TRANSCRIPTION
2016, VOL. 7, NO. 1, 1–13
http://dx.doi.org/10.1080/21541264.2015.1128518

Taylor & Francis
Taylor & Francis Group

REVIEW

OPEN ACCESS

A model for genesis of transcription systems

Zachary F. Burton[a], Kristopher Opron[b], Guowei Wei[b], and James H. Geiger[c]

[a]Department of Biochemistry and Molecular Biology, Michigan State University, E. Lansing, MI, USA; [b]Department of Mathematics, Michigan State University, E. Lansing, MI, USA; [c]Department of Chemistry, Michigan State University, E. Lansing, MI, USA

ABSTRACT

Repeating sequences generated from RNA gene fusions/ligations dominate ancient life, indicating central importance of building structural complexity in evolving biological systems. A simple and coherent story of life on earth is told from tracking repeating motifs that generate α/β proteins, 2-double-$\Psi-\beta$-barrel (DPBB) type RNA polymerases (RNAPs), general transcription factors (GTFs), and promoters. A general rule that emerges is that biological complexity that arises through generation of repeats is often bounded by solubility and closure (i.e., to form a pseudo-dimer or a barrel). Because the first DNA genomes were replicated by DNA template-dependent RNA synthesis followed by RNA template-dependent DNA synthesis via reverse transcriptase, the first DNA replication origins were initially 2-DPBB type RNAP promoters. A simplifying model for evolution of promoters/replication origins via repetition of core promoter elements is proposed. The model can explain why Pribnow boxes in bacterial transcription (i.e., $^{-12}$TATAATG^{-6}) so closely resemble TATA boxes (i.e., $^{-31}$TATAAAAG^{-24}) in archaeal/eukaryotic transcription. The evolution of anchor DNA sequences in bacterial (i.e., $^{-35}$TTGACA^{-30}) and archaeal (BRE$_{\text{up}}$; BRE for TFB recognition element) promoters is potentially explained. The evolution of BRE$_{\text{down}}$ elements of archaeal promoters is potentially explained.

ARTICLE HISTORY
Received 1 September 2015
Revised 29 November 2015
Accepted 30 November 2015

KEYWORDS
cradle-loop barrel metafold; double-$\Psi-\beta$-barrels; general transcription factors; LECA (the last eukaryotic common ancestor); LUCA (the last universal common cellular ancestor); replication; RIFT barrels; RNA polymerase; Rossmann folds; σ factors; TATA-binding protein (TBP); the carboxy terminal domain (CTD) of RNA polymerase II; TIM barrels; transcription; transcription factor B (TFB); α/β protein folds

After the advent of coding, ancient evolution of life on earth becomes a starkly simple story of replication errors or, perhaps more likely, RNA gene fusions/ligations resulting in repeating RNA sequences encoding repeating protein motifs.[1] A trend toward increased biological complexity was largely driven by generation of repeating sequences, for which there appears to have been strong positive selection. The number of repeats often appears to be limited by solubility and structural closure, and also some dimeric repeats or true dimers were selected for nucleic acid binding (i.e., TBP and helix-turn-helix dimers). Because of relatively weak initial competition for enzyme specificity and functionality from ribozymes, many of the earliest successful protein folds were strongly selected for structure, solubility, and complexity. The ancient and ubiquitous α/β protein fold that supports most of fundamental metabolism and energy transduction appears to have been initiated by repetition of a $\beta-\alpha-\beta-\alpha$ motif (Fig. 1). During emergence from the RNA-protein world (~4.1 billion years ago) through LUCA (the last universal common cellular ancestor of bacteria, archaea and eukaryotes; ~3.5 to 3.8 billion years ago), 2-double-$\Psi-\beta$-barrel (DPBB) type RNA polymerases (RNAPs) remained a major replicating polymerase.[2-6] LUCA evolved to become one of the first cellular organisms with a unified DNA genome. Replication of the first DNA appears to have been initiated using 2-DPBB type RNAPs followed by DNA synthesis using reverse transcriptase.[7] Because 2-DPBB type RNAPs dominated LUCA replication and transcription, divergence of bacteria and archaea was

CONTACT Zachary F. Burton burton@cns.msu.edu Department of Biochemistry and Molecular Biology, Michigan State University, 603 Wilson Rd., E. Lansing, MI, 48824-1319, USA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/ktrn.

Supplemental data for this article can be accessed on thepublisher's website.

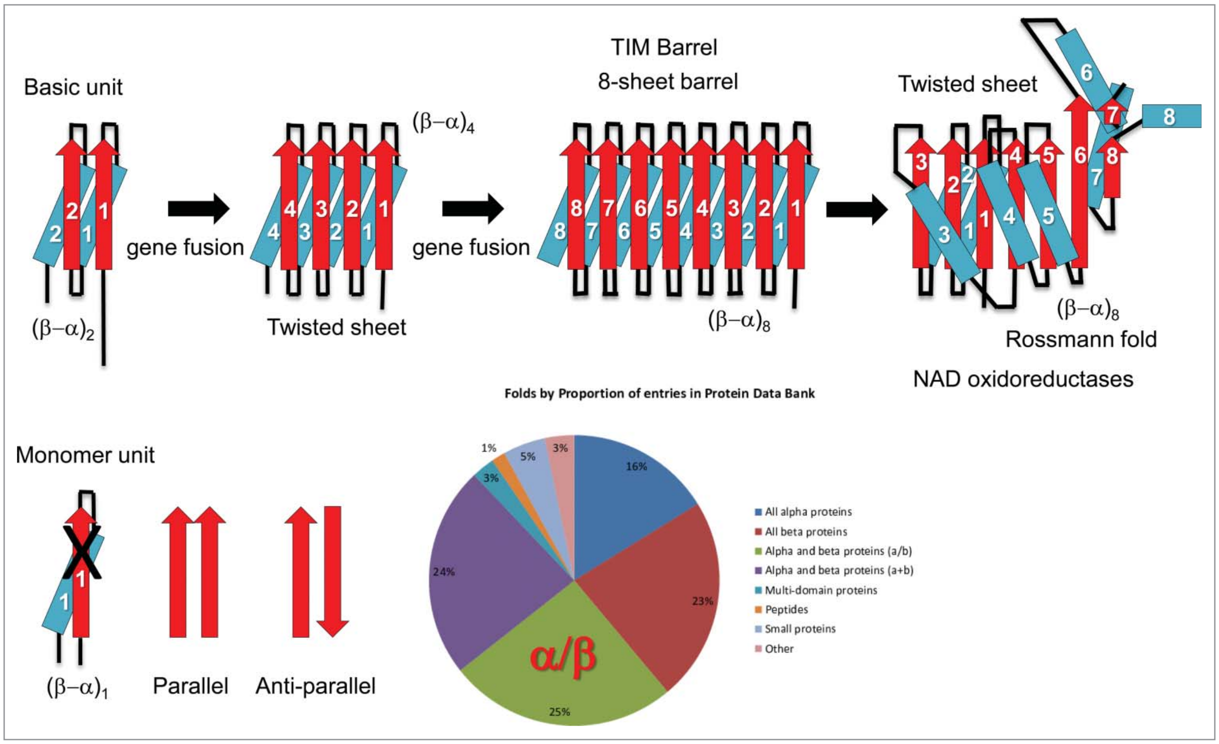


Figure 1. α/β folds are simple $(\beta{-}\alpha)_n$ repeat proteins. The pie chart indicates that ~25% of all structures in the RCSB protein data bank are α/β fold proteins. A model is shown for evolution of TIM barrels $(\beta{-}\alpha)_8$ and Rossmann folds $(\beta{-}\alpha)_8$.

driven by coevolution of 2-DPBB type RNAPs, RNAP general transcription factors (GTFs) and RNAP promoters. By contrast, DNA polymerases (DNAPs) and distinct promoters and replication origins were not yet dominant.[7] In support of this ancient replication mechanism, non-homologous DNAPs that exist today in bacteria and archaea appear to have arisen separately after divergence of bacteria and archaea.

Eukaryotes are generally more complex than bacteria and archaea, and the tortured path to eukaryotic evolution explains increased genomic, functional and organismal complexity. At LECA (the last eukaryotic common ancestor; ~1.6 to 2.2 billion years ago), eukaryotes resulted from endosymbiosis and genetic fusion of a Lokiarchaeota phylum archaea[8,9] and an α-proteobacterium.[10,11] Although many modern archaea have lost the capacity to engulf a bacterial endosymbiont, Lokiarchaeota has the ESCRT I, II and III (endosomal sorting complexes required for transport) endocytosis/phagocytosis systems and also actin and tubulin, which are also required for endocytosis (Fig. S1).

