Exon coded polypeptides as primordial enzymes

A dissertation
submitted by

Yulia
Ivanova

In partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

in

Chemistry

Tufts
UNIVERSITY
School of Arts
and Sciences

February, 2012

Thesis Advisor: Professor Krishna Kumar

List of Figures

List of Tables

List of Schemes

Abstract

Proteins are diverse and very complex molecules capable of delivering enormous rate accelerations over uncatalized reactions. Questions regarding origin and evolution of protein activity only recently started to receive experimental answers. Two theories have been put forward that propose possible mechanisms for the origin of protein activity: "The Origin of Genes" by Gilbert and "Exons as Microgenes" by Knowles. Both theories indicate the crucial role of exon shuffling in generation of primordial genes coding for first proteins. However, Knowles suggests an additional statement that each exon-coded polypeptide was independently translated and primordial enzymes were non-covalent assemblies of these protein fragments. In our study we aimed at providing experimental evidence to support "Exons as Microgenes theory". Two fusion enzymes were generated where flexible linker was incorporated at the position defined by exon/exon boundaries of human adenylate kinase 1 (AK1). The length of this linker was such that allowed for sufficient spatial separation of protein fragments thus preserving their identities as individual protein fragments. The crucial importance of the exon/exon boundary for introduction of linker was demonstrated on the example of control fusion proteins, where the linker was moved away from the exon/exon boundary by a few amino acids. Enzymatically competent reassemblies were only observed in the case of fusion proteins with linkers situated at exon/exon boundaries. This correlation was demonstrated to be relevant both *in vitro* and *in* vivo. Relative to control wild type enzyme, specific activities were only moderately lower, 2 and 9 times for fusion proteins containing one linker and 2 linkers (at two different exon/exon boundaries) respectively. Designed protein constructs

were structurally characterized via circular dichroism (CD) analysis. All protein constructs retained folding patterns similar to that of control full-length enzyme, though their folding was found to be diminished. Binding constants of both substrates were found to be higher than those of wild type enzyme, which could account for observed lower activity of fusion proteins. The broader utility of the linker mediated exon-coded polypeptide reassembly approach is yet to be shown on other proteins.

Chapter 1

INTRODUCTION

1. Protein evolution

1.1 Evolutionary puzzle: protein diversity

Proteins are diverse in structure and activity. Our knowledge of a number of different protein sequences present in living organisms has improved dramatically over the past couple decades. The sequencing of the human genome in 2001 revealed the presence of little over 3 billion base pairs within it, and an estimated 25000-30000 genes that code for proteins. Additional level of protein diversity comes from the fact that many eukaryotic proteins undergo post-translational modifications, such as phosphorylation, glycosylation, disulfide formation, side chain oxidation, N- and C-terminal truncations[1]. As the amount of sequencing information grows exponentially, it is becoming clear that proteins found in living organisms represent only a small fraction of the 20^n amino acid combinations that are theoretically possible (n being the number of amino acids). The question is why is so much of potential amino acid space unoccupied? One answer to the question is considering physical, chemical and functional constraints that must be accommodated in order to produce an active enzyme. Only limited number of amino acid combinations can give rise to stable secondary structures, such as α-helices and β-sheets. The necessity of generating a hydrophobic core where catalysis can take place and the precise positioning of catalytically active residues resulting in enormous acceleration of enzymatic rates over uncatalyzed reactions places additional limitation on the nature of amino acids used in specific locations[2]. Such explanations interpret the limited occupancy of protein space as a result of an underlying fitness of a protein sequence, implying that the sequences that are not found are those that simply do not work due to either physical, chemical or biological reasons. These explanations however assume that all of the sequences have been tested and only functional ones retained. However, sequences evolve from other sequences. It is practically impossible for nature to go though 20^{200} (assuming length of a protein to be 200 amino acids on average) different possibilities for each protein before it resides to choosing the best one. Additionally, considering that total number of atoms in visible universe was estimated to be 10^{80} there is not enough carbon on Earth to go though all $20^{200}=10^{260}$ sequence possibilities for each protein.

The ways in which nature searches though active protein space always leaves a footprint on the modern time proteins sequences. "Exon shuffling" is one of such identified footprints[3]. The discovery of introns within coding sequences of eukaryotic genes quickly led to assumption that exon exchange or "exon shuffling" into new combinations could be the mechanism for evolution of new protein structures and functions. Examples of exon shuffling are continuously being uncovered[4] (Figure 1).

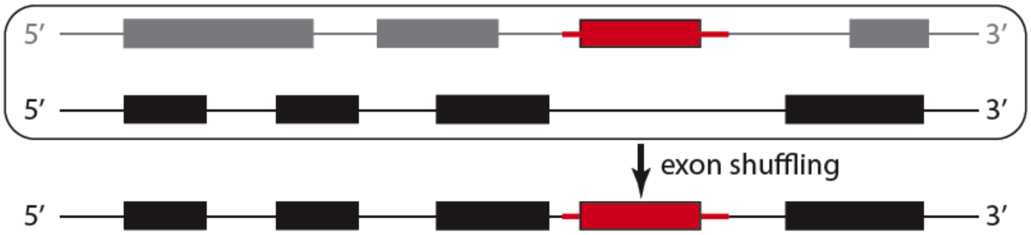


Figure 1. Schematic representation of exon shuffling. The exon highlighted in red has been transferred from the parent gene to the recipient gene. The intron sequence bordering shuffled exon (colored in red) is used to allow for genetic transfer and is carried over in part into the sequence of recipient gene.

The protease family involved in blood clotting cascade, the low density lipoprotein receptor family, the serine/threonine kinases and the broad immunoglobulin and immunoglobulin-like protein superfamily all appear to have evolved at least in part due to utilization of exon shuffling[4]. An exact exon sequence coding for a 40 amino acids long cysteine rich sequence found originally in low-density lipoprotein (LDL) was also identified in the sequence of unrelated human EGF precursor as well as in the protease factor IX[5].

The expanding number of mosaic proteins suggests that the structural and functional diversity of proteins has been achieved by the combinatorial rearrangement of a limited number of basic motifs i.e. exons. Knowing a frequency of exon reuse, the size of the universe of different exons can be estimated. A global study on the determination of exon reuse frequency among different genes was undertaken in 2003[6]. Two databases were generated using sequences submitted to NCBI. One contained 56276 internal exons from putatively unrelated genes (determined as the ones with less than 20% sequence homology) and the second contained 8917 internal exons from regions of these genes that are homologous with prokaryotic genes. It was demonstrated that with 95% significance level, 3500 exon-sequences matches in the large database and 500 matches in the second ancient conserved regions database. Since the size-rank relationship for these databases follows a power law with size falling off as the inverse square root of the rank, the size of the universe of exons was estimated at 15000-30000. The size of each exon having a peak at 105-120 bp (35-40 amino acids) was estimated for *Homo sapiens* in a separate study[7]. Even though the estimates on the order of 30000 may appear to be rather large, they are nothing in comparison to the above-mentioned 20^{200} possibilities nature had to go through if it were to construct a 200 amino acid protein by a single amino acid search. If, however, it were to construct a protein by a linear search of 10^5 exons, it would only need to try 10^6 different possibilities. This allows for faster exploration of a set of potential proteins and explains rapid evolutionary appearance of new forms, such as the 'explosion' of the Cambrian era.

Several attempts have been made to recreate the evolutionary path taken by nature. A critical step in constructing an *in vitro* protein evolution system is the generation of a large pool of random DNA sequences containing long open reading frames (ORFs). Random sequences made from polymerization of four nucleotides have been successfully used as library sources for new RNA and DNA molecules, however, this approach suffers from high frequency of termination codon generation during polymerization, which would lead to "corrupt" protein sequences[8,9]. One way to avoid this obstacle is to construct biased DNA sequences rather than random ones. Using short stretches of DNA (microgenes) instead of using nucleotides as building blocks leads to minimal formation of in frame termination codons. Such an approach would relate to the previously discussed point that new sequences evolve from already existing ones, as opposed to having to go through all the random possibilities from scratch. A library containing long open reading frames was created in 1997 by Noda group[10]. A single microgene was used as a template for generation of entire library. A microgene was polymerized in a head-to-tail manner, with nucleotide insertion or deletion randomly occurring at the end-joining positions (Figure 2).

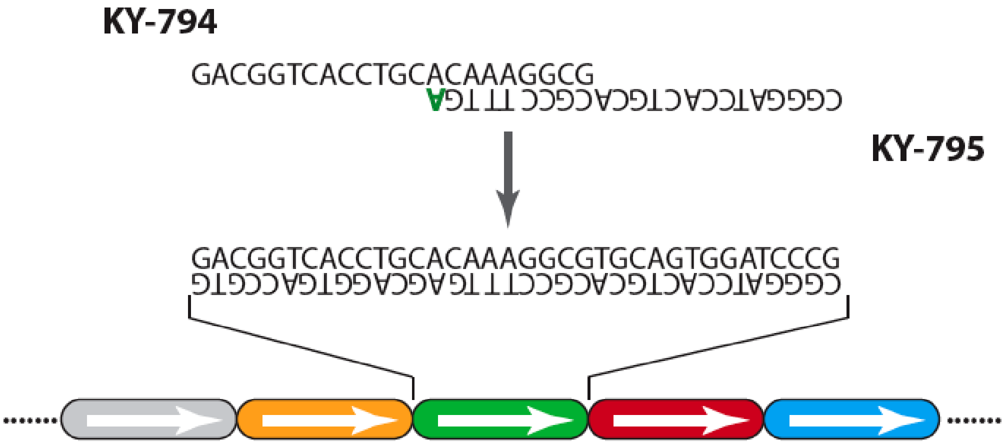


Figure 2. Primer set used to accomplish head-to-tail polymerization of the microgene target. The diversity in the microgene portions of the final open reading frame is derived due to the presence of a mismatched nucleotide at 3'-OH end (highlighted in green) in the sequence of primers.

This changes the reading frame of the microgene at the juncture, thus producing new coding sequence. The resulting microgene polymers are combinatorial libraries of 2 × 3 reading frames of the microgene. When a starting microgene does not contain termination codons within its six frames, the resultant polymer would also not contain any stop codons, allowing for formation of large open reading frames. A 120 amino acid protein library made from a 36 bp microgene could have $2 \times 3^{10} = 1.2 \times 10^{5}$ items contributing to its molecular diversity. The choice of starting microgene will influence the nature of the resulting library. The chances of obtaining folded soluble proteins increases if a microgene coding for a short unit of secondary structure is used as a microgene template. When a β-strand forming sequence was used as a microgene template, six out of ten randomly selected polymers produced a large number of proteins and three out of four proteins investigated further were shown to be soluble, indicating presence of ordered structures.

Another way for generation of new genes is random fragmentation of a pool of related genes, followed by reassembly of the fragments in a self-priming polymerase reaction. Four 1.6 kilobase genes coding for class C cephalosporinases, (58-82% identical at the DNA sequence level) were chosen from four different bacterial species[11]. The four genes were shuffled either individually or as a pool (Figure 3).

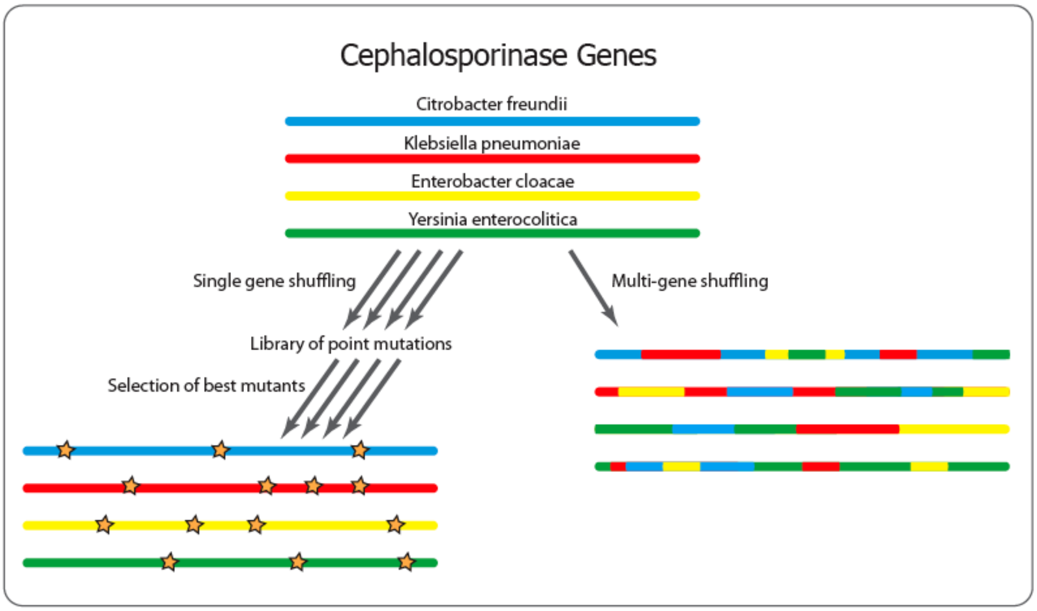


Figure 3. Comparative overview of two methods used to evolve new proteins: Single gene shuffling and multi-gene shuffling. In both cases genes are generated in a single assembly reaction from 60-mer synthetic oligonucleotides, however in case of single gene shuffling, these 60-mer synthetic oligonucleotides cover the sequence space of only one gene, while in multi gene shuffling, a mixture of 60-mer synthetic oligonucleotides is used thus covering the sequence space of all four originally used genes within each assembly reaction.

Equal numbers of *E. coli* transformants were plated on selective media and colonies showing the highest resistance were identified and further characterized. Clones originating from four single gene libraries showed up to an eightfold increase in selective antibiotic resistance compared to those expressing wild type protein. The best clone originating from the gene family library showed a 540-fold increase in resistance compared to the wild type protein. The two most resistant clones were sequenced and analysis of their DNA sequence revealed that both clones were chimeras of the three out of four originally used proteins containing a few single point mutations. This observation demonstrates that the highest evolutionary acceleration rate is observed in case when domains of parent proteins are exchanged in a random manner.

Several attempts at creating a novel enzyme by domains fusion from different proteins have been made, such as the fusion of dihydrofolate reductase to the N-terminus of thymidylate synthase to mimic a naturally occurring bifunctional enzyme found in protozoa and plants; fusion of a carboxy-terminal truncated aminocyclopropane-1-carboxylate (ACC) synthase to amino-terminal truncated ACC oxidase to create an active bifunctional enzyme capable of cleaving S-adenosylmethionine to directly produce ethylene[12]. Other examples include bifunctional enzyme aspartokinase-homoserine dehydrogenase I from *E. coli* that catalyzes non-consecutive reactions in aspartate biosynthesis pathway was split into two domains[13]. Each of these domains, kinase and dehydrogenase, were fused to a non-native domain of same function. The aspartokinase I domain was fused to the enzyme aspartate semialdehyde dehydrogenase, while dehydrogenase I was fused to aspartokinase III from *E.coli*. In both cases fusion led to formation of properly folded catalytically active bifunctional enzymes with kinetic parameters comparable to that of the native enzyme.

These experiments demonstrated the general ability of exon-shuffling to generate new genes in the laboratory setting on a faster time scale than regular evolution.

1.2 Primordial enzymes and evolutionary age of introns

As demonstrated above, “exon shuffling” is an efficient method of creating novel functional genes from pre-existing sub-units: exons. In addition to providing a plausible framework for recent molecular evolution the idea also suggests an avenue for the creation of the primordial genes for the first proteins. Proposed by Walter Gilbert this idea has come to be known the “Origin of genes” theory[14]. It suggests that the primordial genes were made of small pieces: exons. These small exons encoded short polypeptide chains ~15-20 amino acids long and the basic evolutionary method to generate new genes is to shuffle the exons. Intron sequences were used to facilitate the shuffling of exons, thus placing the evolutionary origin of introns prior to the divergence of prokaryotes and eukaryotes. The major trend of evolution was to lose introns and to fuse small exons to form longer modern day exons.

The role of introns in the history of genes has been a subject for a debate between two extreme positions. One side postulates that introns were used to assemble the first genes and is called introns early or The Origin of Genes. The other side states that introns were added during the evolutionary break between prokaryotes and eukaryotes to split previously continuous genes[15]. The later approach is known as introns late. Evidence supporting both theories are present but are interpreted differently by two warring sides[15]. For example, the absence of introns in prokaryotes is one of the arguments used by introns-late supporters – that the insertion of introns took place after prokaryotes and eukaryotes diverged. However, the loss of introns in prokaryotes can be explained by increased genome pressure in prokaryotes requiring faster transcription rates. It was also demonstrated that the protist *Trichomonas vaginalis* has a splicing apparatus as demonstrated by organism's ability to splice out short introns engineered into a reporter gene sequence[16]. It was also shown that an ORF coding for putative poly(A) polymerase (PAP1) is interrupted by 94nt long sequence flanked with canonical splicing site sequences on 5' and 3' ends. Additionally, it has been experimentally demonstrated that introns can be lost in mammals. Alignment of more then 17000 genes from human, mouse, rat and dog to perform a genome-wide analysis of intron loss and gain indicated no intron gain and 122 incidences of intron loss, most of which occurred in the rodent lineage[17]. The intron losses occurred mostly in the highly expressed, housekeeping genes indicating that intron loss takes place via germline recombination of genomic DNA with intronless cDNA, a process known as retroposition. These results imply that intron gain has not played a role in protein evolution during last 95 million years and most likely dates back to more ancient history of protein evolution. On the other hand, when different set of proteins was analyzed (alpha and beta tubulins as well as actins) from different eukaryotic organisms, the observed phylogenetic organization of 33 out of 35 introns was better explained by a single intron gain event[18]. An alternative explanation requires five or more independent intron losses. Gilbert has shown that in triosephosphate isomerase (TIM) a majority of introns were found in at least two eukaryotic kingdoms, indicating that at least some of the introns were older, and dated back to at least the plant-animal kingdom split. Analysis of the same sequence data set by Dibb and Newman offered an alternative explanation. Analysis of intron flanking regions indicated presence of a conserved sequence on both sides of intron, thus allowing them to argue that there is a sequence bias towards where the intron would be inserted. This consensus sequence was termed a "proto-splice site"[19]. Exon shuffling is governed by splice frame rules[4]. Introns may interrupt the open reading frame of a gene between two consecutive codons (phase zero introns), between the first and second nucleotide of a codon (phase one introns), or between the second and the third nucleotide of a codon (phase 2 introns). Based on the phase of the flanking intron exons can also be classified into different groups. These include symmetric exons that are interrupted by an intron of the same phase on its 3’ and 5’ end: 0-0, 1-1 and 2-2, and asymmetric exons that are interrupted by introns with different phases on 3’ and 5’ ends. Asymmetric exons are named 0-1, 0-2, 0-3, 1-2, 2-0, and 2-1 (Figure 4).

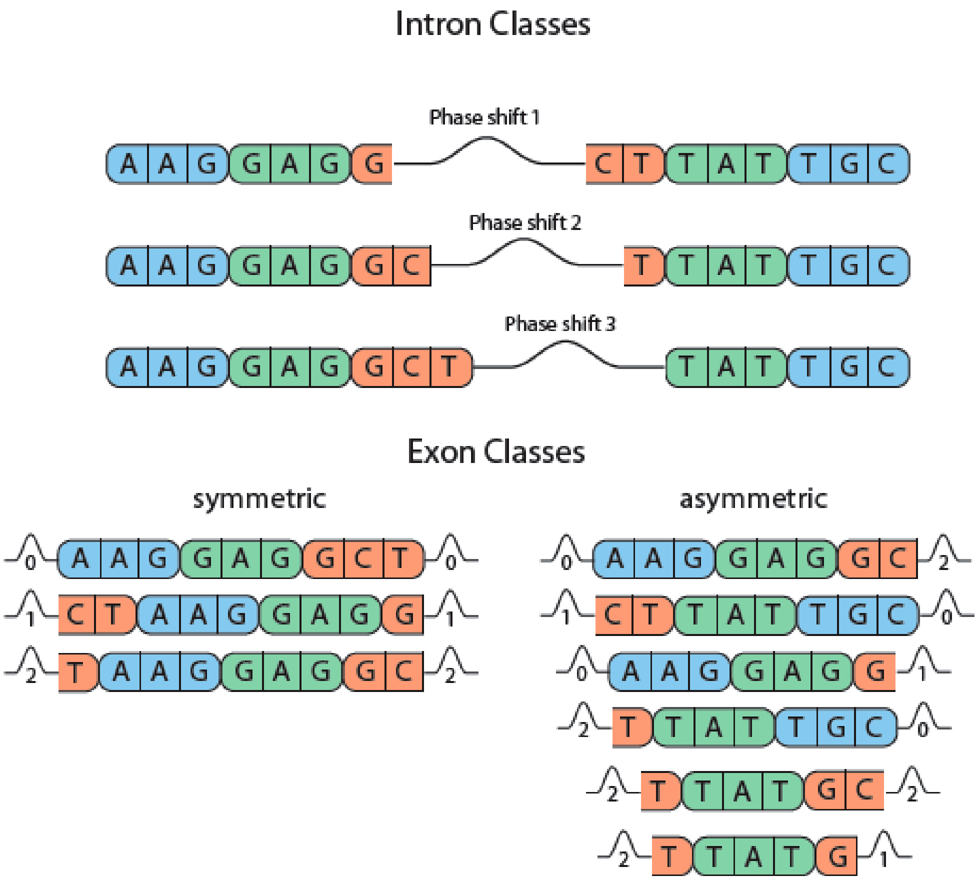


Figure 4. Various classes of exons and introns phases. Phase 1: insertion takes place after first codon nucleotide, Phase 2: insertion takes place after second codon nucleotide, Phase 3/0: insertion takes place after third codon nucleotide. Depending on the phase of the intron bordering the exon in question, exons are classified in symmetric and asymmetric, depending on whether or not the intron phases on both sides of the exon are in the same phase.

The symmetric exons are the only exons or sets of exons that can be shuffled and inserted into introns of the same phase without disturbing the open reading frame. If introns were to be inserted into the previously continuous gene sequences one would expect to observe equal statistical distribution of phase 1, phase 2 and phase 3 or 0 introns, resulting in 33% of each present. However, analysis of the genomic sequences of ancient proteins (having high sequence homology to prokaryotic analog) indicates bias towards phase 0 introns with 56% of introns being present in phase 0, 24% in phase 1 and

21% in phase 3[14]. Most of the mosaic protein exon-coded domains are represented by symmetric exons or symmetric sets of asymmetric exons. While it is clear that number of modern time introns is rather large and have been added to the extant gene sequences there is also a number of introns that are ancient that were present prior to the divergence of prokaryotes and eukaryotes as demonstrated by their conserved positions. Thus, the consensus is slowly being reached at the point of a synthetic theory of intron evolution where some introns are agreed to be modern and thus added to the sequence of a previously un-split gene, while others are regarded as ancient and pre-dating the prokaryotic/eukaryotic split[20]. This synthetic theory is supported by recent observation of significant correlation between symmetric units of shuffling and the age of protein domains[21]. Ancient domains, defined as domains shared between prokaryotes and eukaryotes, are more frequently bounded by phase 0 introns and their distribution is biased towards the central part of protein. Modern domains, defined as those found exclusively in eukaryotes, are more frequently bounded by phase 1 introns and are present predominantly at the C- or N- terminus of a protein. The ExInt database was created for this study consisting of 134,442 sequences from GenBank with annotated introns positions and phases indexed with respect to amino acid sequences. "Domain database" (Pfam) generated was used. In a dataset of 537,987 introns, a significant statistical bias of 49 % was observed towards phase 0 introns, followed by 28% phase 1 and 23% phase 2 introns. A significant excess of symmetrical exons was also observed, 10% excess for 0-0 exons, 23% excess for phase 1-1 and 12% excess for phase 2-2 exons. The excess of phase 0-0 exons was predominantly located in the central portion of a protein, while 1-1 and 2-2 exons were mostly situated at C- or N-terminus. Thus, it is the fraction of modern day introns that are considered ancient that were potentially involved in generation of primordial genes via exon shuffling.

Though previously described as DNA-based introns and exons theory, The Origin of Genes also nicely flows out of the RNA world view. In this picture, RNA genetic material creating by splicing RNA enzymes do all the biochemistry, introduce activated amino acids, one by one to build up short polypeptides to support ribozyme function and finally utilize all 20 amino acids, short exons, and mRNA splicing creating protein catalysts[14].

1.3 Exon-intron boundaries correlation to domain distribution

Additional supporting evidence for the role of exon shuffling mechanism in creation of primordial genes comes from high degree of correlation between the intron/exon boundaries on the genomic DNA level (or exon/exon boundaries at the protein level) to domain distribution along ancient protein sequences. As previously discussed only a portion of extant modern day present introns could have been involved in generation of new genes as exon shuffling has a limitation on the phase of introns allowed to be shuffled next to other exons without disturbing open reading frame sequence, thus, in order to be relevant the subset of ancient proteins has to be analyzed for such correlation[4]. However, the idea of establishing correlation between domain distribution and exon/intron boundaries does not only intrigue researches investigating the role of exon shuffling in primordial gene formation. Establishing such correlations would be helpful in answering protein folding related questions as well as general questions about enzyme catalysis. There have been a few approaches at establishing such correlations in different, though not necessarily ancient proteins. A short discussion of one of the methods developed to investigate presence of such correlation is described below.

Inspection of molecular models derived from X-ray studies performed in 1973 by Wetlaufer showed that in all but the smallest globular proteins, the polypeptide chain folds into several globular units, which may sometimes be loosely connected[22]. These units are commonly called “domains” or “structural domains” to indicate that they are derived from three-dimensional protein structures. Domains often carry out specific functions and are re-used in the context of various proteins as discussed earlier. Domains are units of protein evolution as well as protein structure and multidomain proteins performing complex biochemical reactions have been generated via fusion of genes coding for individual domains in a process known as domain shuffling. The location of structural domains by visual inspection of models suffers from subjectivity, thus objective definitions and algorithms to detect structural domains within a protein sequence are required. An algorithm based on the surface area calculations was worked out by Wodak *et al.* in 1981[23]. This approach defined structural domains as regions of protein structure where most of the interactions between atoms or residues occur within the regions and least without. This algorithm was tested on glycogen phosphorylase, thermolysin and several other proteins. The sizes of domains formed during this analysis nicely correlated with size and position of proteolytically generated domains. No correlation of domain position to the position of exon/intron boundaries was made using this method.

Mitiko Gö proposed an alternative method for identification of domains within proteins[24]. She defined a domain as a continuous region of a polypeptide chain with all α-carbons being less than a defined distance apart. This defined distance was set at 27 Å. The distances between i^{th} and j^{th} α-carbon atoms of hemoglobin were plotted in a triangular space in which both *y*- and *x*- axes are the residue sequence number. The grey areas represent the pairs of C^{α} atoms within 9 Å, the white areas represent pairs whose distances are 9-27 Å, and black regions are the pairs separated by more then 27 Å (Figure 5).

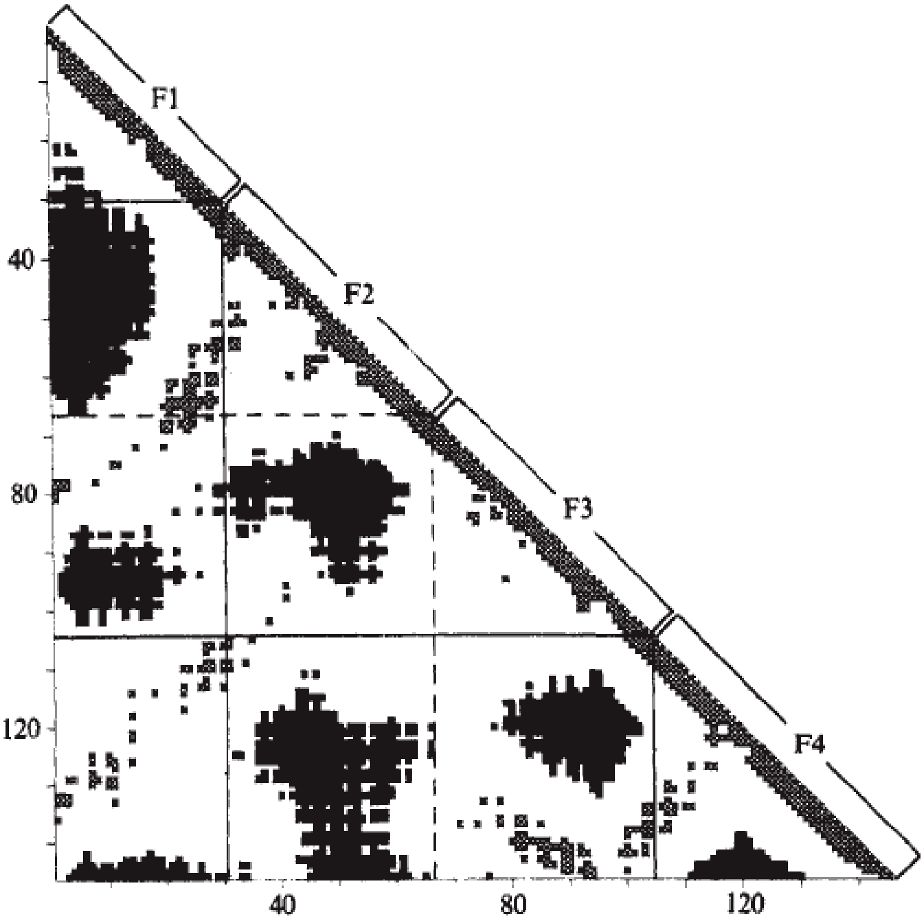


Figure 5. Distance map or Gö plot of hemoglobin β-chain[24]. Along x- and y- axis are the amino acid numbers. Structural units of hemoglobin are shown on diagonal axis. Solid black crossing lines are positions of existing hemoglobin exon/exon boundaries, black dashed line indicates the position of predicted fourth exon/exon boundary later discovered in related leg hemoglobin. Reproduced from reference 24 with permission. © Nature 1981

Latter regions are grouped into seven major continuous areas. This method for identification of protein domains is very illustrative and simple to execute. It has since gained popularity in the community and was used to demonstrate the domain distribution along the sequences of various proteins. The most remarkable discovery made after careful inspection of the various patterns of the distance maps or "Gö plots" is the correlation between the position of the domain boundary and the position of the exon/exon boundary in a protein sequence. Even more remarkable, Gö was able to predict the position of a new exon/exon boundary in hemoglobin that was later found to exist in a related leghemoglobin. Such an astonishing discovery led researchers to further investigate the potential importance of such correlation. If found to be spread throughout the proteome, such correlation would further support the notion that domain shuffling may have been accomplished by random combination of preexisting exonic regions of a gene.

In 1996, Walter Gilbert did a thorough evaluation of existing correlation between intron positions and domain boundaries on a subset of 32 ancient proteins that have intronless prokaryotic homologs[25]. With a module size set at 28 Å it was determined that 216 out of 570 introns lie in the 28 Å linker-region, almost 34 more than expected based on a random basis, where linker regions are defined as overlaps between the 28 Å modules (Figure 6A).

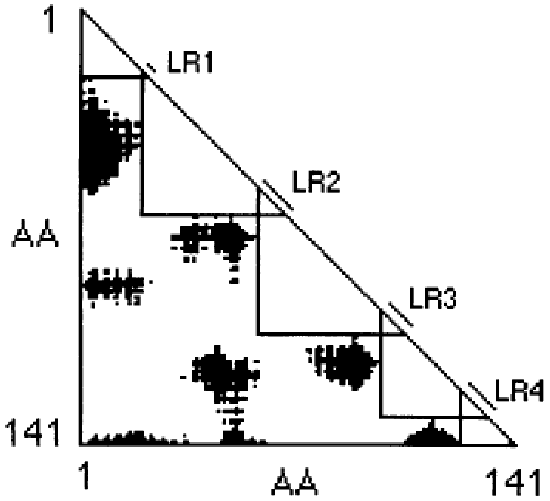


Figure 6A. Gö plot of hemoglobin β-chain with indicated linker regions[25]. Five modules are identified by large triangles along the diagonal. The size of the modules is limited by adjoining black regions, that are population of amino acids with *i*th and *j*th carbon separated by more than 27 Å. The linker regions are indicated as LR and are defined as the region of overlap of module triangles. Reproduced from reference 25 with permission. © PNAS 1996.

This correlation becomes even more pronounced as one reduces the number of tested proteins down to 20 “old” proteins, where “old” describes proteins whose origination dates prior to the phylogenetic divergence of eukaryotes from prokaryotes. In the latter case, 14 out of 20 introns were located within the linker region. The INTER-MODULE software used to predict linker regions allowed for variation in a module size. When the module diameter was allowed to vary between 6 and 50 Å, three major peaks for intron positions were identified. These peaks correlated with module sizes of 21 Å, 28 Å and 33 Å (Figure 6B and 6C).

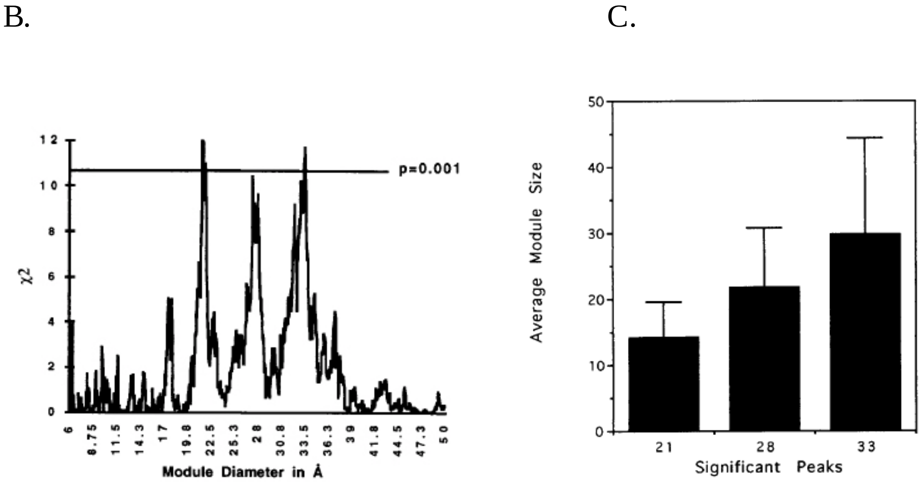


Figure 6B. Excess intron positions in linker regions as a function of module diameter[25]. Three statistically relevant linker region sizes dominate the distribution of introns. **C.** Lengths of predicted module sizes in comparison to their converted amino acid lengths. Reproduced from reference 25 with permission. © PNAS 1996

From these module sizes the average internal exon length can be calculated, assuming “exons” to be defined as lying between the midpoints of the linker regions, for each of these peaks. The lengths of exon products are approximately 15, 22 and 30 amino acids. These numbers correlate well with the statistical analysis of data on the *H. sapiens* exon length distribution[7] and supports the hypothesis that short exons were used to assemble the ancient conserved genes and served as modules in protein evolution.

1.4 Other approaches to generation of new proteins

The variation in gene number among organisms indicates that there is a general process for new gene origination. Exon shuffling is considered to be the most powerful method for generation of protein diversity within short time frame. There are however several other different molecular mechanisms that are known to be involved in the creation of new genes[3]. The details of these mechanisms are understood to varying degrees. For example, gene duplication is a classical mode of new gene generation and was considered to be the leading cause for protein diversity prior to the discovery of exon-shuffling. During this process, a gene is duplicated; the new copy is free to evolve new function, while original gene retains its function. Retroposition is another mechanism in which the generation of new gene is achieved via duplication of a parent gene. During retroposition, the duplication of a parent gene is done through reverse transcription of expressed parental mRNA. Because a retroposed gene copy usually does not retropose a promoter copy of the parent gene, this mechanism requires the new gene to develop its own promoter regulatory sequence in order to be functional. Mobile elements, such as ALU elements that have the ability to self-insert into coding portions of previously existing genes, were found be a standard phenomenon in formation of new genes in the human genome. In prokaryotic organisms, lateral gene transfer, which denotes the horizontal gene transfer from one organism to another, has been found to be frequent. In some cases, lateral gene transfer can lead to generation of a new gene in the recipient organism; for example, converting bacteria into pathogens. Gene fusion and fission are two other mechanisms that allow for new gene formation via either fusion of two adjacent genes or fission of a single gene into two. In 2000, Thompson *et al.*

identified a human fusion gene, KUA-UEV, in which the ubiquitin E2 domain of tumor susceptibility gene (UEV) and a newly identified gene known as KUA were fused together by read-through transcription, followed by alternative spicing of their coding sequences[26] (Figure 7).

