Biomedical Research and Reviews

Mini Review

ISSN: 2515-9186

Toward an elucidation of the process underlying the evolution of gene regulation several billion years ago

Masami Horikoshi*

Laboratory of Developmental Biology, Institute for Quantitative Biosciences, The University of Tokyo, Japan

Abstract

Because the DNA of primeval organisms no longer remains, the earliest ages of world history are generally considered off limits to molecular evolutionary analysis. However, a new method can measure evolutionary distances between existing and ancestral organisms by focusing on repeating sequences. This approach utilizes the similarity in intramolecular direct repeats present in the transcription initiation factors TBP and TFIIB as an evolutionary measure revealing the degree of similarity between the genes of present offspring and those of their ancestors. The approach also reveals the properties of the ancestors and the order of emergence of TBP and TFIIB. While both archaeal and eukaryotic transcription initiation systems utilize TBP and TFIIB, eukaryotic systems include larger numbers of initiation factors. It remains uncertain how eukaryotic transcription initiation systems have evolved. Inter-repeat sequence dissimilarity indicates that the asymmetry of two repeats in TBP and TFIIB has gradually increased during evolution. Interspecies sequence diversity indicates that the resultant asymmetric structure, which is related to the ability to interact with multiple factors, diverged in archaeal TBP and archaeal/eukaryotic TFIIB during evolution. This suggests that eukaryotic TBP initially acquired multiple Eukarya-specific interactors through asymmetric evolution of the two repeats. After the asymmetric TBP generated the complexity of the eukaryotic transcription initiation systems, its diversification halted, and its asymmetric structure spread throughout eukaryotic species. This approach provides a new way to discuss mechanistic and system evolution quantitatively.

Among the many questions that challenge the limits of human wisdom, perhaps none are more central to our existence than "How did the universe begin?" and "How was the first life created?" Human being has long endeavored to answer these two questions. Charles Darwin introduced his theory of evolution after finding the rudimentary traces of evolution by comparing the phenotypes of organisms [1]. More than a hundred years later, the neutral theory of evolution was proposed by Motoo Kimura to explain the evolutionary process at the molecular level. His idea was that mutations on genes would not always lead to a change in function or phenotype [2]. Susumu Ohno argued that gene duplication would play an important role in the process of biological diversification [3]. To describe the intricate and countless processes from molecular evolution to phenotypic evolution, we must know how biochemical and biological reactions are entwined and evolve as they form systems. This has been one of the longstanding problems in the study of evolution.

In the study of evolution, fossils are often used to estimate the dates when organisms diverged from their common ancestors. In the field of molecular evolution, scientists study and compare gene and protein sequences in organisms' molecules and speculate as to when the organisms branched out from their common ancestors. However, these methods have inherent drawbacks: 1) we cannot obtain genes of ancient organisms; 2) we cannot measure the evolutionary distances between ancient organisms and their offspring; and 3) when we try to group genes into families based on the genetic material of present offspring, a part of the family's genes has to be set aside as an outer group and excluded from analysis. As a result, the evolutionary process of the whole gene family can never be determined, giving rise to the need for a new approach.

A new analytical method has been developed to measure the evolutionary distances between existing and ancestral organisms by focusing on sequences that repeat themselves within a gene [4]. This method does not require an outer group to calculate the evolutionary distance between the present offspring genes and their ancestors [4]. With the newly *defined evolutionary indicator d_{DR}, which represents the distance between direct repeats*, the shortcomings of the pre-existing methods described above were overcome [4]. Because the DNA of primeval organisms no longer remains, molecular evolutionary analysis is often considered to be confined to the later ages of world history. However, the introduction of this reliable, repeat-based approach promises to overcome a problem in molecular evolution that has remained unsolved for over half a century, while yielding more accurate findings on the evolution of biological systems than the estimations by any of the existing methods. Using the newly defined evolutionary indicator, molecules involved in transcription, that is, in the transfer of genetic information from DNA to messenger RNA-were studied as described below, and the transcription systems of organisms from nearly 3 billion years ago were found to differ from those of existing archaea and eukaryotes [4].

