Nucleic Acids Research, 2002, Vol. 30 No. 17 **3643–3652**

SURVEY AND SUMMARY

Novel domains and orthologues of eukaryotic transcription elongation factors

Chris P. Ponting*

MRC Functional Genetics Unit, Department of Human Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, UK

Received May 28, 2002; Accepted July 2, 2002

ABSTRACT

The passage of RNA polymerase II across eukaryotic genes is impeded by the nucleosome, an octamer of histones H2A, H2B, H3 and H4 dimers. More than a dozen factors in the yeast *Saccharomyces cerevisiae* are known to facilitate transcription elongation through chromatin. In order to better understand the evolution and function of these factors, their sequences have been compared with known protein, EST and DNA sequences. Elongator subcomplex components Elp4p and Elp6p are shown to be homologues of ATPases, yet with substitutions of amino acids critical for ATP hydrolysis, and novel orthologues of Elp5p are detectable in human, and other animal, sequences. The yeast CP complex is shown to contain a likely inactive homologue of M24 family metalloproteases in Spt16p/Cdc68p and a 2-fold repeat in Pob3p, the orthologue of mammalian SSRP1. Archaeal DNA-directed RNA polymerase subunit E" is shown to be the orthologue of eukaryotic Spt4p, and Spt5p and prokaryotic NusG are shown to contain a novel 'NGN' domain. Spt6p is found to contain a domain homologous to the YqgF family of RNases, although this domain may also lack catalytic activity. These findings imply that much of the transcription elongation machinery of eukaryotes has been acquired subsequent to their divergence from prokaryotes.

INTRODUCTION

Chromatin decompaction is required for efficient RNA polymerase II (RNAP-II)-mediated transcription of eukaryotic protein coding genes (1). Transcription is divided into an initiation stage, during which transcription factors and RNAP bind to promoter sites and RNA synthesis commences, followed by an elongation stage, during which RNAP traverses along the DNA assembling an RNA transcript. Transcription elongation through chromatin is severely hindered by the nucleosome, a structure containing DNA wrapped around two copies each of histones H2A, H2B, H3 and H4. Disruption of the structural integrity of the nucleosome, by histone acetylation and/or methylation, by DNA unwinding, or by histone translocation, allows passage of the RNAP-II complex along the gene (1–3).

In the yeast *Saccharomyces cerevisiae* at least a dozen factors are known to facilitate elongation through chromatin (3) (Table 1). These are Rad26p, CP (a heterodimeric factor of Cdc68p/Spt16p and Pob3p), Elongator (containing two subcomplexes, each of three subunits), the Spt4p–Spt5p heterodimer and Spt6p. The molecular functions of these 12 differ greatly. Human DSIF, containing orthologues of yeast Spt4p–Spt5p, functionally interacts with other elongation factors as well as physically with the largest subunit of RNAP (4). Both FACT, the human version of the yeast CP complex, and Spt6p bind histones directly (5,6) whereas Elongator Elp3p acts as a histone acetyltransferase (7). An additional Rad26p-associated factor called Def1p/YKL054Cp enables ubiquitin-mediated proteolysis of RNAP (8).

Human orthologues are known for all of the 12 *S.cerevisiae* elongation factors, with the exceptions of Elp5p and Elp6p, and Def1p/YKL054Cp. Consequently, transcription elongation processes in mammals and yeast are likely to be highly similar. In contrast, only three of the 12 factors, namely Rad26p, Spt5p and Spt6p, have highly sequence-similar homologues in bacteria, and archaea have likely orthologues only of Rad26p, Spt5p and Elp3p. The paucity of candidate orthologues of eukaryotic transcription elongation factors in archaea is curious since they are thought to possess chromatin-like structures (9).

This study sought to determine whether previously undetected homologues, orthologues and domains of *S.cerevisiae* transcription elongation factors could be detected using in-depth sequence database searches. Its aims included the prediction of molecular function using the homology paradigm, and the identification of candidate orthologues of yeast elongation factors in mammals, bacteria and archaea. Sequence data from diverse sources, including incompletely sequenced genomes and expressed sequence tags, were found to be valuable in identifying previously unforeseen evolutionary relationships.

*Tel: +44 1865 272175; Fax: +44 1865 272420; Email: chris.ponting@anat.ox.ac.uk

Table 1. Human, archaeal and bacterial orthologues or homologues (in parentheses) of 12 *S.cerevisiae* transcription elongation factors

Saccharomyces cerevisiae protein	Human orthologue (homologue)	Archaeal orthologue (homologue)	Bacterial orthologue (homologue)	Molecular function/features
Rad26p	CSB	(e.g. APE0413)	(e.g. *E.coli* hepA)	Transcription-coupled DNA repair
Cdc68p/Spt16p	FACT p140	(Xaa-Pro dipeptidase)	(Xaa-Pro dipeptidase)	*Metalloprotease homologue*
Pob3p	SSRP1	None	None	*Novel repeats*
Elp1p	IKBKAP	(WD40 repeat proteins)	(WD40 repeat proteins)	Unknown
Elp2p	Elp2	(WD40 repeat proteins)	(WD40 repeat proteins)	Unknown
Elp3p	FLJ10422	e.g. MJ1136	None	Histone acetyltransferase
Elp4p	Paxneb	(e.g. *AF0352*)	(*ATPases*)	*Inactive ATPase homologues*
Elp5p	*Rai12*	None	None	Unknown
Elp6p	*FLJ20211*	(e.g. *AF0352*)	(*ATPases*)	*Inactive ATPase homologues*
Spt4p	Spt4	*rpoE*"	None	Binds Spt5/NusG?
Spt5p	Supt5h	NusG	NusG (RfaH)	*Novel NGN domain*
Spt6p	Supt6h	None	Tex	*Novel YqgF domain*
Duf1p	(*Etl-1 CUE domains*)	None	None	Recruits UBCs?

Novel findings are given in italics.

MATERIALS AND METHODS

PSI-BLAST, TBLAST-N and BLASTX searches (10) were undertaken at the National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov/blast/) using NCBI databases, including the non-redundant protein sequence database (nr; ftp.ncbi.nlm.nih.gov/blast/db/) currently containing approximately 900 000 sequences. PSI-BLAST searches employed an *E*-value inclusion threshold of 2×10^{-3} and composition-dependent statistics (11), except where stated. The *E*-value corresponding to an alignment score *x* is the number of false positive sequences that are expected to be aligned with scores *x*, or higher, in that search by chance. Additional BLAST searches used the VGE (www.vge.ac.uk) and NCBI unfinished genomes' (www.ncbi.nlm.nih.gov/Microb_blast/unfinishedgenome.html) sites, and organism-specific sites such as dicty.sdsc.edu/annot-blast.html (for *Dictyostelium discoideum*) and www.sanger.ac.uk/Projects/C_briggsae/blast_server.shtml (for *Caenorhabditis briggsae*). Other searches used the nrdb90 protein sequence database (12) (ftp://ftp.ebi.ac.uk/pub/databases/nrdb90/) for which no pair of sequences has greater than 90% pairwise identity. This database contained 474 487 sequences. Pairwise comparison of sequences was achieved using Blast-2-Sequences (http://www.ncbi.nlm.nih.gov/gorf/bl2.html) (13).

Multiple alignments were initially constructed using Clustal-W (14) and manually edited using Seaview (15) according to the guidelines of Bork and Gibson (16). Alignments were presented using the CHROMA tool (17). Hidden Markov model (HMM) searches of protein sequence databases used HMMER2 (18) and an *E*-value inclusion threshold of 0.1. Domain-based analyses used SMART (smart.ox.ac.uk) (19), Pfam (www.sanger.ac.uk/Pfam/) (20) and CDD (www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) (21). Detection of distantly related repeats used Prospero (22). Conservation of gene order in completely sequenced genomes was investigated using the COG database (www.ncbi.nlm.nih.gov/COG/) (23). Comparison of a conserved alignment block of Def1p orthologues with nrdb90 used MoST, the motif search tool (24), and parameters $I = 80\%$ and $E = 0.05$.