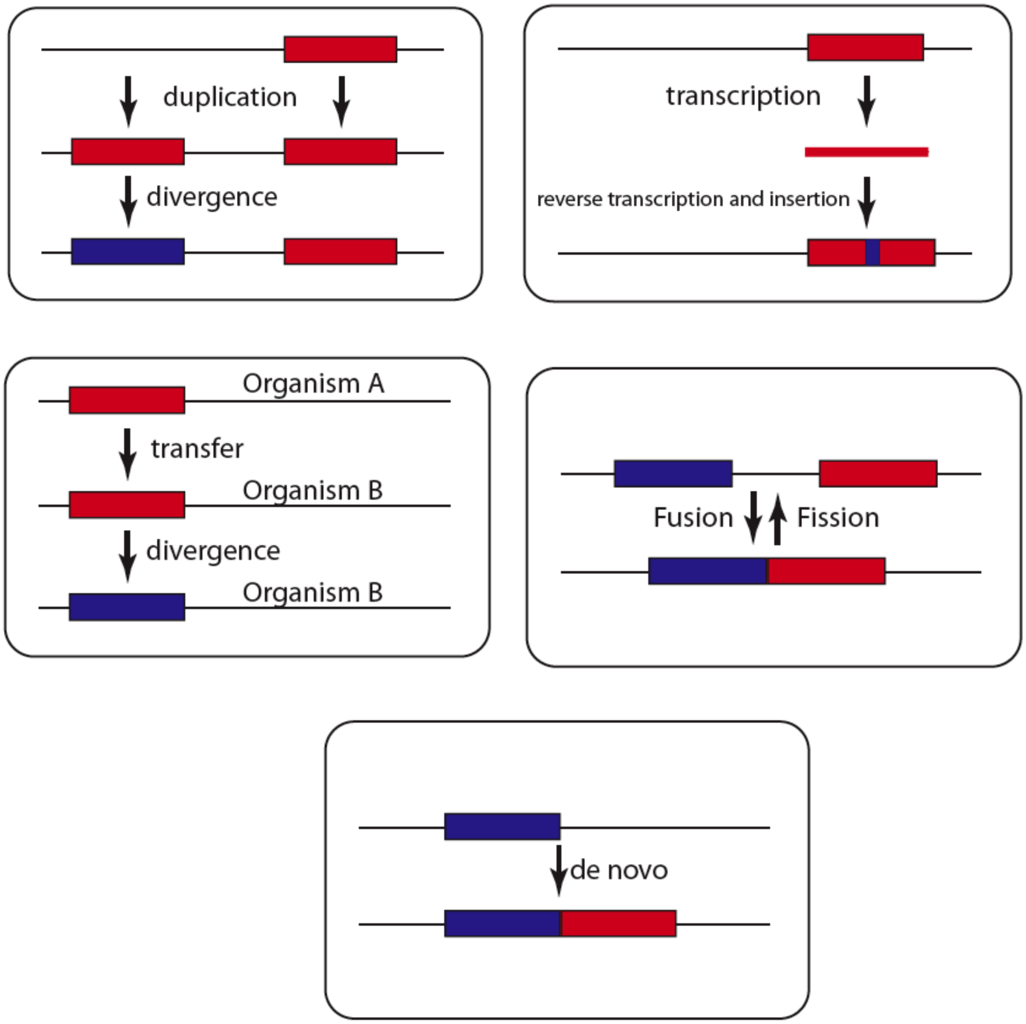


Figure 7. Mechanisms of creation of new gene structures. Each box illustrates a different mechanism leading to the formation of new gene. Clockwise from top left, the mechanisms include: gene duplication, transposition, lateral gene transfer, gene fusion/fission and de novo origination of new coding sequences from previously non-coding sequences.

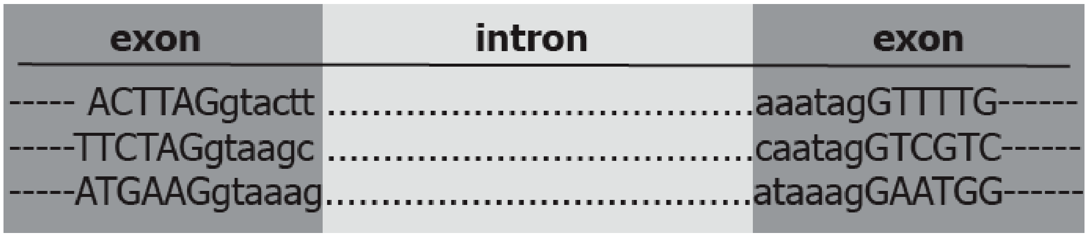
There are many interesting questions that could be asked about mechanisms behind the origin of a new gene. Among these are the following: Once the new gene originates in an individual, how does it spread throughout an entire population of species and become fixed? How often do new genes originate? Though intriguing, these questions are subjects of multiple reviews and are out of the scope of this dissertation.

2. Primordial genes and enzyme

2.1 “Exon as microgenes”

The key role of exon shuffling in generation of primordial genes coding for proteins is well accepted. The fact that exons were re-used multiple in the context of different proteins is well documented and the correlation between the exon/intron boundaries and domains, possessing certain function or activity that is utilized during formation of another protein, is established by example of ancient proteins[11,27,28]. Intron sequences of the primordial genes were used in order to shuffle around functional exons and the footprint of statistical bias towards symmetrical exons and phase 0 introns that lead to no change in open reading frames is easily detectable from sequence analysis of ancient proteins. The removal of introns from the coding sequence of genes using spliceosomal machinery would lead to formation of continuous gene sequences coding for functional proteins. However, during sequence analysis of the gene coding for phosphoenolpyruvate mutase from *Tetrahymena*, a new observation suggesting an alternative route to active protein assembly was made[29].

Phosphoenolpyruvate mutase from *Tetrahymena* is an enzyme that catalyzes the interconversion of phosphoenolpyruvate (PEP) and phosphonopyruvate. This enzyme is thought to be responsible for the formation of carbon-phosphorus bonds in nearly all naturally occurring phosphonates. Although organisms that metabolize phosphonates are rare, they appear across the evolutionary spectrum. The biosynthesis of phosphonates is therefore thought to be an ancient metabolic process. Surprisingly, the cDNA clone for phosphoenolpyruvate mutase contains two “in-frame” amber codons (TAG) that are utilized by the majority of eubacteria and eukaryotic organisms as one of the three stop codons. Upon examination of the genomic sequence of 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, the amber codon[29] (Figure 8).



exon	intron	exon
----- ACTTAGgtactt	....	aaatagGTTTTG------
-----TTCTAGgtaagc	....	caatagGTCGTC------
-----ATGAAGgtaaag	....	ataaagGAATGG------

Figure 8. Exon/intron boundaries in the DNA sequence of phosphoenolpyruvate mutase from *Tetrahymena*.

Given that the probability of two introns occurring right after the two in frame TAG codons is approximately 1 in 10^5, it seems likely that the location of these amber codons is the reflection of an early functional role for the sequences. It was suggested that exons were once microgenes that ended with amber codon and encoded relatively short independently translated polypeptides, which could assemble spontaneously into an active protein. In this scenario, from various segments of primordial RNA, initiation and termination of protein synthesis would produce a library of short polypeptides, some combinations of which would combine to form multichain protein assemblies having enzymatic activity. Splicing of these microgenes by using their common amber termini as a recognition element could then bring the appropriate microgenes in proximity, improving the chances for linker inheritance of all the fragments required to generate catalytic activity.

This notion is consistent with the exon-shuffling based generation of primordial genes. Within the "exons as microgenes" theory, exons that code for polypeptides are required to non-covalently assemble proximal to each other to ensure proper inheritance of a set of exons required to produce functional proteins. Read-through rather than transcriptional termination at the end of each microgene after the initial splicing out of the introns would then produce a continuous polypeptide chain that would fold into a protein with greater thermal stability. There are only two major approaches to how such read-through can occur: either termination codon gets mutated to a coding sequence or the meaning of a stop codon is changed[29]. *Tetrahymena* employed the second approach to generate read-through sequences. The sequence of a termination codon TAG was retained but now assigned to code for attachment of Glu to a growing polypeptide chain[29].

Such an approach allows for generation of a variety of different polypeptide building blocks that could be utilized for different functions. If a produced polypeptide was found to not contribute to catalytic activity, it could instead have been used for building encapsulated structures and served as a highly ordered template for the assembly of nucleotides. These functions could have potentially allowed for formation of the current biochemical machinery that combines the elaborate and coordinated interaction between nucleic acids and proteins to allow for the functioning of living organisms[30].

The discovery of amber codons terminating sequences of exons as well as introns in phopshoenolpyruvate mutase from *Tetrahymena* implies that in primordial times, at least in the case of some genes their original structure was of such a sequence of several microgenes, represented by each individual exon terminating with an amber codon. **The non-covalent assembly of individually translated polypeptides would lead to formation of active protein** (Figure 9).

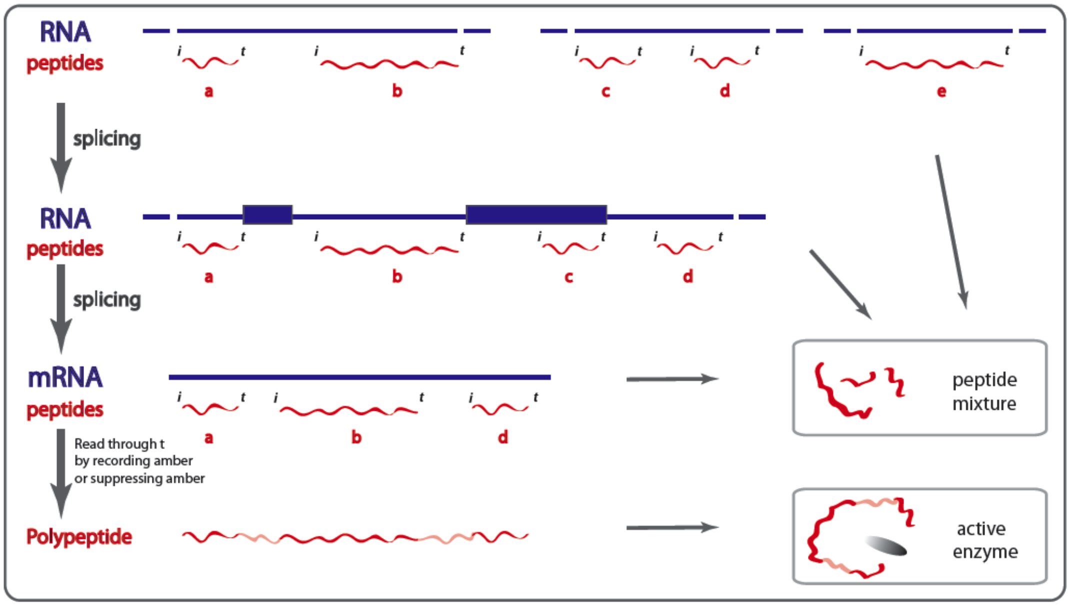


Figure 9. Schematic representation of "exons as microgenes" theory. Exon coded polypeptides are separately produced off of the respective short RNAs, followed by non-covalent assembly in solution to produce an active protein. Exons or coding RNA that are involved in generation of active polypeptides are brought in proximity by splicing. Evolution towards modern day enzymes takes places either through codon reassignment (of termination codon) or by conserving mutations in it leading to functional reassignment. Suppression of termination codon leads to generation of a continuous covalent polypeptide chain.

The goal of this dissertation is to provide experimental support to the exons as microgenes theory by examining whether it is possible to generate enzymatically active proteins through assembly of exon-coded polypeptides.

2.2 Effective routes to protein re-assembly

The ability of proteins to recover enzymatic activity after a full-length enzyme is split into two or more polypeptide fragments that are allowed to re-assemble in solution has been investigated and shown to be successful in many different proteins[31]. Such an experimental approach to enzymatic activity recovery is most commonly known as protein complementation assay or PCA. Various enzymes including cytochrome c[32], RNase A[33], the $(\beta\alpha)_8$-barrel enzyme[34,35,36], phosphoribosylanthranilate isomerase[37], tRNA synthetase[38], alanine racemase[39], and the pituitary growth hormone[40] have been subjected to protein complementation assays.

There are two major approaches that allow one to generate separate domains. The first method involves incorporation of a stop codon within a gene sequence cloned on a certain plasmid at a site where one wishes to terminate synthesis of the first domain. Correspondingly, a start codon, along with the ribosome binding site, is inserted at the site where one would like to initiate protein synthesis of the second domain. This allows for both protein domains to be produced off of the same plasmid, thus making the experimental set up easier. However, unequal expression levels are frequently observed in case of these constructs. An alterative route is to use a site-specific proteolytic cleavage to produce the desired protein fragments, which can then be further purified. The proteolytic site might be native or it can be introduced genetically using one of the many commercially available site-mutagenesis kits[41] (Figure 10).

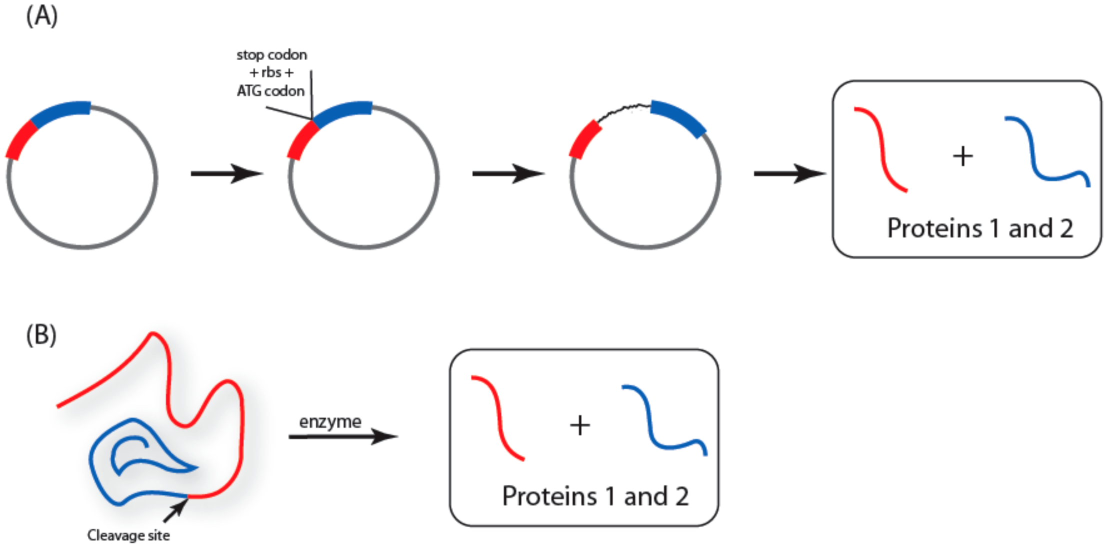


Figure 10. Methods of generation protein complementation fragments. **A**. Genetic modification of original gene consisting of introduction of a stop codon at the end of the desired protein sequence of fragment 1 followed by incorporation of a ribosome binding site (RBS) and start codon (ATG) immediately preceding the sequence of the second protein fragment. **B**. Presence of cleavage site or incorporation of such using chemical/genetic methods allows for enzymatic (proteolytic) cleavage of the full-length protein generates two protein fragments.

All of the above mentioned proteins were split into two fragments using one of the methods discussed above. The amount of recovered activity via protein complementation assay varies anywhere from 3% of the wild-type enzymatic activity, as observed in the case of split alanine racemase[39], to 100% in the case of phosphoribosylanthranilate isomerase[37], judging by either structural characterization of the complement using CD or NMR, or kinetic parameters of the complement. The phenomenon of catalytic activity recovery from non-covalent complementation of protein fragments has found use in the field of studying protein-protein interactions[42] and protein folding[43]. Below utilities of protein fragment complementation are illustrated with experimental examples.

Ever since its discovery in 1950s by Fred Richardson, the protein complementation system of S-peptide and S-protein, components of the bovine poncreatic ribonuclease have served as a model system for studying protein-protein interactions[42] (Figure 11).

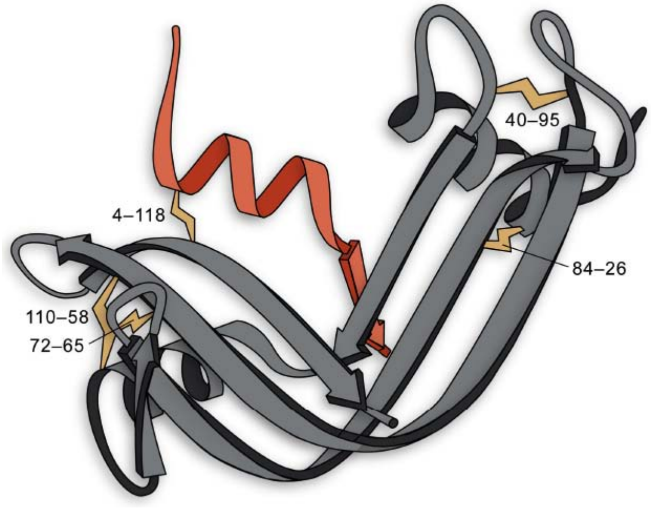


Figure 11. Structure of "RNase-S"[42]. Protein S is shown in grey, peptide S is shown in red. Disulfide bonds are in yellow and are formed between cysteine residues indicated by numbers. Reproduced from reference 42 with permission. ©ChemComm 2011.

In this system, a S-peptide analog is synthesized and its interaction with S-protein is investigated. The importance of introduced mutations can be estimated through measurement of kinetic parameters of complex formation (Kd) as well as the tightness of binding (K_m). That system has served as a basis for generation of the widely popular S-tag based fusion protein purification system. In this system S-peptide sequence is incorporated into the sequence of the parent protein and S-peptide affinity for S-protein is used as the basis for protein purification. However, there are certain drawbacks. One of these drawbacks is due to inefficient proteolytic generation of S-peptide and S-protein from RNase A. Inefficient proteolysis leaves uncleaved RNase A behind, thus contaminating the S-peptide/S-protein system with uncleaved parental protein. Another limitation is the inherent low K_d of the S-protein-S-peptide-complex. This complication hinders investigation of systems with K_ds in the middle to low nanomolar range due to the extended time frame required for formation and investigation of complexes. These drawbacks are currently being addressed with a goal of making this system more robust.

Protein fragment complementation has recently been used in creation of novel biosensors for dsDNA detection[44]. Detection of dsDNA is enabled by a method named sequence-enabled reassembly (SEER) (Figure 12).

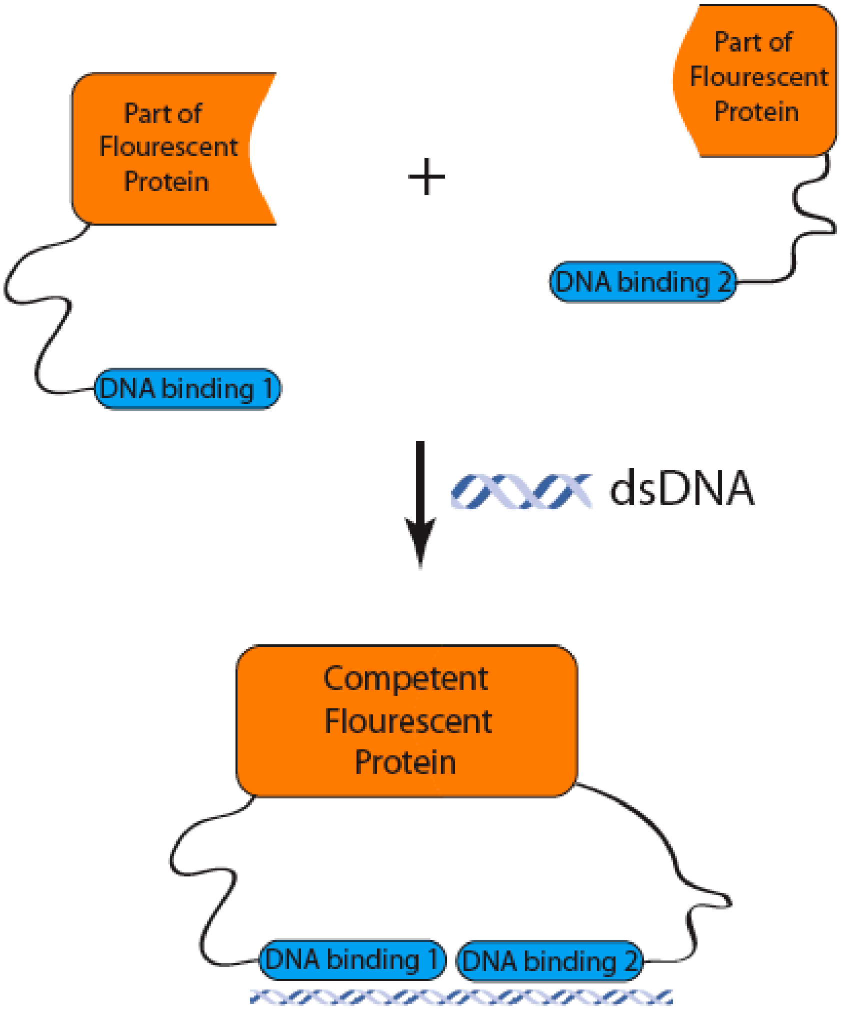

Part of Flourescent Protein
+
Part of Flourescent Protein
DNA binding 2
DNA binding 1
dsDNA
Competent Flourescent Protein
DNA binding 1
DNA binding 2

Figure 12. Sequence-Enabled Reassembly. Split fluorescent protein fusion constructs are originally non-fluorescent. The binding of the dsDNA target to DNA binding domains **1** and **2** results in reassembly of the fluorescent protein.

In this approach, fragments of split-reporter proteins are appended to DNA detection domains. The binding event is then monitored by signal generation arising from conditional reassembly of split-protein halves. By fusing split-GFP to sequence-specific zinc finger domains, recognition of target DNA will induce GFP reassembly. Detection of single nucleotide sequence is possible in any one experiment. However, since the GFP reporter affords the opportunity to use other GFP variants with distinct spectroscopic properties, there is room for generation of new SEER systems that would be able to simultaneously detect multiple targets. This would be beneficial for multiple applications, including DNA profiling and ratiometric analysis. Many other proteins have been used in protein fragment complementation assays (PCAs) and have a number of biotechnological applications. Among these are dihydrofolate reductase, ubiquitin, glycinamide ribonucletide transformylase, β-lactamase, β-galactosidase and others.

It has been amply demonstrated that recovery of enzymatic activity via complementation is possible and robust enough to find applications in biotechnology as well as basic research such as elucidation of protein-protein interactions. Even though demonstrated possible, none of fragments generated for complementation within experiments described above correlated with exon/exon boundaries of an original protein. *The goal of this dissertation is to experimentally demonstrate that complementation of protein fragments defined by exon boundaries leads to effective re-assembly of protein and recovery of enzymatic activity.*

2.3 Protein "cut-points" and correlation to structure

Much like correlations established between exon/intron boundaries and the domain distribution along protein sequences of ancient enzymes, attempts at establishing correlation between locations of cut sites to the recovered activity have been made[45,46]. Even though the number of successful protein fragment complementation (PFC) experiments is rather large, judging from the number of publications on the subject, one has to be very careful when choosing a site for successful protein surgery. In theory, cut sites should not fall between well-defined domains or structural units or within conserved regions, as well as within secondary structure elements such as α-helices. For most locations, a cut does not lead to protein fragment complementation; proteins presumably do not re-assemble due to inefficient assembly or improper folding of the fragments. For example, isoleucyl-tRNA synthetase was found to tolerate cuts at 11 of 18 sampled locations[47].

For some sites in certain proteins this might be overcome by fusion of the fragments to dimerization domains, which could facilitate correct assembly as seen in case of split green fluorescent protein dsDNA reporter assisted protein assembly[44]. Incremental truncation, a method for rapidly generating a DNA library of every 1 base pair deletion of a gene or gene fragment, can be used to evaluate “protein fragment complementation space”[31]. Within this methodology, the gene of interest is divided into two overlapping non-active fragments. Each fragment is cloned into compatible vectors designed specifically to allow for creating incremental truncation libraries. In short, the vector containing the fragment of a gene is digested with restriction enzymes A and B, where A leaves a 3’ recessed end susceptible to ExoIII digestion, whereas restriction enzyme B leaves a 5’ overhang resistant to ExoIII digestion. Following the addition of ExoIII, short samples are taken after discrete time intervals in this manner allowing for generation of varying lengths of 3’ DNA terminus. Single-stranded tails left from ExoIII digestion are removed by incubation with mung bean nuclease and the ends are blunted by the Klenow fragment. The plasmids are then re-circularized using T4 DNA ligase under dilute conditions and ready for a transformation into a desired host such as *E. coli*[31] (Figure 13).

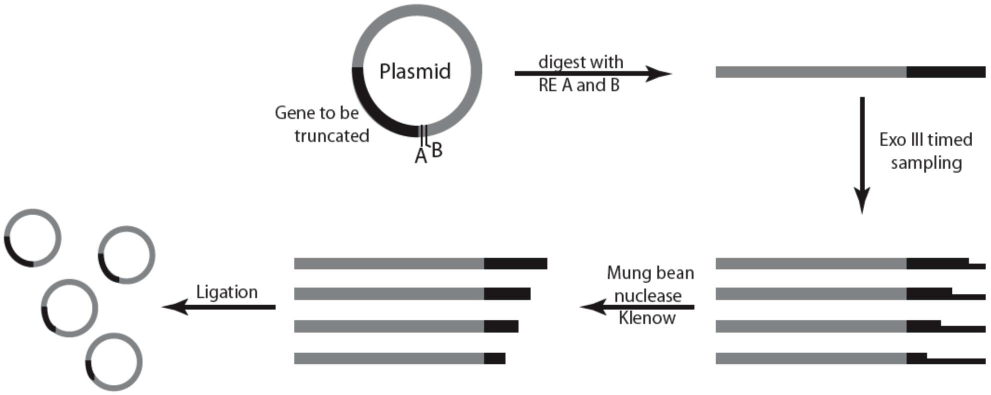


Figure 13. Gene fragment generation through incremental truncation. The original plasmid coding for a full length protein is digested using two restriction enzymes: A and B. Enzyme B leaves an overhang resistant to Exo III digestion, while enzyme A does not. Timed digestion with Exo III nuclease produces a number of genes varying in length, shorter gene fragments are produced by longer incubation times. The generated set of truncated genes is then blunt ended and ligated.

Plasmids containing both fragments of the gene are treated the same way and are transformed into the same host cell. An engineered screening assay for detection and identification of active complements is used. This methodology was applied to explore the structural foundation for the catalytic activity of *E. coli* glycinamide ribonucleotide formyltransferase (PurN)[48]. Based on the structure of PurN, the *purN* gene was divided into two non-active overlapping fragments. The N-terminal fragment consisted of DNA coding for residues 1-144 and the C-terminal piece consisted of DNA coding for residues 63-212. By experimental design, the cut-points were limited to the region of overlap between the two fragments. Two fragments were cloned and subjected to incremental truncation thereby generating libraries (Figure 14).

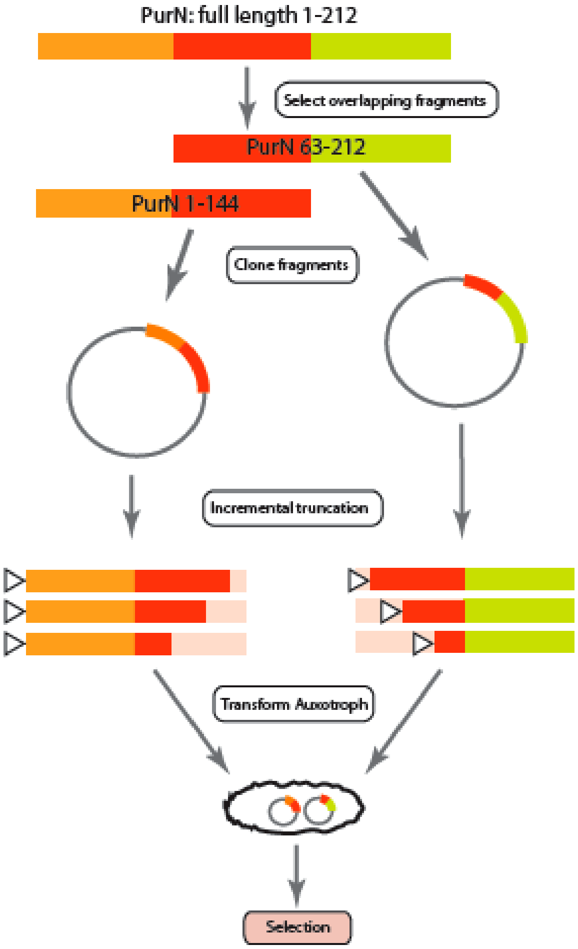


Figure 14. Generation of an incremental library of overlapping PurN protein fragments. Two fragments coding for the overlapping regions of PurN protein (1-144 and 63-212) were separately cloned. The plasmids so obtained were subjected to incremental truncation, and sets of two plasmids coding for overlapping but truncated protein fragments were transformed into *E. coli* on selective media. Viable colonies were identified and the sequence of protein fragments leading to formation of active protein reassembly was elucidated through sequencing.

Truncation size diversity was confirmed by picking 10 random library members and examining the size of the truncated gene by a restriction enzyme or PCR method. These libraries were transformed into an auxotrophic *E. coli* strain unable to grow on minimal media lacking purines, unless GAR transformylase activity was supplied to the bacterium. Active PurN heterodimers are then selected by plating the crossed library onto selective minimal plates. The frequency of functional heterodimer generation was found to be 0.67%. Active heterodimers were found with cut-points within non-conserved as well as conserved regions of a protein. Two of the established active heterodimers were purified for further characterization. Their kinetic parameters (K_m values) were found to be very close to those of native wild-type enzyme. Another poorly active heterodimer was purified with a goal of establishing the minimal specific activity required to be delivered by a heterodimer in order to be identified as a positive in this screen. It turned out that the lowest detection limit of specific activity was 500-fold lower then the wild type enzyme. Even though this method provides us with a tool to generate the complementation fragments of any protein of any size and permits us to establish the correlation between obtained activity and the location of "cut-point", such work would be technically challenging as it would require individual isolation and sequencing of all the plasmids generated during formation of incremental truncation libraries (provided that the average size of the libraries generated is 10^5-10^6). Additionally, the proposed screening technique would only allow identification of active heterodimers with loss of all information on the locations of the cuts that led to formation of inactive heterodimers.

Another attempt at elucidation of tolerated split positions was undertaken by Bertolaet and Heitmeyer and discussed in their review on functional protein evolution[41]. They postulated that by randomly interrupting genes with inserted sequence encoding peptide "looplets" and then assaying the resulting "interrupted" protein for catalytic activity one could identify "functional" cuts. These loops were inserted by randomly digesting the gene of interest with DNase I, that leaves two-base overhang; further generating flush ends with mung bean nuclease thereby eliminating the two base overhang, and ligating in an eight-nucleotide insert so that the net result is the insertion of six nucleotides into the gene. While the codons located at the site of the digestion may be mutated, the net insertion of six nucleotides would not lead to any shifts in the open reading frame of the parent gene. The model system proposed for this study was *E. coli* chloramphenicol acetyltransferase (CAT) type III. The crystal structure for this protein has been solved and used as a critical guide to these studies. The CAT enzyme confers resistance to the antibiotic chloramphenicol, which inhibits peptidyl transferase activity of the 50S subunit of prokaryotic ribosomes. Thus, growth selection on chloramphenicol media was used as a powerful and convenient assay for catalytic activity. The colonies that grew on the selective media are the ones containing the plasmid coding for an active CAT enzyme. Plasmid DNA were isolated from selected transformants and digested with appropriate restriction enzymes, leading to liberation of the inserted genes. The size of these genes was determined by comparison to a standard DNA ladder by polyacrylamide gel electrophoresis (PAGE). Analysis of the first amino-terminal quarter of the promoter showed that 16% of the positions (1 out of every 6.3 amino acids) tolerated the interruption of its sequence and structure with a dipeptide insert and remains catalytically active. This establishes that interactions occur mainly between structural elements (i.e. between a β-strand and α-helix), with some occurring in the last turn of a α-helix. All of the identified allowed cut-points were located on the surface of the protein, however, no comments on the correlation between observed growth rate to the cut-point position on the protein were made.

It is obvious from presented data that an intuitive approach to definition of cut positions cannot be relied on. While it feels like common sense to not introduce cuts in the middle of the secondary structure units, such as α-helicies, it does not correlate well with some of the experimentally observed data for various engineered PurN fragmented proteins. While Bertolaet and Heitmeyer indicate that majority of the cut positions leading to the formation of active protein fragment complements are between structural elements, the major point that is made is that all of identified bisection positions localize on the protein surface. That result is consistent with the major premise of the exons as microgenes theory. If the rudimentary enzymes were in fact non-covalent assemblies of short exon-coded polypeptides the charged C and N termini of these polypeptides would be best thermodynamically accommodated at protein surface positions.

The following chapter will be dedicated to overview of the currently available experimental data to support the “exons as microgenes” theory. A few attempts have been made to demonstrate recovery of enzymatic activity of protein reassembly though non-covalent complementation of fragments defined by exon/intron boundaries. These partially successful experiments laid foundation of our novel approach to confirmation and characterization of successful exon-coded polypeptide reassembly.

2.4 References

(1) Nedelkov, D. *Macedonian Journal of Chemistry and Chemical Engineering* **2008**, *27*, 99-106.

(2) Dorit, R. L.; Gilbert, W. *Current Opinion in Genetics and Development* **1991**, *1*, 464-469.

(3) Long, M.; Betr√°n, E.; Thornton, K.; Wang, W. *Nature Reviews Genetics* **2003**, *4*, 865-875.

(4) Kolkman, J. A.; Stemmer, W. P. C. *Nature Biotechnology* **2001**, *19*, 423-428.

(5) Sudhof, T. C.; Goldstein, J. L.; Brown, M. S.; Russell, D. W. *Science* **1985**, *228*, 815-822.

(6) Saxonov, S.; Gilbert, W. *Genetica* **2003**, *118*, 267-278.

(7) Peng, L.; Sakharkar, K. R.; Sakharkar, M. K. *International Journal of Integrative Biology* **2009**, *5*, 87-91.

(8) Ellington, A. D.; Szostak, J. W. *Nature* **1990**, *346*, 818-822.

(9) Dube, D. K.; Loeb, L. A. *Biochemistry* **1989**, *28*, 5703-5707.

(10) Shiba, K.; Takahashi, Y.; Noda, T. *Proceedings of the National Academy of Sciences of the United States of America* **1997**, *94*, 3805-3810.

(11) Crameri, A.; Raillard, S. A.; Bermudez, E.; Stemmer, W. P. C. *Nature* **1998**, *391*, 288-291.

(12) Trujillo, M.; Duncan, R.; Santi, D. V. *Protein Engineering* **1997**, *10*, 567-573.

(13) James, C. L.; Viola, R. E. *Biochemistry* **2002**, *41*, 3726-3731.

(14) Gilbert, W.; De Souza, S. J.; Long, M. *Proceedings of the National Academy of Sciences of the United States of America* **1997**, *94*, 7698-7703.

(15) Roy, S. W. *Genetica* **2003**, *118*, 251-266.

(16) Vanacova, S.,Yan, W.; Carlton, J. M.; Johnson, P. J. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102*, 4430-4435.

(17) Coulombe-Huntington, J.; Majewski, J. *Genome Research* **2007**, *17*, 23-32.

(18) Straus, D.; Gilbert, W. *Molecular and cellular biology* **1985**, *5*, 3497-3506.

(19) Dibb, N. J.; Newman, A. J. *EMBO Journal* **1989**, *8*, 2015-2021.

(20) De Souza, S. J. *Genetica* **2003**, *118*, 117-121.

(21) Vibranovski, M. D.; Sakabe, N. J.; De Oliveira, R. S.; De Souza, S. J. *Journal of Molecular Evolution* **2005**, *61*, 341-350.

(22) Wetlaufer, D. B. *Proceedings of the National Academy of Sciences of the United States of America* **1973**, *70*, 697-701.

(23) Wodak, S. J.; Janin, J. *Biochemistry* **1981**, *20*, 6544-6552.

(24) Go, M. *Nature* **1981**, *291*, 90-92.

(25) De Souza, S. J.; Long, M.; Schoenbach, L.; Roy, S. W.; Gilbert, W. *Proceedings of the National Academy of Sciences of the United States of America* **1996**, *93*, 14632-14636.

(26) Thomson, T. M.; Lozano, J. J.; Loukili, N.; Carrio, R.; Serras, F.; Cormand, B.; Valeri, M.; Diaz, V. M.; Abril, J.; Burset, M.; Merino, J.; Macaya, A.; Corominas, M.; Guigi, R. *Genome Research* **2000**, *10*, 1743-1756.

(27) De Souza, S. J.; Long, M.; Schoenbach, L.; Roy, S. W.; Gilbert, W. *Gene* **1997**, *205*, 141-144.

(28) Fedorov, A.; Roy, S.; Cao, X.; Gilbert, W. *Genome Research* **2003**, *13*, 1155-1157.

(29) Seidel, H. M.; Pompliano, D. L.; Knowles, J. R. *Science* **1992**, *257*, 1489-1490.

