BIOORGANIC CHEMISTRY 23, 355–368 (1995)

REVIEW

The Evolution of Functional Proteins[1]

BONNIE L. BERTOLAET*,[2] AND DAVID P. HEITMEYER†

*Cancer Center, Departments of Pharmacology and Medicine, University of California, San Diego, La Jolla, California 92093-0684; and †Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138

Received July 24, 1995

How did enzyme catalysts evolve? First, a single catalytic group of rudimentary effectiveness could have been incorporated into a single short peptide. In the second stage, several peptides would bind together, providing a multichain assembly with improved catalytic effectiveness. These peptides would eventually be joined together in a single chain to increase thermal stability and ensure the linked inheritance of all the elements of the complex. Finally, this rough protein would be refined and improved by classical selection processes. Evidence for the second step comes from investigation of the cDNA sequence and the genomic sequence of a gene from *Tetrahymena* and forms the basis of the exon microgene theory (*1*). This hypothesis suggests that in the early stages of the evolution of functional proteins RNA consisted of exon microgenes, each of which terminated in an amber codon (UAG) and encoded independently translated peptides. These peptides would form catalytic, multichain assemblies that were the rudimentary precursors of the enzymes we know today. In support of these ideas, complementation studies have shown that a protein "refragmented" at its exon–exon boundaries produces a functional multichain protein complex. These studies have been expanded to a more general investigation of the resilience of protein catalytic function in the face of insults to protein structural integrity.

Two approaches have been taken to understand the formidable efficiency of enzymes as catalysts. One approach has been to study the physical and chemical origins of the large rate enhancements mediated by enzymes and to probe the relationship between protein structure and catalytic function. The second approach has been to focus on the question of how enzyme function evolved and how it has been refined to the present level of selectivity and efficiency. These two approaches are related, for if the development of catalytic power can be traced, then the nature of enzyme catalysis as it is found today will be better understood.

The evolution of functional proteins can be considered to have occurred in three phases. The first stage involved the incorporation of catalytically functional groups into peptides. In the second, these peptides were assembled together to create a catalytically active species—a rudimentary enzyme—in which at least some of the necessary functional groups are present in the "active site." Finally, the last stage involved enzyme optimization: the refinement of the position and orientation of

[1] Supported by the National Institutes of Health, GM37007.

[2] To whom correspondence should be addressed at the Scripps Research Institute, 10666 N. Torrey Pines Rd., MB7, La Jolla, CA 92037.

0045-2068/95 $12.00

each of the catalytic groups, the improvement of the binding of the transition state for the reaction to be catalyzed, and the enhancement of the substrate selectivity to the point that yields the efficient catalyst known today. This discussion focuses on the second phase of this proposed development of functional proteins, in which individual peptides are recruited to form a primordial catalytic assembly.

The central dogma of molecular biology in the 1950s asserted that DNA is transcribed into RNA which is then translated into protein (*2*). While this simple statement provides the basic outline of the modern biosynthesis of proteins in prokaryotes, the analogous process is somewhat more complex in eukaryotes. In 1977, it was discovered that eukaryotic genes are not continuous but are interrupted by noncoding sequences known as introns (*3, 4*). Introns are the segments of DNA found between the coding regions, known as exons, and it is just the exons that are ultimately expressed into protein (*5*). After the genomic DNA is transcribed into RNA, the RNA segments complementary to the introns are excised, resulting in a continuous message that is translated into protein.

It is not yet understood why eukaryotic genes are interrupted by introns, nor is it understood what determines the position of these intervening sequences. There are many theories regarding the origin and function (or absence of function) of introns. Although there are many variations, each with its own subtleties, there are basically two views regarding the origin of introns. One view holds that introns are ancient, predating the divergence of eukaryotes and prokaryotes, and that ancestral genes contained interrupting sequences. The other view attests that introns have "invaded" eukaryotic genes that were at one time continuous, relatively recently in evolutionary time. It seems clear that some introns (such as the self-splicing group I and II introns) are mobile elements; however, it is not clear when this mobility was acquired, nor how the current role of introns relates to their original function (*6*). An alternative hypothesis has been proposed on the ancient origin of exons and introns specifically in genes encoding functional proteins, the exon microgene theory (*1*). This theory suggests that exon–intron and intron–exon boundaries were originally determined early in evolution by terminating amber codons (TAG) of exons that each encoded independently translated peptides which assembled to form catalytically active complexes.

