Appendix 3

Protein Basics

Proteins are a class of organic compounds that are the most abundant molecules in living systems. They are hugely important, since numerous hormones, neurotransmitters, and messengers of the immune system are made of protein; ditto for the receptors that respond to those messengers, the enzymes that construct or degrade them,* the scaffolding that shapes a cell, and so on.

A key feature of proteins is their shape, since the shape of a protein determines its function. Proteins that form the scaffolding of a cell have the shape of the different crossbars in scaffolding at construction sites (sort of). A protein hormone will have a distinctive shape that is unique and distinctively different from the shape of a hormone that has different effects.* And a protein receptor must have a shape that is complementary to the shape of the hormone or neurotransmitter that it binds (back to the time-honored cliché of appendix 1, namely that a messenger like a hormone fits into its receptor like a key into a lock).

Some proteins change their shape, typically moving between two conformations. Suppose you have an enzyme (again, a protein) that synthesizes a molecule of sucrose by linking a molecule of glucose to a molecule of fructose. The enzyme must have one conformation that resembles the letter *V*, where one end binds a glucose molecule at a particular angle, the other fructose. The binding of both triggers the enzyme to shift to its other conformation, where the two ends of the *V* move close enough for the glucose and fructose to be linked. The sucrose floats off, and the enzyme flips back to its original conformation.

What determines the shape and function of a protein? Any given protein is made of a string of amino acids. There are about 20 different types of amino acids—including some familiar ones like tryptophan and glutamate. Each protein's string of amino acids is unique—like the string of letters that composes a word. Your typical protein is about 300 amino acids long, and with 20 different amino acid types, there are nearly 10^{400} possible sequences (that's ten followed by four hundred zeros)—more atoms than there are in the universe.* The amino acid sequence of a protein influences the unique shape(s) of that protein. Dogma used to be that amino acid sequence *determines* the shape(s) of that protein, but it turns out that the shape is also subtly altered by things like temperature and acidity—in other words, environmental influences.

And what determines the sequence of amino acids that are strung together to form a particular protein? A particular gene.

DNA AS THE BLUEPRINT FOR CONSTRUCTING PROTEINS

DNA is another class of organic compounds, and just as there are roughly 20 different types of amino acids, there are 4 different "letters" (called nucleotides) that make up DNA. A sequence of 3 nucleotides (called a codon) codes for a single amino acid. If there are 4 different types of nucleotides, and each codon is 3 nucleotides long, there can be a total of 64 different codons (4 possibilities in the first place × 4 in the second × 4 in the third = 64). A few of those 64 are reserved to signal the end of a gene, and after eliminating those "stop codons," there are 61 different codons coding for 20 different amino acids. Therefore, there is "redundancy"—almost all amino acids can be specified by more than one unique codon (an average of about 3, i.e., 61/20). Typically the different codons coding for the same amino acid differ by only a single nucleotide. For example, four different codon sequences code for the amino acid alanine: GCA, GCC, GCG, and GCT (A, C, G, and T are the abbreviations for the four types of nucleotides).* Redundancy will be important for understanding gene evolution.

The entire stretch of nucleotides that codes for a single type of protein is called a gene. The entire collection of DNA is called the genome, coding for all of the tens of thousands of genes in an organism; "sequencing" the genome

means determining the unique sequence of the billions of nucleotides that make up that organism's genome. That stretch of DNA is so long (containing roughly twenty thousand genes in humans) that it has to be broken into separate volumes, called chromosomes.

This produces a spatial problem. The DNA library is found in the center of the cell, in the nucleus. Proteins, however, occur all over the cell, are constructed all over it (just think of proteins in the axon terminals of a spinal neuron in a blue whale, terminals that are light-years away from that neuron's nucleus). How do you get the DNA information out to where the protein is made? There is an intermediary that completes the picture. The unique nucleotide sequence in DNA that codes for a particular gene is copied into a string of similar nucleotide letters in a related compound called RNA. Any given chromosome contains a staggeringly long stretch of DNA, coding for one gene after another; in contrast, this stretch of RNA is only as long as the particular gene. In other words, a more manageable length. That RNA is then shipped to wherever it is supposed to be in the cell, where it then directs which amino acids are strung together in which sequence to form a protein (and there are amino acids floating around in a cell, ready to be grabbed for the protein-construction project). Think of RNA as a photocopy of a single page out of this vast twenty-thousand-page-long DNA encyclopedia. (And multiple copies of the cognate protein can be made from the instructions in a photocopy page of RNA. This sure helps in circumstances where copies of the protein must wind up in each of the thousands of a single neuron's axon terminals.)

This produces what is termed the "central dogma" of life, a concept first framed in the early 1960s by Francis Crick, half of the renowned Watson and Crick, who discovered the "double helix" structure of DNA (with more than a little purloined help from Rosalind Franklin, but that's another story). Crick's central dogma holds that the nucleotide sequence of DNA that composes a gene determines how a unique stretch of RNA is put together . . . which determines how a unique stretch of amino acids are put together . . . which determines the shape(s) of the resulting protein . . . which determines that protein's function. DNA determines RNA determines protein.* And implicit in that central dogma is another critical point: one type of gene specifies one type of protein.

