**FACULTÉ DES ÉTUDES SUPÉRIEURES
ET POSTDOCTORALES**

uOttawa

L'Université canadienne
Canada's university

**FACULTY OF GRADUATE AND
POSTDOCTORAL STUDIES**

Robert Carter

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Biology)

GRADE / DEGREE

Department of Biology

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Evolution of the Eukaryotic RNA Polymerases

TITRE DE LA THÈSE / TITLE OF THESIS

Guy Drouin

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

Stéphane Aris-Brosou

Michel Dumontier

Gary Olsen (University of Illinois)

Howard Rundle

Xuhua Xia

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

uOttawa

L'Université canadienne
Canada's university

Evolution of the Eukaryotic RNA polymerases

Robert Carter

Thesis submitted to the
Faculty of graduate and Postdoctoral studies
In partial fulfillment of the requirements
For the PhD degree in the Department of Biology.

Department of Biology
Faculty of Science
University of Ottawa

©Robert Carter, Ottawa, Canada, 2009

Library and Archives
Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-63684-8
Our file Notre référence
ISBN: 978-0-494-63684-8

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Contents

List of Tables.....	iv
List of figures.....	v
Abstract.....	vi
Résumé.....	viii
Acknowledgements.....	x
Chapter 1	1
Introduction	1
Chapter 2	14
Structural differentiation of the three eukaryotic RNA polymerases	14
Abstract	15
Introduction.....	16
Results	19
Discussion	28
Methods.....	34
Acknowledgements	42
Tables	43
Figure legends.....	46
Figures	47
Chapter 3	51
The Evolutionary Rates of Eukaryotic RNA Polymerases and of Their Transcription Factors Are Affected by the Level of Concerted Evolution of the Genes They Transcribe	51
Abstract	52
Introduction.....	53
Materials and Methods	58
Results	63
Discussion	66
Acknowledgements	68
Tables	69
Figure legends.....	72
Figures	73
Chapter 4	74
The increase in the number of subunits in eukaryotic RNA polymerase III relative to RNA polymerase II is due to the early recruitment of general transcription factors.....	74
Abstract	75

Introduction.....	76
Materials and Methods	80
Results	81
Discussion	86
Acknowledgements	92
Figure legends.....	93
Figures	95
Chapter 5	99
Discussion, Conclusion, and Future Directions	99
Literature Cited	107

List of Tables

Table 1.1 Composition of cellular RNAPs and their GTFs.....	12
Table 2.1 Cleft loop functions and amino acid locations in the two largest subunits of RNA polymerase II.....	43
Table 2.2 Length variations between paralogous cleft loops of eukaryotic RNAPs..	44
Table 2.3 Parameter estimates and model results for fixed-site models.....	45
Table 2.4 Probabilities that the type 1 and type 2 sites identified by the Knudsen <i>et al.</i> (ref. 45) method are randomly distributed.....	45
Table 3.1 Proteins of the three eukaryotic transcriptional machineries analyzed in this study	69
Table 3.2 Mean nonsynonymous substitution rate ($\pm$ standard error) of group 1 paralogous RNA polymerase subunits and comparisons between RNA polymerase paralogs.....	70
Table 3.3 Nonsynonymous evolutionary rates of TFs and non-paralogous subunits.....	71

List of figures

Figure 1.1 Steps in the transcription cycle by cellular RNAPs.....	13
Figure 2.1 Sequence features of the two largest subunits of the eukaryotic RNAPs.....	47
Figure 2.2 Mean normalized relative substitution rates of cleft loops.....	48
Figure 2.3 Distribution of type 1 and type 2 sites in the eukaryotic RNAPs determined using the method of Knudsen <i>et al.</i>	49
Figure 2.4 Distribution of sites that have evolved via relaxed selection along the ancestral branch in all three polymerases according to the branch-site method of Zhang <i>et al.</i>	50
Figure 3.1 Schematic illustration of the molecular coevolution hypothesis.....	73
Figure 4.1 Regions of homology between the RNAPIII-specific subunits and the RNAPII general transcription factors TFIIIE and TFIIIF.....	95
Figure 4.2 Regions of homology between RPC37 and RPC53 and the human TFIIIF heterodimer corresponding to the dimer binding interface of TFIIIF α and β	96
Figure 4.3 Region of homology between the RPC82 C-terminal domain and the <i>Sulfolobus solfataricus</i> N-terminal extended winged helix domain of TFE.....	97
Figure 4.4 Parsimonious scenarios for the evolution of the cellular RNAPs, emphasizing the homologies between different cellular RNAP subunits and their GTF proteins.....	98

Abstract

All eukaryotes contain at least three homologous RNA polymerases (RNAPI, RNAPII, and RNAPIII), which are directed towards their target genes by interactions with RNAP-specific general transcription factors (GTFs). Although the RNAPs transcribe a subset of nuclear genes, it is unknown if they have evolved altered functionality. An *in silico* approach to identify putative differences in the three RNAPs was used to identify whether any functional differences likely exist by identifying amino acids that have experienced shifts in evolutionary rates and by identifying length differences between the cleft loops of the RNAPs. The difference in the levels of concerted evolution experienced by the genes transcribed by RNAPs was then exploited to test the molecular coevolution hypothesis of Dover and Flavell (1984. *Cell*. 38:622-623.). According to the molecular coevolution hypothesis, concertedly evolving DNA increases the evolutionary rate of any interacting proteins. We thus compared the evolutionary rates of the three RNAPs and their GTFs, since their target genes undergo different levels of concerted evolution. Finally, the origins of the 5 subunits that are specific to RNAPIII were examined, since no homologous relationships have been identified thus far for any of these subunits.

Several sites that have experienced shifts in substitution rates in the ancestral RNAPs were found in all three enzymes and these sites were clustered near the active sites in all cases. Several cleft loops with different lengths (in amino acids) between the three RNAPs were also identified. The validity of the molecular coevolution hypothesis was largely confirmed via the demonstration that most

subunits of RNAPI evolve faster than those of RNAPIII and most subunits of RNAPIII evolve faster than RNAPII. This is consistent with the molecular coevolution hypothesis because RNAPI experiences higher levels of concerted evolution than the target genes of RNAPIII. The evolutionary rates of the GTFs of RNAPI and RNAPIII were also higher than those of RNAPII. Finally, four of the five RNAPIII-specific subunits were identified as homologs of RNAPII GTFs, indicating that several of the GTF paralogs existed in ancestral RNAPIII and were subsequently recruited to the enzyme before the diversification of eukaryotes.

Résumé

Tous les eucaryotes contiennent au moins trois ARN polymérases (les ARNP I, II, et III), qui sont dirigées vers leurs gènes cibles par des interactions avec des facteurs généraux de transcription (FGT). On ignore si ces trois ARNP ont évolué des fonctionnalités différentes. Nous avons utilisé une approche *in silico* pour déterminer si des différences fonctionnelles existent entre ces trois ARNP. Pour ce faire, nous avons analysé les différences de taux de substitutions des acides aminés et les différences de longueur dans les régions en boucles flexibles. Nous avons aussi testé l'hypothèse de coévolution moléculaire de Dover et Flavell. Selon cette hypothèse, l'évolution concertée de l'ADN augmente le taux d'évolution de toutes les protéines qui interagissent directement ou indirectement avec les gènes qui subissent différents taux d'évolution concerté. Nous avons donc comparé les taux d'évolution des trois ARNP et leurs FGT, étant donné que leurs gènes cibles subissent différents niveaux de l'évolution concertée. Enfin, les origines des 5 sous-unités qui sont spécifiques à l'ARNPIII ont été examinées, car aucun homologue n'avait, jusqu'à présent, été identifié pour ces sous-unités.

Nous avons identifié plusieurs sites qui ont connu des fluctuations des taux de substitution dans les trois ARNP ancestrales et ces sites sont regroupés à proximité des sites actifs. Plusieurs régions en boucles flexibles ont aussi des longueurs différentes chez les trois ARNP. La validité de l'hypothèse de coévolution moléculaire a été largement confirmée par la démonstration que la plupart des sous-unités d'ARNPI évoluent plus rapidement que celles des ARNPIII et que la plupart des sous-unités d'ARNPIII évoluent plus vite que ceux des ARNPPII. Ceci est

cohérent avec l'hypothèse de coévolution moléculaire, car les gènes transcrits par L'ARNPI subissent des niveaux plus élevés d'évolution concertée que les gènes transcrits par l'ARNPIII. Les taux d'évolution des FGT de l'ARNPI et l'ARNPIII étaient également plus élevés que ceux de l'ARNPII. Enfin, quatre des cinq sous-unités spécifiques à l'ARNPIII ont été identifiés comme étant des homologues des FTG de l'ARNPII, indiquant que plusieurs des FTG qui existaient dans l'ARNPIII ancestrale et ont par la suite été incorporés dans cette enzyme avant la diversification des eucaryotes.

Acknowledgements

I would like to thank the following people for their support: Dr. Guy Drouin for acting as my thesis supervisor, my fiancée Michele Megson for her helpful suggestions and for acting as a sounding board for many of my ideas and presentations, and my parents Ruth and Ken for always encouraging me throughout my education. I would also like to thank Stéphane Aris-Brosou for his insights and helpful advice and, with Michel Dumontier, for agreeing to be a member of my graduate committee. Finally, I would like to thank all of my fellow graduate students that helped me in various ways throughout my PhD degree.

Chapter 1

Introduction

The DNA-dependent RNA polymerases comprise a subset of the more general RNA polymerase enzymes. RNA polymerases refer to the set of all proteins that catalyze the addition of nucleoside monophosphates in the 5' to 3' direction from nucleotide precursors. All RNA polymerases require a template, which specifies the order of addition of nucleotides. These templates are either RNA or DNA. Consequently, RNA polymerases are categorized as either RNA-dependent RNA polymerases or DNA-dependent RNA polymerases. Structural comparisons show that these two classes of RNA polymerases share a core Double-*Psi* beta barrel fold and are thus related, although they are highly divergent (Iyer *et al.* 2003). RNA-dependent RNA polymerases are found in RNA viruses and in RNAi pathways in many eukaryotes (Siomi and Siomi, 2009), where they are used to transcribe antisense genomes into translatable sense mRNA and to amplify the silencing signals of exogenous RNA, respectively. DNA-dependent RNA polymerases are responsible for all transcription in cells, including transcription of cytoplasmic DNA viruses. Thus, they transcribe target genes into either mRNA or non-coding RNAs, thereby permitting expression of genomic information. Here we are concerned with aspects of the evolution of the eukaryotic DNA-dependent RNA polymerases, henceforth called RNAPs.

Before proceeding, we briefly review the function and regulation of transcription by RNAPs in order to put this work into perspective. Although RNAPs themselves are able to catalyze the addition of nucleoside monophosphates to an elongating RNA

transcript, they require a suite of other proteins to direct RNAPs to initiate transcription of the correct genes and at the correct time. While transcription of a particular gene involves a complex interplay between gene-specific activators, these activators inevitably recruit a collection of general transcription factors (GTFs) that are specific to each RNAP and are required for directed transcription of target genes. The GTFs recognize either a) specific DNA elements in the neighbourhood of the gene, b) other GTFs, or c) the RNAP with which they associate (Paule and White, 2000; Thomas and Chiang, 2006). These GTFs essentially lay the foundation for an incoming RNAP, directing it to the appropriate position upstream from a transcription start site (figure 1.1). The ensemble of GTFs that recruits the RNAP is termed the Pre-Initiation Complex (PIC). The PIC recruits the appropriate RNAP, which is then poised for transcription initiation. Transcription initiation involves the transition from the closed to the open complex, which is accomplished with the aid of some of the GTFs. The open complex is formed when the non-template and template strands of the DNA are melted (figure 1.1). Upon melting of the DNA, the template strand is fed into the active site of the RNAP while the non-template strand straddles the outside of the RNAP. The resulting transcription bubble has an unpaired template strand which is used as the template for directing synthesis of a new transcript. The RNAP then undergoes abortive transcription/initiation, which is characterized by the repeated synthesis and release of many short transcripts, until a transcript length of up to 15 nts is reached (Dvir, 2002; Hsu, 2002). At this point, the RNAP becomes highly processive, usually elongating the transcript until the transcript is terminated (figure 1.1).

The processive elongation phase is coordinated by a large number of mobile regions called the cleft loops, many of which can be seen in the structure of a transcribing RNAPII in figure 2.1 (Cramer *et al.* 2001; Gnatt *et al.* 2001; Naji *et al.* 2008). These are a series of mobile regions up to 35 amino acids in length. When the template strand is fed into the active site during the transition from closed to an open complex, several of these adopt an ordered structure and interact with the template strand, presumably stabilizing its presence in the RNAP active site (Cramer *et al.* 2001; Gnatt *et al.* 2001). During the transition from abortive to processive elongation, several additional cleft loops become ordered via interactions with the RNA:DNA hybrid, likely playing a role in the stabilization of the hybrid and facilitating the transition to a processive elongation complex (Westover *et al.* 2004). The addition of nucleoside monophosphates to the 3' end of the elongating RNA is largely driven by two additional cleft loops, the bridge helix and the trigger loop, via a Brownian ratchet mechanism (Bar-Nahum *et al.* 2005; Brueckner and Cramer, 2008). Although the details of this mechanism are still uncertain, several aspects of this mechanism are clear. The 3' nucleoside of the RNA in the RNA:DNA hybrid cycles between a pre- and post-translocated state (Nudler, 2008). In the pre-translocated state, the 3' end of the RNA occupies the $i+1$ site and is abutted against the bridge helix, a cleft loop that occupies the position where the next templated base would otherwise lie. In the post-translocated state, the RNA:DNA hybrid has shifted away from the bridge helix by one nt into the i site, revealing the $i+1$ site. The hybrid cycles between the pre- and post-translocated state until the next incoming nucleotide binds the $i+1$ site, thereby locking the hybrid into the post-translocation state. This cycle of events involves a number of structural

changes in both the bridge helix and the trigger loop, although the exact mechanism is still unknown (Nudler, 2009). Subsequently, nucleotide addition occurs via catalysis with the aid of the trigger loop (Wang *et al.* 2006). These actions result in the increase of the length of the RNA by one nt and the cycle repeats. Other cleft loops are described in chapter 2 along with their functions.

Eukaryotes have evolved three RNAPs compared to the single enzyme present in prokaryotes (table 1.1). Bacterial RNAP is the simplest cellular RNAP. It has four different subunits, one of which forms a homodimer, to make a five subunit RNAP (Zhang *et al.* 1999). Most Archaeal RNAPs contain 11 subunits and are thus compositionally more complex than bacterial RNAPs, although they have orthologs of all four bacterial subunits (table 1.1)(Hirata *et al.* 2008; Kusser *et al.* 2008; Korkhin *et al.* 2009). All sequenced eukaryotes contain three RNAPs, although plants have an additional two RNAPs (Pontier *et al.* 2005). These are RNA polymerase I (RNAPI), RNA polymerase II (RNAPII), and RNA polymerase III (RNAPIII). These are composed of 12, 14, and 17 subunits in yeast, respectively. The 12 subunits of RNAPII either have homologs in the other two eukaryotic RNAPs or are physically shared between two or three of the RNAPs (table 1.1). Thus, the evolution of the cellular RNAPs is characterized by the addition of subunits to a conserved core that is maintained in all three domains of life.

Prokaryotes use specific members of related GTFs to recruit the prokaryotic RNAP to GTF-specified genes, whereas each of the three eukaryotic RNAPs along with its associated GTFs transcribes a subset of the nuclear genes. Although there is only a single bacterial RNAP, a collection of homologous σ GTFs functionally partitions

bacterial genes into separate classes that are each recognized by a specific σ -RNAP complex called the holoenzyme (Wösten, 1998). Finer regulatory control of bacterial gene expression is acquired by an array of gene-specific inducers and repressors. Archaeal transcription regulation is less well understood, although recent studies have demonstrated a similar mechanism of partitioning the genes into classes as seen in the multiple- σ method of bacteria. Many archaea contain several paralogs of the general transcription factors TBP and TFB and specific combinations of these two GTFs are used to recruit RNAP to specific genes (Facciotti *et al.* 2007; Lu *et al.* 2008). Unlike prokaryotes, the multiple RNAPs of eukaryotes each transcribe a subset of the nuclear genes. Each of these RNAPs has its own suite of GTFs. RNAPI transcribes rRNA, RNAPII transcribes the miRNA and mRNA, and RNAPIII transcribes 5S rRNA and tRNAs.

The functional divergence of paralogous subunits and the acquisition of RNAP-specific subunits have granted each of the eukaryotic RNAPs unique functionality. The RNAPI and RNAPIII paralogs of RPB9 (RPA9 and RPC9, respectively) contribute to the transcriptional fidelity of their respective polymerases by proofreading RNA and removing misincorporated nucleotides from the 3' end (Alic *et al.* 2007; Kuhn *et al.* 2007). Proofreading is essential for these RNAPs since deletion mutants of RPC9 are non-viable (Chedin *et al.* 1998) and those of RPA9 are thermosensitive (Nogi *et al.* 1993). In contrast, deletion mutants of their RPB9 paralog are viable (Nesser *et al.* 2006). In addition to its role in RNA proofreading, RPC9 is also required for facilitated recycling of RNAPIII, in which an RNAPIII that has completed transcription re-initiates transcription from the same template (Landrieux *et al.* 2006). Facilitated recycling is

specific to RNAPIII, highlighting another functional innovation of a particular eukaryotic RNAP.

In addition to the fidelity requirements of the eukaryotic RNAPs conferred by RPB9 and its RNAPI and RNAPIII paralogs, RNAPI and RNAPIII each contain several subunits that are lacking homologous counterparts in RNAPII and thus provide unique functionality to these RNAPs. The RNAPI-specific heterodimer RPA34.5-RPA49 displays elongation stimulatory activity (Kuhn *et al.* 2007) and cooperates with RPA7 (an RNAPI subunit paralogous to RPB7) in the recruitment of RNAPI to the preinitiation complex at the rRNA gene promoter (Beckouet *et al.* 2008). RNAPIII contains 2 separate sub-complexes, RPC31-RPC34-RPC82 and RPC37-RPC53, both considered part of the RNAPIII protein complex since they remain bound to the RNAPs throughout the transcription cycle. Note that RPC31 is only present in unikonts (Proshkina *et al.* 2006). Each member of the RPC31-RPC34-RPC82 heterotrimer is required for cell viability (Mosrin *et al.* 1990; Chiannikulchai *et al.* 1992; Stettler *et al.* 1992), although functions have only been assigned to RPC34 and RPC31. In *S. cerevisiae*, RPC34 is implicated in both the recruitment of RNAPIII to the promoter via interactions with TFIIB-BRF (Brun *et al.* 1997; Andrau *et al.* 1999; Kenneth *et al.* 2008), as well as in the subsequent step of open complex formation (Brun *et al.* 1997). RPC31 is also required for transcription initiation (Thuillier *et al.* 1995). The RNAPIII-specific subcomplex RPC37-RPC53 is involved in transcript elongation rate and is required for terminator recognition (Landrieux *et al.* 2006).

Although the RNAPs each display some unique behaviours such as PIC-recruitment mechanisms, termination mechanisms, and fidelity, it is unknown whether

the eukaryotic polymerases are otherwise functionally equivalent. The two largest subunits of the RNAPs are each paralogous across the three RNAPs. These subunits make up the bulk of these enzymes, each of them totaling over 1000 amino acids in length, and act as binding platforms for the other RNAP subunits (Ishiguro *et al.* 1998). Thus, a key function of the largest subunits of each RNAP is to bind and recruit the remaining polymerase subunits, which are then able to associate with the correct PIC. In addition to their role in recruiting the correct RNAP subunits, the two largest subunits of the RNAPs shape the active site clefts and coordinate most of the transcription reaction (Cramer *et al.* 2001; Jasiak *et al.* 2006; Kuhn *et al.* 2007). Thus, it is a possibility that the two largest subunits have evolved RNAP-specific functions by fine-tuning the amino acid composition near their active sites concomitantly with the remodeling of the outer regions of these subunits required for the recruitment of the appropriate subunits.

We test this hypothesis in chapter 2 by using an *in silico* approach to identify amino acids that have experienced changes in substitution rates in the two largest subunits of the three RNAPs. Since shifts in amino acid substitution rates implies changes in functional constraints, the presence of such sites in the 2 largest subunits would suggest functional differences near the active sites of the RNAPs. In addition to identifying rate-shift sites, we compare the substitution rates and the length (in amino acids) of homologous cleft loops in the three RNAPs. Since the cleft loops are integral to RNAP function, differences in either their lengths or substitution rates between the three RNAPs would also be indicative of functional differences.