As with other arguments used to support the notion that introns are evolutionarily ancient, this theory relies on the supposition that "the faint outline of ancient patterns" (*7*) can give insight into the evolution of functional proteins (*1*). The theory derives from several observations made during the cloning of the gene that encodes phosphoenolpyruvate mutase from *Tetrahymena* (*8*). This enzyme catalyzes the interconversion of phosphoenolpyruvate (PEP) and phosphonopyruvate, and is thought to be responsible for the formation of carbon–phosphorous bonds in nearly all naturally occurring phosphonates. Although organisms that metabolize phosphonates are rare, they appear across the evolutionary spectrum, and the biosynthesis of phosphonates is therefore thought to be an ancient metabolic process (*9*).

Surprisingly, the cDNA clone for phosphoenolpyruvate mutase contains two in-frame amber condons (TAG). Most organisms such as eubacteria and eukaryotes utilize amber (TAG) codons as one of three "stop" signals in protein synthesis:

TAG (amber), TAA (ochre), and TGA (opal). However, these signals are not universal and examples of deviations are known in mitochondria (in which opal codons are read as tryptophan) and ciliated protozoa such as *Tetrahymena* (in which amber and ochre codons are read as glutamine; *10, 11*).

Upon examination of the *genomic* sequence for PEP mutase, it was discovered that two of the three introns in the gene were precisely located after the two in-frame amber codons. Moreover, the two introns that followed the exons with amber termini also ended with TAG. All three sets of exon–intron and intron–exon junctions conform to the conserved consensus splice junction sequences found in the protein-encoding genes of all eukaryotes (*12*). There has been little speculation regarding the origin of these consensus sequences (*13*), but it should be noted that these sequences are distinct from the junction sequences used in the splicing mechanisms for nonprotein encoding genes, such as those for rRNA and tRNA (*14*). Because the splicing mechanisms for the rRNA and tRNA genes were available early in evolution (i.e., in the "RNA world"), it seems unlikely that the existence of consensus splice site sequences unique to protein-encoding genes is an historical accident. Moreover, self-splicing introns have been found predominantly in rRNA and tRNA genes and in organellar genes, but largely not in nuclear genes that encode functional proteins. Most of the evidence supporting the view that introns are modern is derived from studies on self-splicing introns such as group I and II introns. Based on the differences in splicing site recognition sequences and splicing mechanisms for self-splicing introns and nuclear introns in protein-encoding genes, it thus seems unlikely that all introns originated from a common ancestor.

Given that the probability[3] of two introns occurring precisely after the two in-frame TAG codons is approximately 1 in 10^5, and given that the exon–intron and intron–exon boundaries conform to the consensus splicing sequences for eukaryotic protein-encoding genes, it seems likely that the location of the amber codons at these boundaries is a reflection of an early functional role for these sequences. Based on all of these observations, the exon microgene hypothesis was generated in which the location of the introns in the *Tetrahymena* mutase gene was used to explain the origin of the consensus 3′ terminal sequence of eukaryotic exons and introns (*1*). It was suggested that exons were once "microgenes" that ended with an amber codon and encoded relatively small, independently translated peptides.

The hypothesis runs as follows. From various segments of RNA early in evolution, initiation and termination of protein synthesis (the latter signaled by amber codons) would produce a library of peptides, some of which would spontaneously combine to form multichain assemblies having low catalytic activity

[3] This calculation is based on the following: There are two in-frame amber codons within the PEP mutase gene-coding region of 900 nucleotides. Thus, there are two positions out of 900 nucleotides in the PEP mutase gene that occur immediately after an in-frame amber codon. The probability of randomly placing two of the three introns immediately after the amber codon is given as: $(2/900) \times (1/899) \times 3$ (the probability of "choosing" immediately after one of the amber codons times the probability of choosing after the second amber codon times the number of different ways that three introns can be placed at two specific positions). Thus, the chance that both in-frame amber codons are followed immediately by an intron is 1/134,850, or 1 in 134,850.

