Comparative Genomics and Evolution of Proteins Associated with RNA Polymerase II C-Terminal Domain

Zhenhua Guo and John W. Stiller

Howell Science Complex N108, Department of Biology, East Carolina University

The C-terminal domain (CTD) of the largest subunit of RNA polymerase II provides an anchoring point for a wide variety of proteins involved in mRNA synthesis and processing. Most of what is known about CTD-protein interactions comes from animal and yeast models. The consensus sequence and repetitive structure of the CTD is conserved strongly across a wide range of organisms, implying that the same is true of many of its known functions. In some eukaryotic groups, however, the CTD has been allowed to degenerate, suggesting a comparable lack of essential protein interactions. To date, there has been no comprehensive examination of CTD-related proteins across the eukaryotic domain to determine which of its identified functions are correlated with strong stabilizing selection on CTD structure. Here we report a comparative investigation of genes encoding 50 CTD-associated proteins, identifying putative homologs from 12 completed or nearly completed eukaryotic genomes. The presence of a canonical CTD generally is correlated with the apparent presence and conservation of its known protein partners; however, no clear set of interactions emerges that is invariably linked to conservation of the CTD. General rates of evolution, phylogenetic patterns, and the conservation of modeled tertiary structure of capping enzyme guanylyltransferase (Cgt1) indicate a pattern of coevolution of components of a transcription factory organized around the CTD, presumably driven by common functional constraints. These constraints complicate efforts to determine orthologous gene relationships and can mislead phylogenetic and informatic algorithms.

Introduction

Transcription of protein-encoding genes in eukaryotes is an elaborate process involving a myriad of coordinated functions. Growing evidence indicates that this coordination occurs in large subnuclear complexes called "transcription factories" or "transcriptosomes" (Halle and Meisterernst 1996; Iborra et al. 1996), where the diverse proteins required for synthesis of a complete and functional mRNA are brought together (Howe 2002; Zorio and Bentley 2004). These subnuclear compartments are dynamic but structurally stable, recruiting active genes into preassembled transcription and processing centers (Osborne et al. 2004). At the core of the machinery of the transcription factory is RNA polymerase II (RNAP II) and its repetitive C-terminal domain (CTD).

The RNAP II CTD comprises a series of seven amino acid repeats; different organisms have varied numbers of repeats (e.g., 52 in mammals, 26–28 in yeast), but the "canonical" sequence (tandemly arrayed heptapeptides with the consensus Y_1-S_2-P_3-T_4-S_5-P_6-S_7) is conserved across diverse eukaryotes (Corden 1990; Stiller and Hall 2002). The CTD has been described alternatively as a "landing pad" or a "symphony conductor"; it coordinates a wide variety of interactions with proteins needed to transcribe a typical eukaryotic gene and then process the resulting message (Bentley 1999, 2002; Hirose and Manley 2000; Howe 2002; Zorio and Bentley 2004). In this role it is a key organizing center for the transcriptosome (Carty and Greenleaf 2002; Howe 2002). Evidence of additional and highly specific protein interactions with the CTD in both mammals and yeast, a number not directly related to transcription, suggests that CTD is a center for the organization of general nuclear function (Carty and Greenleaf 2002; Phatnani, Jones, and Greenleaf 2004).

Key words: capping enzyme, comparative genomics, evolution, CTD-associated proteins, C-terminal domain, RNA polymerase II.

E-mail: stillerj@mail.ecu.edu.

Mol. Biol. Evol. 22(11):2166–2178. 2005
doi:10.1093/molbev/msi215
Advance Access publication July 13, 2005

Given the importance of the CTD for coordinating RNAP II transcription and very likely other essential functions it comes as no surprise that its consensus sequence, repetitive nature, and many of its key roles have been conserved from yeast to mammals (Corden 1990; Riedl and Egly 2000; Maniatis and Reed 2002; Palancade and Bensaude 2003). It is also not surprising that canonical CTD heptads are conserved strongly in many other eukaryotic groups, for which transcriptional processes are not well characterized (Stiller and Hall 2002; Stiller and Cook 2004). Broad-scale comparisons of the RNAP II largest subunit (RPB1), however, show that the CTD has been permitted to degenerate in a number of protistan groups; in some cases, little or no evidence of a repetitive heptapeptide structure remains. In phylogenetic analyses of RPB1 sequences, eukaryotes in which the CTD is conserved generally are recovered as a unique evolutionary group, suggesting that they share a single common ancestor (Stiller and Hall 2002). Moreover, the level of degeneration of the CTD found in most organisms outside this grouping is incompatible with essential RNAP II function in yeast (Stiller and Cook 2004). These combined evolutionary and genetic analyses suggest that major differences in RNAP II transcription, perhaps the coalescence of the transcriptosome itself, are responsible for the phylogenetic pattern of CTD conservation observed; however, the specific protein interactions responsible for these evolutionary differences have not been investigated.

Most CTD-associated proteins have been identified and investigated functionally only in animals and/or yeast and very little is known for other eukaryotes. Thus far, more than 50 proteins have been reported to interact with the CTD, mostly from biochemical and genetic studies (see Supplementary Table 1, Supplementary Material online). In addition, structures of three proteins in complex with CTD heptapeptides have been determined: capping enzyme (CE) guanylyltransferase (GTase) (Cgt1) (Fabrega et al. 2003), pre-mRNA cleavage complex II protein (Pcf11) (Meinhart and Cramer 2004), and peptidyl-proline isomerase (Pin1) (Verdecia et al. 2000). These combined data provide the basis for a comprehensive investigation of

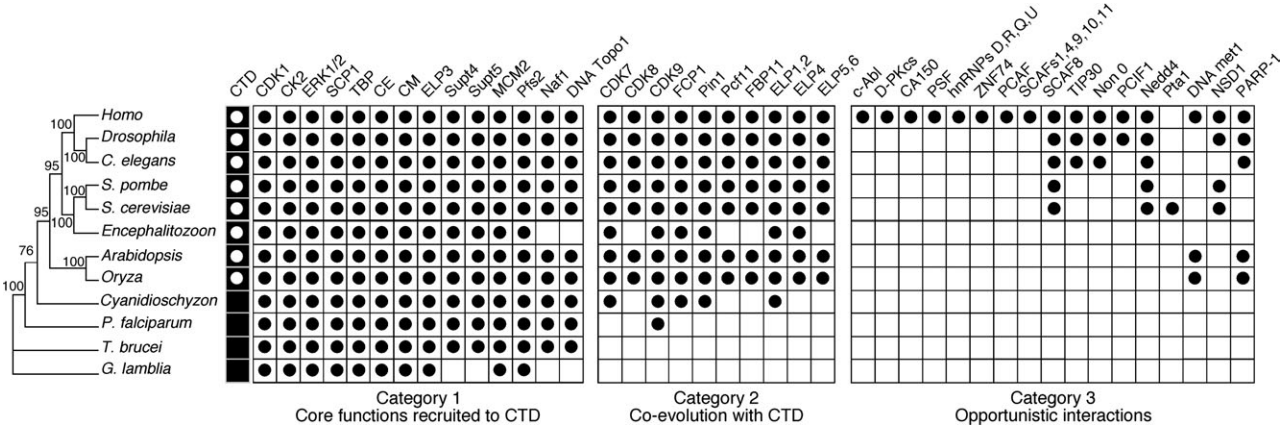

FIG. 1.—The apparent distribution of the 50 CTD-associated proteins in genomes of 12 organisms compared to the pattern of conservation of the RNAP II CTD. Some of these proteins are grouped based on their combined interactions with the CTD, resulting in 41 columns. The phylogenetic tree of the organisms was recovered from sequences of RPB1. White dots in the first column indicate conservation of the CTD in that organism and in its broader associate taxonomic lineage when known. Black dots in all other columns indicate that the specific sequence was recovered from the respective genome. The three designated categories for patterns of distribution among these proteins are discussed at length in the text.

the evolutionary distribution and conservation of CTD-associated proteins and their relationship to the differential pattern of conservation of the CTD across the eukaryotic domain. Here we report a survey of 12 completed or nearly completed eukaryotic genomes for putative homologs of all genes with products known to interact with the CTD in animals or yeast. This includes eight genomes from lineages in which the CTD is conserved strongly and four from organisms outside these groups. These sequences are examined for specific evidence of correlated evolution that can help to explain the phylogenetic distribution of conserved CTD heptads and provide insights into the overall evolution of the RNAP II transcription factory.

Materials and Methods

Database Searches

Twelve complete or nearly complete eukaryotic genomes were investigated, including animals (*Homo sapiens* [Hs], *Drosophila melanogaster* [Dm], and *Caenorhabditis elegans* [Ce]), green plants (*Arabidopsis thaliana* [At] and *Oryza sativa* [Os]), yeasts (*Saccharomyces cerevisiae* [Sc] and *Schizosaccharomyces pombe* [Sp]), microsporidia (*Encephalitozoon cuniculi* [Ec]), red algae (*Cyanidioschyzon merolae* [Ce]), and three parasitic protists (*Plasmodium falciparum* [Pf], *Trypanosoma brucei* [Tb], and *Giardia lamblia* [Gl]).