(30) Carny, O.; Gazit, E. *FASEB Journal* **2005**, *19*, 1051-1055.

(31) Paschon, D. E.; Ostermeier, M. In *Methods in Enzymology* 2004; Vol. 388, p 103-116.

(32) Wu, L. C.; Laub, P. B.; Elove, G. A.; Carey, J.; Roder, H. *Biochemistry* **1993**, *32*, 10271-10276.

(33) Jackson, D. Y.; Burnier, J.; Quan, C.; Stanley, M.; Tom, J.; Wells, J. A. *Science* **1994**, *266*, 243-247.

(34) Sancho, J.; Fersht, A. R. *Journal of Molecular Biology* **1992**, *224*, 741-747.

(35) Hocker, B.; Beismann-Driemeyer, S.; Hettwer, S.; Lustig, A.; Sterner, R. *Nature Structural Biology* **2001**, *8*, 32-36.

(36) Kallenbach, N. *Proceedings of the National Academy of Sciences of the United States of America* **2001**, *98*, 2958-2960.

(37) Eder, J.; Kirschner, K. *Biochemistry* **1992**, *31*, 3617-3625.

(38) Burbaum, J. J. *Biochemistry* **1991**, *30*, 319-324.

(39) Galakatos, N. G.; Walsh, C. T. *Biochemistry* **1987**, *26*, 8475-8480.

(40) Li, C. H.; Bewley, T. A. *Proceedings of the National Academy of Sciences of the United States of America* **1976**, *73*, 1476-1479.

(41) Bertolaet, B. L.; Heitmeyer, D. P. *Bioorganic Chemistry* **1995**, *23*, 355-368.

(42) Watkins, R. W.; Arnold, U.; Raines, R. T. *Chemical Communications*, *47*, 973-975.

(43) De Prat-Gay, G. *Protein Engineering* **1996**, *9*, 843-847.

(44) Furman, J. L.; Badran, A. H.; Shen, S.; Stains, C. I.; Hannallah, J.; Segal, D. J.; Ghosh, I. *Bioorganic and Medicinal Chemistry Letters* **2009**, *19*, 3748-3751.

(45) Traut, T. W. *Proceedings of the National Academy of Sciences of the United States of America* **1988**, *85*, 2944-2948.

(46) Blake, C. *Trends in Biochemical Sciences* **1983**, *8*, 11-13.

(47) Shiba, K.; Schimmel, P. *Proceedings of the National Academy of Sciences of the United States of America* **1992**, *89*, 1880-1884.

(48) Ostermeier, M.; Nixon, A. E.; Shim, J. H.; Benkovic, S. J. *Proceedings of the National Academy of Sciences of the United States of America* **1999**, *96*, 3562-3567.

Chapter 2

PRIMORDIAL ENZYMES

2.1 INTRODUCTION

Enzymes are essential to life as we know it, serving as catalysts for nearly all chemical reactions that define cellular metabolism. Their enormous rate accelerations and high selectivities make them extremely valuable in multiple research as well as industrial applications. The properties of enzymes are determined by their precise three-dimensional structures, however, our knowledge of structure–function relationships and the detailed rules that govern protein folding is at best incomplete. It follows that the design of enzymes from first principles remains a challenging proposition. In order to be able to create a functional protein from scratch one needs to understand the evolution of protein complexity. Once the principles by which nature evolved these complex "machines of execution" are understood, modern research can model the path of natural evolution to create functional proteins.

In order to shed light on how modern proteins could have evolved, the fundamental question is how functionality was embedded into proteins? What perturbations to protein structure can be handled without losing enzymatic activity? Many studies are dedicated to answering this question and are mostly done through a site-directed mutagenesis approach[1,2]. This approach allows sequential alterations of amino acids and correlates individual amino acid modifications to changes in overall protein structure and function. While this provides information on which parts of protein are essential for catalysis, this approach yields almost no information on how nature generated the protein under study.

The idea of "exons as microgenes" (after the initial discovery of introns as intervening, non-coding parts of DNA) provided a long-sought hypothesis for protein evolution. Within this hypothesis, exons were proposed to independently code for short polypeptides that would non-covalently assemble in solution[3]. If a successful combination of exon-coded polypeptides was obtained, it would be selected. The idea of recovery of protein activity via fragment complementation has been extensively studied since. Many proteins, as discussed in Chapter 1, have been cut at favorable positions and then allowed to reassemble in solution. The activity generated through complementation spans the entire range depending on particular pieces and the protein. However, none of these experiments support the "exons as microgenes" theory as the generated polypeptide fragments were not defined by exon/intron boundaries. The boundaries that are imprinted on the DNA level indicating the end of one exon and the beginning of the other can be easily translated onto the protein level. Successful complementation of exon-coded polypeptides derived from an ancient protein would provide direct experimental support for the "exon as microgenes" theory. Several attempts to test this theory were undertaken previously and will be discussed herein as precedence for the work in this thesis.

2.2 Exon-coded-polypeptide-guided protein re-assembly

Though "exons as microgenes" idea was proposed 35 years ago, however, only three studies have attempted to provide experimental evidence. Experimental challenges associated with generation of protein fragments defined by exon boundaries were the major reason behind such delay in experimental support. Absence of sequencing data hence knowledge on exon/intron structure of genes as well as tools to generate and purify protein fragments hindered these studies 35 years ago. One of the very first proteins to be studied was chicken egg lysozyme. The exon-coded polypeptides were attached to a much larger polypeptide chain and allowed to reassemble in solution. Even though some restored activity was observed (on the order of 50,000 less than compared to the full-length enzyme), the presence of the much larger protein tag may have interfered with proper complementation causing either a positive or negative effect.

There are two important studies that tried to answer this question; firstly, the examination of O_2 binding properties of the product of the central exon of β-globin gene by Craik and Beychok[4] and secondly, complementation of fragments of triosephosphate isomerase defined by exon boundaries by Knowles and Bertolaet[5]. Because results of these studies were crucial in the design of our experiments, they are discussed in more detail.

2.2.1 O_2 binding properties of the product of the central exon of β-globin gene

Many crystal structures of hemoglobin from a variety of sources have been determined, including human hemoglobin. Regardless of the source, hemoglobin is a protein with complex architecture. It consists of four polypeptide chains: two α chains and two β chains, coded for by a total of three exons. The hemoglobin tetramer or hemoglobin A is best described as a dimer of two identical αβ subunits ($\alpha_1\beta_1$ and $\alpha_2\beta_2$) with a principal role of oxygen transport[6,7]. Hemoglobin reversibly binds oxygen in its ferrous state. The stability of the FeII-oxygenated species seems to be achieved at least in part by accommodation of the heme group in a hydrophobic pocket of the protein with the majority of stabilization coming from interaction of side chain groups that are projected into the heme binding pocket[4]. The structural complexity may be due to the many binding events that have to be orchestrated in order for the protein to perform its function. The binding of heme and oxygen as well as protons, carbon dioxide and 2,3-diphosphoglycerate that are involved in allosteric regulation of hemoglobin's activity need to be accommodated for effective functionality.

Interestingly enough the distribution of the residues that are involved in these binding events along the amino acid sequence is not random and majority of heme binding residues (16 out of 20) have been shown to be located on the polypeptide chain coded for by a central exon of hemoglobin[8]. Such correlation for location of heme binding residues was observed in the case of both α- and β-globin chains. Residues involved in intersubunit contacts between $\alpha_1\beta_1$ and $\alpha_2\beta_2$ dimers and binding of allosteric regulators show a strong distribution correlation as well. For example, 12 out of 15 residues in the β-subunit that form intersubunit bonds are located in the central exon coded polypeptide and 13 out of 18 in case of α. The majority of the residues involved in stabilization of interactions between α_1 and β_1 are situated on the polypeptide coded for by a third exon. Such correlation of functional residues that are segregated between different exon-coded polypeptides promoted the suggestion that different exon-coded polypeptides of hemoglobin correspond to the major functions of hemoglobin.

In order to address this hypothesis all three exon-coded polypeptides of mouse β hemoglobin chain were prepared[9]. Rapid generation of these polypeptides was facilitated by coincidental localization of arginine residues to the exon/exon boundaries of mouse hemoglobin and a facile digestion of the intact β hemoglobin chain with the arginine specific protease clostripain lead to formation of the three desired polypeptides (Figure 15).

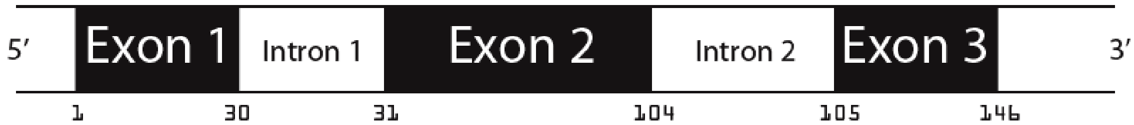


Figure 15. Schematic representation of the β-globin intron/exon gene organization. Exon 1, Exon 2 and Exon 3 are the exonic regions of the gene. Intron 1 and Intron 2 represent the two introns of the gene. Numbers below the exonic regions translate into the amino acid sequence position of the β-globin protein. E2 is the central exon-coded polypeptide responsible for binding heme. Amino acid residues 31 and 99 are arginines that were utilized for generation of exon-coded polypeptide through arginine-specific protease cleavage.

A minor complication was observed during generation of the central polypeptide due to the presence of an additional arginine residue close to the N-terminus of central polypeptide. The reactivity of this N-terminal proximal arginine is lower than the reactivity of the arginine located at exon1/exon2 boundary during protease digestion, however formation of both versions of the central polypeptide (31-104 full length version and 40-104 doubly digested version) was observed as judged by SDS-PAGE analysis of Sephadex-50 purified polypeptides. These two versions of central exon (exon 2) coded polypeptide were used as a mixture of equimolar amounts of each in the following experiments. The spectral profile of the Soret band and the visible region of the untreated β hemoglobin chain and central β hemoglobin fragment mixture (31-104 and 41-104) appeared to look identical with the Soret maxima coinciding at 420 nm. The spectral profile of a peptic digest of β globin in the presence of hemin dicyanide was unchanged from that of hemin dicyanide reagent, confirming that it is the fragment defined by exon/intron boundaries of the central exon of β hemoglobin that is required in order to exhibit heme-binding function.

The ability of the β globin central fragment to bind heme was very encouraging and prompted studies of its' ability to reversibly bind oxygen[4]. If demonstrated to be unable to perform reversible oxygen binding, the authors set out a goal of establishing minimal requirements for delivery of this function. The distinctive spectral characteristics of various ligand derivatives of hemoglobin have been well documented and provide a sensitive means of evaluating the environment of the prosthetic group in the protein. The ability to identify various heme and reversibly or irreversibly bound oxygen species based on spectral analysis was utilized in this study as well. The ability to bind oxygen reversibly is the functional hallmark of hemoglobin and only upon restoration of reversible binding can an exon-coded polypeptide be called fully functional.

To follow oxygen binding to the hemoglobin central domain, both by itself and in combination with other exon-coded fragments of hemoglobin, the α- and β-chains were proteolytically cleaved with an arginine-specific protease clostripain to generate fragments corresponding to exon-coded regions.

Comparison of the visible and Soret spectra of heme reconstituted β-globin, isolated central exon-coded polypeptide, and β-globin digest revealed no observable difference. Reduction with dithionite followed by subsequent removal of residual dithionite led to the production of oxyhemoglobin in the case of the undigested sample. However, hemichromogen, which is diagnostic of a heme complex that does not reversibly bind oxygen, was produced in case of isolated central exon-coded polypeptide and β-globin digest. When the complementary heme-containing α subunit was combined with either the central exon-coded polypeptide of β-globin or β-globin digest, a typical oxy spectrum was generated (Figure 16).

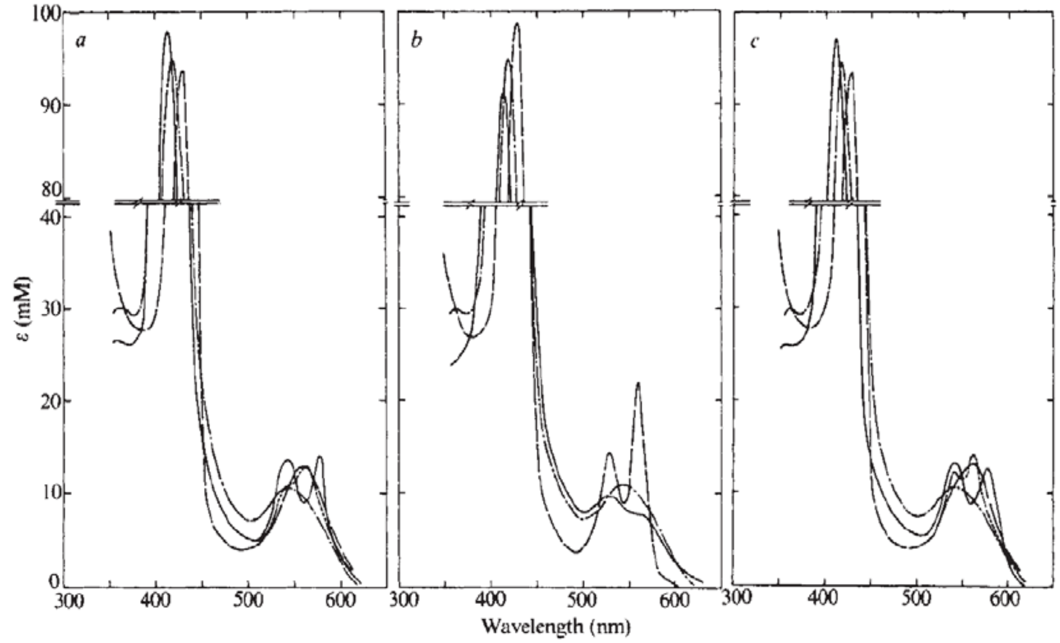


Figure 16. Spectral profiles of various species of β-globin. Proteolytic digest, full-length and α- β-subunits complement in the presence/absence of heme or oxygen[4] **a.** Cyanmet spectra of an equimolar mixture of β-globin with haemin cyanide followed by a transient ferrocyanide spectrum immediately after dithionite addition and the eventual deoxy-spectrum, followed by oxy-spectrum after dithionite removal. Clear spectral differences are observed in the 600-500 nm range for different β-globin-heme bound species. **b.** Similar spectra were recorded for digested β-globin. Note the different spectral profile for the species generated upon removal of dithionite from solution (spectrum with two peaks). Such a profile is characteristic of irreversibly bound oxygenated species. **c.** Similar spectra as in **a** with digested β-globin complemented with heme containing intact α-subunit. Notice the restoration of the oxy-species profile (two peaks are observed of same intensity in 500nm-600nm range) as in the case of undigested β-globin. Reversible oxygen binding is observed in case when digested β-globin is complemented with the intact hemoglobin α-subunit. Reproduced from reference 4 with permission. © Nature 1981

The observed ability of a central exon coded polypeptide to bind oxygen in the presence of heme is a crucial finding, as it demonstrates the ability of an exon 2-coded polypeptide outside of the protein content to bind oxygen as well as heme, although irreversibly. Also, upon addition of complementary heme containing α-subunit, full recovery of the reversible oxygen-binding pattern was observed. This observation supports the notion that exon-coded polypeptides can non-covalently re-assemble in solution leading to recovery of catalytic activity.

2.2.2 Complementation of fragments of triosephosphate isomerase defined by exon boundaries

Knowles revisited this problem with an eye towards explaining the origin of consensus RNA splice site sequences in protein-coding genes at exon-intron junctions. His proposal suggested that the conserved splice site AG could have been derived from TAG of a termination codon originally present at this location, leading to production of independently translated peptides. Ability of these peptides to reassemble in solution and form an active enzyme was tested via *in vivo* complementation of fragments of triosephosphate isomerase defined by exon boundaries[5].

Triosephosphate isomerase (TIM) is an ancient, ubiquitous and a highly conserved enzyme that is believed to have evolved prior to the archaebacteria-prokaryotic-eukaryotic branching. Much is known about its structure, mechanism and stability, making it a very favorable subject for study[10]. Chicken muscle TIM is the quintessential paradigm for the 8 $\alpha\beta$-barrel structure. The enzyme is a homodimer with a subunit molecular weight of 26500. The chicken gene coding for this protein has seven exons interrupted by six introns. Non-random distribution of chicken TIM exon sizes lead Gilbert to hypothesize that the TIM gene was originally constructed by the multiplication of the “genetic units” encoding the $\alpha\beta$-barrel motif. All but two of the splice junctions in the genomic sequence for chicken TIM conform to the eukaryotic consensus sequence of …AG/.. for the exon/intron juncture and …/ag… for the intron/exon one. Surface locations of exon-intron junctions in the TIM monomer are consistent with the need of charged N- and C-termini of independently translated peptides to be solvated. Such position of boundaries would also be beneficial when a stop-codon read-through occurred, allowing peptide fragments to be joined by a peptide loop that would be readily accommodated at the surface of the folded polypeptide (FIGURE 17).

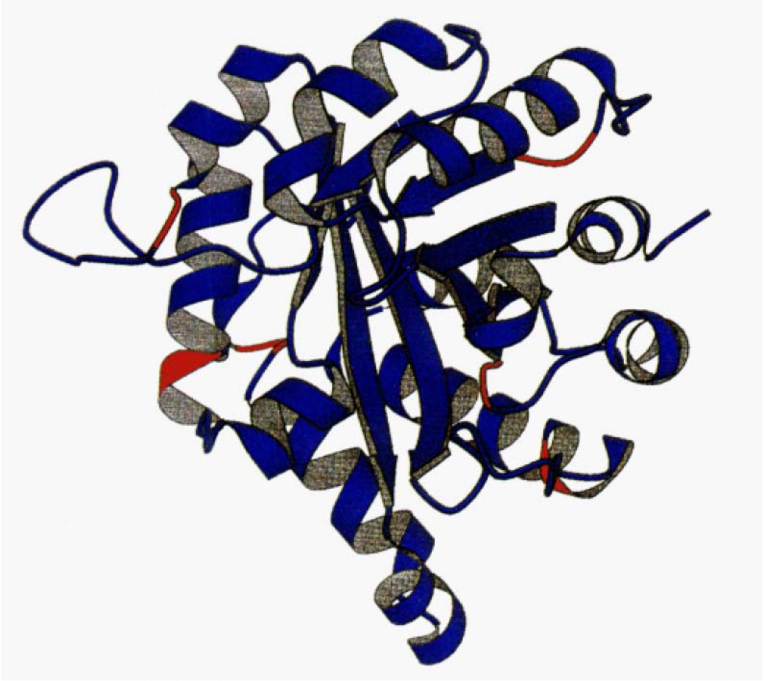


Figure 17. Cartoon ribbon diagram of chicken TIM monomer[5]. Locations of exon/exon boundaries of TIM are indicated in red. Reproduced from reference 5 with permission. © Biochemistry 1995

TIM is highly conserved in terms of sequence, structure and kinetic characteristics. Many of the conserved residues lie in, or near the active site residues, most of which are encoded by different exons[11]. Therefore, the codons for active site residues are distributed among the exons of the TIM gene. For this reason, only when a proper complementation of fragments defined by exon boundaries was achieved would one observe TIM activity, in other words all active site residues must re-assemble to create an active enzyme. Because the nature and location of the exon-intron junctions, rather than domains of the protein are the subject of this work, the choice of the cleavage sites was of great importance. Proteolytic cleavage would not have generated appropriate peptides, so the TIM gene was modified to encode polypeptides corresponding precisely to those defined by exon/intron boundaries. The continuous chicken TIM gene was split at three exon/intron junctions to encode a set of TIM proteins, each fragmented into two polypeptides. The splits were made by inserting a sequence that includes a stop codon at the end of the boundary-determining exon and a start codon at the beginning of the next exon. The ribosomal binding site was introduced upstream of the latter exon , creating an antercistronic sequence that is not a multiple of three. Such an arrangement eliminates the possibility of read-through translation. The designed plasmids code for individual polypeptides that are translationally coupled (FIGURE 18).

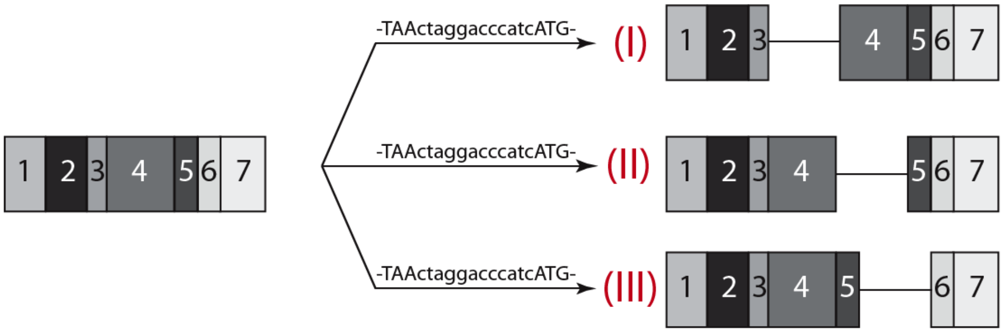
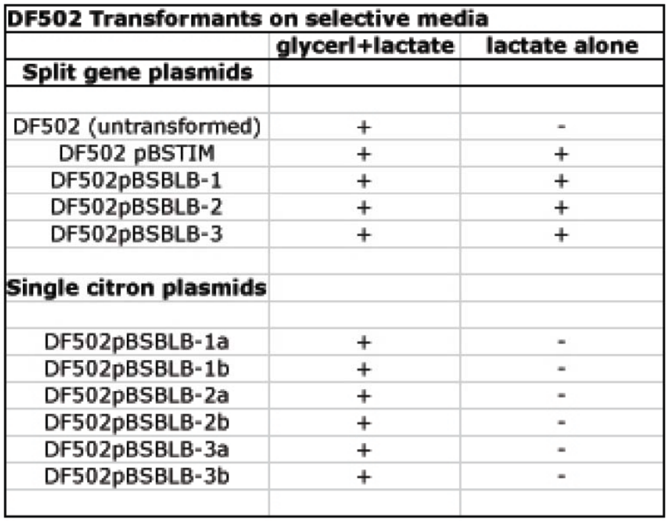


Figure 18. Schematic representation illustrating generation of "split" TIM genes. Split sequences containing stop codon-ribosome binding site-start codon were introduced at three different exon/intron junctions. The total length of the split sequences was 18nt or 17nts. The ribosomal binding site is 'AGGA'.

The *in vivo* selection for complemented TIM activity utilized the genetically modified *E. coli* strain - DF502. This strain lacks the endogenous TIM gene and is therefore unable to survive solely on either lactate or glycerol, that are precursors for glyceraldehyde-3-phosphate and dihydroxyacetone phosphate respectively. The DF502 strain can live on either lactate or glycerol only if it carries a plasmid expressing a functional TIM gene[12]. DF502 cells were transformed with engineered plasmids coding for split TIM proteins. Plasmids coding for one of the two TIM polypeptides were used as a control. After having supplemented the colonies with a required antibiotic and lactate on the LB agar plate, an analysis of the velocity of bacterial growth and the shape of the colonies was performed. Such data were used as indicators of *in vivo* enzymatic activity of complemented TIM. No colonies were observed after 10 days of incubation in the case of DF502 transformed with control plasmids. Five days following transformation all three DF502 cell lines expressing split TIM proteins produced viable colonies. The growth rates and shapes of the colonies obtained were compared to those obtained in DF502 transformed with plasmids coding for intact isomerases carrying point mutations. The *in vitro* activities of these isomerases was calculated in earlier studies upon purification of a protein[13]. In comparison, the positive control plates where DF502 cells were transfected with the plasmid coding for intact TIM enzyme only took 1 day to grow on the selective media LB agar plate (TABLE 1).



DF502 Transformants on selective media		
	glycerl+lactate	lactate alone
Split gene plasmids		
DF502 (untransformed)	+	-
DF502 pBSTIM	+	+
DF502pBSBLB-1	+	+
DF502pBSBLB-2	+	+
DF502pBSBLB-3	+	+
Single citron plasmids		
DF502pBSBLB-1a	+	-
DF502pBSBLB-1b	+	-
DF502pBSBLB-2a	+	-
DF502pBSBLB-2b	+	-
DF502pBSBLB-3a	+	-
DF502pBSBLB-3b	+	-

Table 1. Assessment of catalytic activity of TIM protein using reassembly of fragments. Two different medium conditions were used (glycerol+lactate, and lactate alone) to grow *E. coli* transformed with either plasmids bearing the coding sequence for split proteins (split gene plasmids) or plasmids bearing a coding sequence of one portion of a gene (single citron plasmids). Presence or absence of growth are indicated with + and – respectively.

However, since Western Blotting experiments performed demonstrated unequal expression of two polypeptides with at least a 10-fold difference, the correlation of growth rate could not be used to accurately determine specific catalytic activity. Thus, previous research could only conclude that the specific activity of the complemented TIM enzyme was on the order of the specific activity of a TIM enzyme carrying the H95N point mutation. With respect to the wild type enzyme, specific activity was estimated to be about 10^4 times lower.

Exon-coded polypeptide complementation of the chicken TIM enzyme is the first example in the field that addresses the question of generating functional proteins via complementation of peptide fragments defined by exon/intron boundaries. The results obtained clearly demonstrate the feasibility of such a complementation on the qualitative level. However, quantitative analysis was not conclusive due to difficulties with isolating individual fragments. Regardless of these drawbacks, this experiment provided significant support for the “exon as microgenes” theory.

2.3 Motivation to create a novel approach to protein engineering

The studies performed by the Beychok and Knowles groups provide compelling experimental data demonstrating that exon-coded polypeptides have the potential to function when not present in the context of a full-length protein. They also have the potential for creating functional enzymes via non-covalent complementation of polypeptide fragments defined by exon-intron boundaries. The observation that the cental exon-coded polypeptide of hemoglobin is capable of reversibly binding oxygen in the presence of the α-subunit provides additional support for the possibility that complementation of exon-coded polypeptides may have been nature’s way of evolving new proteins and new activities. The exon shuffling process is commonly agreed upon as one of the ways that nature searches for new catalytically active protein assemblies.

The studies discussed above are of crucial importance. Nonetheless, they are incomplete in that they do not include rationally designed negative control experiments to demonstrate that the observed complemented activities are in fact due to the exon-intron boundary within the context of the protein examined. Beychok’s paper studies only reassembly of exon-coded polypeptides coming from the β - globin gene, failing to investigate the influence of addition of the exon coded fragments of α-globin chain. The Knowles effort went deeper and analyzed three different exon/intron boundaries. However, these studies were performed via two-fragment complementation only and did not deliver a structural and enzymatic characterization of the constructs obtained. The ability of exon-coded polypeptides to reassemble in solution and form an active enzyme had to be demonstrated *in vitro*. Moreover, the lower enzymatic activities required further investigation. A potential hypothesis proposed was that ineffective protein folding leads to poor substrate binding. A proper negative control has to be used in order to establish the crucial role of the exon/intron boundary for reassembly experiments. While attempts to address these questions were done in the Knowles group, technical difficulties such as insufficient amounts of pure chicken TIM fragments, improper folding, and a questionable purity level hindered these studies in 1995.

It became evident that novel rational design had to be developed to engineer exon-coded polypeptide reassembled proteins. Such design should provide us with properly folded fragments that would allow for sufficient purification, and permit execution of a comprehensive study to correlate the importance of exon-intron boundaries in search of new and active proteins.

2.4 Design principles for linker mediated exon-coded polypeptide guided protein complementation

Several methods permitting protein complementation exist; these were discussed in detail in Chapter 1. Many methods rely on enzymatic cleavage of the purified full-length protein. Such an approach, as indicated by Knowles, could not be utilized to create protein fragments defined by exon/intron boundaries due to lack of proteases that would be required for the specific position of the cut. An alternative route is genetically fusing the exons that code for the protein fragments of interest to a secondary structure unit that would drive the assembly of the complementing fragments. However, these secondary structure units, such as a coiled coil motif or any other oligomerazing motifs[14], have the potential to interfere with recovery of catalytic activity. As demonstrated by Knowles study, the straightforward approach of separately cloning, expressing and purifying the exon-coded polypeptides is technically difficult due to problems with proper protein folding as well as purification. Charging protein sequences with His tag was recently developed to facilitate protein purification and it has been implicated in observed increased solubility of a test set of expressed proteins. Majority of these His tagged proteins were present in the cytoplasm of the host cells[15]. Fusion of a His tag could be an elegant way of generating soluble polypeptide fragments with minimal possibility of the tag interference with protein activity recovery. It is important to notice that fusion of individual exon-coded polypeptides of TIM to GST tag, which is similar to His tag resulted in increased solubility of engineered proteins, did not lead to production of increased amounts of properly folded fragments[5].

Ability to generate sufficient amounts of properly folded exon-coded polypeptides is crucial for our ability to perform comprehensive *in vitro* studies on protein activity recovery from exon-coded fragment reassembly. Protein folding is a complex matter; even now after at least 30 years of research we do not have a complete picture of all the rules and forces that govern it[16]. Through examination of multiple crystal structures of soluble proteins it has been shown that often times secondary structure units from one exon-coded polypeptide would be stabilized through interactions with amino acids coming from another exon-coded polypeptide. This is not to say that such interaction cannot be re-created once all the protein fragments re-assemble in solution, but in order to achieve the same interaction pattern as observed in full length protein each individual exon-coded polypeptide has to adopt same conformation on its own as it does in the contest of full length protein. Statistical analysis of the exon length distribution in *Homo sapiens* indicated a peaked distribution with a center around 105-120 nts[17], making an average exon-coded polypeptide of 35-40 amino acids in length. The question that now beckons attention is whether such length is sufficient for formation of folded polypeptides in solution? Depending on a sequence of a polypeptide secondary structures such as α-helices and β-turns and strands can be observed on the peptides of such lengths[18,19]. One of the stabilizing forces behind alpha-helix formation is the presence of the residues with side chains that can make hydrogen bonds with main chain amide or carbonyl groups when located at the beginning of the end of a peptide[20]. Given that on average at least 5 of exon-coded polypeptides would have to re-assemble to form an active protein, chances for such stabilization being present in all the protein fragments are rather low. This problem does not exist in the context of full-length protein as the N- and C- termini of corresponding polypeptides are covalently linked to the first and the last amino acids of the following and preceding exon-coded polypeptide respectively. Additionally, the entropic penalty that needs to be paid in order to bring five or more fragments together in solution would hinder fragment reassembly.

It has been noted that at least some of proteins follow the hierarchic folding pathway, meaning that folding begins with formation of structures that are local in sequence and marginal in stability, once formed these local structures interact to produce intermediates of increased complexity and stability and ultimately collapse into their native conformation[21]. α-lactalbumin, apomyoglobin, RNase H, barnase and cytochrome *c* all follow this hierarchic folding mechanism[22,23]. Such a mechanism implies that protein fragment folding preceded formation of the globular native fold. On the other hand, there is another set of proteins, known as class II within the protein folding community that do not follow such a mechanism and in this case tertiary interactions not only stabilize local structures, but actually determine them.

These observations imply that even though some of the protein fragments that represent simple units of secondary structure are capable of folding on their own, the final stabilization leading to formation of a native full-length globular protein fold involves interactions between these folded local structures For many other proteins formation of these local structures is actually secondary and is determined by tertiary interactions. Thus, an exon-coded polypeptide reassembly approach that allows for keeping all the fragments on the same covalent chain will not only remove the need to pay the entropic penalty for bringing multiple fragments together in solution but also has the potential for a greater probability of obtaining properly folded fragments. It would also facilitate our ability to generate substantial amounts of the desired constructs *in vivo* due to lower susceptibility of longer polypeptides chains towards protealytic degradation.

2.5 A novel approach to exon-coded polypeptide guided reassembly of active protein

To overcome the problems associated with improper protein folding and the entropy penalty for having to complex multiple polypeptide fragments we used a different approach: linking the exon encoded polypeptides with flexible linker. In our approach, a catalytically inert and flexible linker was used to covalently connect the exon-coded protein fragments. The length of the linker would have to be such that individual protein fragments are separated by sufficient distance to ensure their independent identities (Figure 19).

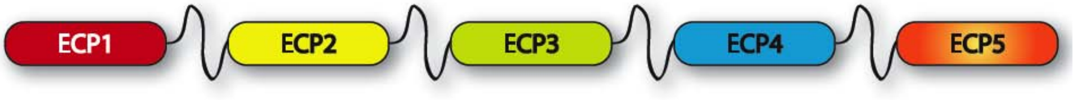


Figure 19. Schematic representation of a linker-mediated reassembled enzyme. Reassembly of exon-coded polypeptides is guided by the presence of a flexible linker. ECPs are exon-coded polypeptides of the corresponding number; the linker region is shown using curved black lines.

Use of linkers was previously demonstrated within studies aimed at generating novel enzymes though the fusion of domains belonging to different proteins[24]. Much like in case of protein fragment complementation, in some cases attempts at gene fusion have not been successful because of the instability or inability of proper folding post fusion of the enzymes. The inclusion of a linker between the respective pieces has been examined as a means to provide a flexible tether to join two enzyme domains. For example, citrate synthase has been fused to malate dehydrogenase by a flexible in frame three-amino acid, Gly-Ser-Gly linker[25]. Very little change in kinetic activities of each enzymatic domain was observed and this new construct was shown to be able to restore the growth ability of yeast that is deficient in both catalytic activities. A more detailed investigation of the nature and length of the linker that permits formation of an active fusion protein was done in the fusion of beta-galactosidase gene to galactose dehydrogenase gene[26]. Several linkers ranging from 3 to 13 amino acids were examined and the correlation between the length of the linker and the level of catalytic activity of each domain within fusion protein was established. Both *in vivo* and *in vitro* activity assays were used to characterize catalytic activity of domains. Lower stabilities and higher K_m values were demonstrated in the case of fusion proteins with short linkers, while optimal activity was achieved with a 9 amino acid linker. The observed kinetic properties nicely correlated with growth rate experiments with recombinant cells carrying plasmids encoding various fusion proteins using lactose as sole carbon source. It was observed that cells containing plasmids coding for hybrid enzymes with longer linkers grew slower (3-5%) than cells with hybrid enzymes with the shortest linker. The product of the coupled reaction from lactose by the hybrid enzymes, galactonolactone, represents the end product in *E. coli* metabolism. The formation of the lactone by these bifunctional enzymes instead of galactose-1-phosphate by endogenous galactokinase of the host cell would result in slower growth rate. Thus, the slower growth rate is indicative of more efficient transfer of the galactose produced to the galactosidase domain of a fusion protein in the cells with longer spacing between the catalytic business ends. Sequences of all of the linkers studied are rather complex and no comments on the possible influence of the linker side chains on enzymatic catalysis performed were made. No data are available to indicate presence of absence of the secondary structure for any of the linkers studied. The presence of the secondary structure in the linker may influence folding pattern of the fusion enzyme.

In the design of our linker enabled exon-coded polypeptide reassembly, we chose to use catalytically inert, flexible and unstructured linker. Use of such a linker would automatically eliminate the possibility of linker interference with catalysis as well as minimize the possibility of changes to the folding pattern of individual exon-coded polypeptides. For our study a Gly_7 linker fits these criteria and provides the desired flexibility between connecting protein fragments.

Short segments of polyglycine exhibit a strong tendency to be in an extended conformation in solution[27]. A model peptide system compromised of two different tripeptide units, Tyr-Glu-Ser and Ala-Thr-Asp, between which 1 to 18 glycine residues were inserted, was analyzed by NMR and small-angle X-ray scattering (SAXS). NMR studies of polyamino acids are frequently compromised by serious resonance overlaps. The constructs mentioned above has also showed severe overlaps in the Gly_n linker region, but resonances from the other six amino acid residues bordering the linker were nicely resolved. The backbone ^1H-^{15}N and $^1H_\alpha$-$^{13}C_\alpha$ residual dipolar couplings (RCDs) from these residues were measured at natural abundance level and compared between the various constructs with different linker lengths. It was demonstrated that RCD values of the two tripeptide units are insensitive to the variations in linker lengths, suggesting that the extension of the linker does not interfere with overall angular relationship of the two tripeptides, indicating the extended conformation of all of the various linker lengths probed. In was noted, however, that the solubility of AcYES-Gn-ATD was diminished in case when *n* was greater than 9. In order to confirm the extended conformation of the polyglycine linkers, dimensional characterization via small-angle X-ray scattering was performed. Such analysis allows one to get at the exact numbers for the length and the radius of gyration of the peptides in question, thus directly supporting the preference of polyglycine for an extended conformation. For example, the fitting analysis of the scattering profile of Ac-YES-G6-ATD yielded a radius of gyration and the maximum dimension of 9.1 Å and ~41 Å respectively (Figure 20).