Preliminary sequence data were obtained from the Institute for Genomic Research website at http://www.tigr.org.

RESULTS AND DISCUSSION

Archaeal homologues of Spt4p

Initial attempts to identify non-eukaryotic homologues of Spt4p using PSI-BLAST searches of protein sequence databases were unsuccessful. However, Spt4p homologues from incompletely sequenced eukaryotic genomes, such as *D.discoideum*, were identified from TBLAST-N searches of expressed sequence tag (EST) databases (Fig. 1). The conceptual protein sequences were then used to query nr using PSI-BLAST. A database search using, as query, *D.discoideum* Spt4p, derived from ESTs AU038537 and AU073920, yielded significant similarity ($E = 9 \times 10^{-3}$) to the DNA-directed RNA polymerase subunit E" (rpoE") from the hyperthermophilic archaeon *Methanococcus jannaschii* after one iteration. Sequence similarity extends over a N-terminal C4 zinc ribbon and a C-terminal α/β-containing region. Similar rpoE" homologues occur in all other completely sequenced archaea, including a version in *Sulfolobus acidocaldarius* that is fused to DNA-directed RNA polymerase subunit E (Fig. 2).

Despite claims of a RNA polymerase architecture common to both archaea and eukarya (25), previous studies had identified neither archaeal counterparts of eukaryotic Spt4p nor eukaryotic versions of archaeal DNA-directed RNA polymerase subunit E". The identification of these two molecules as homologues resolves these discrepancies and further implicates archaeal subunit E" as a Spt4p-like transcription elongation factor. This might account for the apparent absence of subunit E" in RNAP complexes from *Methanobacterium thermoautotrophicum* (26), since it would be expected to bind archaeal NusG (which binds RNAP) rather than binding RNAP directly (see below).

A novel NusG N-terminal (NGN) homology domain

The C-terminal regions of NusG and Spt5p contain one, and multiple, KOW motifs, respectively (27). A PSI-BLAST search with an N-terminal region of yeast Spt5p (amino acids

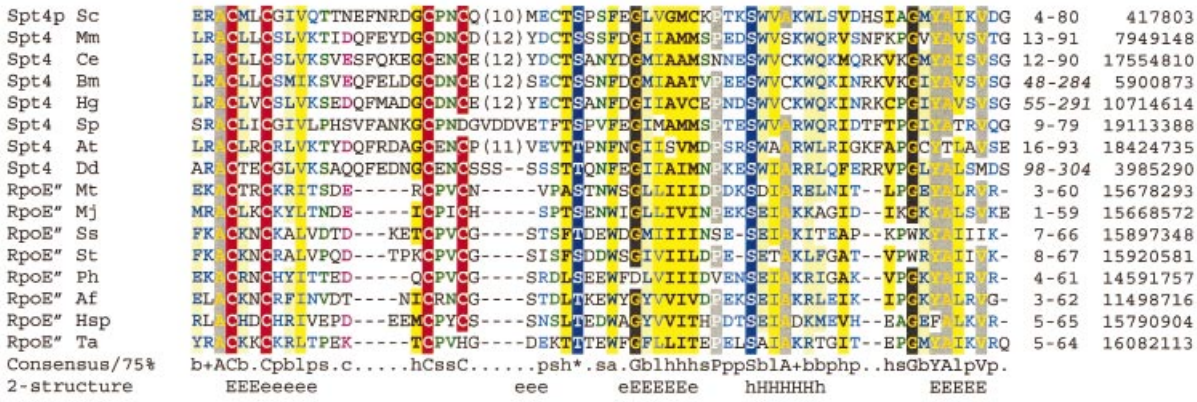

Figure 1. Multiple sequence alignment of eukaryotic Spt4p and archaeal DNA-directed rpoE" represented using CHROMA and a 75% consensus. Cys residues that are likely to bind Zn^{2+} are shown in white-on-red, whilst conserved Ser/Thr residues that might act as phosphorylation sites are shown as white-on-blue. Secondary structures predicted (56) at expected accuracies of >82% (E, H) or >72% (e, h) are indicated below the alignment (E/e, extended or β-strand structure; H/h, α-helical structure). Numbers in parentheses represent amino acids that have been excised from the alignment. Amino acid numbers (or, in italics, nucleotide numbers for ESTs) and GenInfo numbers are given following the alignment. Species abbreviations: Af, *A.fulgidus*; At, *A.thaliana*; Bm, *Brugia malayi*; Ce, *C.elegans*; Dd, *D.discoideum*; Hg, *Heterodera glycines* (soybean cyst nematode); Hsp, *Halobacterium* sp. NRC-1; Mj, *M.jannaschii*; Mm, *Mus musculus*; Mt, *M.thermoautotrophicum*; Ph, *P.horikoshii*; Sc, *S.cerevisiae*; Sp, *S.pombe*; Ss, *Sulfolobus solfataricus*; St, *Sulfolobus tokodaii*; Ta, *T.acidophilum*.

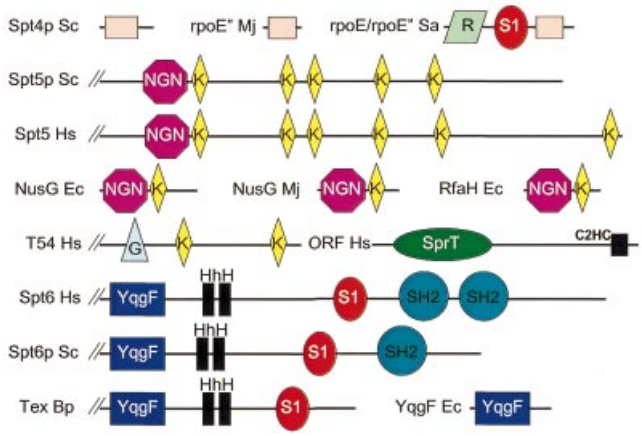

Figure 2. Schematic representation of the domain architectures of molecules discussed in this paper, approximately to scale. Domain abbreviations: C2HC, uvs-2-like C2HC zinc finger; G, G-patch domain; HhH, helix–hairpin–helix motifs; K, KOW motif; R, Rpb7-like N-terminal domain; S1, ribosomal protein S1-like RNA-binding domain. Species abbreviations: Bp, *B.pertussis*; Ec, *E.coli*; Hs, *Homo sapiens*; Mj, *M.jannaschii*; Sa, *S.acidocaldarius*; Sc, *S.cerevisiae*.

220–380, thereby lacking a highly acidic region 1–219) revealed additional significant similarity to NusG homologues [*M.jannaschii* NusG ($E = 2 \times 10^{-4}$) and *Pyrococcus horikoshii* NusG ($E = 6 \times 10^{-4}$)] after two rounds. These NGN domains appear to occur in all Spt5p and NusG homologues in archaea, bacteria and eukarya (Fig. 3). Thus, Spt5p and NusG contain two distinct regions of homology: an NGN domain and one or more KOW motifs (Fig. 2). Database searches using HMMs detected both an NGN domain and a KOW motif in NusG-like bacterial proteins RfaH (28). This is consistent with the known role of RfaH in regulating bacterial transcription elongation (28).

The newly identified NGN domain in human Spt5p may possess an affinity for Spt4p. Yamaguchi *et al.* (4) determined the Spt4p-binding region of Spt5p as amino acids 176–313. This region overlaps both its NGN domain (amino acids 176–267) and its first KOW motif (amino acids 272–299). A natural corollary to the finding that an NGN–KOW region of Spt5p binds Spt4p is that their archaeal homologues, rpoE" and NusG, may also associate. An Spt4p/rpoE"-binding role for NGN–KOW cannot be a universally applicable function since bacteria lack apparent Spt4p orthologues.