The proteins included in this study are shown in figure 1 with details provided in Supplementary Table 1 (Supplementary Material online). TBlastN, BlastP, and PSI-Blast searches for these sequences (Altschul et al. 1997) were undertaken at the National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov/blast) using NCBI databases at default parameters, including the nonredundant protein sequence database. Searches employed an *E* value inclusion threshold of 0.01, and composition-dependent statistics were used with PSI-Blast (Schaffer et al. 2001). Additional Blast searches used organism-specific sites to help in identifying potentially divergent homologs by reducing the size of target databases and increasing optimization of search parameters for a given genome; these included http://merolae.biol.s.u-tokyo.ac.jp/blast/blast.html (for *C. merolae*), http://www.plasmodb.org/plasmodb/servlet/sv?page=blast (for *P. falciparum*), http://www.genedb.org/genedb/tryp/blast.jsp (for *T. brucei*), and http://gmod.mbl.edu/perl/site/giardia?page=intro (for *G. lamblia*).

Both yeast and human sequences were used in separate queries to increase the probability of identifying divergent orthologs and to confirm that the two sequences did not produce different nearest matches. To support putative homologies detected by the initial Blast searches, inferred amino acid sequences were used as queries in reciprocal TBlastN searches to verify that they preferentially retrieved the original query sequences. Each unannotated sequence recovered was also queried against established domain models using the conserved domain search in PSI-Blast (Marchler-Bauer and Bryant 2004) to assure that it most closely matched the protein family used to find it initially. Multiple sequence alignments (available upon request) using ClustalW (Thompson, Higgins, and Gibson 1994) were performed to further support putative homologies by confirming conservation of core protein domains and that Blast similarities did not simply reflect single-domain homologies, but rather extended across the broader range of the target sequence.

Sequence Comparisons and Phylogenetic Analyses

Protein pairwise distances were calculated with Protdist (Felsenstein 1989, PHYLIP version 3.573) under an invariable site + discrete Γ rate model (four discrete category estimate) and a Jones-Taylor-Thornton (JTT) substitution matrix with maximum-likelihood parameters calculated in Tree-Puzzle (Strimmer and von Haeseler 1996). Phylogenetic analyses on universally present proteins were performed using Bayesian inference when needed (MrBayes 3.0 b4, Huelsenbeck and Ronquist 2001). Four simultaneous Markov chains were run, also under an invariable + Γ rate model and a JTT substitution matrix. Four chains, one heated, were run for 10^6 generations, beginning with random a priori

trees. Trees were sampled from the posterior probability distribution every 100 generations. The first 10^5 generations were considered the required "burn-in," after which the tree probabilities clearly had converged on a stable range of values. They were excluded from analyses of Bayesian posterior probabilities; thus, a total of 9,000 trees were examined to determine the 50% majority-rule consensus tree and Bayesian support values. In addition, 1,000 distance (Protdist + Neighbor) bootstrap replicates were performed in PHYLIP version 3.573 (Felsenstein 1989), also using a JTT substitution matrix.

Homology Modeling of CE Structures

Protein tertiary structure of CE was predicted using the alignment mode in Swiss-Model (http://swissmodel.expasy.org//SWISS-MODEL.html), an automated protein homology modeling server (Schwede et al. 2003). In producing alignments for inference of structural homology, CE sequences from each of the 12 genomes examined were specified as respective target sequences and the sequence of RNA GTase (Cgt1) from *Candida albicans* as the corresponding template (accession number P78587, Protein Data Bank (PDB) code lp16). The server then built the best tertiary model for each of the 12 sequences, and the structure coordinate files of the predicted models were visualized, analyzed, and manipulated using the DeepView program (Swiss-PDB Viewer). Three-dimensional structural images were produced and homologous positions (based on the best iterative alignment) to those known to interact with the CTD in *C. albicans* were identified on each tertiary model.

To access reliability and quality of predicted models, WHATCHECK reports (WHAT IF package: Vriend 1990; Hooft et al. 1996) were obtained from the Swiss-model server. As a further control for prediction reliability, we used the *C. albicans* Cgt1 structure to predict the *Chlorella* virus GTase structure, and then compared the predicted model to the experimentally determined structure, accession number Q84428, PDB code 1CKM (Håkansson et al. 1997). All predicted models were analyzed by distance-matrix alignment (DALI: Holm and Park 2000) to ascertain the similarity of each to the *C. albicans* template; values were determined for Z score (the strength of similarity of the best domain-domain alignment), the number of structurally aligned residues, the root mean square deviation (RMSD) over *lali* equivalenced C-alpha atoms in rigid-body superimposition, and sequence identity over the aligned positions.

Models generally were good fits to the template based on WHATCHECK statistics (Supplementary File 1, Supplementary Material online), and root mean square differences between predicted and template structure (Supplementary Table 2, Supplementary Material online) were all within the range of empirical comparisons of known protein homologs (Chothia and Lesk 1986). The lP16 template was able to resolve the core domains of *Chlorella* virus GTase; however, based on RMSD scores the match between predicted model and template was closer than between the two experimentally determined structures (Supplementary Table 2, Supplementary Material online), and some differences can be seen between the predicted and actual structures (Supplementary Fig. 1, Supplementary Material online). This reflects the limitation of the RMSD calculation, which compares only protein regions in which the general fold of the polypeptide chains is similar (common core); regions where they differ substantially are not reflected in RMSD scores (Chothia and Lesk 1986). For these reasons, homology models may not fully predict the tertiary structure of a given protein, but they do provide clear measures of how well each sequence can be fit to the known structure of CE from *C. albicans*, a member of the "CTD clade."

Results and Discussion

Overall Genome Content and Relationship to Conservation of the CTD

Proteins known to have biochemical or genetic interactions with the RNAP II CTD were collected through a systematic survey of the literature. We identified 50 proteins or protein complexes that have been reported to bind or interact with the CTD in humans and/or yeast. A number of these interact as groups, resulting in a total of 41 targets for our comparative analyses (fig. 1). This list is necessarily incomplete as additional CTD-protein interactions continue to be identified (e.g., Phatnani, Jones, and Greenleaf 2004). Detailed descriptions, including nomenclature and functional information, are provided in Supplementary Table 1 (Supplementary Material online). Database searches were used to recover potential homologs from all 12 genomes studied. Cyclin-dependent kinases (CDKs) represent a large family of proteins with a complex evolutionary history and were the subject of a prior detailed investigation to estimate orthologous relationships (Guo and Stiller 2004). Other than CTD-directed CDKs and with the exception of Naf1 from *Plasmodium* (see *Discussion* below), all other searches either returned a clear nearest match well above our designated threshold (poorest E value was capping enzyme guanylyltransferase (CEG) from *T. brucei* at 6×10^{-4}) or produced no match close to the $E = 0.01$ cutoff value. Putative homologies were further examined as described under *Materials and Methods*, and the phylogenetic distribution of each sequence was compared to the pattern of CTD conservation; the results are summarized in figure 1. Based on these comparative analyses, we grouped proteins associated with the CTD into three loose categories: (1) proteins that appear to be conserved across the broad range of eukaryotes, with or without a conserved CTD; (2) those with a pattern of distribution similar to that of the conserved CTD; and (3) those identified in only one or a narrow range of genomes.

Category 1: Universal Proteins

Among the CTD-associated proteins analyzed, 10 were found universally in the 12 eukaryotic genomes examined (fig. 1). Four others, Supt4, Supt5, Naf1, and DNA Topo 1, appear to be missing from one or several genomes but with no obvious correlation to degeneration or loss of the RNAP II CTD. Supt4 and Supt5 form a conserved protein complex known as DRB sensitivity inducing factor in humans and are associated with RNAP II transcription elongation (reviewed in Hartzog, Speer, and Lindstrom 2002). Previous characterizations of these proteins have suggested that homologs of Supt5 also are present in most prokaryotes, based on significant similarity to NusG sequences in Bacteria and Archaea;

however, no Supt4 homolog has been identified in prokaryotes (Ponting 2002; Hartzog 2003). In our Blast searches, putative homologs of both Supt5 and Supt4 were found in all eukaryotic genomes, with one exception; no sequence with significant similarity to either protein was found in *Giardia*. Likely explanations for the absence of Supt5 are either a high level of sequence divergence of a *Giardia* homolog from other eukaryotes or complete loss of the gene from *Giardia*. Either explanation could also apply to Supt4; however, *Giardia* has been argued to be among the earliest diverging eukaryotic lineages, displaying putatively transitional stages in its transcriptional systems interpreted as "prokaryotic properties" (Best et al. 2004). In this light, an absence of Supt4 in *Giardia* could represent the ancestral eukaryotic condition; genomic sampling of other excavate taxa (Dacks et al. 2001) will help to determine which of these explanations is most likely.

DNA Topo I also was not detected in *Giardia* or in the microsporidian *Encephalitozoon*; the latter has been argued to be closely related to fungi (Baldauf et al. 2000; Keeling and Fast 2002; fig. 1). Strongly conserved homologs of this topoisomerase were recovered from all other eukaryotes. Eukaryotic DNA Topo I belongs to a different subfamily of enzymes from eukaryotic DNA Topo III and bacterial Topos I and III (Corbett and Berger 2004). Blast searches of all bacterial and archaeal sequences available in GenBank, using human DNA Topo I as the query, recovered no significant matches. Reciprocal searches, querying all eukaryotic genomes with both *Escherichia coli* DNA Topos I and III, returned significant matches to homologs of human DNA Topo III but not to the putative DNA Topo I included in our investigation. An absence of DNA Topo I from *Giardia* could be interpreted as further evidence of this organism's ancestral position among eukaryotes; however, the fact that DNA Topo I was not recovered from *Encephalitozoon* suggests that it could have been lost independently from both of these highly reduced and modified parasites.

