have exploited extensively for over thirty years. In that way, it represents **a prototype for** the more sweeping engineering that characterize **synthetic biology**, and nowadays its components play important roles in synthetic biology projects. The classic analysis of the system was carried out in the 1950s and 1960s by **Francois Jacob and Jacques Monod.**<sup>62</sup> Significantly, their work depended on genetic analysis, not biochemical experiments. However, subsequent molecular characterization of the genes and proteins involved spectacularly confirmed their conclusions.

Figure 19.1 diagrams the features of the *lac* operon. Three "structural" genes Z, Y and A code for (1)  $\beta$ -galactosidase, the enzyme that breaks lactose, a disaccharide, into galactose and glucose, its constituent monosaccharides; (2) lactose permease, a transport protein that allows lactose to enter the cells; and (3) thiogalactoside acetyltransferase (GAT), an enzyme that plays a still-only-partly-understood role in eliminating toxic thiogalactosides taken up by lactose permease. The fourth gene (*lacI*) codes for the *lac* repressor, a DNA binding protein similar to the trp repressor described in Section 18. In addition, as mentioned above, under conditions of low intracellular glucose concentration the CAP-binding protein plays an activating role for the *lac* operon. Its binding site lies just upstream from the promoter.

The **key mechanism** for regulating expression of the *lac* operon involves **binding of the** *lac* **repressor protein to its operator** site adjacent to (and overlapping) the beginning of the *lac*Z gene. In contrast to the trp repressor described in Section 18, the *lac* repressor binds to its operator in the *ligand-free* **state**. As described in the legend to Figure 19.1, **two weaker operators also participate** in shutting down this set of genes. **In the absence of lactose** (or another suitable  $\beta$ -galactoside) the **repression is** *nearly* **100%**.<sup>63</sup> When lactose concentrations rise in the cell, small ("basal") levels of  $\beta$ -galactosidase convert lactose into **the actual trigger** of gene expression, an isomer called **allolactose**.

<sup>&</sup>lt;sup>62</sup> Who also postulated the existence of mRNA, see Section 14.

<sup>&</sup>lt;sup>63</sup> The system is "leaky," however, with occasional mRNA synthesis occurring that leads to the production of enough  $\beta$ -galactosidase to convert any incoming lactose to allolactose.

This compound binds tightly to the repressor, causing a conformation change that sharply lowers its affinity for DNA. The **ligand-bound repressor** then **dissociates from the DNA**, and RNA polymerase can productively bind the promoter and start to transcribe mRNA from the *lacZ*, *lacY* and *lacA* genes. For this to be an efficient process in wild-type cells, however, **the CAP protein must also be bound to the promoter**. That requires glucose levels to fall so that adenylate cyclase produces sufficient cyclic-AMP to convert CAP into its DNA-binding conformation (See Section 18).



**Figure 19.1** Diagram of the *lac* operon genes. The blue arrows indicate the transcription start sites for the *lac*I repressor mRNA and the *lac*ZYA enzyme-coding mRNA. The red bar (t) marks the transcription termination site for the second mRNA. Operator O<sub>1</sub> overlaps the promoter and provides a binding site for a dimeric repressor. Operator O<sub>2</sub> lies 400 nts downstream from O<sub>1</sub>, *within the coding sequence* for *lacZ*, and can bind a second repressor dimer. These two dimers can interact to stabilize repressor binding at the promoter and prevent transcription. The intervening DNA must form a loop to allow this to occur.<sup>64</sup> The lengths of genes *lacI*, *lacZ*, *lacY* and *lacA* are 1040, 3510, 780 and 825 nts, respectively. The CAP site lies to the left of the P*lac* promoter, 6-7 nts away from the end of the *lac*I gene.

 $<sup>^{64}</sup>$  Another weak operator site lies 90 nts upstream of O<sub>1</sub> and a repressor dimer bound there can also associate with the primary dimer bound to O<sub>1</sub>. In that case the O<sub>2</sub>-bound dimer must first dissociate.