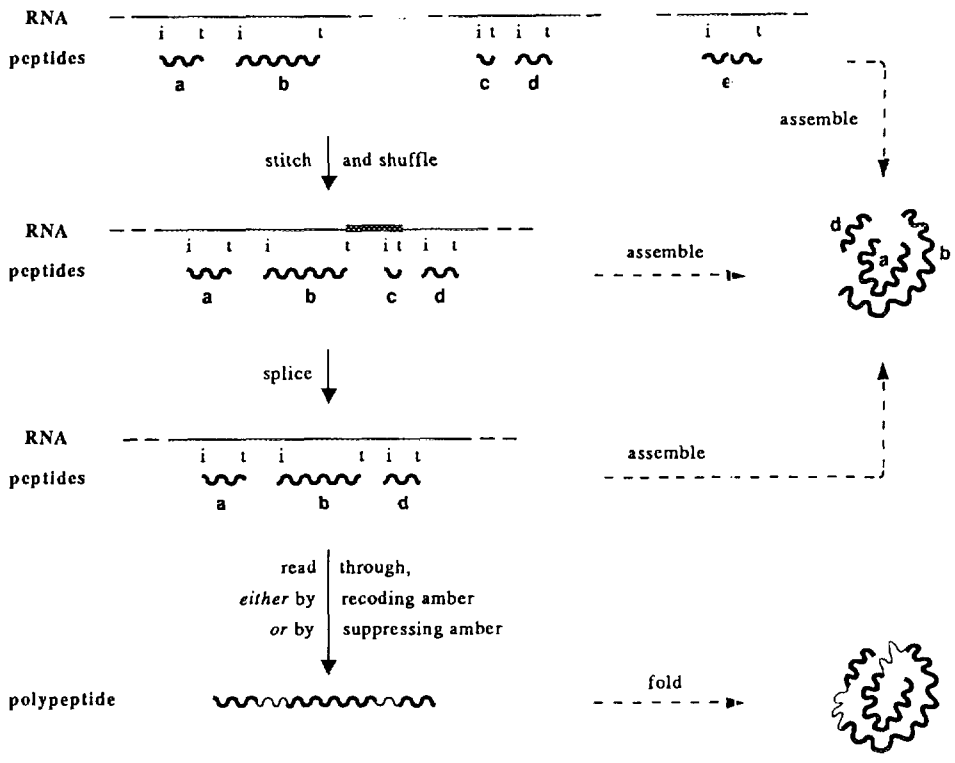

FIG. 1. Putative development of functional proteins. From a variety of segments of RNA, initiation of protein synthesis (at i) and termination of protein synthesis (at t) produces a library of peptides (a, b, c, etc.) from which catalytic activity is generated by spontaneous peptide assembly. Use of the termination amber codons (t) as a recognition element for stitching useful RNA segments together and for shuffling these microgenes [by mechanisms analogous to those of present-day RNA splicing, which is illustrated by the excision of the intron (heavy bar)] allows the rough juxtaposition and (therefore) the linked inheritance of the microgenes that encode the components of a particular functional assembly. Read-through of the terminating amber codons by one of the two mechanisms indicated (recoding amber, or suppressing amber) then produces a continuous polypeptide that has greater thermal stability than the segmental assembly of smaller peptides. This figure was adapted from Seidel *et al.* (*1*).

(Fig. 1). These would be the rudimentary precursors of enzymes. Intron-mediated shuffling of these exon microgenes using their common amber 3′ termini as recognition elements would bring the appropriate microgenes together, thereby increasing the chances of linked inheritance of the individual microgenes from generation to generation. That is, the chances of retaining all the microgenes encoding peptide fragments that contribute to a particular catalytic activity would be increased if they were close together on the same stretch of RNA. Excision of the introns, by means presumably analogous to modern splicing mechanisms, would lead to a spliced message that still encoded individually translated peptides. In some cases, these excised introns can be thought of as "failed" exons in the sense that they encoded peptides that did not add to the overall stability or

function of a particular catalytic complex.[4] Finally, translational "read-through" of the termination codon for exons that were nearby and in-frame would produce a continuous polypeptide, increasing the thermal stability of the resulting functional protein. This single-chain protomer would then be the substrate for catalytic refinement and improvement over evolutionary time.

Read-through can occur by two mechanisms. First, the amber or stop codon can mutate, so that it encodes an amino acid. Alternatively, the meaning of the stop codon can be suppressed, so that the amber codon now encodes an amino acid. This is the case with the amber codons in *Tetrahymena* (*16*). According to this hypothesis, read-through of the originally terminating amber codons would produce "extra" peptide sequences encoded by the intervening RNA between the amber codon of one exon and the initiation codon of the downstream exon. These extra peptide loops would be least disruptive to the overall structure if they were located at the protein surface, where, indeed, exon junctions are predominantly found (*17*). The surface location of exon boundaries solves the problem of how to connect the termini of exon products that are not near to one another in the initial peptide assemblies. Finally, a surface location, and exposure to the polar, aqueous solvent, would best accommodate the charged amino- and carboxyl-termini of the original exon products in the rudimentary multichain assemblies. The association of splice junction sites with former amino- and carboxyl-termini has been previously proposed (*18*). This suggestion, along with the observation that these junctions map to the surface, is consistent with the existence of an early catalytic complex composed of exon-encoded peptides. Moreover, the circular permutation of phosphoribosylanthranilate isomerase (*19*), in which the amino- and carboxyl-termini are moved to different locations within the protein while maintaining its functional activity, is consistent with the notion that "primitive" proteins contained several potential termini locations, other than those known today. Those termini would have resulted from the original complexation of independently translated peptide fragments, each with its own set of amino- and carboxyl-termini at the later exon boundaries.