In eukaryotes, genetic duplications generated RNAPs I, II and III and the carboxy terminal domain (CTD) repeat on RNAP II, which facilitated nuanced RNAP II regulation required to support complexity and multicellularity.[4] The story of the inception of biological complexity, therefore, includes recurrent cases in which repeating sequences were generated.[1] Most surprisingly, however, after up to ~3.5 to 4 billion years, initial repeats can remain recognizable, sometimes in sequence but more often in secondary structure. Examples include α/β proteins, the RIFT barrel, DPBBs, bacterial σ transcriptional initiation factors, TFB (transcription factor B), TBP (TATA-binding protein) and the RNAP II CTD. Core promoter elements are posited also to be generated via repetition of motifs, potentially explaining the similarity between archaeal/eukaryotic TATA boxes (i.e., $^{-31}$TATAAAAG^{-24}) and bacterial Pribnow boxes (i.e., $^{-12}$TATAATG^{-6}) (see below).

α/β *folds: generating complexity via RNA gene fusions/ligations*

Remarkably, glycolysis, the citric acid cycle and the glyoxylate cycle are catalyzed by ancient α/β fold proteins (Fig. 1, Figs. S2-S8). In addition to core metabolism and redox potential, ATPases, GTPases and kinases are supported by α/β folds. So, much or all of the most ancient metabolism and also redox and chemical energy transduction are supported by α/β fold proteins that date from the RNA-protein world. The α/β proteins can be described as $(\beta{-}\alpha)_n$ repeat proteins, generated from $\beta{-}\alpha{-}\beta{-}\alpha$ repeats.[12,13] Because to support an extended chain structure a β-sheet requires hydrogen bonding to a second β-sheet, the basic unit for repeat generation appears to be $\beta{-}\alpha{-}\beta{-}\alpha$ rather than a monomeric $\beta{-}\alpha$ unit. A β-sheet interacts with a neighboring β-sheet in either a parallel or antiparallel orientation, and, without a partner, a β-sheet cannot hydrogen bond to maintain its characteristic extended β-sheet conformation. Most ancient proteins with many parallel β-sheets are α/β fold proteins. As shown in Figure 1, the α/β fold comprises about 25% of proteins represented in the RCSB protein data bank. A model is shown for fusion of $(\beta{-}\alpha)_n$ repeats to generate TIM barrels $(\beta{-}\alpha)_8$ and Rossmann folds $(\beta{-}\alpha)_8$. We conclude that ubiquitous α/β fold proteins are generated from simple $(\beta{-}\alpha)_n$ repeats.

Glycolytic enzymes are of the TIM (triose phosphate isomerase) barrel fold $(\beta{-}\alpha)_8$ (Fig. S2).[14-21] This ancient and ubiquitous fold was generated by duplication (probably via RNA gene fusion or ligation) of a $(\beta{-}\alpha)_4$ unit, which was itself generated by earlier duplication of a $(\beta{-}\alpha)_2$ unit. Formation of the 8-parallel β-sheet TIM barrel gives closure to the structure. Further polymerization beyond 8 sheets breaks the barrel and can create a larger, more flexible horseshoe structure. Compared to TIM barrels, Rossmann folds are generated from $(\beta{-}\alpha)_8$ repeats that are rearranged into a twisted sheet (Figs. 1 and S3).[15,19,22-24] Starting from a TIM barrel, in order to generate a Rossmann fold required rearrangement of $\beta4$ to pair with $\beta1$ rather than with $\beta3$. Such rearrangement is possible because $\alpha3$ and its surrounding loops can span the distance required for repositioning $\beta4$. A second rearrangement pairs both $\beta7$ and $\beta8$ with $\beta6$, as indicated in Figures 1 and S3. We posit that Rossmann folds $(\beta{-}\alpha)_8$ may have resulted from rearrangement of a TIM barrel $(\beta{-}\alpha)_8$. In a TIM barrel, $\beta1{-}\beta8$ curvature supports closure of the barrel. Because of necessary rearrangements, the Rossmann fold linear, twisted sheet is supported by the opposite curvature of $\beta1{-}\beta3$ and $\beta4{-}\beta7/\beta8$. Similarly, TOPRIM domains $\sim(\beta{-}\alpha)_{4\text{-}5}$ (Fig. S4) appear to be generated from a larger repeat such as a Rossmann-like fold. Many ATPases, GTPases and kinases are Rossmann-like folds that emerged in the RNA-protein world (Fig. S5-S6). Swi-Snf ATPases include 2 $(\beta{-}\alpha)_6$ domains (Figs. S7-S8). Some of these folds require rearrangements of β-sheets. Ubiquitous α/β fold proteins that account for essentially all ancient metabolism and energy transduction, therefore, were generated on a scaffold formed by simple repetition of a $\beta{-}\alpha{-}\beta{-}\alpha$ unit.

Importantly, α/β protein folds provide both structure and solubility. The $\beta{-}\alpha{-}\beta{-}\alpha$ unit and its larger repeats create structure through interaction of parallel β-sheets, and these folds are soluble, because β-sheets, which by themselves might form amyloid-like interactions, are each paired with an α-helix. The solubility of the β-sheet fold, therefore, appears to have been promoted by the associated helices. Early and enduring success of the α/β fold, therefore, is explained by structure, sufficient functionality and solubility. Remarkably, so far as we are aware, the active site in α/β fold proteins is always located to the C-terminal side of the β-sheets that dominate the fold. It appears that this relation may have been established ~4 billion years ago on a ~4.6 billion year old earth and maintained via powerful co-evolutionary forces. So far as we are aware, in a TIM barrel, a Rossmann fold or a Rossmann-like protein, no active site or allosteric site locates to the N-terminal end of the β-sheets.

A simple model for evolution of RNAPs, general transcription factors and promoters: Complexity generated via repeated sequences

Two-DPBB type RNAPs

Multi-subunit RNAPs are of the 2-DPBB type.[2,4-6] DPBBs are 6-β-sheet barrels of the ancient cradle-loop barrel metafold.[25] In Figure 2, schematic diagrams are shown of the DPBB and its parent, the RIFT barrel (RIFT for its occurrence in riboflavin synthases, F1 ATPase and translation factors).[25-27] RIFT barrels were initially generated from dimerization of a $\beta1{-}\beta2{-}\alpha1{-}\beta3$ motif, which subsequently became ligated to form a $\beta1{-}\beta2{-}\alpha1{-}\beta3{-}\beta1'{-}\beta2'{-}\alpha1'{-}\beta3'$ barrel. Unlike the RIFT barrel, the DPBB has a complex looped and pseudo-knotted fold. Monomeric RIFT barrels formed from dimeric RIFT barrels through duplication probably via ligation of 2 identical RNAs. Dimeric RIFT barrels gave rise to 8-β-sheet swapped hairpin barrels (not shown). The coloring of the schematic in Figure 2 was chosen to emphasize the

