A Study on the Structural and Evolutionary Aspects of Protein Modules

野坂, 通子
九州大学理学研究科生物学専攻

https://doi.org/10.11501/3054130

出版情報：九州大学, 1990, 理学博士, 課程博士
バージョン：
権利関係：

A Study on the Structural and Evolutionary Aspects of Protein Modules

野　坂　通　子

(1)

A Study on the Structural and Evolutionary Aspects of Protein Modules

by

Michiko Nosaka

Department of Biology,
Faculty of Science,
Kyushu University,
Fukuoka 812, JAPAN

ABSTRACT

A structural domain of a protein can be decomposed into modules, which are defined as compact segments. It has been established that in at least number of proteins the intron positions of an eukaryotic gene correspond well to some of the module boundaries of its coding protein. The module organization of a protein from one species is usually the same as that of the same protein from other species. These facts suggest that modules are fundamental units of protein structure and that some introns disappeared from module boundaries during evolution. In this study, modules of various globular proteins are identified by the most advanced method and are then analyzed as to their structure and their connection to the evolution of proteins.

First, adenylate kinase, an enzyme essential for life, is examined, and it is discovered that its module organization had changed during its protein evolution. Since this enzyme is assumed to have existed during early evolution, it is expected that this enzyme protein might provide some evidence of early protein structures. The modules of this enzyme are identified and the amino acid sequences of its isozymes are compared with each other, resulting in the location of a large gap on one of the module boundaries. The result means that the insertion or deletion of modules occurred during this protein evolution. This example definitively proves the evolutionary significance of modules.

Second, the distribution of module size, as determined by the number of amino acids in the module, is analyzed by examining 85 proteins with various lengths of from 36 to 498 residues. The average module size, which is found to be independent of the

protein length, is 15 residues. This size coincides with the length of its ancestral polypeptide, which has been inferred from experimental results. Furthermore, the size distribution of modules is found to be independent of whether the proteins are produced by eukaryotes or by prokaryotes. This result suggests that modules existed as structural units of proteins before the divergence of the two urkingdoms. In addition, a comparison of the size distribution of modules with that of exons derived from the 210 available genes demonstrates that most of the contemporary exons consist of two or three modules.

Third, correlations between modules and secondary structures of proteins are studied. Two clear tendencies are discovered. (1) Module boundaries occur more frequently on the β -structure than on the other secondary structures. (2) The average module size of a protein has a positive correlation to the helix ratio of the protein. Explanations for these tendencies are useful for the study of tertiary protein structure.

Fourth, the correspondence between the module boundaries and the intron positions of 24 proteins is statistically examined. The extensive results confirm the close relationship between the intron positions of a gene and the module boundaries of the coded protein.

This study makes it clear that modules are important both as structural and as evolutionary units of proteins. Moreover, it strongly supports the hypothesis that a number of introns of ancestral genes were lost during the evolution of proteins.

CONTENTS page

I. INTRODUCTION

During the evolution of protein, more tertiary structures (polypeptide folds) than primary structures (amino acid sequences) survived. Therefore, studies of these tertiary structures are likely to provide clues to the early stages of protein evolution. Hence, an analysis of the module structures of these tertiary structures could potentially be very effective for the study and understanding of protein evolution.

In the last decade, it has been established that the gene structures of eukaryotes consist of exons and introns. After transcription from a DNA sequence, introns are spliced out from the messenger RNA and only exons are connected to the mature messenger RNA, which is then translated into a protein. After the discovery of introns, Gilbert(1978) conjectured that introns have been useful in creating new proteins through their role in exon-shuffling. He hypothesized that exons are the functional units which assume various combinations in order to produce different proteins and that introns are the mediators of exon-shuffling. Blake (1978) pointed out the possibility that the split structures of eukaryotic genes might be reflected in the mosaic structures of proteins.

In 1981, Gō discovered the correspondence between a split gene structure and the protein module structure encoded by the gene. Distance maps, which express the amount of distance between every residue pair in a protein, permitted Gō's discovery of modules and provided the original identification method of modules. Modules are the structurally-compact units which compose the globular domains of proteins. Gō identified the four modules of the human hemoglobin subunits. Two of the three boundaries of

these four modules corresponded to the positions of introns in mouse hemoglobin genes. If each module corresponds to an ancestral exon, one intron which is absent in the contemporary gene must have existed in the ancestral gene as one of the module boundaries. Later, a study of a homologous protein, leghemoglobin, which is produced in the nodules of soy beans, disclosed that introns of this protein gene existed in these three module boundaries. (Jensen, et. al., 1981).

In the proteins analyzed up to now, the positions of introns and module boundaries generally correspond, and in the same proteins derived from different species, the module organizations are common. These results suggest that modules and corresponding introns are universal as to protein and gene structures, respectively and that most modules would appear to resemble their ancestors. Therefore, it is suggested that modules must have been one of the fundamental units of protein structures in the evolution and the refinement of protein functions. Gō's module hypothesis is that: (1) each ancestral coding gene, which would have been short, corresponds to a protein module, (2) introns have existed since the earliest stages of protein evolution, (3) introns have helped the evolution of protein by means of exon shuffling, and (4) some introns have disappeared, presumably because of the loss of their roles (Gō, 1985). In addition to the proposal that modules corresponded to ancestral "mini" genes, which were selected and gathered to produce a new protein, Gō also discussed the other structural characteristics of modules. For example, Gō pointed out that the hydrophobic and the hydrophilic residues of chicken egg lysozyme are respectively

localized within each module. This proved that the hydrophobic interactions between modules should assist in the assembly of modules (Gō, 1983). Gō also suggested the possibility that modules might have an effect on other observable features in native proteins, such as the balanced stability of the protein conformation and the flexibility of protein structure (Gō and Nosaka, 1987). Furthermore, the internal location of modules in several proteins was studied in order to understand the manner in which modules make up these protein (Gō, 1984, Gō and Nosaka, 1987, 1989). The functionally-important sites of proteins were also carefully inspected to determine how they localized on modules. These findings about individual proteins should be useful in understanding the general nature of protein architecture.

