Fifty Years of Sequence Analysis: What Have We Learned?

Russell F. Doolittle
Center for Molecular Genetics
Univ. California, San Diego

In 1955, not quite fifty years ago, Sanger's group published the first amino acid sequence of a protein, bovine insulin. Later that year, they reported sequences from sheep and pig, each of which showed a small number of differences from the bovine. Even earlier, sequences of several polypeptide hormones had been worked out, and their similarities and differences had already caught the eye of evolutionists. From this tiny amount of data sprang the hope, considered unrealistic by many at the time, that the histories of all living organisms might be reconstructed. Shortly thereafter, the notion that gene duplications were the source of most proteins was given a boost when the sequences of the alpha and beta chains of hemoglobin were reported to be more than 40 percent identical. By the 1960's, amino acid sequence analysis had taken hold as the major tool for determining the divergences of both creatures and their proteins. The obvious similarities of bacterial and eukaryotic enzyme sequences lent hope that the routes leading to all the major groups of organisms could be established, and the observation that proteins with different functions could have similar sequences showed that radical changes of function were possible. The paradigm that "all new proteins came from old proteins" became the dogma of the day. By the 1970's, the RNA sequencing of certain RNA moieties was almost routine, and the sequencing of numerous small subunit rRNAs revealed unexpectedly that there were three domains of life, now referred to as Archaea, Bacteria and Eukarya. The finding framed one of the major mysteries of all biology: which group came first and how are the others related to it? To this day the question has not been satisfactorily resolved. The advent of DNA sequencing in the late 1970's ushered in a new era, the deluge of data making all previous work seem trivial, even if fundamental questions remained. The campaign in the 1980's to sequence the human genome changed the nature of biological inquiry in a major way. Comparative genomics may indeed reveal the patterns of how the major life forms have evolved. Among recent successes using the genomic approach has been the unraveling and intertwining of proteins involved in photosynthesis and nitrogen fixation and their spread through the prokaryotic world. On

another front, the history of protein folds is being analyzed with surprising clarity. The technological advances that have made all this possible—computers, robotics, photochemistry, and much more—are as mind-numbing as the results. But where are we going? What do we want to know? Although some of the evolutionary questions framed fifty years ago have been answered in part, many remain. What were the primordial proteins and how did they arise? What was the nature of the first cells? Although sequence analysis alone may not tell all, it has given us a great beginning.