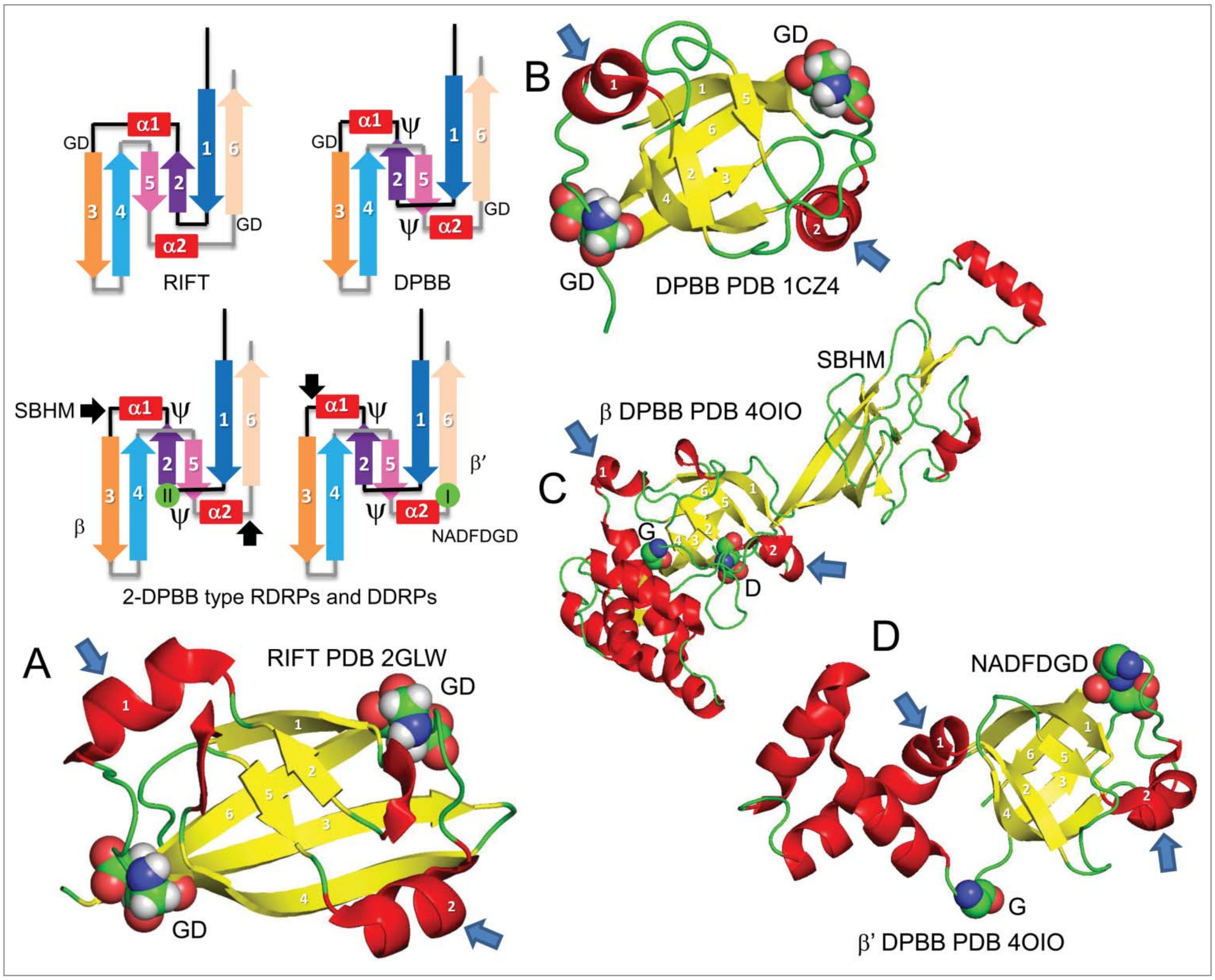


Figure 2. Cradle-loop barrels: RIFT barrels and DPBBs. A) PHS018 RIFT barrel (PDB 2GLW).[26] B) VatN-N DPBB (PDB 1CZ4) (a AAA+ ATPase).[54] C) RNAP β' DPBB (PDB 4OIO).[55] D) RNAP β DPBB (PDB 4OIO). Small blue arrows indicate $\alpha 1$ and $\alpha 2$. Small black arrows in the schematics indicate insertions in RNAP DPBBs. Conserved GD motifs and possible GD relics in RNAP DPBBs are indicated in sphere representation. A signature motif of RNAPs, NADFDGD that binds Mg-I (Mg-A), ends in a conserved GD box. Molecular graphics images were made using Pymol (https://www.pymol.org/).

duplication or gene fusion. In molecular graphic images (Figs. 2A-D), features of the specialized RNAP DPBBs are emphasized. In RIFT barrels and DPBBs, the conserved GD (glycine-aspartic acid) box is found after $\alpha 1$ and $\alpha 2$ and just before $\beta 3$ and $\beta 6$. The signature motif of multi-subunit RNAPs, NADFDGD that binds the catalytic Mg-I (Mg-A) through 3 aspartic acids, appears to end in a GD box.[25] In RIFT barrels and DPBBs, $\beta 2$ and $\beta 5$ lie in a "cradle" formed by $\beta 1$, $\beta 6$, $\beta 3$ and $\beta 4$. The $\beta 1$-$\beta 2$ and $\beta 4$-$\beta 5$ loops form the "cradle-loops" of the cradle-loop barrel fold.

In RNAPs, 2-DPBBs border the 2-Mg active site, and loops from the barrels bind active site Mg-I and Mg-II (Mg-A and Mg-B) (Fig. 3). Opposite from the DPBBs, the bridge helix and the mobile trigger loop also enclose active site Mg-I and Mg-II. Because multi-subunit RNAPs distribute to all cellular life, 2-DPBB type RNAPs must have been present at LUCA (~3.5 to 3.8 billion years ago).[2,4] Because both DNA template-dependent and RNA template-dependent RNAPs of the 2-DPBB type exist, 2-DPBB type RNAPs appear to be rooted in the RNA-protein world (up to ~4.1 billion years ago) prior to LUCA (>3.5 billion years ago). Interestingly, in transcription and replication of Hepatitis δ virus, which has a RNA genome, human RNAP II, which is normally a DNA template-dependent RNAP, can function as a RNA template-dependent RNAP.[28,29] Unique to DNA template-dependent RNAPs and inserted between $\beta 2$ and $\beta 3$ of the β-subunit type DPBB is a sandwich barrel hybrid motif (SBHM) that permits utilization of a DNA template for initiation and elongation.[2,4] To illustrate this point, in bacterial RNAP, the SBHM is also termed the "flap" domain. The "flap tip" helix

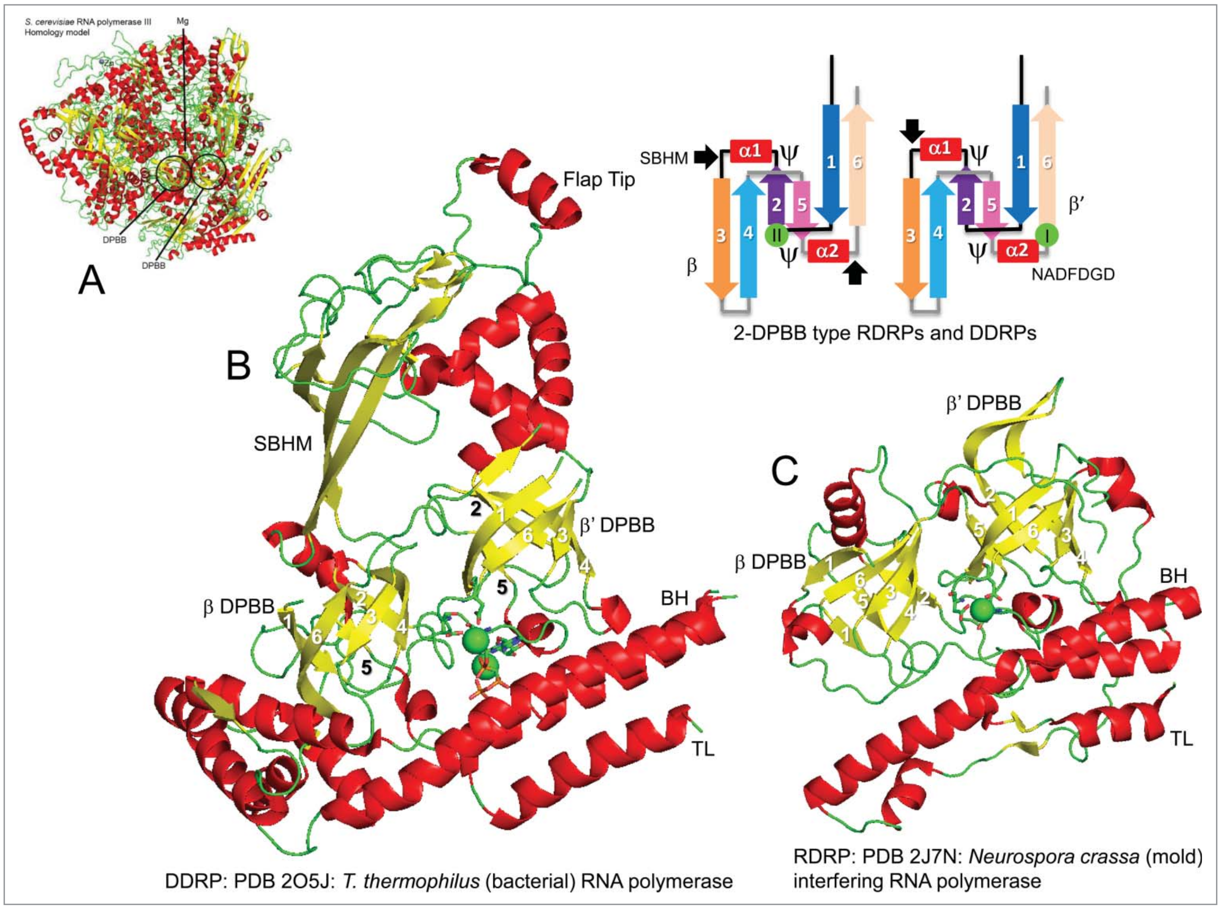

S. cerevisiae RNA polymerase III
Homology model
Mg
Zn
DPBB
DPBB
A
Flap Tip
SBHM
α1
ψ
1
6
2
5
3
4
II
ψ
α2
β
I
β'
NADFDGD
2-DPBB type RDRPs and DDRPs
B
SBHM
β' DPBB
β DPBB
BH
TL
C
β' DPBB
β DPBB
BH
TL
DDRP: PDB 2O5J: *T. thermophilus* (bacterial) RNA polymerase
RDRP: PDB 2J7N: *Neurospora crassa* (mold) interfering RNA polymerase