# 20. Regulation of gene expression by RNA

The importance and versatility of RNA in cellular functions has become increasingly appreciated in recent years in the wake of **major new discoveries**. This section summarizes our present knowledge of control systems based primarily on RNA rather than protein.

In addition to the transcription control imposed by protein repressors and activators, bacteria have **a form of RNA-based control called attenuation**. The tryptophan biosynthetic genes (discussed in Section 19) provide an example. In the "leader sequence" of the mRNA (the first 140 nts) there are three possible hairpin sequences, the third of which constitutes a **terminator for RNA polymerase**. The leader sequence encodes a "leader peptide" with 14 amino acid residues, two of which are tryptophan. The coding region for the leader peptide includes the first half of the first hairpin of the mRNA, and **if tryptophan is in short supply**, the ribosome stalls while synthesizing the peptide. This prevents formation of the first hairpin with the result that **the terminator hairpin** also **cannot form** – the second (exposed) half of the first hairpin interacts with the first half of the terminator and prevents it from participating in the terminator structure (see Figure 18.1). At least **half a dozen amino acid biosynthetic operons in** *E. coli* **have attenuation-regulated transcription;** this mechanism has been found in other organisms as well.

Domain 1 Domain 2 Domain 3 Domain 4 5 tp leader region 5 tp lead

# Attenuation of the trp operon mRNA

**Figure 18.1** The mechanism of transcription attenuation.

Within the past six years, a related, but **more complex type of built-in**, **mRNA-based gene expression control** has been discovered. Just as the sequences of some RNA molecules fold into tertiary structures with catalytic activity (ribozymes), others can fold into structures that display **highly specific binding sites for small molecules**<sup>65</sup>. If an mRNA molecule has such a sequence upstream of the translation start position,<sup>66</sup> **binding of a small molecule can produce a conformational change** in the folded RNA structure that either causes formation of a terminator hairpin<sup>67</sup> or otherwise alters the structural state of mRNA so as to affect **either transcription or translation**. **Most reported cases of these "riboswitches" involve inhibition** of the expression of the affected genes, but in at least one case the switch triggers activation.

Synthetic biologists have already begun to exploit riboswitches in several ways. They are **readily available parts to use in engineering novel systems**. Furthermore, they can be artificially evolved to bind to non-natural ligand molecules. Part II of this primer will examine the engineering aspects of this topic in greater detail.

Another conformational trick used by some RNA molecules to regulate gene expression involves a **segment of mRNA just ahead of the translation start site** that folds into a **tertiary structure resembling the ultimate RNA binding partner of the encoded protein**. In *E. coli*, for example, at least six ribosomal proteins control their operons by this mechanism. Since the completely synthesized, folded protein normally binds to rRNA, the messenger that encodes it displays a **mimic of the rRNA binding site** to which the protein binds, blocking further translation. As the ribosomal proteins are used to assemble new ribosomes, the mRNA-bound molecules dissociate, releasing the allowing the messenger to direct the synthesis of more (now-needed) protein.

A very dramatic discovery by Andrew Fire (1959) and Craig Mello (1960) in 1998 has revolutionized biologists' views on the regulatory role of RNA. Working in *C. elegans*, Fire and Mello discovered that **double-stranded RNA** with a sequence corresponding to a piece of mRNA **can lead to blocking translation** of that mRNA or even cause its **total degradation**. This **RNAi** ("RNA interference") phenomenon has now been found in a wide range of organisms, including humans. Special enzymes **process the double-stranded RNA** into short **single-stranded pieces (21-25 nts long)** and then facilitate their

<sup>&</sup>lt;sup>65</sup> Especially those with flat "aromatic" rings similar to the nucleobases.

<sup>66</sup> The 5'-untranslated region of "5'-UTR."

<sup>&</sup>lt;sup>67</sup> As in the attenuation process described above for the trp operon.

interaction with target mRNAs. In some cases the RNAi molecules even migrate to the nucleus and interfere with transcription by modifying the cognate "chromatin" (protein-packaged DNA – see Section 21).