Likewise, no potential homolog of Naf1 (Dez et al. 2002) was found in *Encephalitozoon* or *Giardia*. One sequence from *P. falciparum* was retrieved at $E = 0.002$, above our cutoff for initial Blast screens; however, a conserved domain search using this sequence identified it as containing the Gar1 domain, and reciprocal PSI-Blast and TBlastN searches recovered Gar1 homologs from the other organisms in our study with E values between 2×10^{-20} (*Arabidopsis*) and 2×10^{-9} (*Giardia*). Gar1 is a component of box H/ACA small nucleolar ribonucleoprotein particle, and Naf1 was identified originally based on core domain similarities to Gar1 (Fatica, Dlakic, and Tollervey 2002). No other significant match for Naf1 was found in the *P. falciparum* genome, suggesting that its homolog is absent or has diverged even beyond the point of the more distantly related Gar1 protein family.

Although some of these category 1 "universal" proteins were not found in one or several highly divergent parasites, they appear to be present in all other completed eukaryotic genomes examined (fig. 1), including those outside the CTD clade. Consistent with this strong conservation, most of these proteins function in processes that are central to the biology of the cell (Supplementary Table 1, Supplementary Material online). Considering their core functions and widely conserved distributions, these CTD-associated proteins must have existed before the origin of the CTD or at least before its primary structure of tandemly repeated canonical heptads came under intense stabilizing selection. When the CTD and its integrated function in RNAP II transcription emerged, these proteins appeared to have been incorporated into CTD-based transcription and processing centers. Once this occurred, they likely came under different functional constraints as they coevolved with the CTD and other proteins in the transcriptosome, perhaps even becoming codependent over the course of their common evolutionary history. A further investigation of such evolutionary coconservation with the CTD is presented and discussed in a separate section below.

One additional noteworthy result of our investigation of this category of proteins involves Pfs2, which is a critical factor for 3′ end mRNA processing in yeast (Zhao, Hyman, and Moore 1999). CstF50 in humans was proposed to be homologous to Pfs2 based on their comparable functions and the common presence of WD repeat sequences (Gross and Moore 2001). However, another WD repeat protein (WDC146) from humans is recovered as the sequence with greatest similarity to Pfs2 (with E value 1×10^{-70}) in our Blast searches. In contrast, human CstF50 is not detected within the cutoff E value of 0.01. Furthermore, when we used human CstF50 in a reciprocal Blast search, yeast Pfs2 was not recovered. We also detected apparent orthologs of Pfs2 in all organisms sampled in this study, whereas CstF50 was found only in animals and plants. WDC146 has been implicated in cytodifferentiation and/or DNA recombination, but no function in pre-mRNA cleavage has been identified thus far (Ito et al. 2001). Combined with results from the mechanistic studies discussed above, our genome comparisons suggest that the functional homologs of CstF50 in humans and plants are not the evolutionary homologs of yeast Pfs2.

Interestingly, this is the same evolutionary pattern of orthology seen in eukaryotic CEs (Shuman 2002). The fungal CE comprises a separately encoded triphosphatase (TPase) and GTase; however, metazoan and plant CEs consist of an RNA TPase fused to a GTase. Furthermore, the primary structure and mechanisms of CE TPase in animals and plants are quite different from those in fungi and other eukaryotes (Shuman 2002; Hausmann et al. 2005). Broader scale comparative genomics suggest that these unique architectures of Pfs2 and CE, shared between animals and green plants, reflect a generally greater similarity of proteins involved in RNA metabolism in these two groups (Anantharaman, Koonin, and Aravind 2002).

Category 2: Sequences That Roughly Correlate with a Conserved CTD

Twelve proteins or complexes have the same or very similar apparent phylogenetic distributions as the conserved RNAP II CTD (fig. 1). Consistent with this similarity, all these proteins are specifically related to transcription or mRNA processing. At least several, CDK7, CDK8, and Fcp1, appear to be CTD specific in their functions (Licciardo et al. 2001; Prelich 2002) (see details and references in Supplementary Table 1, Supplementary Material online).

The general level at which this group of proteins correlates with a conserved CTD is illustrated by the family of CTD-directed CDKs (fig. 1). Both CDK7 and CDK9 were identified in all organisms containing a canonical RNAP II CTD but were generally not found in most organisms outside that group. For example, although the CTD appears to be under relaxed selection and has been allowed to degenerate in most red algae (Stiller and Hall 1998), *Cyanidioschyzon* contains an apparent ortholog of CDK7. A CDK7 homolog also was proposed for *Giardia* based on its nearest similarity to yeast CDK7 in a comparison among CDKs of the two species (Liu and Kipreos 2002); however, neither this sequence nor any other from *Giardia*, *Trypanosoma*, or *Plasmodium* shows affinity for CDK7 in more broad-scale phylogenetic analyses of CDKs (Guo and Stiller 2004). In contrast, the putative CDK7 ortholog from red algae branches at the base of the CDK7 family in updated phylogenetic analyses (unpublished data).

Based on both Blast similarities and detailed phylogenetic analyses (Guo and Stiller 2004), apparent orthologs of CDK9 are present in both *Cyanidioschyzon* and *Plasmodium*. In contrast, the third CTD-directed kinase, CDK8, appears to be restricted to organisms in which the CTD is strongly conserved (see first column in fig. 1); however, it is not found in the microsporidian parasite *Encephalitozoon*, which does have a conserved CTD and nests tightly within the CTD clade (fig. 1; see Stiller and Cook 2004). Thus, the presence or absence of CTD-directed CDKs is not sufficient to explain differences in conservation of the RNAP II CTD among different eukaryotic taxa.

The remaining proteins in this category show similar patterns of loose but imperfect correlation with the conserved CTD. In fact, of the entire set of protein sequences analyzed in this study, the only one that strictly correlates with a conserved CTD is ELP4, a component of the Elongator complex (Kim, Lane, and Reinberg 2002; Shilatifard, R. C. Conaway, and J. W. Conaway 2003). Other protein sequences that make up Elongator were identified in all eukaryotes (ELP3) or at least in red algae (ELP1–2), and two components (ELP5 – 6) were not detected in *Encephalitozoon* (fig. 1). One of the most interesting aspects of this category of sequences is that, with the exception of ELP4, the same sequences appear to be present/absent in red algae and the microsporidia. The microsporidia are an unusual group of organisms; they have the most highly reduced genomes of all eukaryotes and an extreme reduction of their core molecular machinery (Peyretaillade et al. 1998; Keeling and Fast 2002). For example, the small subunit of the ribosome has lost all but its most essential components and is considerably smaller than the 16S subunit present in prokaryotic cells (Vossbrinck et al. 1987). And yet, despite their accelerated evolutionary rates and the apparent selective pressure to miniaturize their molecular machinery, all microsporidians examined to date have a well-conserved RNAP II CTD.

It is tempting to use the microsporidia as a baseline for determining which CTD-protein interactions are indispensable and therefore directly responsible for conferring a strong stabilizing selection on CTD structure; however, the similar distribution of CTD-related proteins in red algae makes this straightforward argument untenable. Rhodophytes display relatively slow rates of molecular evolution (Stiller, Riley, and Hall 2001), yet show no evidence of stabilizing selection on CTD primary structure (Stiller and Hall 1998). Thus, our comparative data suggest that at least two factors have been important in shaping the current distribution of CTD-related proteins in eukaryotes: one is stabilizing selection on certain key CTD-protein functions or on protein-protein interactions indirectly associated with the CTD (see below) and the other, evolutionary forces that lead to accelerated overall rates of genome evolution. The antithetical effects of these factors may confound attempts to draw any direct correlation, at least from comparative genomic data, between the presence of a conserved CTD and a clearly defined set of core functions that define a CTD-based transcription system. The interplay between shared functional constraints and differences in substitution probabilities also has important implications for deep-level phylogenetic analyses, which are addressed in a separate section below.

Category 3: Proteins with Opportunistic CTD Interactions

We identified 24 CTD-related proteins that exhibit a limited distribution in only one or several genomes. Many of these sequences were identified in humans through extensive biochemical searches for "phospho-CTD–associated proteins" (PCAPs) (Carty and Greenleaf 2002). Consequently, with the exception of Pta1 (detected only in budding yeast), all these proteins are found in the human genome; a number of them also were identified in one or both of the other animals examined (fig. 1). Five of the sequences have somewhat contradictory phylogenetic distributions. Poly(ADP-ribose) polymerase 1 (PARP-1) was found in all animals and plants and DNA Met1 in humans and plants. In contrast, Scaf8, Nedd4, and NSD1 appear restricted to animals and fungi (plus *Encephalitozoon*). These conflicting distributions are symptomatic of a phylogenetic bipolarity of evidence regarding relationships among animals, plants, and fungi (Baldauf and Palmer 1993; Stiller 2004). Molecular evidence tends to support a relationship between animals and fungi or animals and plants but seldom between plants and fungi.