Further investigation of KOW motif sequences resulted in their identification in KIN17, a component of the ultraviolet (UV)-C response (29), eukaryotic homologues of human T54 (see also Pfam family PF00467), and in eukaryotic ribosomal S4 proteins (Fig. 2). Finally, although there has been some disagreement in the literature concerning the number and position of KOW motifs in Spt5p (4,30), this study resulted in detection of five KOW motifs in fungal Spt5p and six KOW motifs in mammalian Spt5p (Fig. 2).

The functions and structures of KOW motifs remain enigmatic. From the known structure of the large ribosomal subunit (31), the KOW sequence motif of L24 occurs in three β-strands within a larger src homology 3 (SH3) domain fold, which lies at the exit of the polypeptide tunnel. L24 interacts with several RNA domains in the ribosome, which is in agreement with the original proposal of KOW as an RNA-binding motif (27).

Spt6p domain homologues

Little is known about individual functional domains of yeast Spt6p. Maclennan and Shaw (32) identified a src homology 2 (SH2) domain in a C-terminal region whereas, more recently, Doerks *et al.* (33) described a 'CSZ domain' encompassing most of the N-terminal remainder of Spt6p. The CSZ region encompasses two tandem helix–hairpin–helix (HhH) motifs likely to bind DNA (34). Although Spt6p homologues in eukaryotes other than fungi contain a readily identifiable ribosomal S1-like RNA-binding domain in a region C-terminal to the CSZ, pairwise alignment of Spt6p

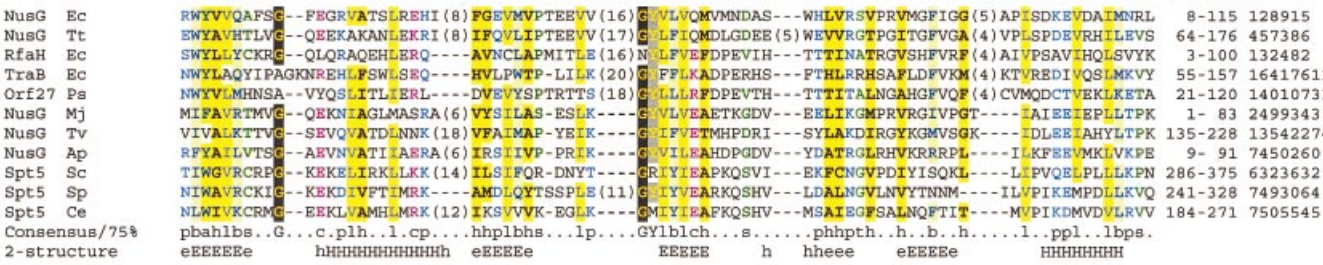

Figure 3. Multiple sequence alignment of NGN domains represented using CHROMA and a 75% consensus. Secondary structures predicted (56) at expected accuracies of >82% (E, H) or >72% (e, h) are indicated below the alignment (E/e, extended or β-strand structure; H/h, α-helical structure). Numbers in parentheses represent amino acids that have been excised from the alignment. GenInfo and amino acid numbers are given following the alignment. Species abbreviations: Ap, *Aeropyrum pernix*; Ce, *C.elegans*; Ec, *E.coli*; Mj, *M.jannaschii*; Ps, *Pseudomonas syringae* pv. *maculicola*; Sc, *S.cerevisiae*; Sp, *S.pombe*; Tt, *Thermus thermophilus*; Tv, *Thermoplasma volcanium*.

homologues demonstrates that the S1 domain is present also in fungi Spt6p (data not shown) (Fig. 2). Eukaryotic Spt6p has previously been shown to be homologous, over the CSZ-S1 domains' region, to the *Bordetella pertussis Tex* (toxin expression) gene product (35,36). Tex is an essential and ubiquitous factor in bacteria and is hypothesised to regulate transcriptional processes (35).

Iterative database searches revealed that the region of CSZ that is N-terminal to the tandem HhH motifs contains a domain predicted to possess a RNase H fold (37). For example, a search for conserved domains in *Xylella fastidiosa* Tex (GenInfo code 11278678) using CDD (21) revealed a possible Pfam-FGGY-like RNase H fold domain (amino acids 330–391; $E = 1 \times 10^{-3}$). A PSI-BLAST search using this sequence as the query identified *Synechocystis* sp. PCC 6803 sll0832 as significantly similar to Tex ($E = 3 \times 10^{-4}$ in round 2). Sll0832 is a member of the YqgF domain family of RNases that includes the eponymous *Escherichia coli* Yqgf. A reciprocal search with *E.coli* YqgF as the query yielded the expected significant similarity ($E = 1 \times 10^{-3}$) with *Bacillus subtilis* Tex in four rounds.

These findings demonstrate that YqgF-homologous domains occur in bacterial Tex orthologues and eukaryotic Spt6p orthologues within their CSZ regions. Thus, CSZ represents a domain and motif combination ('architecture') that is preserved in Tex/Spt6p homologues rather than being a single large domain. This is the first observation of a protein containing both YqgF and other domain types (37). Although previously thought to be absent in archaea (37), YqgF homologous domains are detectable in this kingdom. A PSI-BLAST search with the YqgF domain of *Synechocystis* sp. (strain PCC6803) sll0832 (amino acids 16–141) identified *M.thermoautotrophicum* MTH839 amino acids 23–129 as homologous in nine search rounds with $E = 7 \times 10^{-4}$ (Fig. 4).

YqgF domains in Tex and TexL orthologues are likely to possess a nuclease function. The residues Asp (twice), Glu, and Ser or Thr are absolutely conserved (Fig. 4) in both Tex orthologues and YqgF-like proteins at positions that are thought to contribute to nuclease activity (37). The substrate of this Tex nuclease domain is tentatively suggested to be RNA since Tex negatively regulates transcription when overexpressed, and also since it contains a C-terminal RNA-binding (S1) domain (35).

Spt6p paralogues and transcription elongation

It has been suggested (38) that Spt6p is the eukaryotic orthologue of bacterial Tex. Although Spt6p is a Tex homologue, it is not the most sequence-similar Tex homologue in eukaryotes. The hypothetical proteins human FLJ10379, *Drosophila melanogaster* LD12377p/CG5253 and *Caenorhabditis elegans* ZK973.1 are significantly more sequence similar to bacterial Tex (~35% pairwise sequence identity) than are Spt6p homologues (~25%). Moreover, these three eukaryotic proteins share the same CSZ and S1 domain architecture and predicted catalytic residues in their YqgF domains as bacterial Tex (Fig. 4). Thus, it is predicted that bacterial Tex and eukaryotic homologues such as human FLJ10379 are orthologous and may have comparable cellular functions. In this paper, human FLJ10379, *D.melanogaster* LD12377p/CG5253 and *C.elegans* ZK973.1 homologues, which from searches of EST databases appear to be widespread in eukaryotes, will be described as TexL (Tex-like) genes.

The molecular functions of Tex and TexL orthologues are unknown. In some cases, the repeated co-occurrence of genes in prokaryotic genomes has been used to accurately predict function (39). Consequently, the genomic contexts of Tex orthologues were investigated using the COG database (23). Viewing these contexts (http://www.ncbi.nlm.nih.gov/cgi-bin/COG/coogtik?COG2183) demonstrated that the GreA transcription elongation factor (COG0782) was the most proximal 5′ gene to Tex in four completely sequenced genomes. This is likely to be significant since in *E.coli* the genes are encoded on the same strand whereas in *Vibrio cholerae*, *Haemophilus influenzae* and *Pasteurella multocida* they are on complementary strands. Bacterial GreA is known to promote efficient RNA polymerase transcription elongation past template-encoded arresting sites (40). This suggests that bacterial Tex and eukaryotic TexL, in common with GreA, are transcription elongation factors.