The main purpose of the present study is to provide the necessary biological and statistical evidence in order to clarify still further the general significance of modules to the structure and the evolution of proteins. Any changes in the module organization of related proteins will be examined to demonstrate the biological evidence that proteins have evolved altering the combination of their modules. Module size will be examined to evaluate the universality of different proteins. To understand the structural meaning of modules, a survey will be made of any relationship between modules and secondary structures, which are the sub-elements of a domain in the hierarchy of protein structure. Finally, the correspondence between module boundaries and intron positions will be checked by statistical testing to reinforce this correlation with the possible number of data and to grasp the practical degree of this correspondence. It is hoped that a study of the modules of

various proteins from these different perspectives will provide some clues to existing questions about modules and will suggest some topics for further examination.

II. METHODS

II-1 Methods of Module Identification

Since the discovery of modules, another method for module identification have been developed (Gō & Nosaka, 1987). The original method employs a distance map, which shows the distance relations of all residue pairs in a protein according to three degrees of distance: close, intermediate and distant (Gō 1981, 1983, 1985). Unlike the distance map, the refined method uses centripetal and extension profiles, which are calculated in different ways from the coordinate data set of proteins. The two profiles are called CP and EP, respectively (Gō & Nosaka, 1987).

(1) The Distance Map Method

A distance map of a protein represents two-dimensionally the distance relations between each pair of alpha carbons, the central atoms in amino acid residues. Modules of a protein are identified accounting the distant pairs on the map (Gō, 1981).

This method has limitation of its utility. In the case of relatively small proteins, such as Hemoglobin and Lysozyme, the module boundaries are located in the center of those proteins, i.e., residues on module boundaries are not distant from any other residues. Therefore, the algorithm of this method is useful in identifying the module boundaries of these small proteins, such as mono-layer proteins. However, this algorithm is not always effective to the module boundaries in larger proteins. Since there are many additional interactions between the modules of larger proteins, the distance relations of larger proteins are more complicated than those of smaller proteins. Therefore, some modification of the distance map algorithm is needed for

module identification of larger proteins which have either core modules (as found in carboxypeptidase A, Gō, 1984) or multi-domains. Another problem with this original method is that it is inevitably accompanied by some arbitrariness in describing module boundaries with residue numbers of proteins. These two limitations of the distance map method are resolved in a refinement of this method known as the centripetal and extension profile method.

(2) The Centripetal and Extension Profiles Method

To eliminate the general arbitrariness of the distance map method and to deal adequately with large proteins, centripetal and extension profiles were introduced (Gō and Nosaka, 1987). A centripetal profile indicates the central locality of module boundaries in a protein; by contrast, an extension profile characterizes the compactness of modules. Each of these profiles has been defined according to the following observations: (1) module boundaries exist in either the center or the local center of a protein, so they are not distant from other neighboring residues; and (2) because the residues within an identified module are not distant from each other, modules have structural compactness. Therefore, module boundaries should show relative extendedness.

To calculate these profiles, the "window length", or the searching range of distance relations along a peptide main chain, must be specified. Ideally, the window length should determined in a self-consistent way to get the most accurate results possible. The best window length should be defined by considering with each module size to be identified; however, the

module sizes of a protein vary considerably. Hence, when the window length is applied to numerous proteins which diverge both as to functions and degrees of specificity, a series of window lengths must be applied in both profiles in order to avoid missing the module boundaries of the proteins.

(i) An Explanation of the Two Profiles;

The Centripetal Profile

Gō defined the centripetal character of the i-th residue, F_i, as the average of the squared distances between the i-th residue and every residue existing within a range of (2k+1) residues along a peptide chain, that is from the i-k to the i+k residue:

$$F_i = \sum_{j_1 \leq j \leq j_2} r_{ij}^2 \; / \; (j_2 - j_1) \qquad (1)$$

where j_1 = MAX(1,i-k), j_2 = MIN(n,i+k), and n is the total residue number of the protein. The centripetal profile (CP) is the graph of F_i versus the location of residue i. The residue at which F_i is the local minimum indicates that this residue is not far from other neighboring residues along the peptide main chain, suggesting that the i-th residue is in the local center of the protein. Every local minima of the function F (summation of F i) is, hence, a potential module boundary. Here, adequately-smoothed profiles are used to eliminate the effect of trivial or irregular changes in the profile. Figure 1-(a) shows a series of smoothed centripetal profiles of TIM (triose phosphate isomerase) protein. The horizontal and the vertical axes represent the residue number and the centripetal index F, respectively. The arrows indicate the local minima of centripetal profiles. The local minima of the profiles represent the results produced by the distance map in Figure 2 to within a few residue differences.

The Extension Profile

Observations from distance maps indicate that modules are expected to show structural compactness. In other words, module boundaries are relatively extended. Gō (1987) introduced an extension profile for a protein. The extension profile consists of the extension indexes which are defined to each residue. This index E_i is the average of the weighted square distances, where the average calculation involves the distances between every pair of residues along a peptide main chain that are within a limited span of the i-th residue. The extension index for the i-th residue, E_i, is defined as follows;

$$E_i = \frac{1}{(j_2-j_1)(j_2-j_1+1)} \sum_{j_1 \leq j \leq j_2} G_{mj} \qquad (2)$$

and

$$G_{mj} = \begin{cases} r_{mj}^2 & \text{for } j - m \leq k \\ r_{mj}^2 \ / \ (j - m) & \text{for } j - m > k \end{cases} \qquad (3)$$

where: j_1 = MAX(1,i-k), j_2 = MIN(n,i+k), n is the total number of residues of the protein, and k is the number of windows; and r_{mj} is the distance between the alpha carbons of the j-th and the m-th residues. The extension profile (EP) is the graph of E_i in comparison to the location of residue i. Since module boundaries have an extended form, they are near the local maxima of the extension profile. In other words, identified modules would not have an extended form in the middle of their structures.