**RNAi proves to be remarkably specific and efficient**. Only a few molecules per cell can shut down a target gene – presumably because mechanisms exist to amplify the RNAi molecules. This means that it can be used as a powerful **tool to "knock-down" specific genes**. That possibility obviously tempts the pharmaceutical industry as well as basic researchers, and **"RNAi-based therapeutics"** is a currently hot area<sup>68</sup>. **No form of RNAi has been discovered in bacteria**, but in principle an RNAi-based system could be adapted for use in bacterial cells.<sup>69</sup> Synthetic biologists will probably soon unveil applications of this potent and elegant gene-control system.<sup>70</sup>

Another important RNA-based form of gene regulation involves mRNA splicing (see Section 15). Suppose a gene contains n introns and n + 1 exons. Typically n is a relatively small number (1-30, perhaps).<sup>71</sup> Even so, this means that **if some of the exons are skipped** when the mRNA is assembled, a very **large number of alternative protein sequences can be made**. This phenomenon, called alternative splicing, plays a large role in human biology – the current estimate is that **60% of all human protein-coding genes are alternatively spliced**. The decision of whether or not a splicing reaction will occur resembles the control of transcription initiation. Just as promoter strength varies, the **splice site affinity for the splicing machinery differs from site to site**. There are also **splicing repressors** – proteins analogous to gene repressors and **splicing enhancers** (see Section 18 for the transcription-control counterparts).

Table 20.1 (see next page) summarizes the various forms of RNAmediated gene regulation, including some not discussed here.<sup>72</sup> Clearly as synthetic biology moves towards more complex systems – especially in eukaryotes – we will need to take these phenomena into account. Obviously they afford both opportunities to employ RNA control mechanisms, but also pitfalls when an engineered part inadvertently interferes with a key natural process.

<sup>&</sup>lt;sup>68</sup> Clinical trials are underway (as of early 2007).

<sup>&</sup>lt;sup>69</sup> At least one patent application has already been filed (May 2005).

<sup>&</sup>lt;sup>70</sup> A recent report (B.A. Janowski, *et al* (2007) *Nat. Chem. Biol.* DOI: 10.1038/nchembio860). describes another role for dsRNA: *up-regulation* of gene transcription for some important mammalian genes. This amounts to an inverse of RNA!

<sup>&</sup>lt;sup>71</sup> The human titin gene (*cf.* Section 7) has 363 introns!

<sup>&</sup>lt;sup>72</sup> We also need to expect that new discoveries will add to the list!

<b>Tuble 20.1</b> IN MI Inculated regulation of gene expression,	Table 20.1	<b>RNA-mediated</b>	regulation of	gene expression.
--	------------	---------------------	---------------	------------------

Туре	Target	Mechanism	Example
mRNA attenuation	transcription	favored formation of	trp operon in
		a terminator hairpin	E. coli
translational	translation	proteins block the	ribosomal pro-
repression		RBS in mRNA	tein S7 in E. coli
riboswitch	transcription	favored formation of	riboflavin bio-
		a terminator hairpin	synthetic genes
			in many bacteria
riboswitch	translation	block the ribosome	adenine genes in
		binding site	V. vulnificus
RNAi	translation	binds to mRNA;	120 C. Elegans
		sometimes triggers	genes; 250 hu-
		cleavage	man genes
alternative splicing	transcription	splicing activator and	sex-determina-
		repressor proteins	tion in Drosophila
transcriptional	transcription	RNA hairpin binding	TAT protein of
activation		by TAT localizes RNA	HIV
		polymerase at start site	
translational	translation	protein liberates a	tmRNA
activation		stalled ribosome	
transcription	transcription	RNA duplex binds to	progesterone
activation		DNA for 5'-UTR	receptor gene
RNA editing	transcription	deamination or guide-	human intestine
		RNA directed inser-	apolipoprotein
		tion/deletion	
chromatin remo-	transcription	?	?
deling			

# 21. Gene regulation by DNA packaging

In bacteria the DNA resides in a central cluster called a **nucleoid** <sup>73</sup> (as seen in Figure 1.1). This region of the cell has **non-specific DNA-binding proteins** that help compact the genome, and, of course, it contains the enzymes that replicate, repair and recombine the DNA as well as those that synthesize RNA, the nascent RNA molecules themselves, and activator and repressor proteins. In contrast with eukaryotes, however, the bacterial nucleoid has very limited high-order structure – essentially **loops of** *ca.* **100 kb attached to a flexible protein scaffold**. Whether or not RNA or protein molecules participate in regulation related to DNA packaging in bacteria, however, is not presently known.