In addition, Tex was found to be the neighbouring gene to *E.coli* sprT homologues in four bacterial genomes: *Lactococcus lactis* (L86677), *Streptococcus pyogenes* (SPy0581), *Bacillus halodurans* (BH0532) and *B.subtilis* (ydcK) (all same strand). These bacterial sequences are homologous to eukaryotic proteins, including human ACRC, since they are found within five PSI-BLAST rounds using the human ACRC sequence (amino acids 411–691) as query and

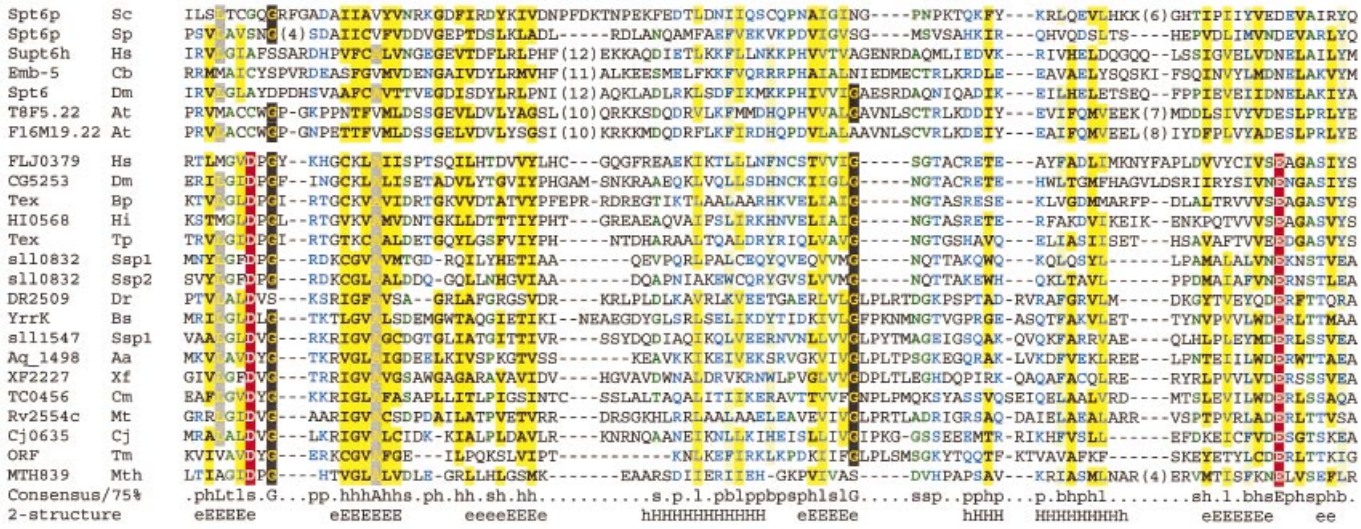

Figure 4. Multiple sequence alignment of YqgF-homologous domains, including those in Tex and Spt6p orthologues, represented using CHROMA and a 75% consensus. Predicted active site residues in YqgF nucleases (37) are shown as white-on-red. Secondary structures predicted (56) at expected accuracies of >82% (E, H) or >72% (e, h) are indicated below the alignment (E/e, extended or β-strand structure; H/h, α-helical structure). Species abbreviations: Aa, *Aquifex aeolicus*; At, *A.thaliana*; Bp, *B.pertussis*; Bs, *B.subtilis*; Cb, *C.briggsae*; Ce, *C.elegans*; Cj, *Campylobacter jejuni*; Cm, *Chlamydia muridarum*; Dm, *D.melanogaster*; Dr, *Deinococcus radiodurans*; Hi, *H.influenzae*; Hs, *H.sapiens*; Mt, *Mycobacterium tuberculosis*; Mth, *M.thermoautotrophicum*; Sc, *S.cerevisiae*; Sp, *S.pombe*; Ssp1, *Synechocystis* sp. PCC6803; Ssp2, *Synechocystis* sp. PCC7002; Tm, *Thermotoga maritima*; Tp, *Treponema pallidum*; Xf, *X.fastidiosa*. GenInfo numbers and amino acid numbers are: Spt6p Sc: 6321552, 750–858; Spt6p Sp: 11359284, 590–687; T8F5.22 At: 7487801, 772–890; F16M19.22 At: 10092249, 506–625; Supt6h Hs: 1136386 779–894; Emb5 Cb: 1669619, 747–862; Spt6 Dm, 7290693, 783–898; FLJ10379 Hs: 11430242, 194–293; CG5253 Dm: 7300889, 526–627; Tex Bp: 2501166, 344–443; HI0568 Hi: 2501167, 327–424; Tex Tp: 7445176, 332–426; sll0832 Ssp1: 7469941, 16–103; sll0832 Ssp2: 13924476, 2–89; DR2509 Dr: 11136112, 16–113; YrrK Bs: 6226496, 1–103; sll1547 Ssp1: 6226451, 5–105; Aq_1498 Aa: 6226434, 1–99; XF2227 Xf: 1135885, 11–111; TC0456 Cm: 11278851, 8–110; Rv2554c Mt: 6226486, 22–123; Cj0635 Cj: 11135891, 1–97; ORF Tm: 7462233, 12–100; MTH839 Mth: 7482690, 23–115.

an *E*-value inclusion threshold of 0.002. The *ACRC* gene maps to the Dystonia parkinsonism critical interval in Xq13.1 (41). It is inferred from their genomic co-occurrence with Tex that SprT and ACRC also function in transcription elongation. Three viral SprT homologues are known, in *Mamestra configurata* nucleopolyhedrovirus and *Leucania separata* nucleopolyhedrovirus. These are the only viral homologues of eukaryotic transcription elongation factors known. Widespread conservation of a HExxH motif and His and Cys residues indicates that SprT homologues are metalloproteases (Fig. 5).

Homologues of Spt16/SSRP1 (FACT)

The human orthologue of yeast Spt16p/CDC68p is one subunit of FACT, a heterodimer which is a chromatin-specific transcription elongation factor (5). A PSI-BLAST database search with yeast Spt16p/CDC68p as query revealed that it is a member of the metallopeptidase family M24: in the first search round, *Staphylococcus aureus* subsp. *aureus* N315 Xaa-Pro dipeptidase (GenInfo code 15927110) was found with $E = 2 \times 10^{-5}$. Interestingly, Spt16p/CDC68p orthologues lack amino acids that are known to be essential for catalysis (data not shown). Thus, Spt16p/CDC68p is predicted to adopt the fold of the peptidase M24 family, but not possess its protease activity. The Spt16p/Cdc68p metalloprotease-homology domain is within an N-terminal region known to affect chromatin structure thereby inhibiting transcription (42). In the absence of catalytic residues, it might be thought that the molecular function of Spt16p/CDC68p is as a DNA-binding factor. This would be consistent with the observation that a *Schizosaccharomyces pombe* metallopeptidase M24 family member has been shown to preferentially bind curved DNA (43). However, as human Spt16 has been reported not to bind unmodified DNA (44) its function still remains to be determined.

The sequences of SSRP1 (single-stranded recognition protein 1), the second subunit of FACT, were also investigated for distant homology. Although no previously unknown SSRP homologues were detected, a tandem repeat within all animal SSRP1s was detected (Fig. 6). For example, a search for repeats in *D.melanogaster* SSRP1 using Prospero (22) revealed significant internal sequence similarity ($P = 1.6 \times 10^{-3}$). This was consistent with the results of PSI-BLAST searches. For example, a search with *S.cerevisiae* Ynl206c, using an *E*-value inclusion threshold of 0.002, indicated the presence of a second repeat in *Xenopus laevis* SSRP1 (DUF87) with $E = 3.2$.

