

# Foreward

*Joshua Lederberg*

Historically rich in novel, subtle, often controversial ideas, Molecular Biology has lately become heir to a huge legacy of standardized data in the form of polynucleotide and polypeptide sequences. Fred Sanger received two, well deserved Nobel Prizes for his seminal role in developing the basic technology needed for this reduction of core biological information to one linear dimension. With the explosion of recorded information, biochemists for the first time found it necessary to familiarize themselves with databases and the algorithms needed to extract the correlations of records, and in turn have put these to good use in the exploration of phylogenetic relationships, and in the applied tasks of hunting genes and their often valuable products. The formalization of this research challenge in the Human Genome Project has generated a new impetus in datasets to be analyzed and the funds to support that research.

There are, then, good reasons why the management of DNA sequence databases has been the main attractive force to computer science relating to molecular biology. Beyond the pragmatic virtues of access to enormous data, the sequences present few complications of representation; and the knowledge-acquisition task requires hardly more than the enforcement of agreed standards of deposit of sequence information in centralized, network-linked archives.

The cell's interpretation of sequences is embedded in a far more intricate context than string-matching. It must be conceded that the rules of base-complementarity in the canonical DNA double-helix, and the matching of codons

to the amino acid sequence of the protein, are far more digital in their flavor than anyone could have fantasized 50 years ago (at the dawn of both molecular biology and modern computer science.) There is far more intricate knowledge to be acquired, and the representations will be more problematic, when we contemplate the pathways by which a nucleotide change can perturb the shape of organic development or the song of a bird.

The current volume is an effort to bridge just that range of exploration, from nucleotide to abstract concept, in contemporary AI/MB research. That bridge must also join computer scientists with laboratory biochemists—my afterword outlines some of the hazards of taking biologists's last word as the settled truth, and therefore the imperative of mutual understanding about how imputed knowledge will be used. A variety of target problems, and perhaps a hand-crafted representation for each, is embraced in the roster. There is obvious detriment to premature standardization; but it is daunting to see the difficulties of merging the hardwon insights, the cumulative world knowledge, that comes from each of these efforts. The symposium had also included some discussion of AI for bibliographic retrieval, an interface we must learn how to cultivate if we are ever to access where most of that knowledge is now deposited, namely the published literature. Those papers were, however, unavailable for the printed publication.

It ends up being easy to sympathize with the majority of MB computer scientists who have concentrated on the published sequence data. Many are even willing to rely on neural-network approaches that ignore, may even defeat, insights into causal relationships. But it will not be too long before the complete sequences of a variety of organisms, eventually the human too, will be in our hands; and then we will have to face up to making real sense of them in the context of a broader frame of biological facts and theory. This book will be recalled as a pivotal beginning of that enterprise as an issue for collective focus and mutual inspiration.