The target genes of the three transcriptional systems each have unique promoter organization and unique mechanisms of evolution that have the potential to affect the evolution of the RNAPs and GTFs with which they interact. To recapitulate the target genes of each transcriptional system, RNAPI transcribes rRNA, RNAPII transcribes mRNA and miRNA, and RNAPIII transcribes 5S rRNA, tRNA, and several small non-coding RNAs including a subset of miRNAs. Eukaryotic rRNA genes are organized as one to several tandem arrays per genome of usually over 100 copies each (Prokopowich *et al.* 2003). These tandem arrays evolve largely by unequal crossing-over (Liao *et al.* 1999). This mechanism of concerted evolution tends to rapidly homogenize the individual units of a tandem array. Since the promoters of each rRNA gene unit are contained within the tandem arrays, they are also rapidly homogenized. The tRNA target genes of RNAPIII are dispersed throughout the genome, whereas the 5S genes can be either dispersed in the genome or found in the rRNA gene tandem arrays. Although the tRNA genes are dispersed, they also evolve by concerted evolution due to the high similarity between different tRNA genes (Amstutz *et al.* 1985; Leung *et al.* 1991). The homogenization of tRNA genes is not as rapid as that of rRNA gene because they undergo gene conversion instead of unequal crossing over, which tends to occur at a slower rate (Szostak and Wu 1980; Amstutz *et al.* 1985). A unique feature of RNAPIII-transcribed genes is that they contain internal promoters. These promoters are bound by the general transcription factors TFIIC and TFIIIA, which ultimately function to position TFIIIB upstream from the transcription initiation site (Paule and White, 2000). Consequently, the tRNA promoters also evolve largely by gene conversion, which causes homogenization of tRNA, although at a rate that is slower

than that of unequal crossing over (Szostak and Wu 1980; Amstutz *et al.* 1985). In contrast to class I and III genes, the protein-coding targets of RNAPII do not typically evolve by concerted evolution. This is because RNAPII transcribes thousands of genes in a genome, most of which are unrelated.

The molecular coevolution hypothesis of Dover and Flavell (1984) predicts that proteins that interact with DNA elements that are subject to rapid homogenization are likely to evolve at a higher rate than proteins that interact with DNA elements that evolve by purely vertical means. Consequently, this hypothesis predicts that RNAPI and its GTFs should evolve faster than RNAPIII and its GTFs. Likewise, RNAPIII and its GTFs should evolve faster than RNAPII and its GTFs.

In chapter 3, we test the molecular coevolution hypothesis using RNAP and GTF proteins from each of the three transcriptional systems. Specifically, we calculate the mean number of non-synonymous substitutions per site of each protein in pairwise comparisons of up to 25 ascomycete species. We then test whether the evolutionary rates of proteins from the three transcriptional systems are consistent with the molecular coevolution hypothesis, after controlling for known effectors of evolutionary rates such as gene expression level, structure, and functional constraints (Graur and Li, 2000; Pal *et al.* 2001; Drummond *et al.* 2006; Bloom *et al.* 2006).

As noted above, both RNAPI and RNAPIII contain subunits that have no counterparts in RNAPII. Although strictly true, the RNAPI-specific RPA34.5-RPA49 heterodimer has recently been shown to be homologous to the binding interface between the N-termini of the subunits of the TFIIF heterodimer, which is a general transcription factor of RNAPII (Kuhn *et al.* 2007). Heterospecific complementation of S.

cerevisiae RNAPIII subunits revealed that none of the five RNAPIII-specific proteins of *S. pombe* were able to replace their *S. cerevisiae* orthologs *in vivo* due to their rapid evolution, in contrast to 10 of the 12 remaining RNAPIII proteins (Proshkina *et al.* 2006). Structural and mutation studies indicate that the two RNAPIII-specific sub-complexes occupy similar positions as TFIIE and TFIIF of RNAPII (Proshkina *et al.* 2006; Chen *et al.* 2007; Fernandez-Tornero *et al.* 2007). Since the RNAPIII-specific subunits evolve quickly and occupy similar locations as two of the RNAPII GTFs, it is possible that the RNAPIII-specific subunits are in fact homologous to these GTFs, but that this homology has gone unrecognized due to a lack of sensitive homology detection methods as well as their rapid evolution.

In chapter 4, we use a sensitive homology detection method to identify homologs of the RNAPIII-specific subunits to determine whether they are indeed homologous to proteins of the RNAPII GTFs. The method used, HHpred, relies on comparing a Hidden Markov Model (HMM) of a query protein against a set of target gene family HMMs. We conclude the chapter with a summary of the evolution of the eukaryotic RNAPs (Soding *et al.* 2005).

Collectively, the three investigations presented here in chapters 2-4 address several aspects of the evolution of the eukaryotic RNAPs. They address i) the functional consequences of the paralogous relationship of the two largest subunits between the three RNAPs, ii) the effects of concerted evolution of target genes on the evolutionary rates of the three transcriptional systems, and iii) the relationship between the RNAP-specific subunits of RNAPIII and the general transcription factors of RNAPII. This work is concluded with chapter 5, which discusses several implications of the findings

reported in chapters 2-4. These include potential homologies between the GTF repertoires of the three RNAPs, the true eukaryotic ortholog of the single archaeal RNAP, and potential scenarios for the formation of the three eukaryotic RNAPs.

Table 1.1. Composition of cellular RNAPs and their GTFs. Rows with subunits in red text contain homologous subunits in some or all RNAPs. Subunits in the same row that have grey backgrounds denote subunits that are shared between polymerases. Subunits in black text have no homologs. Some GTFs are composed of multiple proteins, whereas RNAP entries are all individual proteins.

		RNAP			
	Bacterial	Archaeal	RNAPI	RNAPII	RNAPIII
GTF	σ	TBP	TBP	TBP	TBP
		TFB		TFIIB	TFIIB-Brf
		TFE		TFIIE	
			CF	TFIIA	TFIIIA
			UAF	TFIIF	TFIIIC
				TFIIS	TFIIB-Bdp
RNAP	β'	Rpo1	RPA1	RPB1	RPC1
	β	Rpo2	RPA2	RPB2	RPC2
	α	Rpo3	RPAC3	RPB3	RPAC3
		Rpo4	RPA4	RPB4	RPC4
		Rpo5	RP5	RP5	RP5
	ω	Rpo6	RPA6	RPA6	RPA6
		Rpo7	RPA7	RPB7	RPC7
		Rpo8	RP8	RP8	RP8
			RPA9	RPB9	RPC9
		Rpo10	RP10	RP10	RP10
	α	Rpo11	RPAC11	RPB11	RPAC11
		Rpo12	RP12	RP12	RP12
				RPA34.5	RPC37
				RPA49	RPC53
				RPC31	
			RPC34		
			RPC82

Figure 1.1 Steps of transcription from the formation of a pre-initiation complex (PIC) through processive elongation. Grey ellipses represent GTFs; connected white ellipses represent an RNAP; orange curved lines inside RNAP denote disordered cleft loops, straight orange lines represent ordered cleft loops; the black rectangle represents the promoter region; the blue double line represents DNA; short green lines denote RNA abortive transcripts, whereas the long green line represents a growing RNA transcript; black arrow represents transcription initiation site. See Chapter 1 text for details.

Chapter 2

Structural differentiation of the three eukaryotic RNA polymerases

Robert Carter and Guy Drouin

Département de biologie et Centre de recherche avancée en génomique
environnementale, Université d'Ottawa, Ottawa, Ontario, Canada, K1N 6N5

Running title: Functional differentiation of RNA polymerases

Correspondance to: Guy Drouin, Département de biologie, Université d'Ottawa, 30
Marie Curie, Ottawa, Ontario, Canada, K1N 6N5. Tel.: (613) 562-5800 ext. 6052, FAX:
(613) 562-5486, E-mail: gdrouin@science.uottawa.ca

(Published in *Genomics*. 2009 Aug 29. PMID: 19720141. Epub ahead of print)

Author Contributions: RC and GD designed the study, RC performed all analyses, and RC and GD wrote the paper.

Abstract

All eukaryotes have three different RNA polymerases (RNAPs) which transcribe different types of genes. RNA polymerase I transcribes rRNA genes, RNA polymerase II transcribes mRNA, miRNA, snRNA, and snoRNA genes, and RNA polymerase III transcribes tRNA and 5S rRNA genes. Here, we use an *in silico* approach to identify putative functional differences between these three RNAPs. Our results show that cleft loops are the most conserved regions of RNAPs, but that these regions have several significant length differences between the three RNAPs. There were also significant shifts in the rates of evolution of some amino acid sites in all three enzymes prior to the diversification of eukaryotes and these amino acids are preferentially located near the catalytic center of three enzymes. These results suggest that functional differences exist between the three RNAPs and that these differences are all clustered near the active sites.

Keywords: RNA polymerase, transcription, eukaryotic, adaptation, selective pressures

Introduction

The cellular DNA-dependent RNA polymerases (RNAPs) generate RNA transcripts from genes encoded in the genome. RNAPs are composed of between five and 17 subunits, at least four of which are paralogs across all three domains of life (Cramer, 2002; Best *et al.* 2004). A defining characteristic of all eukaryotes is the presence of at least three nuclear RNAPs in their genomes, which were created by a series of gene duplication events sometime before the diversification of eukaryotes. Each of these RNAPs transcribes a particular subset of nuclear genes. RNA polymerase I (RNAPI) transcribes rRNA genes, RNA polymerase II (RNAPII) transcribes mRNA, miRNA, snRNA, and snoRNA genes, and RNA polymerase III (RNAPIII) transcribes tRNA and 5S rRNA genes. This is in contrast with prokaryotes where a single RNA polymerase is responsible for the transcription of all genes.

In over 1.5 billion years of evolution since their divergence from each other, the three transcription machineries discussed here have adapted themselves to efficiently transcribe their respective target genes (Izban and Luse, 1992; Lang and Reeder, 1995; Peyroche *et al.* 2000; Orlicky *et al.* 2001; Siaut *et al.* 2003; Zhang *et al.* 2005; Landrieux *et al.* 2006; Alic *et al.* 2007; Kuhn *et al.* 2007). These adaptations have occurred by both modification of existing subunits and general transcription factors (GTFs) as well as the recruitment of novel ones. These subunits recognize and form a complex with their appropriate RNAP by binding either to the largest (RNAPI-1, RNAPII-1 and RNAPIII-1) or second largest (RNAPI-2, RNAPII-2 and RNAPIII-2) subunit of each RNAP, which each exist as paralogs across the three RNAPs. One of the roles of these paralogous

subunits is thus to act as polymerase-specific interaction hubs for the remaining subunits, allowing each subunit to interact with its correct RNAP.

In addition to providing unique substrates that polymerase-specific subunits bind to give each of the RNAPs their specific functionality, the two largest subunits also shape the active site cleft of the enzymes where the transcription reaction occurs. The RNAPII enzyme contains 11 cleft loops in the two largest subunits of the RNAPs that are critically involved in polymerase functionality (Cramer *et al.* 2001; Gnatt *et al.* 2001; Epshtein *et al.* 2002; Kettenberger *et al.* 2004; Westover *et al.* 2004; Bar-Nahum *et al.* 2005; Suenaga *et al.* 2006; Wang *et al.* 2006; Brueckner and Cramer, 2008; Naji *et al.* 2008; Tan *et al.* 2008). These cleft loops, 7 of which are in the largest subunit and 4 in the second largest subunit, are loops between 8 and 35 residues in length in *Saccharomyces cerevisiae* (Table 2.1). Although they are spread throughout the primary sequence of the two largest subunits of the RNAPs, they are all oriented towards the active site (i.e., the main enzyme channel) in the actual protein complex (Figure 2.1). These regions are typically highly mobile due to a lack of structural contacts with non-neighboring sites in the nucleic-acid free RNAPs, but become ordered in the actively transcribing DNA-bound enzyme, the ternary elongation complex (TEC). Their lack of structural contacts in the nucleic acid-free enzyme makes them susceptible to thermal motion, so they are likely to fluctuate between conformations until ordered by their interactions in the TEC.

It is well known that the RNAPs have essentially the same two-ion catalytic transcription mechanism and that subunits specific to each provide polymerase-specific functions, but it remains unknown whether residues of the two largest subunits

themselves contribute to polymerase-specific functionality of the three RNAPs. Here we address this issue by using two different *in silico* approaches to identify potential regions in the two largest subunits of RNAPI, II, and III that might contribute to polymerase-specific functions. Our analysis includes sequences from 17 eukaryotic organisms from diverse phyla including fungi, plants, metazoans, alveolates, and amoebozoans (listed in the Methods section below). We compare the relative substitution rates and length variations between cleft loops of the three RNAPs, as these are known to be critically involved in polymerase functionality (Cramer *et al.* 2001; Gnatt *et al.* 2001; Epshtein *et al.* 2002; Kettenberger *et al.* 2004; Westover *et al.* 2004; Bar-Nahum *et al.* 2005; Suenaga *et al.* 2006; Wang *et al.* 2006; Brueckner and Cramer, 2008; Naji *et al.* 2008; Tan *et al.* 2008). We also examine amino acids in the two largest subunits that underwent ancestral shifts in amino acid substitution rates following their birth by duplications, telltale signs of functional differences between paralogs (Wang and Gu, 2001; Gu *et al.* 2002). We identify several cleft loops with significantly different lengths between at least two of the three RNAPs. We also find that many sites near the active site in the two large subunits of the RNAPs have experienced shifts in substitution rates following their formation by duplication events but preceding the diversification of eukaryotes. Together, these results suggest that the two largest subunits of each of the RNAPs have evolved specific functions above-and-beyond those conferred by their polymerase-specific subunits and transcription factors.

Results

Length variation in the cleft loops of the RNAPs

We examined our alignments for the presence of the cleft loops in the largest and second largest subunits of each of the three RNAPs (figure 2.1) and then determined whether the cleft loops of the different RNAPs had polymerase-specific length deviations using non-parametric Kruskal-Wallis tests (Zar, 1999). Kruskal-Wallis statistical tests were used because the lengths were discrete and non-normally distributed. Of all the differences between cleft loops, the most evident is the variation in the flap loop lengths between the three polymerases (Table 2.2). Whereas most RNAPII-2 and RNAPIII-2 subunits contain flap loops that are at least 12 residues in length, the flap loop is non-existent in most RNAPI-2 subunits. The RNAPI of the three earliest diverging protists contain an 8 residue flap loop, but this loop is absent from other phyla. All other species examined here, with the exception of one insertion in *Neurospora crassa*, are entirely lacking this cleft loop in RNAPI. Thus, non-protist RNAPI lacks the flap loop, whereas it has been maintained in RNAPII and RNAPIII of all species, although often truncated in the latter. Low-resolution cryo-electron micrographs of free RNAPI indicate that the domain where the flap loop should be located is highly flexible in this RNAP and can swivel between open and closed states, in contrast to that of RNAPII (De Carlo *et al.* 2003). Thus, the domain that contains the flap loops in RNAPII and RNAPIII but not in RNAPI also displays unique behavior in RNAPI relative to the other RNAPs.

Switches 2 and 3 have remained a constant length in all three RNAPs (Table 2.2). Only switch 1, the longest of the switch regions, showed a significant length deviation. Although switch 1 has remained 23 residues in length in RNAPII and RNAPIII of nearly all species examined, it is one residue shorter in almost all RNAPI sequences. Aside from these switch regions, only the bridge and trigger loop showed no significant length variations (Table 2.2). Like the switch regions, the bridge and trigger loop are grouped together in the actual protein complex. The bridge helix can be seen traversing the main enzyme channel underneath the nucleic acid in Figure 2.1; the trigger loop lies just underneath towards the front of the enzyme (bottom of structure in Figure 2.1) but tends to resist crystallization efforts and is thus not visible.

All other cleft loops examined here showed significant length variation between the three eukaryotic RNAPs (Table 2.2). These include the rudder, zipper, and lid of the largest subunit and the two fork loops of the second largest subunit. Thus, significant length variations are found between homologous cleft loops involved in interactions with the DNA:RNA hybrid. Additionally, length variations are also found in the zipper and fork loop 2, which are likely responsible for maintenance of the transcription bubble (Cramer *et al.* 2001; Gnatt *et al.* 2001; Westover *et al.* 2004).

Stronger purifying selection is acting on the cleft loops compared to the remaining residues of the two largest subunits of the RNAPs

We examined the selective pressures exerted on codons of the cleft loops to determine their relative importance to the RNAPs compared to the remaining codons of the two largest subunits. Due to their functional involvement in the transcription process, we expected these to evolve slower than the average residues in the RNAPs. A series of nested models were fitted to each subunit to determine if the evolutionary process is significantly different between these cleft loops and the remainder of these proteins (Yang and Swanson, 2002). Model B fit the data significantly better than the null model for both subunits 1 and 2 ($p < 0.001$, Table 2.3). Model B differs from the null model by allowing the evolutionary rates to differ between the two partitions while maintaining a constant nonsynonymous/synonymous (ω) ratio across the two partitions. Codon substitution rates were lower in the cleft loops in both subunits, when compared to the remaining codons. We next tested model D versus the null model B, thereby allowing the evolutionary rates and ω to differ between the two partitions. This alternate model fit the data significantly better than model B for both subunits ($p < 0.01$, Table 2.3). According to model D, the cleft loops of both subunits evolved with lower rates and lower ω ratios than the remaining codons. Thus, substitution rates and ω ratios are lower in the cleft loops, indicating that these regions are relatively slow to evolve and are resistant to non-synonymous changes. This is despite their lack of defined structural constraints.

Relative rates of evolution of amino acids in the cleft loops of the three RNAPs

Since the cleft loops experience strong purifying selection, we examined these regions in the two largest subunits of the RNAPs to see if they were uniformly resistant to evolutionary change in each individual RNAP. Different levels of constraint exerted on homologous regions might result from different functionality at these regions, especially since these regions are largely devoid of any structural constraints. According to this hypothesis, we would expect to find significantly different selective pressures in homologous cleft loops if these sites are responsible for functional differences between the RNAPs. To measure these rates, we compared the mean relative rate of residues from each cleft loop relative to the average amino acid substitution rate of the whole protein.

The flap loops of RNAPII and RNAPIII evolve very rapidly compared to the average substitution rate of amino acids in these proteins (Figure 2.2). Indeed, based on the normalized distribution of relative rates, these residues evolve between 1 and 1.5 standard deviations above the average amino acid residue in each paralog. No relative rates could be determined for the flap loop of RNAPI, as this region does not exist in non-protist RNAPs (see below).

The slowest evolving cleft loops are the bridge helix, the trigger loop, and switch 2 of all three RNAPs, as well as the fork loop 2 of RNAPIII (Figure 2.2). All of these regions are integral to the actual transcription mechanism, which consists of ribonucleotide addition, polymerase translocation, and pyrophosphate release. The

average relative rates of the remaining cleft loops are all within one standard deviation of the average substitution rate of the protein in which they are encoded (Figure 2.2).

Phylogenetic identification of amino acid residues involved in functional divergence and their locations in the RNAPs

We constructed maximum likelihood phylogenies of the two largest subunits of 17 diverse eukaryotic RNAPs (Supplementary Figure 2.2). Using these phylogenies, we examined the coding sequences of the two large paralogous subunits for sites that have undergone shifts in substitution rates along specific lineages of the phylogeny. We first used the method of Knudsen *et al.* (2003) to identify both type 1 and type 2 sites in amino acid sequences. Type 1 sites are sites which are variable in one polymerase but fixed in the other two (or vice versa). Type 2 sites are sites which are fixed for one amino acid in one polymerase and fixed for a different amino acid in the other two polymerases (or vice versa). Type 1 sites therefore represent higher substitution rates in extant sequences whereas type 2 sites represent higher substitution rates in ancestral sequences (see the Materials and Methods section below for further discussion of type 1 and 2 sites). The predicted residues can be found in Supplementary Table 2.2. The number of significant ($\alpha = 0.05$) type 1 sites per RNAP (RNAPI, 77; RNAPII, 74; RNAPIII, 90) was greater than the number of significant ($\alpha = 0.05$) type 2 sites per RNAP (RNAPI, 28; RNAPII, 28; RNAPIII, 15). We mapped these residues onto the structure of an actively polymerizing *S. cerevisiae* RNAPII TEC (ref. 16; PDB code

1R9T) because a) it contains many of the cleft loops that were absent from earlier structures and b) it is of a TEC and thus the majority of the residues are near the locations where they act during the transcription process (Figure 2.3).

To quantify the spatial distribution of type 1 and type 2 sites in the two large paralogous subunits of each of the three RNA polymerases, we used computer simulations to generate an inter-residue distance measure for randomly placed rate-shift sites in the protein (see Methods). We tested the null hypothesis that these rate-shift sites occur randomly throughout the structure by comparing the empirical spatial distribution of rate-shift sites against the null distribution to assess whether these residues were significantly clustered compared to randomly placed residues. Significance was assessed according to the null distribution using two-sided tests with $\alpha = 0.05$.

In all three RNAPs, the null hypothesis of randomly distributed type 1 sites could not be rejected (Table 2.4). Type 1 sites are therefore nonclustered throughout the structure (Figure 2.3). In contrast, the null hypothesis of randomly distributed type 2 sites was rejected for all three RNAPs (table 2.4). Type 2 sites are therefore clustered and, as shown in Figure 2.3, preferentially located at the center of the enzyme where the nucleic acid and polymerase interact and where the transcription reaction itself takes place.

We also examined whether the type 1 and type 2 sites are overrepresented in the cleft loops. Using chi-square tests, we tested the null hypothesis that a residue in a cleft loop has the same probability of being a rate-shift site as any other residue in the protein. Although 10% of type 1 compared to 15% of type 2 sites were found in the cleft

loops, which make up roughly 10% of the protein residues, we were unable to reject the null hypothesis for either type 1 ($\chi^2 = 0.25$, $p = 0.62$) or type 2 ($\chi^2 = 0.45$, $p = 0.50$) sites.