It is not apparent from this analysis what these repeats' function might be. However, it is unlikely to be DNA binding since this function is conveyed by the high-mobility group (HMG) domain present in most of the SSRP1 homologues. An alternative hypothesis is that the repeats in SSRP1 mediate its affinity for Spt16, the other FACT subunit. It is notable that although the isolated HMG domain of SSRP1 binds DNA, it cannot do so in the full-length molecule except when in the presence of Spt16 (44). Thus, one of the functions of the SSRP1 repeats may be to regulate its multidomain conformational change that is induced by Spt16-binding.

Elongator subunits

The histone acetyltransferase complex holo-elongator can be isolated as two subcomplex factors that associate with

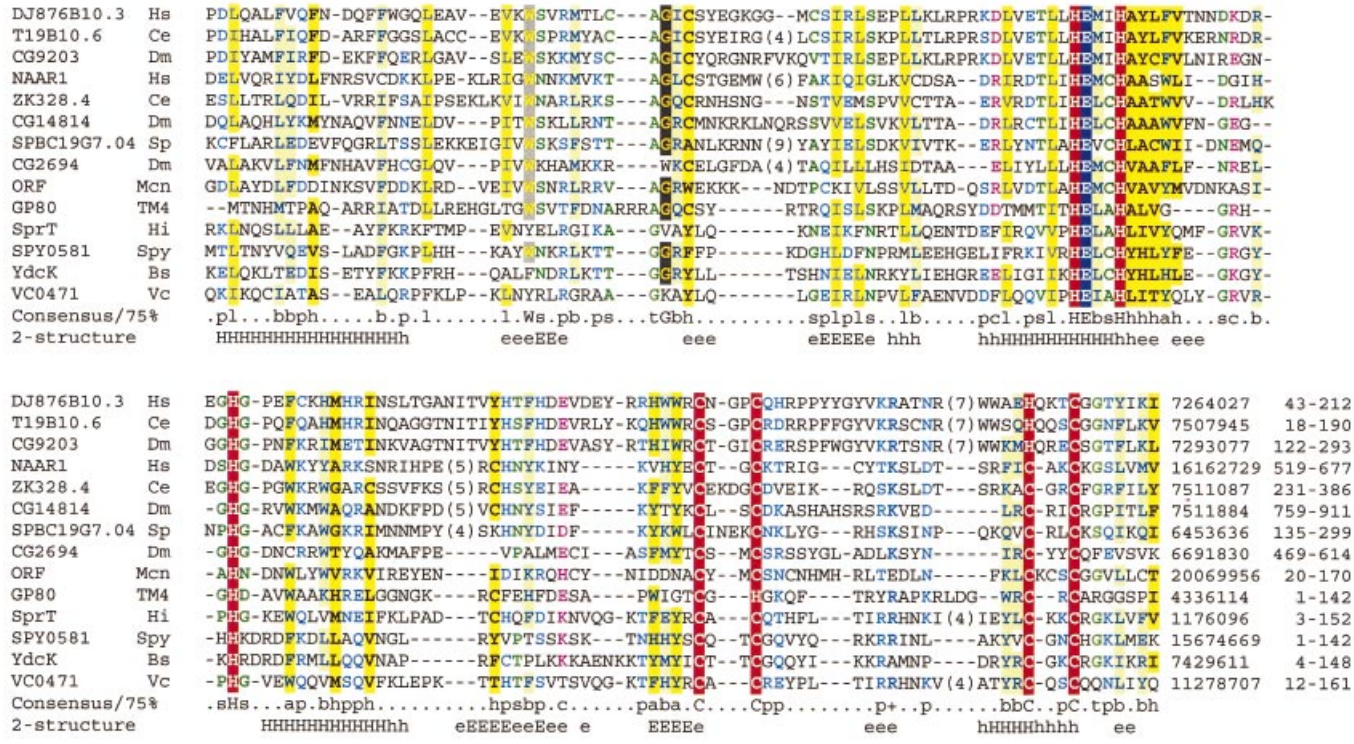

Figure 5. Multiple sequence alignment of SprT homologues represented using CHROMA and a 75% consensus. His and Cys residues that are likely to bind divalent metal ions are shown in white-on-red, whilst a likely catalytic Glu residue is shown as white-on-blue. Secondary structures predicted (56) at expected accuracies of >82% (E, H) or >72% (e, h) are indicated below the alignment (E/e, extended or β-strand structure; H/h, α-helical structure). Numbers in parentheses represent amino acids that have been excised from the alignment. GenInfo and amino acid numbers are given following the alignment. Species abbreviations: Bs, *B.subtilis*; Ce, *C.elegans*; Dm, *D.melanogaster*; Hi, *H.influenzae*; Hs, *H.sapiens*; Mcn, *M.configurata* nucleopolyhedrovirus; Sp, *S.pombe*; Spy, *S.pyogenes*; TM4, Mycobacteriophage TM4; Vc, *V.cholerae*.

Figure 6. Multiple sequence alignment of SSRP1 repeats represented using CHROMA and a 75% consensus. The top two tiers of sequences represent the two repeats in SSRP1 orthologues. Secondary structures predicted (56) at expected accuracies of >82% (E, H) or >72% (e, h) are indicated below the alignment (E/e, extended or β-strand structure; H/h, α-helical structure). Species abbreviations: Ce, *C.elegans*; Dm, *D.melanogaster*; Hs, *H.sapiens*; Sc, *S.cerevisiae*; Sp, *S.pombe*; Xl, *X.laevis*; Zm, *Zea mays*. Amino acid numbers are given following the alignment and GenInfo numbers are: SSRP1 Hs: 730840; SSRP1 Dm: 12644386; SSRP1 Xl: 4586285; SSRP1 Ce: 1174454; C32F10.5 Ce: 7496859; YML069W Sc: 2497082; SPBC609.05 Sp: 2497082; SSRP1 Zm: 8920409; YNL206C Sc: 732190; SPAC6G9.03c Sp: 2842694.

RNAP-II (45,46). For the first subcomplex, Elp1p and Elp2p contain WD40 repeats, whilst Elp3p is a histone H3 and H4 acetyltransferase and possible histone demethylase (47). The functions of the three proteins, Elp4p, Elp5p and Elp6p, in the second subcomplex remain poorly understood. The sequence-based approaches used in this study, however, demonstrate that both Elp4p and Elp6p are inactive homologues of P-loop ATPases/GTPases.

Likely orthologues of yeast Elp4p were previously identified in vertebrates and invertebrates (45). PSI-BLAST

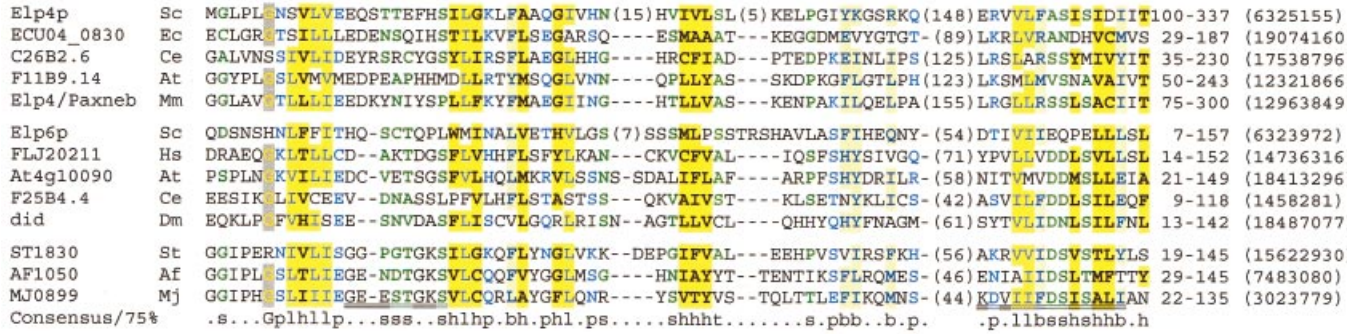

Figure 7. Multiple sequence alignment of Elp4p orthologues (top five sequences) and Elp6p orthologues (next five sequences) with archaeal probable ATPases (last three sequences) represented using CHROMA and a 75% consensus. Walker A and B motifs are double underlined beneath the alignment. Numbers in parentheses represent amino acids that have been excised from the alignment. Amino acid limits and GenInfo numbers are shown at the end of the alignment. Species abbreviations: Af, *A.fulgidus*; At, *A.thaliana*; Ce, *C.elegans*; Dm, *D.melanogaster*; Ec, *Encephalitozoon cuniculi*; Mj, *M.jannaschii*; Mm, *M.musculus*; Hs, *H.sapiens*; Sc, *S.cerevisiae*; St, *S.tokodaii*.