Just for everyone's sanity, I'm going to mostly ignore RNA. For our purposes, what is interesting is what genes, the starting point, have to do with their end products—proteins and their functions.

MUTATIONS AND POLYMORPHISMS

Genes are inherited from your parents (half the genes from each [not entirely true, as covered in the main text]). Suppose that when someone's DNA genome is being copied for inclusion in their egg or sperm, a mistake is made in the copying of one single nucleotide; with billions of nucleotides, that's bound to happen sometimes. As a result, unless corrected, the gene, now with its nucleotide sequence erroneously differing in one spot, is passed on to an offspring. This is a mutation.

In classical genetics there are three types of mutations that can occur. The first is called a point mutation. One single nucleotide is copied incorrectly. Will this change the amino acid sequence of the protein coded for? It depends. Back to redundancy in the DNA code, from a few paragraphs ago. Suppose there is a codon in a gene with the sequence GCT, coding for alanine. But there has been a mutation, yielding GCA instead. No problem—that still codes for alanine. It's an inconsequential, "neutral" mutation. But suppose the mutation instead was GAT. This codes for a completely different amino acid called asparagine. Uh-oh.

In actuality, though, this may not be a big deal, if the new amino acid looks a lot like the one that was lost. Suppose you have a nucleotide sequence coding for the following metaphorical amino acid sequence:

"I/am/now/going/to/do/the/following"

Thanks to a subtle mutation, there is a change of one amino acid, but one without a ton of consequences:

"I/am/now/going/ta/do/the/following"

This would still be comprehensible to most people; the protein would merely be perceived as coming from New York. Translated into protein-ese, the protein has a slightly different shape and does its usual task a bit differently (maybe a little slower or faster). Not the end of the world.

But if the mutation codes for an amino acid that produces a protein with a dramatically different shape, the consequences can be enormous (even fatal).

Back to

"I/am/now/going/to/do/the/following"

What if there is a mutation in a nucleotide helping to code for the first *w*, a mutation with a big consequence?

"I/am/not/going/to/do/the/following"

Trouble.

The next type of classical mutation is called a deletion mutation. In this scenario a copying error is made during the inheritance of a gene. But instead of a nucleotide being miscopied, it is deleted. For example, in a case where the seventh nucleotide is deleted,

"I/am/now/going/to/do/the/following"

becomes

"I/am/now/oingt/od/ot/hef/ollowing"

This can frameshift everything over to generate gibberish, or even a different message (e.g., "For dessert I'd like the mousse" mutating to "For dessert I'd like the mouse").

Deletion mutations can involve the loss of more than a single nucleotide. At an extreme, this can involve the deletion of the entire gene, or even a stretch of genes on a particular chromosome. Definitely not good.

Finally, there are insertion mutations. During copying of the DNA to pass on to the next generation, a nucleotide is inadvertently copied twice, duplicated. Thus:

"I/am/now/going/to/do/the/following"

becomes

"I/am/now/ggoin/gt/od/oth/efollowin"

Gibberish, or perhaps a different message, as in the following case, where an *e* has been inserted near the end of the string of letters: "Mary turned John down for a date because she did not enjoy boweling." In some cases an insertion mutation can involve the insertion of more than a single nucleotide. At an extreme, this can even involve the duplication of an entire gene.

Point, deletion, and insertion mutations are most of what mutations are about.* Deletion and insertion mutations often have major consequences, usually deleterious, but sometimes produce a new, interesting protein.

Back to point mutations. Consider one that results in the substitution of a single amino acid in the protein, one that works a bit differently from the correct amino acid. As noted above, as a result, the protein still does its old job, but maybe does it a bit faster or slower. This could be the grist for evolutionary change—if the new version is disadvantageous, reducing the reproductive success of anyone who carries it, it will be gradually selected against, removing it from a population. If instead the new version is more advantageous, it will gradually replace the old one in a population. Or if the new version works better than the original in some circumstances but worse in others, it may reach equilibrium in the population with the original version, where a certain

percentage of people have the old version, the remainder the new. In this case the particular gene would be described as coming in two different forms or variants, as coming with two different "alleles." Most genes come with multiple alleles. And the result is individual variation in the functioning of genes (this is covered in far more complexity in chapter 8).

Finally, a clarification of the confusion where two sound bites about genetics collide. The first is that, on the average, full (non–identical twin) siblings share 50 percent of their genes.* The other is that we share 98 percent of our genes with chimps. So are we more related to chimps than to our siblings? No. Comparisons between humans and chimps are about *types* of traits—we both have genes coding for traits related to having, for example, eyes, muscle fibers, or dopamine receptors, and both lack genes related to having, for example, gills, antennae, or flower petals. So there's 98 percent overlap at that level of comparison. But comparison between any two humans is about *versions* of those traits—both have a gene that codes for, say, this thing called eye color, but do they share the version that codes for the same particular color? Same for blood type, type of dopamine receptor, and so on. We have 50 percent overlap with siblings at this level of comparison.