In addition to the method of Knudsen *et al.* (2003), we used the branch-site codon model of Zhang *et al.* (2005) to identify amino acids that were most likely involved in functional divergence of the three paralogs. We applied both branch-site test 1 and test 2, which test for relaxed constraints and positive selection along ancestral branches of the RNAPs, respectively (Zhang *et al.* 2005). Only test 1 produced significant results (likelihood ratio test, $p < 0.0001$ for both subunits) with 27 codons common to the three RNAPs having a posterior probability greater than 50%. Among these 27 sites, only site 472 of the largest subunit had a greater than 95% posterior probability. Test 2 gave likelihood ratio test p -values of 0.999998 for both subunits. This model only identified 27 codons that evolved by relaxed purifying selection along the 3 basal branches and did not identify any positively selected codons. Note that, since these 27 codons had increased substitution rates in the three basal branches, they encode residues that are equivalent to the type 2 sites of the Knudsen *et al.* (2003) method.

We mapped all 27 codons identified by test 1 onto the structure of a TEC and quantified the distribution of these sites in the same fashion (Figure 2.4). Although all but one of these sites had only between 50% and 95% posterior probability of having evolved under relaxed purifying selection in the 3 basal branches, we nevertheless mapped them on a 3D structure because the likelihood ratio test used to test for significance of the branch-site codon model demonstrated that such sites do in fact exist. As can be seen in Figure 2.4, these sites are clustered around the active site of

the RNAPs. Furthermore, like the results of the Knudsen *et al.* (2003) method, the null hypothesis of randomly distributed sites was rejected for these codons ($p = 0.006$).

Finally, as with the sites identified by the Knudsen *et al.* (2003) method, the sites identified by test 1 of the Zhang *et al.* (2005) method are not overrepresented in the cleft loops ($p = 0.42$).

There is some overlap between the sites predicted by the Knudsen *et al.* (2003) and Zhang *et al.* (2005) methods (Supplementary Table 2.2). Site 472 in the largest subunit was predicted by both methods. Site 527 in the largest subunit and site 835 in the second largest subunit of RNAPII and RNAPIII were also predicted by both methods, as was site 781 in RNAPI and RNAPII. Sites 341, 777, and 1400 of the largest subunit and site 226 of the second largest subunit were all predicted in the codon model and in only one RNAP using the method of Knudsen *et al.* (2003)

The limited overlap between the Zhang *et al.* (2005) and the Knudsen *et al.* (2003) methods, given that they each identified some 30 ancestral sites involved in the differentiation of the three RNAPs, is likely the result of how these methods were used to identify these sites. We used the Zhang *et al.* (2005) method to identify codons that have evolved by relaxed purifying selection (or positive selection) in all three lineages. Thus, for a site to be identified by this method, the same codon must have evolved by relaxed purifying selection (or positive selection) independently along the basal branch of all three RNAPs. In contrast, the amino acid-based method of Knudsen *et al.* (2003) identifies sites that had increased ancestral substitution rates in only one RNAP (i.e., relative to the other two RNAPs). Furthermore, the Zhang *et al.* (2005) method lacks power. Identifying *which* sites have undergone relaxed purifying selection is a much

more difficult problem than determining *whether* relaxed purifying selection has been a factor during the evolution of homologs (Zhang *et al.* 2005). Since the alternate hypothesis of the branch-site codon model for relaxed purifying selection fit our dataset better than the null hypothesis ($p < 0.0001$), only the identity of these sites is in question, not whether or not they exist.

Discussion

Our results show that each of the three eukaryotic RNAPs examined in this study contains cleft loop length differences and amino acids near their active sites that have experienced ancestral shifts in substitution rates. This suggests that the paralogous largest and second largest subunits of the eukaryotic RNAPs do more than just act as interfaces that allow recruitment of polymerase-specific subunits. It suggests that the active sites of these enzymes have additionally been modified in ways that likely have subtle effects on transcription, although this awaits experimental verification.

We compared the cleft loops of the three RNAPs and found differences in both the lengths and evolutionary rates of many of these regions (table 2.2 and figure 2.2). Since these regions are free to sample conformational space via thermal motion, we reasoned that length differences between these regions would alter the possible conformational space that they could visit, while simultaneously increasing or reducing their reach within the nucleic acid-binding and catalytic central channel of the RNAPs. Differences in function could easily arise by extending or reducing the reach of these regions to allow them to acquire conformations needed for particular interactions.

The cleft loops responsible for the transcription mechanism itself are highly conserved in length and are the slowest evolving of the cleft loops, which attests to the conserved transcriptional mechanism of the three RNAPs examined here (table 2.2 and figure 2.2). The bridge helix and trigger loop are invariable in length and evolve slowly compared to the other cleft loops. Transcriptional properties such as fidelity and transcriptional rate are easily altered by amino acid changes introduced into these

regions (Tan *et al.* 2008) so it not unexpected that these regions are resistant to change. With the exception of a one residue shorter switch 1 in RNAP1 compared to the other two RNAPs, all switch regions evolve slowly and show no significant length variation between RNAPs (table 2.2). The switch regions interact with both downstream DNA and the RNA-DNA hybrid and are believed to couple clamp closure to the presence of the hybrid (Gnatt *et al.* 2001). Thus, this aspect is likely conserved in the three RNAPs as well.

The cleft loops toward the rear of the enzyme near the upstream end of the RNA-DNA hybrid show the greatest variability in lengths both between and within a particular RNAP. The clamp-associated zipper, lid, and rudder cleft loops evolve at rates that are similar to the average residue of the encoding RNAPs (table 2.2 and figure 2.2). The high variability in length and the relatively fast evolutionary rate of these regions suggests that these cleft loops are likely to behave differently in the three RNAPs. The *S. cerevisiae* lid is believed to act as a wedge that separates the RNA-DNA hybrid at its upstream terminus (Westover *et al.* 2004). However, the relatively high substitution rate of this region (figure 2.2), and the variability in lengths (table 2.2) suggest that a unique sequence is not necessary, although a loop of some sort is required to act as a wedge to separate the heteroduplex nucleic acid. The length variations in fork loop 1 and the rudder, which promote stability of the RNA-DNA hybrid together with the lid, indicate that the RNAPs might differ in their ability to maintain a stable RNA-DNA hybrid.

The relatively high substitution rates of flap loop in RNAPII and RNAPIII is puzzling, especially since the flap loop is entirely missing in non-protist RNAPI (Figure 2.2; Table 2.2). This suggests that the flap loop is of little importance in any of the

RNAPs. However, this loop has been retained in the bacterial RNAP (Zhang *et al.* 1999; Cramer *et al.* 2001), which suggests otherwise. That the flap loop has been maintained in all bacterial and eukaryotic RNAPs, except for in non-protist RNAPI, suggests that the role that this region plays in the transcription reaction is either not required in RNAPI, or has been co-opted by another protein or region of this polymerase. A cryo-EM structure of RNAPI shows additional density at the dock domain of RNAPI relative to RNAPII (De Carlo *et al.* 2003). The RNAPII general transcription factor TFIIB interacts with the dock domain and traverse the region above where the flap loop is located in RNAPII (Bushnell *et al.* 2004; Chen and Hahn, 2004). RNAPIII has a homologous transcription factor, TFIIB, whereas RNAPI does not. Since TFIIB traverses the region where the flap loop is located in RNAPII and a homolog is present in RNAPIII, the lack of the flap loop in RNAPI might be due to differences in the way that RNAPI interacts with its GTFs.

The analyses of the amino acids experiencing ancestral shifts in substitution rates in the three eukaryotic RNAPs not only support the existence of such amino acids but also show that they are clustered near the active sites in all three RNAPs. The results from the analyses performed using the Knudsen *et al.* (2003) and Zhang *et al.* (2005) methods both support the existence of some 32 ancestral sites involved in the three RNAPs (Supplemental Table 2.2). Although the type 2 sites predicted by these two methods are not perfectly congruent, they are both significantly clustered near the active sites of the enzymes (Figures 2.3 and 2.4; Table 2.4) whereas the type 1 sites predicted by the Knudsen *et al.* method (2003) are nonclustered (Figure 2.3). The biological relevance of the clustering of type 2 sites as opposed to type 1 sites is

consistent with published results of well-characterized protein families, where it has been demonstrated that type 2 sites are better indicators of functional divergence than type 1 sites (Lopez *et al.* 2002; Gribaldo *et al.* 2003). Thus, following the duplication events that formed these new RNAPs, selection pressure near the active site of these enzymes was temporarily relaxed, allowing exploration of novel functionality. The exploration and subsequent fixation of new residues preceding the diversification of the eukaryotes in each of these polymerases at these sites suggests that these duplicates had acquired useful differences, as renewed selective pressure on these novel polymerase subunits prevented subsequent sequence change of these residues. In the case of type 1 sites, their random distribution indicates, as suggested by previous studies (Gribaldo *et al.* 2003), that such sites are not likely to be functionally important.

Although the biological significance of the type 2 sites and their location towards the active sites of the three RNAPs is uncertain, many of these changes might be due to interactions with other proteins. By definition, the amino acid changes of the type 2 sites occurred in the newly duplicated RNAPs of the earliest eukaryotes, coinciding in time with the origin of most of the polymerase-specific subunits and GTFs. Thus, it is possible that these sites in each polymerase interact with some of these polymerase-specific GTFs and accessory proteins, and thus coevolved. Consistent with this possibility, many polymerase-associated proteins are known or thought to interact with the active site of the RNAPs. Structural studies clearly show that TFIIB (Bushnell *et al.* 2004; Chen and Hahn, 2004) and TFIIS (Kettenberger *et al.* 2003) have finger-like projections that probe deep within the active site of RNAPII. Additionally, the mediator complex has recently been shown to contain many functionally important disordered

regions, some of which might probe the active site of RNAPII (Tóth-Petróczy *et al.* 2008). This possibility is supported by the extensive contacts observed via cryo-EM between the yeast mediator complex and RNAPII (Davis *et al.* 2002). Finally, the TFIIF α - β interface has a flexible arm domain that potentially interacts with the active site (Gaiser *et al.* 2000). The location of this interface near the active site of RNAPII is supported by the demonstration that the TFIIF α subunit, which encodes the arm domain as part of the binding interface, crosslinks to the TFIIB linker and B finger in the active site of RNAPII (Bushnell *et al.* 2004). Furthermore, mutagenesis screens have identified TFIIF α mutations at this binding interface that suppress TFIIB mutants known to interfere with normal RNAPII start site selection (Sun and Hampsey, 1995; Freire-Picos *et al.* 2005).

In conclusion, our sequence-based analyses of the differentiation of eukaryotic RNAPs demonstrate 1) that cleft loops are the most conserved regions of RNAPs, 2) that the cleft loops of the different RNAPs have several significant length differences which may endow each with specific functionality, 3) that there have been significant shifts in the rates of evolution of some 32 amino acids in all three enzymes prior to the diversification of eukaryotes, but concurrent with the recruitment of GTFs and accessory proteins, and 4) that these amino acids are clustered towards the center of the enzyme where the nucleic acid interacts with the polymerase, the transcription reaction occurs, and where many accessory factors and GTFs interact with the RNAPs to modulate aspects of the transcription process. Together, these results suggest that the two largest subunits of the three RNAPs contribute to polymerase-specific functionality beyond merely acting as polymerase-specific binding platforms for accessory factors

and GTFs. They also provide avenues for future experimental research into the divergence of the three RNAPs.

Methods

Sequences and alignments

The protein sequences of RNAPs are highly conserved. For example, the only differences between the mouse and human RNA polymerase II largest subunit protein sequences, which diverged from one another some 80 MYA, consists in a single amino acid difference and a four amino acid insertion over their 1970 amino acid long alignment (results not shown). For our data set, we therefore mainly used sequences from more distantly related species in order for our data set to contain enough sequence variability. On the other hand, we did not include the highly divergent sequences from *Trichomonas* and *Giardia* because their high degree of sequence divergence made it difficult to align them with the other eukaryotic sequences.

Protein sequences were obtained for the largest and second largest subunits of RNAPI, RNAPII, and RNAPIII of 17 eukaryotes using BLAST searches of the NCBI NR database with default settings. The *S. cerevisiae* protein sequences of the largest (CAA65619) and second largest (CAA99357) subunits were used as queries. Coding DNA sequences were obtained from the GenBank files of each protein sequence. The species used were as follows: *Arabidopsis thaliana*, *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Dictyostelium discoideum*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Neurospora crassa*, *Entamoeba histolytica*, *Trypanosoma brucei*, *Theileria parva*, *Plasmodium falciparum*, *Cryptococcus neoformans*, *Encephalitozoon cuniculi*, *Tribolium castaneum* and

Kluyveromyces lactis. The accession numbers of all eukaryotic sequences used are listed in Supplementary Table 2.1. We used the manual alignments of the two largest subunits constructed by Cramer *et al.* (2001) as profiles to which we aligned all paralogs of each subunit (Supplementary Figure 2.1) using the profile mode of T-Coffee, with all other parameters set at their default values (Notredame *et al.* 2000). DNA coding sequences were aligned to the amino acid alignments using a PERL script.

Throughout the chapter and in the supplemental information, amino acid sites are numbered relative to the sequence of the homologous (aligned) yeast RNAPII-1 and RNAPII-2 subunits (accession numbers CAA65619 and CAA99357, respectively). For example, amino acid site 472 of RNAPI1 corresponds to amino acid 472 of the yeast RNAPII1 sequence and not to amino acid 472 of RNAPI1. This numbering system was used in order to be able to determine the position of any given site in the 3D structure of the enzymes. Since high resolution RNAP structural data exists only for yeast RNAPII, all sites were numbered according to its amino acid sequences.

Detecting Homologous Regions

We defined homologous regions as stretches of at least 5 consecutive reliably aligned amino acid residues from the full Cramer-based alignments with no more than one gap at any position in each of the RNAP paralogs. Homologous regions separated by only one unreliably aligned column were connected into a contiguous homology region. We used these alignments of homologous regions (Supplementary Figure 2.1), which we call homologous alignments, for all analyses except the comparisons of length

deviations of the cleft loops (see below). These homologous alignments were used instead of the full Cramer-based alignments (Supplementary Figure 2.1) because of the poor alignment quality between many regions in the full alignments (caused by the large divergence time between the three RNAPs).

Phylogenetic Inference

We constructed maximum likelihood phylogenies of the two largest subunits of the RNAPs. These were required for the subsequent identification of functionally important residues in these proteins, which are characterized using the phylogenetic relationships between taxa. Phylogenies were constructed using the protein sequences of the homologous alignments, as described above, and PhyML (Guindon and Gascuel, 2003). We used eight discrete rate categories in our analyses, with the shape parameter α and the proportion of invariable sites estimated from the data. The substitution model used was that of JTT (Jones *et al.* 1992).

Comparison of cleft loops

Since cleft loops are likely the most functionally important regions of RNAPs, we examined the relative evolutionary rates of residues and the variation in lengths in the cleft loops between the three RNAPs to quantify the functional constraints operating on these regions. None of these regions are buried in the protein structure and only the bridge helix has continuous secondary structure throughout its length. The cleft loops are mostly free from the structural constraints that affect the evolution of other amino acids in the protein. Thus, resistance to evolutionary change in these regions is due almost entirely due to functional constraints (Cheng *et al.* 2005).

We used two different methods to measure the functional constraint acting on the cleft loops: one DNA based and one amino acid based. For the DNA based method, we used a fixed-site codon model (Yang and Swanson, 2002) to determine if these cleft loops experience stronger purifying selection than the remainder of the proteins. The fixed-site model estimates different parameter values for each partition in a data set. Since these partitions must be defined *a priori*, these models are useful to test hypotheses about different evolutionary constraints acting on different portions of a gene (Yang and Swanson, 2002). The codon alignments of each of the two largest subunits of the RNAPs were split into two partitions, one containing codons coding for residues of the cleft loops and the other containing all remaining codons of the alignment. We used likelihood ratio tests to compare the nested models A, B, and D (Yang and Swanson, 2002) to determine if the cleft loops were under stronger functional constraints than the remainder of the protein as measured by the non-

synonymous/synonymous substitution rate ratio (ω). Model A treats the two partitions as one, assigning them equal rates, transversion/transition rate ratios (κ), codon frequencies (π), and ω values; model B allows different rates in each partition but is otherwise identical to model A; model D extends model B by allowing different κ and ω values in each partition. Since different ω in each partition are included only in model D, model D must fit the data significantly better than models A or B before it can be concluded that the two partitions have evolved under different selective pressures.

The rate4site program (Mayrose *et al.* 2004) was used to identify the relative amino acid substitution rates between residues of each cleft loop of the three RNAPs. We compared the mean relative rate of residues from each cleft loop relative to the average amino acid substitution rate of the whole protein. Since this method calibrates the rate of a site against the average rate of the whole protein, it is not affected by differences in overall protein substitution rates when comparing homologous proteins that evolve at different rates.

We split our full Cramer-based amino acid alignments of each of the two largest subunits into the three component RNAP alignments, generating 6 alignments in total: two RNAPI, RNAPII, and RNAPIII alignments, one for each of the two largest subunits. We estimated the relative substitution rates of the cleft loops of each RNAP's subunit via the empirical Bayesian method of rate4site (Mayrose *et al.* 2004), using default program parameters, our maximum likelihood phylogenies and the full Cramer-based amino acid alignments as input. To measure the importance of each functional region, we calculated the mean relative rate of its component residues. This gives an indication of the relative importance of a region when compared to the rest of the protein,

assuming that the average substitution rate decreases with the increasing importance of the protein region.

We also examined length variations of the cleft loops between the three RNAPs. To identify length variations, we used non-parametric Kruskal-Wallis tests (Zar, 1999) for identifying differences in the number of residues found in each cleft loop between the three RNAPs. This procedure tests the likelihood that a group of samples come from the same population based on rank comparisons of a variable from each member of the samples. For each cleft loop, we treated the three RNAPs as different samples and the lengths of these regions in each species as the variable being measured. We therefore tested whether each cleft loop was of uniform length in all three eukaryotic RNAPs.

Detecting rate-shift sites and analysis of their spatial distributions

We used our phylogenetic trees (Supplementary Figure 2.2) to examine amino acid sequences of the two large subunits of the RNAPs for sites that have undergone shifts in substitution rates along specific lineages of the phylogeny. These exist as two different types of sites. Type 1 sites are characterized by having different substitution rates along different lineages near the leaves of a phylogenetic tree and type 2 sites are characterized by high ancestral substitution rates along the basal branches of a tree, followed by a decrease and homogenization of substitution rate towards the leaves (Gu, 1999; 2001). Both type 1 (Wang and Gu, 2001; Gu *et al.* 2002) and type 2 (Gribaldo *et al.*, 2003) sites have been used to identify amino acid sites that are likely responsible for functional divergence between homologous proteins, with the assumption that changes

in substitution rates result from changes in functional constraints. We examined the three RNAPs for both type 1 and type 2 sites using the method of Knudsen *et al.* (2003). This method takes as input a multiple sequence alignment and a phylogenetic tree with an *a priori* defined branch that separates the two functionally diverged clades in the phylogenetic tree. The null hypothesis at each site assumes a constant amino acid substitution rate along the tree; the type 1 site alternate hypothesis assumes a different rate for each of two subtrees in the phylogeny; the type 2 site alternate hypothesis allow the extension of a basal branch of one of the subtrees, representing an increased ancestral rate following formation of one of the clades defined by the subtrees. The latter two hypotheses each add one additional parameter to the null hypothesis. The significance of these hypotheses is determined using likelihood ratio tests. In cases where both alternative hypotheses explain the evolution of a site significantly better than the null hypothesis, the hypothesis with the largest likelihood ratio is used to explain the evolution of that site. For each of the two largest subunits of the RNAPs, we performed three comparisons: RNAPI compared to RNAPII and RNAPIII, RNAPII compared to RNAPI and RNAPIII, and RNAPIII compared to RNAPI and RNAPII. These three comparisons were required since only two clades can be compared at a time.

The branch-site codon method of Zhang *et al.* (2005) was used to detect sites evolving by relaxed purifying selection or positive selection along ancestral lineages of the RNAPs. This method distinguishes between two types of branches, the foreground and background branches, both of which have to be defined *a priori*. They differ only in that sites in the foreground branches are able to have an ω parameter greater than one. Thus, foreground branches are capable of having non-synonymous substitution rates

that are higher than the synonymous substitution rates. We used the protein phylogenies of the two largest RNAP subunits for the branch-site method, fixing only the topology for the analyses, and assigned each ancestral branch leading from the central trifurcating node to each of the three RNAP clades as foreground branches (Supplementary Figure 2.2). We applied both branch site test 1 and test 2 of Zhang *et al.* (2005) which test for relaxed purifying and positive selection of sites along the foreground branches, respectively. These tests have the same alternate hypothesis, which assumes $\omega > 1$ at certain sites along the foreground branches, but differ in their null hypotheses. The null hypothesis of test 1 assumes two site classes, $\omega = 1$, and $0 < \omega < 1$; the null hypothesis of test 2 is the same as that of test 1, but with one additional site class that allows foreground branches with $\omega = 1$ and background branches with $0 < \omega < 1$.