searches with these Elp4p orthologues provide evidence that Elp4p are ATPase homologues. For example, a search with the *Arabidopsis* orthologue (GenInfo code 12321866) reveals significant similarity ($E = 4 \times 10^{-4}$) in two rounds to the ATPase domain in the *X.fastidiosa* 9a5c radA-like protein (Fig. 7).

Saccharomyces cerevisiae Elp6p is not apparently similar to any sequence in the nr, although an orthologue is readily apparent from a search of the *Candida albicans* unfinished genome (TBLAST-N $E = 9 \times 10^{-18}$ using http://www.ncbi.nlm.nih.gov/Microb_blast/unfinishedgenome.html). Detailed searches did reveal a family of orthologues in other eukaryotes, such as human FLJ20211, *S.pombe* SPBC3H7.10, *D.melanogaster* diminished discs (DID), *D.discoideum* ORF (GenInfo code 12007287), and *Arabidopsis thaliana* F28M11.10, and two lines of evidence indicated that these represent Elp6p orthologues. First, a PSI-BLAST search found significant similarity between the *A.thaliana* F28M11.10 sequence and *C.albicans* Elp6p with $E = 9 \times 10^{-3}$ after three rounds. This search used amino acids 21–242 of F28M11.10, but corrected to reflect sequence differences manifest between F28M11.10, and ESTs AV552362 and AU226546. The search also employed a nrdb90 database that was supplemented by the *C.albicans* Elp6p orthologue. Secondly, the family is represented in all free-living eukaryotic genomes sequenced to date except *S.cerevisiae*. Consequently, Elp6p is the most likely candidate as the *S.cerevisiae* orthologue of this family.

Surprisingly, proposed Elp6p family members are also likely homologues of Elp4p. A PSI-BLAST search with *D.melanogaster* DID (amino acids 13–245) and an *E*-value inclusion threshold of 5×10^{-3} revealed marginal similarity ($E = 6 \times 10^{-3}$) in five rounds to a human hypothetical protein (GenInfo code 15214765) that is, in turn, homologous to Elp4p. This search was undertaken using nrdb90 supplemented by Elp6p orthologues' sequences from *C.albicans*, *Lycopersicon esculentum* and *Ciona intestinalis* taken from EST and genome sequencing projects. Construction of a multiple alignment of Elp4p and Elp6p homologues with ATPases demonstrates that the elongator components lack the phosphate-binding P-loop (Fig. 7). This implies that these proteins lack ATPase activities.

In the absence of ATPase activities the functions of Elp4p and Elp6p orthologues remain to be clarified. Whatever these functions might be, it is possible that archaea possess analogous functions since ATPase homologues that are similar in sequence to Elp4p and Elp6p and have substitutions within their P-loops are apparent for example in *Archaeoglobus fulgidus* (AF0352, AF0518 and AF1172), *Halobacterium* sp. NRC-1 (HtlC; GenInfo code 15790668), *P.horikoshii* (PH1120) and *Thermoplasma acidophilum* (Ta0084).

No orthologues of Elp5p (also known as YHR187w and Iki1p) outside of the fungi are readily apparent from BLAST database searches using composition-dependent statistics (11). However, a reciprocal PSI-BLAST search with a *D.discoideum* protein sequence (GenInfo code 19570052) and composition-dependent statistics revealed significant similarity ($E = 2 \times 10^{-4}$) to *S.pombe* Elp5p (SPBC18E5.05c) in three rounds. Such searches determined that Elp5p orthologues are present across the eukaryotes, in mammals, *Drosophila* and *C.elegans* (Fig. 8). Little is known of these proteins except that expression of *Rai12*, the mouse orthologue gene, is induced by retinoic acid (48). However, the identification of likely Elp5p orthologues should assist in the investigation of this Elongator subunit in mammals.

Homologues of *S.cerevisiae* Def1p

Rad26p facilitates UV-light-induced DNA damage and appears to protect RNAP-II from degradation during the repair process (8). In contrast, the association of Def1p with Rad26p in chromatin appears to enable ubiquitination of RNAP-II and leads to its proteolysis by the proteasome (8). As noted elsewhere (8), UV-induced RNAP-II ubiquitination and degradation has been observed in fungi and mammals, yet Def1p orthologues have not been detected in standard protein sequence databases.

In order to search for candidate Def1p orthologues, the *S.cerevisiae* Def1p sequence was compared with unfinished genome, EST and protein sequences using the NCBI BLAST web resources. This resulted in the identification of likely orthologues from four fungi: *C.albicans* (on contig 6–2503), *Aspergillus fumigatus* (on fragment 2283), *S.pombe* (gene SPBC354.10) and *Coccidioides immitis* (encoded in ESTs BF251037 and BF252062). A MoST search (24) using these

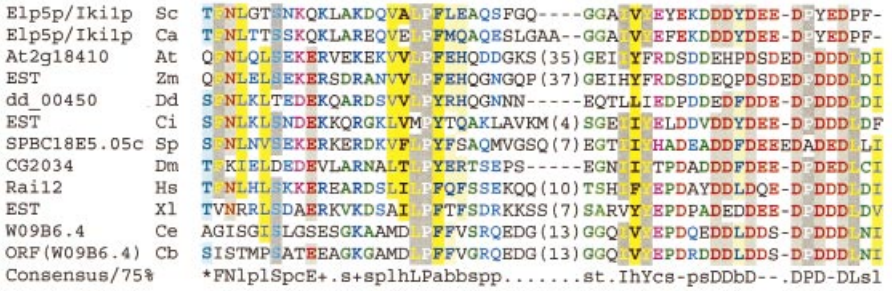

Figure 8. Multiple sequence alignment of the C-terminal regions of Elp5p homologues represented using CHROMA and a 75% consensus. Sequence conservation is detectable throughout these likely orthologues but accurate multiple sequence alignment is problematic. Numbers represent amino acids that have been excised from the alignment. Species abbreviations: At, *A.thaliana*; Ca, *C.albicans*; Cb, *C.briggsae*; Ce, *C.elegans*; *Ci*, *C.intestinalis*; Dd, *D.discoideum*; Dm, *D.melanogaster*; Hs, *H.sapiens*; Sc, *S.cerevisiae*; Sp, *S.pombe*; Xl, *X.laevis*; Zm, *Z.mays*. GenInfo numbers and amino acid numbers (when known) are: Elp5/Iki1p Sc: 6321981, 257–309; At2g18410 At: 15224168, 303–392; EST Zm: 6721029 (conceptual translation); dd_00450 Dd: 19570052, 201–253; EST Ci: 16853181 (conceptual translation); SPBC18E5.05c Sp: 19112643, 252–314; CG2034 Dm: 7292214 210–262; Rai12 Hs: 20086425, 244–307; EST Xl: 13166929 (conceptual translation); W09B6.4 Ce: 17536805, 295–361.

sequences identified the N-terminal of two CUE domains (49) in mouse Enhancer-trap-locus-1 (Etl-1; amino acids 271–300) (50) as being similar to these sequences with $E = 4.8 \times 10^{-2}$ (Fig. 9). CUE domains in Etl-1 were identified using Pfam (20). Additionally, *C.albicans* Def1p was the highest scoring sequence, albeit with a non-significant *E*-value (0.74), in a search of known sequences using the SMART CUE domain HMM.