The window size for the extension profile should also be optimally chosen. As in the case of centripetal profiles, a series of extension profiles with ten window size (k) varying

from 10 to 20 residues is monitored. The window sizes used for this profile are smaller than those used for the centripetal profile. The compactness of a local segment is checked directly by examining the local maxima of this profile. Figure 1-(b) shows a series of extension profiles of TIM protein, where the horizontal and the vertical axes show the residue number and index E, respectively. The arrows indicate the location of typical local maxima of the profiles, which are in accord with the local minima of centripetal profiles (a). Here, identified segments do not have strongly extended form in the middle of their chains.

Refinements of This Method

As mentioned before, it was difficult to choose the best window length in the module identification procedures because it was needed further investigation to the characters of these profiles. Thereby, the module identification of new proteins by this method were achieved by a lot of searching procedures with various window lengths of these two profiles before this study. Hence, a standardized procedure is required for this method in the next progressing step.

(ii) Procedures for Module Identification; Refinements

The comparisons between these profiles of various window length and module boundaries identified on their respective distance map is surveyed in order to refine this method and the following progresses are completed. First, each of a series of optimal window length of these two profiles is determined Second, a new index for the module determination from local minima of centripetal profiles is introduced. Finally, a

standardized procedure for module identification is established. With these refinements module boundaries are identified more objectively and more rapidly.

The conditions of these two profiles for various proteins were examined in detail. The window size of a centripetal profile is chosen which covers twice the length of any module. This length is based on the earlier study of the size distribution of modules and on the investigation of the CPs of several proteins. A series of seven window lengths (15, 20, 25, 30, 35, 40, 45 residues) is used as the standard.

Since the positions of the local minima of the centripetal profile vary a little according to the searching window length, the most probable positions are selected. For this purpose, an index I(i) is introduced which is defined as the total number of local minima counted over the examined profiles (of k windows) within three residues, the i-th residue itself and its two nearest neighbors:

$$I(i) = \sum_{k} \sum_{i-1}^{i+1} \text{(the number of local minima counted in CPs)}$$

The residues with indexes of larger than four are candidates for module boundaries. If these candidates are close enough to each other, they are further combined into one according to their index numbers. (If necessary, the value of index F is taken into secondary consideration.)

The comparison between the distance map method and the centripetal and extension method provides that the centripetal profile is essential for the identification of modules. Therefore, the centripetal profile is applied first and the extension profile is monitored. Candidates for module boundaries

are selected from all of the local minima observed in the series of centripetal profiles. They are then checked as to the compactness of their tertiary structures by means of a series of extension profiles of the protein. Although most of these candidates can be readily identified by means of centripetal profiles, there are some cases in which it is difficult to locate clear boundaries. This situation occurs either when more than two stable local minimum points are detected within a six-residue span or when the index F of centripetal profile is relatively high. In such a case, the lowest number of module boundaries are defined by selecting the most reliable points from neighboring stable residues and, by regarding the stable points with a higher value of index F as non-candidates.

The differences between the results from the distance map method and the results from the refined method are easily understood. The distance map method selects module boundaries by weighing the distant relations over the total length of a protein, whereas the refined method deals equally with the distance relations of a finite number of neighboring residues (by using window sizes). Additional boundaries, which can not be distinguished from the originally identified boundaries on a distance map, can be detected by the refined method because of the clarity and stability of CP minima. To satisfy the criteria for module boundaries in various proteins is so difficult in some cases that only clear and stable minima of CPs are employed as module boundaries. Therefore, only the most certain module boundaries and modules are discussed in this study.

To account for variations from the results of the distance

map method, a new method is developed for module identification. The applicable lengths of given parameters of the two profiles are established. A new index of centripetal profile is also introduced in order to locate module boundaries with total objectivity from a series of the window lengths. With these refinements, the centripetal and extension profiles method not only can deal with larger proteins but also complete the identification procedure more objectively and more rapidly than the distance map method used originally.

II-2 Calculation of Phylogenetic Distance

According to the alignment of amino acid sequences, the evolutional distance D between every pair sequences is calculated by using Jukes and Cantor's formulation (Jukes and Cantor, 1969):

$$D = - \frac{L - 1}{L} \ln \{ (L*s - 1) / (L - 1) \}$$

where L is taken as 21, the number of amino acid species plus one, regarding the insertion or deletion the another kind of amino acid. "s" is the similarity which is expressed by the ratio of the counted number of invariant residues to the total number of aligned residues. The phylogenetic tree of aligned AK sequences is constructed from these calculated distances.

II-3 The Modified UPGMA Method for the Phylogenetic Tree

In order to estimate the time when the insertion or deletion of modules occurred in adenylate kinase family, the phylogenetic tree is constructed by a modified UPGMA method, in which no assumption of constant evolutionary rate is made (Tajima and Nei, 1984, Lee, 1981).

III. RESULTS

Using the refined method, research has been undertaken in four areas in order to better understand the roles of modules in the structure and the evolution of proteins. The first study establishes the evolutional change of module organizations, confirming the importance of modules to the evolution of protein. The second survey accomplishes the distribution of module sizes over 85 proteins, demonstrating the universality of modules in protein structure, and then it compares these results with the distribution of exon sizes over 210 genes, suggesting the most probable combination of module-size segments. The third investigation detects two correlations between modules and the secondary structures of proteins, providing any of other possible structural meaning of modules. The final study statistically confirms the correspondence between module boundaries and intron positions of the 24 proteins currently available.