**Eukaryotic cells have a much larger amount of DNA** to manage than bacteria – and **a larger proportion of their genes will not be expressed** in any given cell. Thus the packing of the DNA is much more extensive and regular. Double-helical molecules get wrapped around small **spools of "histone" proteins** in segments **146 base pairs** long to **form "nucleosomes,"** connected by "linker" segments of about 50-60 base pairs. Because of its appearance in the electron microscope, this arrangement is usually described as "beads-on-a string" (See Figure 21.1). Successive additional levels of wrapping lead to the

**Figure 21.1** Electron micrograph of *Drosophila* (fruit-fly) chromatin showing the "beads-on-a-string nucleosome pattern. Individual nucleosomes are *ca*. 11 nm in diameter.



<sup>&</sup>lt;sup>73</sup> Note that this is *not* a nucleus; there is no surrounding membrane.



**Figure 21.2** The second stage of chromatin compaction – formation of the 30-nm fiber. Histone proteins are colored blue (core) and yellow (linker). This artist's conception does not exactly represent the most probable current model of the structure.

more compact forms seen in the light microscope, including the familiar "metaphase" chromosome structure. Figure 21.2 shows an approximate model for the **second stage of the compaction process** (to form the **"30-nm fiber")** that illustrates how the DNA in such folded structures will be sequestered – and difficult to transcribe.

About half of the mass of a chromosome consists of DNA, the remainder is mostly protein, with some RNA included. The **DNA/protein complex** is generally referred to as **"chromatin."** One of the most important factors in eukaryotic gene regulation consists of the processes that unpack genes to be expressed from the appropriate segment of chromatin. Much progress in understanding these processes has been made in recent years, but for our present purpose this need not be described in detail.<sup>74</sup> Suffice it to say, that **chromatin structure and "remodeling" plays a very great role in how eukaryotic gene expression is regulated**, and ultimately will be a challenge for synthetic biologists to master.

<sup>&</sup>lt;sup>74</sup> Chemical modification of the extended "tails" of histones that project outward from the nucleosome cores plays the major role. Methyl groups and acetyl groups are attached and removed from these tails, leading to greater or lesser degrees of compaction.

# 22. Epigenetic control

Epigenetics refers to **heritable characteristics** transmitted from one cell to its daughters<sup>75</sup> that are *not encoded in the nucleotide sequence* of its DNA *per se*. They may be carried by **modifications of the DNA bases** made after replication has occurred, or they may be the result of the presence of **specific regulatory molecules (usually proteins)** passed from parent to daughter cells.

*E. coli* bacteriophage  $\lambda$  (lambda) provides a well-understood example of epigenesis<sup>76</sup> that has inspired synthetic biologists. Normally bacteriophage infect their hosts, take over the cellular gene expression machinery and – after substantial intracellular multiplication to form new phage particles – lyse (break open) the host cells and disperse into the medium where they may find new bacteria to infect. All phage have this lytic pathway. Phage  $\lambda$ , however, belongs to a subset of bacteriophages that can follow an alternate route. Once inside the cell, these viruses may follow the lytic path, but they also can **splice their DNA** into the bacterial chromosome where it gets replicated along with the bacterial genes. This means that all the descendents of the original infected cell carry the  $\lambda$  genome in a cryptic state (called a lysogen). Maintenance of this state depends upon the expression of a phage gene that encodes the  $\lambda$  repressor. The repressor shuts down the sites for transcription of the genes needed for lytic growth and moreover, *stimulates* transcription of its own gene. Thus cells in this lysogenic state produce more than adequate amounts of the repressor protein to insure continuation of that state after cell division. Daughter cells not only inherit the parental genes via DNA, but also **inherit the lysogenic state** epigenetically via repressor protein molecules.<sup>77</sup>