Figure 3. 2-DPBB type RNAPs. A) *S. cerevisiae* (yeast) RNAP III (a homology model). The 2-DPBBs border the active site. B) *Thermus thermophilus* RNAP catalytic core including 2-DPBBs, the SBHM, the bridge helix (BH) and trigger loop (TL) (PDB 2O5J) (a DNA template-dependent RNAP).[31] C) *N. crassa* (mold) interfering RNAP catalytic core including 2-DPBBs, BH and TL (PDB 2J7N) (a RNA template-dependent RNAP).[34]

interacts with the bacterial σ factor (HTH$_4$) to support transcription initiation[30] and, during elongation, interacts with the exiting RNA.[31,32] To overcome transcription stalls, the flap tip helix also interacts with the Swi-Snf type ATPase HepA/RapA to promote RNAP backtracking.[33] 2-DPBB type RNA template-dependent RNAPs include a bridge helix and a trigger loop, but lack an inserted SBHM, which is required for utilization of a DNA template but not an RNA template.[34]

General transcription factors (GTFs)

TBP and σ/TFB

Here we describe evolution of TBP (TATA-binding protein) and a primordial transcription initiation factor that gave rise to σ factors in bacteria and to TFB (Transcription Factor B) in archaea, driving divergence of bacteria and archaea.[35] TBP includes 2 TBP-fold repeats. TFB, with 2-HTH (helix-turn-helix) repeats, appears to be derived from a 4-HTH primordial initiation factor. Both TBP and the 4-HTH primordial initiation factor are posited to have existed at LUCA.[35] As described above, at LUCA, GTFs can also be considered replication origin binding factors, because replication on the first DNA templates initiated via transcription followed by reverse transcription.[7] TBP was generated via duplication of a TBP fold, and, consistent with its near 2-fold symmetry, TBP lands within the DNA minor groove at the TATAAAAG box.[36] As we have described elsewhere, the primordial initiation factor that gave rise to σ/TFB is posited to be a regular repeat of 4-HTH domains (Fig. 4). In sequence, the primordial initiation factor is most similar to the 2-HTH domains (historically termed "cyclin-like" repeats) of archaeal TFB. Because of cooperation and compensation from TBP and TFE, and because of addition to TFB of a N-terminal Zn-ribbon extension, 2 of 4-HTH units are thought to

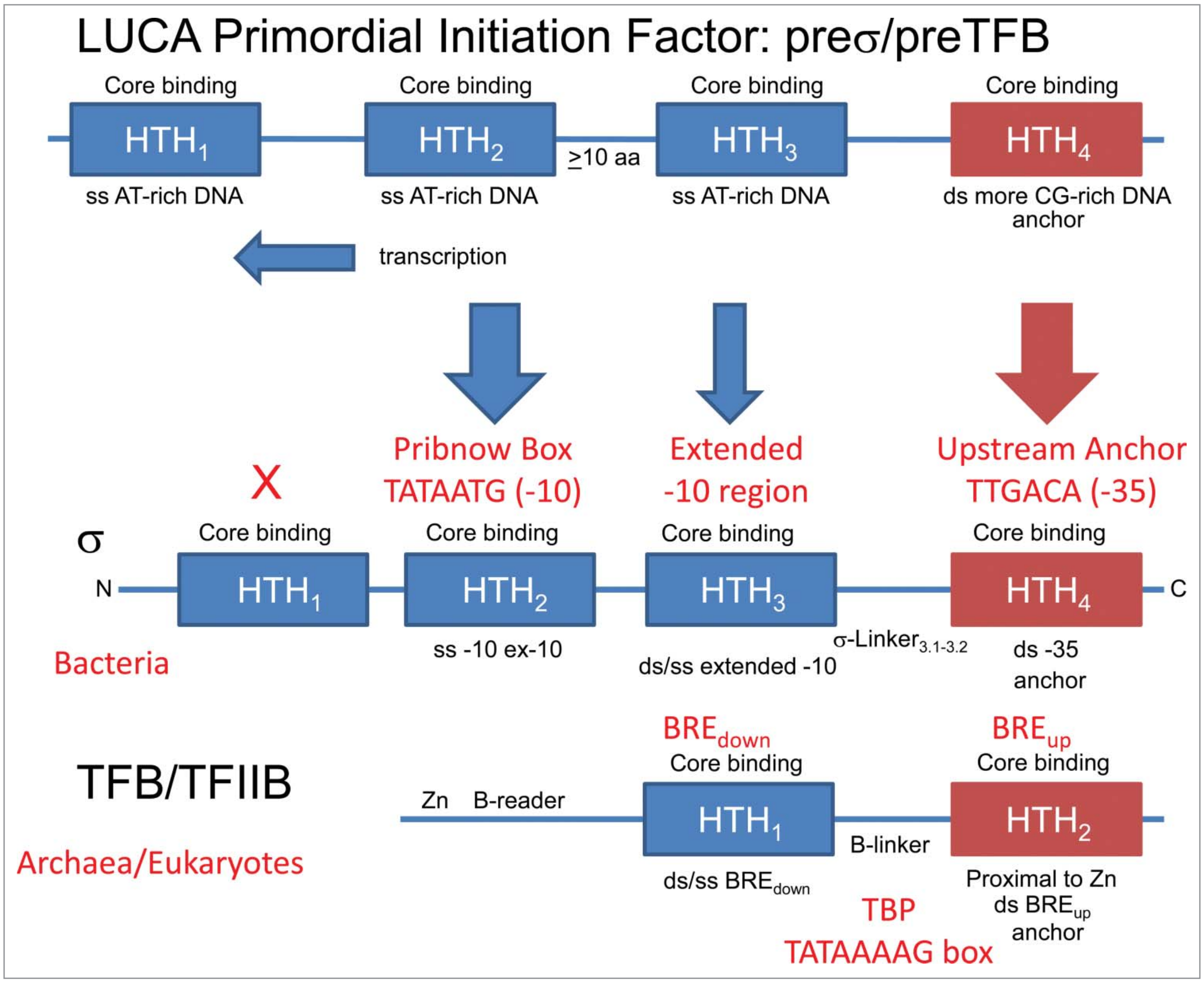


Figure 4. A model for evolution of bacterial σ factors and archaeal TFB from a 4-HTH primordial initiation factor at LUCA. Classic σ homology regions overlap with HTH$_{1\text{-}4}$.[35]

have been lost from TFB in evolution. Bacterial σ factors are derived from a repeat of 4-HTH domains, but, because of coevolution with RNAP and promoters, the 4-HTH repeat structure in σ, although recognizable, is degenerate in sequence.[35]

The evolutionary model for σ combined with recent x-ray structures of initiating RNAP describes σ functions in initiation. In Figure 5, a RNAP-promoter complex (PDB 4XLN) is shown with RNAP core subunits removed from the image to visualize the intact transcription bubble and DNA interacting with σ.[37] Because HTH$_1$ appears vestigial, HTH$_1$ is not shown. HTH$_4$ binds the "anchor" DNA -35 region of the bacterial promoter ($^{-35}$TTGACA^{-30}). This contact anchors RNAP at the promoter upstream and specifies the direction of bubble opening and downstream transcription. HTH$_4$ is a typical HTH unit with a 8-10 amino acid H2 that makes canonical contact, via the N-terminal end of its "recognition helix" H3, to the major groove of the DNA. HTH$_3$ is also a typical HTH that binds upstream of the initiation site where the bubble opens. HTH$_3$ has a typical 8-10 amino acid H2. In the structure with an open bubble, the N-terminal end of HTH$_3$ H3 does not bind in the DNA major groove, as would be expected during an initial encounter with double-stranded (ds) DNA, but no structure of HTH$_3$ on dsDNA is currently available. Perhaps, as the bubble opens, HTH$_3$ twists from an initial typical HTH-major groove contact. HTH$_2$ is a highly specialized HTH-derived fold that opens the -10 region of the bacterial promoter ($^{-12}$TATAATG^{-6}). HTH$_2$ has an elongated H2 (20 amino acids) and bulky hydrophobic residues on H3 (i.e., KFSTYATWWIR) judged inconsistent with binding to dsDNA.[38-40] HTH$_2$ aggressively attacks dsDNA, flips out $^{-11}$A of the nontemplate DNA strand (NDS) and then flips out $^{-7}$T