The limited apparent distributions of proteins in this category suggest that most have functions that are specific to a restricted group of eukaryotes in which they are found. For example, some are related to human diseases; P300/CBP-associated factor (PCAF) is a coactivator of the tumor suppressor P53 (Cho et al. 1998), and TIP30 is involved in human immunodeficiency virus regulation (Xiao et al. 1998). As noted earlier, Carty and Greenleaf (2002) identified many of the sequences that are unique to humans based on their highly specific binding affinities for the phospho-CTD. Recent comparable analysis of the yeast proteome resulted in a similar yield of PCAPs (Phatnani, Jones, and Greenleaf 2004), and investigations of other CTD-containing eukaryotes, such as *Arabidopsis*, undoubtedly will uncover specific PCAP interactions in these organisms as well. Their limited phylogenetic distributions and more narrow functions suggest that many of these interactions are opportunistic, originating after the CTD was

canalized at the core of RNAP II transcription. Once present, they may help to maintain stabilizing selection on CTD structure; however, as a group they clearly cannot be invoked as a primary explanation for conservation or degeneration of the CTD across the broad diversity of eukaryotes.

Evolution of the CTD and Its Related Proteins

As discussed in *Introduction*, the primary motivation for this investigation was an exploration, through comparative genomics, of the differential pattern of conservation of the RNAP II CTD across the eukaryotic domain. A canonical CTD (tandem repeats of consensus sequence YSPTSPS) is conserved strongly in only a subset of eukaryotic groups; all descended from a common ancestor in phylogenetic trees based on RPB1 sequences (Stiller and Hall 2002; Stiller and Cook 2004). If this CTD clade represents a monophyletic evolutionary lineage, the origin of a distinct set of coadapted CTD-protein interactions, perhaps the transcriptosome itself, could be responsible for conferring strong stabilizing selection on CTD structure. In this case, we might expect to recover a core group of CTD-protein interactions common to all members of the CTD clade but absent from groups in which the CTD has degenerated. Although the general correlation of a conserved CTD with our category 2 proteins (fig. 1) is suggestive of such a coalescent event, no definitive core set of interactions emerged from our investigation.

Determining what set of protein interactions may be responsible for evolutionary stabilization of CTD structure is complicated by the fact that parasitic organisms have been the primary focus of protistan genomics. These parasites tend to have much higher probabilities of molecular substitution, which can lead to erroneous results from phylogenetic, Blast, and other informatic approaches to pairwise sequence comparisons. For example, in our prior investigation of CTD-directed kinases, we uncovered a pattern of correlated evolution of CDKs with the CTD (Guo and Stiller 2004). It is not clear, however, whether this correlation represents shared evolutionary history, shared functional constraints, or a combination of the two. Both *CDK*7 and *CDK*8 have roughly the same apparent distribution as the conserved RNAP II CTD (fig. 1). No clear ortholog of either was found in the genomes of "non-CTD" parasites *Giardia*, *Trypanosoma*, or *Plasmodium* (fig. 1). Nevertheless, all these genomes contain sequences that are not easily assigned to any *CDK* family. Some of them could be orthologs of *CDK*7 or 8 that have undergone extreme sequence divergence (similar to the degenerated CTDs of these organisms) and which prevents them from positioning properly in phylogenetic analyses of CDK sequences (Guo and Stiller 2004).

In this investigation, a number of CTD-related proteins appear to be missing from the same three parasitic taxa. There are three explanations for this finding: first, these organisms diverged from the eukaryotic tree before the origin of each of the respective CTD-related proteins; second, the proteins are still present but accelerated rates of evolution mask their orthologous relationships to sequences of more slowly evolving eukaryotes; and finally, the proteins were once present but have been lost from these genomes completely, perhaps due to the absence of strong stabilizing selection conferred by coadapted evolution with the CTD. The similar distribution of category 2 sequences in the red alga (relatively slow rate of molecular evolution, but no conservation of CTD structure) and microsporidia (fast molecular evolution, but with a strongly conserved CTD) suggests that all these factors could help to explain the imprecise correlations we found between the CTD and its related proteins.

Because category 2 proteins could not be identified in most "non-CTD" genomes, further comparative analyses of coevolution with the CTD cannot be investigated with these sequences. The same is not true for CTD-related proteins in category 1, which generally were retrieved from all available genomes. Therefore, we examined this set of sequences for evidence of coconservation with the RNAP II CTD.

Correlated Conservation of the CTD and Its Related Proteins

Homologs of CDK1 appear to be present in all eukaryotic genomes, sometimes as multiple copies, but reliable orthologous relationships have not been established (Guo and Stiller 2004). CE TPase in animals and plants is not orthologous to capping enzyme triphosphatase (CET) from fungi and other eukaryotes (Shuman 2002); as noted above, we found the same to be true of CstF50/Pfs2 sequences. In addition, homologs Naf1 and DNA Topo I were not detected in several genomes (fig. 1). Consequently these four proteins were excluded from our more detailed analyses of category 1 sequences.

To examine whether overall evolutionary rates appear to correlate with a conserved CTD, a matrix of maximum-likelihood distances was calculated for each of the 10 putatively orthologous proteins identified (fig. 2*A*). Similar to the pattern found for presence/absence of category 2 proteins, there is some evidence of a generally slower rate of evolution in organisms with a strongly conserved CTD. Just as for presence/absence data, however, the correlation is complicated by apparently accelerated sequence evolution in all four parasitic taxa. Although most sequences of the non-CTD parasites *Giardia*, *Trypanosoma*, and *Plasmodium* are less conserved than those of CTD-clade organisms, so are nearly all sequences from *Encephalitozoon*. In contrast, sequences from the non-CTD red alga *Cyanidioschyzon* generally show comparable rates of evolution to those in CTD-clade genomes (fig. 2*A*). Thus, an increase in the overall rate of primary sequence evolution in CTD-related proteins appears to correlate more with a parasitic lifestyle than with conservation of CTD structure.

There are interesting differences in the patterns of sequence conservation among the 10 proteins examined. Abd1, Spt4, Elp3, TBP, and MCM2 are strongly conserved in all CTD-clade genomes, including *Encephalitozoon*, but more highly divergent in the three protistan parasites. In contrast, Erk1, SCP1, and CK2 have more constant mean divergent rates in all organisms. This suggests that Abd1, Spt4, Elp3, TBP, and MCM2 may be coevolving with the CTD, under comparable functional constraints imposed by interactions during the RNAP II transcription cycle. In contrast, the rates of evolution of Erk1, SCP1, and CK2 do not

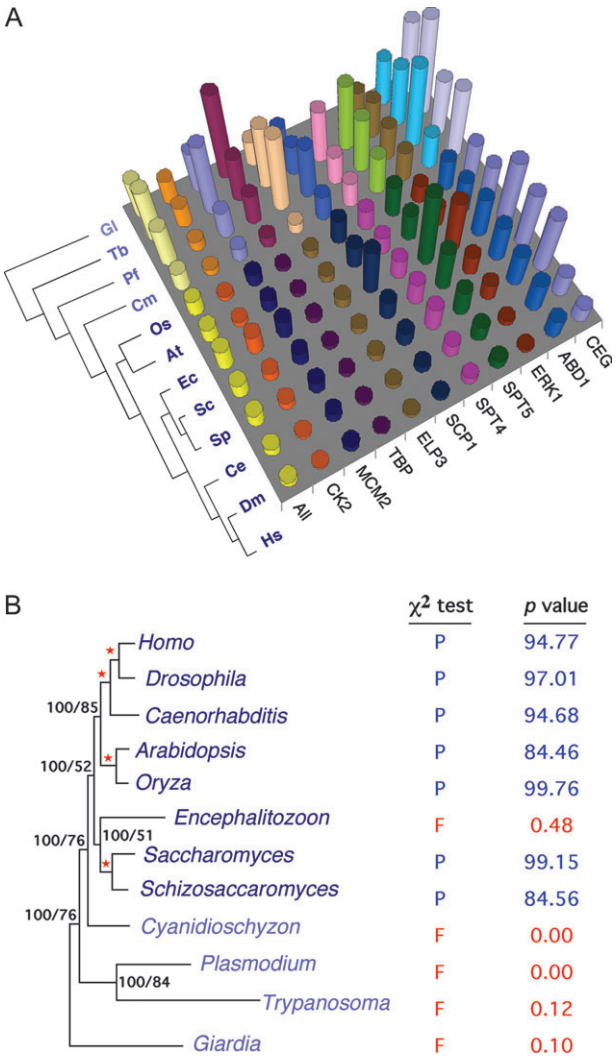

FIG. 2.—Sequence evolution of CTD-related proteins from category 1. (*A*) The matrix of maximum-likelihood distances from 10 of category 1 proteins calculated with Protdist (PHYLIP version 3.573) under an invariable site + discrete Γ rate model (four discrete category estimate) and a JTT substitution matrix with maximum-likelihood parameters calculated in Tree-Puzzle. For convenience, divergences are plotted against the human ortholog, which accounts for the small relative divergence of sequences from the other two animals. The non-CTD organisms are shown in lighter shades to distinguish from the CTD-clade organisms. (*B*) Consensus tree based on nine category 1 proteins recovered from Bayesian and maximum-likelihood analyses, with results of a χ^2 test for deviation from average amino acid composition. Support values for each node from Bayesian inference and distance bootstrap are shown on the left and right of the slash, respectively. A star designates where both values are 100%. Organisms with sequences that failed χ^2 test are shown in red.

appear to correlate at all with conservation of the CTD (fig. 2*A*). This is reasonable considering the transcription-specific functions of the former five proteins, as opposed to more universal functions of the latter three (see Supplementary Table 1, Supplementary Material online). It also suggests that interactions of the CTD with Erk1, SCP1, and CK2 may not be conserved across all members of the CTD clade.