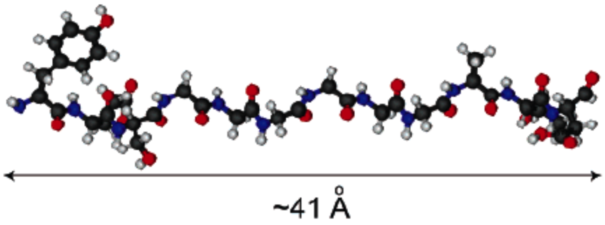


Figure 20. Graphic representation of Gly$_6$ linker in its extended conformation. Estimated length of entire peptide is ~41 Å including the bordering amino acids.

The average hydrodynamic diameter of a globular protein can be determined using different mathematical models[28,29]. These models utilize the X-ray data and use algorithms to arrive at the values for hydrodynamic diameters of proteins. For example, an ellipsoidal model is considered to be a classical approach for modeling the protein shape. Within this model, the protein is regarded as a prolate or oblate ellipsoid. Prior to wide availability of X-ray data, analytical methods such as analytical ultracentrifugation and intrinsic viscosities were routinely used to determine the conformation and shape of a protein. As a result of one of these studies, the diameter of a smaller globular protein of a 25.6 kDa size (chymotrypsinogen) was calculated to be 18.1 Å, while catalase, a protein of 230.3 kDa size measured at ~ 40 Å[28].

The length of the polyglycine part of described above Gly$_6$ linker is ~30 Å which would allow for sufficient spatial separation of the exon-coded polypeptides. The peptidic nature of the linker is experimentally favorable because it allows for a simple incorporation of this linker into the cDNA sequence of the studied gene. The introduction of additional stop or start codons is not required and the engineered protein is created as a single polypeptide chain coding for a predetermined number of polypeptide fragments defined by exon/intron boundaries covalently connected by Gly$_7$ linkers.

Design of negative controls for such a study is challenging task. A straightforward assumption would be to move the location of Gly_7 linker away from the exon/exon boundaries and if resulting protein were to not exhibit enzymatic activity it would explicitly demonstrate the crucial importance of the exon/exon boundary position of a linker for successful protein reassembly. In this optimistic scenario, the interpretation of results becomes relatively simple. However, in a situation when such negative control protein exhibits enzymatic activity even though position of the linker is no longer defined by exon/exon boundaries interpretation of results becomes much more complex. One could either argue that chances of recovering catalytic activity through protein fragment complementation are independent of the exon/exon structure of a protein of interest or that the ancestral gene coding for this protein used to have an intron in this position that was lost over the course of evolution. In our design linkers were moved away from the exon/exon boundaries by moderate distance. Chances that two introns were separated by less than 10 amino acids are rather slim, given that the exon length distribution peaks around 35-40 amino acids[17]. The Origin of Genes theory suggests that primordial time exon-coded polypeptides were probably only on the order of 15-20 amino acids long[30] and longer exons observed in modern time proteins were created by fusion of two or three shorter exons. But even considering 15 amino acids as the shortest number of amino acids representing an exon coded polypeptides, the shift of a linker position by 15 or fewer amino acids should still result in the introduction of a linker into the neighboring exon-coded polypeptide omitting the possibility of locating the linker to a potential primordial exon/intron boundary.

2.5 References

(1) Bornscheuer, U. T.; Pohl, M. *Current Opinion in Chemical Biology* **2001**, *5*, 137-143.

(2) Pohl, M. *Chemical Engineering and Technology* **2001**, *24*, 17-20.

(3) Seidel, H. M.; Pompliano, D. L.; Knowles, J. R. *Science* **1992**, *257*, 1489-1490.

(4) Craik, C. S.; Buchman, S. R.; Beychok, S. *Nature* **1981**, *291*, 87-90.

(5) Bertolaet, B. L.; Knowles, J. R. *Biochemistry* **1995**, *34*, 5736-5743.

(6) Sharonov Yu, A.; Sharonova, N. A. *Molecular Biology* **1975**, *9*, 119-140.

(7) Curtin, P.; Kan, Y. W. *Annals of the New York Academy of Sciences* **1989**, *565*, 1-12.

(8) Eaton, W. A. *Nature* **1980**, *284*, 183-185.

(9) Craik, C. S.; Buchman, S. R.; Beychok, S. *Proceedings of the National Academy of Sciences of the United States of America* **1980**, *77*, 1384-1388.

(10) Marchionni, M.; Gilbert, W. *Cell* **1986**, *46*, 133-141.

(11) Straus, D.; Gilbert, W. *Molecular and cellular biology* **1985**, *5*, 3497-3506.

(12) Fraenkel, D. G. *Annual Review of Biochemistry* **1986**, *VOL. 55*, 317-337.

(13) Hermes, J. D.; Blacklow, S. C.; Knowles, J. R. *Proceedings of the National Academy of Sciences of the United States of America* **1990**, *87*, 696-700.

(14) Furman, J. L.; Badran, A. H.; Shen, S.; Stains, C. I.; Hannallah, J.; Segal, D. J.; Ghosh, I. *Bioorganic and Medicinal Chemistry Letters* **2009**, *19*, 3748-3751.

(15) Loughran, S. T.; Walls, D. *Methods in molecular biology (Clifton, N.J.)*, *681*, 311-335.

(16) Anfinsen, C. B. *Science* **1973**, *181*, 223-230.

(17) Peng, L.; Sakharkar, K. R.; Sakharkar, M. K. *International Journal of Integrative Biology* **2009**, *5*, 87-91.

(18) Munoz, V.; Serrano, L. *Current Opinion in Biotechnology* **1995**, *6*, 382-386.

(19) O'Neil, K. T.; DeGrado, W. F. *Science* **1990**, *250*, 646-651.

(20) Shi, Z.; Olson, C. A.; Bell Jr, A. J.; Kallenbach, N. R. *Biopolymers - Peptide Science Section* **2001**, *60*, 366-380.

(21) Baldwin, R. L.; Rose, G. D. *Trends in Biochemical Sciences* **1999**, *24*, 26-33.

(22) Matouschek, A.; Kellis Jr, J. T.; Serrano, L.; Fersht, A. R. *Nature* **1989**, *340*, 122-126.

(23) Baldwin, R. L.; Rose, G. D. *Trends in Biochemical Sciences* **1999**, *24*, 77-83.

(24) Hakansson, M.; Linse, S. *Current Protein and Peptide Science* **2002**, *3*, 629-642.

(25) Lindbladh, C.; Brodeur, R. D.; Lilius, G.; Bulow, L.; Mosbach, K.; Srere, P. A. *Biochemistry* **1994**, *33*, 11684-11691.

(26) Carlsson, H.; Ljung, S.; Bolow, L. *Biochimica et Biophysica Acta - Protein Structure and Molecular Enzymology* **1996**, *1293*, 154-160.

(27) Ohnishi, S.; Kamikubo, H.; Onitsuka, M.; Kataoka, M.; Shortle, D. *Journal of the American Chemical Society* **2006**, *128*, 16338-16344.

(28) GarciÃÅa De La Torre, J.; Huertas, M. L.; Carrasco, B. *Biophysical Journal* **2000**, *78*, 719-730.

(29) Yan, B. C.; Yan, J. F. *International Journal of Biological Macromolecules* **1999**, *24*, 65-67.

(30) Gilbert, W.; De Souza, S. J.; Long, M. *Proceedings of the National Academy of Sciences of the United States of America* **1997**, *94*, 7698-7703.

Chapter 3

CONVERTING HUMAN ADENYLATE KINASE INTO PRIMORDIAL ENZYME

3.1 Choice of target protein

To aid studies of protein evolution and gene prediction research, Saxonov and Gilbert have created an extensive Exon-Intron database (EID)[1]. The database was generated based on the sequencing data from eukaryotic organisms provided in GenBank release 112. It contains 51289 protein-coding genes with characterized gene structure, resulting in total of 287209 exons present. Along with the vast amount of sequencing data it also contains an extensive amount of information about each gene, such as its' DNA and protein sequence as well as splicing site motifs. Due to inherent redundancy (~17%) carried over from GenBank, purging of the sequences with 99% identity level reduced the database size down to 42460 genes or 243,89 exons. Additional subdatabase containing exons with experimentally confirmed exon/intron boundaries had been established by means of comparison between genomic DNA and mRNA of each gene. This subdatabase was smaller in size and was composed of only 11242 genes or 62474 exons.

Given the significant number of genes that have experimentally characterized exon/intron gene structures and thus are perfectly suitable for analysis though our linker mediated exon-coded polypeptide reassembly approach an additional subset of filters had to be created to allow for selection of the best protein targets.

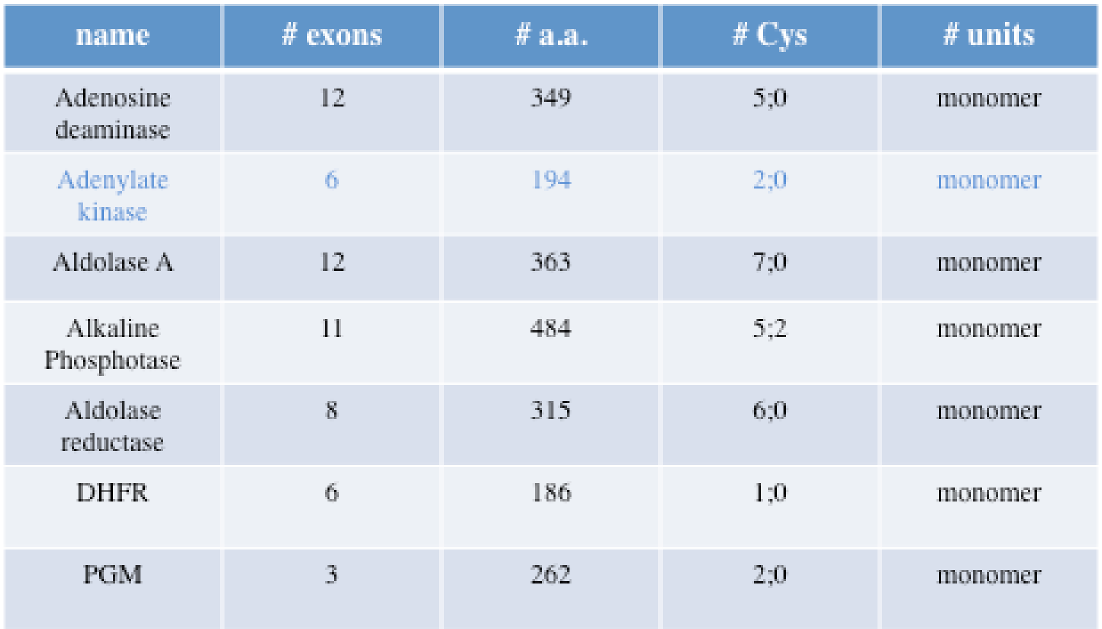
Firstly, the protein had to be of ancient origin as determined by a high amino acid homology level (≥ 75%) between eukaryotic and prokaryotic versions of protein in question. As discussed in detail in Chapter 1, origination of new genes based on exon-shuffling is heavily dependent on the intron phase and only introns that do not interfere with the open reading frame (phase 0 introns) of shuffled exons would be useful in creating new primordial genes. Overrepresentation of phase 0 introns as well as symmetrical exons have been shown to be more pronounced in the case of ancient proteins, whose origination predates divergence of prokaryotes and eukaryotes. The subset of ancient genes was separated into yet another subdatabase by Saxonov *et al.* Generated database contains an even smaller number of genes thus significantly decreasing the number of possible candidates for our study.

Secondly, we imposed the stipulation that the crystal structure of the protein would have to be determined and available through The Protein DataBank (www.pdb.org) in order to assess the location of the exon/exon boundaries on the protein surface. It was noticed by Knowles during their analysis of complementation of TIM fragments defined by exon/intron junctions that all of the boundaries are located to the protein surface[2]. Such localization of exon/exon protein boundaries could be the rudimentary signature of ancient time protein assembly and organization. If it is true that exon-coded polypeptides were produced independently from the exons or microgenes and then assembled non-covalently in solution their charged C- and N-termini would be best stabilized if located at the protein surface, thus solvent exposed. Thus it is very useful to be able to have a handle on the location of the exon/exon protein boundary of the protein under investigation.

Additionally, the total number of exons present in the gene or in general the length of a protein had to be limited to avoid technical difficulties that may be faced during synthesis and expression of large genes in a host cell line. An additional limitation was applied on the oligomeric state of an enzymatically active protein. We envisioned that the target protein had to be functional as a monomer; otherwise the number of fragments that had to reassemble in solution would be artificially increased, thereby increasing the entropic penalty of complementation. Such an increase in entropic penalty would be observed even in case of our experimental design, because despite the fact the exon-coded polypeptides are located on the same covalent chain, if the protein were active as a dimer two of those chains would have to interact in solution.

Finally, the number of inter-disulfide bonds was used as a filter to screen for an optimal protein candidate. Formation of multiple disulfide bonds within the protein scaffold could pose an additional challenge in obtaining properly folded exon-coded polypeptides. Such a difficulty in folding may arise due to improper disulfide bond formation. This phenomenon was observed during complementation studies of hen egg lysozyme performed by Bradley Crothy as well as Dr. Pamuk-Turner in our lab. Thus, in our study we aimed at using a protein that possessed no disulfide bonds. With these considerations, the presence of cysteines was allowed as long as they did not participate in the formation of disulfide bonds.

A scan of the IED subdatabase of genes coding for ancient proteins though these additional filters resulted in only a limited number of potential protein candidates. Information about their compliance with these filters is summarized in Table 2.



name	# exons	# a.a.	# Cys	# units
Adenosine deaminase	12	349	5;0	monomer
Adenylate kinase	6	194	2;0	monomer
Aldolase A	12	363	7;0	monomer
Alkaline Phosphotase	11	484	5;2	monomer
Aldolase reductase	8	315	6;0	monomer
DHFR	6	186	1;0	monomer
PGM	3	262	2;0	monomer

Table 2. Resulting list of protein candidates for linker mediated protein reassembly. Short summary on the hits obtained for each protein when passed through filters is shown. # exons – indicates total number of exons present in the gene, #a.a. is the length of protein in amino acids, # cys provides information on both total number of cys residues within protein sequence (first number) and the number of cys residues involved in disulfide bond formation (second number), # units represent enzymatically active oligomeric state of the protein

Out of six potential candidates, we chose human adenylate kinase 1, which is an ancient enzyme with high (>75%) homology between eukaryotic and prokaryotic protein analogs of relatively short length (194 nt, 5 exons), that contains two cysteine residues not involved in formation of disulfide bonds within protein scaffold and has a well established crystal structure. The exon/intron structure of this gene has been experimentally determined and was used as a guide to placing Gly$_7$ linker.

3.2 A closer look at a target: human adenylate kinase 1

Adenylate kinases (AKs) have a very important role in nucleotide metabolism in all organisms by means of phosphotransfer networks[3]. AKs catalyze the reversible transfer of the γ-phosphate group from a phosphate donor (normally ATP) to a phosphate acceptor (normally AMP), with the concomitant formation of two molecules of ADP. In addition to playing a crucial role in the homeostasis of adenine nucleotide metabolism, AKs are involved in cellular energetics through complex phosphotransfer networks regulating intracellular ATP-producing processes[4]. There are six currently known isoforms of human adenylate kinase that are characterized by their specific tissue distribution. AK1, the target of our investigation, is a cytoplasmic enzyme mainly expressed in skeletal muscle, brain and erythrocytes. AK2 is located both in the cytosol and mitochondrial space and is expressed mainly in the liver, kidneys, and to a lesser extent, in the spleen and heart. AK3 is ubiquitously expressed in all tissues, whereas AK4 is mainly expressed in the kidneys, heart and liver. AK5 is found in brain tissue and is a cytosolic enzyme. Finally, the recently discovered AK6 is a nuclear enzyme; the pattern of its tissue distribution is yet to be determined[5].

The AK1 isoform of adenylate kinase is highly conserved and shares ≥75% of sequence identity among AK1 sequences from human, mouse, rat, yeast and *E. coli* adenylate kinase. AK1 deficiency in erythrocytes is a rare genetic disorder associated with hemolytic anemia. More recently it was demonstrated that AK1 gene deletion in mice leads to disruption of muscle energetic inventory despite metabolic rearrangements implemented by mice to compensate for AK1 deletion[4]. Human AK1 (EC 2.7.4.3) is a ubiquitous monomeric enzyme of 194 aa length[6]. The gene coding for this protein was cloned and characterized by Nakazawa's group in 1989[7]. AK1 enhances the reaction between AMP and MgATP by a factor of >10^{12} and is a nearly perfect enzyme with $k_{cat}/K_m \geq 10^7$ [$s^{-1}M^{-1}$][8].

A large body of work has been informative in elucidation of the exact enzymatic mechanism utilized by the enzyme to perform the phosphate transfer reaction. An iso-random bi-bi mechanism for phosphate transfer by adenylate kinase from rabbit muscle was originally proposed by Sheng *et al* [9]. The suggested model involves the binding of the substrates, AMP and MgATP in no specific order, by one of the forms of enzyme whereas the other conformational form of enzyme can bind two molecules of MgADP. Substrate inhibition by both AMP and MgATP is well established for adenylate kinase in the high concentration range of substrates: 1mM and higher based on the termination of linear dependence between 1/v and 1/[S] in the high concentration range. The transition from one form of the enzyme to another is accompanied by a drastic conformational change, going from an open state of the protein indicative of its ability to bind substrates to a closed form with the active site residues involved in binding. Such a conformational change is possible due to the presence of a P-loop motif, common to ATP and GTP binding proteins. The P-loop is located close to the N-terminus situated between an α-helix and a β-strand and has a conserved sequence of GXPGXGKGT (FIGURE 21). During the conformational change upon binding of substrates, changes in the P-loop allow for formation of a closed form of an enzyme[10].

Ten years later the presence of a separate binding site for Mg^{2+} ions was established in *E. coli* adenylate kinase, transitioning the well accepted random bi-bi model of enzymatic action into what is commonly referred to as modified random bi-bi model. A comprehensive study of the forward reaction of adenylate kinase revealed that previously the unreported Mg^{2+} dependence allows for a 23±3-fold increase in the reaction rate[11]. This finding not only clarified the details of enzymatic mechanism of adenylate kinase action but also raised an important question of the role of divalent magnesium ions in the regulation of energy homeostasis.

The human AK1 gene contains 5 exons and 4 introns that vary in length, with the third intron being the largest, spanning 3228 nt[12]. The crystal structure of human AK 1 in complex with bis-adenosine-pentaphosphate was resolved and is in the Protein Databank with an accession number 1z83 (FIGURE 21).

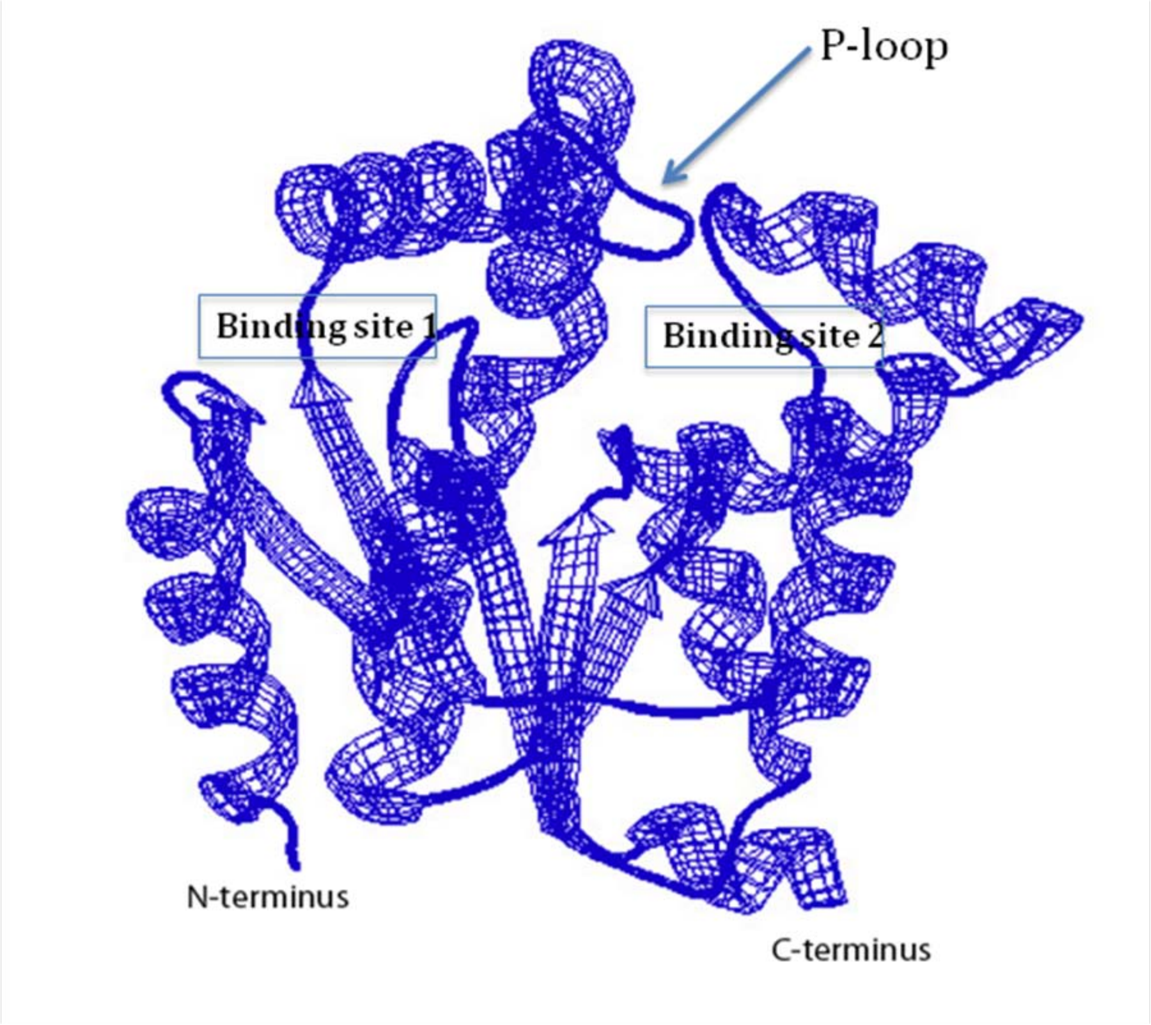
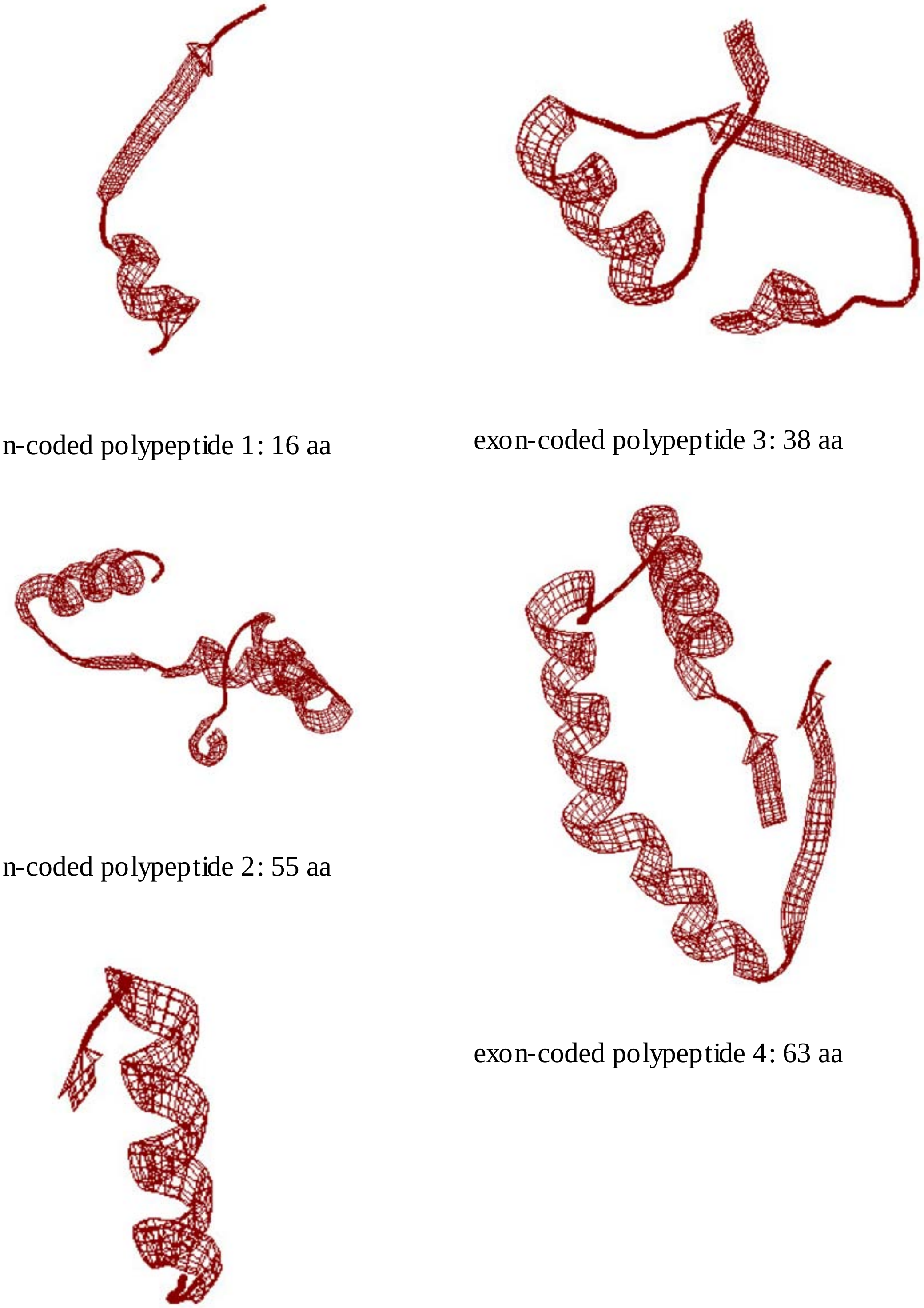


Figure 21. Cartoon ribbon mesh diagram of human Adenylate kinase 1 (PDB ID 1z83) generated using SwissPDB viewer. The N- and C-termini are indicated. Note the location of P-loop and proposed substrate binding sites based on the binding position of bis-adenosine-pentaphosphate.

Since the information about exon/intron organization of the gene is known, we were able to analyze the exon-coded polypeptide distribution along the protein sequence. All five exon-coded polypeptides represent a set of secondary structures: such as α-helices and β-strands (Figure 22).



exon-coded polypeptide 1: 16 aa

exon-coded polypeptide 3: 38 aa

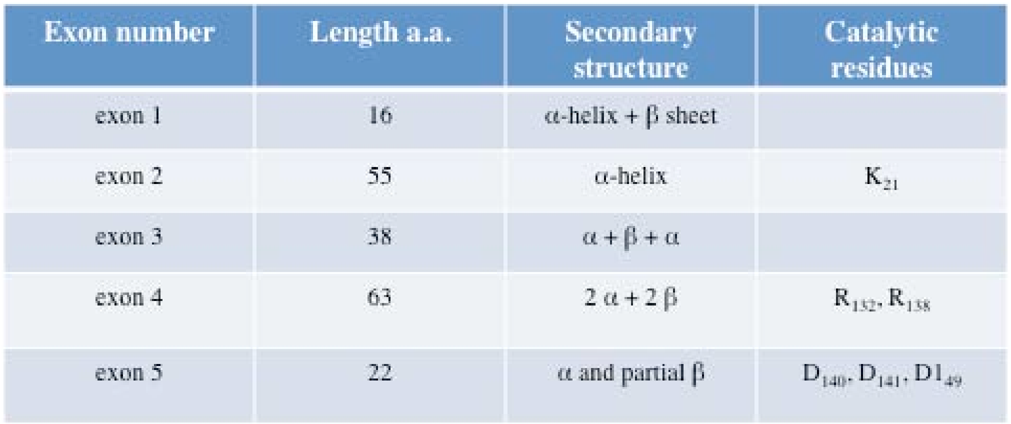
exon-coded polypeptide 2: 55 aa

exon-coded polypeptide 4: 63 aa

exon-coded polypeptide 5: 22 a

Figure 22. Schematic representation of five exon coded polypeptides of human AK1. The crystal structure of human AK1 was used and portions of amino acid sequence corresponding to exon-coded polypeptides are highlighted in red.

In all but one case (exon4/exon5 boundary) the location of exon/exon boundary falls in between the units of secondary structure. The termini of exon-coded polypeptides much as in the case of TIM exon defined fragments are located at the protein surface. Structural analysis of exon-coded polypeptides is summarized in a Table 3.



Exon number	Length a.a.	Secondary structure	Catalytic residues
exon 1	16	α-helix + β sheet	
exon 2	55	α-helix	K_{21}
exon 3	38	α + β + α	
exon 4	63	2 α + 2 β	R_{132}, R_{138}
exon 5	22	α and partial β	D_{140}, D_{141}, $D1_{49}$

Table 3. Structural analysis of human AK1 exon coded polypeptides. Length of each exon coded polypeptide as well as the secondary structure composition is indicated. Distribution of catalytic residues potentially involved in binding AMP and ATP along the exon-coded polypeptides is summarized in last column[12,13].

3.3 Design and generation of constructs

The nucleotide sequence as well as protein sequence for human AK1 gene is readily available through the NCBI and is published elsewhere[7,6]. A simple in-frame addition of 21 nucleotides coding for seven Gly residues in the desired exon/exon juncture within cDNA sequence of human AK1 would produce a chimeric gene coding for exon-defined polypeptide fragments separated by ~30 Å by the Gly_7 linker[14].

In our study, we designed two fusion genes. The first gene contained a sequence coding for Gly_7 at the exon 3/exon 4 junction leading to production of a gene referred to as G1. Presence of the Gly_7 linker led to generation of two adenylate kinase fragments defined by the exon 3/exon 4 boundary. Addition of a second linker coding sequence at the exon4/exon5 boundary generates a gene containing two Gly_7 linkers. The two linkers separate wild-type human AK1 into three complementing fragments defined by exon 3/exon 4 and exon 4/exon 5 boundaries (FIGURE 23).

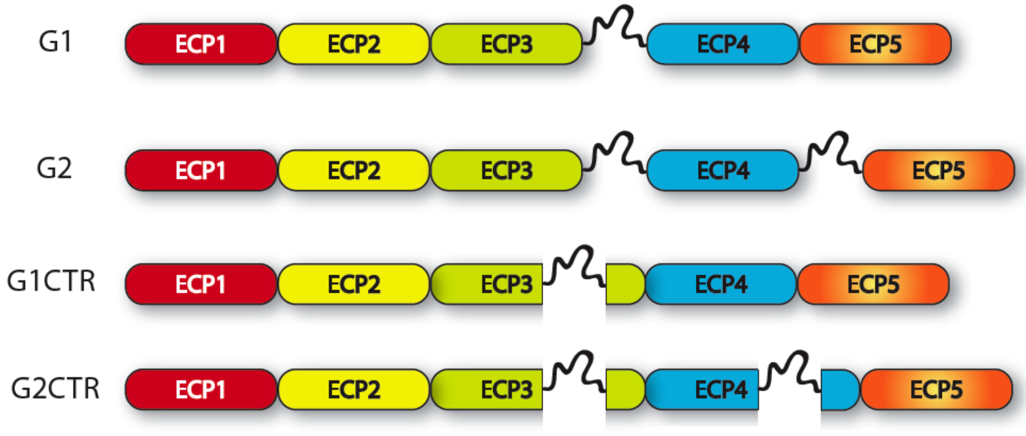


Figure 23. Engineered human adenylate kinase 1 fusion proteins. Different exon coded polypeptides are highlighted in different color. Gly_7 linker is shown as black wiggly line introduced at the end of exon-coded polypeptide 3 in case of G1 fusion protein as well as at the end of exon coded polypeptide 4 in case of G2 fusion protein. Both control proteins (G1CTR and G2CTR) have their linker positions changed, their positions are no longer defined by exon/exon boundary of human adenylate kinase.

The third fusion protein containing the Gly_7 linker at the exon2/exon3 boundary as well as exon 3/exon 4 and exon 4/exon 5 boundaries was generated. However, conditions for production of this protein in the host *E. coli* cell line were not identified. Limited number of variations to standard protein expression conditions was tested to achieve acceptable protein expression levels, such as temperature, concentration of inducing reagent IPTG, and duration of overexpression. None of the tested conditions resulted in successful protein expression. The boundary between exon1 and exon was never split by addition of the linker, due to the small size of the exon coded polypeptide 1. We anticipated potential difficulties with folding for a 16 amino acid long peptide.

Negative controls were generated to test the importance of the exon/exon boundary for successful reassembly. In order to address this question the Gly_7 linker position on the cDNA of human AK1 was changed in a way that the location of the Gly_7 coding sequence was no longer dictated by exon/exon boundary on cDNA. Both linkers were moved away from the exon/exon boundary by several amino acids (4 in case of a linker originally incorporated between exon 3 and exon 4 and by 23 amino acids in case of a linker incorporated between exon 4 and exon 5). The resultant control genes are named G1CTR and G2CTR.

All the generated cDNA sequences, including that of wild type human adenylate kinase 1 were submitted to Integrated DNA Technologies (IDT DNA) for synthesis and later cloned into pET11a and pET30a Xa/LIC vector, allowing for protein expression in a *E. coli* [BL21pLysS(DE3)] host cell line. Use of the pET30a Xa/LIC vector system not only allowed for high efficiency cloning, but also permitted the use of affinity based (His-tag) chromatography for facile purification of the engineered proteins, as the His tag coding sequence is incorporated in the expression portion of the plasmid. An additional advantage confirmed on our system by use of the His tag is the increase in solubility of the designed constructs. Such an influence of His tag on solubility of engineered is reported in the literature[15]. The rather small size of the tag relative to the length of the fusion protein, has minimal potential to interfere with the activity of reassembled protein. The used of the pET11a cloning system results in production of naked fusion proteins, referred to as G1.3 (tagles analog of G1) and G2.1 (tagles analog of G2), thus eliminating the potential complication of catalysis interference due to presence of taged amino acids. However, such gene design greatly complicates fusion protein purification strategy.

Since the BL21pLysS(DE3) host cell line was used to overexpress the engineered fusion proteins, we had to ensure the absence of *E. coli* adenylate kinase in our protein preparations. Presence of *E. coli* adenylate kinase could lead to false positive read outs in enzymatic activity assays. An additional anion exchange chromatography step was utilized for protein purification. This purification step takes advantage of the drastic difference in the isoelectric point of human AK1 and *E. coli* AK. The estimated pI of human AK1 is 8.9 while *E. coli* AK has a pI of 5.3. The addition of a Gly_7 sequence does not alter the pI of engineered protein, thus allowing us to use this purification step uniformly on all engineered proteins, both the taged and tagless versions.

Two different purification strategies were developed. The fusion protein cloned into pET11a produced tagless constructs that were first partially purified through size exclusion chromatography (fractionation range is 10,000-200,000 Da). It allowed to separate the target proteins from the higher molecular weight proteins of the host cells. Fractions containing proteins of interest were pooled together and further purified using anion exchange column to ensure removal of *E. coli* AK. Fusion proteins generated using the pET30a Xa/LIC vector result in fusion proteins containing tag sequence (His_6) on their N-termini. Presence of the His tag was utilized as it permitted affinity purification. Since wild type *E. coli* AK does not have a His tag in its sequence this purification step also facilitated the removal of undesired contaminant. Denaturing affinity purification was the first step applied to purify His-tagged fusion proteins trapped in inclusion bodies within the host cells, followed by anion exchange chromatography to further ensure absence of *E. coli* AK in protein preparations. Native affinity purification was utilized as a last purification step to increase the overall purity level of engineered proteins.

Final purity levels of fusion proteins preparation were tested using Western Blot with a primary *E. coli* AK antibody to ensure there was no *E. coli* AK contamination. The lowest detection limit of *E. coli* AK was established at 0.5 ng of protein through generation of a standard curve for *E. coli* AK protein detection using Western Blot methodology (FIGURE 24).

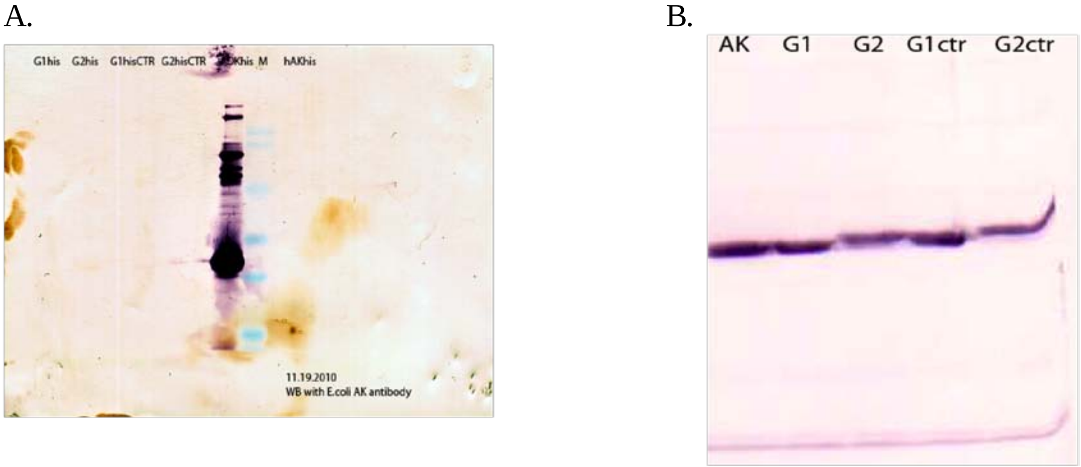


Figure 24.Purity levels and identity of fusion proteins established by Western Blot analysis. A. Western Blot analysis of purified engineered human AK 1 proteins using *E. coli* AK primary antibody. **B.** Western Blot analysis of engineered human AK1 proteins using human AK1 primary antibody.