The introduction of a new evolutionary indicator, d_{DR}, enabled us to estimate the evolutionary distances between the EA (earliest

Correspondence to: Masami Horikoshi, Laboratory of Developmental Biology, Institute for Quantitative Biosciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan, Tel: +81-(0)3-5841-8469, Fax: +81-(0)3-5841-8468, E-mail: horikosh@iam.u-tokyo.ac.jp

Key words: *direct repeats, evolutionary indicator (d_{DR}), mechanistic evolution, system evolution, TATA box-Binding Protein (TBP), Transcriptional Factor IIB (TFIIB)*

Received: August 10, 2018; **Accepted:** August 22, 2018; **Published:** August 27, 2018

ancestral) gene and the genes of present offspring for both TBP (TATA box-Binding Protein) and TFIIB (Transcription Factor IIB) with direct repeats without setting an outer group [4]. Moreover, the combination of d_{DR} and a phylogenetic tree allowed us to determine the clade containing the genes most similar to the EA gene at a glance. The d_{DR} value was calculated for each TBP and TFIIB gene. Archaeal TBP and TFIIB genes have lower d_{DR} values than eukaryotic genes [4]. Further, *Mj* (*Methanocaldococcus jannaschii*) TBP and TFIIB and its closely related genes are more similar to the EA gene than to the other genes.

Under three assumptions-that the d_{DR} values of TBP and TFIIB can be compared directly, that the relationship between the d_{DR} values of TBP and TFIIB is linear throughout evolution, and that the best fitting line can be extrapolated to the y-axis-a surprising result was obtained. Since the coefficient of the best fitting line is approximately 1.0 and mutations in the TBP and TFIIB genes are accumulated at a nearly similar rate [4], the *positive y-intercept* of the best fitting line implies that mutations had already been accumulated on the TFIIB gene when the TBP gene was generated. This in turn suggests that the TFIIB gene was generated before the TBP gene. The d_{DR} analysis could predict the order of emergence of the TBP and TFIIB genes for the first time. Further development of a method to quantitatively measure molecular evolution would provide detailed insights into the evolutionary development of the transcription apparatus at the system level. I believe that this approach may ultimately lead to a new field of molecular evolution, which might be called "mechanistic and system evolution".

Considering that TFIIB is a pol II (RNA Polymerase II)-interacting factor and forms a complex with pol II [5], the functional modulation of pol II might be evolutionarily initiated by direct interaction with TFIIB. This is one possible hypothesis to explain the development of the early transcription apparatus and its regulation. It is interesting that the crystal structures of the pol II-TFIIB complex [5] and the eubacterial RNAP-σ complex [6,7] suggested a functional relationship between TFIIB and σ factor. Pol II/RNAP and its interacting factor(s), such as TFIIB and σ factor, would be earlier forms of the transcription apparatus in evolution, as the results of the d_{DR} analysis suggested. On the other hand, TBP does not form a stable complex with pol II [5] but forms a complex with various GTFs (General Transcription Initiation Factors) for transcriptional activation. The ability of TBP to interact with other GTFs may have been acquired in earlier archaea and eukaryotes along with the change in the surface properties of TBP. Since the environment surrounding eubacteria and archaea is known to affect the amino-acid compositions of their molecules [8], a large difference in the environment between earlier archaea and eukaryotes could have facilitated the change in the surface properties of eukaryotic TBP, leading to the association of other GTFs to establish a more complicated regulatory system of transcription.

In addition, it was revealed that the mechanism by which the system that initiates the copying of genetic information, known as transcription, became more complex as eukaryotic cells arose from the archaea domain of organisms about 2 to 2.5 billion years ago, as described below [9]. In the eukaryotic system, three RNA polymerases, each of which contains >10 subunits, are required along with several initiation factors. Pol I, II, and III mainly synthesize rRNA, mRNA, and tRNA, respectively. The identified initiation factors in the pol II system are TFIIA, TFIIB, TFIID (TBP and TAFs, TBP-associated factors), TFIIE, TFIIF, and TFIIH. Although the archaeal system is similar to the eukaryotic pol II system, it requires only one enzyme that synthesizes all types of RNA and a smaller number of initiation factors: TBP, TFIIB, and TFIIE. In archaeal and eukaryotic systems, TBP (TFIID) first interacts with DNA, leading to the recruitment of RNA polymerase (pol II) and other initiation factors. It is a critical question how a simple ancestral system acquired the complexity found in Eukarya during evolution. In other words, the driving force behind the development of complexity in Eukarya must be identified.

Two distinct methodologies were used to investigate how mutations have accumulated in TBP and TFIIB during evolution: the d_{DR} method [4] and a typical method used for molecular phylogenetic analysis [10]. The Phylogenetic Diversity (PD) value, which is the mean branch length connecting a species to other species, is calculated using the evolutionary distance [10]. The d_{DR} calculations showed that archaeal TBP and TFIIB exhibit lower d_{DR} values than their eukaryotic counterparts, suggesting that archaeal TBP and TFIIB maintain more of their original molecular function than eukaryotic TBP and TFIIB [5]. The neighbor-joining phylogenetic tree revealed that both archaeal TBP and TFIIB show diversity. In contrast, eukaryotic TBP except Protists are homogeneous. Eukaryotic TFIIB except Animals show more diversity than eukaryotic TBP. The PD values of eukaryotic TBP were typically lower than those of archaeal TBP [5]. In the case of TFIIB, the PD values for eukaryotic kingdoms were in most cases higher than those of Archaea [5]. It is worth noting that the trend in the distribution of the PD values of TBP was completely different from that for TFIIB, as expected from the visual inspection of the phylogenetic trees of TBP and TFIIB.