We assessed the spatial locations of all rate-shift sites using simulations to test for the probability that these sites are randomly distributed throughout the structures of the two largest subunits of the RNAPs. For the site predictions of both the Knudsen *et al.* (2003) and the Zhang *et al.* (2005) methods, we simulated the distribution of sites that would result if rate-shift sites were randomly distributed throughout the two largest subunits of the RNAPs. For each test, we counted the number of these rate-shift sites that were identified in each subunit and performed 1000 simulations of this number of randomly placed sites in these two subunits of the structure (pdb code 1R9T). We grouped the results of both subunits for a given test because these subunits are tightly associated and cooperate to perform the functions of the RNAPs. In each simulation, we calculated the sum of the nearest neighbor distances for each site and used these

as our test statistic for the null distribution. We then calculated the sum of nearest neighbor distances for each set of predicted rate-shift sites and compared these to the corresponding null distributions at the 0.05 significance level. In total, 7 of these cluster tests were performed: one type 1 and type 2 test for each of the three RNAPs using the amino acids identified by the Knudsen *et al.* (2003) method, and one test using the amino acids identified by the method of Zhang *et al.* (2005).

Acknowledgements

We thank Stéphane Aris-Brosou (Biology Department, University of Ottawa) for his advice and comments. We also thank the anonymous referees for their useful and constructive comments on a previous version of this chapter. This work was supported by a Discovery Grant from the Natural Science and Engineering Research Council of Canada to G.D

Tables

Table 2.1. Cleft loop functions and amino acid locations in the two largest subunits of RNA polymerase II.

Subunit	Cleft Loop	Location (amino acids)	Function	Reference
Largest	Zipper	39-51	Unknown function, but proposed to be involved in maintenance of upstream end of transcription bubble	[1]
	Lid	248-260	Binds and stabilizes upstream portion of DNA-RNA hybrid; participates in formation of RNA exit groove via interaction with the flap loop	[1-3]
	Rudder	304-324	Binds and stabilizes upstream portion of DNA-RNA hybrid; required for transition from closed to open preinitiation complex in archaea	[1-4]
	Switch 2	328-346	Couples presence of nucleic acid in PIC to clamp closure by interactions with bridge helix and DNA template strand	[1-2]
	Bridge	811-845	Alpha helix that traverses the active site just upstream of catalytic Mg ions; interacts with switch regions in TEC and is integral to the transcription reaction	[1-2], [4-9]
	Trigger Loop	1084-1091	Remains disordered, but is integral to transcription reaction in all proposed models of the transcription reaction	[4], [7-9]
	Switch 1	1384-1406	Couples presence of nucleic acid in PIC to clamp closure by interactions with bridge helix and DNA template strand	[1-2]
	Fork Loop 1	466-478	Binds and stabilizes upstream portion of DNA-RNA hybrid	[1-4]
	Fork Loop 2	500-512	Facilitates separation of template and non-template strand just downstream from the bridge helix and catalytic Mg ions	[2], [11], [10]
	Flap Loop	920-933	participates in formation of RNA exit groove via interaction with the flap loop	[1]
Second largest	Switch 3	1108-1129	undergoes disordered-to-ordered transition in TEC; Ordered region likely stabilizes the RNA-DNA hybrid by interacting with its 3' end	[1-2]

Note – The locations are with respect to the largest subunit (GenBank number CAA65619) and second largest subunit (GenBank number CAA99357) of the RNA polymerase II protein sequence from *S. cerevisiae*. Reference numbers refer to the following citations: [1] Cramer *et al.* (2001); [2] Gnatt *et al.* (2001); [3] Westover *et al.* (2004); [4] Naji *et al.* (2008); [5] Epshtein *et al.* (2002); [6] Bar-Nahum *et al.* 2005; [7] Suenaga *et al.* (2006); [8] Brueckner and Cramer (2008); [9] Tan *et al.* (2008); [10] Wang *et al.* (2006); [11] Kettenberger *et al.* (2004).

Table 2.2. Length variations between paralogous cleft loops of eukaryotic RNAPs.

Cleft Loop	Length Variation			p-value
	RNAPI	RNAPII	RNAPIII	
Zipper	11(1), 12(14), 15(1), 21(1)	11(11), 12(2), 13(4)	7(1), 12(2), 13(6), 14(6), 15(1), 16(1)	1×10^{-4}
Lid	7(1), 13(1), 12(14), 16(1)	12(3), 13(12), 14(1), 16(1)	12(3), 13(5), 14(9)	7×10^{-5}
Rudder	9(1), 11(1), 13(1), 14(1), 15(3), 16(2), 18(5), 19(3)	20(1), 21(16)	17(2), 18(1), 19(3), 20(7), 23(2), 24(1), 26(1)	8×10^{-8}
Switch 2	19(17)	19(17)	19(17)	1
Bridge	35(17)	35(17)	35(17)	1
Trigger Loop	6(1), 8(16)	8(17)	8(17)	0.4
Switch 1	22(15), 23(2)	23(16), 24(1)	23(17)	2×10^{-9}
Fork Loop 1	13(2), 14(14), 16(1)	10(1), 12(4), 13(12)	12(17)	2×10^{-9}
Fork Loop 2	13(1), 14(16)	13(17)	13(17)	1×10^{-10}
Flap Loop	0(13), 8(3), 11(1)	4(1), 12(1), 13(1), 14(12), 15(1), 159(1)	2(1), 3(1), 5(1), 10(1), 12(2), 13(1), 14(2), 15(6), 17(1), 23(1)	2×10^{-7}
Switch 3	27(1), 22(16)	22(17)	22(17)	0.4

Notes. – This table lists the number of times (in parentheses) a cleft loop of a given length of amino acids was observed among the 17 species analyzed. The *p*-values represent the significance of the Kruskal-Wallis test of differences in the length distributions of the functional regions of the three RNAPs. All tests have 2 degrees of freedom

Table 2.3. Parameter estimates and model results for fixed-site models.

Model	Parameter / model result	SU1		SU2	
		OCLEFT LOOP ^a	CLEFT LOOP ^b	OCLEFT LOOP ^a	CLEFT LOOP ^b
A (null)	ω^d	0.01533		0.02139	
B	rates ^c	1	0.5867	1	0.7057
	ω^d	0.01483		0.02135	
	<i>p</i> -value ^e	<i>p</i> < 0.001		<i>p</i> < 0.001	
D	rates	1	0.76596	1	0.95016
	ω^d	0.0159	0.01127	0.02167	0.01582
	<i>p</i> -value	<i>p</i> < 0.001		<i>p</i> < 0.01	

Notes. a) OCLEFT LOOP, partition consisting of amino acids found outside cleft loops; b) CLEFT LOOP, partition consisting of amino acids found in cleft loops; c) rates, relative rates of the two partitions; d) ω , nonsynonymous/synonymous ratio; e) *p*-values refer to the statistical significance of a given model relative to the model above it.

Table 2.4. Probabilities that the type 1 and type 2 sites identified by the Knudsen *et al.* (ref. 45) method are randomly distributed.

	Type 1 sites	Type 2 sites
RNAPI versus RNAPII & RNAPIII	0.209	<0.001
RNAPII versus RNAPI & RNAPIII	0.151	0.047
RNAPIII versus RNAPI & RNAPII	0.227	<0.001

Figure legends

Figure 2.1. Sequence features of the two largest subunits of the eukaryotic RNAPs. The top and bottom figures represent the primary sequences of the largest and second largest subunits, respectively. Alphabetically labeled black boxes represent previously identified homology regions that are present in both eukaryotic and bacterial RNAPs. Grey boxes extend these regions and represent homologous regions present in the three eukaryotic RNAPs (see text for inclusion criteria). Colored regions represent poorly aligned regions. The locations of the cleft loops relative to these homology regions are indicated below and above the primary structures of the largest and second largest subunits, respectively. The middle figure is a representative RNAP structure of a ternary elongation complex (pdb code 1R9T). In this structure, cleft loops are drawn in yellow, the template DNA strand is drawn in light blue, RNA is drawn in dark green, dark blue regions are the largest subunit, red regions are the second largest subunit and gray regions are the other subunits. The numbers in parentheses (black or white) are the eight cleft loops visible in this view and correspond to the numbers shown in the primary sequence representations. Cleft loops that are not numbered denote regions that are either not present in the structure or are not visible due to the viewing angle.

Figure 2.2. Mean normalized relative substitution rate of cleft loops indicating, for each region and paralog, the average residue's deviation from the mean substitution rate of the whole protein.

Figure 3.3. Distribution of type 1 and type 2 sites in the eukaryotic RNAPs determined using the method of Knudsen *et al.* (2003). This structure (pdb code 1R9T) is from Westover *et al.* (2004). Top, distribution of type 1 sites; bottom, distribution of type 2 sites. RNAPs are shown the following order: RNAPI (left), RNAPII (middle), and RNAPIII (right). The view is from the front of the enzyme along the DNA axis, through the main enzyme channel and into the active site. The template strand can be seen exiting along the back of the enzyme at a roughly a 90° angle with the incoming double-stranded DNA. Dark gray regions are the largest subunit, light gray regions are the second largest subunit, blue regions are DNA and rate-shift sites are indicated in red.

Figure 2.2. Distribution of sites that have evolved via relaxed selection along the ancestral branch in all three polymerases according to the branch-site method of Zhang *et al.* (2005). The view and color coding are identical to those of Figure 3.

Figures

Figure 2.1.

Figure 2.2

Figure 2.3

Figure 2.4.

Chapter 3

The Evolutionary Rates of Eukaryotic RNA Polymerases and of Their Transcription Factors Are Affected by the Level of Concerted Evolution of the Genes They Transcribe

Robert Carter and Guy Drouin

Département de biologie et Centre de recherche avancée en génomique
environnementale, Université d'Ottawa, Ottawa, Ontario, Canada, K1N 6N5

Type: Research Article

Running title: Evolutionary Rates of Eukaryotic RNA Polymerases

Keywords: RNA polymerase, concerted evolution, transcription factors, eukaryotic

Correspondance to: Guy Drouin, Département de biologie, Université d'Ottawa, 30

Marie Curie, Ottawa, Ontario, Canada, K1N 6N5. Tel.: (613) 562-5800 ext. 6052, FAX:

(613) 562-5486, E-mail: gdrouin@science.uottawa.ca

(Published in *Mol Biol Evol.* 2009. Jul 24. PMID: 19633229. Epub ahead of print)

Author Contributions: RC and GD designed the study, RC performed all analyses, and RC and GD wrote the paper.

Abstract

A defining characteristic of all eukaryotes is the presence of three RNA polymerases, each of which transcribes a particular subset of nuclear genes. RNA polymerase I transcribes rRNA genes, RNA polymerase II transcribes mRNA, miRNA, snRNA, and snoRNA genes, and RNA polymerase III transcribes 5S rRNA and tRNA genes. Here we use the sequences of up to 25 Ascomycete species to show that the type of genes transcribed by each RNA polymerase affects their evolutionary rates and those of their transcription factors. The RNA polymerase subunits and transcription factors of genes whose promoters experience higher levels of concerted evolution evolve significantly faster than those experiencing lower levels of concerted evolution. The rate of evolution of RNA polymerase genes and their transcription factors are therefore not only the result of diverse selective constraints but are also influenced by the level of concerted evolution of the genes they transcribe.

Introduction

Amino acid substitution rates in proteins vary by several orders of magnitude and much of the underlying cause for this variability has until recently been attributed to different levels of functional constraints (e.g., Graur and Li 2000). However, more recent studies have shown that the most significant selective constraint experienced by yeast genes is selection for translational robustness where highly expressed proteins evolve more slowly than less expressed proteins (Pál *et al.* 2001; Drummond *et al.* 2005, 2006; McInerney 2006). This suggests that the rate of yeast protein evolution is mostly the result of purifying selection pressure to minimize the misfolding of proteins and that this effect is stronger for more abundant proteins due to their higher number of translation events. In fact, the recent study of Drummond and Wilke (2008) demonstrated that selection against the toxicity of misfolded proteins generated by ribosome errors is sufficient to explain this effect. Other factors, such as protein structure, local mutation rates, gene length, dispensability and the number of protein-protein interactions may also influence the evolutionary rates of proteins (Graur and Li 2000; Drummond *et al.* 2005; Bloom *et al.* 2006; McInerney 2006; Zhou *et al.* 2008).

Here we test another hypothesis, called the molecular coevolution hypothesis, which predicts that the evolutionary rate of different protein-coding genes will be positively correlated with the amount of homogenization (concerted evolution) experienced by the promoters of the genes they interact with (Dover and Flavell 1984). These authors based this hypothesis on the observation that the strict species specificity of ribosomal RNA genes for the hosts' transcriptional machinery was due to

an incompatibility between transcription factors and ribosomal RNA gene promoters of different species (such as between human and mouse). They suggested that this incompatibility was the result of the rapid turnover of rRNA gene promoters by mechanisms such as unequal crossing over and gene conversion, which lead to the rapid evolution of rRNA transcription factors; hence the species-specificity of these transcription factors (Dover and Flavell 1984). In contrast, human mRNA genes can be transcribed by the RNA pol II of species as distantly related as yeasts because the promoters of mRNA genes, since they are not subject to DNA turnover mechanisms, do not evolve quickly. This hypothesis therefore predicts that as the level of concerted evolution experienced by promoters increases, so too will the evolutionary rates of the proteins interacting either directly or indirectly with these promoters. In other words, the faster evolution of the promoters of genes being homogenized by unequal crossingover and/or gene conversion is expected not only to select for alleles of the transcription factor genes better able to interact with the new promoter variants being homogenized but to also select for alleles of polymerase subunits which bind these transcription factors (figure 3.1).

In the case of RNA polymerase genes, the molecular coevolution hypothesis predicts that the genes coding for RNA polymerase I (RNAPI), and those coding for its transcription factors, should evolve faster than those coding for RNA polymerase III (RNAPIII) which, in turn, should evolve faster than those coding for RNA polymerase II (RNAPII). These predictions are based on the observation that the promoters, like the coding regions, of the ~140 tandemly repeated yeast genes coding for the 18S, 5.8S and 28S ribosomal are almost all identical (Szoztak and Wu 1980; Ganley and

Kobayashi, 2007). This high sequence identity is the result of the high degree of DNA turnover by mechanisms such as unequal crossingovers and gene conversions which quickly homogenize this tandemly repeated gene family (Szostak and Wu 1980; Dover 1982; Arnheim 1983; Schlötterer and Tautz 1994). The homogenization of the promoters of the ribosomal genes transcribed by RNA polymerase I is not detrimental because all these genes have the same function. However, this high degree of turnover does mean that these promoter sequences evolve faster than those not affected by concerted evolution (figure 3.1a). The internal promoters of the 5S rRNA and tRNA genes transcribed by RNA polymerase III also evolve under the effect of homogenizing forces but this homogenization is not as quick as those of rRNA genes because these genes are often found dispersed throughout the genome and, consequently, are more likely to be homogenized by gene conversion events rather than unequal crossover (figure 3.1b; Morzycka-Wroblewska *et al.* 1985; Schlötterer and Tautz 1994; Graur and Li 2000; Marck *et al.* 2006). For example, whereas the frequency of crossover between the rRNA gene units of *Saccharomyces cerevisiae* has been calculated to be 10^{-2} per generation, the frequency of gene conversion between the dispersed serine tRNA of *Schizosaccharomyces pombe* is only of 10^{-5} per progeny spore (Szostak and Wu 1980; Amstutz *et al.* 1985). Note that recent work on the evolution of 5S ribosomal genes in filamentous fungi has shown that the concerted evolution of 5S genes in some of these species is due to birth-and-death evolution under strong purifying selection and with different genera having different rates of birth and death (Rooney and Ward 2005). Although these birth-and-death rates were not measured, their effect on the rate of concerted evolution of these dispersed 5S genes is likely lower than the fast-evolving

clustered and tandemly repeated genes transcribed by RNA polymerase I (figure 3.1a). Finally, the promoters of the diverse mRNA-coding genes transcribed by RNA polymerase II are not expected to evolve under the effect of homogenizing forces because RNA polymerase II transcribes thousands of different genes and homogenizing their diverse promoters would be imminently detrimental (figure 3.1c). Therefore, the molecular coevolution hypothesis predicts that the transcription factors that bind to fast evolving RNA polymerase I promoters, and the proteins that bind these transcription factors, should evolve faster than those of RNA polymerase III which, in turn should evolve faster than those of RNA polymerase II. These predictions are the opposite of those made by the gene expression hypothesis. If one assumes (but, as we show below, this assumption is not supported by the available gene expression data) that the gene expression levels of the proteins involved in transcribing the different types of RNAs are directly related to transcript abundance levels (i.e., more proteins are needed to make up more RNA molecules), then, given that RNA polymerases I, III and II transcripts respectively make up 80%, 15% and 5% of the RNAs in exponentially growing yeast cells (Warner 1999), the gene expression hypothesis would predict that the evolutionary rate of the proteins making up RNA polymerase I and its transcription factors should evolve slower than those of RNA polymerase III which, in turn should evolve slower than those of RNA polymerase II. The fact that we observe the opposite evolutionary rates supports the molecular coevolution hypothesis and demonstrates that the evolutionary rate of RNA polymerase genes and of their transcription factors genes is positively correlated with the amount of homogenization experienced by the genes they transcribe. The fact that the DNA turnover processes involved in the

homogenization of gene families have a significant effect on the evolutionary rate of proteins also demonstrates that non-adaptive processes can affect the evolutionary rates of proteins.

Materials and Methods

Species, sequences and alignments

We used the sequences of 25 Ascomycete species for which complete genome sequences were available when this study was initiated and for which phylogenetic relationships are well established (Fitzpatrick *et al.* 2006). These 25 species are listed in supplementary table 3.1 (Supplementary Material online). We only used Ascomycete sequences in order to minimize the problem of substitution saturation when comparing sequences from different fungal phyla (results not shown) and because the gene expression hypothesis was originally derived using yeast data (Pál *et al.* 2001; Drummond *et al.* 2005, 2006; Bloom *et al.* 2006).

In order to retrieve all sequences from the three eukaryotic transcriptional machineries (listed in Table 3.1), we first retrieved them from the *Saccharomyces cerevisiae* genome (available at NCBI) because they are well annotated and are associated with experimental data. For each of these genes, we then searched the NR protein database for Ascomycete orthologs using the reciprocal best hit strategy. According to this strategy, two proteins are considered orthologs if each is the others top hit in reciprocal BLAST searches (Altschul *et al.* 1997). We defined orthologs as reciprocal best hits with a minimum of 25% sequence identity and 75% sequence length with the query sequence, as well as a reciprocal E-value less than 1×10^{-5} . This strategy allowed us to retrieve about 40% of the sequences. We then used PSI-BLAST searches to retrieve more sequences (Altschul *et al.* 1997). We first used the protein

sequences retrieved using the reciprocal best hit strategy to build PSI-BLAST profiles which were then searched against the Ascomycete protein database using BLASTP. For sequences that were not identified by these methods, we used a combination of TBLASTN and synteny analysis to identify the remaining sequences from Ascomycete whole-genome shotgun sequences, RefSeq genomes, and the RefSeq nucleotide collection, using orthologs from closely related species as queries. For each sequence obtained from genomic data using TBLASTN, the start codon, stop codon and intron boundaries were identified by alignment with a close ortholog using GeneWise2 (Birney *et al.* 2004).

Since the closest BLAST hit is often not the nearest neighbor, we used alignments and phylogenetic analyses to ensure that all automatically retrieved sequences were indeed orthologous and functional (Koski and Golding 2001). Alignments of protein sequences were performed using the iterative refinement with weighted sum-of-pairs and consistency scores (L-INS-i) mode of MAFFT (Katoh *et al.* 2002). Protein phylogenies for each of the proteins were constructed using PhyML (Guindon and Gascuel 2003), WAG substitution matrices (Whelan and Goldman 2001), allowing a proportion of sites to be invariant and allowing rates to vary across sites with 4 gamma distributed rate categories. The shape parameter, topology, branch lengths, and the proportion of invariant sites were estimated from the data. We examined the resulting topologies and branch lengths of each protein for consistency with the Ascomycete phylogeny of Fitzpatrick *et al.* (2006) and the taxonomy classification scheme of NCBI. Protein sequences that deviated significantly from the accepted topology or had branch lengths several times longer than expected were removed. The

correct sequences were re-acquired manually using PSI-BLAST and TBLASTN and the PhyML phylogeny was reconstructed. This process was re-iterated until all phylogenies had similar topologies with no abnormally long branches, thus ensuring that all retrieved sequences were orthologous and functional.

In almost all cases, we found the complete set of the 7 RNA polymerase paralogous subunit sequences listed in Table 3.1 in all 25 Ascomycete species (supplemental table 3.1; Supplementary Material Online). The only exceptions are the RPA4 and RPA9 sequences which were retrieved for only 12 and 22 species, respectively. We also found each transcription factor (TF) and non-paralogous subunit sequence in at least 24 of the 25 Ascomycete species (Supplemental Table 3.1) with 20 species having a full complement of TFs and non-paralogous subunits. We did not include subunit TFIIC60 sequences in our analyses because they had unusually high mean number of nonsynonymous substitutions per nonsynonymous site when compared with other RNA polymerase III TFs (see below; results not shown). The nucleotide coding sequences were aligned to the protein alignments using PAL2NAL (Suyama *et al.* 2006).