None of the purified protein constructs contained detectable amounts of *E. coli* AK and thus can be assigned to contain <0.5 ng of *E. coli* AK. When assayed for catalytic activity, 0.5 ng of *E. coli* AK exhibited no detectable phosphoryl transfer activity. On the other hand, all designed constructs demonstrated strong interactions on the Western Blot with primary human AK1 antibody. *E.coli AK* was shown to not interact with primary human AK1 antibody (FIGURE 25).

A.

B.

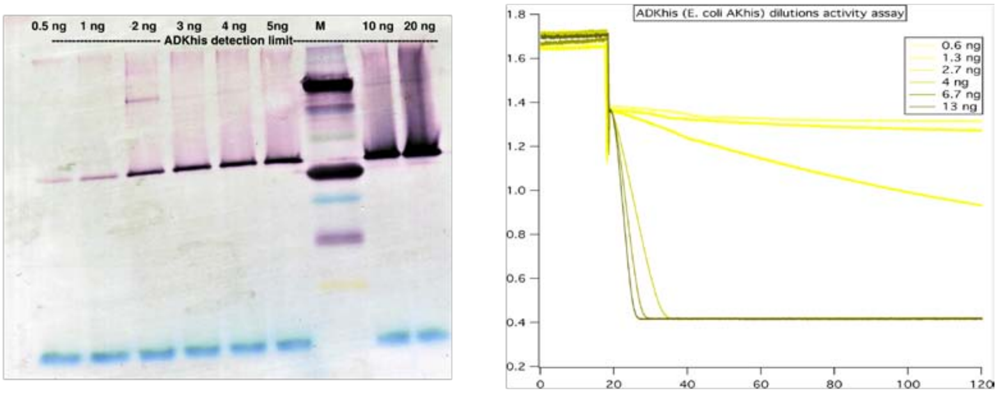


Figure 25. Characterization of *E. coli* AK enzymatic activity and detection limits. A. Western blot analysis of various amounts of purified *E. coli* AK protein using *E. coli* AK primary antibody to establish lower detection limit of *E. coli* AK protein in engineered protein preparations. **B.** Enzymatic activity assay of different *E. coli* AK amounts.

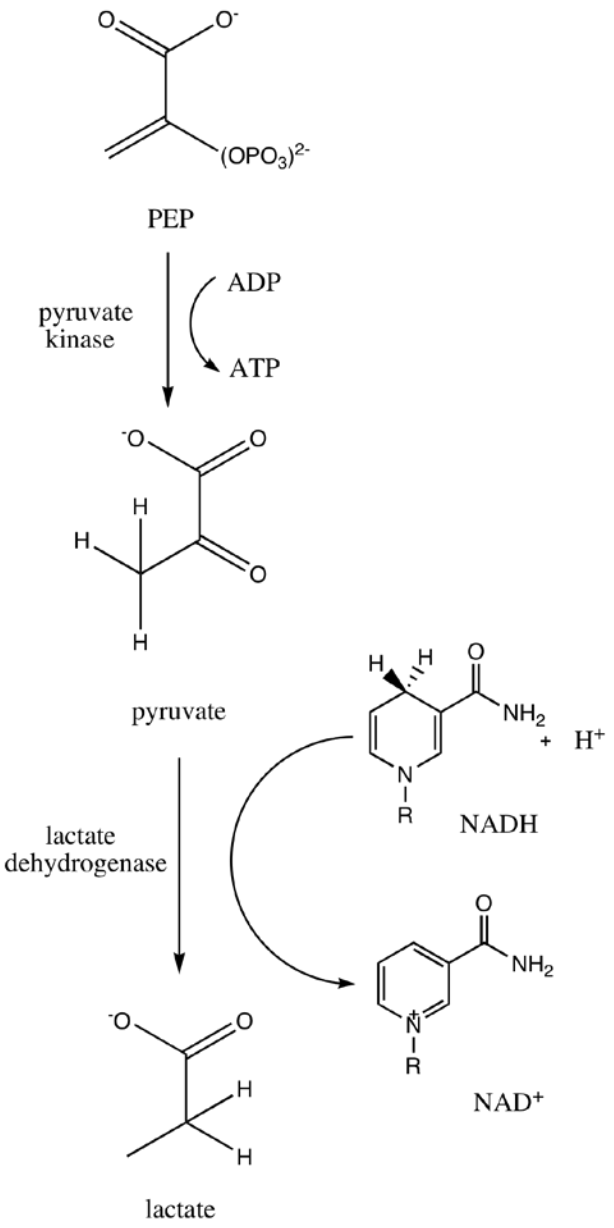
3.4 Functional and Structural characterization of engineered proteins *in vitro*

Our design strategy allowed us to produce relatively large amounts of protein constructs, thereby permitting both structural and functional characterization of the fusion proteins. Such analysis was hindered by quantities of engineered proteins in earlier studies by Knowles and Beychok[16,2]. Functional characterization focused on establishing leves of enzymatic activity exhibited by fusion proteins in comparison to wild-type human adenylate kinase 1. A glance at a change in binding affinities of both substrates (AMP and MgATP) among fusion and wild type enzymes yielded data that allowed us to estimate structural changes brought upon the protein structure due to the introduction of linker. Additional structural characterization using circular dichroism permited us to estimate the overall folding of the engineered proteins.

3.4.1. Functional characterization of engineered proteins in vitro

Enzymatic activity

The enzymatic activity of engineered proteins was assessed by using a coupled spectrophotometric ATPase assay. The ATPase assay was first described by Kreuzer and Jongeneel and later further modified[17]. ATPase assay itself is a sequence of two reactions. In the first step pyruvate kinase converts one molecule of phosphoenolpyruvate (PEP) into pyruvate, while also converting one molecule of ADP to ATP. In the second step lactate dehydrogenase converts one molecule of pyruvate to molecule of lactate while oxidizing NADH to NAD^+. Oxidation of NADH can be easily monitored as decrease in absorbance at 340 nm (Scheme 1).



Scheme 1. Coupled spectrophotometric assay. ADP production due to activity of engineered fusion proteins in coupled to conversion of phosphoenolpyruvate to pyruvate using pyruvate kinase. Produced pyruvate is converted into lactate using lactate dehydrogenase, which utilizes NADH as a cofactor to carry out this transformation. During pyruvate to lactate conversion NADH gets oxidized to NAD^+, which is accompanied by decrease in absorbance at 340nm.

It is the second chemical conformation in the first coupled step (conversion of PEP to pyruvate) that is coupled to the enzymatic assay of a protein in question. Only in the presence of the active reassembled protein or a wild-type adenylate kinase would there be ADP production. Without produced ADP, pyruvate kinase is unable to convert phosphoenolpyruvate into puryvate, thus preventing the initiation of cascade reactions. Oxidation of NADH is used as a read-out in the assay. The rate of change in absorbance at 340nm is directly correlated to the rate of ADP produced by the engineered enzyme.

Enzymatic activity assays were performed with wild-type human AK1 and all the fusion proteins: G1 and G2 (tagless and taged versions), G1CTR and G2CTR. Lysozyme, various concentrations of imidazole and formulation buffer served as the assay's quality and specificity controls. Enzymatic activities of all proteins were determined at 37 °C. Relative specific activities were calculated according to Pan *et al.* [9]:

$$\text{Specific activity} = \nu \times V/m,$$

where ν is the slope, V the reaction volume and m the mass of the enzyme in Da.

Given that definition of specific activity, relative specific activities were determined as ratio of specific activity of protein A over specific activity of protein B:

$$SA_1/SA_2 =(\nu_1 \times c_2 \times MW_2)/(\nu_2 \times c_1 \times MW_1)$$

where c is the concentration of a protein determined using Lowry assay, MW is the molecular weight of the proteins in question in Da.

Both G1his and G2his proteins were found to be active *in vitro* and their relative specific activities, as compared to the activity of human AK, were 1/2 and 1/9 respectively (Figure 26).

Figure 26. Graphic representation of generated enzymatic curves of fusion proteins. The decrease of absorbance at 340nm was monitored after reaction was initiated by addition of indicated fusion protein in case of each shown trace. The highest enzymatic activity correlates with the steeper slope. Concentrations of all fusion proteins used in assay were kept constant.

Control proteins were found to be inactive as no detectable decrease in absorbance was observed. A control experiment in which AMP was omitted from the mixture indicated no detectable activity exhibited by either of the constructs or the wild-type enzyme; this indicates that no non-specific ATP hydrolysis was occurring to produce ADP.

In an experiment where ATP was omitted no decrease in absorbance was observed, indicating that there was no P_i binding in the ATP binding site, which could lead to potential phosphate transfer onto AMP and subsequent ADP formation (Figure 27).

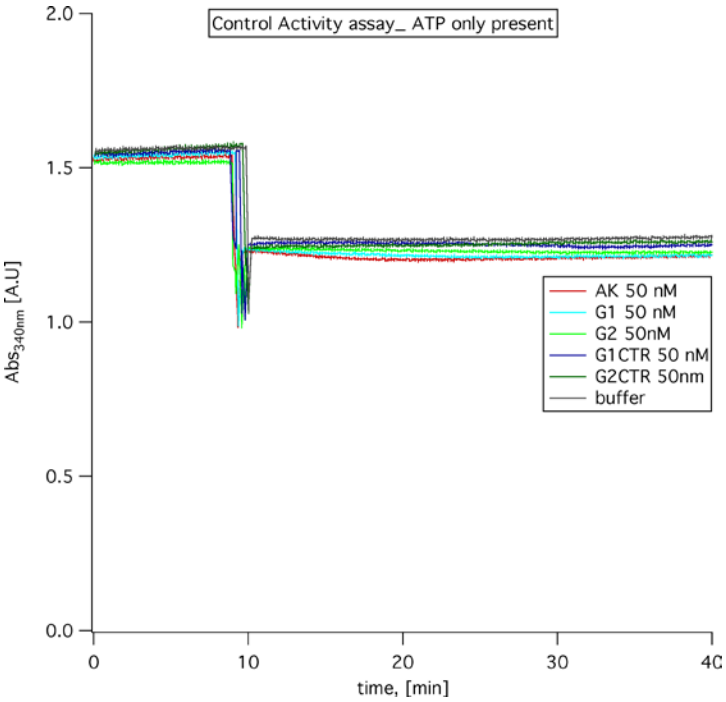


Figure 27. Graphic representation of generated enzymatic curves of fusion proteins in the absence of AMP. The noticeable absence of decrease in 340nm absorbance is indicative of enzymatically inert reaction.

Together these data strongly indicate that location of the Gly_7 linker is crucial to allow successful protein reassembly as no detectable activity was observed in the case of control proteins G1CTR and G2CTR. Relatively small decrease in activity was observed for engineered exon-coded-polypeptide complementing proteins, indicating successful protein reassembly via complementation of fragments defined by exon boundaries.

The untaged versions of fusion proteins were assayed for activity next. The observed relative specific activities were lower compared to wild-type human adenylate kinase as well as the His-tagged fusion proteins. Approximately, a 60-fold decrease in relative activity was observed in the case of G1 and G2 proteins (Figure 28).

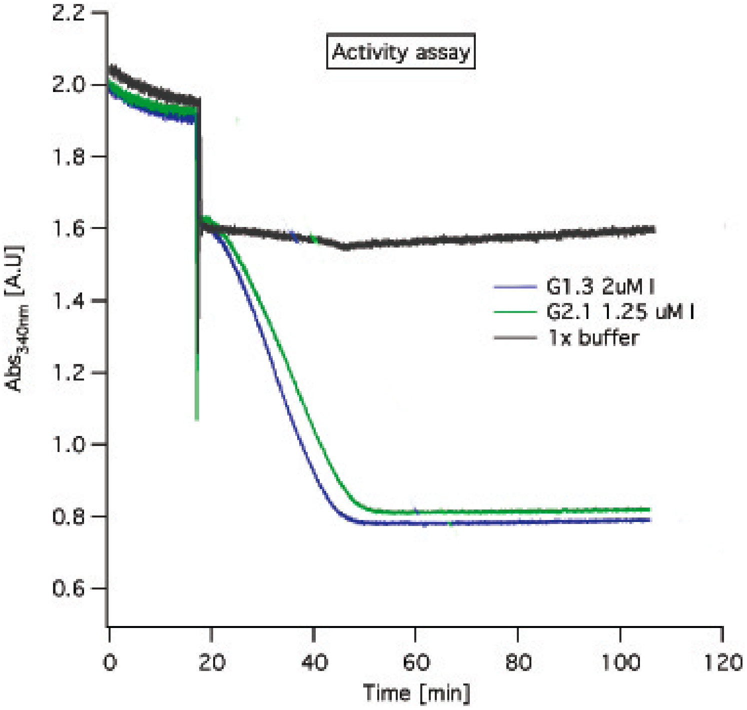
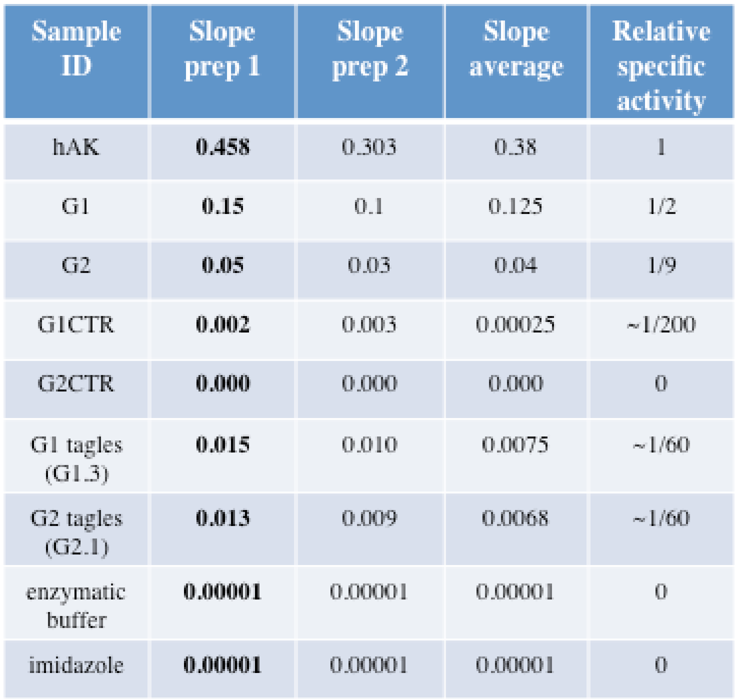


Figure 28. Graphic representation of generated enzymatic curves of the tagles fusion proteins. The decrease of absorbance at 340nm was monitored after reaction was initiated by addition of indicated fusion protein in case of each shown trace. The highest enzymatic activity correlates with the steeper slope. Relative specific activities were calculated in respect to the activity of full-length adenylate kinase observed in set of experiments performed for tagged fusion proteins.

These proteins suffered from poor solubility and it was technically challenging to manipulate protein solutions as they were found to be stable for only a few days as determined by decrease in the protein band intensity by SDS-PAGE gel. Dissolution of lyophilized proteins did not allow proper folding as significant precipitation was observed during dialysis. Due to these experimental challenges, we did not pursue the characterization of tagless protein constructs further and concentrated our efforts on the His-tagged fusion proteins.

Activities of fusion proteins as well as wild-type human adenylate kinase changed slightly from one protein preparation to another (± 30%), which is expected for a protein purified after overexpresion. Summary of obtained enzymatic slopes for two different protein preparation is presented in a Table 5 below.



Sample ID	Slope prep 1	Slope prep 2	Slope average	Relative specific activity
hAK	**0.458**	0.303	0.38	1
G1	**0.15**	0.1	0.125	1/2
G2	**0.05**	0.03	0.04	1/9
G1CTR	**0.002**	0.003	0.00025	~1/200
G2CTR	**0.000**	0.000	0.000	0
G1 tagles (G1.3)	**0.015**	0.010	0.0075	~1/60
G2 tagles (G2.1)	**0.013**	0.009	0.0068	~1/60
enzymatic buffer	**0.00001**	0.00001	0.00001	0
imidazole	**0.00001**	0.00001	0.00001	0

Table 4. Variations in enzymatic slopes observed between two protein preparations. Different protein preparations indicate that proteins were overexpressed and purified on separate days as a part of accessing the reproducibility of the enzymatic reassembly. Raw reaction slopes measured using IgorPro software and normalized to represent same enzymatic concentrations are shown and converted over to relative specific activities where activity of wild type AK1 is assigned to be 1.

A quick glance at a change in substrate binding

K_m values for both substrates of human AK1 (AMP and MgATP) were determined using the coupled spectrophotometric assay. It has been previously demonstrated that adenylate kinase utilizes a complex mechanism with binding of three substrates taking place independently. It is worth noticing however, that these studies were performed on *E. coli* adenylate kinase and have not been shown to hold true in case human adenylate kinase 1. In a thorough kinetic analysis of active protein reassemblies (G1 and G2) concentration of all three substrates would have to be varied at the same time followed by fitting of obtained kinetic data to the complex model describing binding of all three potential substrates. However, the goal of this study is not to get a comprehensive overview of kinetic profile of engineered proteins, but to get a quantitative handle (Km values) on the binding affinities of substrates to fusion proteins, which would allow us to estimate the structural changes brought due to the introduction of the linker sequence into the proteins. In order to carry out such an analysis a simplified kinetic profiling was performed. Within this analysis, the concentration of one of the two (AMP and MgATP) substrates was held constant while varying the concentration of the other substrate. Such an approach does not allow us to investigate whether or not fusion proteins follow the same mechanism for phosphonyl transfer as the wild type enzyme, but it provides us with quantitative data on the binding affinities of the substrates.

The enzyme activity was assayed at five different concentrations of each substrate. The obtained K_m values are reported in Table 5.

Sample ID	**KmATP, mM**	**KmAMP, mM**
AK1	0.12 ± 0.01	0.13 ± 0.01
G1	0.27 ± 0.02	0.1 ± 0.007
G2	0.1 ± 0.003	0.36 ± 0.01

Table 4. Experimentally determined Km values for taged versions of fusion proteins. Concentration of AMP was held constant at 0.5 mM while concentration of ATP was varied for Km determination. Concentration of ATP was held constant at 1mM while concentration of AMP was varied for Km determination.

Both G1 and G2 proteins were characterized by somewhat lower Km values for both substrates, which is indicative of linker induced structural changes to the binding pockets of the substrates. The further interpretation of data is hindered by the absence of information on exact sites for binding AMP and MgATP on human AK1. Observed lower specific activity of the protein constructs might be however at least in part attributed to lower binding affinities of substrates.

3.4.2 Structural characterization of protein constructs in vitro

A priori, one might expect the tertiary structure of the designed constucts to be significantly perturbed by the introduction of the Gly_7 linker. Human adenylate kinase has now been converted into a set of two or three polypeptide fragments defined by exon/exon boundaries. The size of the linker was intentionally chosen to be large so that it would allow for proper separation of generated fragments, but such a structurally modified constructs could potentially lead to compromised folding if multiple tertiary interactions determining formation of local structures are affected. In order to assess changes in protein structure brought upon the introduction of the Gly_7 linker, circular dichroism (CD) analysis of all engineered proteins as well as wild type human AK1 was performed. Human AK1 is 58% α-helical and exhibited a nice α-helical profile during CD analysis. Both His-tagged G1 and G2 protein constructs had CD traces indicative of overall α- helical structure though they were folded to a lesser extent based on observed molar ellipticities (FIGURE X). Mean residue ellipticities (deg $\times$ cm^2 $\times$ $dmol^{-1}$) were calculated using the relation:

$$[\theta] = [\theta]_{obs} \times \mathrm{MRW}/10 \cdot l \cdot c$$

where θ_{obs} is the measured signal (ellipticity) in millidegrees, l is the optical pathlength of the cell in cm, c is the concentration of the protein construct in mg/ml and

MRW is the mean residue molecular weight (molecular weight of the peptide divided by the number of residues).

The estimated α-helicities were lower in the case of G1 and G2 by approximately 20%. G1CTR and G2CTR proteins indicated a slightly worsened overall folding and had degrees of α--helicities a bit lower then those of G1 and G2 proteins, although a noticable qualitative difference between the CD traces of G1, G2 and G1CTR, G2CTR proteins was observed (Figure 29).

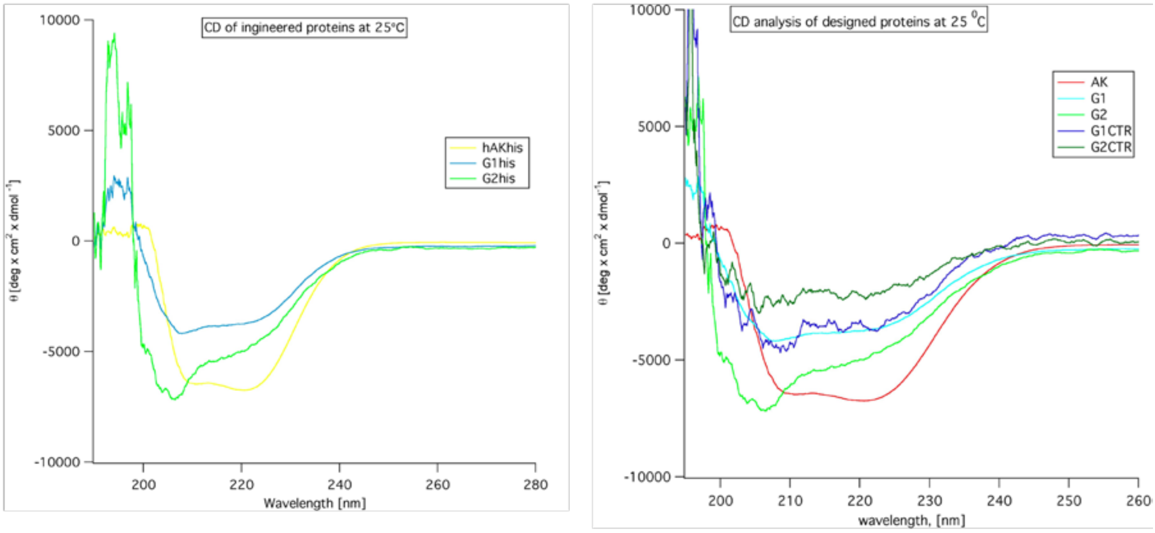


Figure 29. Comparative overlay of CD spectra of fusion proteins. All analyses were performed at 25 °C. Characteristic alpha helical profiles were observed in case of all fusion proteins, however degrees of helicities were lower for fusion proteins when compared to wild-type enzyme. Control proteins exhibited only slightly lower helicity values.

Such an observation is important to our studies since it indicates that even though the misplacement of the Gly_7 linker leads to enzymatically diminished proteins, the reason for the decreased activity is not due to complete disruption of secondary structure. The latter supports the notion that exon-coded polypeptides have a certain function, because only in case when exon-coded polypeptides are complemented does one recapitulate the activity. This observation agrees well with notion proposed within "The Origin of Genes" as well as "Exons as Microgenes" theories that exon-coded polypeptides had a certain function and the complementation of these fragments whether via exon-shuffling or the non-covalent complementation of individually translated polypeptides leads to recovery of enzymatic activity.

The observed diminished folding of G1 and G2 in comparison to wild type human AK1 provides additional explanation for the differences observed in specific activities between wild type enzyme and engineered fusion proteins.

An attempt to characterize the conformational change that takes place within adenylate kinase structure upon binding of ATP and AMP was made. Adenylate kinase adopts an "open" conformation when it is not bound to the substrates while transitioning into a "closed", more compact conformation upon substrate binding. Changes in the shape of proteins can be in principle observed by the native polyacrylamide gel electophoresis (PAGE), because the open form of a protein would migrate slower relative to the more compact substrate bound form. Unfortunately, regardless of the conditions under which native PAGE was conducted the bands for both samples (substrate bound and free forms of proteins) appeared as smears of the gel when stained, complicating analysis of relative positions to each other.

Attempt at addressing same question using size exclusion chromatography was made. However, when the test set of proteins covering the broad range of molecular weights from 100 kDa down to 14 kDa was analyzed through the size exclusion chromatography using Superdex 200, it was observed that at least this particular resin would not be useful in resolving the mass differences anticipated between "open" and “close” forms of adenylate kinase. This particular resin in our handsd was unable to resolve protein bands of 92 kDa, 53 kDa and 29 kDa, as fractions containing proteins corresponding to those masses were observed during SDS-PAGE analysis of collected fractions.

3.5 Functional characterization of engineered proteins *in vivo*

The experimental data obtained through *in vitro* functional and structural assays provided us with valuable information on activity and folding of our constructs, implicating crucial role of exon/exon boundaries in creation of successful reassembly. To test the ability of the fragments generated via introduction of the Gly_7 linker into the full-length enzyme to have activity *in vivo*, complementation was performed using *E. coli* as a host organism. The choice of *E. coli* as a host for *in vivo* complementation is that a prokaryote might produce environmental conditions potentially similar to those present in primordial times. Unlike an *in vitro* assay of purified engineered proteins, which is performed in an artificial environment devoid of competing packing interactions and proteolytic processes leading to protein degradation, the *in vivo* complementation assay provides a more realistic environment to test for catalytic activity. Whether the environment of the prokaryotic cell is the optimal proxy of the primordial time conditions is questionable, but since we do not have reliable information about actual environmental conditions of a time when exon-shuffling based gene formation evolved, we assume that *E. coli* AK cell composition provides what is currently possible.

3.5.1 Design of the in vivo complementation system

DH5α *E. coli* cells were used as a host organism for *in vivo* complementation. The rationale behind our *in vivo* complementation assay is the following: since human adenylate kinase as well as Gly_7 fusion proteins were shown to convert ATP to ADP via phosphate transfer, the decrease in intracellular amount of ATP should be observed short after expression of the fusion proteins is initiated [18](Figure 30).

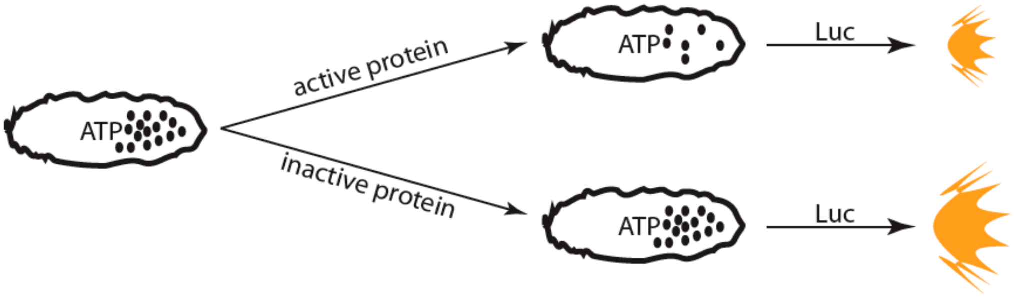


Figure 30. Schematic representation of *in vivo* activity assay system. The host *E. coli* cell is drawn as a circle with a certain intracellular amount of ATP present (indicated as black dots). Upon activation of protein expression off of the transformed plasmid the intracellular amount of ATP is expected to go down in case of active fusion proteins and remain the same in case of inactive or control proteins, as represented by either decreased or unchanged number of black dots in transformed cells. After 2 hours of protein expression, cells were lyzed and a cocktail of luciferase and luciferase substrate were added to the lyzate. Since luciferase utilizes ATP, lower luminescence signal was expected to be observed in case of cells transformed with plasmids coding for active fusion proteins.

Regulation of AK1 activity is complex in eukaryotic organisms[3], but becomes simpler in prokaryotic organisms such as *E. coli* which only has one isoform of adenylate kinase. Its major role is to maintain the nucleotide ratios while using the feedback regulation mechanism to sense changes in the amounts of substrates present. Enzymatic activities of AMP-sensitive metabolic enzymes rely heavily on the proper execution of adenylate kinase function.

DH5α is not a protein overexpression cell line and only small amounts of fusion proteins were expected to be produced due to weak recognition of the T7 promoter by endogenous *E. coli* RNA polymerase[19,20,21]. The need for low *in vivo* concentration of proteins comes from the observed low solubility of proteins *in vivo*. This observation was made during protein overexpression experiments performed in the BL21pLysS(DE3) *E. coli* strain. In that instance all fusion proteins were found localized in inclusion bodies, indicating that at least at the concentration reached during overexpression these proteins are not soluble in the cytoplasm of the cell. The majority if not all generated protein fusion material was found in the cell debri of the BL21pLysS(DE3), as judged by the SDS-PAGE analysis of cell debris and cytoplasmic fractions. On the other hand wild-type human adenylate kinase was present as soluble cytoplasmic protein even in the high concentration range achieved during overexpression. We chose not to use BL21pLysS(DE3) strain for *in vivo* complementation assay due to potential solubility problems associated with high level of protein expression. Specific activities of fusion proteins as determined though the *in vitro* activity assay are high, allowing us to assume that we could detect the changes in intracellular ATP levels even when low concentrations of fusion proteins are produced *in vivo*. Moderate decrease in specific activity, 2 and 9 fold in the case of G1 and G2 respectively, would correspond to catalytic rate enhancement over uncatalyzed reaction of ~10^5 taking into account observed differences in K_m values for the substrates. As indicated earlier in Chapter 3 $(k_{cat}/k_{uncat)}/K_m$ for wild type adenylate kinase is > 10^7.

Changes to intracellular ATP levels due to introduction of the fusion or wild type adenylate kinase were monitored using an *in vivo* ATP measuring kit from Calbiochem

(FIGURE 31). In short, after transformation of DH5α with desired plasmids and fusion protein expression for 2 hours, cells were lysed and supplemented with both luciferase and luciferin (provided with the kit). Luciferase converts one equivalent of luciferin into oxyluciferin while emitting light thus enabling luminescent read-out. Formation of oxyluciferin is coupled to convertion of one equivalent of ATP to AMP and PPi, thus if the intracellular ATP levels are lowered, one would expect lower luminescent read-out as less ATP is available for the luciferase to catalyze the reaction. The expected decrease in intracellular ATP levels was observed in the case of cell lines transformed with plasmids coding for enzymatically active human AK1, G1 and G2 proteins. Cell lines carrying G1CTR and G2CTR genes exhibited the same levels of ATP as control cell lines transformed with an empty vector. The transformation procedure itself had no affect on the intracellular ATP levels as the amount of ATP detected were comparable between wild type DH5α and DH5α transformed with an empty vector (Figure 31).

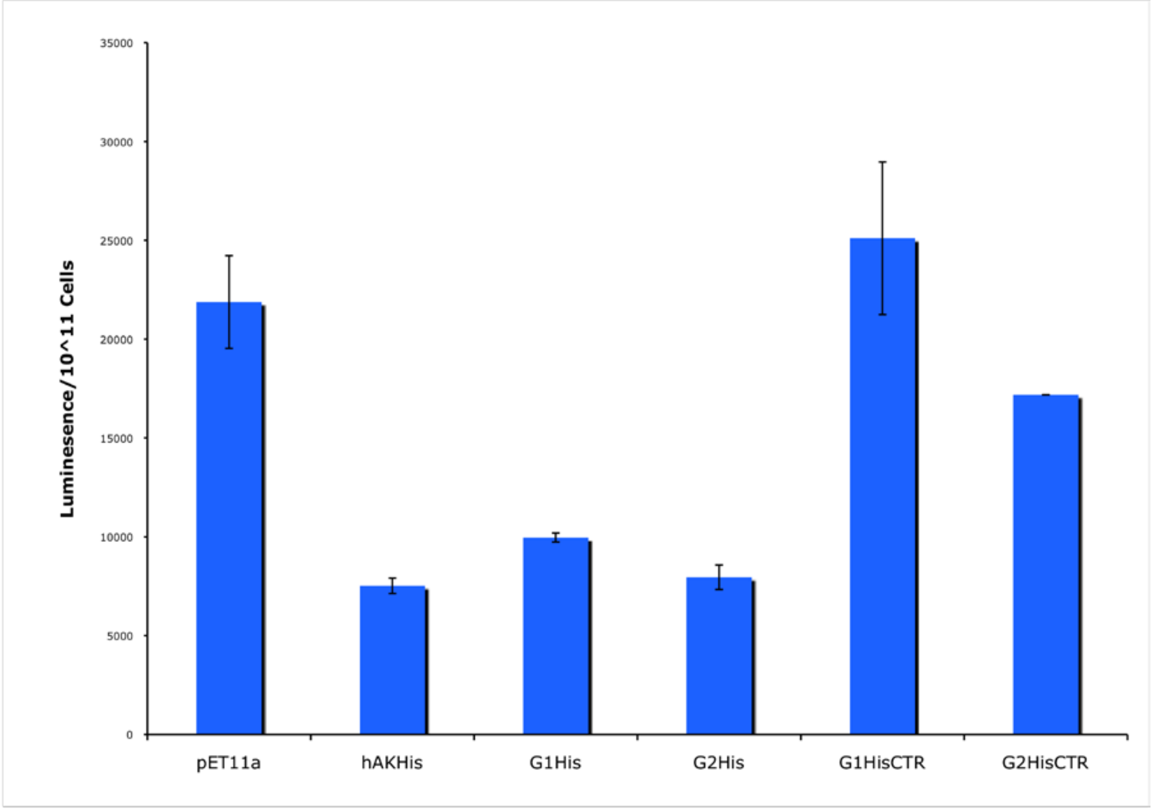


Figure 31. Bar graph representation of the luminescence read-outs observed in *in vivo* complementation assays.

Levels of *in vivo* activity for G1 and G2 are comparable to those of hAK protein. When compared to *in vitro* activity, higher *in vivo* activities observed for G1 and G2 may perhaps be due to a better folding of protein constructs in the intracellular environment. Comparison of protein expression profiles of several *E. coli* proteins was performed using Western blot methodology with broad-spectrum affinity serum for *E. coli* proteins. Comparison of various protein bands pre and post fusion proteins induction indicated little to no change in levels of expression of endogenous proteins (Figure 32).

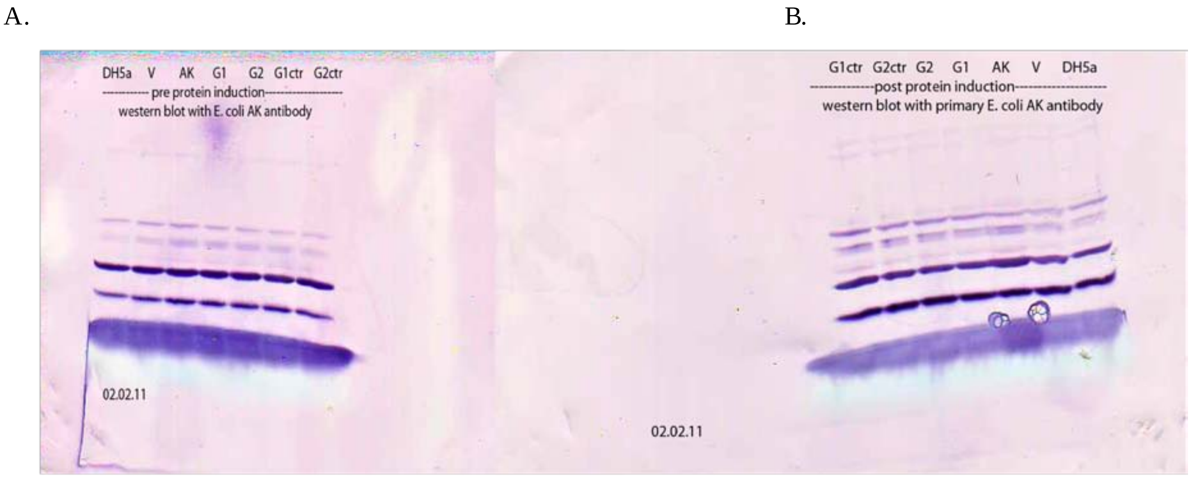

A.
B.
DH5a V AK G1 G2 G1ctr G2ctr
----------- pre protein induction-----------------
western blot with E. coli AK antibody
02.02.11
G1ctr G2ctr G2 G1 AK V DH5a
--------------post protein induction--------------------
western blot with primary E. coli AK antibody
02.02.11

Figure 32. A. Western Blot analysis of cell culture aliquots prior to ITPG protein induction. Broad specificity anti *E. coli* AK serum was used as primary antibody. Visualization is performed via AP-linked secondary antibody **B.** Western Blot analysis of cell culture aliquots post ITPG protein induction. Broad specificity anti *E. coli* AK serum was used as primary antibody. Visualization is performed via AP-linked secondary antibody

This supports the notion that observed changes in intracellular ATP levels are due to the presence of fusion proteins and not because of the changes in expression levels of other host proteins that potentially may utilize ATP either directly as a substrate or indirectly by requiring ATP for their synthesis.