III-1 The Insertion or Deletion of Modules in the Adenylate Kinase Family; Structural Differences Based on Modules

Adenylate kinase is a ubiquitous protein in nature (Noda, 1973). This enzyme catalyzes the transition of the phosphoryl group from an ATP (adenine-tri-phosphate) to an AMP (adenine-mono-phosphate) and produces two ADP (adenine-di-phosphate) molecules (although in one case, a GTP (guanine-tri-phosphate) is substituted for one ATP). ATP molecule is the material of genetic nucleotides and, at the same time, it is the typical energy carrier for organisms. An ATP releases free energy through the hydration of a phosphoryl moiety, where an ATP become an ADP and

an free phosphoryl molecule. An AMP is the form taken when another phosphoryl group are further released from an ADP molecule. ATP, ADP, and AMP molecules work also as the control signals for cell metabolism. It is well known that each of these three molecules have allosteric effects on the enzymes of the glycolytic pathway. It should be recognized that the ratio of these three adenine nucleoside contents in a cell have a strong influence on the cooperative control of cell metabolism.

$$ATP + AMP \xleftrightarrow{Mg^{++}} 2\ ADP$$

Under biological conditions, adenylate kinase catalyzes this reversible reaction with magnesium cation (Mg^{++}); therefore, this enzyme can accommodate the balance of these nucleotides contents according to the cell situation. In energy carrying system, adenylate kinase can catalyze the reproduction of molecules from ADP either AMP or ATP depending upon their relative levels of concentrations. Conversely, it also has the ability to create an ATP and an AMP molecules from two ADP molecules in a low concentration of ATP. Since adenylate kinase plays such a key role in the control of life metabolism, it is believed that this enzyme has been an essential protein to life since the beginning of evolution.

(1) Isozymes and their Amino Acid Sequences

There are two isozyme groups of adenylate kinases. These are different as to amino acid length. Although the kinetics of these groups are almost the same in higher organisms, these groups are coded in independent genes and are expressed differently (Frank, et.al., 1984, Povey, et.al., 1976, shows, et.al., 1975). Short types of adenylate kinases exist abundantly in higher organisms

in cytosols of muscle cell, brain cell, and red blood cell, while long type enzymes localize in either the inter-membrane or the matrix of the mitochondria of the other cells or in the protoplasms of primitive organisms. It should be noted that the mitochondria are the energy-producing organella, while muscle, blood, and brain cells are energy-consuming rather than energy-producing system.

Ten amino acid sequences of adenylate kinases have been reported; they are located: in the cytosol of bovine muscle (Kuby, et.al., 1984), in the inter-membranes of bovine mitochondria (Frank, et.al., 1984), in the matrix of bovine mitochondria (Tomasselli, et.al., 1980, Wieland, et.al., 1984), in the cytosol of yeast (Prova, et.al., 1987, Tomasselli,et.al., 1986), in the cell of E. coli (Brune, et.al., 1985), in the cytosol of human muscle (Von Zabern, et.al., 1976), in the cytosol of rabbit muscle (Kuby, et.al., 1984), in the cytosol of porcine muscle (Heil, et.al., 1974), in the cytosol of chicken muscle (Kishi, et.al., 1986), and in the cytosol of carp muscle (Reuner, et.al., 1988). These adenylate kinases will be referred to as: AK1B, AK2B, AK3B, AKY, AKE, AK1H, AK1R, AK1P, AK1C and AK1F, respectively. Except for AKY, the cytosolic AKs are short enzymes existing in muscle, while the other AKs, which belong to the long isozyme group, either exist within the mitochondria in common cells of higher organisms or within the cells of E. coli. These ten amino acid sequences are compared with each other and the evolutional relations among them are evaluated in the study of the module structure of porcine muscle cytosolic adenylate kinase.

(2) The Module Structure of Adenylate Kinase

Porcine adenylate kinase (AK1P), which is the only enzyme that has been submitted to the tertiary structural data bank, is composed of at least 14 modules, possibly 16. Figure 3 shows the centripetal profiles of porcine adenylate kinase, where the horizontal and the vertical axes show the residue number and the index F, respectively. The arrows indicate the identified module boundaries, and the two white arrow heads with dashed lines illustrate possible additional boundaries. The conditions of these two positions should be observed carefully. According to the importance of modules in protein evolution, it is expected that the position of a large alternation would occur at some of these identified module boundaries.

(3) The Alignment of the Ten Sequences in the Adenylate Kinase Family

Two alignments between long and short types of the amino acid sequences had been reported earlier (Brune, et.al., 1985, Frank, et.al., 1986). They determined sequences in the long type of isozyme, and they noted a large gap (an insertion or deletion of amino acids) in the middle of the sequences. However, the reported positions of this gap differed each other. In the present study, thereby, an alignment of these ten sequences is achieved and the position of a large gap is located at residue number of porcine AK. This can be confirmed by means of a simple classification of all amino acids according to the universal codons in table 1. Because the process searched here is as old as the establishment of the genetic codons, this grouping of all amino acids into four categories is assumed to be sufficient to

confirm the position of the gap.

This classification is based on the four species of the second nucleotide in the universal codons. It should be noted that the chemical characteristics of these amino acids are strongly coincident with the groups discriminated in the second codons. If the second codons are U, only hydrophobic residues (phenylalanine, leucine, isoleucine, valine and methionine) are referred to, and if the second codons are A, almost all of hydrophilic and potentially hydrophilic residues (histidine, glutamate, glutamine, aspartate, asparagine and lysine) are coded. All of the degenerations of codons for amino acids except serine are coincident with this applied classification. The mutational tendencies between two amino acids during evolution also seems to support this grouping (Dayhoff et. al., 1978). Except for some of the chemically-important changes to contemporary proteins, the mutation rate between two amino acids within a group is generally higher than the mutation rate between two amino acids from different groups. These are good reasons for this grouping category based on the second codon selectivity for amino acids. The first codons do not show such a distinctive correlation as to either the chemical similarity or the evolutional tendency of amino acids. The third codons, as is known as the wobble of codons, provides only a very weak specificity for amino acids.