These marginal similarities may not have provided sufficient evidence for the presence of a CUE domain in Def1p orthologues, except for the existence of strong functional similarities between yeast Def1p and Cue1p in the literature. Cue1p is known to recruit the soluble ubiquitin-conjugating enzyme (UBC) Ubc7p to the endoplasmic reticulum (ER) membrane prior to the ubiquitination of products that undergo ER-associated degradation (51). Def1p coordinates the ubiquitination of RNAP-II, presumably when transcription is stalled at a site of DNA damage (8). Consequently, similarities in both sequence and function indicate that these two proteins contain a conserved CUE domain. Like the CUE domain in Cue1p, the predicted Def1p CUE domain may recruit UBC E2 to the transcription complex.

Interestingly, among mammalian CUE domains, the yeast putative Def1p CUE domains are most similar to those in Etl-1, a member of the SNF2/SWI2 family of transcriptional regulators. Since yeast Rad26p, the interaction partner of Def1p, is also a member of this family, the domain architecture arising from the conceptual fusion of Def1p and Rad26p is almost equivalent to that of Etl-1. Using the concept of Etl-1 as a 'Rosetta Stone protein' (52), this suggests that mammalian Etl-1, whose cellular function remains ill determined, may lie in regulating transcription elongation.

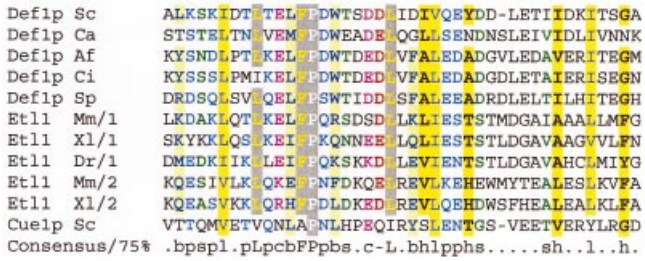

Figure 9. Multiple sequence alignment of CUE domains in Def1p, Enhancer-trap-locus-1 (Etl-1) orthologues and yeast Cue1p, represented using CHROMA and a 75% consensus. Species abbreviations: Af, *A.fumigatus*; Ca, *C.albicans*; Ci, *C.immitis*; Dr, *Danio rerio*; Mm, *M.musculus*; Sc, *S.cerevisiae*; Sp, *S.pombe*; Xl, *X.laevis*. GenInfo numbers and amino acid numbers (when known) are: Def1p Sc: 6322796, 21–62; Def1p Sp (SPBC354.10): 19112026, 77–119; Etl1 Mm: 1082208, 271–313 and 362–404; Cue1p Sc: 6323920, 65–106.

Conclusions: evolution of eukaryotic transcription elongation factors

These data suggest that of all the modern components of the eukaryotic transcription elongation machinery only NusG/Spt5 and RNAP itself were present in the last common ancestor of the three kingdoms of cellular life, archaea, eubacteria and eukarya. The eukaryotic transcription elongation machinery appears to have appropriated components from other cellular processes such as protein degradation (Spt16p/Cdc68p is a metalloprotease homologue), ATP-dependent chromatin remodelling (Elp4p and Elp6p are ATPase homologues) and nucleic acid hydrolysis (Spt6p contains a YqgF nuclease domain homologue). In each of these three cases appropriation is associated with apparent losses in enzymatic activities, with substitutions of known active site residues.

Apart from NusG/Spt5, the only factor that appears to have survived *in situ* since the common ancestor of eukaryotes and archaea is Spt4/rpoE", whereas the bacteria and eukaryotes only otherwise share Tex/TexL and possibly sprT. Eukaryotic Spt6 is a specialisation of bacterial Tex with accretions of a single SH2 domain in fungi and a pair of consecutive SH2 domains in animals, and with a loss of YqgF nuclease activity. Eukaryotes have also evolved a transcription elongation apparatus that has no demonstrable homologues in the prokaryotes. This includes subunits of the Paf1 complex (53–55) and domains in SSRP1/Pob3p and Def1p that are only currently found elsewhere in other eukaryotic proteins.

Comparison of eukaryotic Spt5 with prokaryotic NusG shows that it too has acquired structural additions. It has accreted many additional domains, in particular multiple KOW motifs (Fig. 2), since it diverged from the archaeal and

bacterial NusG lineages. This may reflect the numerous physical interactions with the eukaryotic-specific Paf1 and Spt16p/CDC68p/Pob3p/FACT complexes (53).

ACKNOWLEDGEMENTS

I would like to thank Abigail Lazzerine and Nick Dickens for assistance in searching databases for Spt16p and Spt6p homologues, respectively. Preliminary sequence data was obtained from the Institute for Genomic Research website at http://www.tigr.org.