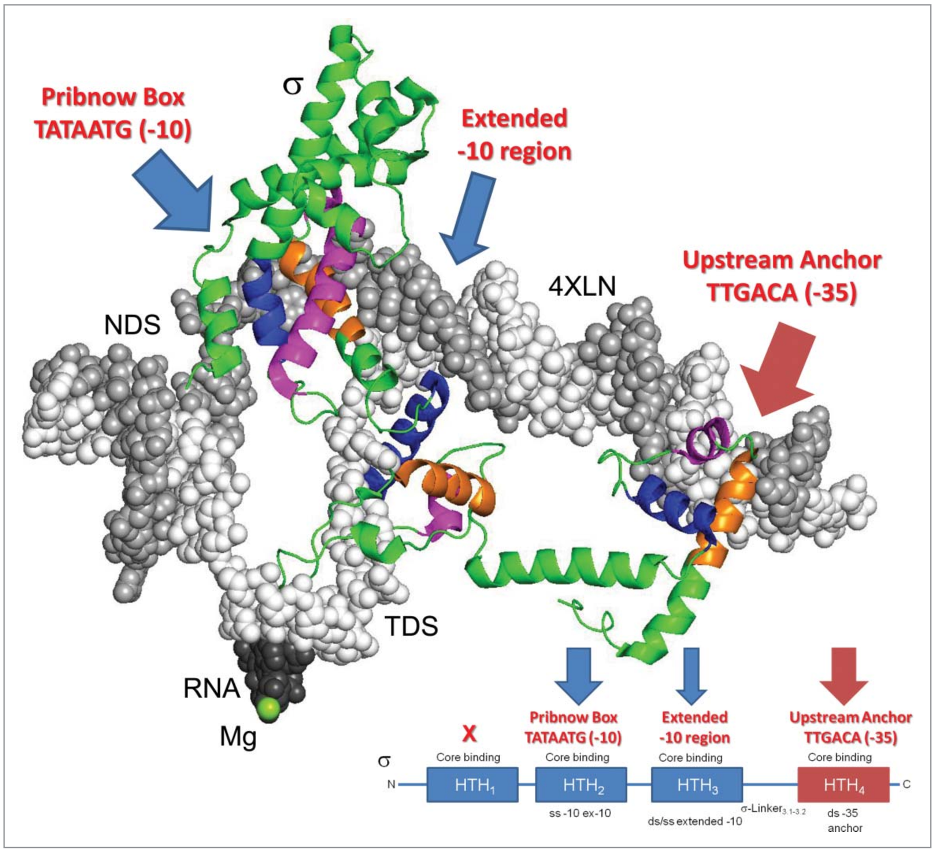


Figure 5. Bacterial σ factor interactions with promoter DNA in initiating complexes with an open transcription bubble (PDB 4XLN).[37] RNAP was removed from the image in order to visualize σ (green except at HTH motifs) interactions to promoter DNA. HTH units are colored blue (H1), magenta (H2) and orange (H3; N-terminal end only). Only σ HTH_4, HTH_3 and HTH_2 were colored. In this view, HTH_1 is obscured by HTH_2. To locate the RNAP active site, RNA and Mg-I are shown.

on the NDS to help unzip DNA to +1 for initiation on the template DNA strand (TDS).[41] Degeneracy of the 4-HTH units of σ relative to the proposed regular 4-HTH repeat LUCA primordial initiation factor, from which σ was derived, is explained by powerful coevolution of RNAP, σ HTH repeats and promoter DNA. To describe the evolutionary pressures on σ, each HTH repeat is selected for: 1) specific promoter recognition; 2) RNAP binding; 3) autoinhibition of promoter binding off of RNAP; 4) solubility off of RNAP; and 5) release from core RNAP during elongation. In σ factors, the 2 most important of the 4-HTH repeats are generally HTH_4 that binds the anchor DNA (-35) and the highly specialized HTH_2 that opens the -10 region.[39] By contrast to σ, the 2 remaining archaeal TFB HTH repeats function cooperatively with TBP and TFE and more independently of RNAP, and, consistent with somewhat relaxed evolutionary pressures, TFB has maintained a very recognizable repeat structure (the 2 "cyclin-like" repeats) that through evolution have become partly obscured in σ. Although bacteria no longer utilize TBP, bacterial RNase HIII includes a TBP fold, indicating that a bacterial ancestor (i.e. LUCA) may have included TBP, as proposed here (Figure S9).[36,42]

TFE

In addition to TBP and TFB, archaea also utilize the GTF TFEα/β.[43] TFEα includes a winged HTH (WHTH) motif and a Zn ribbon. The WHTH motif in TFEα is most similar to the ArsR family of WHTH transcription factors dispersed to bacteria and archaea. The WHTH motif in TFEβ is most similar to the MarR family of WHTH transcription factors dispersed to bacteria and archaea. It is clear that the WHTH motif is ancient and, because it is distributed to both bacteria and archaea, the WHTH motif may have been present at LUCA. It is unclear when TFEα/β arose in archaea (i.e., before or after divergence of bacteria and archaea). Because of the difficulty in opening B-form DNA, it is also possible that a helicase may have aided promoter and replication origin opening at LUCA, analogous (not homologous) to the eukaryotic TFIIH helicases.

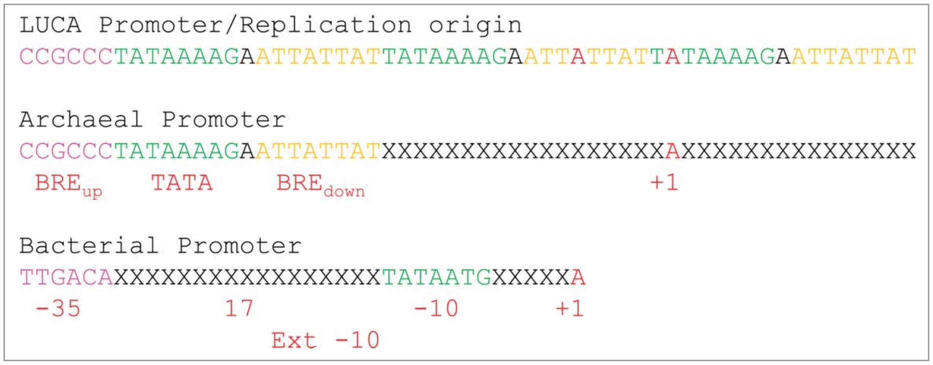
According to these simple and simplifying models, LUCA transcription and replication, on the first DNA templates, were supported by a mechanism that is very recognizable today. We posit that TBP bound multiple TATAAAAG boxes. A primordial initiation factor with 4-HTH repeats ("cyclin-like" repeats) bound to surrounding BREs (TFB-recognition elements). TFE may have been present, or TFE may have evolved separately in archaea after divergence. Also, a helicase may have facilitated promoter/replication origin opening. This model is simplifying, because it roots the tree of life at LUCA and describes the radiation of bacteria and archaea. Bacteria and archaea are posited to have diverged because they evolved to interpret, transcribe and replicate their genomes using distinct RNAP-GTF-promoter combinations. In bacteria, GTFs, RNAP and promoters became much more tightly coupled and co-dependent than in archaea, and this difference is seen comparing the degenerate HTH units of σ factors in bacteria and the more recognizable cyclin-like HTH repeats of TFB in archaea.

Model for a LUCA promoter: Proposed TATAAAAG and BRE repeats, followed by simplification via coevolution in bacteria and archaea

Repeating sequences generated α/β folds, RIFT barrels, DPBBs, TBP (2 TBP-fold repeats), σ (4-HTH repeats) and TFB (2-HTH repeats). Similarly, promoters at LUCA are posited to have evolved via repetition of an AT-rich sequence. As a possible example, a hypothetical LUCA promoter is posited to be generated by alternating repeats of a TATAAAAG box and an AT-rich BRE_{down} (TFB-recognition element downstream of TATA) (Fig. 6). Three repeats are shown with an upstream anchor sequence (a GC-rich BRE_{up}). In the LUCA promoter/replication origin, there may have been many more than 3 repeats, but 3 repeats is sufficient to generate a model. From the proposed LUCA promoter, an archaeal promoter with a BRE_{up}, a TATAAAAG box and a BRE_{down} can be generated via simplification, and simplification is also posited for archaeal TFB, which is posited to have lost 2-HTH repeats from a 4-HTH primordial initiation factor. Similarly, a bacterial promoter can be derived, with a $^{-35}\text{TTGACA}^{-30}$ -35 anchor region, an extended -10 region, and a Pribnow box ($^{-12}\text{TATAATG}^{-6}$) -10 region. In this model, the Pribnow box of the bacterial promoter is derived from a downstream TATAAAAG box of the LUCA promoter, explaining why Pribnow boxes, which are bound by a specialized HTH-derived domain (σ HTH_2), resemble TATAAAAG boxes, which are bound by TBP (not present in bacteria) binding in the DNA minor groove. Other features of promoters can be generated by coevolution of interacting factors and promoters. For instance, alternate σ factors with different promoter recognition can be generated via co-evolutionary forces. As we have previously shown, some alternate σ factors are most similar to archaeal TFB in sequence, particularly in σ HTH_4 T1-H2-T2 (i.e., RRT-QREIAKAL-GIS) and TFB HTH_2 T1-H2-T2 (i.e., RRT-QREVAEVA-GVT).[35]

Model for LUCA GTFs on promoter DNA

From X-ray structures, a model for GTFs on a LUCA promoter/replication origin can be constructed (Fig. 7). PDB 1AIS shows archaeal TBP and TFB HTH_1-HTH_2 on promoter DNA.[44] The LUCA GTF-



```
LUCA Promoter/Replication origin
CCGCCCTATAAAAGAATTATTATTATAAAAGAATTATTATTATAAAAGAATTATTAT

Archaeal Promoter
CCGCCCTATAAAAGAATTATTATXXXXXXXXXXXXXXXXXXAXXXXXXXXXXXXXXX
 BREup  TATA    BREdown                  +1

Bacterial Promoter
TTGACAXXXXXXXXXXXXXXXXXTATAATGXXXXXA
 -35         17          -10      +1
                Ext -10
```

Figure 6. A model for a LUCA promoter sequence generated as an AT-rich repeat of TATAAAAG boxes and BREs. The repeat sequence simplifies to an archaeal and a bacterial promoter.