In the sequence comparison, the relationships between two amino acids from different sequences are described with four situations according to this classification: being identical, belonging to the same class, belonging to either one of two classes (only in the

case of Serine), or belonging to different classes. Each alignment between two sequences is confirmed on a contrast map, which expresses the similarity of every pair of amino acids in comparing sequences (data are not shown). Though the position of the large alternation can be assigned at 132 or at 138 in the residue number of porcine AK, the position of 132 is preferable according to the super imposed-analysis of the two structures of porcine AK (short type) and yeast AK(long type) (Egner, et.al., 1987). These sequences are aligned regarding the conservation of functionally important residues. Figure 4 shows the alignment of the ten sequences, in which either a large deletion or a large insertion exists on residue number 132 of porcine AK. Here, AK3B, AK2B, AKY, AKE, AK1F, AK1C, AK1R, AK1P, AK1B, and AK1H are adenylate kinases in: bovine mitochondria matrix, bovine mitochondria inter-membrane, yeast cytosol, E. coli, carp muscle, chicken muscle, rabbit muscle, porcine muscle, bovine red cells, and human muscle, respectively. Interestingly, there are deletions of more than four residues at the two possible module boundaries (102 and 138). This suggests that these possible two might have been the clear boundaries.

Figure 5 summarizes the module organization of porcine AK, in which the position (at residue number 132) and the size of the inserted or deleted segment (26 residues) is shown. The enzyme consists of at least 14 modules (M1 - M14). Stick and ball models of this enzyme is drawn in Figure 6 from BNL atomic coordinate data set (in stereo views from different directions). Each arrow in (a) and (b) indicates the position of a large alteration which is near the top of the wall forming a large cleavage. The 26 residue segment, which is included only in the long isozymes,

covers a part of this cleavage in AKY (Egner, et.al., 1987). An example of structural change of protein evolution based on module structure is, therefore, proposed in adenylate kinase family.

(4) The Intron Position of an Isozyme as Support for the Insertion or Deletion of Modules

None of the introns in the AK1 genes of chicken and human, which have been available up to now, is located at the position of the large gap. However, an intron of long type isozyme (AK3B) gene does exist on the boundary. This explains the participation of introns in the evolution of module organization, i.e., the shuffling of small exons. Therefore, the existence of this intron supports the possibility of insertion or deletion of modules during the evolution of protein. (Suminami, et. al., 1988, Matsuura, et. al., 1989) (Nakazawa, et. al., personal communication)

(5) Estimation of the Time of the Incident

According to the alignment of the ten sequences, the identity of each pair from comparing sequences is calculated according to the distances. Table 2 shows the identity of each sequence pair in the lower half and the compared residue number of each pair in the upper half. Small deletions are taken into consideration in the calculation of its identity as another kind of amino acid. The calculated parts of these sequences which are common in the ten sequences are about 80 percent of the total length of the short type sequence. Because the lowest identity is still more than 28 percent, it is concluded that all of these sequences have a common ancestor. As a result, the phylogenetic tree of these ten AK sequences can be constructed by the modified UPGMA method in order to estimate the time when this situation

emerged.

Figure 7 shows the phylogenetic tree of of the adenylate kinases. AK3B, AK2B, AKY, AKE, AK1F, AK1C, AK1R, AK1P, AK1B, and AK1H are the same as explained in Figure 4. The short type enzymes make a cluster on this tree and are supposed to have diverged from the long isozymes by gene duplication in early stages of this protein evolution. Divergent point 1 shows the gene duplication and it, as well as the other points with numbers (2, 3 and 4), is a possible divergent point of prokaryotes and eukaryotes. The short type adenylate kinases are in accord with the species divergence, whereas the long type enzymes have some complexity. Adenylate kinase of bovine AK2(AK2B) is nearer to that of yeast cytosol(AKY) than to any other AK. If AKY is originally coded by the mitochondrial gene in yeast, an organella which had come from a prokaryote, the divergent points of the two urkingdoms is 1. Otherwise, the divergence time is at any point of 2, 3 or 4. Therefore, the gene duplication of the two AK isozymes occurred before the divergence of eukaryotes and prokaryotes or happened at about the same time as the divergence of the two kingdoms.

(6) Function of the Additional Modules in the Long Isozymes

In spite of the structural differences, the kinetics of all isozymes are almost the same. In their structural data where AK binds a substrate-analog, Ap5A (P1,P5-di(adenosine-5'-) penta-phosphate), Egner, et.al.(1987) discussed the meaning of this additional part of yeast AK which is not included in Short type AK that. Since this substrate analog was buried in yeast enzyme, they theorized that this segment might cover the

substrate after induced fit motion of the enzyme. Their observation makes sense. If true, it means that modules must participate in its induced fit action. Furthermore, the chemically-active function of this segment can be expected because the amino acid sequences of the additional modules in long adenylate kinases are well conserved. The existence of Histidine in this segment seems to be very important. Histidine is such a weak base (PK=6.0) that it reacts as an electron donor only in strongly acidic conditions, suggesting that histidine is optionally active in the hydrophobic environment formed after the induced fit of this enzyme. This idea seems to be supported by the alignment of this region presented in Figure 4. The neighboring lysine residues, which are strong electron acceptor and supposed to provide a stabilizing effect for phosphorous anions, are not so strictly conserved in long AKs as the same residues in short AKs. Therefore, the difference between long and short isozymes seems to depend on their induced fit forms and on their environments.

(7) Classification of the Large Alteration as Insertion or Deletion

In considering whether the large alternating part was inserted or deleted in the AK family, it seems probable for several reasons, that the large segment had deleted from a long type AK to merge with short type AKs. No short AK has yet been found in any primitive organism, while it does exist in the cytosols of muscle, brain and red blood cells of animals where the biological conditions are highly specified to consume energy. In addition, all of the cytosols are completely isolated or

localized from the energy delivery system, or digestive organs. That is, they are localized in the periphery of animal bodies. These cells are so specific that the conditions within these cells may be constant and/or simply compared to those of other cells. Therefore, the functional mechanism of short adenylate kinase would be simpler than that of the isozyme in primitive cells. Moreover, this situation can explain the covering function of the additional modules in long AK. Since long isozymes show higher affinities to ATP molecules than do short isozymes, this substrate seems to be bound immediately and rapidly isolated from a mixture of various molecules in the cells.