REFERENCES

1. Orphanides,G. and Reinberg,D. (2000) RNA polymerase II elongation through chromatin. *Nature*, **407**, 471–475.
2. Richards,E.J. and Elgin,S.C. (2002) Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. *Cell*, **108**, 489–500.
3. Svejstrup,J.Q. (2002) Chromatin elongation factors. *Curr. Opin. Genet. Dev.*, **12**, 156–161.
4. Yamaguchi,Y., Wada,T., Watanabe,D., Takagi,T., Hasegawa,J. and Handa,H. (1999) Structure and function of the human transcription elongation factor DSIF. *J. Biol. Chem.*, **274**, 8085–8092.
5. Orphanides,G., Wu,W.H., Lane,W.S., Hampsey,M. and Reinberg,D. (1999) The chromatin-specific transcription elongation factor FACT comprises human SPT16 and SSRP1 proteins. *Nature*, **400**, 284–288.
6. Bortvin,A. and Winston,F. (1996) Evidence that Spt6p controls chromatin structure by direct interaction with histones. *Science*, **272**, 1473–1476.
7. Wittschieben,B.O., Otero,G., de Bizemont,T., Fellows,J., Erdjument-Bromage,H., Ohba,R., Li,Y., Allis,C.D., Tempst,P. and Svejstrup,J.Q. (1999) A novel histone acetyltransferase is an integral subunit of elongating RNA polymerase II holoenzyme. *Mol. Cell*, **4**, 123–128.
8. Woudstra,E.C., Gilbert,C., Fellows,J., Jansen,L., Brouwer,J., Erdjument-Bromage,H., Tempst,P. and Svejstrup,J.Q. (2002) A Rad26–Def1 complex coordinates repair and RNA pol II proteolysis in response to DNA damage. *Nature*, **415**, 929–933.
9. Zlatanova,J. (1997) Archaeal chromatin: virtual or real? *Proc. Natl Acad. Sci. USA*, **94**, 12251–12254.
10. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
11. Schäffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
12. Holm,L. and Sander,C. (1998) Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423–429.
13. Tatusova,T.A. and Madden,T.L. (1999) Blast 2 sequences—a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.*, **174**, 247–250.
14. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (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.
15. Galtier,N., Gouy,M. and Gautier,C. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.*, **12**, 543–548.
16. Bork,P. and Gibson,T.J. (1996) Applying motif and protein searches. *Methods Enzymol.*, **266**, 162–184.
17. Goodstadt,L. and Ponting,C.P. (2001) CHROMA: consensus-based colouring of multiple alignments for publication. *Bioinformatics*, **17**, 845–846.
18. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
19. Letunic,I., Goodstadt,L., Dickens,N.J., Doerks,T., Schultz,J., Mott,R., Ciccarelli,F., Copley,R.R., Ponting,C.P. and Bork,P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, **30**, 242–244.
20. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
21. Marchler-Bauer,A., Panchenko,A.R., Shoemaker,B.A., Thiessen,P.A., Geer,L.Y. and Bryant,S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **30**, 281–283.
22. Mott,R. and Tribe,R. (1999) Approximate statistics of gapped alignments. *J. Comput. Biol.*, **6**, 91–112.
23. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
24. Tatusov,R.L., Altschul,S.F. and Koonin,E.V. (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl Acad. Sci. USA*, **91**, 12091–12095.
25. Best,A.A. and Olsen,G.J. (2001) Similar subunit architecture of archaeal and eukaryal RNA polymerases. *FEBS Microbiol. Lett.*, **195**, 85–90.
26. Darcy,T.J., Hausner,W., Awery,D.E., Edwards,A.M., Thomm,M. and Reeve,J.N. (1999) *Methanobacterium thermoautotrophicum* RNA polymerase and transcription *in vitro*. *J. Bacteriol.*, **181**, 4424–4429.
27. Kyrpides,N.C., Woese,C.R. and Ouzounis,C.A. (1996) KOW: a novel motif linking a bacterial transcription factor with ribosomal proteins. *Trends Biochem. Sci.*, **21**, 425–426.
28. Bailey,M.J.A., Hughes,C. and Koronakis,V. (1997) RfaH and the *ops* element, components of a novel system controlling bacterial transcription elongation. *Mol. Microbiol.*, **26**, 845–851.
29. Angulo,J.F., Rouer,E., Mazin,A., Mattei,M.G., Tissier,A., Horellou,P., Benarous,R. and Devoret,R. (1991) Identification and expression of the cDNA of KIN17, a zinc-finger gene located on mouse chromosome 2, encoding a new DNA-binding protein. *Nucleic Acids Res.*, **19**, 5117–5123.
30. Hartzog,G.A., Wada,T., Handa,H. and Winston,F. (1998) Evidence that Spt4, Spt5 and Spt6 control transcription elongation by RNA polymerase II in *Saccharomyces cerevisiae*. *Genes Dev.*, **12**, 357–369.
31. Ban,N., Nissen,P., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.
32. Maclennan,A.J. and Shaw,G. (1993) A yeast SH2 domain. *Trends Biochem. Sci.*, **18**, 464–465.
33. Doerks,T., Copley,R.R., Schultz,J., Ponting,C.P. and Bork,P. (2002) Systematic identification of novel protein domain families associated with nuclear functions. *Genome Res.*, **12**, 47–56.
34. Doherty,A.J., Serpell,L.C. and Ponting,C.P. (1996) The helix–hairpin–helix DNA-binding motif: a structural basis for non-sequence-specific recognition of DNA. *Nucleic Acids Res.*, **24**, 2488–2497.
35. Fuchs,T.M., Deppisch,H., Scarlato,V. and Gross,R. (1996) A new gene locus of *Bordetella pertussis* defines a novel family of prokaryotic transcriptional accessory proteins. *J. Bacteriol.*, **178**, 4445–4452.
36. Kaplan,C.D., Morris,J.R., Wu,C.T. and Winston,F. (2000) Spt5 and Spt6 are associated with active transcription and have characteristics of general elongation factors in *D. melanogaster*. *Genes Dev.*, **14**, 2623–2634.
37. Aravind,L., Makarova,K.S. and Koonin,E.V. (2000) Holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. *Nucleic Acids Res.*, **28**, 3417–3432.
38. Anantharaman,V., Koonin,E.V. and Aravind,L. (2002) Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res.*, **30**, 1427–1464.
39. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
40. Borukhov,S., Polyakov,A., Nikiforov,V. and Goldfarb,A. (1992) GreA protein: a transcription elongation factor from *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **89**, 8899–8902.
41. Nolte,D., Ramser,J., Niemann,S., Lehrach,H., Sudbrak,R. and Muller,U. (2001) ACRC codes for a novel nuclear protein with unusual acidic

repeat tract and maps to DYT3 (dystonia parkinsonism) critical interval in Xq13.1. *Neurogenetics*, **3**, 207–213.
42. Evans,D.R.H., Brewster,N.K., Xu,Q., Rowley,A., Altheim,B.A., Johnston,G.C. and Singer,R.A. (1998) The yeast protein complex containing Cdc68 and Pob3 mediates core-promoter repression through the Cdc68 N-terminal domain. *Genetics*, **150**, 1393–1405.
43. Yamada,H., Mori,H., Momoi,H., Nakagawa,Y., Ueguchi,C. and Mizuno,T. (1994) A fission yeast gene encoding a protein that preferentially associates with curved DNA. *Yeast*, **10**, 883–894.
44. Yarnell,A.T., Oh,S., Reinberg,D. and Lippard,S.J. (2001) Interaction of FACT, SSRP1, and the high mobility group (HMG) domain of SSRP1 with DNA damaged by the anticancer drug cisplatin. *J. Biol. Chem.*, **276**, 25736–25741.
45. Winkler,G.S., Petrakis,T.G., Ethelberg,S., Tokunaga,M., Erdjument-Bromage,H., Tempst,P. and Svejstrup,J.Q. (2001) RNA polymerase II elongator holoenzyme is composed of two discrete subcomplexes. *J. Biol. Chem.*, **276**, 32743–32749.
46. Krogan,N.J. and Greenblatt,J.F. (2001) Characterization of a six subunit holo-elongator complex required for the regulated expression of a group of genes in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, **21**, 8203–8212.
47. Chinenov,Y. (2002) A second catalytic domain in the Elp3 histone acetyltransferases: a candidate for histone demethylase activity? *Trends Biochem. Sci.*, **27**, 115–117.
48. Spanjaard,R.A., Lee,P.J., Sarkar,S., Goedegebuure,P.S. and Eberlein,T.J. (1997) Clone 10d/BM28, an early S-phase protein, is an important growth regulator of melanoma. *Cancer Res.*, **57**, 5122–5128.
49. Ponting,C.P. (2000) Proteins of the endoplasmic reticulum-associated degradation pathway: domain detection and function prediction. *Biochem. J.*, **351**, 527–535.
50. Soininen,R., Schoor,M., Henseling,U., Tepe,C., Kisters-Woike,B., Rossant,J. and Gossler,A. (1992) The mouse *Enhancer trap locus 1* (*Etl-1*): a novel mammalian gene related to *Drosophila* and yeast transcriptional regulator genes. *Mech. Dev.*, **39**, 111–123.
51. Biederer,T., Volkwein,C. and Sommer,T. (1997) Role of Cue1p in ubiquitination and degradation at the ER surface. *Science*, **278**, 1806–1809.
52. Marcotte,E.M., Pellegrini,M., Ng,H.-L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
53. Squazzo,S.L., Costa,P.J., Lindstrom,D.L., Kumer,K.E., Simic,R., Jennings,J.L., Link,A.J., Arndt,K.M. and Hartzog,G.A. (2002) The Paf1 complex physically and functionally associates with transcription elongation factors *in vivo*. *EMBO J.*, **21**, 1764–1774.
54. Mueller,C.L. and Jaehning,J.A. (2002) Ctr9, Rtf1, and Leo1 are components of the Paf1/RNA polymerase II complex. *Mol. Cell. Biol.*, **22**, 1971–1980.
55. Pokholok,D.K., Hannett,N.M. and Young,R.A. (2002) Exchange of RNA polymerase II initiation and elongation factors during gene expression *in vivo*. *Mol. Cell*, **9**, 799–809.
56. Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.