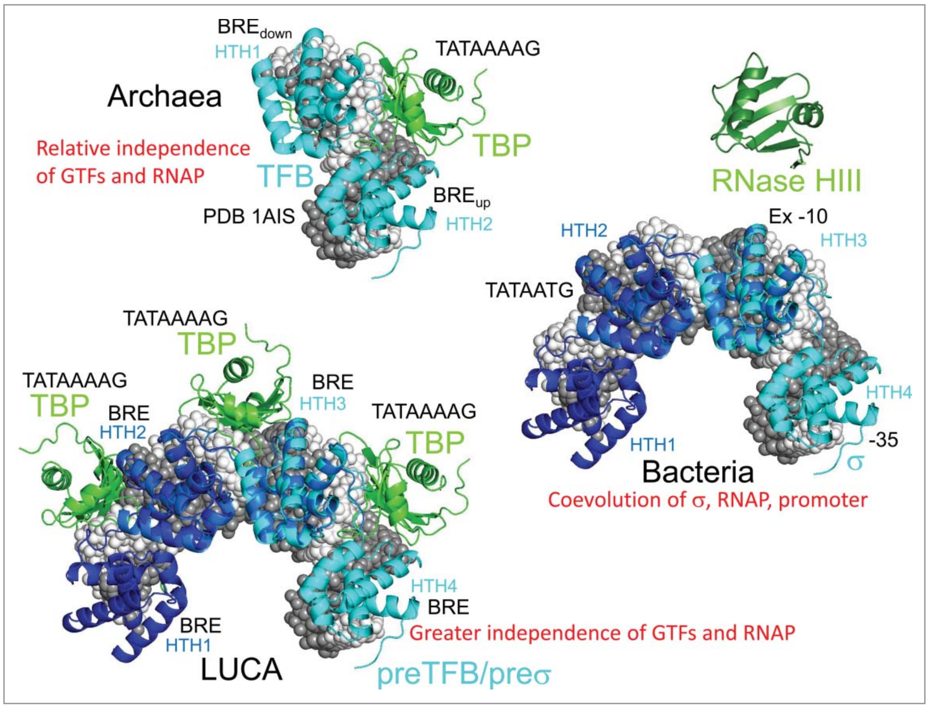


Figure 7. A model for primordial GTFs on a LUCA promoter and for radiation to archaea and bacteria. The model was constructed by superimposing 3-PDB 1AIS structures (archaeal TBP-TFB-promoter DNA).[44] The 4-HTH primordial initiation factor was generated by sequential alignment of 3 2-HTH repeats of TFB HTH_1-(HTH_2/HTH_1)-(HTH_2/HTH_1)-HTH_2. Bacterial systems are posited to have lost TBP and to have made the σ factor more strongly coevolved with promoter DNA and RNAP than in archaea or at LUCA.

promoter model is obtained by superimposing 3-PDB 1AIS structures with TFB HTH_1-(HTH_2/HTH_1)-(HTH_2/HTH_1)-HTH_2. Because in the model 4 of 6-HTH units are superimposed (HTH_2/HTH_1), 6-HTH units reduce to 4-HTH units, as are found in bacterial σ factors. What results is a promoter repeat of 3-TATAAAAG boxes and 4-BREs. Three TBP molecules bind the 3-TATAAAAG boxes and the 4-HTH repeats each bind a BRE. Because the archaeal TBP-TFB-promoter 1AIS structure was used to generate the LUCA TBP-primordial initiation factor-promoter structure, archaeal transcription is obtained from the LUCA model by simplification. To suppress unwanted transcription starts in archaeal promoters, selection against multiple TATA boxes is expected, leading to degeneration of the initial repeat structure (Fig. 6). Bacteria do not utilize TBP, although bacteria encode RNase HIII, a TBP structural homolog (Fig. S9).[36,42] As described above, bacterial σ factors are derived from a 4-HTH repeat primordial initiation factor. In archaeal transcription, GTFs are more independent of RNAP in the initiation mechanism and are not as powerfully coevolved as they are in bacteria, in which RNAP, the promoter and the σ factor are more strongly interacting and codependent. The degeneracy and specialization of the σ 4-HTH repeats, therefore, are explained by co-evolutionary pressures. In the LUCA model, TBP and the 4-HTH primordial initiation factor function as agents to facilitate DNA opening and appear more independent of RNAP than archaeal and bacterial GTFs. Because, without a mechanism to open DNA, DNA genomes cannot evolve from RNA genomes, TBP and the hypothesized 4-HTH primordial initiation factor are posited to be the most central components of this ancient mechanism, and TBP and the 4-HTH primordial initiation factor are posited to have been present at LUCA.[35]

To test the feasibility of this model, one might combine a LUCA promoter, LUCA GTFs and a 2-DPBB type RNAP on negatively supercoiled DNA and search for evidence of promoter opening and/or accurate initiation. A very similar experiment was successfully done decades ago using TBP, TFIIB and RNAP II on a supercoiled DNA template.[45]

LECA: the RNAP II CTD repeat

LECA is the last eukaryotic common ancestor (~1.6 to 2.2 billion years ago), and a surprisingly simple and compelling model for genesis of eukaryotes is available from extensive phylogenomic studies.[9-11,46-49] A story of LECA, moreover, can be related by focusing on multi-subunit 2-DPBB type RNAPs and their associated factors (Figure S10). A current model is that eukaryotes arose from endosymbiosis between a eukaryote-like Lokiarchaeota phylum archaea[8,50] and a resident population of α-proteobacteria. The mitochondria and mitochondrial DNA are relics of the α-proteobacteria. Many bacterial genes were transferred to what was initially the archaeal genome. Remarkably, a massive attack was launched against the archaeal genome by a α-proteobacterial group II self-splicing intron mobile genetic element.[10,47,51,52] Eukaryotic splicing and widely dispersed introns developed from jumping and insertions of group II intron elements. Eukaryotic genome complexity, therefore, results (in part) from group II intron invasion, Lokiarchaeota DNA and α-proteobacterial DNA. Many other bacterial genomes are represented in eukaryotic DNA, presumably acquired from many horizontal gene transfers (i.e., from virus-mediated horizontal gene transfer, plasmid-mediated horizontal gene transfer, endosymbiosis and natural horizontal gene transfer). Of course, there are additional contributions to eukaryote genome complexity such as genome duplications, transposons and insertion elements. The nucleus arose as a defense against translation of intron sequences that invaded genes. The splicing apparatus is posited to have arose to restrain and regulate self-splicing of group II introns. Many other eukaryote-specific genes and functions evolved as a result of novel pressures from the initial genome fusion, intron invasion, new cell architectures and resulting chaos.

More complex eukaryotic genomes allowed for duplication of genes and diversification of functions (Figure S10). Two-DPBB type RNAPs diversified into RNAP I, II and III, and RNAP II acquired the carboxy terminal domain (CTD), a heptapeptide (7aa; consensus ^{1}YSPTSPS^{7}) repeat, on the largest (β' type) subunit.[4,53] RNAPs I, II and III and the CTD on RNAP II appear to be rooted in LECA. The CTD is thought to have initially evolved to couple splicing of introns to transcription, making the CTD YSPTSPS repeat another evolutionary defense against group II intron invasion.[53] Subsequently, the CTD became a much more general scaffold for evolutionary innovation linked to RNAP II transcription. Ultimately, an extensive CTD interactome coupled transcription to many related processes. The CTD interactome interfaces with the chromatin interactome, linking transcription with epigenetics, and complex eukaryotic signaling is also coupled to the CTD and chromatin interactomes. Because of regulation by cyclins and cyclin-dependent kinases, the RNAP II transcription cycle, which is regulated by the CTD interactome, resembles a eukaryotic cell cycle, indicating that the RNAP II transcription cycle and the eukaryotic cell cycle were coevolved.[4] Consistent with the complexity of the CTD and its interactome, regulation of RNAP II promoter-proximal pausing by the CTD and the CTD interactome appears to be a primary marker of animal body plan complexity (i.e., the pre-Cambrian and Cambrian Explosion). The simple idea is that ever more nuanced RNAP II regulation, centered in eukaryotes on the CTD interactome, licenses higher order organismal complexity. Consistent with this idea, the number of repeats found in the CTD in different eukaryotes tends to correlate with overall organismal complexity.[53] According to this view, the CTD repeat initially evolved to cope with group II intron invasion and then became a scaffold for evolutionary innovation that was essential to support, and helped to drive, increasing eukaryote complexity. Landmark innovations in animal complexity, therefore, appear to track with innovations in RNAP II regulation via the CTD repeat interactome.