III-2 Module Size

The size distribution of modules is studied by the further improved method which includes the additional smoothing. This additional smoothing has been developed for the first step of the automatic identification of modules. Before beginning this study, this additional smoothing process is checked, using 29 non-homologous proteins, whether or not it represents the original results obtained by the distance map method.

(1) Additional Smoothing and Comparison between Original and Improved Methods

The module boundaries which are difficult to locate are more reasonably and rapidly dealt with by an additional smoothing procedure. The detailed procedures are explained elsewhere (S. Tomoda, M. Nosaka, and M. Gō, in preparation). Only the clear boundaries identified with this additional smoothing are analyzed in this section.

In order to check the adequacy of this newly improved method, the least numbers of module boundaries identified by the distance map method are compared with the least number of module boundaries identified by the improved method. Table 3 lists the 29 proteins examined. They are all globular, with less than 200 residues. Sixteen of these proteins are proteins from eukaryotes, twelve are those from prokaryotes, and one is the protein from bacterio-phage.

Table 4 summarizes a comparison of the module boundaries of the 29 proteins determined by the original method with those identified by the improved method. When the most strict criteria were applied, only 146 boundaries were identified by the distance map method, while 200 boundaries were detected by this further improved method. Of the 146 boundaries detected by the original method, 143 are also detected and the other three were weakly identified by the new method. After sufficient consideration, the additional 57 boundaries detected by the new method but missed by the distance map method have now all been accepted as module boundaries by us. As shown in Table 4, average module size of these 29 proteins becomes smaller than the old estimation. Each of the three exceptions exists in the middle of polypeptide chains of three different proteins (ribonuclease A, aspartate carbamoyle transferase and cytochrome S-C-2). Figure 8 shows the frequency of the absolute differences in the common results of the original and of the refined method. The horizontal axis shows the difference expressed in the number of residues and the vertical axis shows the number of differences. The new results represent 89 percent of the old results within an error range of $\pm$ 2 residues and the average difference of these corresponding

143 boundaries is 1.2 residues. This degree of accuracy suggests the suitability of representing the old results by the refined method.

New results includes almost all boundaries which are identified as the most certain case on their each distance map. This method identifies more additional module boundaries, which are accepted by the newest consideration. Hence, I conclude that the refined method is more efficient in detecting module boundaries than the old method and that this additional smoothing is useful for the statistical analyses of modules.

(2) Distribution of Module size

The variation or the uniformity of module lengths is surveyed for the 85 proteins whose peptide lengths varies from 36 to 498 residues. Table 5 is the list of these proteins, where code is the BNL code name of each protein and size is the total length of it. The number of proteins from eukaryotes, from prokaryotes and from virus and phages are 49, 30 and 6, respectively. Some proteins with the same name, such as cytochrome C, are different from one another both in their amino acid sequences and in peptide lengths.

Figure 9 shows the distributions of module size, where the horizontal and the vertical axes represent the module length and the frequency of each length, respectively. Here, (a) is the total distribution of identified 1065 modules, (b) is the total distribution of internal 650 modules which do not contain N and C terminal modules, (c) is the distribution of 347 modules from eukaryotes, (d) is the distribution of 68 modules from prokaryotes, and (e) is the distribution of modules from virus

and phage, respectively. Since this improved method detects additional module boundaries, the size distribution of modules is smaller than in the previous study (Gō and Nosaka, 1987). All of these distribution patterns are similar to one another. Table 6 summarizes the results from each of these three sources. Although virus and phage proteins have smaller modules than the other two protein groups, it is proper that the modules of all proteins are universal.

The following two results can be deduced from the observations of the size distribution of modules; 1. Module size of these 85 proteins vary from 5 residues to 34 residues, indicating the uniformity of module size distribution. The average size of these modules is 15 residues, possibly implying the original state of ancestral modules. 2. The pattern of module distribution is common among the three protein groups, prokaryotes, eukaryotes, and viruses or phages. Taking into account that similar proteins from different species have the same organizations of modules, this fact suggests that modules are fundamental units of protein structure and that modules and corresponding introns existed before the divergence of prokaryotes and eukaryotes.

The variation of module size in a protein often shows a smaller range than the variation of total module size. Hence, It is worthwhile to examine whether the average module size of a protein depends on any characteristics of its protein. Figure 10 shows the relation of a protein module's average size to the protein's total length. The horizontal and the vertical axes show the total peptide length and the average module size of the protein,

respectively. Both are expressed in number of residues. Although there is no significant correlation between them, this does suggest that modules are relatively uniform as to their size and are independent of the protein's length. In smaller proteins of less than about 100 residues the average module size varies from 10 to 22 residues, while in larger proteins the average module size is within the more narrow range of from 14 to 18 residues. Since the results indicate that some proteins with larger modules, which are marked with a circle in Fig. 10, are extremely helix-rich, the correlations between the average module length of proteins and the secondary structure contents in the proteins are studied in section 3.

(3) A Comparison between the Size Distributions of Modules and Exons

The size distribution of modules has been compared with the size of distribution of exons compiled from 210 genes of independent exon organizations. Table 7 lists these genes, which codes various proteins of from 60 to 1772 residues as to amino acid length, showing their reference and author. Since the selection is not limited to only non-homologous proteins, some proteins of the same super family are included. However, only the gene mostly divided by introns have been selected from homologous proteins which have identical functions.

Figure 11-1 shows the size distribution of 1056 exons from the 210 genes, where the horizontal axis shows the exon length in amino acid numbers and the vertical axis represents the frequency of each exon length. Only the peptides coding exons are compiled and the N- and C-terminal exons which include untranslated parts

are not accumulated. Exons which are longer than 600 nucleotide length are not shown because such exons are small in number and exist dispersively. Only 13 of large exons are in the middle of these genes, 10 are N-terminal and 13 are C-terminal exons. Naora(1984) and Hawkins(1988) also reported that large exons of more than 600 nucleotide are rare. It is worthwhile to notice that the size distribution of the exons is in accord with the size distribution of the modules. The small exon part of this distribution is similar to the distribution of modules and the larger exon part can be regarded as the combination of the several distributions of the connected modules.