If the juxtaposition of a pair of microgenes in the shuffling process were to result in neighboring exons being out-of-frame, then read-through would be selected against, and the intervening sequence would have to be excised later to result in a functional, single-chain protein. This process, of course, survives today. The spliceosome would have evolved after the constraint to recognize AG had been established, originally with AG in frame, but later less constrained, due to junctional sliding, allowing for the distribution of intron phases observed today (*20*). The migration of the splice sites evidently did not result in the complete randomization of intron phases but retained the greatest frequency for phase 0 introns (i.e., those falling between codons), although to varying extents depending on the species (*20*).

The theory that exons were microgenes suggests that primitive proteins were assembled by exon product complementation. According to this view, the exon products need not fold independently: rudimentary enzymes could equally well be

[4] The notion that the amber codon, TAG, is the origin for the 3′ splice site of introns as well as for exons is supported by the analysis of exon–intron and intron–exon boundary sequences of human introns (*15*). This study has found that these splice sites are similar, thereby suggesting that both derive from a common ancestor (i.e., TAG).

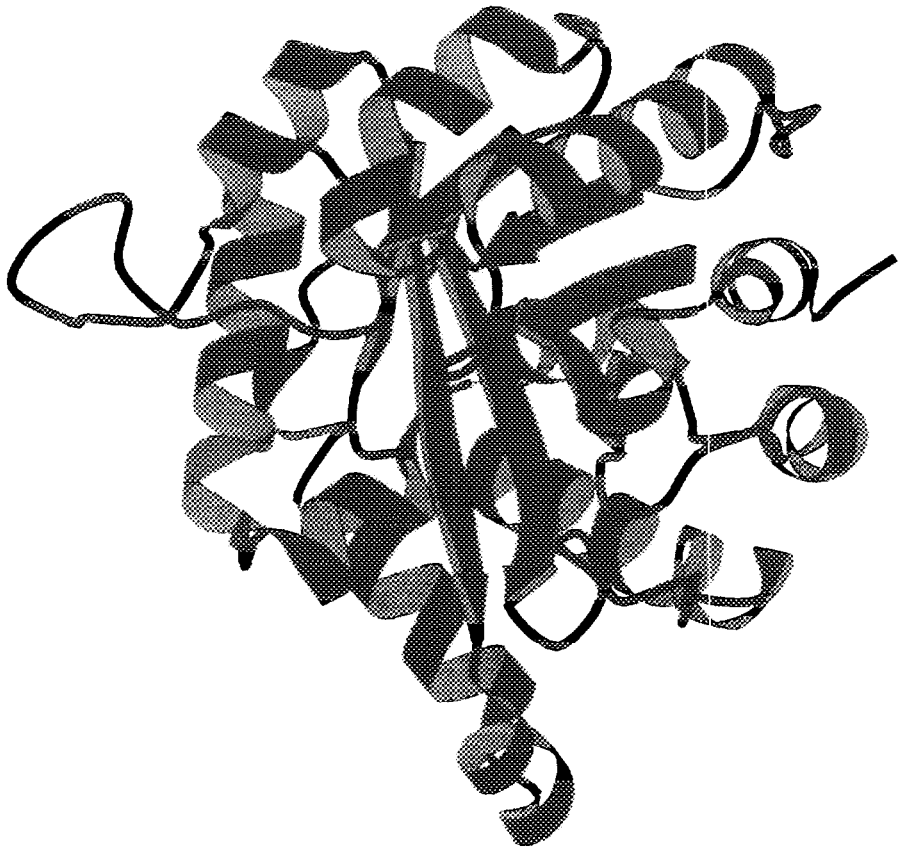

FIG. 2. Side view of a monomer of chicken triosephosphate isomerase. Surface locations of exon–exon junctions are highlighted in red.

assembled by the stable complementation of exon products that alone have no defined three-dimensional structure. That such complementation is possible is well documented, and there are many examples that illustrate the not on that multichain complexes formed from fragmented proteins can exhibit catalytic activity. Fragments have been produced from proteins by chemical (*21, 22*) and proteolytic cleavage (*23–26*), and at the DNA level by means of polycistronic messages that encode separate segments of the protein of interest (*27–29*). While some of these studies have focused on producing intact protein domains, others have yielded peptide fragments that alone have no distinct folded identity.

