Richard V. Eck (1922–2006)

Bioinformatics: In the beginning

History is often set by those who write it, and this is clearly evident in the many books recounting the “heroic” period of the biological sciences, the late fifties and early sixties, marking the genesis of the fields of molecular biology and bioinformatics. The visibility by numbers of contributions from the major players, regardless of their subsequent correctness, can often render out of focus the sometimes more accurate and carefully constructed efforts of less prolific researchers. One should beware of the oft given advice “standing on the shoulders of giants”; these may not be so firm.

The 50’s–60’s was a time of great excitement in biology; the prize was to crack the genetic code, or to solve the “protein cryptogram,” as Richard Eck labeled it in a 1962 paper in the *Journal of Theoretical Biology*. This theoretical question attracted people from a variety of scientific backgrounds. Back in 1954, George Gamoff, a cosmologist and hero to this writer as a young student, had proposed that the combination of the four bases in DNA, A, G, T, and C, taken three at a time, would produce a total of 64 sets, sufficient to code for the 20 amino acids commonly occurring in proteins. Two at a time would only produce 16 sets, and four at a time, unnecessarily too much information to be an economical code.

Arguments ranged back and forth in the theoretical realm with little experimental guidance because of the dearth of accurate protein sequence information at that time, and of course, no nucleic acid sequences. Gamow’s simple hypothesis, labeled by some less generously as “numerology,” was that it would be economical for the code to be overlapping, but the biologist Sidney Brenner argued against this on statistical grounds. There was uncertainty whether the bases taken three at a time should be overlapping or not. The nonoverlapping code, now known to be correct, was preferred by Crick (1958), a physicist, and then was supported by some chemical evidence from Morowicz’ lab and by some other biological considerations. The nonoverlapping code became the general belief at that time. That a biological property could be explained at the molecular level from chemistry, led to the later designation of this practice as “Molecular Biology.”

Richard Eck, with his training as a mathematician, argued in a *Nature* paper (1961) that the popular belief that the code was nonoverlapping involved a number of unstated assumptions, which were not clearly justified. One of these was that there were no external factors acting in addition to the base triplets, determining the identity of the amino acid that they coded for. He likened the question to a jigsaw puzzle, where given a sufficient number of pieces, the code was overlapping, and the control of the protein sequences was nonrandom, it should be possible to make progress in solving the “protein cryptogram.” However, the question first to be addressed was whether protein sequences occurred at random with no external constraints, the consequence being that there would not be a possibility of solving the jigsaw puzzle from published protein sequences alone. Eck examined the case of a number of alleles, that is, proteins having the same function from different species, such as insulin, containing a different amino acid residue at one locus. At the time of that writing, there were 74 alleles with published sequences, and Eck showed by statistical analysis that of the 20 amino acids, and these 74 cases, a random substitution could only yield 22 duplications. Therefore, the pairs are not a random sample; the same pairs are found repeatedly, as would be expected from an overlapping code “in the context of adjacent links.” Of course, almost 50 years later we now know that although the DNA codes by nonoverlapping triplets, the possibility of contributions from “external factors” is evident from the discovery of alternative RNA splicing, which leads in effect to an overlapping code that can be extracted from the one DNA sequence.

Eck demonstrated the concept of the “jigsaw puzzle” idea by using a punched card sorting machine, where a randomly constructed protein sequence was divided into a number of theoretical “hydrolysis products.” Eck showed that the original sequence could be completely determined provided the number of fragments obtained was more than twice the number of links in the protein. The requirement for this number of fragments became much less if the information was available as to the amino acid composition of the complete protein as well as for each fragment. Several years later, Dayhoff and Eck (1970)

extended this theoretical argument to show that, from a determination of the exact mass of each fragment by mass spectroscopy, the primary protein sequence could be reconstructed. In other words, this was the method of “shotgun sequencing” to be rediscovered and applied to nucleic acid sequencing more than 20 years later in connection with the Human Genome project.

In 1961, Nirenberg and Matthaei showed that a ribosome preparation would produce polyphenylalanine in the presence of a synthetic nucleic acid polyU. This was the first identification of a code, the triplet UUU, as corresponded to the amino acid phenylalanine. In a short time, numbers of similar experiments allowed about 45 of the 64 possible triplets to be assigned to amino acids, some amino acids belonging to more than a single triplet. The order within each triplet had to be guessed at, and it was not until several years later that the defined oligonucleotide syntheses carried out by Nirenberg’s group proved the correct order of the bases in each triplet that we know today. In a 1963 paper in *Science*, Eck pointed out that the 64 possible triplets divided into 32 pairs, such that in many of these pairs, one member differed from the other only by the change of a pyrimidine base for the other pyrimidine, that is C for U, or of the purine bases, A for G. Using this rule he predicted the assignment of the remaining triplets, and these turned out to be mostly correct. We now know, however, that three of the triplets code for protein chain termination instead of an amino acid. A further implication of these earlier experiments was the discovery of the mechanism of transcription of the DNA signal into messenger RNA, which subsequently bound to the ribosome for the synthesis of the protein. As a consequence of his coding rules, Eck theorized about a structural basis for code recognition by the adapter molecules (transfer RNAs) postulated by Crick (1958), one RNA for each amino acid, brought each in turn to the messenger as needed for protein chain extension. Each transfer RNA contained the “codon” postulated by Crick, corresponding to that amino acid, so the complementarity of hydrogen bonding as recognized in the

structure of the DNA double helix must also be the case for the recognition process between messenger and adapter RNAs.

In 1966, Eck and Dayhoff analyzed the amino acid sequence of ferredoxin and demonstrated the nonrandomness of sequence repeats, evidence this protein must have evolved its function by doubling shorter sequences, in support of the developing idea of gene duplication in evolution. The simple idea of Eck and Dayhoff was to place the two halves of the ferredoxin sequence under one another with the insertion of a gap:

AYKIADSCVSCGACASECPVNAIS QGDSI 29
30 FVIDADTCIDCGNCANVCPVGAPVQE 55

The procedures now employed with great sophistication in sequence comparisons among proteins and nucleic acids, all derive from this initial demonstration. An incidental footnote in this seminal work was the very convenient and well-known one-letter abbreviations for the amino acids, now an essential tool in the field of Bioinformatics. Russell Doolittle (University of California, San Diego) brought to my attention that a one-letter code was first suggested by Keil in 1963. This same year, 1966, also marked the birth of what was really the first database in the field of Bioinformatics, the “Atlas of Protein Sequence and Structure” by Eck and Dayhoff. In this publication the methodology of protein sequence comparison and the proposal for notations were outlined in detail.

Richard Eck graduated from the University of Maryland in 1943. Beginning in 1955, he was employed as a statistician at the National Cancer Institute, and then from 1960 at the National Biomedical Research Foundation. In 1969, he relocated to the University of Georgia.

A fitting epitaph for Dick Eck would be:

Arg-Ile-Cys-His-Ala-Arg-Asp-Val-Glu-Cys-Lys, 1922–1986.

JOHN LEE

*Department of Biochemistry and Molecular Biology,
University of Georgia, Athens, Georgia 30602, USA*