The size distribution of exons shows a broad and symmetric shape, whose peak is near 40 residues and whose average length is 46.7 residues. Attention should be paid to the fact that only a small part of this exon distribution is in accord with the size distribution of modules. As a result, the exon distribution can be explained as the distribution of exons made from segments which correspond from one module to several number of connected modules. Assuming that all exons are composed of small segments which code several number of modules, the best fit distribution of segments is calculated using the connected module distributions (Figure 11-2). These distributions are produced by the convolutions of the size distribution of modules. Figure 11-2 shows the main part of the size distribution of exons (a) and the model distributions derived from the size distribution of modules. Each of the horizontal and the vertical axes are the same as in Fig.11-1. Here, (b) is the best fit combination of five distributions of segments which are convoluted from one to five modules and (c) is the distribution of the convoluted

segments using a random numbers of modules. A comparison between (a) and (c) suggests that modules did not randomly accumulate to make an exon and introns did not randomly delete. The most effective distribution is that of three modules. Therefore, most abundant exons code a peptides of three module length. This number is coincide to the ratio obtained from the comparison between the intron positions and the confirmed module boundaries.

This study supports the idea that exons have been produced by the ligature of small segments which had coded ancestral modules of proteins at an early stage of evolution. Furthermore, the existence of the smallest exons of about eight residues in contemporary genes gives hint about one unresolved aspect in module identification, how to deal with small segments selected from the module identification procedure. Small modules which have emerged from the new, or additional, boundaries must surely be accepted as modules because small exons of about eight residues exist in nature. Yet, it should not be forgotten the fact that other small modules which might be produced under certain additional conditions should also be considered carefully in module identification.

The fine structure of the size distribution of exons actually appears to embrace several small distributions. Because some evolutional changes, such as the deletion of amino acids, might well have happened as a result of the fusion of modules and/or the specification of protein structures. It is hard to analyze the connected exons of the larger exons because the more detailed conditions of this fine structure have not been here.

III-3 Modules and the Secondary Structures of Proteins

Proteins have a hierarchy in their structures, which is presumably an adaptive feature of organisms to assist their protein synthesis. Studies of the correlation between module and this hierarchy and may well be important in studies of the protein evolution. The hierarchy of protein structures consist of: (1) the primary structure (amino acid sequence), (2) the secondary structures (helices, β -structures, turns and random coil), (3) the tertiary structure (the folding of a peptide chain), and (4) the fourth structure (the configuration of all peptide chains constructing a whole protein). It could be claimed that a new aspect of this hierarchy must be recognized with the discovery of modules, which construct domains of proteins. Gō described modules as compact elements, whose position in the hierarchy was under the tertiary structure and above the primary structure. Although it has been agreed that modules were different from the secondary structures, a stricter relationship between modules and secondary structures had not yet been made clear. It is likely that the formulation of a clear relationship between modules and some secondary structures would bestow useful information on studies of the structure and the evolution of proteins.

Several correlations between modules and the secondary structures have been surveyed in order to understand the structural importance of modules to the protein hierarchy. To avoid any unnecessary differences in secondary structures created by procedural inconsistency among the authors, each secondary structure is identified by using the Miyazawa-Gō method, which

gives each secondary structure a numerical definition. As a result, two interesting correlations between modules and secondary structures are revealed: module boundaries occur on β -structures more frequently than in other secondary structures, and there is a positive correlation between the average module size of a protein and the helix ratio of the protein. The former correlation is clearly observed in 24 proteins whose modules were identified strictly in section III-4, while the latter correlation is found in the 85 proteins studied in section III-2.

(1) Module Boundaries and the Secondary Structures

Table 8 lists those 24 proteins whose tertiary structure and whose gene structure have been already established. The correlations between module boundaries and the secondary structures of 24 proteins are expressed in Figures 12. The horizontal and the vertical axes show the secondary structure ratios of the proteins and the corresponding secondary structure ratios only on the module boundaries in the proteins, respectively. The helix ratios are shown in (a), the β -structure ratios are shown in (b), the turn ratios are shown in (c), and the random coil ratios are shown in (d). Lines with slope 1 indicate the standard situation where the module boundaries has no preference for the secondary structure. In these 24 proteins, the helix and the random coil have no preference to module boundaries, while β -structure has about double the preference for the boundaries as is likely in a case of standard situation. Turn seems to avoid becoming module boundaries. The tendencies of turn and β -structure can be explained by the observation that, while module boundaries often occur in buried or less accessible

area of proteins, turns usually exist on the surface of proteins and β -structures usually exist in the core parts of proteins. Since the ratio of turns to module boundaries is extremely low in a protein and its degree of the succession in a protein is also smaller than in other secondary structures, however, it can not be concluded that this tendency on the part of turns is significant. It is, however, conceivable that β -structures are more dominant in occupying module boundaries than are the other secondary structures.

(2) Average Module Size and the Ratios of the Secondary Structures

The correlations between the average size of one protein modules and the secondary structure ratios of the protein listed in Table 5 are demonstrated in Figures 13. The module boundaries of these 85 proteins are identified by further improved method. Each of the horizontal axes shows the average module size of a protein and each of the vertical axis shows the helix ratio of the protein (a), the β -structure ratio of the protein (b), the turn ratio of the protein (c), and the random coil ratio of the protein (d), respectively. The average size of one protein modules correlates positively to the helix ratio of the protein. This correlation function is 0.56, which is entirely significant for all 85 proteins. However, the average size of one protein modules shows only weak negative correlations to the ratios of turns and random coils and shows no correlation to the ratio of β -strand.

In the previous study of small and globular proteins whose module boundaries were defined by the distance map method, the

tendencies of the secondary structures other than the helix were different from these results. The ratios of the β -structure and the random coil in a protein demonstrated a weak and negative correlations to the average module size of the protein and the ratio of turn showed no correlation to the average module size (the data is not shown). Taking into account the fact that helices and other secondary structures in proteins exist in contrary conditions to one another in protein structures, it can be concluded that only the helix ratio of a protein correlates significantly to the average module size of the protein.