The general evolutionary rates for *Cyanidioschyzon* sequences are similar to those for members of the CTD clade. In nearly all cases, the category 1 ortholog from red algae is less divergent from humans than is the ortholog from *Encephalitozoon* (fig. 2*A*), although the latter is believed to have a closer evolutionary relationship to the animal kingdom (Baldauf et al. 2000; Keeling and Fast 2002). Despite the generally increased rates of molecular evolution in parasites, which predominate over any stabilizing selection conferred by interactions with the CTD, some evidence of correlated evolution appears to be present. A Bayesian phylogenetic tree recovered from the nine category 1 proteins also recovers the CTD clade (fig. 2*B*). This suggests one of two things: these CTD-related proteins and RPB1 display the same actual historical pattern of evolution or both have fallen under parallel functional constraints that result in comparable but incorrect tree topologies. Two factors suggest that functional constraints rather than historical signal may be driving both tree topologies. First, despite the relatively slow rate of overall divergence of red algal sequences (fig. 2*A*), they deviate significantly from mean amino acid composition in a χ^2 test (fig. 2*B*). Second, the tree topologies recovered from both RPB1 and CTD-related proteins differ from widely accepted eukaryotic relationships based on phylogenetic analyses of other gene sequences (see Baldauf 2003 for a thorough review).

The correlated evolution suggested by our bioinformatics comparisons requires more detailed analyses to determine whether common functional constraints are, indeed, leading to potentially erroneous phylogenetic results. In particular, empirical evidence of conservation of direct interactions, which parallel evolutionary patterns recovered in comparative analyses, could provide clear evidence of coevolutionary functional constraint between the CTD and its related proteins. To date, three-dimensional structure and specific physical interactions with the CTD have been resolved for only one protein that is present universally in the genomes under investigation: it is, CE GTase.

Comparative Evolution of CE and the CTD

The primary sequences of CEs are relatively variable among all organisms in our investigation (fig. 2*A*). Their core functions, however, are highly conserved. Therefore, we undertook an investigation of structural conservation at both the primary and tertiary levels and whether they correlate with conservation of the CTD. The crystal structure of RNA GTase (Cgt1) from *C. albicans*, complexed with the RNAP II CTD heptapeptide repeats, has been solved to 2.7 Å (Fabrega et al. 2003). This provides a template for predicting CE structures in other organisms. We used homology modeling to examine how well CEs from the 12 organisms included in our study conform to the known *C. albicans* Cgt1 template structure.

Ultimately, reliable three-dimensional structures must be determined empirically for each protein. Nevertheless, our computational comparisons yielded provocative results. Primary structures, including residues forming known CTD docking sites (CDS), do not show a correlation of conservation with the RNAP II CTD; however, the overall fit of predicted tertiary structures do. Twenty specific CDS residues were determined from the crystal structure of *C. albicans* RNA Cgt1 complexed with the CTD (Fabrega et al. 2003). We aligned Cgt1 orthologs from *C. albicans*

and all organisms sampled in this study (fig. 3*A*); in most cases CDS residues lie near strongly conserved domains and can be aligned with reasonable confidence. This means that the relative locations of these residues within each predicted tertiary structure of Cgt1 are identifiable as well. The tertiary position of each CDS residue identified in *C. albicans* (Fabrega et al. 2003) is shown in figure 3*B*.

Despite the lack of evidence for correlated evolution between the CTD and specific CDS residues, homologs from all members of the CTD clade fit the known structure of *C. albicans* Cgt1 across the full lengths of their sequences. The same is not true, however, for non-CTD organisms. A clear overall correlation between conservation of tertiary structure and presence of a canonical CTD can be seen in figure 4. The predicted structures of all eight CTD-clade CEs can be superimposed simultaneously onto the *C. albicans* structure, with little evident deviation (individual superimposed structures are provided in Supplementary Fig. 2, Supplementary Material online). This includes the structure from the microsporidian *Encephalitozoon*, which is among the fastest evolving sequences at the primary level (fig. 2*A*). All four sequences from "non-CTD" organisms, including several that display lower apparent rates of primary sequence divergence than *Encephalitozoon* (fig. 2*A*), show deviations from the predicted three-dimensional structure (fig. 4). Additional loops, sheets, or longer unmodeled segments in each of these sequences prevent complete superimposition on the *C. albicans* three-dimensional model (fig. 4*A*) or onto each other (fig. 4*B*). Interestingly, the overall architecture of the regions of Cgt1 that contact CTD residues (Fabrega et al. 2003), as well as spatial arrangements of CDS residues themselves, is generally conserved in all organisms whether or not the RNAP II CTD is present (fig. 3*C*).

Despite the overall conservation of fit among predicted CE tertiary structures from CTD-clade organisms, there is no obvious connection between chemical properties of CDS residues and the presence of conserved CTD. On the sequence alignment showing CDS positions, only a few sites (R140 and L163 in fig. 3*A*) are even reasonably conserved. This is true universally, as well as among sequences from CTD-clade organisms. Moreover, with the possible exception of R140, there is no indication of a correlation between chemical properties of CDS residues and the presence of a canonical CTD (fig. 3*B*). In fact, despite the remarkable overall match of predicted Cgt1 tertiary structures among CTD-clade organisms (fig. 4), at least one CDS residue from *C. albicans* (D175) appears to have been lost in animals (fig. 3).

The suggestion of stronger tertiary conservation of CE in members of the CTD clade is intriguing, particularly given the lack of conservation of individual CDS sites known from *C. albicans*. Both genetic and structural evidences indicate a large amount of flexibility in binding of the CTD by associated proteins (Verdecia et al. 2000; Fabrega et al. 2003; Greenleaf 2003; Meinhart and Cramer 2004; Stiller and Cook 2004). Conformation of heptapeptides bound in solution by yeast Pcf11 show that the phospho-CTD does not present a preformed structure, but rather is bound by an induced fit to the specific structure of the Pcf11 docking site (Noble et al. 2005). This flexibility is consistent with the requirement of the CTD to bind a diverse array of proteins.

Ironically, the need for remarkable flexibility may actually help to explain strong conservation of the canonical CTD sequence. Large numbers of individual substitutions or certain individual changes could cause the CTD to take on an ordered tertiary structure, even when it is not bound to one of its protein partners. Such a preordered structure would reduce or eliminate CTD flexibility and could prevent induced fit binding to one or more CTD-related proteins. A requirement for induced binding to many different proteins may account for strong selection on a tandemly repeated YSPTSPS sequence, which can be phosphorylated and dephosphorylated without losing structural flexibility. Thus, presence of significant deviations from this sequence in some organisms could reflect reduced CTD-binding requirements and therefore reduced selection on CTD flexibility.

What Accounts for Differential Conservation of the CTD?

The overall results of our bioinformatic investigations of CTD-related proteins and CE specifically indicate that mechanical flexibility extends to the overall evolution of CTD-protein interactions. There is general evidence of evolutionary conservation correlated with the presence of a canonical CTD; however, the specific bases of this correlation are not clear. The only CTD-related protein that appears strictly correlated with the canonical CTD is ELP4, a component of Elongator; but other components of Elongator also are found in *Cyanidioschyzon* or, in the case of ELP3, in other non-CTD organisms as well. Although it would seem too simplistic an explanation, it is possible that the coalescence of Elongator as a CTD-related complex was the final piece of the puzzle responsible for conferring strong stabilizing selection on CTD structure. A determination of the roles of ELP1–3 in non-CTD organisms could verify whether a shift in their functions accompanied strong evolutionary conservation of the CTD.

Given the apparent flexibility of CTD-protein interactions, both functionally and through time, we think the explanation for evolutionary conservation of the CTD is unlikely to be the canalization of one particular function. It is far more likely that the CTD came under strong stabilizing selection only when its cumulative role in RNAP II transcription and processing reached some critical mass of coadapted functions. Once so fully integrated into the RNAP II transcription cycle, strong selection on these coadapted functions, as a group, would have prevented the CTD from breaking down. There is confirmatory evidence for this hypothesis from observations that higher order collective interactions increase the efficiency of protein functions using the CTD as a docking platform (Noble et al. 2005). This suggests that the structure of the CTD is strongly conserved only when it lies at the core of a transcription factory or transcriptosome, orchestrating a large group of interdependent protein-protein interactions. Elaborations may be added in some organisms and individual parts lost in others, but the infrastructure of the factory must remain intact to confer viable RNAP II transcription.