Evolutionary rates, statistical analyses and gene expression

We used the maximum likelihood method implemented in the codeml program of the PAML package to measure the number of nonsynonymous substitutions per nonsynonymous site (dN) between all pairs of sequences in each DNA alignment (using the options seqtype = 1, runmode= -2 and CodonFreq = 2 in the codeml.ctl files; Yang

2007). Although we used the default ω value of 0.4 as the starting value, using an ω value of 1.4 did not significantly affect our results (the dN values obtained were at most 1.6% different; results not shown).

We analyzed two sequence categories, paralogous sequences and non-paralogous sequences. The first type of analyses included RNA polymerase subunits RPB1, RPB2, RPB3, RPB4, RPB7, RPB9, and RPB11 sequences and the paralogs of each of these proteins in RNAPI and RNAPIII (table 3.1). We grouped these paralogs to control for the effect that protein structures have on evolutionary rates (Bloom *et al.* 2006). Since paralogs tend to have both conserved structures and similar functions, these effects will be minimal when comparing their evolutionary rates (Chothia and Lesk 1987; Cramer *et al.* 2008). In other words, by comparing the mean pairwise nonsynonymous evolutionary rates between paralogs we are largely controlling for both structural effects and functional constraints on the evolutionary rates.

For paralogous RNA polymerase sequences, we used *t*-tests to test whether the nonsynonymous rates of RNAPI paralogs are larger than those of RNAPIII paralogs, and whether the nonsynonymous rates of RNAPIII paralogs are larger than those of RNAPII paralogs. These two tests were performed for each set of paralogs in the three RNA polymerases except RP3 and RP11. For the RP3 and RP11 subunits, there are only 2 paralogs: one for RNAPII and one that is shared between RNAPI and RNAPIII (table 3.1). For these subunits, we tested whether the mean synonymous rate of the RPAC subunit is greater than the mean nonsynonymous rate of its RNAPII paralog. We used one-sided *t*-tests to assess whether mean nonsynonymous values were

statistically different. Mean nonsynonymous values and standard errors were calculated from all $(n \times (n-1))/2$ pairwise sequence comparisons within each group using R version 2.4.1 (R Development core team 2006). For each analysis, the same set of species was used in each *t*-test to control for divergence times. This ensures that the same genes are compared in different species and allows comparing evolutionary rates without having to know the divergence times of the species being compared.

The non-paralogous sequences category included all TFs and non-paralogous RNA polymerase subunits (table 3.1). We used Kruskal-Wallis tests to test whether the nonsynonymous rates of TFs and subunits specific to RNA polymerase I are larger than those of RNA polymerase III and whether the nonsynonymous rates of TFs and subunits specific to RNA polymerase III are larger than those of RNA polymerase II. The sample sizes for this second category were 10 RNA polymerase I proteins, 13 RNA polymerase II proteins, and 13 RNA polymerase III proteins.

Gene expression levels were obtained from Holstege *et al.* (1998) and are reported in mRNA molecules/cell.

Results

The list of the 25 yeast species from which we retrieved gene sequences is shown in supplementary table 3.1 (Supplementary Material online). The genes we used belong to the following three groups: RNA polymerase paralogs, i.e., the homologous sequences specific to each RNA polymerase (such as the largest subunit of each polymerase), RNA polymerase non-paralogs, i.e., the non-homologous sequences specific to given polymerases (such as the RPC31 subunit sequence specific to RNA polymerase III) and transcription factors (TFs) specific to each RNA polymerase (table 3.1). We grouped these sequences into two different categories which we call the paralogous and non-paralogous data set. The former is composed uniquely of RNA polymerase paralogs whereas the later is composed of the RNA polymerase non-paralogs and the TFs. The GenBank Accession numbers of the sequences we used are shown in supplementary table 3.1 (Supplementary Material online).

The evolutionary rates of paralogous subunits of the three RNA polymerases and the *t*-tests performed to determine whether the RNAPI subunits evolve faster than their RNAPIII paralogs, and whether the RNAPIII paralogs evolve faster than their RNAPII paralogs, are shown in table 3.2. Although 4 out of 12 *p*-values are not significant, most comparisons show RNAPI paralogs evolve significantly faster than RNAPIII paralogs and that RNAPIII paralogs evolve significantly faster than RNAPII paralogs. Two of the non-significant values (for RPAC40 > RPB3 and RPAC19 > RPB11), where RPAC values are actually smaller than RNAPII, are rate comparisons between a subunit shared between RNA polymerases I and III and a subunit specific to RNA polymerase

II. The lower evolutionary rates of these RPAC subunits may simply be the result of the higher evolutionary constraints they experience because they are shared between RNA polymerases I and III. The other two non-significant values are for subunit 9, one of the smallest subunits, and may reflect stochastic variation. Overall, after controlling for highly similar structures and functional constraints, the nonsynonymous substitution rate of paralogous RNA polymerase subunits is therefore positively correlated with the extent of concerted evolution experienced by the promoters they bind to. This correlation is not due to differences in expression levels because the mean expression levels of the paralogous subunits of the three RNA polymerases are not statistically different (ANOVA, $p = 0.87$, 14 total df). The mean expression level of the paralogous RNA subunits of RNA polymerase I (i.e., that of the RPA1, RPA2, RPA4, RPA7 and RPA9 subunits) is 2.84 ± 0.70 mRNA molecules/cell, that of the corresponding RNA polymerase II subunits is 2.90 ± 0.43 mRNA molecules/cell and that of the corresponding RNA polymerase III subunits is 2.46 ± 0.73 mRNA molecules/cell.

The mean pairwise nonsynonymous values for the non-paralogous sequences (i.e., the RNA polymerase-specific subunits and TFs) are shown in Table 3.3. Although the average nonsynonymous evolutionary rates of RNAPI (0.89 ± 0.08) and RNAPIII (0.80 ± 0.07) genes are not statistically different (Kruskal-Wallis test, p -value = 0.2148, df = 1) those of RNAPI and RNAPIII are both larger than that of RNAPII (0.47 ± 0.04 ; Kruskal-Wallis test, p -values of 3.5×10^{-4} and 8.5×10^{-5} , respectively, with 1 degree of freedom for each test). Thus, RNAPI and RNAPIII TFs and RNA polymerase-specific subunits have similar mean pairwise nonsynonymous values but these values are significantly larger than those of RNAPII. This observation is not due to differences in

expression levels because the mean expression levels of the non-paralogous proteins of the three RNA polymerases are not statistically different (ANOVA, $p = 0.33$). The mean expression level of the non-paralogous sequences of RNA polymerase I (table 3.1) is 1.70 ± 0.59 mRNA molecules/cell, that of RNA polymerase II sequences is 1.42 ± 0.24 mRNA molecules/cell and that of RNA polymerase III sequences is 0.99 ± 0.22 mRNA molecules/cell.

Discussion

Our results support the hypothesis that the amount of homogenization (concerted evolution) experienced by promoters affects the evolutionary rates of RNA polymerases and those of their transcription factors. All significant differences in nonsynonymous evolutionary rate observed between the paralogs of the three yeast RNA polymerase genes all show the paralogs of RNA polymerase I evolving faster than those of corresponding RNA polymerase III paralogs, which, in turn, evolve faster than those of RNA polymerase II. Although the average nonsynonymous evolutionary rates of the RNA polymerase I and III genes coding for non-paralogous subunits and transcription factors are not statistically different from one another, they are both larger than those of RNA polymerase II. The fact that the results based on non-homologous subunits and transcription factors are not as clear-cut as those based on paralogous subunits is not surprising because they are based on the rate of evolution of unrelated (non-paralogous) genes whereas those based on paralogous genes are based on homologous genes having very similar structures and functions in all three RNA polymerases. The results based on non-paralogous subunits and transcription factors nevertheless show that they evolve faster when involved in transcribing genes whose promoters experience concerted evolution (non-paralogous subunits and TFs of RNA polymerases I and III) than when involved in transcribing genes not subject to concerted evolution (non-paralogous subunits and TFs of RNA polymerase II). Interestingly, the faster rate of evolution of some RNA polymerase genes cannot be explained by the gene expression hypothesis because both the paralogous subunits and non-paralogous

proteins of all three RNA polymerases are expressed at the same level ($p = 0.87$ and 0.33 , respectively).

In conclusion, our results clearly show that the evolutionary rates of eukaryotic RNA polymerases and of their transcription factors is affected by the level of concerted evolution of the genes they transcribe. Thus the homogenization of gene families caused by mechanisms such as unequal crossovers and gene conversions can affect the evolutionary rate of the genes involved in their transcription (Dover 1982; Dover and Flavell 1984). The DNA turnover processes involved in the homogenization of gene families are therefore another example of the emerging evolutionary “world-view that gives much more prominence to non-adaptive processes” (Koonin 2009; also see Lynch 2007a, 2007b) because the effect of these processes on the evolutionary rates of RNA polymerase genes and their transcription factors is clearly non-adaptive.

Supplementary Material

Supplementary table 3.1 is available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgements

We thank Stéphane Aris-Brosou (Biology Department, University of Ottawa) for his advice and comments. We also thank the anonymous referees for their constructive comments on a previous version of this chapter. This work was supported by a Discovery Grant from the National Science and Engineering Research Council of Canada to G. D.

Tables

Table 3.1. Proteins of the three eukaryotic transcriptional machineries analyzed in this study. Proteins are organized by analysis category subunit relationship group and protein complex membership.

Category	Group	Complex	Subunits							
1	RNA polymerase paralogs	RNAPI	RPA1	RPA2	RPAC40	RPA4	RPA7	RPA9	RPAC19	
		RNAPII	RPB1	RPB2	RPB3	RPB4	RPB7	RPB9	RPB11	
		RNAPIII	RPC1	RPC2	RPAC40	RPC4	RPC7	RPC9	RPAC19	
2	RNA polymerase non-paralogs	RNAPI	RPA34.5	RPA49						
		RNAPII								
		RNAPIII	RPC31	RPC34	RPC37	RPC53	RPC82			
	A-CF	RRN3	RRN6	RRN7	RRN11					
	A-UAF	RRN5	RRN9	RRN10	UAF30					
	B-TFIIA	TFIIAab	TFIIAg							
	B-TFIIB	TFIIB								
	B-TFIIIE	TFIIIEa	TFIIIEb							
	B-TFIIIF	TFIIIFa	TFIIIFb	TFIIIFc						
	B-TFIIH	TFB1	TFB2	TFB4	RAD3	SSL1				
	B-TFIIS	TFIIS								
	C-TFIIIA	TFIIla								
	C-TFIIIB	TFIIIB-BRF	TFIIIB-BDP							
	C-TFIIIC	TFIIIC55	TFIIIC91	TFIIIC95	TFIIIC131	TFIIIC138				

Table 3.2. Mean nonsynonymous substitution rate ($\pm$ standard error) of group 1 paralogous RNA polymerase subunits and comparisons between RNA polymerase paralogs.

Paralog	RNAPI subunit and rate	RNAPIII subunit and rate	P-value of RNAPI > RNAPIII	RPAC subunit and rate	RNAPII subunit and rate	P-value of RNAPIII > RNAPII or RPAC > RNAPII
1	RPA1 0.390 $\pm$ 0.008	RPC1 0.280 $\pm$ 0.005	<2.20 x 10 ⁻¹⁶ (25, 598 df)	-	RPB1 0.242 $\pm$ 0.004	5.39 x 10 ⁻⁶ (25, 598 df)
2	RPA2 0.240 $\pm$ 0.004	RPC2 0.225 $\pm$ 0.004	0.02 (25, 598 df)	-	RPB2 0.152 $\pm$ 0.002	<2.20 x 10 ⁻¹⁶ (25, 598 df)
3	-	-	-	RPAC40 0.315 $\pm$ 0.004	RPB3 0.326 $\pm$ 0.004	0.87 (25, 598 df)
4	RPA4 0.650 $\pm$ 0.033	RPC4 0.544 $\pm$ 0.026	0.01 (11, 108 df)	-	RPB4 0.365 $\pm$ 0.006	<2.20 x 10 ⁻¹⁶ (25, 598 df)
7	RPA7 0.642 $\pm$ 0.011	RPC7 0.524 $\pm$ 0.010	6.73 x 10 ⁻⁹ (25, 598 df)	-	RPB7 0.367 $\pm$ 0.007	<2.20 x 10 ⁻¹⁶ (25, 598 df)
9	RPA9 0.401 $\pm$ 0.011	RPC9 0.394 $\pm$ 0.011	0.34 (22, 460 df)	-	RPB9 0.413 $\pm$ 0.012	0.29 (25, 598 df)
11	-	-	-	RPAC19 0.295 $\pm$ 0.006	RPB11 0.463 $\pm$ 0.012	1 (25, 598 df)

Notes. P-values are based on one-sided *t*-tests and numbers in parentheses are the number of genes used for each comparison. The number of genes used for each comparison varies because each test can only be performed when sequences are available for both subunits being compared.

Table 3.3. Nonsynonymous evolutionary rates of TFs and non-paralogous subunits.

Subunit or TF	Mean $\pm$ SE
RPA34	0.978 $\pm$ 0.029
RPA49	0.713 $\pm$ 0.026
RPC53	0.864 $\pm$ 0.026
RPC31	0.773 $\pm$ 0.025
RPC82	0.924 $\pm$ 0.032
RPC34	0.602 $\pm$ 0.019
RPC37	0.680 $\pm$ 0.023
RRN3	0.611 $\pm$ 0.023
RRN11	1.022 $\pm$ 0.037
RRN7	1.031 $\pm$ 0.033
RRN6	0.901 $\pm$ 0.026
RRN5	1.400 $\pm$ 0.054
RRN9	1.048 $\pm$ 0.034
RRN10	0.724 $\pm$ 0.024
UAF30	0.479 $\pm$ 0.017
TFIIAg	0.286 $\pm$ 0.010
TFIIAab	0.474 $\pm$ 0.016
TFIIB	0.443 $\pm$ 0.016
TFIIE α	0.594 $\pm$ 0.021
TFIIE β	0.583 $\pm$ 0.019
TFIIF β	0.593 $\pm$ 0.020
TFIIF α	0.628 $\pm$ 0.021
TFIIF γ	0.387 $\pm$ 0.012
SSL1	0.346 $\pm$ 0.010
TFB1	0.671 $\pm$ 0.022
RAD3	0.180 $\pm$ 0.006
TFB2	0.394 $\pm$ 0.013
TFB4	0.514 $\pm$ 0.017
TFIIS	0.518 $\pm$ 0.018
TFIIIA	0.657 $\pm$ 0.023
TFIIB-BRF	0.535 $\pm$ 0.019
TFIIB-BDP	0.852 $\pm$ 0.025
TFIIIC138	1.561 $\pm$ 0.110
TFIIIC55	0.599 $\pm$ 0.019
TFIIIC95	0.745 $\pm$ 0.022
TFIIIC131	0.701 $\pm$ 0.021
TFIIIC91	0.947 $\pm$ 0.027

Figure legends

Figure 3.1. Schematic illustration of the molecular coevolution hypothesis. (A) Concerted evolution by unequal crossing over of homologous tandemly repeated genes (white rectangles) on homologous chromosomes causes rapid homogenization of promoter mutants (symbols upstream of the genes) and is responsible for the rapid evolution of coevolving transcription factors (light-grey spheres) and RNA polymerase subunits (dark grey spheres). (B) Concerted evolution by gene conversion of dispersed homologous genes has similar effects on interacting proteins as unequal crossing over, but occurs at a slower rate. (C) The promoters of non-homologous genes (rectangles of different shades) are not homogenized. The proteins that bind them therefore evolve more slowly. The number of arrows (generations) is proportional to the amount of time.

Figures

Figure 3.1.

Chapter 4

The increase in the number of subunits in eukaryotic RNA polymerase III relative to RNA polymerase II is due to the early recruitment of general transcription factors.

Robert Carter and Guy Drouin

Département de biologie et Centre de recherche avancée en génomique
environnementale, Université d'Ottawa, Ottawa, Ontario, Canada, K1N 6N5

Type: Research Article

Running title: Increasing number of subunits in eukaryotic RNA polymerases

Keywords: RNA polymerase, subunits, transcription factors, eukaryotic

Correspondance to: Guy Drouin, Département de biologie, Université d'Ottawa, 30
Marie Curie, Ottawa, Ontario, Canada, K1N 6N5. Tel.: (613) 562-5800 ext. 6052,
FAX: (613) 562-5486, E-mail: gdrouin@science.uottawa.ca

(Submitted to *Molecular Biology and Evolution*)

Author Contributions: RC and GD designed the study, RC performed all analyses,
and RC and GD wrote the paper.

Abstract

The number of subunits of RNA polymerases increases during evolution from 5 in eubacteria to 12 in archaebacteria. In eukaryotes, which have three RNA polymerases, the number of subunits expanded from 12 in RNA polymerase II to 14 in RNA polymerase I and to 17 in RNA polymerase III. It was recently demonstrated that the increase in subunit complexity of RNA polymerase I relative to RNA polymerase II was due to the permanent recruitment of general transcription factors (GTFs), which associate with polymerases to direct much of transcription. Here, we extend this finding by demonstrating that 4 of the 5 RNA polymerase III-specific subunits were acquired in the same fashion, by the permanent recruitment of GTFs. We also use the available evidence to propose an evolutionary history of the cellular RNA polymerases in the three domains of life.

Introduction

RNA polymerases are responsible for all cellular transcription. In prokaryotes, a single RNAP handles the full transcription burden of the cell, whereas transcription in eukaryotes is partitioned between the three polymerases that are common to all eukaryotes. RNA polymerase I (RNAPI) transcribes ribosomal RNA (rRNA), RNA polymerase II (RNAPII) transcribes protein-coding genes and some microRNAs, and RNA polymerase III (RNAPIII) transcribes several small RNAs including 5S rRNA and tRNA.

All cellular RNAPs are multi-subunit protein complexes that share a core set of homologous subunits across the three domains of life. The eubacterial RNAP is the simplest of the cellular RNAPs. It consists of the active-site forming β' and β subunits, an α homodimer, and a ω subunit (Zhang *et al.* 1999). In archaea, the core bacterial polymerase homologs are retained, but additional subunits have been recruited. It is composed of the following 12 subunits, named according to the scheme used by Korkhin *et al.* (2009): Rpo1 (homologous to β'), Rpo2 (homologous to β), Rpo3 (homologous to α), Rpo4 and Rpo7 (homologous to the RPB4-RPB7 dissociable heterodimer), Rpo5, Rpo6 (homologous to ω), Rpo8 (homologous to Rpb8), Rpo10, Rpo11 (homologous to α), Rpo12, and Rpo13 (Hirata, Klein, and Murakami 2008; Kusser *et al.* 2008, Korkhin *et al.* 2009). In many Archaea, however, either or both of the two largest subunits are split into two different genes (Rpo1N and Rpo1C, and Rpo2N and Rpo2C) that together encode the counterpart of the full eukaryotic subunit. Furthermore, Rpo8 is not necessary for *in vitro*

transcription by archaeal RNA polymerases and is only present in Crenarchaeaota (Werner and Weinzierl, 2002; Koonin, Makarova, and Elkins 2007; Hirata, Klein, and Murakami 2008; Kwapisz, Beckouët, and Thuriaux 2008). The eukaryotic RNAPs are a further elaboration of the prokaryotic RNAPs. Eukaryotic RNAPs have homologs of all 12 archaeal RNAP subunits except Rpo13 (Korkhin *et al.* 2009) and each of the eukaryotic RNAPs additionally contain a homolog of the archaeal GTF TFS/M, except that these homologs are permanent subunits in the eukaryotic RNAPs. RNAPII is thus composed of 12 subunits, 11 that are homologous to archaeal polymerase subunits (RP1 through RP8 and RP10 through RP12) and one (RP9) that is homologous to an archaeal GTF (TFS/M). Each of the 12 subunits of RNAPII is either shared between the three RNAPs or is encoded by a paralog in one or both of the other RNAPs. Eukaryotic polymerase-specific subunits are unique to RNAPI and RNAPIII, which are composed of 14 and 17 subunits, respectively (Proshkina *et al.* 2006; Jasiak *et al.* 2006; Kuhn *et al.* 2007).

Each eukaryotic RNA polymerase has a large number of general transcription factors (GTFs). The GTFs are a group of proteins and multi-protein complexes that are required for directed transcription *in vivo*. Each polymerase has its own set of GTFs, although the functions they perform are analogous. In essence, they direct the polymerases to the promoter, help form the transcription bubble, and allow the polymerases to initiate transcription. Unlike the majority of the polymerase subunits, most GTFs are not homologous across the three eukaryotic RNA polymerases, implying that they were recruited early in eukaryotic evolution following the formation of their associated polymerases. Based on the extent of protein homology, RNAPII

appears to be the ortholog of the prokaryotic transcriptional machinery since it is the only eukaryotic RNAP that has homologous counterparts to all three archaeal GTFs, namely TBP, TFB, and TFE, which are homologous to TBP, TFIIB, and TFIIE α in RNAPII, respectively.