Relative reproducibility of designed *in vivo* complementation assay was demonstrated by obtaining consistent trends of lower detected ATP levels in cell lines carrying fusion genes coding for active enzymes. While absolute values obtained in different replicates of same experiment were not exactly the same, the observed ATP levels trend remained unchanged (FIGURE 33).

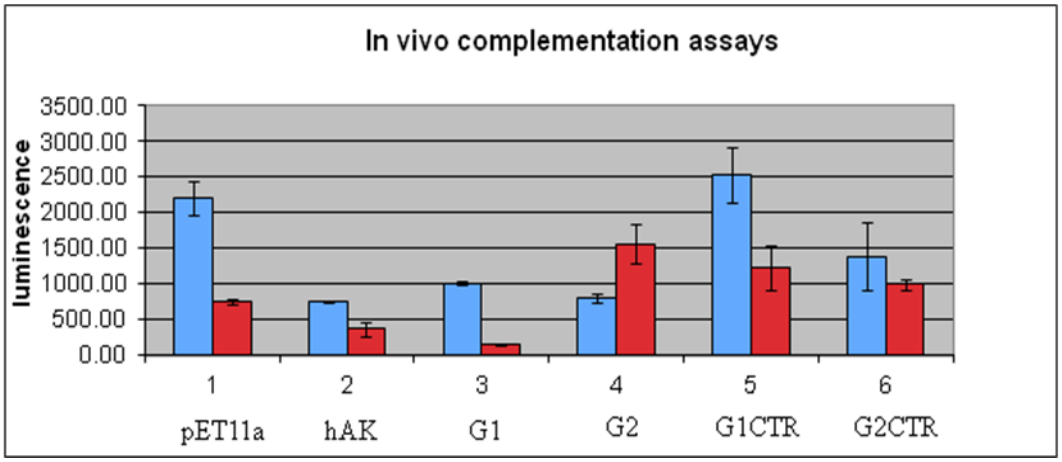


Figure 33. Comparative representation of *in vivo enzymatic* activity of two replicates. Raw luminescence read-out data was scaled to represent the *in vivo* levels of ATP on the same amount of host cells.

Varying amounts of *in vivo* ATP present in the host cell line on different days when replicate experiments were performed could be attributed to a different energetic state of the cell as well as overall level of cell viability which can change from day to day.

3.6 Non-covalent protein reassembly

We previously attempted to perform non-covalent complementation of exon-coded polypeptides from human adenylate kinase 1. As a starting point, the wild-type enzyme was split into two fragments as defined by the exon 3/exon 4 boundary. cDNAs coding for each fragment were submitted to Integrated Dna Technologies (IDT DNA) for gene synthesis and cloned into a pET30 Xa/LIC vector. Fragments were successfully overexpressed in *E. coli* BL21pLysS(DE3) cell line and purified in two steps: denaturing affinity chromatography (Ni-NTA column) was followed by anion exchange chromatography (DEAE resin) to remove any adventitious *E. coli* AK contaminant. The resulting adenylate kinase fragments were subjected to enzymatic assay analysis both as independent fragments and as a pre-incubated mixture of two fragments. In neither of the cases were we able to detect enzymatic activity, as judged by the zero slope of the generated curve (Figure 34).

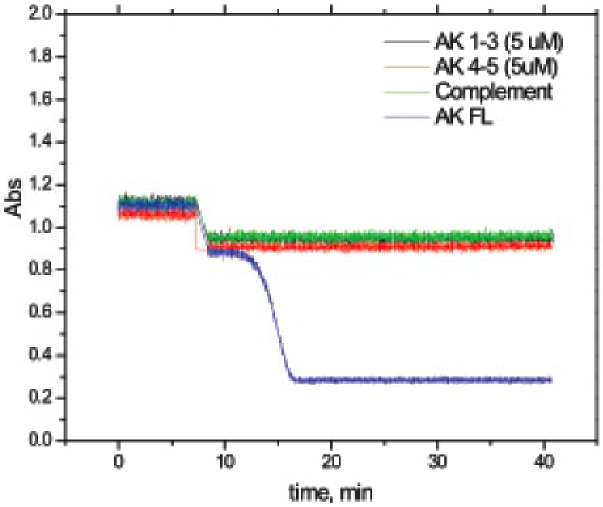


Figure 34. Graphic representation of generated enzymatic curves of the non-covalent protein reassemblies. Enzymatic activity assay traces of hAK1-3, hAK4-5, non-covalent complement of hAK1-3 + hAK4-5 and wild type Adenylate kinase (AK FL). Enzymatic activity assays were initiated by addition of a respective protein after original equilibration of reaction mixture at 37°C for ~ 5 minutes. No enzymatic activity was detected as judged based on the absence of decrease of absorbance at 340nm.

The reason for the failure to detect enzymatic activity coming from non-covalent reassembly may be due to the detection limit of the coupled spectrophotometric assay. Some effort was invested in order to establish fluorescence based activity assay, where oxidation of NADH to NAD$^+$, the final step in the coupled spectrophotometric assay, is coupled to a sensor that generates fluorescent product. Amplite fluorometric NADH assay kit is available though ABD Bioquest Inc (catalog number 15257). In principle, higher consumption of NADH would indicate lower activity of a protein studied, as the coupled spectrophotometric assay was never initiated (no ADP is produced by inactive fragment) thus allowing the entire amount of NADH present in the assay reaction mixture to be eventually consumed by the amplite sensor leading to generation of a high fluorescence signal. If, however, the coupled spectrophotometric assay was initiated due to formation of ADP, some of the NADH would be consumed by lactate dehydrogenase thus leaving less NADH for consumption by the sensor- reagent. Additionally, such assay is based on the competitive consumption of leftover NADH not used by lactate dehydrogenase to convert pyruvate to lactate, so not knowing the exact kinetic parameters of the sensor reagent conversion to red fluorescence product hindered analysis of the activity data. The kinetic parameters of the reaction between amplite reagent and NADH could have been further developed, but the progress with linker mediated protein reassembly encouraged us to halt further investigation into standardizing the fluorescence-based activity assay.

3.7 Conclusion

We have demonstrated the possibility for reconstitution of protein enzymatic activity via a linker-mediated strategy for protein reassembly on the example of two fusion enzymes designed from the human AK1. Utilizing peptidic linker to bring exon-coded polypeptides together allowed us to overcome the entropy penalty for assembling multiple protein components in solution. It also simplified purification greatly as all complementing polypeptides were present on a single polypeptide chain. All of these advantages allowed us to perform functional and structural analysis of engineered proteins as sufficient amounts of fusion proteins were produced, something that has hindered similar studies on TIM and β-hemoglobin earlier by Knowles and Beychok.

Most importantly, we have demonstrated that location of the Gly_7 linker at the exon-intron boundary is crucial for proper reassembly of active proteins, thus providing experimental evidence to support the theory of “exons as microgenes”. To our knowledge, this is the first experimental evidence directly supporting the central role of the exon/exon boundary for proper protein reassembly. Multiple successful protein complementation experiments were performed as summarized in Chapter 2 with varying levels of recovered enzymatic activities exhibited by protein complement ranging anywhere between 3% of wild type enzyme activity to 100% recovery of enzymatic activity. In our case, enzymatic activity was only recovered for assembly of protein fragments connected by linkers situated at the exon/exon boundaries, while those protein constructs where the linker was moved away from the boundary were shown to be inactive. The length of the linker is sufficient to spatially separate the complementing fragments of adenylate kinase by a distance almost two times the diameter of a full-length protein ensuring that proposed linker mediated exon-coded polypeptide reassembly approach serves as a realistic model for non-covalent complementation of separately produced exon-coded polypeptides predicted in primordial times as hypothesized in the “exons as microgenes” theory.

The observed specific activities were only moderately lower then those obtained for wild-type human AK1 protein. These values correlate well with the recovered enzymatic activities of a variety of protein complements studied to date. With the range of recovered enzymatic activities shown for various protein complements spanning the range of 3% to 100% our reassembled proteins fall in the middle with respective 11% and 50% of activity recovered. However our data are unique, as the complementing portions of the protein exactly correspond to the exon-coded polypeptides of adenylate kinase. In comparison to the only other study where recovered enzymatic activity was measured on the non-covalent exon-coded polypeptide reassembled TIM, our contracts exhibit much higher recovered enzymatic activity.

Due to the introduction of Gly_7 linker in either of the constructs, disturbances to the tertiary structure of the protein were expected to take place. Both the G1 and G2 constructs exhibited diminished folding (~20% lower molar elipticities were observed) as judged from their CD spectra. Beychok reported similar decrease in overall helicity on the example of reassembled hemoglobin. The non-covalent complement of β- and α-globins showed ~20% lower overall helicities in comparison to the wild type hemoglobin. Diminished structural organization of both the fusion proteins (G1 and G2) is reflected in changes in K_m values for both substrates (ATP and AMP). The change of 1.5-2 fold was observed for each of the substrate in case of both enzymatically active fusion enzymes, G1 and G2. The absence of the published data on the Km values for the wild type human AK1 made it impossible for us to compare our results to the literature values, however measured Km values for wild type AK1 were close to the values published for AK1 analogs from other organisms.

We have taken care that observed *in vitro* enzymatic activities come from our constructs and not due to *E. coli* AK as it could not be detected in purified proteins. The Western blot detection limit was found to be 0.5 ng of pure *E.coli* AK protein. When assayed for enzymatic activity, 0.5 ng of *E.coli* AK showed no detectable activity. Observed *in vivo* enzymatic activities correlate well with the *in vitro* data. As judged by a decrease in the intracellular level of ATP, only hAK, G1 and G2 constructs were active and led to significant decrease in the concentration of intracellular ATP. We showed that the homeostasis of the host cell line was not changed upon introduction of engineered proteins. Thus, the difference in ATP levels could be attributed to either the activity or inactivity of our constructs.

The linker-mediated protein reassembly approach can be broadly applied to any protein in question, as it possesses no inherent limitations on its use. Its applicability to other proteins needs to be tested in order to demonstrate its broader utility. The success of linker-mediated protein reassembly demonstrated by high levels of recovered enzymatic activity could give rise to a common approach to protein complementation. The location of the split will no longer be arbitrarily chosen on the case-to-case bases, but will be dictated by the exon/intron boundaries of gene coding for the protein of interest. Currently, we are analyzing all the published data on the protein complementation with the goal of getting at a dependence of amount of recovered enzymatic activity by the protein complement and the protein split position in respect to the exon/exon boundary of this protein.

3.9 Experimental Section

3.9.1 Gene design, cloning and expression

Sequences coding for all Gly_7 fusion proteins were submitted to Integrated Dna Technologies (IDT DNA) for gene synthesis. Genes were obtained from IDT in a pIDTSMART-AMP vector. Two primers, forward and reverse respectively, were used for PCR amplification of G1, G2, G1CTR and G2CTR genes as well as the wild-type human AK1 gene using respective IDT DNA plasmid as a template. Sequences of these primers were designed in a way allowing for ligation independent cloning into pET30 Xa/LIC vector and are listed below:

5’AGAGGAGAGTTAGAGCCTTACTTAAG3’ 5’GGTATTGAGGGTCGCATGGAAA 3’

A primer set allowing for incorporation of NdeI restriction site at the 5’end of the gene and BamHI restriction site at the 3’-end was used to generate gene sequences coding for tagless versions of G1 and G2 proteins. Plasmids provided by IDT DNA were used as templates in individual polymerase chain reactions (PCR). Sequences of primers are listed below:

5’ CGCGCCCATATGATGGAAGAAAAACT 3’

5’ TTTGGATCCTTACTTAAGCGCATCCA 3’

A primer set allowing amplification of a portion of human adenylate kinase containing the sequence for exon1-exon2-exon3 was designed in a way allowing for ligation independent cloning into pET30 Xa/LIC vector and the sequences are listed below:

5’AGAGGAGAGTTAGAGCCTTACTTAAG 3’

5’ AGAGGAGAGTTAGAGCCTTATTATTT 3’

A plasmid containing the sequence for wild-type human adenylate kinase provided by IDT was used as a template for the PCR reaction.

A primer set allowing amplification of a portion of human adenylate kinase containing sequence for exon4-exon5 was designed in a way allowing for ligation independent cloning into pET30 Xa/LIC vector and the sequences are listed below:

5’ GGTATTGAGGGTCGCATCGGTCAGC 3’

5’ GGTATTGAGGGTCGCATGGAAA 3’

A plasmid containing the sequence for wild-type human adenylate kinase provided by IDTDNA was used as a template for the PCR reaction.

For all PCR reactions conducted, 1 ug of plasmid template was used. A cocktail of 100 pmol/ul forward (1µl of 100 µM primer stock solution) and 100 pmol/ul reverse primers (1µl of 100 µM primer stock solution) in T4 DNA polymerase buffer containing 100 µM dNTPs (1 µl of 20 µM stock solution) was added to the plasmid templates and PCR reaction was initiated upon addition of 1 µl T4 DNA polymerase in a final volume of 100 µl. Nuclease-free water was added as needed to bring up the reaction volume to 100 µl. PCR reaction conditions were the following:

Denaturing – 1min at 94°C

Annealing – 3 min at 60°C

Extending – 2 min at 72°C

Repeat 30 times

Obtained PCR fragments were analyzed on 1% agarose gel for presence of proper sized bands, purified using a Qiagen PCR purification kit and subcloned into either pET30a Xa/LIC vector or pET11a vector following manufacturer’s protocol. In short, for ligation independent cloning into pET30 Xa/LIC vector (product number 69077-3 fromNovagen), PCR purified fragments were digested with LIC-quality DNA polymerase provided with the cloning kit for 0.5 h at 22°C in the presence of 25 mM dATP, 100 mM DTT and T4 DNA polymerase buffer followed by DNA polymerase inactivation at 75°C for 20 min. Digested fragments were incubated with pre-digested vector for 20 min at

22°C. NovaBlue GigaSingles (product number 71127 from Novagen) competent cells (25ul) were used to transform the entire ligation mixture. During the transformation procedure, LB broth supplemented with 50ug/ml kanamycin was used. A more detailed procedure can be found in the users protocol provided with the cloning kit (TB163 Rev. M0211JN). Same protocol was followed for cloning of DNA sequences coding for exon1-exon2-exon-3 and exon4-exon5 portions of human adenylate kinase.

In case of pET11a cloning the following procedure was followed. 1 μg of each plasmid and pET11a were independently subjected to double restriction digestion with NdeI and BamHI in the BamHI specific buffer. 1 μl of each enzyme was used and the reaction mixture volume was brought up to 100 μl with nuclease-free water. Double digestion reaction was allowed to proceed for 16 hours (overnight) at 37 °C. Upon completion, digestion reactions were evaporated to dryness, redissolved in 10 μl of nuclease-free water and loaded on 1% Agarose gel. Bands corresponding to digested genes were cut out from the gel and purified using a Qiagen gel purification kit. Concentrations of obtained DNA sequences were determined using a Nanodrop spectrophotometer. Ligation dependant cloning was set up using 1:10 excess of insert to vector. On average, 0.5 μg of insert was incubated with 25 ng of digested, purified pET11a vector in 1x ligation buffer in the presence of 1 μl of T4 DNA ligase (total volume 20 μl) for 2 hours at room temperature. Entire ligation reaction volume was transformed into 50 μl of Novablue GigaSingles. In short, 20 μl of ligation mixture were added to thawed on ice completent cells, incubated on ice for 5 minutes, followed by heat shock at 42 °C for 30 seconds with subsequent incubation on ice for 2 minutes.

Transformation mixture was plated on LB agar plates supplemented with 100 ug/ml ampicillin.

A single colony was picked from plates incubated overnight at 37°C and was placed in 5 ml of LB broth with 50 ug/ml kanamycin. Cultures were allowed to grow for 16 h at 250rpm and 37°C. Cells were pelleted by centrifugation at 13200 rpm and RT. Cells were lysed and plasmids were purified using mini Quagen or Wizard plus SV miniprep (Promega) DNA purification kits. Purified DNAs were stored at -20°C in TE buffer (10mM Tris-Cl, pH 7.5). The nucleotide concentration of dsDNA was determined by measuring the UV absorbance at 260nm by a NanoDrop spectrophotometer (ND-1000, Thermo Scientific) and using ε=6500 $M^{-1}cm^{-1}$. Obtained plasmids were successfully sequenced at TUCF using T7 primer. A plasmid coding for *E. coli* AK protein was a generous gift of Prof. Mathews from Dept. of Biochemistry and Biophysics, Oregon State University. Raw sequencing data for all engineered plasmids can be found in Appendix 1.

E. coli strain BL21 (DE3)pLysS cells (NovaBlue) were transformed with plasmids coding for G1, G2 both tagless (pET11a vector based) and His-tagged (pET30 Xa/LIC), G1CTR, G2CTR, human AK1 and *E. coli* AK proteins. In short, ~ 2 μg of each plasmid were individually added to 50 μl of thawed on ice BL21 (DE3)pLysS cells. After heat shock at 42 °C for 30 seconds, followed by 2 minutes incubation on ice, cells containing the desired plasmid were spread on plates supplemented with 50 ug/ml kanamycin and 34 ug/ml chloramphenicol (has limited solubility in water, 1000x stock solution was prepared in ethanol) in case of genes cloned into pET30 Xa/LIC vector. For G1 and G2 genes cloned into pET11a LB agar plates supplemented with 100 ug/ml ampicillin and 34 ug/ml chloramphenicol were used. Plates were incubated overnight at 37°C. Isolated colonies were grown in 5 ml LB broth with 50 ug/ml kanamycin or 100 ug/ml ampicilin and 34ug/ml chloramphenicol overnight using the incubator shaker at 37°C and 250 rpm. Fresh LB medium with the same antibiotic composition was inoculated for each cell culture with 1:50 dilution of the overnight culture. Cells were further grown at 37°C and 250rpm until the OD_{600} of medium reached 0.6. The cells were then induced for protein expression by addition of isopropyl-β-D-thiogalactosidase to 0.4 mM in each flask. The cells were harvested by centrifugation using Sorvall RC-5C plus (ThermoScientific) floor centrifuge (rotor SLC1000) for 1h at 19000rpm after 3h of induction period. Cell paste was stored at -20°C if it was not lysed immediately[22,23].

3.9.2 Protein purification and characterization

For His-tag containing fusion proteins the following purification protocol was followed[22,24,25]:

collected cell paste for each of the protein constructs was suspended in 50mM Tris [tris(hydroxymethyl)aminomethane-HCl], pH=8.0, sonicated at 600W for 5 minutes and, centrifuged at 19,000 rpm for 1 hour to separate cytoplasmic and the cell debris portions of the cells. Cell debris containing proteins of interest were dissolved in buffer A (8M urea, 0.1 M sodium phosphate buffer, 0.01 M Tris-Cl, pH=8.0). The obtained solution was applied to a Ni-NTA column pre-equilibrated in buffer A. The column was then washed with buffer B (8M urea, 0.1 M sodium phosphate buffer, 0.01 M Tris-Cl, pH=6.3) to remove non-specifically bound proteins. Desired proteins were eluted with buffer C (8M urea, 0.1 M sodium phosphate buffer, 0.01 M Tris-Cl, pH=4.5). Fractions containing desired protein were determined using sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and the fractions combined. Combined protein containing fractions were concentrated using Amicon filter, dialyzed against buffer D {20 mM Tris-Cl, 300mM NH_4Cl, pH=8.0} and applied on DEAE Sepharose fast flow column equilibrated in buffer D. A gradient from 30% to 100% high salt buffer E (20mM Tris-Cl, 1M NH_4Cl, pH=8) was run for 500 minutes with a flow rate of 1.0 ml/min at 4°C. Fractions containing desired protein were determined by SDS-PAGE and combined. Combined protein containing fractions were concentrated using an Amicon filter, dialyzed against buffer E (20 mM sodium phosphate, 500 mM NaCl, 20mM imidazole pH=7.4) and applied on NiNTA column pre-equilibrated with buffer E for affinity separation under native conditions. The column was washed with buffer F (20 mM sodium phosphate, 500 mM NaCl, 50mM imidazole pH=7.4) to remove non-specifically bound proteins. Desired proteins were eluted in buffer G (20 mM sodium phosphate, 500 mM NaCl, 200 mM imidazole pH=7.4). Fractions containing desired protein were determined by SDS-PAGE and combined. Combined protein containing fractions were concentrated using an Amicon filter and dialyzed against buffer H (1x PBS, 1mM DTT, 5% glycerol). The same purification strategy was applied to all engineered proteins: hAK1-3 (protein containing exon 1 though 3 coded polypeptides), hAK4-5 (protein containing exon 4 though 5 coded polypeptides), G1his and G2his with a minor exception for not performing the last native Ni-NTA purification step on hAK1-3 and hAK4-5. Wild type human AK1 and *E. coli* AK were purified from cytoplasmic portion of the lysed cells via a native Ni-NTA column. In case of human AK1, affinity chromatography was followed by anion exchange chromatography. After centrifugation for 1hr at 4200 rpm, cells were suspended in buffer E (20 mM sodium phosphate, 500 mM sodium chloride, 20 mM imidazole pH = 7.4) and applied onto NiNTA column pre-equilibrated in buffer E. Column was washed with buffer F (20 mM sodium phosphate, 500 mM sodium chloride, 50 mM imidazole pH = 7.4) to remove any non-specifically bound proteins. hAKhis was eluted in buffer G (20 mM sodium phosphate, 500 mM sodium chloride, 200 mM imidazole pH = 7.4). Fractions containing desired protein were determined using SDS-PAGE and combined. Combined protein containing fractions were concentrated using Amicon filter and dialyzed against buffer D (20 mM Tris-Cl, 300mM NH_4Cl, pH=8.0) and applied on DEAE Sepharose fast flow column equilibrated in buffer D. Conditions for the anion exchange chromatography of human AK1 were followed as above for engineered proteins. Fractions containing human AK1 protein were determined by SDS-PAGE and combined. Combined protein containing fractions were concentrated using Amicon filter and dialyzed against buffer H (1x PBS, 1mM DTT, 5% glycerol).

For tagless fusion proteins the following purification protocol was followed:

collected cell paste for each of the protein constructs was suspended in 50mM Tris [tris(hydroxymethyl)aminomethane-HCl], pH=8.0, sonicated at 600W for 5 minutes and, centrifuged at 19,000 rpm for 1 hour to separate cytoplasmic and the cell debris portions of the cells. Cell debris containing proteins of interest were dissolved in buffer A {8M urea, 0.1 M sodium phosphate buffer, 0.01 M Tris-Cl, pH=8.0}. Obtained solution was applied to Superdex 200 column pre-equilibrated in buffer I (50 mM Tris-Cl pH 7.0, 30 mM NaCl, 0.1 mM DTT). Fractions containing desired proteins were determined by SDS-PAGE and combined. Combined protein containing fractions were concentrated using Amicon filter and dialyzed against buffer D (20 mM Tris-Cl, 300mM NH_4Cl, pH=8.0) and applied on DEAE Sepharose fast flow column equilibrated in buffer D. A gradient from 30% to 100% high salt buffer E (20mM Tris-Cl, 1M NH_4Cl, pH=8) was run for 500 minutes with a flow rate of 1.0 ml/min at 4°C. Fractions containing desired protein were determined by SDS-PAGE and combined. Combined protein containing fractions were concentrated using Amicon filter and dialyzed against buffer H (1x PBS, 1mM DTT, 5% glycerol). Raw SDS-PAGE data of column fractions of each purification step for each fusion protein can be found in Appendix 2.

Protein concentration was determined according to Lowry protocol. Detailed protocol can be found in the manufacturer's instruction manual (LIT448 Rev D from Bio Rad). In short, 9 dilutions of a protein standard (BSA) covering the range from 0.1 mg/ml to 1.6 mg/ml were prepared. 5 μl of each of standard protein sample and sample were pipetted into a dry microtiter plate. 25 μl of reagent A was added into each well, followed by addition of 200 μl of reagent B. The plate was gently agitated and was incubated at room temperature for 15 minutes. After 15 minutes, absorbances were read at 750 nm using microtiter plate reader. The standard curve was generated using raw data obtained from various concentrations of BSA protein. The mg/ml concentrations of engineered proteins were determined using a generated standard curve. Raw data containing A750nm readouts for BSA samples as well as fusion protein samples can be found in Appendix 3. Concentrations of hAK1-3 and hAK4-5 protein fragments were determined measuring the UV absorbance at 280nm by a NanoDrop spectrophotometer (ND-1000, Thermo Scientific) after both protein fragments were denatured using buffer A (8M urea, 0.1 M sodium phosphate buffer, 0.01 M Tris-Cl, pH=8.0). 1:10 dilution was affectively achieved as 9 volumes of buffer A were used to effectively denature 1 volume of each protein fragment solution. Extinction coefficients for generated protein fragments were determined using Protein Calculator v3.3 (Scripps Research Institute) and were: ε=3840 $M^{-1}cm^{-1}$ for hAK1-3 and ε=5120 $M^{-1}cm^{1}$ for hAK4-5 protein fragments.

SDS-PAGE analysis was performed to assess the purity of obtained proteins. Obtained SDS-PAGE gels were stained separately with either silver or Coomassie blue stain. Molecular weights of proteins were determined by MALDI-TOF using α-cyano-4-hydroxycinnamic acid as a matrix[26]. Observed masses closely matched theoretical values: G1his m/z $[27116]^+$ exp. $[27002]^+$; G2his m/z $[27446]^+$ exp. [27402]; G1hisCTR m/z $[26981]^+$ exp. $[27002]^+$; G2hisCTR $[27454]^+$ exp. $[27402]^+$, hAKhis m/z $[26639]^+$ exp. $[26603]^+$, G1 tagless m/z^{2+}[11060.9] exp. m/z [22034.1], G2 tagless m/z^{2+}[11201.6] exp. [22433.5]. Images of obtained MALDI spectra can be found in Appendix 4.

Western blot analysis using human AK primary antibody (AbCam) was done on G1, G2 (taggless as well as His-tagged) and hAK proteins to ensure their nature as desired. Western blot using primary *E. coli* adenylate kinase antibody was done to ensure the absence of *E. coli* adenylate kinase in the protein preparations. Protocol for Western Blotting was followed as specified in Current Protocols in Molecular Biology. In short, to establish the lowest detection limit of *E. coli* AK in our protein preparations a standard curve was generated. Varying amounts of purified *E. coli* AK protein covering the range of 0.5 ng to 20 ng pure protein, were loaded on SDS-PADE gel. The gel was run at 120V for 1hour and was followed by western blotting membrane transfer (PVDF membrane). Protein transfer was performed overnight at 14V. Membrane probing was done with 1mg/ml solution of primary *E. coli* AK antibody that was a generous gift from Prof. Matthew’s lab. Alkaline phosphatase linked anti-rabbit secondary antibody was used to visualize the bound primary antibody. Upon incubation with chromogenic substrate the development of dark bands of DTP was observed. The detection limit of *E. coli* AK was established to be 0.5 ng.

Primary *E. coli* AK antibody was used to establish the purity of engineered proteins. Known amounts of purified proteins (~ 2 μg, which is 10 times the amount used in activity assays) were loaded on the SDS-PAGE. The gel was run at 120V for 1 hour, followed by blotting with PVDF membrane. Protein transfer was performed overnight at 14V. Membrane probing was done with 1gm/ml solution of primary *E. coli* AK antibody that was a generous gift of Prof. Mathew's lab. Alkaline phosphotase linked anti-rabbit secondary antibody was used to visualize the bound primary antibody. Upon incubation with chromagenic substrate the development of dark bands of DTP was not observed in the case of any of the studies proteins. The identical procedure was performed on the same set samples (gels run side by side) using human AK1 primary antibody (AbCam ab54824). Upon incubation with chromogenic substrate the development of dark bands of DTP was observed for all engineered proteins.

3.9.3 Coupled spectrophotometric ATPase assay

A spectrophotometric assay that couples formation of ADP to a decrease in absorbance on NADH was used to examine enzymatic activities, as well as establish Km values for both substrates[27,28]. ATP was dissolved in 1M Tris Cl, pH 8.0, concentration was determined by UV measurements at 260nm using $\varepsilon = 15400\ M^{-1}\ cm^{-1}$ and stored in aliquots at -20°C. NADH was dissolved in 10mM Tris Cl, pH 8.0, concentration was determined by UV measurements at 340nm using $\varepsilon = 6250\ M^{-1}\ cm^{-1}$ and stored in aliquots at -80°C.

All reactions were performed at 37°C and contained 20 mM Tris OAc, pH 7.5, 1 mM DTT, 5% glycerol, 7.5 mM phosphoenolpyruvate (PEP), 0.2 mM NADH, 20 U/ml pyruvate kinase (PK) and 20 U/ml lactate dehydrogenase (LDH), 5 mM magnesium acetate, 1 mM ATP, 0.5 mM AMP, and 50nM protein. Varying amount of ATP and Amp were used to determine Km values of indicated substrates: for Km(MgATP), at fixed concentration of 0.5 mM AMP, concentration of ATP was varied from 0.3 mM to 1mM; for Km(AMP), at fixed concentration of 1 mM MgATP concentration of AMP varied from 0.25 mM to 0.9 mM. Enzyme activity was assayed at 5 different concentrations of substrate. All measurements were performed in duplicates. Reaction mixtures were preincubated for 5 minutes and activated by addition of protein. Absorbance data were collected using a Varian Cary 50 spectrophotometer equipped with an 18-cell holder with a PCB 150 peltier-controlled water bath. Relative specific activities were calculated according to Pan et al:

$$\text{Specific activity} = \nu \times V/m,$$

where ν is a slope, V – reaction volume and m – mass of the enzyme. Given the definition of specific activity, relative specific activities were determined as the ratio of specific activity of protein 1 to specific activity of protein 2:

$$SA_1/SA_2 =(\nu_1 \times c_2 \times MW_2)/(\nu_2 \times c_1 \times MW_1)$$

where c – concentration of a protein, MW – molecular weight of a protein.

Reaction rates (in μM/min) were calculated by fitting a straight line tangent to the data, and multiplying the slope by 159. Standard Michaelis – Menten analysis of 1/v v.s. 1/c plots obtained resulted in determination of Km values. Raw fitting data can be found in Appendix 4.

3.9.4 Structural protein characterization

Circular dichroism (CD) measurement were performed on a JASCO J-715 spectropolarimeter equipped with a JASCO PT-423S Peltier temperature controller using 0.1 cm pathlengh cuvette at 25°C. Scans were conducted between 190 nm and 260nm at a speed of 20nm/minute with a spectra bandwidth of 2nm and a sensitivity of 20 millidegrees. The protein concentrations were anywhere between 10 to 40 μM in 100mM Tris buffer pH=8.0, 200 mM NaCl (1xPBS) containing 2% β-mercaptoethanol. Molar elipticies were calculated using the relatio:

$$[\theta] = [\theta]_{obs} * (MWR)/10*lc$$

where $[\theta]_{obs}$ is the measured signal in millidegrees, MWR is the mean residue molecular weight (molecular weight of the peptide divided by the number of residues), l is optical pathlengh of cell in cm, and c is protein concentration in mg/ml.

3.9.5 Detection of change in protein conformational change using native gel

Two different native gel conditions were tested. Given high pI of engineered fusion proteins acidic gels had to be used in order to allow proteins to enter the gels without reversing polarity. Hen egg lysozyme was used as the positive control protein. Gel was prepared in a following manner:

4% staking gel pH 6.8: 4 ml of 0.25 M Acetate-KOH pH 6.8, 1.5 ml 30% Acrylamide 0.8% methylene-bis-acrylamide, 9.6 ml miliQ water, 150 μl of 10% APS, 15 μl TEMED (added last)

15% resolving gel pH 4.3: 6.7 ml of 1.5M Acetate-KOH pH 4.3, 6 ml 50% glycerol, 13.3 ml 30% acrylamide 0.8% methelene-bis-acrylamide, 320 μl of 10% APS, 40 μl TEMED (added last).

The ability of native basic gels to resolve fusion proteins was investigated. The gel was prepared in a following manner:

4% stacking gel pH 6.8: 1.25 ml of 0.5% Tris-Cl pH 6.8, 1 ml of 30% Acrylamide 0.8% methylene-bis-acrylamide solution, 2.75 ml nuclease-free water, 25 μl of 10% APS, 2.5 μl TEMED (added last)

15% resolving gel pH 8.8: 2.0 ml 1.5 M Tris-Cl pH 8.8, 3.75 ml 30% Acrylamide 0.8% methylene-bis-acrylamide solution, 2.25 ml miliQ water, 40 μl of 10% APS, 4 μl TEMED (added last).

Polarities of electrodes were reversed during this run, to ensure proper entrance of fusion proteins into the gel. Images of the native PAGEs can be found in Appendix 6.

3.9.6 In vivo ATP monitoring assay

DH5α *E. coli cells* were separately transformed with plasmids bearing human AK1, G1, G2, G1CTR, G2CTR, pET11a vector. Untransformed host cells were used as an additional control in the *in vivo* protein reassembly assay. Transformation mixtures were plated on LB agar plates supplemented with 50 μg/ml Kan. Plates were incubated overnight at 37°C. A single colony was picked of each plate and used to inoculate 5 ml LB cultures supplemented with 50 μg/ml Kan. Fresh 50 ml LB medium supplemented with 50 μg/ml Kan was inoculated for each cell culture with 1:50 dilution of the overnight culture. Cells were further grown at 37°C and 250 rpm until OD_{600} of medium reached 0.5-0.6 and were induced for protein expression by addition of isopropyl-β-D-thiogalactosidase to 0.4 mM in each flask. Cell cultures were further incubated at 37°C for 2 hrs. Cell culture aliquots were taken (1ml each) prior and post IPTG induction. Cell culture aliquots were subjected to Western blot analysis (performed as described earlier) with primary broad selectivity *E.coli* AK antibody to make sure that intracellular levels of *E. coli* AK and other related proteins did not get altered dramatically.

Cells were harvested after 2 hours post protein induction by centrifugation at 4200 rpm at 4°C for 1hour. Cells were re-dissolved in 5ml of LB broth supplemented with 50 μg/ml Kan and serial dilutions were plated on LB agar plates supplemented with 50 μg/ml Kan (to determine concentration in cells/ml) for each culture. 10 μl of 5ml resuspended broth culture were further used to detect internal amount of ATP. The internal ATP detection was performed according to manufacturer's protocol (Calbiochem catalog number 119107). In short, cells were lysed in the white 96-well plate by addition of 100ul of lysing buffer B supplemented with luciferase substrate. Cells were incubated for 10 mins at room temperature. ATP detection reaction was initiated by addition of 1 μl of luciferase enzyme in darkness. Emission of light (luminescence) was detected using a Microtiter plate reader after 1 minute of addition of enzyme. Each experiment was performed five times. Raw luminescence data observed in this study can be found in Appendix 5.

3.9.10. Detection of ATP binding using P^{31} NMR

~ 80 μg of fusion protein (either G1, G2 or wild type adenylate kinase all His tagged versions) were mixed with ~110 μg ATP in total 400 μl of D_2O. P^{31} traces of fusion protein bound to ATP were recorded and a position of γ-phosphate signal was monitored. P^{31} profile of unbound ATP served as a control. H_3PO_4 was used as in external standard to calibrate phosphate signal. Detectable shifts in position of γ-phosphate signal were observed and are summarized in a table below:

Sample	ATP	AK	G1	G2
γ-phosphate signal	-9.88	-4.81	-5.34	-5.86

The observed shifts in γ-phosphate signal for G1 and G2 proteins were lower then that in case of wild type AK which is indicative of effective binding of ATP by the fusion proteins.

3.10 References

(1) Saxonov, S.; Daizadeh, I.; Fedorov, A.; Gilbert, W. *Nucleic Acids Research* **2000**, *28*, 185-190.

(2) Bertolaet, B. L.; Knowles, J. R. *Biochemistry* **1995**, *34*, 5736-5743.

(3) Dzeja, P.; Terzic, A. *International Journal of Molecular Sciences* **2009**, *10*, 1729-1772.

(4) Janssen, E.; Dzeja, P. P.; Oerlemans, F.; Simonetti, A. W.; Heerschap, A.; De Haan, A.; Rush, P. S.; Terjung, R. R.; Wieringa, B.; Terzic, A. *EMBO Journal* **2000**, *19*, 6371-6381.