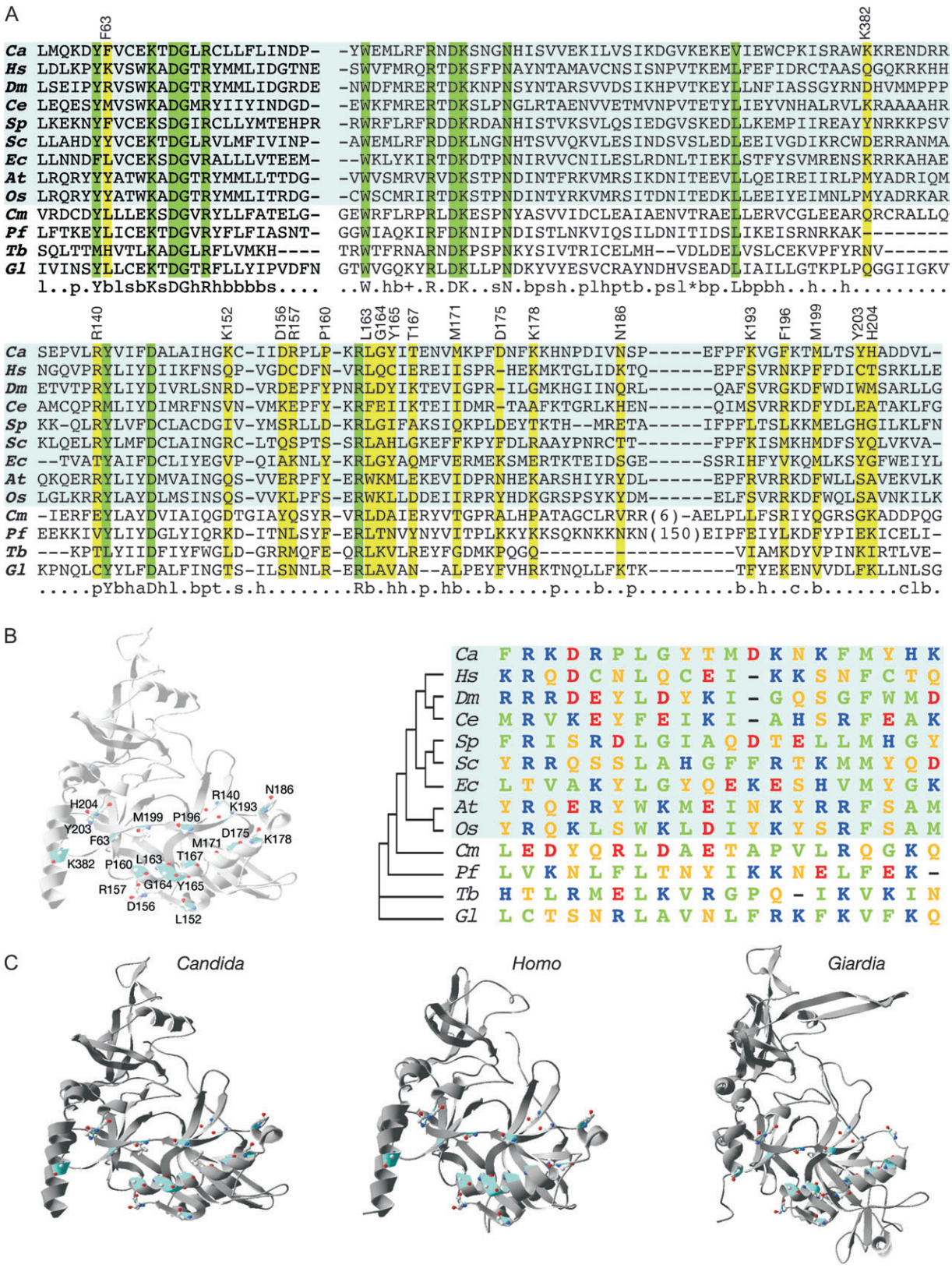

FIG. 3.—Comparative analyses of Cgt1 orthologs. (*A*) Multiple sequence alignment of core domains of *Candida albicans* RNA GTase (Cgt1) and orthologs from all organisms sampled in this study using CHROMA and a 75% consensus. The CTD-clade organisms are shaded light blue. Conserved residues are shown in green and the CDS in yellow. The positions and identities of each CDS annotated on the alignment are based on *C. albicans* Cgt1. (*B*) The tertiary position and chemical properties of each CDS residue identified in *C. albicans* Cgt1. The CTD clade is shaded light blue. Designation of general animo acid properties are as follows: green represents nonpolar, blue represents basic, red represents acidic, and orange represents uncharged polar. (*C*) The predicted structures of human and *Giardia* GTase, with CDS sites highlighted, compared to the *C. albicans* Cgt1 template structure.

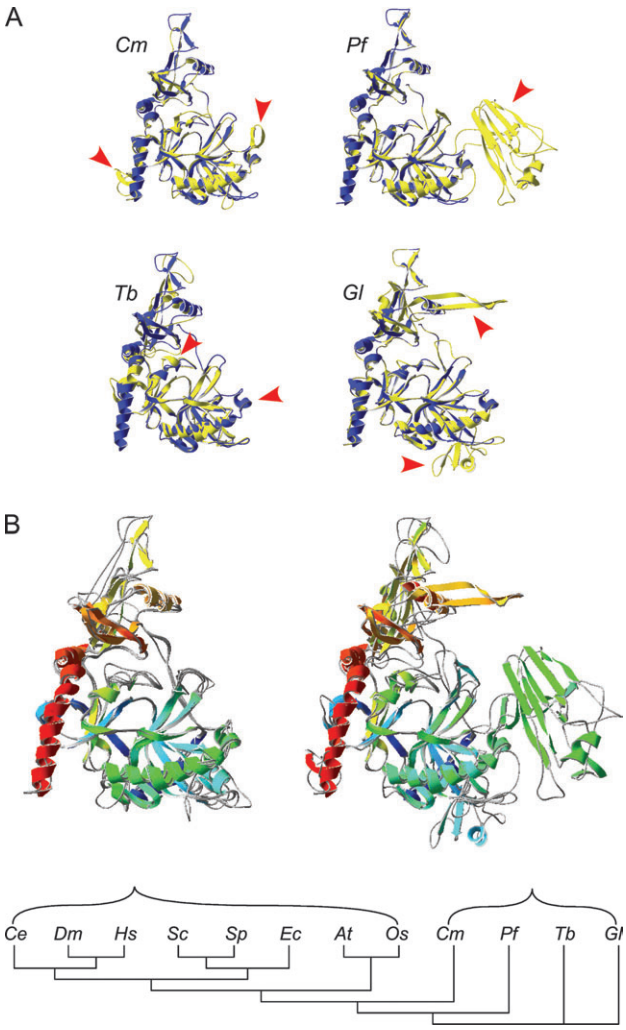

FIG. 4.—Cgt1 structures mapped to the *RPB1* phylogeny. (*A*) Superimposed tertiary structure of *Candida* Cgt1 and predicted structures from each of the four non-CTD–clade organisms included in our study. *Candida* Cgt1 is in blue and the predicted structure from specific organism is in yellow. (*B*) On the left are superimposed structures from all eight CTD-clade organisms onto *Candida* Cgt1, and on the right are superimposed structures from the four non-CTD–clade organisms onto *Candida* Cgt1.

The hypothesis that the presence of a coadapted transcriptosome is responsible for maintaining a canonical CTD is consistent with the results of our comparative analyses of the evolution of CE, that is, the apparent conservation of overall Cgt1 tertiary structure without simultaneous evidence for conservation of CTD docking residues. What is most interesting is that the conservation of the structure correlated with the presence of the CTD is not in CDS regions or in core function domains, but rather in apparent surface regions that might encounter other transcription and processing factors (fig. 4). Thus, direct interactions between the CTD and CE do not appear to drive their correlated conservation but, rather, the fact that both are evolving under the constraints imposed by a myriad of interactions within the transcriptosome. This hypothesis also leads to a straightforward and testable prediction: eukaryotic groups without a strongly conserved RNAP II CTD also should not contain complex transcription factories.

Red algae are an attractive group for testing this prediction. As a whole, they do not exhibit the unusual modes of sequence evolution typical of protistan parasites, and they contain many of the components that collectively form transcription factories when the CTD is present. If our hypothesis is correct, there should be no transcriptosome in red algae or in other eukaryotic groups with a degenerated RPB1 C-termini. Should this prove to be the case, additional puzzles regarding the evolution of the CTD will remain unanswered. What was the original function of tandemly repeated heptapeptides, prior to coalescence of transcriptosomes and why are large numbers of them present in a handful of protists that do not fall within the CTD clade? One possible answer to both questions could be that the original role of repeated heptapeptides was to increase the efficiency of cotranscriptional intron splicing. In groups where heptad repeats were not recruited to coordinate other transcriptional and nuclear functions, loss of introns would have led to relaxed selection on CTD structure. This explanation is consistent with the small amount of available information on organisms with tandemly repeated RPB1 heptapeptides outside the CTD clade (Stiller and Hall 2002), but far more data are needed to demonstrate that such a correlation exists.

Finally, it is important to reiterate that inferences of homology, both of primary sequences through Blast and other search and alignment tools and tertiary structures through homology modeling, must be confirmed by experimental evidence. This seems particularly true for CTD-protein interactions, which may be difficult to predict across even closely related organisms. A further investigation of the possible effects of functional constraints on computational inferences reinforces this caution.