Multicellular eukaryotic organisms undergo a program of **cellular differentiation** that involves activation or repression of sets of genes in response to specific signals, and in many cases the signals themselves are transitory. Nevertheless, the cellular state that they induce endures for many cell generations, perhaps for the lifetime of the organism. This inheritance of gene expression patterns also takes place epigenetically: modification tags on either

<sup>&</sup>lt;sup>75</sup> Or from a multicellular organism to its offspring.

<sup>&</sup>lt;sup>76</sup> The word "epigenesis" has been used in multiple ways in the past and may be confusing. The descriptions given here fit the current usage.

<sup>&</sup>lt;sup>77</sup> The obvious question of how the lysogenic state gets altered to produce a lytic state and generate more phage particles has been answered by many elegant experiments. Space does not permit describing the details here, though its essential feature is simple: a triggering mechanism unleashes a specific protease that degrades the repressor.

the DNA or the DNA-packaging histone proteins are maintained by enzymes that recognize the modified state of the newly synthesized chromatin and perpetuate it. Methyl groups commonly provide such tags, and when chromosome replication occurs, the daughter DNA and histone molecules from methylated sites end up *hemi*methylated (half-methylated). The hemimethylated state provides a good substrate for "maintenance" methylases that add the tagging methyl groups to the newly synthesized DNA and histones, keeping those genetic neighborhoods epigenetically modified (usually shut down).

Two other epigenetic phenomena occur in higher eukaryotes: imprinting and X-inactivation. **Imprinting** refers to modification one copy of a gene such that it is either activated or inactivated while the opposite is true of the second copy. Either the maternal or the paternal copy can carry the imprint which has to be established during the meiotic cell division that gives rise to the egg or sperm respectively. **X-inactivation** occurs in female mammals that have two Xchromosomes; one of these gets inactivated during development so that the expression levels of X-chromosome genes are comparable to those in males (that have only a single X chromosome). This inactivation occurs randomly, so that females are an epigenetic mosaic of cells with either the maternal or paternal X chromosome inactivated.

A final, important point concerns the fact that somatic cell cloning has demonstrated that epigenetic changes can be reversed, *i.e.*, cells can be taken "back to square one" and commence the developmental process from scratch.<sup>78</sup> Future work in the **synthetic biology of multicellular organisms will need to take epigenetic factors into account.** 

<sup>&</sup>lt;sup>78</sup> The memory B cells of the immune system, however, constitute an exception to this statement since self-tolerance requires that sequences encoding autoimmune antibodies have to be deleted.

### 23. Summary

Molecular biology and biochemistry provide a dynamic picture of life's processes at atomic dimensions. The first steps in moving to this high-resolution view of living organisms were taken with the invention of the light microscope about 350 years ago. Visible-light imaging, however, cannot reveal individual atoms and molecules, even with the most powerful lenses. Thus scientists had to turn to chemistry to develop detailed descriptions of the structures and processes that occur living cells. Elaborate 20<sup>th</sup>-century instruments such as X-ray diffractometers, electron microscopes and atomic-force microscopes have helped to bridge the gap between what we can see with our eyes, aided by optical microscopes, and what we can deduce from chemical principles. Now, at the beginning of the 21<sup>st</sup> century, no scale of biological structure – from Ångstroms to meters – escapes our scrutiny. That means that we can confidently engineer living systems, beginning with an increasingly thorough understanding of the natural prototypes and a complete examination of the systems we build.