As mentioned earlier, exon junctions map consistently to the protein surface and

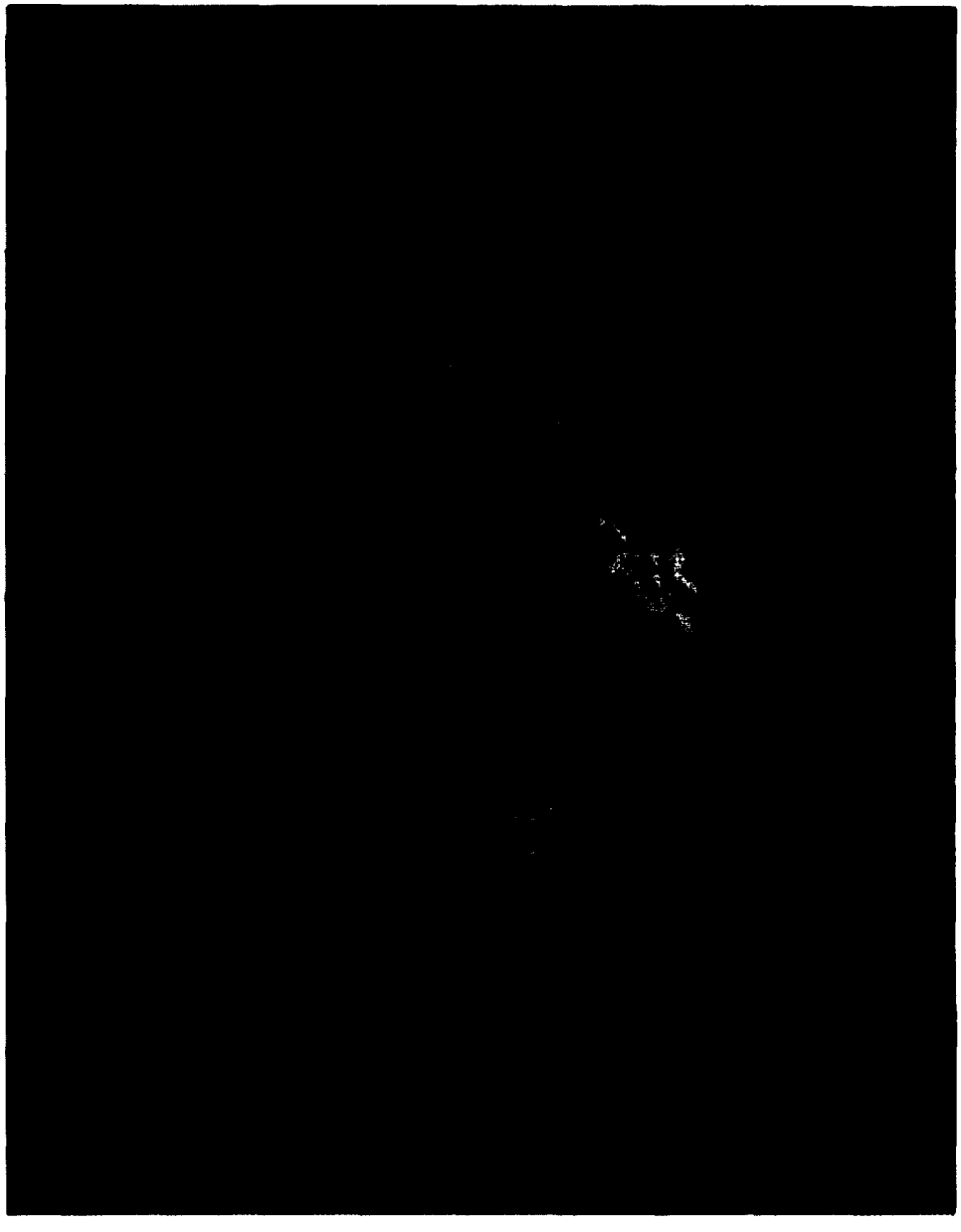

FIG. 7. "Allowed" positions of insertion of a dipeptide into chloramphenicol acetyltransferase (CAT). Trimer of CAT is shown, with allowed or "tolerated" positions in green, positions present in the nonselected library in red, and chloramphenicol in yellow. Coordinates (entry 3CLA, version of July 9, 1990) for chloramphenicol acetyltransferase (*37*) were obtained from the Protein Data Bank (*38, 39*).