The polymerase-specific subunits of RNAPI and RNAPIII give the RNAPs much of their specificity, fine-tuning the transcription needs of their target genes. The two RNAPI-specific subunits RPA34.5 and RPA49 of RNAPI form a heterodimer that provides elongation stimulatory activity (Kuhn *et al.* 2007). In RNAPIII, the RPC37-RPC53 heterodimer is involved in transcription termination and reinitiation (Landrieux *et al.* 2006), whereas the RPC31-RPC34-RPC82 heterotrimer is involved in pre-initiation complex formation and transcription initiation (Brun, Sentenac, and Werner 1997).

Although functions have been attributed to the polymerase-specific subunits of eukaryotic RNAPI and RNAPIII, their origins in early eukaryotes remains a mystery and there exists little structural data describing these proteins. Recently, Cramer and coworkers used HHpred, a Hidden Markov model-based remote homolog detection method (Söding, Biegert, and Lupas 2005), to find homologs of the RNAPI-specific subunits RPA34.5 and RPA49 in sequence profile databases (Kuhn *et al.* 2007). They identified regions of RPA49 and RPA34.5 that were homologous to the N-termini of human TFIIF α and TFIIF β , respectively, which interact to form a novel triple-barrel structure (Gaiser, Tan, and Richmond 2000). They thus demonstrated that the two RNAPI-specific subunits were GTFs that have become permanently associated with RNAPI.

Here, we extend the analysis of Kuhn *et al.* (2007) by using HHpred to identify homologs of the five RNAPIII-specific subunits of *S. cerevisiae* in an effort to determine whether these subunits are also permanently recruited GTFs. Although no homologs of RPC31 were found, we identified regions in RPC82 and RPC34 that are homologous to portions of RNAPII's general transcription factor TFIIE α and β subunits, respectively. Additionally, RPC34 contains a region that is homologous to the sigmaB family of bacterial GTFs. We also identified regions in RPC37 and RPC53 that are homologous to regions in TFIIIF α and β subunits, respectively. The regions of homology between RPC37 and TFIIIF α and RPC53 and TFIIIF β correspond to the heterodimer binding interface between the TFIIIF α and β subunits (Gaiser, Tan, and Richmond 2000), which are the same regions that are homologous between RPA34.5 and RPA49 and the subunits TFIIIF α and β , respectively. Finally, we propose a scenario for the evolution of the nuclear transcription systems in the three domains of life based on the available evidence.

Materials and Methods

We acquired the *S. cerevisiae* coding sequences of the RNAPIII-specific subunits from NCBI and used these as queries in HHpred (Söding, Biegert, and Lupas, 2005). The accession numbers are as follows: RPC31, NP_014248; RPC34, NP_014400; RPC37, NP_012950; RPC53, NP_010131; RPC82, NP_015516. HHpred was used because it is a sensitive remote homology detection method that relies on profile-profile alignments that also considers predicted secondary structural elements in the algorithm (Söding, Biegert, and Lupas, 2005). We queried against all available databases at the HHpred server. For any resulting hits to RNAPII GTFs that have structural data, we determined the region in the target protein structure to which the homologous region corresponded.

In addition to the *S. cerevisiae* single sequence queries, we constructed multiple sequence alignments from these query sequences by including additional hemiascomycete yeast homologs and submitted these alignments as queries to the HHpred server. This was done to control for any alignment errors that might occur during the automated multiple sequence alignment used by HHpred.

Secondary structure prediction of each of the RNAPIII-specific protein query sequences was independently performed using PSIPRED (Jones, 1999; McGuffin, Bryson, and Jones 2000). These predictions were used to complement the secondary structure predictions of HHpred and to support the transfer of structural information from predicted homologs of known structure to the predicted homologous region of the RNAPIII-specific subunits.

Results

RNA polymerase III-specific subunits present in all eukaryotes are homologous to the RNAPII general transcription factors TFII ϵ and TFII F

We used the HHpred server (Söding, Biegert, and Lupas, 2005) to search for regions of homology between RNAPIII-specific subunits and RNA polymerase GTFs found in sequence profile databases. Since the RNAPIII-specific subunits evolve fast compared to the remainder of the RNAPIII subunits (Proshkina *et al.* 2006), we used this sensitive remote homolog detection method to be able to detect potential regions of homology between these subunits and GTFs that are not detectable by traditional homology search methods due to a lack of sensitivity. The RPC37 and RPC53 subunits of the RPC37-RPC53 heterodimer of *S. cerevisiae* showed regions of homology with the human TFII $F\alpha$ (RAP74) and TFII $F\beta$ (RAP30) subunits that form the TFII F heterodimer in RNAPII (figures 4.1a and 4.1b, respectively).

Likewise, subunits RPC82 and RPC34 of the yeast RPC82-RPC34-RPC31 heterotrimer share regions of homology with the TFII $\epsilon\alpha$ and TFII $\epsilon\beta$ subunits of the TFII ϵ GTF of RNAPII (figures 4.1c and 4.1d, respectively). Consequently, RPC82 and RPC34 likely form a heterodimer in organisms that lack RPC31, as do their paralogous subunits in TFII ϵ . The final RNAPIII-specific subunit, RPC31, did not contain regions of homology with GTF HMMs (results not shown).

The regions of homology between RPC37 and TFII $F\alpha$ and RPC53 and TFII $F\beta$ correspond to the binding interface of TFII F (figure 4.2). The binding interfaces of the α and β subunits (RAP74 and RAP30 in humans, respectively) of the human TFII F orthologs form a novel 'triple barrel' dimerization interface that is

encoded by the N-terminal regions of the TFIIF α and β monomers (Gaiser, Tan, and Richmond 2000). The regions of homology correspond to several of the β sheets that form the central barrel. Trypsin digestion suggests that TFIIF α has a flexible linker immediately downstream from the N-terminal binding interface that connects to a C-terminal globular domain (Yong *et al.* 1998). The C-terminal globular domain of TFIIF α has been shown to interact with TFIIB (Fang and Burton, 1996) and RNAPII (Wang and Burton, 1995; Fang and Burton 1996), although its structure remains unsolved. Since the PFAM profile of TFIIF α (PF05793) is constructed from the full protein and since the RPC37 protein has no region that is detectably homologous to the TFIIF α C-terminal region, RPC37 apparently lacks this region that is present in its TFIIF α homolog. The C-terminus of RPC53 is homologous to the TFIIB-interacting N-terminal dimerization interface of TFIIF β (figure 4.1b; Fang and Burton 1996). Since this region is found at the C-terminus of RPC53, RPC53 lacks a downstream region that is homologous to the C-terminal domain of TFIIF β (figure 4.1b). The C-terminal domain of TFIIF β is homologous to region 4 of bacterial σ factors and binds DNA via a cryptic DNA-binding domain (Tan *et al.* 1994) and is required for directed transcription by RNAPII (Garrett *et al.* 1992).

Notably, the regions of both RPC37 and RPC53 that are homologous to the N-termini of TFIIF α and β are also homologous to regions of the RPA49 and RPA34.5 subunits of RNAPI (Kuhn *et al.* 2007). Thus, all three dimers contain this triple β -barrel interface. This clearly indicates that the one-polymerase transcription system in early eukaryotes contained a dimeric GTF that contained this interface. The location of the RPC37-RPC53 binding interface along the RPC37 and RPC53

primary sequences suggests that this heterodimer has a very different structure than its TFIIF and RPA49-RPA34.5 homologs. Secondary structure alignment of both subunits of TFIIF and RPA49-RPA34.5 show that the β -barrel binding interfaces are situated at the N-terminus in all 4 proteins (Kuhn *et al.* 2007). In contrast, the portion of the β -barrel binding interface of RPC53 is situated at its C-terminus (figure 4.1b).

RPC53 also contains two regions that are homologous to other, non-GTF proteins. Part of the N-terminus of RPC53 is homologous to the C-terminal portion of KOG0921 proteins with a probability of 95.2% (figure 4.1b). Proteins comprising KOG0921 are subunits of Dosage compensation complexes. This KOG (eukaryotic orthologous group) is composed of sequences from *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Homo sapiens*. Part of the C-terminus of RPC53 that overlaps with the TFIIF β homology region is also predicted to be homologous to NTR-like domains with a probability 92.17% (figure 4.1b). NTR-like domains also form β -barrel structures, although there is no evidence that they are able to form triple β -barrels structures.

RPC82 and RPC34 share regions of homology with TFIIE α and TFIIE β , respectively (figures 4.1c and 4.1d, respectively). The large RPC82 protein has both an N- and C-terminal region that is homologous to the N-terminal domain of archaeal TFE (figure 4.1c). The TFE protein is the archaeal ortholog of TFIIE α (Meinhart, Blobel, and Cramer 2003). The N-terminal domain of TFE adopts an extended winged helix structure (figure 4.3). This winged helix domain differs from canonical winged helix domains by having an additional α helix at both the N- and C-terminal regions of the domain, extended β -strands that make up wing 1, and a

negatively charged groove between wing 1 and $\alpha 3$ (Meinhart, Blobel, and Cramer 2003). RPC82 also has a region that is homologous to part of the winged helix domain of the penicillinase repressor Blal of *Bacillus licheniformis* (figure 4.1c). This region corresponds to helices 1-3 that form the core of the winged helix domain and the first β -sheet of wing 1. Whether this region of RPC82 forms another winged helix domain is uncertain, as a long α -helix is predicted with high confidence to occupy the position in RPC82 where the second β -strand of the winged helix domain would otherwise lie (results not shown).

The C-terminal half of RPC34 has a low probability of being homologous to TFIIIE β and was only detected when HHpred was initiated with an alignment of hemiascomycete yeast RPC34 sequences, instead of the single yeast query sequence (figure 4.1d). We have included this weakly predicted homology because RPC34 forms a complex with RPC82, which is homologous to the winged helix domain of TFIIIE α . Since TFIIIE is a heterodimer of α and β subunits, it is likely that the RPC82 and RPC34 are homologous to the subunits of this GTF. This region of RPC34 is also predicted to be homologous to the heat shock σ factor SigB (sigmaB) of *Bacillus subtilis* with a probability of 92.39% (figure 4.1d). sigmaB is a member of the σ^{70} family of bacterial general transcription factors, which are responsible for coordinating transcription programs in bacterial cells (Gruber and Gross, 2003). The sigmaB regulon in *B. subtilis* is composed of general stress proteins that are expressed by sigmaB-bound RNA polymerase in *B. subtilis* cells that have encountered an environmental stress (Hecker and Volker, 1998). SigmaB from *B. subtilis* contains 3 domains, which were shown in the *Thermus aquaticus* σ^{70}

homolog to interact with both RNAP and the DNA upstream of the transcription start site (Murakami *et al.* 2002). The region of homology between RPC34 and sigmaB corresponds to the C-terminal portion of the sigmaB 2.3 domain through to the N-terminal third of sigmaB domain 3.2 (figure 4.1d). Structural studies of *Thermus aquaticus* σ^{70} show that the σ^{70} region between domain 2.3 and the beginning of domain 3.3 is positioned on the upstream face of the bacterial polymerase, where it interacts with DNA just upstream from the transcription start site, and where it forms a bridge across the central cleft formed by the β' and β subunits of the bacterial polymerase downstream from the active site (Murakami *et al.* 2002). This region partially overlaps with the predicted location of the RPC31-RPC34-RPC82 heterotrimer on the downstream face of RNAPIII (Proshkina *et al.* 2006).

Discussion

Polymerase-specific subunits are paralogs of RNAPII GTFs

Our results suggest that 4 of the 5 polymerase-specific subunits of RNAPIII are in fact paralogs of subunits of the TFIIIE and TFIIF general transcription factors of RNAPII. Only RPC31 seems to lack homology to other proteins, but this protein is specific to plants and unikonts and is thus not universally present in eukaryotes (Proshkina *et al.* 2006). RPC82 and RPC34 appear to be paralogs of TFIIE α and TFIIE β , respectively, although the confidence in the latter paralogy by HHpred itself is weak (figure 4.1). As discussed above, we believe it to be a true paralog because of its association with RPC82, which is homologous to TFIIE α subunit that associates with TFIIE β (figure 4.3). Finally, we demonstrated that RPC37 and RPC53 are paralogs of TFIIF α and TFIIF β , respectively, and that the region of homology corresponds to the α - β binding interface of TFIIF (figure 4.2). Moreover, this α - β binding interface homology is also present in RNAPI subunits RPA49 and RPA34.5 (Kuhn *et al.* 2007).

Evolution of the cellular transcriptional machinery

Based on the known homologies between transcriptional proteins in the three domains of life, a parsimonious scenario for the evolution of much of the transcriptional machinery emerges (figure 4.4). Structural studies of the RNAPs from the three domains of life indicate that the most recent common cellular ancestor of the three domains of life (MRCAC) had a RNAP with essentially the same subunit

composition of that of modern eubacteria, consisting of the β' and β subunits, an α -homodimer, and an ω subunit (figure 4.4; Zhang *et al.* 1999; Cramer, Bushnell, and Kornberg 2001). The bacterial σ factor is the sole family of general transcription factors in bacteria and both sequence and structural comparisons indicate that this family is homologous to TFIIF β subunit of RNAPII (Sopta, Burton, and Greenblatt 1989; Garrett *et al.* 1992; Chung *et al.* 2003). Remarkably, these two proteins share much of their functionality despite both the vast divergence time separating them and the subsequent elaboration of the transcriptional machinery following the divergence of eubacteria from the lineage leading to archaea and eukaryotes. They both have C-terminal cryptic DNA-binding domains with similar sequences that interact with DNA upstream of the transcription site (Garrett *et al.* 1992; Dombroski *et al.* 1992; Tan *et al.* 1994), they both bind RNAP via a central domain (Lesley and Burgess, 1989; McCracken and Greenblatt, 1991), they both reduce non-specific binding of RNAP to DNA and increase binding to promoter-specific DNA (Hinkle and Chamberlin, 1972; Killeen and Greenblatt, 1992). These shared features of the eukaryotic and bacterial RNAPs indicate that the MRCAC contained a σ -like general transcription factor that was responsible for all initiation steps from promoter recognition through open complex formation. Thus, the transcriptional machinery of MRCAC remained largely unchanged throughout the evolution of eubacteria, but became increasingly elaborate in the lineage leading to archaea and eukaryotes.

The most recent common ancestor of archaea and eukaryotes (MRCAAE) had both a more complex polymerase subunit composition and a larger repertoire of general transcription factors than those of eubacteria. Most of the subunit

complexity that is observed in eukaryotic RNAPII had emerged before the archaea/eukaryote-split, as the archaeal RNAP and eukaryotic RNAPII have almost identical subunit composition (figure 4.4). The only archaeal-specific subunit is Rpo13, which was recently identified as a subunit in the high resolution crystal structure of the Crenarchaeote *Sulfolobus shibatae* (Korkhin *et al.* 2009). Since this subunit is specific to Crenarchaea, it likely arose somewhere along this lineage. The TATA-binding protein (TBP), transcription factor B (TFB), and transcription factor E (TFE) GTFs of Archaea all have homologs in eukaryotes. These homologies suggest that the MRCAAE had essentially the same GTF protein composition as extant Crenarchaea (minus Rpo13), but also contained a σ -like factor, since eukaryotes contain a homolog of this GTF (discussed above). No homologs of σ have been identified in extant archaea, so its functions must have been replaced somewhere along the basal branch separating MRCAAE and the extant archaea, where it was lost. Thus, much like in the bacterial case, the ancestral archaea already contained most of the complement of RNAP subunits and GTFs that appear in extant archaeal cells, and elaborations of the RNAPs occurred mostly along the lineage leading to eukaryotes after their divergence from archaea.

Shared subunit composition of the eukaryotic RNAPs and GTFs evokes a scenario for the evolution of the three-RNAP transcriptional system of extant eukaryotes. The earliest eukaryotes likely contained the full complement of RNAPII subunits, which almost all have orthologs in archaeal RNAPs (figure 4.4). The only exception is the archaeal protein TFS/M, which is a GTF in archaea but became permanently recruited to the ancestral eukaryotic RNAP to form a *bona fide* subunit,

RP9. Early eukaryotes contained TBP, TFB, TFE, and σ as their GTFs. The homology between the triple-barrel binding interface of TFIIF α/β , RPA49/34.5 (Kuhn *et al.* 2007), and RPC37/53 (this work) indicates that the σ factor (homolog of TFIIF β , RPA34.5, and RPC53) of early eukaryotes formed a dimer with another subunit via this domain prior to the expansion of the one-RNAP transcriptional system. This newly added GTF subunit consisted of at least the region of the triple-barrel binding interface, which is known to be involved in transcription start site selection, early bond formation, and stabilization of an early elongating complex in TFIIF of *S. cerevisiae* RNAPII (Khapersky *et al.* 2008). The first RNAP duplication event(s) likely resulted in two RNAPs: RNAPI and the ancestral RNAPII/RNAPIII. This is supported by the existence of three homologous proteins that are found in the RNAPII and RNAPIII transcriptional systems but with no homologs in RNAPI. The first two proteins are TFIIE α and TFIIE β which are found in RNAPII and which are homologous to RPC82 and RPC34 in RNAPIII, respectively. The detected homology between sigmaB, TFIIE β , and RPC34 (figures 4.1 and 4.4) suggests that a duplicated σ factor became associated with TFE of the ancestral RNAPII/RNAPIII, forming a second heterodimeric GTF in this RNAP. In contrast, RNAPI lacks a homolog of either subunit of TFIIE, indicating a loss of TFE in this RNAP. Thus, the ancestral TFE was likely lost along the RNAPI lineage but maintained in both RNAPII and RNAPIII following duplication of the RNAPII/RNAPIII ancestor. The third such protein is the archaeal TFB, which is common to both RNAPII (where it is named TFIIB) and RNAPIII (where it is named TFIIB-BRF) but absent from RNAPI.

Following the evolution of the three eukaryotic RNAPs, RNAPI and RNAPIII each independently expanded their RNAP subunit complement by permanent recruitment of GTFs, and each RNAP expanded their repertoire of GTFs (figure 4.4). These recruitment events resulted in expansion in the number of RNAP subunits from 12 in the ancestral eukaryotic RNAP to 14 in RNAPI (Kuhn *et al.* 2007) and 16 in RNAPIII (17 if we include the unikont- and plant-specific RPC31 protein; Proshkina *et al.* 2006).

Although we interpreted the evolution of the cellular transcription system in the framework of the three domain tree of life (Woese, Kandler, and Wheelis 1990), it is equally consistent (parsimonious) with the eocyte hypothesis in which the Crenarchaea and eukaryotes form a monophyletic clade and are sister taxa of Euryarchaea (Lake *et al.* 1984; Lake, 1988; Rivera and Lake, 1992; Cox *et al.* 2008). The two different topologies are almost equivalent in terms of explaining the origin of RNAP proteins and only differ in terms of the location of the origin of the Rpo8/RP8 subunit and the loss of the σ subunit (figure 4.4). In the three-domain tree, Rpo8 is present in MRAECA and is lost along the Euryarchaeal lineage (figure 4.4), whereas in the eocyte tree it does not originate until the advent of the Crenarchae and does not require loss of this protein in any lineage. However, according to the eocyte tree, two separate losses of σ are required, compared to only one loss in the three domain tree. Further support for the eocyte origin of eukaryotes is the recent identification of a Crenarchaeaote that encodes its ortholog of the large subunit of eukaryotic RNAPs as one gene (as in eukaryotes) instead of

the two typical of other archaea (Cox *et al.* 2008; Kwapisz, Beckouët and Thuriaux 2008; Hallam *et al.* 2006).

The ubiquity of horizontal gene transfer (HGT), which is the transfer of genetic material across mating barriers, could also be responsible for the amplification of the number of RNAPs in eukaryotic genomes (see Philippe and Douady (2003) for review). Accordingly, entire transcriptional systems could have been transferred between early eukaryotes, leading to the three RNAP transcriptional system present in all extant eukaryotes. Although this is of course possible, there are a number of problems with this hypothesis. First, the RNAPs are large multi-subunit protein complexes with no known functions of the individual component proteins. These so-called informational genes are proposed to be resistant to HGT according to the Complexity hypothesis proposed by Lake and colleagues (Jain *et al.* 1999). In contrast to informational genes, The complexity hypothesis proposes that operational genes -genes whose proteins have few interaction partners and typically perform enzymatic reactions- are susceptible to HGT (Jain *et al.* 1999). In fact, the RNAPs themselves were shown to be resistant to HGT in a study involving several bacterial species (Daubin *et al.* 2002). Second, the two largest subunit of the three RNAPs have each been used individually to test the eocyte hypothesis with varying degrees of success, although none of the phylogenies of these proteins supported HGT of these proteins (Cox *et al.* 2008). The fact that gene phylogenies do not support the HGT of at least the two largest subunits of the RNAPs also makes the genome fusion hypothesis of Rivera and Lake (2004) unlikely. This theory posits that the eukaryotes derive from the fusion of

an eocyte (an Archaea) and a bacteria (likely a proteobacteria), which is essentially a massive horizontal gene transfer event (Rivera and Lake, 2004). Even if HGT was responsible for the expansion of the RNAPs in early eukaryotes, its occurrence would not change the recruitment events discussed in chapter 4. The only difference would be that HGT created the additional copies of the RNAPs and GTFs instead of gene duplications.