Part I of this primer has surveyed the contents and processes of naturally evolved biological systems at the molecular level. Recognizing the important distinction between small molecules and macromolecules leads to an understanding of how complex structures can be built from a relatively small set of low-molecular-weight components (amino acids, nucleotides, sugars and lipid monomers). Intermediary metabolism provides these building blocks as well as energy-rich molecules like ATP that can drive the uphill reactions needed to assemble them into the macromolecules. These in turn perform key biological functions such as catalysis, defense, structure-creation, and information storage. Nucleic acids nearly monopolize this latter task, though epigenetic also participate. Most of the remaining functions can be carried out by proteins, but polysaccharides play a major role in forming structures, and lipids are indispensable for providing the membranes that define cells themselves. In eukaryotes they also create important sub-compartments within the cells.

Genetic information, stored in DNA, enables cells to produce specific proteins (via mRNA synthesis) as well as functional RNA molecules. By controlling the ensemble of these "actor" macromolecules, cells manage their own behavior. Synthetic biology aims to modify or extend such behaviors by introducing new genes (and genetic regulatory elements). Ultimately we can envisage reconstructions of cellular genomes that encompass larger and larger numbers of functional genes, perhaps culminating in a totally synthetic cell. Part II of this primer introduces some of the techniques for such engineering.

# REFERENCES to Part I (7/16/07)

#### A. Molecular biology textbooks

Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts and Peter Walker, *Molecular Biology of the Cell*, 4<sup>th</sup> ed., *ca*. 1500 pp. [Garland Science (Taylor and Francis Group), New York, 2002]

A thorough, detailed introduction to molecular biology on a high level.

James D. Watson, Tania A. Baker, Stephen P. Bell, Alexander Gann, Michael Levine and Richard Losick, *Molecular Biology of the Gene*, 5<sup>th</sup> ed., 732 pp. (Pearson Education/Benjamin Cummings with Cold Spring Harbor Laboratory Press, 2004).

Probably the best short introduction to molecular biology (with a strong biochemical component). Has excellent illustrations and clear, up-to-date explanations. The best book for self-study by beginning learners – and a good place for those who want to review and update prior knowledge.

#### **B. Biochemistry textbooks**

David L. Nelson and Michael M. Cox, *Lehninger Principles of Biochemistry*, 4<sup>th</sup> ed., 1100 pp. (W.H. Freeman, San Francisco, 2004).

The most recent edition of the pioneering 1970 textbook by Albert Lehninger that stressed the role of energetics in biochemical transformations. Good coverage of the detailed chemical principles of enzyme-catalyzed reactions. **N.B. A 5<sup>th</sup> edition of this book will be published by early 2008.** 

Jeremy Berg, John Tymoczko, and Lubert Stryer, *Biochemistry*, 6<sup>th</sup> ed., 1120 pp. (W.H. Freeman, San Francisco, 2007)

A now-classic text, originally authored by Lubert Stryer, this book has enviable clarity of exposition and very aesthetically pleasing, scientifically correct illustrations. Many medical applications of biochemistry are discussed.
## C. Cell biology textbooks

Harvey Lodish, Ronald Berk, Paul Matsudaira, Chris A. Kaiser, Monty Krieger, Matthew P. Scott, Lawrence Zipursky and James Darnell, *Molecular Cell Biology*, 5<sup>th</sup> ed., 973 pp. (W.H. Freeman, San Francisco, 2003).

A well-written, detailed classic text.

Geoffrey M. Cooper and Robert E. Hausman, *The Cell: A Molecular Approach*, 4<sup>th</sup> ed., 765 pp. (Sinauer Associates, Sunderland, MA, 2007).

A short, up-to-date cell biology book. Clearly written and abundantly illustrated (with 720 figures!).

## D. Popular Articles [TO BE ADDED]

Selected from *Scientific American* and similar sources.

## E. Reviews [TO BE ADDED]

Selected from *Trends in Biochemical Sciences*, other *Trends* journals and related sources.

## Acknowledgments

Samantha Burke and Mackenzie Cowell made numerous helpful suggestions about this manuscript. Mac also assisted with computer-related problems and he and Meagan Lizarazo answered many questions pertaining to practical issues in synthetic biology/genetic engineering.