Implications for Phylogenetic Analyses at Deep Levels

As noted above, the phylogenetic inference of a CTD clade, including green plants but excluding red algae, is inconsistent with results emerging from broad-scale sequence-based molecular phylogenies (Baldauf 2003). Initially, strong support for an earlier evolutionary divergence of the Rhodophyta in RPB1 phylogenies was considered to be a significant problem (Delwiche and Palmer 1997) for the hypothesis of an all-inclusive Kingdom Plantae (Rhodophyta, Viridiplantae, Glaucocystophyta). This problem has been dismissed as a phylogenetic artifact in RPB1 sequences by most recent reviewers of algal and plant evolution, and at present, the view that red algae and green plants are sister groups is widely accepted; it even appears in most recent biology textbooks (see, for example, Campbell and Reece 2005).

If the hypothesis of a Superkingdom Plantae is correct, then the recovery of a monophyletic RPB1 clade, containing all groups with a strongly conserved RNAP II CTD, must be a phylogenetic "artifact." Such an artifact could be the result of functional constraints placed on the entire RPB1 molecule and perhaps the RNAP II holoenzyme by the presence of the CTD and stabilizing selection on its interactions with multiple protein partners. The common presence of so many coadapted functions could constrain the number and types of individual substitutions permitted in the RPB1 molecule, leading to parallel or convergent

primary sequence evolution in all groups with CTD-based RNAP II transcription. In other words, recovery of a CTD clade in RPB1 phylogenies could reflect parallel functions, rather than a shared evolutionary history.

If functional constraints on CTD-based transcription are causing RPB1 sequences to mislead tree-building algorithms, then our results suggest that the problem extends to proteins that interact with the CTD as well. The cumulative tree-building signal from these proteins also strongly favors the CTD clade in both Bayesian and maximum-likelihood phylogenetic analyses (fig. 2) and extensive analyses of CDKs reveal similar correlated patterns in trees produced from *RPB*1 and CTD-directed CDK sequences (Guo and Stiller 2004). Parallel functional constraints also could explain other incongruities found in comparative analyses of transcription-related genes and general conclusions from phylogenetic analyses of other sequences; for example, recovery of animals and plants as sister groups, with fungi as the outlier in analyses of CE, as well as general proteomics of RNA metabolisms (see fig. 2; Anantharaman, Koonin, and Aravind 2002; Shuman 2002; Stiller 2004). Parallel functional constraints in evolutionarily unrelated taxa can dominate deep phylogenetic trees (Stiller and Hall 1999) and could explain persistent and directional artifacts inferred from a number of investigations of ancient evolution.

The possibility that functional constraints dominate tree reconstruction algorithms raises a broader issue with respect to currently accepted views of ancient evolution. The inference that trees recovered from transcription-related sequences are due to phylogenetic artifacts can be made for two reasons. The first is the presence of detailed information on interactions between the CTD and its protein partners, which provides a mechanistic explanation for parallel constraint. It should be noted, however, that even with such detailed information, the specific constraints on primary sequence evolution are not necessarily evident. In the case of CE, there is a correlation between conservation of the CTD and fit of Cgt1 tertiary models to the solved structure from *C. albicans*; however, there is no obvious pattern of conservation of known CTD docking residues (fig. 3*B*) that should lead to recovery of a CTD clade in phylogenetic analyses of Cgt1 sequences (Stiller 2004).

The second basis for assuming phylogenetic artifacts in RPB1 trees is that they conflict with previously established relationships among major eukaryotic taxa. These prior hypothetical relationships, however, are themselves based almost exclusively on sequence-based phylogenetic analyses. In most cases, little or no detailed information on functional interactions that could affect tree reconstruction has been available. The problem is only exacerbated as data sets increase in size in large phylogenomic studies, where virtually no functional data are available for most of the genes and organisms under investigation. Moreover, computational methods for incorporating such complex patterns of sequence covariation into phylogenetic analyses are under investigation but not yet available (Lockhart and Steel 2005).

Thus, the argument that the trees shown in figures 1 and 2 in this study reflect phylogenetic artifacts becomes somewhat circular. It could also be argued that the CTD clade is a monophyletic evolutionary group and that unknown functional constraints have misled tree-building algorithms in other studies. In either case, the results of our investigation of CTD interactions indicate that incorporating data on functional interactions may be critical for determining the validity of tree topologies recovered from sequence-based phylogenetic algorithms.

Conclusion

The RNAP II CTD has proven to be a remarkably useful and flexible structure for coordinating transcription-related functions. This flexibility and multiplicity of related functions can make it difficult to determine the precise role or even the necessity of the CTD in specific processes. They also complicate efforts to understand the evolution of the CTD across the range of eukaryotic diversity, specifically why it has been so strongly conserved in some groups and permitted to degenerate in others. Nevertheless, our investigation of complete genomes suggests a correlation in conservation of the CTD and proteins with which they interact. This likely reflects functional constraints imposed by complex and largely uncharacterized interactions that result in parallel sequence evolution at primary and/or tertiary levels. In the case of coadapted cellular machinery as diverse as the transcriptosome, this parallel evolution may involve sequences with no other obvious connection. This certainly is true of many of the proteins shown to have highly specific interactions with the phospho-CTD (Carty and Greenleaf 2002; Phatnani, Jones, and Greenleaf 2004). The impact that parallel functional constraints have had on comparative evolutionary investigations is unclear; however, our analyses indicate that understanding these constraints may be essential for accurately interpreting the results of phylogenetic and other informatic analyses.

Supplementary Material

Supplementary Figures 1 and 2, Supplementary Tables 1 and 2, and Supplementary File 1 cited in this study are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/). All alignments and coordinate files of predicted models for CE are available from authors upon request.

Acknowledgments

This research was supported by a grant from the National Science Foundation, MCB#0133295. We thank two anonymous reviewers for very helpful input.

Literature Cited

Altschul, S. E., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Anantharaman, V., E. V. Koonin, and L. Aravind. 2002. Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res. **30**:1427–1464.

Baldauf, S. L. 2003. The deep roots of eukaryotes. Science **300**:1703–1706.

Baldauf, S. L., and J. D. Palmer. 1993. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. Proc. Natl. Acad. Sci. USA **90**:11558–11562.

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle. 2000. A kingdom-level phylogeny of eukaryotes based on combined data. Science **290**:971–977.

Bentley, D. 1999. Coupling RNA polymerase II transcription with pre-mRNA processing. Curr. Opin. Cell Biol. **11**:347–351.

———. 2002. The mRNA assembly line: transcription and processing machines in the same factory. Curr. Opin. Cell Biol. **14**:336–342.

Best, A. A., H. G. Morrison, A. G. McArthur, and M. L. Sogin. 2004. Evolution of eukaryotic transcription: insights from the genome of *Giardia lamblia*. Genome Res. **14**:1537–1547.

Campbell, N. A., and J. B. Reece. 2005. Biology. 7th edition. Benjamin Cummings, San Francisco, Calif.

Carty, S. M., and A. L. Greenleaf. 2002. Hyperphosphorylated C-terminal repeat domain-associating proteins in the nuclear proteome link transcription to DNA/chromatin modification and RNA processing. Mol. Cell. Proteomics **1**:598–610.

Cho, H., G. Orphanides, X. Q. Sun, X. J. Yang, V. Ogryzko, E. Lees, Y. Nakatani, and D. Reinberg. 1998. A human RNA polymerase II complex containing factors that modify chromatin structure. Mol. Cell. Biol. **18**:5355–5363.

Chothia, C., and A. M. Lesk. 1986. The relation between the divergence of sequence and structure in proteins. EMBO J. **5**:823–826.

Corbett, K. D., and J. M. Berger. 2004. Structure, molecular mechanisms, and evolutionary relationships in DNA topoisomerases. Annu. Rev. Biophys. Biomol. Struct. **33**:95–118.

Corden, J. L. 1990. Tails of RNA polymerase II. Trends Biochem. Sci. **80**:1251–1255.

Dacks, J. B., J. D. Silberman, A. G. B. Simpson, S. Moriya, T. Kudo, M. Ohkuma, and R. J. Redfield. 2001. Oxymonads are closely related to the excavate taxon *Trimastix*. Mol. Biol. Evol. **18**:1034–1044.

Delwiche, C. F., and J. D. Palmer. 1997. The origin of plastids and their spread via secondary symbiosis. Plant Syst. Evol. **11**(Suppl.):53–86.

Dez, C., J. Noaillac-depeyre, M. Caizergues-Ferrer, and Y. Henry. 2002. Naf1p, an essential nucleoplasmic factor specifically required for accumulation of box H/ACA small nucleolar RNPs. Mol. Cell. Biol. **22**:7053–7065.

Fabrega, C., V. Shen, S. Shuman, and C. D. Lima. 2003. Structure of an mRNA capping enzyme bound to the phosphorylated carboxy-terminal domain of RNA polymerase II. Mol. Cell **11**:1549–1561.

Fatica, A., M. Dlakic, and D. Tollervey. 2002. Naf1p is a box H/ACA snoRNP assembly factor. RNA **8**:1502–1514.

Felsenstein, J. 1989. PHYLIP—phylogenetic inference package (version 3.2). Cladistics **5**:164–165.

Fong, N., and D. L. Bentley. 2001. Capping, splicing and 3′ processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD. Genes Dev. **15**:1783–1795.

