combine microtechnology, liquid chromatographies, and mass spectrometry techniques. The engineering challenges in making a practical device are significant.

8.5.2. Computational Techniques

The wealth of data emerging from these new experimental techniques is overwhelming. Generally, DNA sequence information could be used to identify which sequences are genes that encode for proteins, what these proteins are, what their biological function is, and ultimately how these genes are regulated.

Deciding what nucleotide sequence corresponds to a gene is not always a clear-cut process. Investigators look for "open reading frames" or ORFs. An ORF is a nucleotide sequence without a "stop" signal that encodes some minimal number of amino acids (about 100). In prokaryotes, identifying ORFs is fairly straightforward. In eukaryotes, because of introns and exons, assignment of ORFs is complicated. Some computer programs can recognize probable (consensus) intron/exon boundaries.

When a prospective gene is identified, we need to know the function of the corresponding protein. In some cases the function can be deduced by comparison to databases with amino acid sequence information on known proteins. In some cases the amino acid sequence and function of a protein is conserved across species.

With a highly conserved protein we might find similiar amino acid sequences—for example, in the fruit fly and humans. In this case we would determine how *homologous* the two amino acid sequences were. We then might infer that the human gene encodes a protein with similiar function. Conservation of amino acids near a catalytic or binding site is particularly critical. Efficient computer algorithms to do such searches are under development.

Ideally one would like to predict protein structure and function solely from the amino acid sequence. Heroic efforts have been made to accomplish this goal, but there is no good general solution. Understanding the folding of proteins into their three-dimensional configuration is a computationally difficult problem.

Even with well-studied organisms, such as *E. coli,* we generally can guess the function of only 50% of the genes. Clearly, knowing the full genome sequence has told us how little we really understand. The process of identifying single genes and the function of the corresponding genes has been the focus of *bioinformatics*.

Even if we knew the identity of every gene in a cell and the function of each corresponding protein, we would have an incomplete understanding of function and cellular physiology. A list of the proteins and their functions needs to be supplemented by an understanding of cellular structure and regulation. A combination of proteins can form metabolic pathways, and there will be a corresponding genetic circuit. Relating the cell's "parts list" to its dynamic, physiological state is an unmet challenge that provides exciting, long-term opportunities for bioengineers.

In particular, models of cells and metabolic circuits as discussed in Chapter 6 provide tools with which to organize data and to understand biological function. Simple stoichiometric models of central metabolism have been used with good success to identify which genes are essential to a cell in a particular environment. Models that also incorporate kinetics and regulatory structure are in development. Such models will be key to