(5) Ren, H.; Wang, L.; Bennett, M.; Liang, Y.; Zheng, X.; Lu, F.; Li, L.; Nan, J.; Luo, M.; Eriksson, S.; Zhang, C.; Su, X. D. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102*, 303-308.

(6) Von Zabern, I.; Wittmann Liebold, B.; Untucht Grau, R. *European Journal of Biochemistry* **1976**, *68*, 281-290.

(7) Matsuura, S.; Igarashi, M.; Tanizawa, Y.; Yamada, M.; Kishi, F.; Kajii, T.; Fujii, H.; Miwa, S.; Sakurai, M.; Nakazawa, A. *Journal of Biological Chemistry* **1989**, *264*, 10148-10155.

(8) Tsai, M. D. *Biochemistry* **1991**, *30*, 6806-6818.

(9) Sheng, X. R.; Li, X.; Pan, X. M. *Journal of Biological Chemistry* **1999**, *274*, 22238-22242.

(10) Saraste, M.; Sibbald, P. R.; Wittinghofer, A. *Trends in Biochemical Sciences* **1990**, *15*, 430-434.

(11) Tan, Y. W.; Hanson, J. A.; Yang, H. *Journal of Biological Chemistry* **2009**, *284*, 3306-3313.

(12) Tian, G.; Yan, H.; Jiang, R. T.; Kishi, F.; Nakazawa, A.; Tsai, M. D. *Biochemistry* **1990**, *29*, 4296-4304.

(13) Sanders Ii, C. R.; Tian, G.; Tsai, M. D. *Biochemistry* **1989**, *28*, 9128-9143.

(14) Ohnishi, S.; Kamikubo, H.; Onitsuka, M.; Kataoka, M.; Shortle, D. *Journal of the American Chemical Society* **2006**, *128*, 16338-16344.

(15) Loughran, S. T.; Walls, D. *Methods in molecular biology (Clifton, N.J.)*, *681*, 311-335.

(16) Craik, C. S.; Buchman, S. R.; Beychok, S. *Nature* **1981**, *291*, 87-90.

(17) Blondin, C.; Serina, L.; Wiesmuller, L.; Gilles, A. M.; Barzu, O. *Analytical Biochemistry* **1994**, *220*, 219-221.

(18) Lasko, D. R.; Wang, D. I. C. *Biotechnology and Bioengineering* **1996**, *52*, 364-372.

(19) Ganguly, A.; Rajdev, P.; Chatterji, D. *Journal of Physical Chemistry B* **2009**, *113*, 15399-15408.

(20) Prosen, D. E.; Cech, C. L. *Biochemistry* **1986**, *25*, 5378-5387.

(21) Steen, R.; Dahlberg, A. E.; Lade, B. N.; Studier, F. W.; Dunn, J. J. *The EMBO journal* **1986**, *5*, 1099-1103.

(22) Abrusci, P.; Chiarelli, L. R.; Galizzi, A.; Fermo, E.; Bianchi, P.; Zanella, A.; Valentini, G. *Experimental Hematology* **2007**, *35*, 1182-1189.

(23) Guarente, L.; Roberts, T. M.; Ptashne, M. *Science* **1980**, *209*, 1428-1430.

(24) Alexandre, J. A. C.; Roy, B.; Topalis, D.; Pochet, S.; P√©rigaud, C.; Deville-Bonne, D. *Nucleic Acids Research* **2007**, *35*, 4895-4904.

(25) Baneyx, F. *Current Opinion in Biotechnology* **1999**, *10*, 411-421.

(26) Beavis, R. C.; Chait, B. T. *Analytical Chemistry* **1990**, *62*, 1836-1840.

(27) Saint Girons, I.; Gilles, A. M.; Margarita, D. *Journal of Biological Chemistry* **1987**, *262*, 622-629.

(28) Hamada, M.; Kuby, S. A. *Archives of Biochemistry and Biophysics* **1978**, *190*, 772-792.

Chapter 4
CONCLUSIONS AND FUTURE DIRECTIONS

4.1 Conclusion

Within this work we were able to develop a novel approach that allowed us to experimentally test the notion that primordial enzymes were non-covalent assemblies of exon-coded polypeptides. A novel aspect of using Gly_7 linker to assist in such protein fragment reassembly led to substantial improvements in our ability to produce substantial amounts of protein constructs with improved folding and removal of entropic penalty associated with reassembly of multiple fragments. Enzymatic activities were recovered in the case of both engineered fusion proteins where position of Gly_7 linker was defined by either exon 3-exon 4 boundary or the exon3-exon4 and exon4–exon5 boundaries together. The latter fusion protein had two linker sequences incorporated. The exon/exon defined position for Gly_7 linker was shown to be crucial for successful protein reassembly. Once the linker was moved away from exon/exon boundary of a protein no detectable enzymatic activity was observed as determined by the coupled spectrophotometric assay. Recovery of enzymatic activity was shown to take place *in vivo* as well.

The origin of new protein function is an important evolutionary question[1]. Attempts at repeating nature's path and creating new protein functionalities in the lab has led to generation of a new field of study referred to as protein engineering[2]. The ability to engineer proteins is widely applied in multiple areas of research. Industrial areas such as agriculture have recently started employing benefits of added use of engineered proteins. The pharmaceutical industry has been harvesting the great potential of engineered proteins in aiding either production of active ingredients or utilizing engineered proteins as drugs themselves, e.g. monoclonal antibodies for quite some time[3].

One of the ways of generation of new proteins was considered within this study with a major focus on protein evolution. The observation that two out of three exons as well as introns in the genomic sequence of *Tetrahymena* terminate with the amber codon[4] allowed Knowles to propose that in the primordial time exons represented microgenes. These microgenes would independently be translated into short polypeptides that would non-covalently assemble to form an active enzymatic complex. Intron would be utilized as splicing units providing mobility to exonic portions of the gene. The observation of statistically higher "in-frame" localized introns implies that at least a large sliver of modern day introns were present prior to eukaryotic/prokaryotic divergence. We believe that compelling evidence has been collected, that exon-coded polypeptides are capable of reassembly with recovery of enzymatic activity to a high degree. One of the future directions of this project is to devise a strategy to provide support to the theory that at least a fraction of modern day introns was present prior to eukaryotic/prokaryotic divergence. One way of doing this is to take a conserved prokaryotic enzyme that has highly homologous eukaryotic counterpart with an established exon/intron gene structure and then introduce Gly_7 linkers at positions within the prokaryotic enzyme that align with exon/exon boundaries of the eukaryotic counterpart. Gly_7 would introduce the split between the "exon" coded portions of prokaryotic enzyme and test it by fragment complementation assay as has been done in this work. This project is currently being investigated in the laboratory on the much discussed example of *E. coli* adenylate kinase. The choice of protein target for this study was mostly dictated by ease of comparability of results to those already acquired. The high homology between the human adenylate kinase and *E. coli* adenylate kinase[5] allow easy identification of Gly_7 incorporation sites within the *E. coli* adenylate kinase protein sequence. Two different human adenylate kinase exon/exon boundaries were translated onto the sequence of *E. coli* adenylate kinase. Two chimeric enzymes will be engineered where either one or two Gly_7 linkers will be incorporated in much the same way as was done with human adenylate kinase. Data obtained could be easily compared to those obtained for Gly_7 split human adenylate kinase and the relationship between recovered activity and structural changes could be made and would shed some light on the evolutionary role of introns.

4.2 References

(1) Long, M.; Betr√°n, E.; Thornton, K.; Wang, W. *Nature Reviews Genetics* **2003**, *4*, 865-875.

(2) Leisola, M.; Turunen, O. *Applied Microbiology and Biotechnology* **2007**, *75*, 1225-1232.

(3) Brannigan, J. A.; Wilkinson, A. J. *Nature Reviews Molecular Cell Biology* **2002**, *3*, 964-970.

(4) Bertolaet, B. L.; Knowles, J. R. *Biochemistry* **1995**, *34*, 5736-5743.

(5) Schulz, G. E.; Schiltz, E.; Tomasselli, A. G. *European Journal of Biochemistry* **1986**, *161*, 127-132.

APPENDIX 1. Sequencing data of engineered proteins

5’ATGGAAGAAAAACTGAAAAAAACCAAGATCATCTTCGTTGTTGGCGGTCCG
GGTTCTGGCAAAGGCACTCAGTGTGAAAAGATTGTCCAGAAATATGGCTATA
CGCATCTGTCTACGGGCGATCTGCTTCGCTCTGAAGTTAGCTCCGGTTCGGCT
CGTGGCAAAAAACTGTCCGAAATTATGGAAAAAGGTCAGCTAGTACCGCTGG
AGACCGTACTGGATATGCTGCGTGATGCGATGGTCGCTAAAGTCAATACTTCT
AAAGGCTTCCTGATTGACGGCTATCCGCGTGAAGTTCAGCAAGGTGAAGAGT
TTGAGCGTCGTATCGGTCAGCCGACCCTCCTGCTGTATGTTGACGCGGGCCCA
GAAACCATGACCCAGCGTCTGTTGAAACGTGGCGAAACCTCTGGTCGTGTCG
ACGATAACGAAGAAACCATTAAAAAACGTCTGGAAACCTACTACAAAGCCA
CGGAACCGGTTATCGCATTCTATGAAAAGCGTGGCATTGTCCGTAAAGTAAA
CGCGGAAGGTAGCGTTGATTCCGTTTTCAGCCAGGTGTGCACCCATCTGGATG
CGCTTAAGTAA3’

Figure A1. DNA sequence of cloned human adenylate kinase

M H H H H H H S S G L V P R G S G M K E T A A A K F E R Q H M D S P D L G T
G G G S G I E G R **M** E E K L K K T K I I F V V G G P G S G K G T Q C E K I V Q K
Y G Y T H L S T G D L L R S E V S S G S A R G K K L S E I M E K G Q L V P L E T
V L D M L R D A M V A K V N T S K G F L I D G Y P R E V Q Q G E E F E R R I G
Q P T L L L Y V D A G P E T M T Q R L L K R G E T S G R V D D N E E T I K K R
L E T Y Y K A T E P V I A F Y E K R G I V R K V N A E G S V D S V F S Q V C T H
L D A L K **Stop**

Figure 2A. Obtained amino acid sequence of cloned human adenylate kinase gene. His tag is highlighted in yellow.

5’ATGCACCATCATCATCATCATTCTTCTGGTCTGGTGCCACGCGGTTCTGGTA
TGAAAGAAACCGCTGCTGCTAAATTCGAACGCCAGCACATGGACAGCCCAGA
TCTGGGTACCGGTGGTGGCTCCGGTATTGAGGGTCGCATGGAAGAAAAACTG
AAAAAAACCAAGATCATCTTCGTTGTTGGCGGTCCGGGTTCTGGCAAAGGCA
CTCAGTGTGAAAAGATTGTCCAGAAATATGGCTATACGCATCTGTCTACGGG
CGATCTGCTTCGCTCTGAAGTTAGCTCCGGTTCGGCTCGTGGCAAAAAACTGT
CCGAAATTATGGAAAAAGGTCAGCTAGTACCGCTGGAGACCGTACTGGATAT
GCTGCGTGATGCGATGGTCGCTAAAGTCAATACTTCTAAAGGCTTCCTGATTG
ACGGCTATCCGCGTGAAGTTCAGCAAGGTGAAGAGTTTGAGCGTCGTGGCGG
TGGCGGTGGCGGTGGTATCGGTCAGCCGACCCTCCTGCTGTATGTTGACGCG
GGCCCAGAAACCATGACCCAGCGTCTGTTGAAACGTGGCGAAACCTCTGGTC
GTGTCGACGATAACGAAGAAACCATTAAAAAACGTCTGGAAACCTACTACAA
AGCCACGGAACCGGTTATCGCATTCTATGAAAAGCGTGGCATTGTCCGTAAA
GTAAACGCGGAAGGTAGCGTTGATTCCGTTTTCAGCCAGGTGTGCACCCATCT
GGATGCGCTTAAGTAA 3’

Figure 3A. DNA sequence of a cloned G1

M H H H H H H S S G L V P R G S G M K E T A A A K F E R Q H M D S P D L G T
G G G S G I E G R **M** E E K L K K T K I I F V V G G P G S G K G T Q C E K I V Q K
Y G Y T H L S T G D L L R S E V S S G S A R G K K L S E I M E K G Q L V P L E T
V L D M L R D A M V A K V N T S K G F L I D G Y P R E V Q Q G E E F E R R G G
G G G G G I G Q P T L L L Y V D A G P E T M T Q R L L K R G E T S G R V D D N
E E T I K K R L E T Y Y K A T E P V I A F Y E K R G I V R K V N A E G S V D S V
F S Q V C T H L D A L K **Stop**

Figure 4A. Amino acid sequence of cloned G1 gene. Gly_7 linker is highlighted in green. His tag is highlighted in yellow. The second bold methionene residues indicates the first amino acid of the G1 protein.

5’ATGCACCATCATCATCATCATTCTTCTGGTCTGGTGCCACGCGGTTCTGGTA
TGAAAGAAACCGCTGCTGCTAAATTCGAACGCCAGCACATGGACAGCCCAGA
TCTGGGTACCGGTGGTGGCTCCGGTATTGAGGGTCGCATGGAAGAAAAACTG
AAAAAAACCAAGATCATCTTCGTTGTTGGCGGTCCGGGTTCTGGCAAAGGCA
CTCAGTGTGAAAAGATTGTCCAGAAATATGGCTATACGCATCTGTCTACGGG
CGATCTGCTTCGCTCTGAAGTTAGCTCCGGTTCGGCTCGTGGCAAAAAACTGT
CCGAAATTATGGAAAAAGGTCAGCTAGTACCGCTGGAGACCGTACTGGATAT
GCTGCGTGATGCGATGGTCGCTAAAGTCAATACTTCTAAAGGCTTCCTGATTG
ACGGCTATCCGCGTGAAGTTCAGCAAGGTGAAGAGTTTGAGCGTCGTGGCGG
TGGCGGTGGCGGTGGTATCGGTCAGCCGACCCTCCTGCTGTATGTTGACGCG
GGCCCAGAAACCATGACCCAGCGTCTGTTGAAACGTGGCGAAACCTCTGGTC
GTGTCGACGATAACGAAGAAACCATTAAAAAACGTCTGGAAACCTACTACAA
AGCCACGGAACCGGTTATCGCATTCTATGAAAAGCGTGGCATTGTCCGTAAA
GGCGGTGGCGGTGGCGGTGGTGTAAACGCGGAAGGTAGCGTTGATTCCGTTT
TCAGCCAGGTGTGCACCCATCTGGATGCGCTTAAGTAA3’

Figure 5A. DNA sequence of a cloned G2

M H H H H H H S S G L V P R G S G M K E T A A A K F E R Q H M D S P D L G T
G G G S G I E G R **M** E E K L K K T K I I F V V G G P G S G K G T Q C E K I V Q K
Y G Y T H L S T G D L L R S E V S S G S A R G K K L S E I M E K G Q L V P L E T
V L D M L R D A M V A K V N T S K G F L I D G Y P R E V Q Q G E E F E R R G G
G G G G G I G Q P T L L L Y V D A G P E T M T Q R L L K R G E T S G R V D D N
E E T I K K R L E T Y Y K A T E P V I A F Y E K R G I V R K G G G G G G G V N A
E G S V D S V F S Q V C T H L D A L K **Stop**

Figure 6A. Amino acid sequence of cloned G2 gene. Gly$_7$ linkers are highlighted in green. His tag is hilighted in yellow.

5’ATGCACCATCATCATCATCATTCTTCTGGTCTGGTGCCACGCGGTTCTGGTA
TGAAAGAAACCGCTGCTGCTAAATTCGAACGCCAGCACATGGACAGCCCAGA
TCTGGGTACCGGTGGTGGCTCCGGTATTGAGGGTCGCATGGAAGAAAAACTG
AAAAAAACCAAGATCATCTTCGTTGTTGGCGGTCCGGGTTCTGGCAAAGGCA
CTCAGTGTGAAAAGATTGTCCAGAAATATGGCTATACGCATCTGTCTACGGG
CGATCTGCTTCGCTCTGAAGTTAGCTCCGGTTCGGCTCGTGGCAAAAAACTGT
CCGAAATTATGGAAAAAGGTCAGCTAGTACCGCTGGAGACCGTACTGGATAT
GCTGCGTGATGCGATGGTCGCTAAAGTCAATACTTCTAAAGGCTTCCTGATTG
ACGGCTATCCGCGTGAAGTTCAGCAAGGTGAAGAGGGCGGTGGCGGTGGCG
GTGGTTTTGAGCGTCGTATCGGTCAGCCGACCCTCCTGCTGTATGTTGACGCG
GGCCCAGAAACCATGACCCAGCGTCTGTTGAAACGTGGCGAAACCTCTGGTC
GTGTCGACGATAACGAAGAANNNNTTAAAAAACGTCTGGAAACCTACTACA
AAGCCACNGAACCGGTTATCGCATTCTATGNNNNGCGTGNCNTTGTCCCGTA
3’

Figure 7A. DNA sequence of G1 CTR gene

M H H H H H H S S G L V P R G S G M K E T A A A K F E R Q H M D S P D L G T
G G G S G I E G R **M** E E K L K K T K I I F V V G G P G S G K G T Q C E K I V Q K
Y G Y T H L S T G D L L R S E V S S G S A R G K K L S E I M E K G Q L V P L E T
V L D M L R D A M V A K V N T S K G F L I D G Y P R E V Q Q G E E G G G G G
G G F E R R I G Q P T L L L Y V D A G P E T M T Q R L L K R G E T S G R V D D
N E E X X K K R L E T Y Y K A X E P V I A F Y X X R X X V P **Stop**

Figure 8A. Amino acid sequence of cloned G1CTR gene. Gly$_7$ linker is highlighted in green. His tag is hilighted in yellow.

5’ATGCACCATCATCATCATCATTCTTCTGGTCTGGTGCCACGCGGTTCTGGTA
TGAAAGAAACCGCTGCTGCTAAATTCGAACGCCAGCACATGGACAGCCCAGA
TCTGGGTACCGGTGGTGGCTCCGGTATTGAGGGTCGCATGGAAGAAAAACTG
AAAAAAACCAAGATCATCTTCGTTGTTGGCGGTCCGGGTTCTGGCAAAGGCA
CTCAGTGTGAAAAGATTGTCCAGAAATATGGCTATACGCATCTGTCTACGGG
CGATCTGCTTCGCTCTGAAGTTAGCTCCGGTTCGGCTCGTGGCAAAAAACTGT
CCGAAATTATGGAAAAAGGTCAGCTAGTACCGCTGGAGACCGTACTGGATAT
GCTGCGTGATGCGATGGTCGCTAAAGTCAATACTTCTAAAGGCTTCCTGATTG
ACGGCTATCCGCGTGAAGTTCAGCAAGGTGAAGAGGGCGGTGGCGGTGGCG
GTGGTTTTGAGCGTCGTATCGGTCAGCCGACCCTCCTGCTGTATGTTGACGCG
GGCCCAGAAACCATGACCCAGCGTCTGTTGAAACGTGGCGAAACCTCTGGTC
GTGTCGACGATAACGAAGAAACCATTAAAAAACGTGGCGGTGGCGGTGGCG
GTGGTCTGGAAACCTACTACAAAGCCACGGAACCGGTTATCGCATTCTATGA
AAAGCGTGGCATTGTCCGTAAAGTAAACGCGGAAGGTAGCGTTGATTCCGTT
TTCAGCCAGGTGTGCACCCATCTGGATGCGCTTAAGTAA3’

Figure 9A. DNA sequence of G2CTR gene

M H H H H H H S S G L V P R G S G M K E T A A A K F E R Q H M D S P D L G T
G G G S G I E G R **M** E E K L K K T K I I F V V G G P G S G K G T Q C E K I V Q K
Y G Y T H L S T G D L L R S E V S S G S A R G K K L S E I M E K G Q L V P L E T
V L D M L R D A M V A K V N T S K G F L I D G Y P R E V Q Q G E E G G G G G
G G F E R R I G Q P T L L L Y V D A G P E T M T Q R L L K R G E T S G R V D D
N E E T I K K R G G G G G G G L E T Y Y K A T E P V I A F Y E K R G I V R K V N
A E G S V D S V F S Q V C T H L D A L K **Stop**

Figure 10A. Amino acid sequence of cloned G2CTR gene. Gly_7 linkers are highlighted in green**.** His tag is hilighted in yellow.

5’ATGGAAGAAAAACTGAAAAAAACCAAGATCATTTTCGTCGTTGGCGGTCCG
GGTTCTGGCAAAGGCACTCAGCGTGAAAAGATTGTCCAGAAATATGGCTATA
CGCATCTGTCTACGGGCGATCTGCTTCGCTCTGAAGTTAGCTCCGGTTCGGCT
CGTGGCAAAAAACTGTCCGAAATTATGGAAAAAGGTCAGCTAGTACCGCTGG
AGACCGTACTGGATATGCTGCGTGATGCGATGGTCGCTAAAGTCAATACTTCT
AAAGGCTTCCTGATTGACGGCTATCCGCGTGAAGTTCAGCAAGGTGAAGAGT
TTGAGCGTCGTGGCGGTGGCGGTGGCGGTGGTATCGGTCAGCCGACCCTCCT
GCTGTATGTTGACGCGGGCCCAGAAACCATGACCCAGCGTCTGTTGAAACGT
GGCGAAACCTCTGGTCGTGTCGACGATAACGAAGAAACCATTAAAAAACGTC
TGGAAACCTACTACAAAGCCACGGAACCGGTTATCGCATTCTATGAAAAGCG
TGGCATTGTCCGTAAAGTAAACGCGGAAGGTAGCGTTGATTCCGTTTTCAGCC
AGGTGTGCACCCATCTGGATGCGCTTAAATAA3’

Figure 11A. DNA sequence of G1 gene (taggless constract)

M E E K L K K T K I I F V V G G P G S G K G T Q R E K I V Q K Y G Y T H L S T G
D L L R S E V S S G S A R G K K L S E I M E K G Q L V P L E T V L D M L R D A
M V A K V N T S K G F L I D G Y P R E V Q Q G E E F E R R G G G G G G G I G Q
P T L L L Y V D A G P E T M T Q R L L K R G E T S G R V D D N E E T I K K R L E
T Y Y K A T E P V I A F Y E K R G I V R K V N A E G S V D S V F S Q V C T H L D
A L K **Stop**

Figure 12A. Protein sequence G1 taggless construct. Gly$_7$ linker is highlighted in green

5’ATGGAAGAAAAACTGAAAAAAACCAAGATCATTTTCGTTGTTGGCGGTCCG
GGTTCTGGCAAAGGCACTCAGTGTGAAAAGATTGTCCAGAAATATGGCTATA
CGCATCTGTCTACGGGCGATCTGCTTCGCTCTGAAGTTAGCTCCGGTTCGGCT
CGTGGCAAAAAACTGTCCGAAATTATGGAAAAAGGTCAGCTAGTACCGCTGG
AGACCGTACTGGATATGCTGCGTGATGCGATGGTCGCTAAAGTCAATACTTCT
AAAGGCTTCCTGATTGACGGCTATCCGCGTGAAGTTCAGCAAGGTGAAGAGT
TTGAGCGTCGTGGCGGTGGCGGTGGCGGTGGTATCGGTCAGCCGACCCTCCT
GCTGTATGTTGACGCGGGCCCAGAAACCATGACCCAGCGTCTGTTGAAACGT
GGCGAAACCTCTGGTCGTGTCGACGATAACGAAGAAACCATTAAAAAACGTC
TGGAAACCTACTACAAAGCCACGGAACCGGTTATCGCATTCTATGAAAAGCG
TGGCATTGTCCGTAAAGGCGGTGGCGGTGGCGGTGGTGTAAACGCGGAAGGT
AGCGTTGATTCCGTTTTCAGCCAGGTGTGCACCCATCTGGATGCGCTTAAGTA
A3’

Figure 13A. DNA sequence of G2 gene (taggless construct)

M E E K L K K T K I I F V V G G P G S G K G T Q C E K I V Q K Y G Y T H L S T G
D L L R S E V S S G S A R G K K L S E I M E K G Q L V P L E T V L D M L R D A
M V A K V N T S K G F L I D G Y P R E V Q Q G E E F E R R G G G G G G G I G Q
P T L L L Y V D A G P E T M T Q R L L K R G E T S G R V D D N E E T I K K R L E
T Y Y K A T E P V I A F Y E K R G I V R K G G G G G G G V N A E G S V D S V F
S Q V C T H L D A L K **Stop**

Figure 14A. Protein sequence of G2 taggless construct. Gly_7 linkers are highlighted in green

5’ATGCACCATCATCATCATCATTCTTCTGGTCTGGTGCCACGCGGTTCTGGTA
TGAAAGAAACCGCTGCTGCTAAATTCGAACGCCAGCACATGGACAGCCCAGA
TCTGGGTACCGGTGGTGGCTCCGGTATTGAGGGTCGCATGGAAGAAAAACTG
AAAAAAACCAAGATCATCTTCGTTGTTGGCGGTCCGGGTTCTGGCAAAGGCA
CTCAGTGTGAAAAGATTGTCCAGAAATATGGCTATACGCATCTGTCTACGGG
CGATCTGCTTCGCTCTGAAGTTAGCTCCGGTTCGGCTCGTGGCAAAAAACTGT
CCGAAATTATGGAAAAAGGTCAGCTAGTACCGCTGGAGACCGTACTGGATAT
GCTGCGTGATGCGATGGTCGCTAAAGTCAATACTTCTAAAGGCTTCCTGATTG
ACGGCTATCCGCGTGAAGTTCAGCAAGGTGAAGAGTTTGAGCGTCGTTAAGG
CTCTAA3’

Figure 15A. DNA sequence of AK1-3 sequence.

M H H H H H H S S G L V P R G S G M K E T A A A K F E R Q H M D S P D L G T
G G G S G I E G R M E E K L K K T K I I F V V G G P G S G K G T Q C E K I V Q K
Y G Y T H L S T G D L L R S E V S S G S A R G K K L S E I M E K G Q L V P L E T
V L D M L R D A M V A K V N T S K G F L I D G Y P R E V Q Q G E E F E R R
Stop

Figure 16A. AK1-3 protein sequence. His tag is highlighted in yellow.

5’ATGCACCATCATCATCATCATTCTTCTGGTCTGGTGCCACGCGGTTCTGGTA
TGAAAGAAACCGCTGCTGCTAAATTCGAACGCCAGCACATGGACAGCCCAGA
TCTGGGTACCGGTGGTGGCTCCGGTATTGAGGGTCGCATCGGTCAGCCGACC
CTCCTGCTGTATGTTGACGCGGGCCCAGAAACCATGACCCAGCGTCTGTTGA
AACGTGGCGAAACCTCTGGTCGTGTCGACGATAACGAAGAAACCATTAAAAA
ACGTCTGGAAACCTACTACAAAGCCACGGAACCGGTTATCGCATTCTATGAA
AAGCGTGGCATTGTCCGTAAAGTAAACGCGGAAGGTAGCGTTGATTCCGTTT
TCAGCCAGGTGTGCACCCATCTGGATGCGCTTAAGTAA3’

Figure 17A. DNA sequence of AK4-5 sequence.

M H H H H H H S S G L V P R G S G M K E T A A A K F E R Q H M D S P D L G T
G G G S G I E G R I G Q P T L L L Y V D A G P E T M T Q R L L K R G E T S G R V
D D N E E T I K K R L E T Y Y K A T E P V I A F Y E K R G I V R K V N A E G S V
D S V F S Q V C T H L D A L K **Stop**

Figure 18A. AK4-5 protein sequence. His tag is highlighted in yellow.

APPENDIX 2. SDS-gel analysis of protein overexpression and purification steps.

His tagged protein constructs

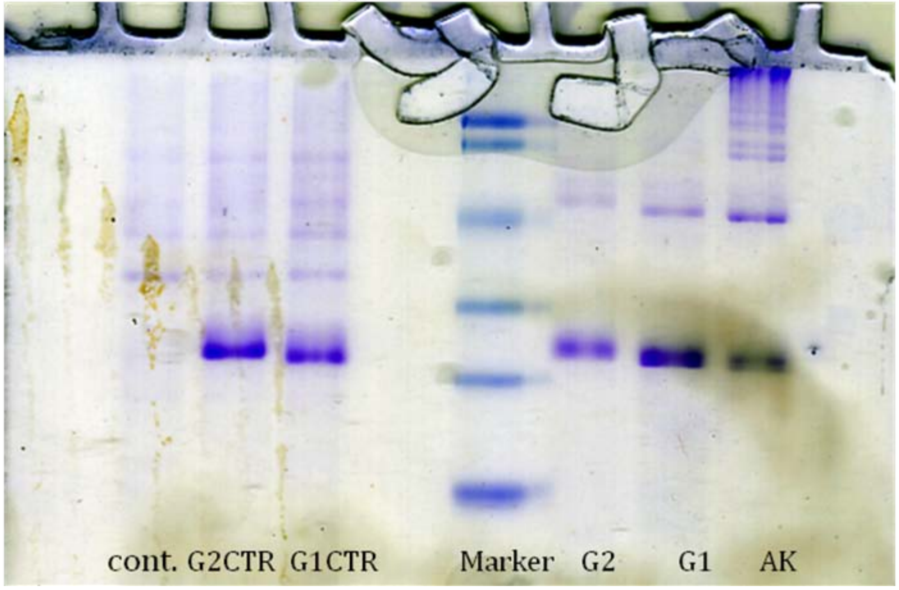


Figure 19A. SDS-PAGE analysis of an aliquot of corresponding 500 mL overexpression LB culture. 1ml was taken out of overexpressed culture and span down at 13200 rpm for 5 min to pellet the cells; supernatant was removed and pelleted cells were lyzed by addition of 1 ml of 8M urea. Cell were physically shredded by passing through 22um gauge niddle 5-10 times each. 3 ul of final 1ml urea solution was loaded into corresponding wells.

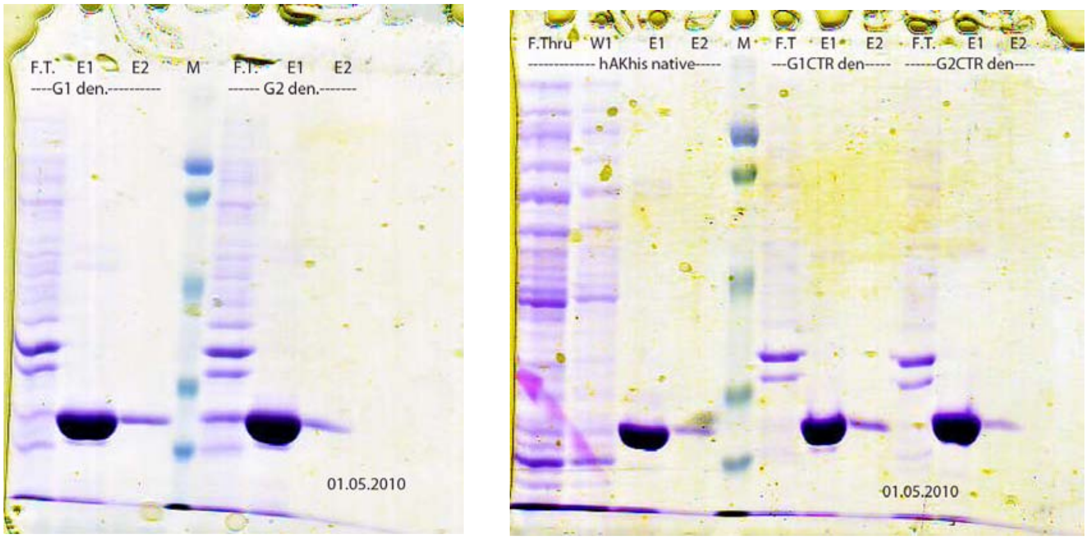


Figure 20A. SDS-PAGE analysis of native affinity chromatography (Ni-NTA) columns. F.T – flow through, E1 and E2 are elution fractions. 5 ul out of 3ml final fraction volume were loaded on the gel for analysis.

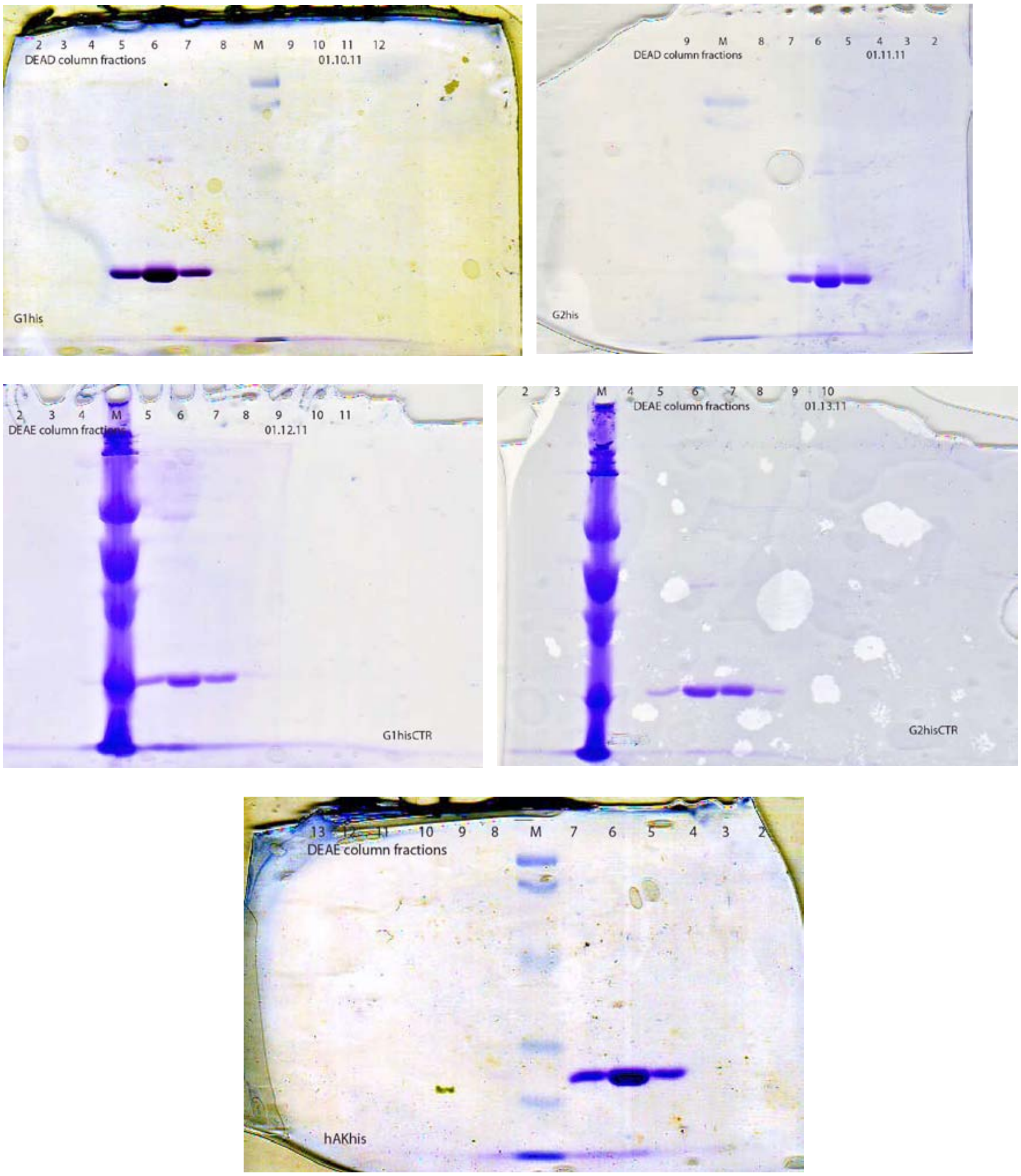


Figure 21A. SDS-PAGE analysis of anion exchange chromatography columns (DEAE resin) of five engineered proteins. In case of all proteins eluate, containing fractions are 5 through 7.

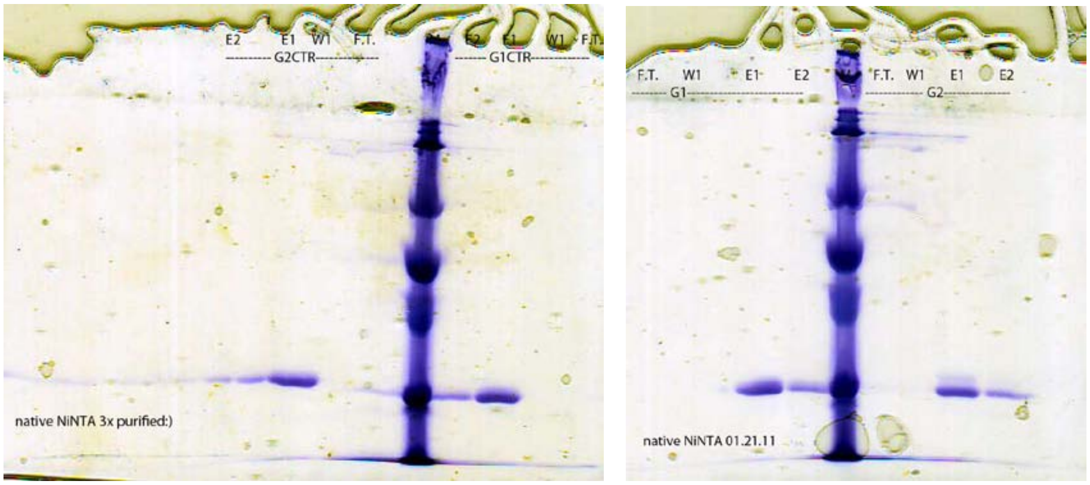


Figure 22A. SDS-PAGE analysis of native affinity column chromatography (final step) for G1, G2 and G1CTR, G2CTR respectively.