Secondary Structures Identified by the Miyazawa-Gō Method

In order to survey the relation between modules and secondary structures universally, the Miyazawa-Gō method was used for the identification of secondary structures (Miyazawa and Gō, 1981). Although this method applies numerical definitions to all of the secondary structures, the identified results of some proteins are extremely different from those of the BNL data record, which are defined according to the personal idiosyncrasies of different researchers of protein structures. This is because of the inexactness of existing definitions of secondary structures. Thereby, the Miyazawa-Gō method has the tendency to ignore some β -structures and terminal residues of each helix structure in the BNL data, and turns are frequently identified instead. As a result, the weak correlations of non-helix structures to the edge of the helix must be carefully evaluated.

III-4 A Statistical Examination of the Correspondence between Module Boundaries and Intron Positions

The correspondence between module boundaries and intron positions have previously been demonstrated for hemoglobin, lysozyme, triose phosphate isomerase and other proteins (Gō, 1981, 1983, 1984; Gō and Nosaka, 1987). As the genetic data and structural data of proteins have increased, many other proteins have been available for study. The 24 proteins in which both the tertiary structure and the coding gene structure are known are examined (see Table 8). The three dimensional structural data was supplied by the BNL data bank and almost all of the gene structures cited come from either EMBL, PROTEIN or GENBANK data bases.

Before their examination, the intron positions of proteins must be renumbered according to the alignment of two sequences from their genetic data and from their tertiary structural data, for the amino acid sequence of genetic data is usually different from the amino acid sequence of the X-ray data. In most cases, species of both the sequences are seldom identical. Here only the proteins with known structure whose sequences can be aligned with genetic sequences are used. In selecting introns from more than two genes of a protein, all the introns are taken into consideration, and introns which are nearer in phylogeny to the source of structural data are weighted against those on corresponding positions of the respective genes. Terminal introns which correspond to the N- or C-terminal of a protein are excluded from the following analysis. There are two types of module boundaries, one of which can be compared with intron positions and the other which introns are not located on in the genes.

(1) Module Boundaries Matching to Intron Positions

Figure 14-(a) shows the distribution of the deviations between module boundaries and each corresponding 125 intron positions of the 24 protein genes. The horizontal axis shows the deviations between the intron positions and the nearest module boundaries. The vertical axis illustrates the counted numbers of each deviation. Introns which interrupt codons after the first and the second bases are plotted together with introns which interrupt codons after the third bases. The most frequent deviation is 1 residue and the deviations smaller than 2 residues are dominant (53 percent of all deviations). The deviations which are larger than 6 residues have very low frequencies.

The preference of intron positions for module boundaries is also supported by the comparison with the distribution of Figure 14-(b). The horizontal and the vertical axes are the same as for Figure 14-(a). This Figure shows the probable frequency of the each deviation in the case that 125 introns are inserted into genes independently of module boundaries. To calculate this probability from simulation, the 11 intron positions of catalase (8CAT in BNL code) and the size distribution of modules in section III-2 are employed as the typical intron positions and the material of randomly connected peptides, respectively. An χ 2-test of these frequency distributions, (a) and (b), significantly rejects at the 1 % level the assumption that introns exist independently of module boundaries.

This study support numerically the intimate relationship between module boundaries and intron positions. It is also notable that, even in the case where the different gene structures of a protein are known for more than two species, most of the

intron positions correspond well to the module boundaries. In other words, not only the common introns but the other introns are located correspondingly to the module boundaries. These phenomena are observed in the genes of triose phosphate isomerase, alcohol dehydrogenase, and many other genes.

Firstly, the introns which exist near the respective module boundaries with a deviation of less than 2 residues are discriminated against the other introns. As shown in Figure 14-a, these deviations are dominant. Furthermore, the fluctuation of module boundaries depending on the identifying methods is $\pm$ 2 residues as estimated from Figure 8 and the average module length is about 15 residues. Thereby, the discrimination of less than 2 residues seems to be preferable.

There are 59 introns which exist more than two residue distant positions from the nearest module boundaries. Of these 59 introns, 20 positions are within a deviation of 4 residues from their respective module boundaries and do not include any other introns in the middle of the modules divided by these 20 introns. Some of these 20 introns may be within the error range of module identification (in the case of their helix structure) and others may have slightly changed for each protein condition. These introns would be explained by the second possibility, i.e., the sliding of introns. 22 other introns are near the respective local minimum points of the CP which have not been selected as module boundaries. It is reasonable to think that these introns satisfy the first possibility by having changed during their protein evolution. These 22 introns might show the positions where module boundaries had become unclear after some structural

changes, insertion/deletion, or substitution of amino acids, as the result of the refinement of protein functions or other factors, such as the fusion of modules. Whether these local minima should be admitted as module boundaries with some additional conditions has not yet been resolved. The Remaining 17 introns do not have any local minima in their centripetal profiles as long as they are examined with standard window lengths. These positions might suggest further changed points which had once been the local minima of centripetal profiles. Otherwise, it is also possible to that these introns might have recently inserted themselves in the genes. Nevertheless, more intensive investigation of these introns is necessary to resolve this question. The discrimination employed here should be useful in pursuing the meaning of introns and the nature of modules.

(2) Module Boundaries without Introns

The ratio of the total number of introns to the total number of module boundaries of these proteins shows that there are about two times the number of the boundaries with no introns as with introns. Differences between these two types of boundaries might explain some interesting features of genetics or biological aspects of proteins. However, the biological meaning of the correlation between module boundaries and intron positions is not clear until today. For instance, no critical differences between the module boundaries with introns and the other boundaries without introns is found in the centripetal profiles In addition, many proteins have demonstrated that their different sources cause little deviation in the positions of introns. Therefore, Many of individual studies for various kind of protein is needed.