Genesis of life on earth

Remarkably, the story of genesis of life on earth is told in stunning detail by tracking α/β fold proteins, 2-DPBB type RNAPs, RNAP GTFs, RNAP promoters, the CTD on RNAP II and the extensive CTD interactome. According to this view, repeating protein motifs (bounded by solubility) form a surprisingly dominant component of the fabric of life, as if, in early evolution, structural complexity may have been positively selected, even before active sites gained refined function and specificity. We suggest that some initial repeats must have had sufficient function to compete successfully with or to collaborate with ribozymes.

α/β proteins ($(\beta{-}\alpha)_n$ repeat proteins) have built-in structure through parallel interacting β-sheets and solubility through pairing of each β-sheet with its α-helix. As such, α/β proteins appear to have won a primordial race to structure, solubility and function. Remarkably, the $\beta{-}\alpha{-}\beta{-}\alpha$ pattern remains discernable after almost ~4.1 billion years of evolution on a ~4.6 billion year old earth (Figs. S2-S8). This indicates, perhaps contrary to intuition, that evolution can be very conservative of core protein repeat motifs over a span of ~4 billion years.

In unexpected ways, evolution describes transcription, and transcription describes evolution. Life originated from a RNA-protein world that included 2-DPBB type RNA-template dependent RNAPs, and, therefore, RNAPs are central to the story of life on earth from: 1) the RNA-protein world to LUCA; 2) from LUCA diverging to bacteria and archaea; and 3) from endosymbiotic fusion of a Lokiarchaeota and an α-proteobacterium at LECA. Thus, a stunningly simple working model for genesis of life on earth is available from evolution of 2-DPBB type RNAPs, RNAP GTFs, RNAP promoters, RNAPs I, II and III, and the RNAP II CTD (Figure S10). The model describes the major branch points in evolution and provides surprising insights into eukaryote and animal complexity. The model provides a deeply conceptual understanding of life on earth.

The complexity of protein structures and, therefore, of living systems can be bounded by the linked issues of: 1) solubility; and 2) closure. The repeating protein sequence folds that win tend to be soluble and to gain closure by forming pseudo-symmetric forms: i.e. dimeric repeats and tetrameric repeats. Examples of dimeric repeats: 1) double-$\Psi-\beta$-barrels; 2) 2-double-$\Psi-\beta$-barrel type RNAPs; 3) TBP (TATA-binding protein); 4) RIFT barrels; 5) TFB (Transcription Factor B); 6) TIM barrels ($\beta-\alpha-\beta-\alpha-\beta-\alpha-\beta-\alpha$ dimer). Examples of tetrameric repeats: 1) the primordial initiation factor (4-HTH); 2) σ factor (4-HTH); 3) TIM barrels ($\beta-\alpha-\beta-\alpha$ tetramer). Closure is attained by formation of barrels and/or pseudo symmetry. What this means is that there are limits to the complexity evolution will achieve, and limits are often approached by simple bounds of solubility and closure. The CTD (52 repeats in humans) is not strongly bounded by these rules because it is largely unstructured, attached to a soluble scaffold (RNAP II) and heavily modified. The length of the CTD is bounded by its functionality.[53]

Abbreviations

CTD	carboxy terminal domain
DNAP	DNA polymerase
DPBB	double-Ψ-β-barrel
GTF	general transcription factor
LUCA	last universal common cellular ancestor of bacteria, archaea and eukaryotes
LECA	last eukaryotic common ancestor
RIFT	occurrence in riboflavin synthases, F1 ATPase and translation factors
TBP	RNA polymerase (RNAP); TATA-binding protein
TFB	Transcription Factor B
BRE	TFB-recognition element
TIM	triose phosphate isomerase

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

Funding

This research was supported by National Science Foundation MCB-1050867 and National Institutes of Health R01 GM 092949.

References

[1] Soding J, Lupas AN. More than the sum of their parts: on the evolution of proteins from peptides. BioEssays 2003; 25:837-46; PMID:12938173; http://dx.doi.org/10.1002/bies.10321

[2] Iyer LM, Koonin EV, Aravind L. Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. BMC Struct Biol 2003; 3:1; PMID:12553882; http://dx.doi.org/10.1186/1472-6807-3-1

[3] Iyer LM, Aravind L. Insights from the architecture of the bacterial transcription apparatus. J Struct Biol 2012; 179:299-319; PMID:22210308; http://dx.doi.org/10.1016/j.jsb.2011.12.013

[4] Burton ZF. The Old and New Testaments of gene regulation: Evolution of multi-subunit RNA polymerases and co-evolution of eukaryote complexity with the RNAP II CTD. Transcription 2014; 5:e28674-1-12.

[5] Lane WJ, Darst SA. Molecular evolution of multisubunit RNA polymerases: sequence analysis. J Mol Biol 2010; 395:671-85; PMID:19895820; http://dx.doi.org/10.1016/j.jmb.2009.10.062

[6] Lane WJ, Darst SA. Molecular evolution of multisubunit RNA polymerases: structural analysis. J Mol Biol 2010; 395:686-704; PMID:19895816; http://dx.doi.org/10.1016/j.jmb.2009.10.063

[7] Koonin EV. The origins of cellular life. Antonie van Leeuwenhoek 2014; 106:27-41; PMID:24756907; http://dx.doi.org/10.1007/s10482-014-0169-5

[8] Spang A, Saw JH, Jorgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJ. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 2015; 521:173-9; PMID:25945739; http://dx.doi.org/10.1038/nature14447

[9] Koonin EV. Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier? Philos Trans R Soc Lond B Biol Sci 2015; 370:20140333; PMID:26323764; http://dx.doi.org/10.1098/rstb.2014.0333

[10] Koonin EV. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Biology Direct 2006; 1:22; PMID:16907971; http://dx.doi.org/10.1186/1745-6150-1-22

[11] Martin W, Koonin EV. Introns and the origin of nucleus-cytosol compartmentalization. Nature 2006; 440:41-5; PMID:16511485; http://dx.doi.org/10.1038/nature04531

[12] McLachlan AD. Gene duplication and the origin of repetitive protein structures. Cold Spring Harb Symp Quant Biol 1987; 52:411-20; PMID:3454271; http://dx.doi.org/10.1101/SQB.1987.052.01.048

[13] McLachlan AD. Repeating sequences and gene duplication in proteins. J Mol Biol 1972; 64:417-37; PMID: 5023183; http://dx.doi.org/10.1016/0022-2836(72)90508-6

[14] Hocker B, Schmidt S, Sterner R. A common evolutionary origin of two elementary enzyme folds. FEBS letters 2002; 510:133-5; PMID:11801240; http://dx.doi.org/10.1016/S0014-5793(01)03232-X

[15] Toth-Petroczy A, Tawfik DS. The robustness and innovability of protein folds. Curr Opin Struct Biol 2014; 26:131-8; PMID:25038399; http://dx.doi.org/10.1016/j.sbi.2014.06.007

[16] Hleap JS, Susko E, Blouin C. Defining structural and evolutionary modules in proteins: a community detection approach to explore sub-domain architecture. BMC structural biology 2013; 13:20; PMID:24131821; http://dx.doi.org/10.1186/1472-6807-13-20

[17] Kim C, Basner J, Lee B. Detecting internally symmetric protein structures. BMC bioinformatics 2010; 11:303; PMID:20525292; http://dx.doi.org/10.1186/1471-2105-11-303

[18] Caetano-Anolles G, Yafremava LS, Gee H, Caetano-Anolles D, Kim HS, Mittenthal JE. The origin and evolution of modern metabolism. Int J Biochem Cell Biol 2009; 41:285-97; PMID:18790074; http://dx.doi.org/10.1016/j.biocel.2008.08.022

[19] Ma BG, Chen L, Ji HF, Chen ZH, Yang FR, Wang L, Qu G, Jiang YY, Ji C, Zhang HY. Characters of very ancient proteins. Biochem Biophys Res Commun 2008; 366:607-11; PMID:18073136; http://dx.doi.org/10.1016/j.bbrc.2007.12.014

[20] Nagano N, Orengo CA, Thornton JM. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 2002; 321:741-65; PMID: 12206759; http://dx.doi.org/10.1016/S0022-2836(02)00649-6

[21] Soding J, Remmert M, Biegert A. HHrep: de novo protein repeat detection and the origin of TIM barrels. Nucleic Acids Res 2006; 34:W137-42; PMID:16844977; http://dx.doi.org/10.1093/nar/gkl130

[22] Hanukoglu I. Proteopedia: Rossmann fold: A beta-alpha-beta fold at dinucleotide binding sites. Biochem Mol Biol Educ 2015; 43:206-9; PMID:25704928; http://dx.doi.org/10.1002/bmb.20849