In conclusion, we demonstrated that subunit expansion of RNAPIII occurred in the same manner as that of RNAPI, by the recruitment of transiently associated GTFs into permanently-associated *bona fide* polymerase subunits. This appears to be a repeating theme throughout RNAP evolution because at least three separate permanent GTF recruitment events occurred: i) the archaeal TFS/M, which is a GTF in archaea, became permanently recruited to the ancestral eukaryotic RNAP to form a *bona fide* subunit, RP9; ii) the RNAPI and RNAPIII homologs of TFIIF were recruited as permanent subunits in early eukaryotes to become RPA49 and RPA34.5 in RNAPI and RPC37 and RPC53 in RNAPIII; iii) TFIIE α and TFIIE β were recruited to become RPC82 and RPC34 in RNAPIII.

Acknowledgements

This work was supported by a Discovery Grant from the National Science and Engineering Research Council of Canada to G. D.

Figure legends

Figure 4.1. Regions of homology between the RNAPIII-specific subunits and the RNAPII general transcription factors TFIIE and TFIIIF. The *S. cerevisiae* query subunits are shown with open boxes and the target HMMs are represented as black bars with thicker regions that represent regions of homology between target and query. The *S. cerevisiae* queries are RPC37 (A), RPC53 (B), RPC82 (C), and RPC34 (D). Protein family targets are labeled as database family ID – description – probability of being a true homolog. Only the target HMMs with a probability greater than 80% of being a true query homolog are displayed; overlapping hits are shown only when more than 50% of its homologous region is non-overlapping; all hits against GTFs are shown, regardless of significance and overlaps. The scale bar represents length in amino acids.

Figure 4.2. Regions of homology between RPC37 and RPC53 and the human TFIIIF heterodimer corresponding to the dimer binding interface of TFIIIF α and β . A) The binding interface between the α (green) and β (yellow) subunits of TFIIIF. The small top left figure shows only the two proteins whereas the larger figure also shows conserved amino acids. Red space-filled amino acids denote high-scoring amino acid columns that are highly-similar between TFIIIF and RNAPIII subunit HMMs (denoted ' | ' in the alignment shown in part B); yellow space-filled amino acids are medium-high scoring amino acid columns (denoted ' + ' in the alignment shown in part B). B) HHpred alignment of *S. cerevisiae* query RPC37 HMM (top) and RPC53 HMM (bottom) with target TFIIIF α and TFIIIF β HMMs, respectively. Regions in grey in top left figure denote amino acids missing in RPC53 HMM.

Figure 4.3. Region of homology between the RPC82 C-terminal domain and the *Sulfolobus solfataricus* N-terminal extended winged helix domain of TFE. A). The small top left figure shows only the structure of the *S. solfataricus* extended winged helix domain whereas the larger figure also shows conserved amino acids. Red space-filled amino acids denote high-scoring amino acid columns that are highly-similar between TFIIIF and RNAPIII subunit HMMs (denoted ' | ' in the alignment shown in part B); yellow space-filled amino acids are medium-high scoring amino acid columns (denoted ' + ' in the alignment shown in part B). B) HHpred alignment of the *S. cerevisiae* RPC82 query HMM C-terminal domain and the HMM of *S. solfataricus*' TFE extended winged helix domain.

Figure 4.4. Parsimonious scenarios for the evolution of the cellular RNAPs, emphasizing the homologies between different cellular RNAP subunits and their GTF proteins. The putative ancestral universal RNAP subunits are denoted in gray and lack black outlines. Proteins with black outlines denote proteins found in only a subset of the RNAPs. Circular proteins denote RNAP subunit. Square proteins in the rounded boxes above RNAPs denote proteins that associate transiently with RNAPs. Square proteins on the RNAPs themselves denote GTFs that have become recruited as RNAP subunits. Proteins of the same color but part of different RNAPs share at least one homologous domain. Proteins with overlayed crosses represent

gene losses. Arrows depict subunit origins. MRCAC, most recent common ancestor of all cells; MRAECA, most recent archaeal and eukaryotic common ancestor; MRECA, most recent eukaryotic common ancestor. Heterodimeric proteins are labeled using a "/" symbol.

Figures

Figure 4.1.

B

RPC37

Q SecStr
Q RPC37
Q Consensus
T Consensus
T pfam05793
T SecStr

111 RYKPKSHLWEIDIPLDEQAFYNKDKAESEWNGVNVQTLKGVGVENNGQYA-AFVKDMQVYLVPPIERVAQLKPFYKY 185 (282)
111 RiKPKs--vEvdvpidt---Yd--kg-----m-kQtLss-----ya-g-----eLHLtPl--ivQLRPsf-y 185 (282)
+++++.-|=|++.-+.-+| +|... |-.++.+|+ -...+|.++..|/+.-.-+|---+|
91 k-k-e-qPWIL---gk--kkf---kG~rE-----Ggv~ena~YYvft---dg-feAyPv~-WYNF~pi~~Y 153 (521)
91 KYRLDNQFWILEVEGKEQRQY---KGIRE-----GGVSENATYYVFTMCNDGEFEAYFVGWYNFTPIAQY 153 (521)

RPC53

Q SecStr
Q RPC53
Q Consensus
T Consensus
T 1f3ua
T SecStr

358 GKVGSIRVHKSG---KLSVKIG-----NVVMDIGKGAETTFLOQDVIALSIADDASSAELLGRVDGKIVV 418 (422)
358 G-iGkl~v-kSG---rv-l~G-----dv~-G---FLOe~v-i-----LG~v~k~vv 418 (422)
|.|||||+|+... .|+|.|. .|.+-+...|.|-...+... ++....-|.|.+=|
37 -evGkirI-----Vsl~L~e-----p-----iPkey~L-----q-----VFse-----kl~-Eg~vv~k~Ec 115 (118)
37 GEVGKLRIAKTQGRTEVSFTLNEDLANIHDIGGKPASVSAPREHPFVLQSVGGQTLTVFTSS-SDKLSLEGIVVQRAEC 115 (118)

Q SecStr
Q RPC53
Q Consensus
T Consensus
T 1f3ua
T SecStr

419 TPQ 421 (422)
419 tP~ 421 (422)
-|. .
116 rP~ 118 (118)
116 RPA 118 (118)

Figure 4.2

Eubacteria

Archaea

Eukaryotes

Figure 4.4.

Chapter 5

Discussion, Conclusion, and Future Directions

In chapters 2 through 4, several aspects of the evolution of the eukaryotic RNAPs were addressed. In chapter 2, we investigated the occurrence and distribution of rate-shift sites and changes in the cleft loops of these enzymes. Type 2 sites were found to cluster in all three enzymes, whereas type 1 sites appeared randomly distributed in all three structures (see table 2.4). Furthermore, several length differences between the slowly evolving cleft loops were identified (see table 2.2). Together, these results suggest that for a short time following their formation, the eukaryotic RNAPs temporarily experienced relaxed functional constraints that resulted in several amino acid changes near the active site of each enzyme.

Concomitant with these amino acid changes was the recruitment to RNAPIII of two GTFs that are paralogous to RNAPII GTFs (see chapter 4). These RNAPIII-specific subunits, RPC37-RPC53 and RPC82-RPC34 are paralogous to TFIIF and TFIIE, respectively (see figure 4.1). No homologs were found for the remaining RNAPIII-specific subunit, RPC31. A parsimonious scenario for the evolution of cellular transcription systems was proposed based on these and the previously identified homologies between the transcriptional machineries of the three domains of life (figure 4.4).

In chapter 3, it was demonstrated that the rate of evolution of RNAP and GTF genes in Ascomycetes is influenced by the rate of concerted evolution affecting their target genes. Accordingly, the average pairwise rate of evolution of RNAPI genes

was higher than that of RNAPIII genes and the average pairwise rate of RNAPIII genes was higher than that of RNAPII genes (figure 3.2). Although no difference in the average rate between RNAPI and RNAPIII GTFs was detected, genes from both systems evolve at higher rates than the GTFs of RNAPII (see text in chapter 3). This is consistent with the molecular coevolution hypothesis of Dover and Flavell (1984) since RNAPI genes evolve in a concerted fashion at a rate that is typically faster than RNAPIII genes. In contrast, RNAPII genes typically do not typically evolve via concerted evolution.

The relationship between the structural differentiation of the RNAPs and the divergence of TFIIF and its paralogs in RNAPI and RNAPIII.

The location of type 2 sites near the active centers of RNAPI, RNAPII and RNAPIII (see chapter 2) might result from interactions with RPA34.5-RPA49, TFIIF, or RPC37-RPC53, respectively. In chapter 4, the homology between RPA34.5/RPA49, TFIIF, and RPC37/RPC53 was described, as was that between RPC82/RPC34 and TFIIE (see figure 4.1). Mutations in the N-terminal binding interface of yeast TFIIF α have been identified that suppress mutations in both TFIIB and RPB1 that affect transcription start site selection (Berroteran *et al.* 1994; Pinto *et al.* 1994; Freire-Picos *et al.* 2005). Consequently, the authors proposed that the TFIIF N-terminal binding interface lies in the active site of RNAPII where it physically compensates for the start site selection mutants (Freire-Picos *et al.* 2005). Since the N-terminal binding interface has been maintained in RPA34.5/RPA49 (Kuhn *et al.*

2007) and RPC37/RPC53 (see figure 4.2), type 2 sites situated near the active sites in RNAPI and RNAPIII might have each coevolved with their TFIIF paralogs RPA34.5/RPA49 or RPC37/RPC53 shortly after the expansion of the ancestral one-RNAP to the three-RNAP transcription system of extant eukaryotes.

The rapidly evolving genes of RNAPI, RNAPIII, and of their GTFs might have obscured other homologous relationships between the three transcriptional systems.

Although we identified the homologous relationships between the RNAPIII-specific subunits and the GTFs of RNAPII (figure 4.1), these relationships went unnoticed until the recent development of sensitive remote homology detection methods such as HHpred (Soding *et al.* 2005). This sensitive method was required due to the rapid evolution of these RNAPIII proteins (Proshkina *et al.* 2006). The question arises as to whether there are other homologous relationships between the GTFs of the three transcriptional systems that have gone unnoticed due to the rapid evolution of RNAPI and RNAPIII GTFs (Proshkina *et al.* 2006; see text in chapter 3) combined with their ancient divergence. If so, we would expect these to be more likely identifiable between RNAPII and either RNAPI or RNAPIII GTFs, since the rapid evolution of RNAPI and RNAPIII GTFs would rapidly obscure any relationships between these two systems compared to relationships between either RNAPI or RNAPIII and RNAPII. This is what is observed, since the only known homologous proteins between the three transcriptional systems are TFIIB and TFIIB-BRF

(Colbert and Hahn, 1992). Although these proteins are homologous, they show divergence in function. Both proteins are involved in RNAP recruitment to the PIC but accomplish this in different ways. TFIIB recruits RNAPII directly, whereas TFIIB-BRF recruits RNAPIII using RPC34 as an intermediate. Thus, there exist functional differences between the only identifiable homologs of the three transcriptional systems. Given a faster rate of evolution of RNAPII GTFs, this relationship might have gone unnoticed due to a lack of sensitivity in detection. That the only detectable homology between the GTFs of the three transcriptional systems is found between RNAPII and RNAPIII and not the rapidly evolving RNAPI and RNAPIII GTFs is consistent with the hypothesis that the rapid evolutionary rates of RNAPI and RNAPIII GTF have obscured any existing homologous relationships. If RNAPII evolved at a higher rate, then perhaps even the evolutionary relationship between TFIIB and TFIIB-BRF would be undetectable.

RNAPII is likely the ortholog of archaeal RNAP

The relatively slow evolution of RNAPII and its GTFs relative to the other two transcriptional systems (chapter 3) has caused this transcriptional system to largely retain its ancestral form. Indeed, RNAPII subunit composition is almost identical to the RNAP of Crenarchaea, although these archaea often contain an additional subunit, Rpo13 (Korkhin *et al.* 2009). Furthermore, all three archaeal GTFs have an orthologous GTF in RNAPII, although the GTF repertoire of RNAPII is much expanded compared to the archaeal system (Best *et al.* 2004). Archaeal TBP, TFB,

and TFE are orthologs of RNAPII GTFs TBP, TFIIB, and TFIIE α . Based on the extent of homologous subunits, the RNAPII transcriptional apparatus appears to be the ortholog of the Archaeal system.

Although it is not possible to confirm this hypothesis because the lack of similarity between either RNAPI or RNAPIII and archaeal RNAP might be due simply to the higher rate of evolution of RNAPI and RNAPIII proteins, evidence from plants hints that RNAPII might be the true ortholog of the archaeal RNAP. Plants encode two additional RNAPs (RNAPIV and RNAPV) compared to other eukaryotes (Pontier *et al.* 2005). The evolution of these RNAPs has been amenable to phylogenetic analysis due to their recent formation relative to either of the ubiquitous RNAPI, RNAPII, and RNAPIII that are present in all eukaryotes. The results from such an analysis regarding the evolution of RNAPIV indicate that the two largest subunits of RNAPIV were duplicated separately in a step-wise fashion and that they derive from duplication of the corresponding RNAPII subunits (Luo and Hall, 2007). Thus a precedent exists for the formation of a novel RNAP via duplication of the largest subunits of RNAPII. Combined with the high similarity between the archaeal and eukaryal RNAPII transcriptional systems, it is likely that the RNAPII is the true ortholog of the archaeal transcriptional system.

Conclusion

Regardless of how the three-RNAP transcriptional system of early eukaryotes arose, it was demonstrated here that RNAP underwent several changes preceding the diversification of eukaryotes. The paralogous relationship between the catalytic two largest subunits of the RNAPs allowed each polymerase to interact with its expanding repertoire of RNAP-specific GTFs, diverging in both regulatory control and function. Simultaneously, the selective pressures on amino acids near the active site in the two largest subunits of the RNAPs underwent changes, implying functional differentiation of the RNAPs themselves near their active sites. This is also evident in the different selective pressures and lengths of the cleft loops from each RNAP. RNAPI and RNAPIII then expanded their subunit composition by recruiting two and four GTF proteins as permanent subunits, respectively. Finally, the nature of the three gene classes that emerged in the three-RNAP transcriptional system of early eukaryotes resulted in each class undergoing distinct levels of concerted evolution. The different levels of concerted evolutions, in turn, affected the evolutionary rates of each transcription system.

Future Directions

The investigations in chapters 2-4 result in several unanswered questions regarding the evolution of the RNA polymerases that are amenable to future study. First, the *in silico* study of chapter 2 identified many type 2 sites near the active sites, as well as several length differences between cleft loops of all three eukaryotic

RNAPs (see figure 2.3-2.4 and table 2.2). These were predicted to be of functional relevance to the properties of each RNAP, potentially affecting the transcriptional properties and interactions between GTFs and the RNAPs (see discussion of chapter 2). Thus, chapter 2 resulted in several predictions that cannot be confirmed without experimental validation. For confirmation, site-directed mutagenesis can be employed to target the type 2 sites and to generate length differences in the cleft loops. The functional consequences of these mutations can be investigated by assessing transcriptional properties of the RNAPs as well as their association with GTFs during PIC formation.

In chapter 3, the molecular coevolution hypothesis was tested using the three transcriptional systems present in all eukaryotes. Our results were consistent with this hypothesis since we identified a correlation between concerted evolution and the non-synonymous substitution rate (dN) of RNAP and GTF genes (see table 3.2 and chapter 3 results section). Since concerted evolution is indifferent to whether a site is non-synonymous or synonymous, dS should also be correlated with the extent of concerted evolution. Thus, further evidence for the validity of the molecular coevolution hypothesis would be the demonstration that the dS of a gene is also correlated with the extent of concerted evolution experienced by the gene. This approach has the advantage that it is not confounded by protein structure effects that can influence dN (Bloom *et al.* 2006). Moreover, recent models allow the separation of mutation bias and codon fitness in calculating dS (Yang and Nielsen, 2008), further controlling for other factors known to affect dS such as codon bias (Hershberg and Petrov, 2008).

The homology between TFIIF, RPA34.5/49, and RPC37/RPC53 as well as between TFIIE and RPC82/34 that was demonstrated in chapter 4 (figure 4.1) illustrates the efficacy of the current generation of homology detection methods such as HHpred. Thus, the possibility that homologies exist between the GTFs of the three eukaryotic transcription systems beyond TFIIB and TFIIB-Brf can be examined. As postulated in chapter 5, such homologies might have gone unnoticed due to the concerted evolution-driven rapid evolution of these genes in RNAPI and RNAPIII. Extending the analysis of chapter 4 to the GTFs of the three eukaryotic RNAPs would therefore also be a fruitful avenue for future research.

Literature Cited

- Alic N, Ayoub N, Landrieux E, Favry E, Baudouin-Cornu P, Riva M, Carles C. 2007. Selectivity and proofreading both contribute significantly to the fidelity of RNA polymerase III transcription. *Proc Natl Acad Sci USA*. **104**:10400-10405.
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25**:3389-3402.
- Amstutz H, Munz P, Heyer WD, Leupold U, Kohli J. 1985. Concerted evolution of tRNA genes: intergenic conversion among three unlinked serine tRNA genes in *S. pombe*. *Cell*. **40**:879-886.
- Andrau JC, Sentenac A, Werner M. 1999. Mutagenesis of yeast TFIIIB70 reveals C-terminal residues critical for interaction with TBP and C34. *J Mol Biol*. **288**:511-520.
- Arnheim N (1983) Concerted evolution of multigene families. In: Nei M, Koehn R (eds) *Evolution of genes and proteins*. Sinauer, Sunderland, pp 38–61.
- Bar-Nahum G, Epshtein V, Ruckenstein AE, Rafikov R, Mustaev A, Nudler E. 2005. A ratchet mechanism of transcription elongation and its control. *Cell*. **120**:183-193.
- Beckouet F, Labarre-Mariotte S, Albert B, Imazawa Y, Werner M, Gadal O, Nogi Y, Thuriaux P. 2008. Two RNA polymerase I subunits control the binding and release of Rrn3 during transcription. *Mol Cell Biol*. **28**:1596-1605.
- Berroteran RW, Ware DE, Hampsey M. 1994. The sua8 suppressors of *Saccharomyces cerevisiae* encode replacements of conserved residues within

- the largest subunit of RNA polymerase II and affect transcription start site selection similarly to sua7 (TFIIB) mutations. *Mol Cell Biol.* **14**:226-237.
- Best AA, Morrison HG, McArthur AG, Sogin ML, Olsen GJ. 2004. Evolution of eukaryotic transcription: insights from the genome of *Giardia lamblia*. *Genome Res.* **14**:1537-1547.
- Birney E, Clamp M, Durbin R. 2004. GeneWise and Genomewise. *Genome Res.* **14**:988-995.
- Bloom JD, Drummond DA, Arnold FH, Wilke CO. 2006. Structural determinants of the rate of protein evolution in yeast. *Mol Biol Evol.* **23**:1751-1761.
- Brueckner F, Cramer P. 2008. Structural basis of transcription inhibition by alpha-amanitin and implications for RNA polymerase II translocation. *Nat Struct Mol Biol.* **15**:811-818.
- Brun I, Sentenac A, Werner M. 1997. Dual role of the C34 subunit of RNA polymerase III in transcription initiation. *EMBO J.* **16**:5730-5741.
- Bushnell DA, Westover KD, Davis RE, Kornberg RD. 2004. Structural basis of transcription: an RNA polymerase II-TFIIB cocystal at 4.5 angstroms. *Science.* **303**:983-988.
- Chédin S, Riva M, Schultz P, Sentenac A, Carles C. 1998. The RNA cleavage activity of RNA polymerase III is mediated by an essential TFIIIS-like subunit and is important for transcription termination. *Genes Dev.* **12**:3857-3871.
- Chen HT, Hahn S. 2004. Mapping the location of TFIIB within the RNA polymerase II transcription preinitiation complex: a model for the structure of the PIC. *Cell.* **119**:169-180.

- Chen HT, Warfield L, Hahn S. 2007. The positions of TFIIF and TFIIE in the RNA polymerase II transcription preinitiation complex. *Nat Struct Mol Biol.* **14**:696-703.
- Cheng G, Qian B, Samudrala R, Baker D. 2005. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Res.* **33**:5861-5867.
- Chiannilkulchai N, Stalder R, Riva M, Carles C, Werner M, Sentenac A. 1992. RPC82 encodes the highly conserved, third-largest subunit of RNA polymerase C (III) from *Saccharomyces cerevisiae*. *Mol Cell Biol.* **12**:4433-40.
- Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**:823-826.
- Chung WH, Craighead JL, Chang WH, Ezeokonkwo C, Bareket-Samish A, Kornberg RD, Asturias FJ. 2003. RNA polymerase II/TFIIF structure and conserved organization of the initiation complex. *Mol Cell.* **12**:1003-1013.
- Colbert T, Hahn S. 1992. A yeast TFIIB-related factor involved in RNA polymerase III transcription. *Genes Dev.* **6**:1940-1949.
- Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaeobacterial origin of eukaryotes. *Proc Natl Acad Sci U S A.* **105**:20356-20361.
- Cramer, P. 2002. Multisubunit RNA polymerases. *Curr Opin Struct Biol.* **12**:89-97.
- Cramer P, Armache KJ, Baumli S, Benkert S, Brueckner F, Buchen C, Damsma GE, Dengl S, Geiger SR, Jasiak AJ, Jawhari A, Jennebach S, Kamenski T, Kettenberger H, Kuhn CD, Lehmann E, Leike K, Sydow JF, Vannini A. 2008. Structure of eukaryotic RNA polymerases. *Annu Rev Biophys.* **37**:337-352.