Tagless fusion protein versions

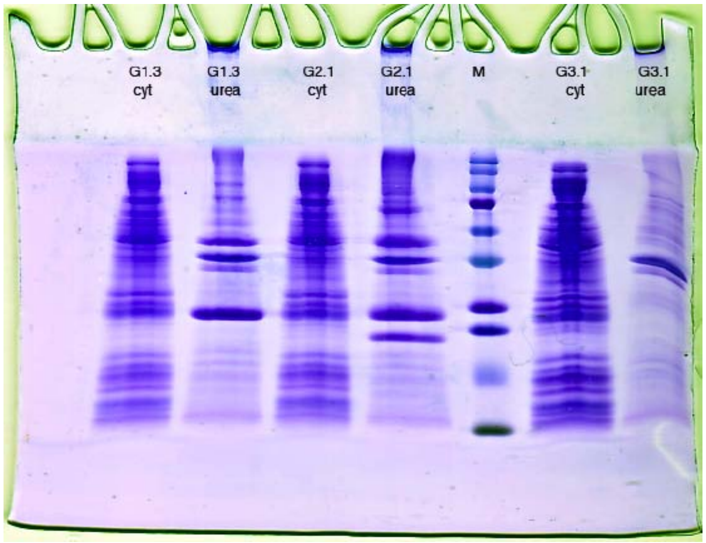


Figure 23A. SDS-PAGE analysis of an aliquot of corresponding 500 mL overexpression LB culture. Overexpression cultures were span down at 19000 rpm for 1 hour to pellet the cells; supernatant was removed (cytoplasm fraction) and pelleted cells were lyzed by addition of 5 ml of 8M urea. 3 ul of final urea solution of cell debri fraction and cytoplasm fraction were loaded into corresponding wells. Lanes named G3.1 can be considered as transformation controls, where protein expression host strain was transformed with empty vector.

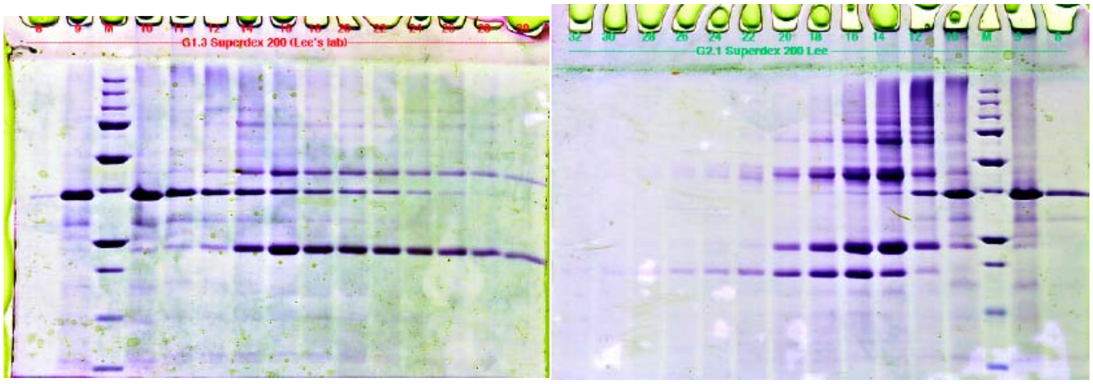


Figure 24A. SDS-PAGE analysis of fractions collected off the size exclusion Superdex 200 columns. 10 ul out of 4 ml final fraction volume were loaded on the gel for analysis.

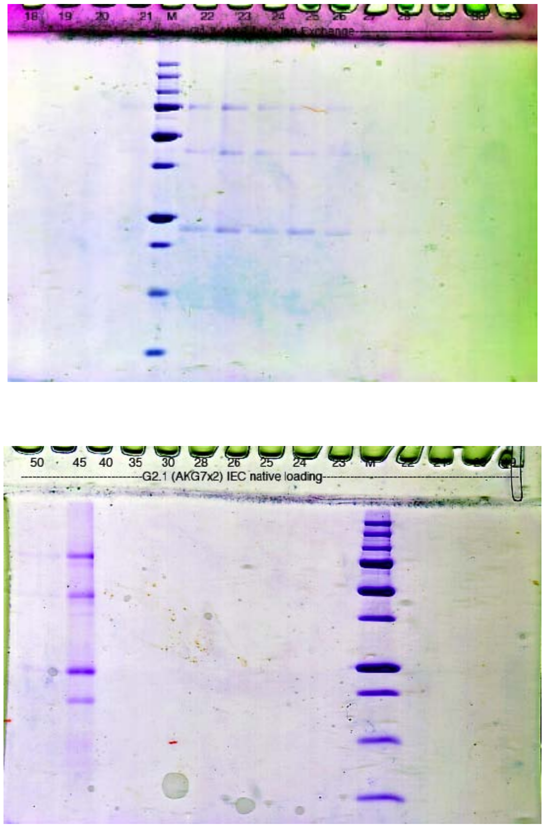


Figure 25A. SDS-PAGE analysis of fractions collected off the anion exchange chromatography columns (DEAE resin) of G1 and G2 engineered proteins.

Adenylate kinase fragments (non covalent assembly)

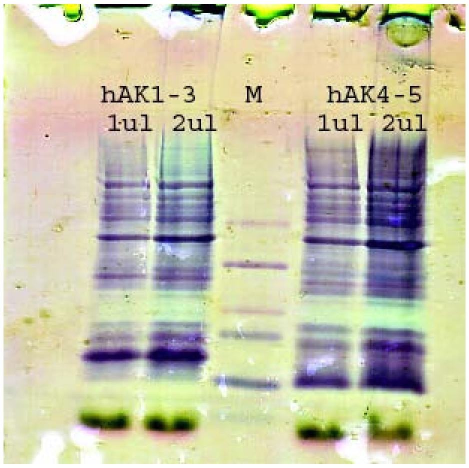


Figure 26A. SDS-PAGE analysis of an aliquot of corresponding 500 mL overexpression LB culture. 1ml was taken out of overexpressed culture and span down at 13200 rpm for 5 min to pellet the cells; supernatant was removed and pelleted cells were lyzed by addition of 1 ml of 8M urea. Cell were physically shredded by passing through 22um gauge niddle 5-10 times each. 1 ul and 2ul of final 1ml urea solution was loaded into corresponding wells.

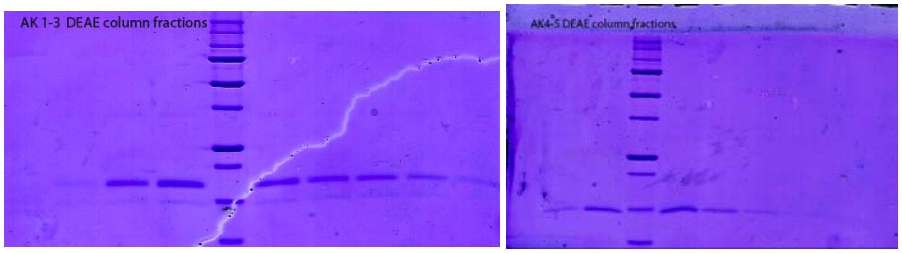
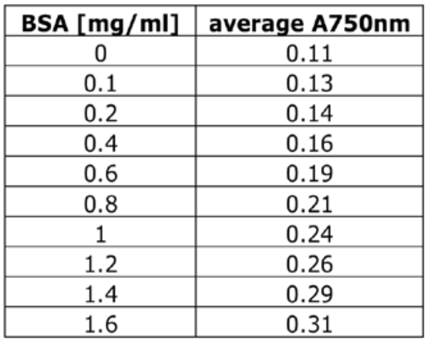


Figure 27A. SDS-PAGE analysis of fractions collected off the anion exchange chromatography columns (DEAE resin) of AK1-3 and AK4-5 protein fragments.

APPENDIX 3. Protein concentration and molecular weight determination

Protein	Molecular weight, Da	pI
hAK	26603.2	8.6
G1	27002.6	8.6
G2	27402.0	8.6
G1CTR	27002.6	8.6
G2CTR	27402.0	8.6
G1 taggless	22087.2	8.6
G2 taggless	22433.5	8.6
AK1-3	16932.2	8.9
AK4-5	14657.5	7.7
E. coli AK	28572.4	5.7

Table 5. Molecular weights and isoelectric points of proteins were determined using protein calculator v3.3 software freely available through Scribbs research institute.



BSA [mg/ml]	average A750nm
0	0.11
0.1	0.13
0.2	0.14
0.4	0.16
0.6	0.19
0.8	0.21
1	0.24
1.2	0.26
1.4	0.29
1.6	0.31

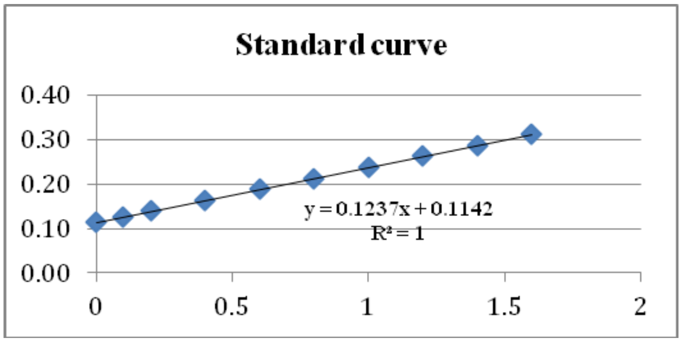


Figure 28A. Standard curve data for BSA concentration determined by Lowry method. Obtained equation for A750nm read-out and protein concentration was used to determine concentration of all proteins.

protein	A750nm 1	A750nm 2	A750nm 3	Average	uM
AK	0.20	0.20	0.19	0.20	25
G1	0.14	0.14	0.14	0.14	8
G2	0.14	0.14	0.14	0.14	8
G1CTR	0.15	0.15	0.15	0.15	11
G2CTR	0.17	0.17	0.17	0.17	15

Table 6. Protein concentrations were determined using Lowry's method. The above indicated standard curve was used to determine mg/ml concentrations according to raw A750nm readings. Using molecular weights calculated via protein calculator v3.3 molar concentrations were determined.

protein	ε $M^{-1}cm^{-1}$	mg/ml read out	μM
AK1-3	5120	0.25	19
AK4-5	3840	0.06	4
G1	8960	0.3	13.6
G2	8960	0.28	12.5
ADKhis	8960	1 mg/ml	38

Table 7. Protein concentrations were determined according to Beer's law. Absorbances at 280nm were determined using NanoDrop spectrophotometer. 1mm pathlength was used. Extinction coefficients were determined using Protein Calculator v3.3 from Scripps research Institute. Each protein sample was denatured using 8 M urea buffer. 10 times excess of urea buffer was used to ensure complete denaturation.

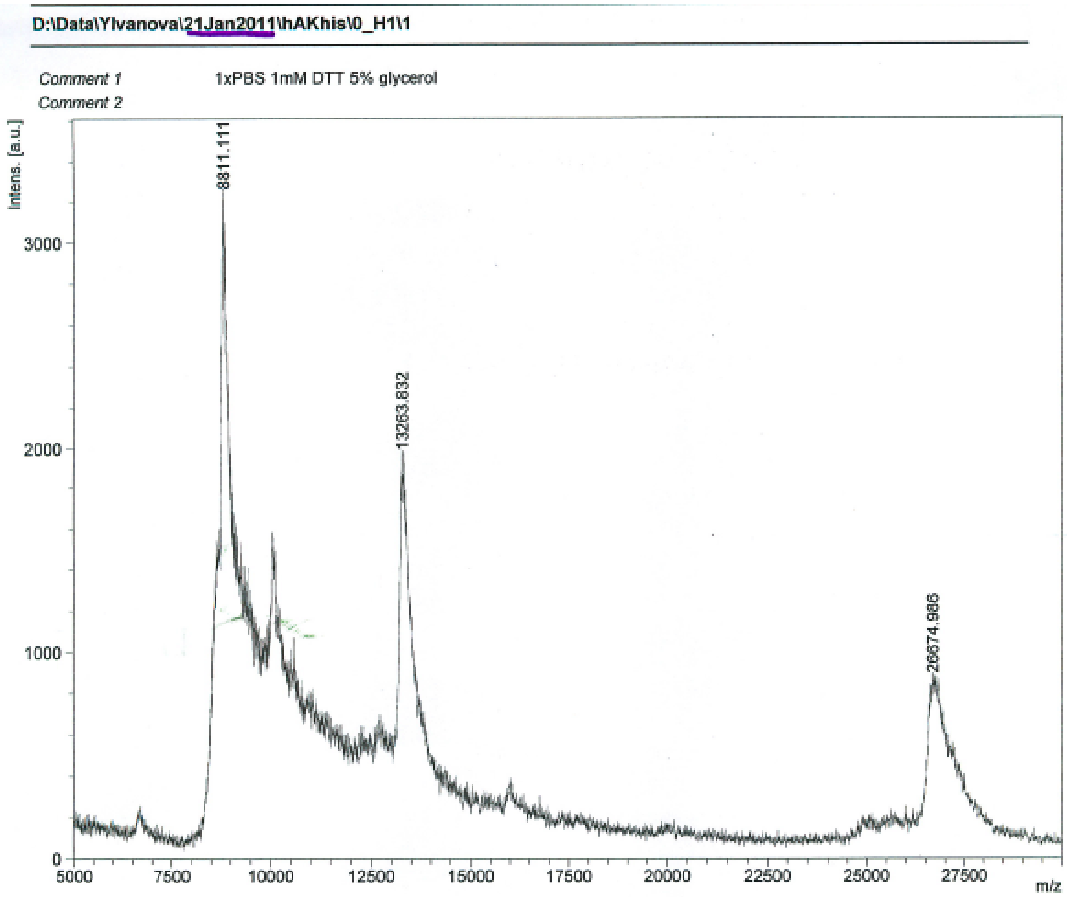


Figure 29A. MALDI data of human adenylate kinase. Expected m/z 26603.2; observed m/z is 26674.9 and m/z^{+2} is 13263.8

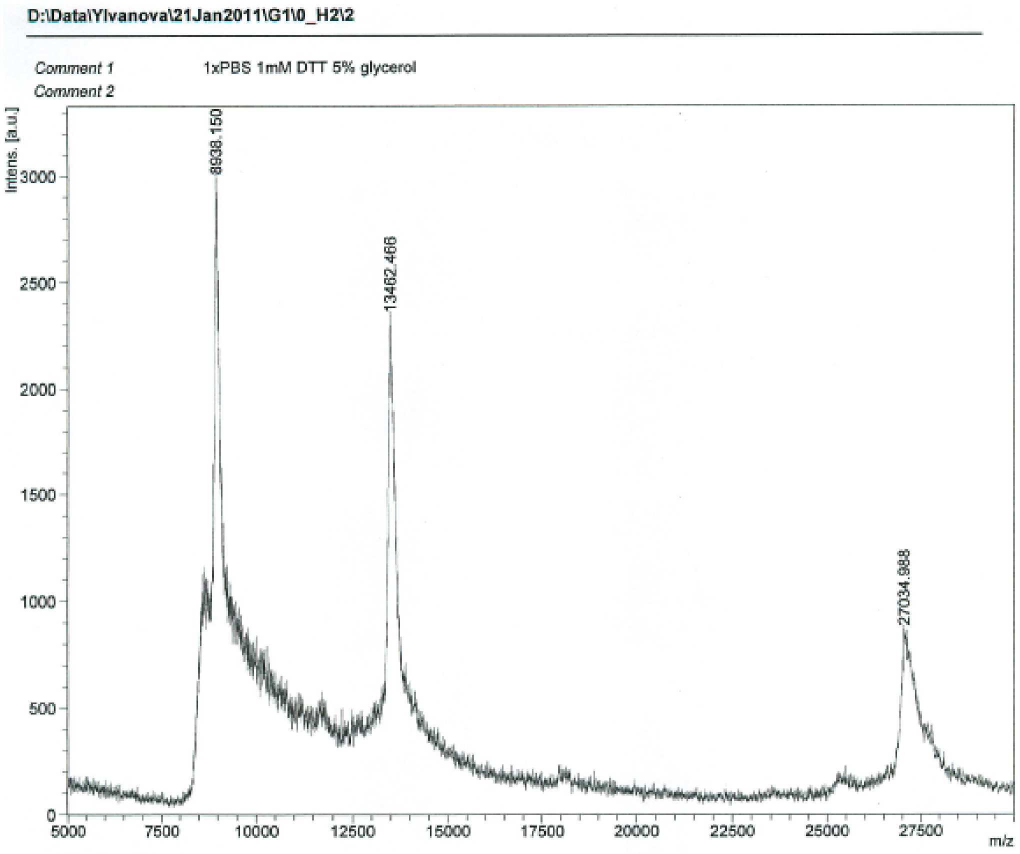


Figure 30A. MALDI data on G1 protein. Expected m/z is 27002.6; observed m/z is 27034.9 and m/z^{+2} is 13462. 5

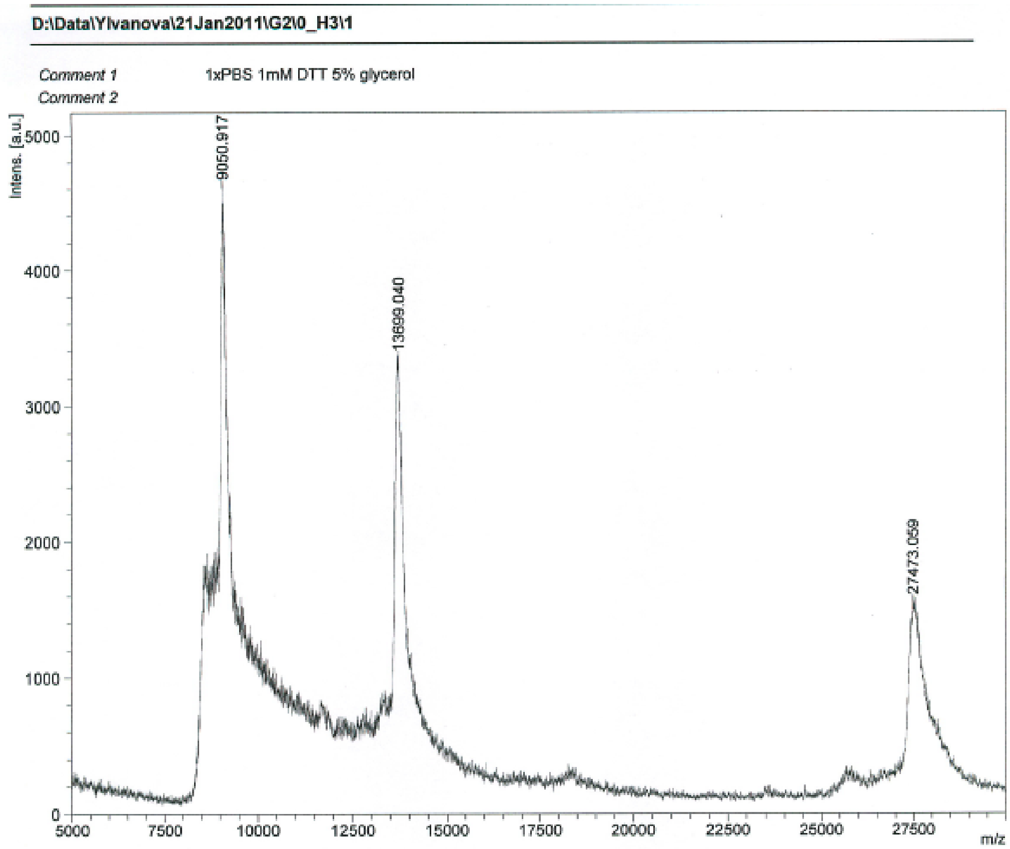


Figure 31A. MALDI data of G2. Expected m/z is 27402.0; observed m/z is 27473. 1 and m/z^{+2} is 13699.0.

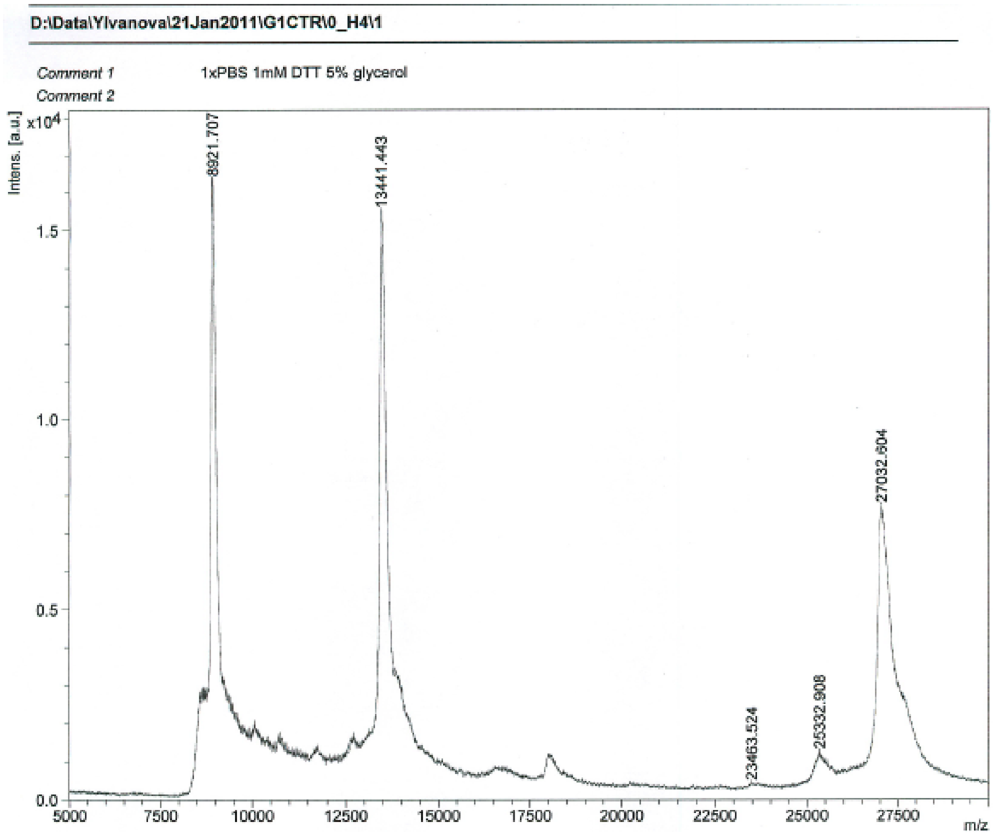


Figure 32A. MALDI data of G1CTR protein. Expected m/z is 27002.6; observed m/z is 27032.6 and m/z^{+2} is 13441.4.

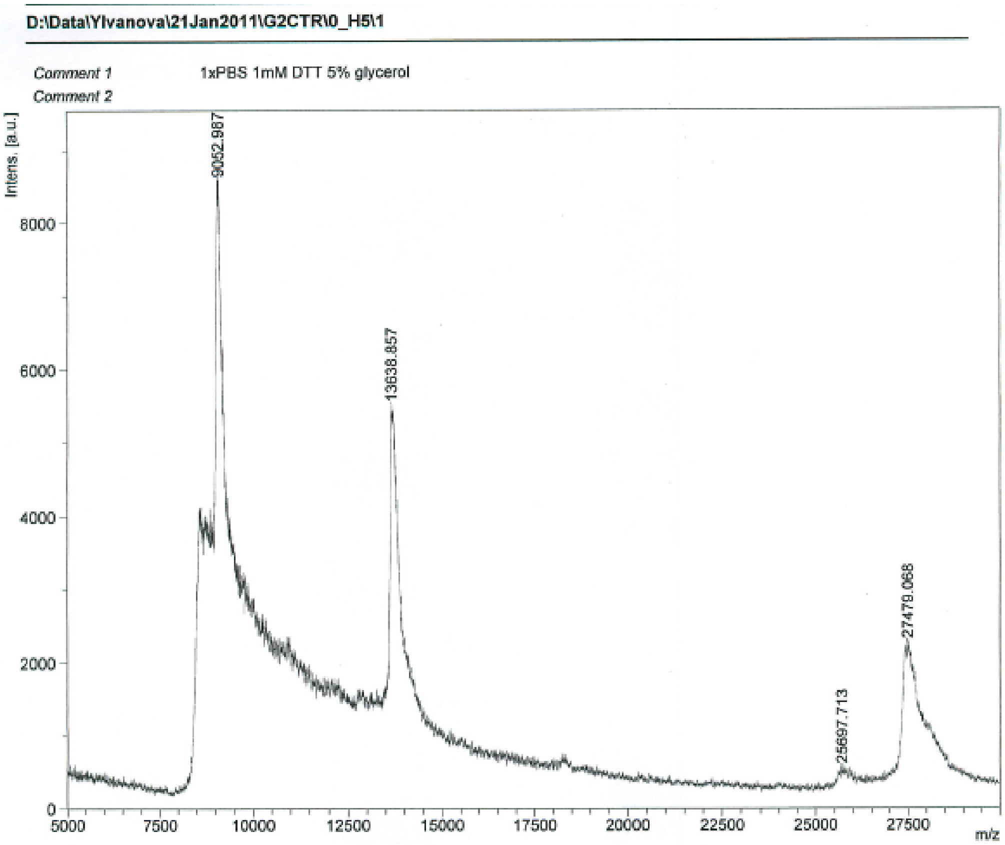


Figure 33A. MALDI data of G2CTR protein. Expected m/z is is 27402.0; observed m/z is27479.1 and m/z^{+2} 13636.9.

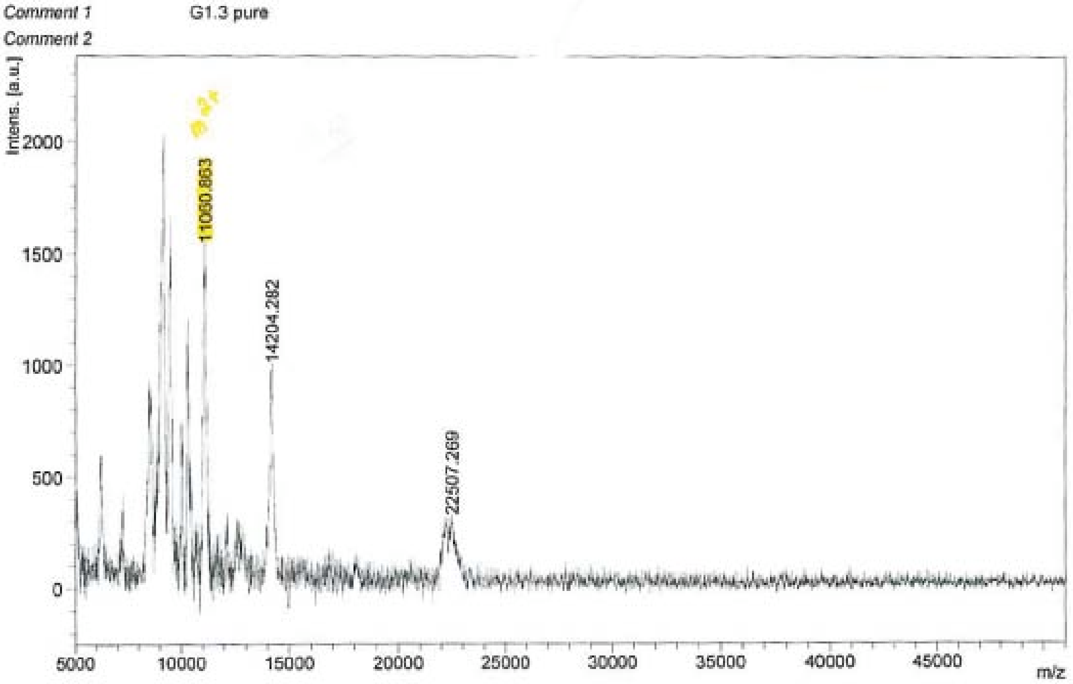


Figure 34A. MALDI data of G1 taggless protein. Expected m/z is is 22087.2; observed m/z^{2+} is 11060.8

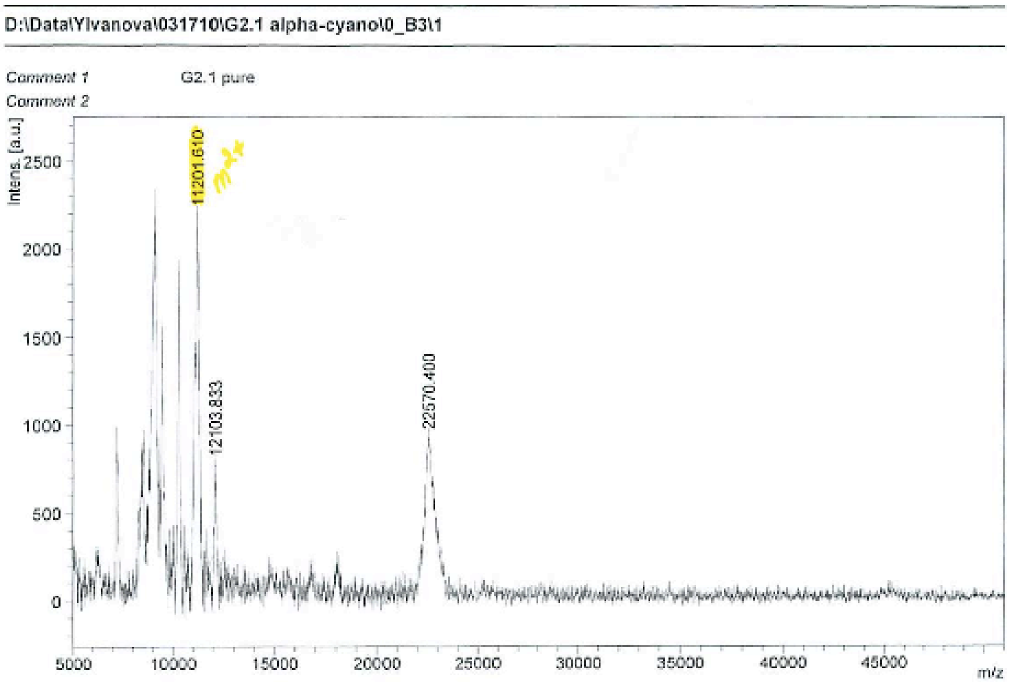


Figure 35A. MALDI data of G2 taggless protein. Expected m/z is 22433.5, observed m/z^{2+} is 11021.6.

APPENDIX 4. Activity assay and kinetics analysis

Components	**1x**	**stock solutions**	**amount of stock used**
protein sample	50 nM	varied from protein to protein	x µl (up to 50 µl with 1x buffer)
$MgCl_2$	5 mM	0.5 M	2 µl
KCl	50 mM	1 M	10 µl
PEP	1 mM	400 mM	0.5 µl
ATP	1 mM	200 mM	1 µl
AMP	0.5 mM	200 mM	0.5 µl
NADH	0.2 mM	17 mM	2.4 µl
PK	20 U/ml	12506 U/ml	0.32 µl
LDH	20 U/ml	12240 U/ml	0.33 µl
10x buffer	1 x	10x	20 µl
di water	n/a	n/a	112.95µl

Table 8. Components of activity and kinetics analysis assays. Concentrations of proteins varied from protein to protein, but final concentration of constructs was maintained 50nM in case of all the proteins. All reactions were initiated by addition of enzyme after 5 minutes of incubating assay reaction mixture at 37°C.

Components	**1x**	**stock solutions**	**amount of stock used**
protein sample	50 nM	varied from protein to protein	x µl (up to 50 µl with 1x buffer)
$MgCl_2$	5 mM	0.5 M	2 µl
KCl	50 mM	1 M	10 µl
PEP	1 mM	400 mM	0.5 µl
ATP	1 mM	263.2 mM	Varied *
AMP	0.5 mM	200 mM	0.5 µl
NADH	0.2 mM	17 mM	2.4 µl
PK	20 U/ml	12506 U/ml	0.32 µl
LDH	20 U/ml	12240 U/ml	0.33 µl
10x buffer	1 x	10x	20 µl
Di water	n/a	n/a	112.95µl

Table 9. Components of kinetic parameters determination for ATP. Concentration of AMP was kept constant 0.5 mM. Different amounts of ATP were added to each individual cuvette. All reactions were initiated by addition of enzyme after initial incubation of assay reaction mixture at 37°C for 5 minutes.

* amounts of ATP added: 0.2 µl (0.26 mM), 0.4 µl (0.53 mM), 0.5 µl (0.66 mM), 0.76 µl (1mM).

Components	**1x**	**stock solutions**	**amount of stock used**
protein sample	50 nM	varied from protein to protein	x µl (up to 50 µl with 1x buffer)
$MgCl_2$	5 mM	0.5 M	2 µl
KCl	50 mM	1 M	10 µl
PEP	1 mM	400 mM	0.5 µl
ATP	1 mM	200 mM	1 µl
AMP	0.5 mM	200 mM	Varied *
NADH	0.2 mM	17 mM	2.4 µl
PK	20 U/ml	12506 U/ml	0.32 µl
LDH	20 U/ml	12240 U/ml	0.33 µl
10x buffer	1 x	10x	20 µl
di water	n/a	n/a	112.95µl

Table 10. Components of kinetic parameters determination for AMP. Concentration of ATP was kept constant at 1 mM. Different amounts of AMP were added to each individual cuvette. All reactions were initiated by addition of enzyme after initial incubation of assay reaction mixture at 37°C for 5 minutes.
* amounts of AMP added: 0.2 µl (0.2 mM), 0.3 µl (0.36 mM), 0.5 µl (0.5 mM), 1 µl (1mM), 1.5 µl (1.5 mM).

APPENDIX 5. *In vivo* activity assay luminescence read out

Sample name	pET11a	hAKhis	G1 His	G2 His	G1 HisCTR	G2 HisCTR
read out 1	20114	10208	9992	8793	29420	17187
read out 2	24632	10977	9925	7646	24461	10441
read ou 3	21285	8671	10347	7803	21814	17354
average	22010	9952	10088	8081	25232	14994
STD.DEV	2345	1174	227	622	3861	3944
%RSD	11	12	2	8	15	26

Table 10. Raw data of the luminescence signals observed in *in vivo* complementation assay

APPENDIX 6. Monitoring of adenylate kinase conformational change using native PAGE.

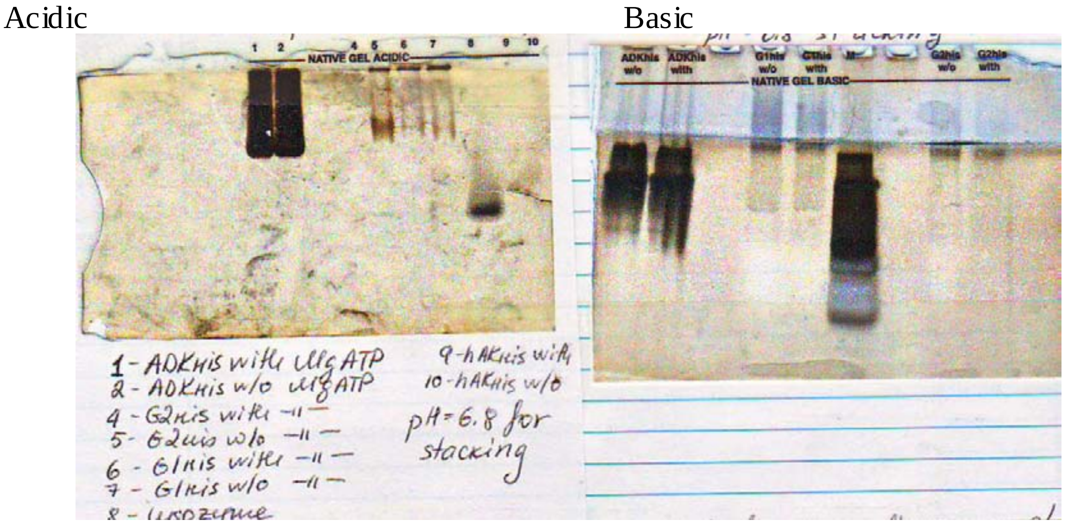


Figure 36A. Analysis of engineered protein constructs on the native PAGE in the presence and absence of the substrates. Basic and acidic gel conditions were tested.