[23] Nath N, Mitchell JB, Caetano-Anolles G. The natural history of biocatalytic mechanisms. PLoS Computational Biol 2014; 10:e1003642; PMID:24874434; http://dx.doi.org/10.1371/journal.pcbi.1003642

[24] Rossmann MG, Argos P. Protein folding. Annu Rev Biochem 1981; 50:497-532; PMID:7023364; http://dx.doi.org/10.1146/annurev.bi.50.070181.002433

[25] Alva V, Koretke KK, Coles M, Lupas AN. Cradle-loop barrels and the concept of metafolds in protein classification by natural descent. Curr Opin Struct Biol 2008; 18:358-65; PMID:18457946; http://dx.doi.org/10.1016/j.sbi.2008.02.006

[26] Coles M, Hulko M, Djuranovic S, Truffault V, Koretke K, Martin J, Lupas AN. Common evolutionary origin of swapped-hairpin and double-psi beta barrels. Structure 2006; 14:1489-98; PMID:17027498; http://dx.doi.org/10.1016/j.str.2006.08.005

[27] Coles M, Djuranovic S, Soding J, Frickey T, Koretke K, Truffault V, Martin J, Lupas AN. AbrB-like transcription factors assume a swapped hairpin fold that is evolutionarily related to double-psi beta barrels. Structure 2005; 13:919-28; PMID:15939023; http://dx.doi.org/10.1016/j.str.2005.03.017

[28] Yamaguchi Y, Mura T, Chanarat S, Okamoto S, Handa H. Hepatitis delta antigen binds to the clamp of RNA polymerase II and affects transcriptional fidelity. Genes Cells 2007; 12:863-75; http://dx.doi.org/10.1111/j.1365-2443.2007.01094.x

[29] Macnaughton TB, Lai MM. HDV RNA replication: ancient relic or primer? Curr Top Microbiol Immunol 2006; 307:25-45; PMID:16903219

[30] Zuo Y, Steitz TA. Crystal Structures of the E. coli Transcription Initiation Complexes with a Complete Bubble. Mol Cell 2015; 58:534-40; PMID:25866247; http://dx.doi.org/10.1016/j.molcel.2015.03.010

[31] Vassylyev DG, Vassylyeva MN, Zhang J, Palangat M, Artsimovitch I, Landick R. Structural basis for substrate loading in bacterial RNA polymerase. Nature 2007; 448:163-8; PMID:17581591; http://dx.doi.org/10.1038/nature05931

[32] Vassylyev DG, Vassylyeva MN, Perederina A, Tahirov TH, Artsimovitch I. Structural basis for transcription elongation by bacterial RNA polymerase. Nature 2007; 448:157-62; PMID:17581590; http://dx.doi.org/10.1038/nature05932

[33] Liu B, Zuo Y, Steitz TA. Structural basis for transcription reactivation by RapA. Proc Natl Acad Sci U S A 2015; 112:2006-10; PMID:25646438; http://dx.doi.org/10.1073/pnas.1417152112

[34] Salgado PS, Koivunen MR, Makeyev EV, Bamford DH, Stuart DI, Grimes JM. The structure of an RNAi polymerase links RNA silencing and transcription. PLoS Biol 2006; 4:e434; PMID:17147473; http://dx.doi.org/10.1371/journal.pbio.0040434

[35] Burton SP, Burton ZF. The sigma enigma: Bacterial sigma factors, archaeal TFB and eukaryotic TFIIB are

homologs. Transcription 2014; 5:e967599; PMID: 25483602; http://dx.doi.org/10.4161/21541264.2014.967599

[36] Brindefalk B, Dessailly BH, Yeats C, Orengo C, Werner F, Poole AM. Evolutionary history of the TBP-domain superfamily. Nucleic Acids Res 2013; 41:2832-45; PMID:23376926; http://dx.doi.org/10.1093/nar/gkt045

[37] Bae B, Feklistov A, Lass-Napiorkowska A, Landick R, Darst SA. Structure of a bacterial RNA polymerase holoenzyme open promoter complex. eLife 2015; 4:1-23.

[38] Darst SA, Feklistov A, Gross CA. Promoter melting by an alternative sigma, one base at a time. Nat Struct Mol Biol 2014; 21:350-1; PMID:24699085; http://dx.doi.org/10.1038/nsmb.2798

[39] Feklistov A, Sharon BD, Darst SA, Gross CA. Bacterial Sigma Factors: A Historical, Structural, and Genomic Perspective. Annu Rev Microbiol 2014; 68:357-76; PMID:25002089

[40] Feklistov A, Darst SA. Structural basis for promoter-10 element recognition by the bacterial RNA polymerase sigma subunit. Cell 2011; 147:1257-69; PMID:22136875; http://dx.doi.org/10.1016/j.cell.2011.10.041

[41] Chen J, Darst SA, Thirumalai D. Promoter melting triggered by bacterial RNA polymerase occurs in three steps. Proc Natl Acad Sci U S A 2010; 107:12523-8; PMID: 20615963; http://dx.doi.org/10.1073/pnas.1003533107

[42] Chon H, Matsumura H, Koga Y, Takano K, Kanaya S. Crystal structure and structure-based mutational analyses of RNase HIII from Bacillus stearothermophilus: a new type 2 RNase H with TBP-like substrate-binding domain at the N terminus. J Mol Biol 2006; 356:165-78; PMID: 16343535; http://dx.doi.org/10.1016/j.jmb.2005.11.017

[43] Blombach F, Salvadori E, Fouqueau T, Yan J, Reimann J, Sheppard C, Smollett KL, Albers SV, Kay CW, Thalassinos K, Werner F. Archaeal TFEalpha/beta is a hybrid of TFIIE and the RNA polymerase III subcomplex hRPC62/39. eLife 2015; 4:e08378; PMID:26067235; http://dx.doi.org/10.7554/eLife.08378

[44] Kosa PF, Ghosh G, DeDecker BS, Sigler PB. The 2.1-A crystal structure of an archaeal preinitiation complex: TATA-box-binding protein/transcription factor (II)B core/TATA-box. Proc Natl Acad Sci U S A 1997; 94:6042-7; PMID:9177165; http://dx.doi.org/10.1073/pnas.94.12.6042

[45] Parvin JD, Sharp PA. DNA topology and a minimal set of basal factors for transcription by RNA polymerase II. Cell 1993; 73:533-40; PMID:8490964; http://dx.doi.org/10.1016/0092-8674(93)90140-L

[46] Koonin EV, Yutin N. The dispersed archaeal eukaryome and the complex archaeal ancestor of eukaryotes. Cold Spring Harb Perspect Biol 2014; 6:a016188; PMID:24691961; http://dx.doi.org/10.1101/cshperspect.a016188

[47] Rogozin IB, Carmel L, Csuros M, Koonin EV. Origin and evolution of spliceosomal introns. Biol Direct 2012; 7:11; PMID:22507701; http://dx.doi.org/10.1186/1745-6150-7-11

[48] Mans BJ, Anantharaman V, Aravind L, Koonin EV. Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex. Cell Cycle 2004; 3:1612-37; PMID:15611647; http://dx.doi.org/10.4161/cc.3.12.1316

[49] Baum DA, Baum B. An inside-out origin for the eukaryotic cell. BMC Biol 2014; 12:76; PMID:25350791; http://dx.doi.org/10.1186/s12915-014-0076-2

[50] Nasir A, Kim KM, Caetano-Anolles G. Lokiarchaeota: eukaryote-like missing links from microbial dark matter? Trends Microbiol 2015; 23(8):448-50; PMID:26112912

[51] Koonin EV, Csuros M, Rogozin IB. Whence genes in pieces: reconstruction of the exon-intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. Wiley Interdiscip Rev RNA 2013; 4:93-105; PMID:23139082; http://dx.doi.org/10.1002/wrna.1143

[52] Koonin EV. Intron-dominated genomes of early ancestors of eukaryotes. J Hered 2009; 100:618-23; PMID: 19617525; http://dx.doi.org/10.1093/jhered/esp056

[53] Yang C, Stiller JW. Evolutionary diversity and taxon-specific modifications of the RNA polymerase II C-terminal domain. Proc Natl Acad Sci U S A 2014; 111:5920-5; PMID: 24711388; http://dx.doi.org/10.1073/pnas.1323616111

[54] Coles M, Diercks T, Liermann J, Groger A, Rockel B, Baumeister W, Koretke KK, Lupas A, Peters J, Kessler H. The solution structure of VAT-N reveals a 'missing link' in the evolution of complex enzymes from a simple betaalphabetabeta element. Curr Biol 1999; 9:1158-68; PMID:10531028; http://dx.doi.org/10.1016/S0960-9822(00)80017-2

[55] Zhang Y, Degen D, Ho MX, Sineva E, Ebright KY, Ebright YW, Mekler V, Vahedian-Movahed H, Feng Y, Yin R, et al. GE23077 binds to the RNA polymerase 'i' and 'i+1′ sites and prevents the binding of initiating nucleotides. eLife 2014; 3:e02450; PMID:24755292