- Cramer P, Bushnell DA, Kornberg RD. 2001. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science*. **292**:1863-1876.
- Daubin V, Gouy M, Perrière G. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res*. **12**:1080-1090.
- Davis JA, Takagi Y, Kornberg RD, Asturias FA. 2002. Structure of the yeast RNA polymerase II holoenzyme: Mediator conformation and polymerase interaction. *Mol Cell*. **10** :409-415.
- De Carlo S, Carles C, Riva M, Schultz P. 2003. Cryo-negative staining reveals conformational flexibility within yeast RNA polymerase I. *J Mol Biol* **329**:891-902.
- Dombroski AJ, Walter WA, Record MT Jr, Siegele DA, Gross CA. 1992. Polypeptides containing highly conserved regions of transcription initiation factor sigma 70 exhibit specificity of binding to promoter DNA. *Cell*. **70**:501-512.
- Dover GA. 1982. Molecular drive: a cohesive mode of species evolution. *Nature*. **299**:111-117
- Dover GA, Flavell RB. 1984. Molecular coevolution: DNA divergence and the maintenance of function. *Cell*. **38**:622-623.
- Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH, 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA*. **102**:14338-14343.
- Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol*. **23**:327-337.
- Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*. **134**:341-352.

- Dvir A. 2002. Promoter escape by RNA polymerase II. *Biochim Biophys Acta*. **1577**:208-223.
- Epshtein V, Mustaev A, Markovtsov V, Bereshchenko O, Nikiforov V, Goldfarb A. 2002. Swing-gate model of nucleotide entry into the RNA polymerase active center. *Mol Cell* **10**:623-634.
- Facciotti MT, Reiss DJ, Pan M, Kaur A, Vuthoori M, Bonneau R, Shannon P, Srivastava A, Donohoe SM, Hood LE, Baliga NS. 2007. General transcription factor specified global gene regulation in archaea. *Proc Natl Acad Sci USA*. **104**:4630-4635.
- Fang SM, Burton ZF. 1996. RNA polymerase II-associated protein (RAP) 74 binds transcription factor (TF) IIB and blocks TFIIB-RAP30 binding. *J Biol Chem*. **271**:11703-11709.
- Fernández-Tornero C, Böttcher B, Riva M, Carles C, Steuerwald U, Ruigrok RW, Sentenac A, Müller CW, Schoehn G. 2007. Insights into transcription initiation and termination from the electron microscopy structure of yeast RNA polymerase III. *Mol Cell*. **25**:813-23.
- Fitzpatrick DA, Logue ME, Stajich JE, Butler G. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol Biol*. **6**:99.
- Freire-Picos MA, Krishnamurthy S, Sun ZW, Hampsey M. 2005. Evidence that the Tfg1/Tfg2 dimer interface of TFIIF lies near the active center of the RNA polymerase II initiation complex. *Nucleic Acids Res*. **33**:5045-5052.

- Gaiser F, Tan S, Richmond TJ. 2000. Novel dimerization fold of RAP30/RAP74 in human TFIIF at 1.7 Å resolution. *J Mol Biol.* **302**:1119-1127.
- Ganley AR, Kobayashi T. 2007. Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Res.* **17**:184-191.
- Garrett KP, Serizawa H, Hanley JP, Bradsher JN, Tsuboi A, Arai N, Yokota T, Arai K, Conaway RC, Conaway JW. 1992 The carboxyl terminus of RAP30 is similar in sequence to region 4 of bacterial sigma factors and is required for function. *J Biol Chem.* **267**:23942-23949.
- Gnatt AL, Cramer P, Fu J, Bushnell DA, Kornberg RD. 2001. Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 Å resolution. *Science.* **292**:1876-1882.
- Graur D, Li W-H. 2000. Fundamentals of molecular evolution, 2nd edn., Sinauer Associates, Sunderland.
- Gribaldo S, Casane D, Lopez P, Philippe H. 2003. Functional divergence prediction from evolutionary analysis: a case study of vertebrate hemoglobin. *Mol Biol Evol.* **20**:1754-1759.
- Gruber TM, Gross CA. 2003. Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu Rev Microbiol.* **57**:441-466.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* **52**:696-704.
- Gu J, Wang Y, Gu X. 2002. Evolutionary analysis for functional divergence of Jak protein kinase domains and tissue-specific genes. *J Mol Evol.* **54**:725-733.

- Gu X. 1999. Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol.* **16**:1664-1674.
- Gu X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. *Mol Biol Evol.* **18**:453-464.
- Hallam SJ, Konstantinidis KT, Putnam N, Schleper C, Watanabe Y, Sugahara J, Preston C, de la Torre J, Richardson PM, DeLong EF. 2006. Genomic analysis of the uncultivated marine *Crenarchaeote* *Cenarchaeum symbiosum*. *Proc Natl Acad Sci USA.* **103**:18296-18301.
- Hecker M, Völker U. 1998. Non-specific, general and multiple stress resistance of growth-restricted *Bacillus subtilis* cells by the expression of the sigmaB regulon. *Mol Microbiol.* **29**:1129-1136.
- Hershberg R, Petrov DA. Selection on codon bias. 2008. *Annu Rev Genet.* **42**:287-299.
- Hinkle DC, Chamberlin MJ. 1972. Studies of the binding of *Escherichia coli* RNA polymerase to DNA. I. The role of sigma subunit in site selection. *J Mol Biol.* **70**:157-185.
- Hirata A, Klein BJ, Murakami KS. 2008. The X-ray crystal structure of RNA polymerase from Archaea. *Nature.* **451**:851-854.
- Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell.* **95**:717-728.
- Hsu LM. 2002. Promoter clearance and escape in prokaryotes. *Biochim Biophys Acta.* **1577**:191-207.

- Ishiguro A, Kimura M, Yasui K, Iwata A, Ueda S, Ishihama A. 1998. Two large subunits of the fission yeast RNA polymerase II provide platforms for the assembly of small subunits. *J Mol Biol.* **279**:703-712.
- Iyer LM, Koonin EV, Aravind L. 2003. 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.* **3**:1.
- Izban MG, Luse DS. 1992. The RNA polymerase II ternary complex cleaves the nascent transcript in a 3'→5' direction in the presence of elongation factor SII. *Genes Dev.* **6**:1342-1356.
- Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA.* **96**:3801-3806.
- Jasiak AJ, Armache KJ, Martens B, Jansen RP, Cramer P. 2006. Structural biology of RNA polymerase III: subcomplex C17/25 X-ray structure and 11 subunit enzyme model. *Mol Cell.* **23**:71-81.
- Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol.* **292**:195-202.
- Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* **8**:275-282.
- Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**:3059-3066.
- Kenneth NS, Marshall L, White RJ. 2008. Recruitment of RNA polymerase III in vivo. *Nucleic Acids Res.* **36**:3757-3764.

- Kettenberger H, Armache KJ, Cramer P. 2003. Architecture of the RNA polymerase II-TFIIS complex and implications for mRNA cleavage. *Cell*. **114**:347-357.
- Kettenberger H, Armache KJ, Cramer P. 2004. Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS. *Mol Cell*. **16**:955-965.
- Khapersky DA, Ammerman ML, Majovski RC, Ponticelli AS. 2008. Functions of *Saccharomyces cerevisiae* TFIIF during transcription start site utilization. *Mol Cell Biol*. **28**:3757-3766.
- Killeen MT, Greenblatt JF. 1992. The general transcription factor RAP30 binds to RNA polymerase II and prevents it from binding nonspecifically to DNA. *Mol Cell Biol*. **12**:30-37.
- Knudsen B, Miyamoto MM, Laipis PJ, Silverman DN. 2003. Using evolutionary rates to investigate protein functional divergence and conservation. A case study of the carbonic anhydrases. *Genetics*, **164**:1261-1269.
- Koonin EV. 2009. Darwinian evolution in the light of genomics. *Nucleic Acids Res*. **37**:1011-1034.
- Koonin EV, Makarova KS, Elkins JG. 2007. Orthologs of the small RPB8 subunit of the eukaryotic RNA polymerases are conserved in hyperthermophilic Crenarchaeota and "Korarchaeota". *Biol Direct*. **2**:38.
- Korkhin Y, Unligil UM, Littlefield O, Nelson PJ, Stuart DI, Sigler PB, Bell SD, Abrescia NG. 2009. Evolution of complex RNA polymerases: The complete archaeal RNA polymerase structure. *PLoS Biol*. **7**:e102.

- Koski LB, Golding GB. 2001. The closest BLAST hit is often not the nearest neighbor. *J Mol Evol.* **52**:540-542.
- Kuhn, C. D., Geiger, S. R., Baumli, S., Gartmann, M., Gerber, J., Jennebach, S., Mielke, T., Tschochner, H., Beckmann, R. & Cramer, P. 2007. Functional architecture of RNA polymerase I. *Cell.* **131**:1260-1272.
- Kusser AG, Bertero MG, Naji S, Becker T, Thomm M, Beckmann R, Cramer P. 2008. Structure of an archaeal RNA polymerase. *J Mol Biol.* **376**:303-307.
- Kwapisz M, Beckouët F, Thuriaux P. 2008. Early evolution of eukaryotic DNA-dependent RNA polymerases. *Trends Genet.* **24**:211-215.
- Lake JA, Henderson E, Oakes M, Clark MW. 1984. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc Natl Acad Sci USA.* **81**:3786-3790.
- Lake JA. 1988. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature.* **331**:184-186.
- Landrieux E, Alic N, Ducrot C, Acker J, Riva M, Carles C. 2006. A subcomplex of RNA polymerase III subunits involved in transcription termination and reinitiation. *EMBO J.* **25**:118-128.
- Lang WH, Reeder RH. (1995). Transcription termination of RNA polymerase I due to a T-rich element interacting with Reb1p. *Proc Natl Acad Sci USA.* **92**:9781-9785.
- Lesley SA, Burgess RR. 1989. Characterization of the *Escherichia coli* transcription factor sigma 70: localization of a region involved in the interaction with core RNA polymerase. *Biochemistry.* **28**:7728-7734.

- Leung J, Sinclair DA, Hayashi S, Tener GM, Grigliatti TA. 1991. Informational redundancy of tRNA(4Ser) and tRNA(7Ser) genes in *Drosophila melanogaster* and evidence for intergenic recombination. *J Mol Biol.* **219**:175-88.
- Liao D. 1999. Concerted evolution: molecular mechanism and biological implications. *Am J Hum Genet.* **64**:24-30.
- Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. *Mol Biol Evol.* **19**:1-7.
- Lu Q, Han J, Zhou L, Coker JA, DasSarma P, DasSarma S, Xiang H. 2008. Dissection of the regulatory mechanism of a heat-shock responsive promoter in Haloarchaea: a new paradigm for general transcription factor directed archaeal gene regulation. *Nucleic Acids Res.* **36**:3031-3042.
- Luo J, Hall BD. 2007. A multistep process gave rise to RNA polymerase IV of land plants. *J Mol Evol.* **64**:101-112.
- Lynch M. 2007a. The origins of genome architecture. Sinauer Associates, Sunderland, MA.
- Lynch, M. 2007b. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci USA.* **104**:8597-8604.
- Marck C, Kachouri-Lafond R, Lafontaine I, Westhof E, Dujon B, Grosjean H. 2006. The RNA polymerase III-dependent family of genes in hemiascomycetes: comparative RNomics, decoding strategies, transcription and evolutionary implications. *Nucleic Acids Res.* **34**:1816-1835.

- Mayrose I, Graur D, Ben-Tal N, Pupko T. 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol Biol Evol.* **21**:1781-1791.
- McCracken S, Greenblatt J. 1991. Related RNA polymerase-binding regions in human RAP30/74 and *Escherichia coli* sigma 70. *Science.* **253**:900-902.
- McGuffin LJ, Bryson K, Jones DT. 2000. The PSIPRED protein structure prediction server. *Bioinformatics.* **16**:404-405.
- McInerney JO. 2006. The causes of protein evolutionary rate variation. *Trends Ecol Evol.* **21**:230-232.
- Meinhart A, Blobel J, Cramer P. 2003. An extended winged helix domain in general transcription factor E/II α . *J Biol Chem.* **278**:48267-48274.
- Morzycka-Wroblewska E, Selker EU, Stevens JN, Metzenberg RL. 1985. Concerted evolution of dispersed *Neurospora crassa* 5S RNA genes: pattern of sequence conservation between allelic and nonallelic genes. *Mol Cell Biol.* **5**:46-51.
- Mosrin C, Riva M, Beltrame M, Cassar E, Sentenac A, Thuriaux P. 1990. The RPC31 gene of *Saccharomyces cerevisiae* encodes a subunit of RNA polymerase C (III) with an acidic tail. *Mol Cell Biol.* **10**:4737-4743.
- Murakami KS, Masuda S, Campbell EA, Muzzin O, Darst SA. 2002. Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex. *Science.* **296**:1285-1290.
- Naji S, Bertero MG, Spitalny P, Cramer P, Thomm M. 2008. Structure function analysis of the RNA polymerase cleft loops elucidates initial transcription, DNA unwinding and RNA displacement. *Nucleic Acids Res.* **36**:676-687.

- Nesser NK, Peterson DO, Hawley DK. 2006. RNA polymerase II subunit Rpb9 is important for transcriptional fidelity in vivo. *Proc Natl Acad Sci USA*. **103**:3268-3273.
- Nogi Y, Yano R, Dodd J, Carles C, Nomura M. 1993. Gene RRN4 in *Saccharomyces cerevisiae* encodes the A12.2 subunit of RNA polymerase I and is essential only at high temperatures. *Mol Cell Biol*. **13**:114-122.
- Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*. **302**:205-217.
- Nudler E. 2009. RNA polymerase active center: the molecular engine of transcription. *Annu Rev Biochem*. **78**:335-361.
- Orlicky SM, Tran PT, Sayre MH, Edwards AM. 2001. Dissociable RNAPII4-RNAPII7 subassembly of RNA polymerase II binds to single-strand nucleic acid and mediates a post-recruitment step in transcription initiation. *J Biol Chem*. **276**:10097-10102.
- Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics*. **158**:927-931.
- Paule MR, White RJ. 2000. Survey and summary: transcription by RNA polymerases I and III. *Nucleic Acids Res*. **28**:1283-1298.
- Peyroche G, Milkereit P, Bischler N, Tschochner H, Schultz P, Sentenac A, Carles C, Riva M. 2000. The recruitment of RNA polymerase I on rDNA is mediated by the interaction of the A43 subunit with Rrn3. *EMBO J*. **19**:5473-5482.
- Philippe H, Douady CJ. 2003. Horizontal gene transfer and phylogenetics. *Curr Opin Microbiol*. **6**:498-505.

- Pinto I, Wu WH, Na JG, Hampsey M. 1994. Characterization of sua7 mutations defines a domain of TFIIB involved in transcription start site selection in yeast. *J Biol Chem.* **269**:30569-30573.
- Pontier D, Yahubyan G, Vega D, Bulski A, Saez-Vasquez J, Hakimi MA, Lerbs-Mache S, Colot V, Lagrange T. 2005. Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in Arabidopsis. *Genes Dev.* **19**:2030-2040.
- Prokopowich CD, Gregory TR, Crease TJ. 2003. The correlation between rDNA copy number and genome size in eukaryotes. *Genome.* **46**:48-50.
- Proshkina GM, Shematorova EK, Proshkin SA, Zaros C, Thuriaux P, Shpakovski GV. 2006. Ancient origin, functional conservation and fast evolution of DNA-dependent RNA polymerase III. *Nucleic Acids Res.* **34**:3615-3624.
- R Development Core Team. 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>
- Rivera MC, Lake JA. 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science.* **257**:74-76.
- Rivera MC, Lake JA. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature.* **431**:152-155.
- Rooney AP, Ward TJ. 2005. Evolution of a large ribosomal RNA multigene family in filamentous fungi: Birth and death of a concerted evolution paradigm. *Proc Natl Acad Sci USA.* **102**:5084-5089.

- Schlötterer C, Tautz D. 1994. Chromosomal homogeneity of *Drosophila* ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. *Curr Biol.* **4**:777-783.
- Siaut M, Zaros C, Levivier E, Ferri ML, Court M, Werner M, Callebaut I, Thuriaux P, Sentenac A, Conesa C. 2003. An RNAPII4/RNAPII7-like complex in yeast RNA polymerase III contains the orthologue of mammalian CGRP-RCP. *Mol Cell Biol.* **23**:195-205.
- Siomi H, Siomi MC. 2009. On the road to reading the RNA-interference code. *Nature.* **457**:396-404.
- Söding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**:W244-248.
- Sopta M, Burton ZF, Greenblatt J. 1989. Structure and associated DNA-helicase activity of a general transcription initiation factor that binds to RNA polymerase II. *Nature.* **341**:410-414.
- Stettler S, Mariotte S, Riva M, Sentenac A, Thuriaux P. 1992. An essential and specific subunit of RNA polymerase III (C) is encoded by gene RPC34 in *Saccharomyces cerevisiae*. *J Biol Chem.* **267**:21390-21395.
- Suenaga A, Okimoto N, Futatsugi N, Hirano Y, Narumi T, Ohno Y, Yanai R, Hirokawa T, Ebisuzaki T, Konagaya A, Taiji M. (2006). Structure and dynamics of RNA polymerase II elongation complex. *Biochem Biophys Res Commun.* **343**:90-98.

- Sun ZW, Hampsey M. (1995). Identification of the gene (SSU71/TFG1) encoding the largest subunit of transcription factor TFIIIF as a suppressor of a TFIIIB mutation in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA*. **92**:3127-3131.
- Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. **34**:W609-612.
- Szostak JW, Wu R. 1980. Unequal crossing over in the ribosomal DNA of *Saccharomyces cerevisiae*. *Nature*. **284**:426-430.
- Tan L, Wiesler S, Trzaska D, Carney HC, Weinzierl RO. (2008). Bridge helix and trigger loop perturbations generate superactive RNA polymerases. *J Biol*. **7**:40.
- Tan S, Garrett KP, Conaway RC, Conaway JW. 1994. Cryptic DNA-binding domain in the C terminus of RNA polymerase II general transcription factor RAP30. *Proc Natl Acad Sci USA*. **91**:9808-9812.
- Thomas MC, Chiang CM. 2006. The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol*. **41**:105-178.
- Thuillier V, Stettler S, Sentenac A, Thuriaux P, Werner M. 1995. A mutation in the C31 subunit of *Saccharomyces cerevisiae* RNA polymerase III affects transcription initiation. *EMBO J*. **14**:351-359.
- Tóth-Petróczy A, Oldfield CJ, Simon I, Takagi Y, Dunker AK, Uversky VN, Fuxreiter M. (2008). Malleable machines in transcription regulation: the mediator complex. *PLoS Comput Biol*. **4**:e1000251.
- Wang BQ, Burton ZF. 1995. Functional domains of human RAP74 including a masked polymerase binding domain. *J Biol Chem*. **270**:27035-27044.

- Wang D, Bushnell DA, Westover KD, Kaplan CD, Kornberg RD. 2006. Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. *Cell*. **127**:941-954.
- Wang Y, Gu X. 2001. Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. *Genetics*, **158**:1311-1320.
- Warner JR 1999. The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci*. **24**:437-440.
- Werner F, Weinzierl RO. 2002. A recombinant RNA polymerase II-like enzyme capable of promoter-specific transcription. *Mol Cell*. **10**:635-646.
- Westover KD, Bushnell DA, Kornberg RD. 2004. Structural basis of transcription: separation of RNA from DNA by RNA polymerase II. *Science*, **303**:1014-1016.
- Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol*. **18**:691-699.
- Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U SA*. **87**:4576-4579.
- Wösten MM. 1998. Eubacterial sigma-factors. *FEMS Microbiol Rev*. **22**:127-50.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. **24**:1586-1591.
- Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol*. **25**:568-579.

- Yang Z, Swanson WJ. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol.* **19**:49-57.
- Yong C, Mitsuyasu H, Chun Z, Oshiro S, Hamasaki N, Kitajima S. 1998. Structure of the human transcription factor TFIIF revealed by limited proteolysis with trypsin. *FEBS Lett.* **435**:191-194.
- Zar JH. 1999. Biostatistical analysis. 4th ed. Prentice Hall, Upper Saddle River, NJ, pp 196-200.
- Zhang G, Campbell EA, Minakhin L, Richter C, Severinov K, Darst SA. 1999. Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution. *Cell.* **98**:811-824.
- Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol.* **22**:2472-2479.
- Zhang Z., Fu J, Gilmour DS. 2005. CTD-dependent dismantling of the RNA polymerase II elongation complex by the pre-mRNA 3'-end processing factor, Pcf11. *Genes Dev.* **19**:1572-1580.
- Zhou T, Drummond DA, Wilke CO. 2008. Contact density affects protein evolutionary rate from bacteria to animals. *J Mol Evol.* **66**:395-404.