The relationship between d_{DR} and PD was examined in order to elucidate the relationship between multifunctionality in two repeats and functional diversification of TBP and TFIIB among species. While the d_{DR} and PD of TFIIB exhibit a positive correlation, those of TBP show the opposite trend [5]. These observations imply that TBP and TFIIB have different evolutionary characteristics. In archaeal TBP, the similarity of the first and second repeats is preserved but, at the same time, the interspecies sequence diversity is increased. In eukaryotic TBP, interspecies sequence diversity is highly restricted but the dissimilarity between the first and second repeats is significantly greater than that observed in Archaea. This observation suggests that the diversification of eukaryotic TBP has been restricted in spite of the differentiation of the two repeats. On the other hand, TFIIB exhibits less similarity between the first and second repeats in accordance with the interspecies sequence diversification throughout Archaea and Eukarya. This information will help to elucidate the distinct functional roles of TBP and TFIIB in the evolutionary diversification of archaeal and eukaryotic transcription initiation systems.

To incorporate the structural and functional views into the d_{DR}-PD relationship, the highly invariant residues in the DR sequence and interspecies conserved residues were mapped onto the tertiary structure of TBP [5]. In addition, the spatial distributions of interacting residues between TBP and various factors were analyzed [5]. TBP forms a saddle-shaped structure with a convex surface composed of α helices and a concave surface formed by a curved antiparallel β sheet. TBP binds to the minor groove of DNA with the β sheet on the concave surface, whereas α helices on the convex surface, the stirrup region, and the sidewall contribute to interactions with other factors, including TFIIA [11,12], TFIIB [13], Mot1 [14], and NC2 [15]. The spatial distributions of residues that are invariant in archaeal TBP but not in eukaryotic TBP were analyzed. These residues are located mainly at the surface, except for the concave surface of TBP. Moreover, these residues are individually well conserved throughout eukaryotic TBP, suggesting that they are involved in the Eukarya-specific functional

acquisition of TBP. Indeed, their tertiary structures show that each stirrup region binds repeat-specific-binding factors such as TFIIA, TFIIB, Mot1, and NC2.

Since the residues of the second stirrup region of eukaryotic TBP are similar to those of Archaea, the second stirrup regions of archaeal and eukaryotic TBP would share a function. Indeed, the tertiary structures indicate that the second stirrup region interacts with TFIIB in both Archaea and Eukarya. In contrast to the three residues in the second stirrup region, those in the first stirrup region differ between archaeal and eukaryotic TBP. This difference might be one of the reasons that eukaryotic TBP acquires an interaction with TFIIA that is specific to Eukarya. The emergence of the variant residues in the first stirrup region in Eukarya seems to cause a loss of the functional redundancy in Archaea and to generate new interactions with TFIIA. Mot1 and NC2 are also Eukarya-specific factors, which inhibit the basal transcription activity of TBP. It was found that the three residues in the first stirrup region are also utilized for the interactions with Mot1 and NC2. It is worth noting that these residues are invariant in archaeal TBP but are not invariant in eukaryotic TBP. As found in the TFIIA-binding site, the variant residues of TBP seem to be utilized to make new interactions with Mot1 and NC2 in Eukarya. In contrast to the case with TBP, no tertiary structure of TFIIB in complex with Eukarya-specific factors has been analyzed. Taken together, the above findings show that residues in the DR sequence that are highly invariant in Archaea but not in Eukarya are key to the functional interaction of eukaryotic TBP with Eukarya-specific factors. This suggests that TBP plays a role in establishing complexity in the eukaryotic transcription initiation system.

Using d_{DR} and PD, we have tried to understand how the complexity of the transcription initiation system was generated during evolution. Since d_{DR} reflects multifunctionality in DR, the smaller number of highly and less-invariant residues in the DR sequence of eukaryotic TBP compared to the DR sequence of archaeal TBP implies that eukaryotic TBP would accommodate the ability to interact with Eukarya-specific factors. Eukarya-specific TBP-Interacting Factors (TIF) would lead to further association with TIF-interacting factors, resulting in the complexity of the eukaryotic transcription initiation system. As in the case of TBP, the smaller number of highly and less-invariant residues in the DR sequence of eukaryotic TFIIB compared to the DR sequence of archaeal TFIIB implies that eukaryotic TFIIB would also accommodate the ability to interact with the Eukarya-specific factors.

