

## Notes to Educators

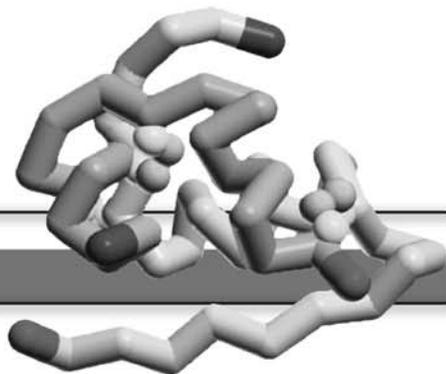
### Amino Acid Properties

Amino acids share a common backbone, through which the amino acids are joined by peptide bonds to build proteins. The amino acid sidechains, however, vary in structure and chemical properties. The properties of these sidechains (non-polar, polar but uncharged, positively charged or negatively charged) impact the final shape and function of the protein. You may wish to have your students explore these properties using the Amino Acid Starter Kit. These concepts can be further reinforced through the Amino Acid Properties activity.

### Numbering Conventions

As your students work through this activity, they might notice that sequences are numbered – but that the numbering strategy changes as they transition from DNA to mRNA to processed mRNA to the protein. It can be a little confusing! But understanding the conventions will make it a little easier.

- DNA sequences are numbered from the 5' phosphate end to the 3' OH end (sugar end) of the sequence. Human chromosomes range in size from 51 to 245 million base pairs, so scientists chop up DNA into smaller pieces to determine the base sequence. Even so, DNA sequences can be hundreds of thousands of base pairs and contain many genes. Typically you'll see DNA sequences starting somewhere in the middle of the sequenced DNA. A gene includes both the coding sequence of the protein, and the regulatory sequences both before ('upstream' or lying on the 5' side of the gene) and after ('downstream' or on the 3' end of the gene). Incidentally, the ends of the DNA sequences from a single chromosome overlap, so scientists can piece them together to get the sequence of an entire chromosome.
- mRNA sequences for each gene are also numbered from the 5' end to the 3' end of the sequence, starting with the number 1, and include both the coding region of the gene and translation regulatory sequences both upstream and downstream of the coding region. These regulatory sequences help the ribosome align correctly on the mRNA. Eukaryotic mRNA initially contains introns that are removed before the sequence leaves the nucleus to be translated by ribosomes found in the cytoplasm. Removal of the introns is called splicing, and a spliced mRNA will have gaps in the numbering, indicating where introns have been removed.



## Notes to Educators (continued)

- Protein sequences are numbered from the N (amino) terminus to the C (carboxy) terminus, beginning with 1. Since all proteins start by reading the AUG codon, all proteins are made with methionine at the beginning, which is numbered 1. But if you looked at a number of protein sequences, you might discover that not all proteins begin with methionine. One reason that not all proteins begin with methionine is that some proteins may contain a signal sequence that directs it to the ER. This signal indicates that the protein is either a membrane protein or a protein to be secreted from the cell; the signal peptide is later removed, or processed, in the ER. As the signal sequence is cut off of the protein, a new amino acid (not methionine) is exposed on the new N-terminal end of the protein.

### **How does the ribosome know where to start? Discovering the Kozak Sequence**

As scientists began exploring the mechanism of protein synthesis, they discovered that all proteins start with methionine, encoded by the single codon AUG. (Sometimes proteins undergo modifications in which the initial methionine is removed after the protein is made. Insulin is an example of this.) In 1986, Marilyn Kozak looked at thousands of nucleotide sequences, lining up the start AUG for each protein, and compared the nucleotide sequences surrounding the start codon. She discovered that there was a strong preference for specific nucleotides surrounding the start codon. This consensus sequence surrounding the start codon is called the Kozak sequence in her honor. Dr. Kozak further demonstrated the importance of the surrounding nucleotides by mutating the region in the human insulin gene and measuring the effects of changing the sequence. When the sequence was closer to the consensus sequence, the protein was made more often; the further from the consensus, the less often the protein was made. We now know that the consensus sequence varies slightly among eukaryotic species, and that some genes that don't match the consensus well are nonetheless able to make proteins. There are additional factors that help to align the ribosome on these mRNA transcripts to allow protein synthesis to occur.