Goldstrohm, A. C., T. R. Albrecht, C. Suñe, M. T. Bedford, and M. A. Garcia-Blanco. 2001. The transcription elongation factor CA150 interacts with RNA polymerase II and the pre-mRNA splicing factor SF1. Mol. Cell. Biol. **21**:7617–7628.

Greenleaf, A. 2003. Getting a grip on the CTD of Pol II. Structure **11**:900–902.

Gross, S., and C. Moore. 2001. Five subunits are required for reconstitution of the cleavage and polyadenylation activities of *Saccharomyces cerevisiae* cleavage factor I. Proc. Natl. Acad. Sci. USA **98**:6080–6085.

Guo, Z., and J. W. Stiller. 2004. Comparative genomics of cyclin-dependent kinases suggest co-evolution of the RNAP II C-terminal domain and CTD-directed CDKs. BMC Genomics **5**:69.

Håkansson, K., A. J. Doherty, S. Shuman, and D. B. Wigley. 1997. X-ray crystallography reveals a large confromational change during guanyl transfer by mRNA capping enzymes. Cell **89**:545–553.

Halle, J. P., and M. Meisterernst. 1996. Gene expression: increasing evidence for a transcriptosome. Trends Genet. **12**:161–163.

Hartzog, G. A. 2003. Transctiption elongation by RNA polymerase II. Curr. Opin. Genet. Dev. **13**:119–126.

Hartzog, G. A., J. L. Speer, and D. L. Lindstrom. 2002. Transcript elongation on a nucleoprotein template. Biochim. Biophys. Acta **1577**:276–286.

Hausmann, S., M. A. Altura, M. Witmar, S. M. Singer, H. G. Elmendorf, and S. Shuman. 2005. Yeast-like mRNA capping apparatus in *Giardia lamblia*. J. Biol. Chem. **280**: 12077–12086.

Hirose, Y., and J. L. Manley. 2000. RNA polymerase II and the intergration of the nuclear events. Genes Dev. **8**:637–648.

Holm, L., and J. Park. 2000. DaliLite workbench for protein structure comparison. Bioinformatics **16**:566–567.

Hooft, R. W. W., G. Vriend, C. Sander, and E. E. Abola. 1996. Errors in protein structures. Nature **381**:272.

Howe, K. J. 2002. RNA polymerase II conducts a symphony of pre-mRNA processing activities. Biochim. Biophys. Acta **1577**:308–324.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**:754–755.

Iborra, F. J., A. Pombo, D. A. Jackson, and P. R. Cook. 1996. Active RNA polymerases are localized within discrete transcription "factories" in human nuclei. J. Cell Sci. **109**:1427–1436.

Ito, S., A. Salai, T. Nomura, Y. Miki, M. Ouchida, J. Sasakai, and K. Shimizu. 2001. A novel WD 40 repeat protein, WDC146, highly expressed during spermatogenesis in a stage-specific manner. Biochem. Biophys. Res. Commun. **280**:656–663.

Keeling, P. J., and N. M. Fast. 2002. Microsporidia: biology and evolution of highly reduced intracellular parasites. Annu. Rev. Microbiol. **56**:93–116.

Kim, J. H., W. S. Lane, and D. Reinberg. 2002. Human elongator facilitates RNA polymerase II transcription through chromatin. Proc. Natl. Acad. Sci. USA **90**:1241–1246.

Licciardo, P., L. Ruggiero, L. Lania, and B. Majello. 2001. Transcription activation by targeted recruitment of the RNA polymerase II CTD phosphatase FCP1. Nucleic Acids Res. **29**:3539–3595.

Liu, J., and T. Kipreos. 2002. The evolution of CDK-activating kinases. Pp. 99–111 *in* P. Kaldis, ed. The CDK activating kinases (CAK). Kluwer Academic/Plenum Publishers, London, UK.

Lockhart, P., and M. Steel. 2005. A tale of two processes. Syst. Biol. (in press).

Maniatis, T., and R. Reed. 2002. An extensive network of coupling among gene expression machines. Nature **416**:499–506.

Marchler-Bauer, A., and S. H. Bryant. 2004. CD-search: protein domain annotations on the fly. Nucleic Acids Res. **32**: W327–W331.

Meinhart, A., and P. Cramer. 2004. Recognition of RNA polymerase II carboxy-terminal domain by 3′-RNA processing factors. Nature **430**:223–226.

Noble, C. G., D. Hollingworth, S. R. Martin, V. Ennis-Adeniran, S. J. Smerdon, G. Kelly, I. A. Taylor, and A. Ramos. 2005. Key features of the interaction between Pcf11 CID and RNA polymerase II CTD. Nat. Struct. Mol. Biol. **12**:144–151.

Osborne, C. S., L. Chakalova, K. E. Brown et al. (11 co-authors). 2004. Active genes dynamically colocalize to shared sites of ongoing transcription. Nat. Genet. **36**:1065–1071.

Palancade, B., and O. Bensaude. 2003. Investigating RNA polymerase II carboxyl-terminal domain (CTD) phosphorylation. Eur. J. Biochem. **270**:3859–3870.

Peyretaillade, E., C. Biderre, P. Peyret, F. Duffieux, G. Metenier, M. Gouy, B. Michot, and C. P. Vivares. 1998. Microsporidian Encephalitozoon cuniculi, a unicellular eukaryote with an unusual chromosomal dispersion of ribosomal genes and a LSU rRNA reduced to the universal core. Nucleic Acids Res. **26**:3513–3520.

Phatnani, H. P., J. C. Jones, and A. L. Greenleaf. 2004. Expanding the functional repertoire of CTD kinase I and RNA polymerase II: novel phosphoCTD-associating proteins in the yeast proteome. Biochemistry **43**:15702–15719.

Ponting, C. P. 2002. Novel domains and orthologues of eukaryotic transcription elongation factors. Nucleic Acids Res. **30**: 3643–3652.

Prelich, G. 2002. RNA polymerase II carboxy-terminal domain kinases: emerging clues to their function. Eukaryot. Cell **1**: 153–162.

Riedl, T., and J. M. Egly. 2000. Phosphorylation in transcription: the CTD and more. Gene Expr. **9**:3–13.

Schaffer, A. A., L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin, and S. F. Altschul. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. **29**:2994–3005.

Schwede, T., J. Kopp, N. Guex, and M. C. Peitch. 2003. SWISS-MODEL: an automated protein homology-modeling sever. Nucleic Acids Res. **31**:3381–3385.

Shilatifard, A., R. C. Conaway, and J. W. Conaway. 2003. The RNA polymerase II elongation complex. Annu. Rev. Biochem. **72**:693–715.

Shuman, S. 2002. What messenger RNA capping tells us about eukaryotic evolution. Nat. Rev. Mol. Cell Biol. **3**:619–625.

Stiller, J. W. 2004. Emerging genomic and proteomic evidence on relationships among the animal, plant and fungal kingdoms. Genomics Proteomics Bioinformatics **2**:69–75.

Stiller, J. W., and M. S. Cook. 2004. The conserved functional unit of the RNAP II C-terminal domain lies within heptapeptide pairs. Eukaryotic Cell **3**:735–740.

Stiller, J. W., and B. D. Hall. 1998. Sequences of the largest subunit of RNA polymerase II from two red algae and their implications for rhodophyte evolution. J. Phycol. **34**:857–864.

———. 1999. Long-branch attraction and the rDNA model of early eukaryotic evolution. Mol. Biol. Evol. **16**:1270–1279.

———. 2002. Evolution of the RNA polymerase II C-terminal domain. Proc. Natl. Acad. Sci. USA **99**:6091–6096.

Stiller, J. W., J. Riley, and B. D. Hall. 2001. Are red algae plants? A critical evaluation of three key molecular data sets. J. Mol. Evol. **52**:527–539.

Strimmer, K., and A. von Haeseler. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13**:964–969.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. Clustal W improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Verdecia, M. A., M. E. Bowman, K. P. Lu, T. Hunter, and J. P. Noel. 2000. Structural basis for phosphoserine-proline recognition by group IV WWW domains. Nat. Struct. Biol. **7**:639–643.

Vossbrinck, C. R., J. V. Maddox, S. Friedman, B. A. Debrunner-Vossbrinck, and C. R. Woese. 1987. Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. Nature **326**:411–414.

Vriend, G. 1990. WHAT IF: a molecular modeling and drug design program. J. Mol. Graph. **8**:52–56.

Wiesner, S., G. Stier, M. Sattler, and M. J. Macias. 2002. Solution structure and ligand recognition of the WW domain pair of the yeast splicing factor Prp40. J. Mol. Biol. **324**:807–822.

Xiao, H., Y. Tao, J. Greenblatt, and R. Roeder. 1998. A cofactor, TIP30, specifically enhances HIV-1 Tat-activated transcription. Proc. Natl. Acad. Sci. USA **95**:2146–2151.

Zhao, J., L. Hyman, and C. Moore. 1999. Formation of mRNA 3′ ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol. Mol. Biol. Rev. **63**:405–445.

Zorio, D. A. R., and D. L. Bentley. 2004. The link between mRNA processing and transcription: communication works both ways. Exp. Cell Res. **296**:91–97.

Peter Lockhart, Associate Editor

Accepted July 4, 2005