On the other hand, the d_{DR}-PD analyses showed a clear difference between the evolutionary characteristics of TBP and those of TFIIB. In contrast to TBP, TFIIB exhibits a monotonic increase of inter-repeat sequence dissimilarity (d_{DR}) in accordance with interspecies sequence diversity (PD). These opposite characteristics of TBP and TFIIB might be explained in part by the degree of structural and functional constraints resulting from differences in the surface area and binding strength of the DNA-binding region between the two initiation factors. There is enough space on the opposite side of the DNA-binding surfaces of both TBP and TFIIB but binding with other factors has not been elucidated at the tertiary structure level. Considering that eukaryotic TBP and TFIIB have been reported to functionally and biochemically interact with a variety of regulatory transcription factors, each interaction would involve specific residues. The development of the complexity of the eukaryotic transcription initiation system during evolution would be further elucidated by additional structural and functional analyses of complexes containing TBP or TFIIB.

The evolutionary trajectory of TBP was speculated upon on the basis of the characteristics of the d_{DR} and PD values of TBP. Namely, through the asymmetric evolution of the intramolecular direct repeat, eukaryotic TBP acquired the ability to interact with multiple Eukarya-specific factors, and then the resultant asymmetric structure was strongly conserved among Eukarya. The multiple factors interacting with eukaryotic TBP were derived from its asymmetric structure and gave rise to the evolutionary complexity of eukaryotic transcription initiation systems. This strategy was named *DECS (DR-mediated establishment of a complex system)*. The development of a method to quantitatively measure molecular evolution offers not only broader and more detailed insights into the evolutionary process of life systems that support cell proliferation and differentiation, but also sheds light on the transformation of the evolutionary mechanism underlying sophisticated biological systems, thereby pointing to the potential application of this knowledge in various fields. Moreover, information gained from such studies could also be applied to the "evolutionary process" of artificial intelligence and precision machinery.

References

1. Darwin C (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life, (1st Edn.), Murray. [Crossref]
2. Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217: 624-626. [Crossref]
3. Ohno S (1970) Evolution by gene duplication, Springer-Verlag. https://link.springer.com/book/10.1007/978-3-642-86659-3
4. Adachi N, Senda T, Horikoshi M (2016) Uncovering ancient transcription systems with a novel evolutionary indicator. *Sci Rep* 6. [Crossref]
5. Kostrewa D, Zeller ME, Armache KJ, Seizl M, Leike K, et al. (2009) RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature* 462: 323-330. [Crossref]
6. Murakami KS, Masuda S, Darst SA (2002) Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 A resolution. *Science* 296: 1280-1284. https://www.ncbi.nlm.nih.gov/pubmed/12016306. [Crossref]
7. Vassylyev DG, Sekine S, Laptenko O, Lee J, Vassylyeva MN, et al. (2002) Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 A resolution. *Nature* 417: 712-719. [Crossref]
8. Reed CJ, Lewis H, Trejo E, Winston V, Evilia C (2013) Protein adaptations in archaeal extremophiles. *Archaea* 2013: 373275. [Crossref]
9. Kawakami E, Adachi N, Senda T, Horikoshi M (2017) Leading role of TBP in the Establishment of Complexity in Eukaryotic Transcription Initiation Systems. *Cell Rep* 21: 3941-3956. [Crossref]
10. Faith DP (1992) Conservation evaluation and phylogenetic diversity. *Biol Conserv* 61: 1-10. https://www.sciencedirect.com/science/article/pii/0006320792912013
11. Tan S, Hunziker Y, Sargent DF, Richmond TJ (1996) Crystal structure of a yeast TFIIA/TBP/DNA complex. *Nature* 381: 127-151. [Crossref]
12. Geiger JH, Hahn S, Lee S, Sigler PB (1996) Crystal structure of the yeast TFIIA/TBP/DNA complex. *Science* 272: 830-836. [Crossref]
13. Nikolov DB, Chen H, Halay ED, Usheva AA, Hisatake K, et al. (1995) Crystal structure of a TFIIB-TBP-TATA-element ternary complex. *Nature* 377: 119-128. [Crossref]
14. Wollmann P, Cui S, Viswanathan R, Berninghausen O, Wells MN, et al. (2011) Structure and mechanism of the Swi2/Snf2 remodeller Mot1 in complex with its substrate TBP. *Nature* 475: 403-407. [Crossref]
15. Kamada K, Shu F, Chen H, Malik S, Stelzer G, Roeder RG, Meisterernst M, Burley, SK (2001) Crystal structure of negative cofactor 2 recognizing the TBP-DNA transcription complex. *Cell* 106: 71-81. https://www.ncbi.nlm.nih.gov/pubmed/11461703. [Crossref]