



Transcription-associated protein families are primarily taxon-specific

Richard M.R. Coulson, Anton J. Enright and Christos A. Ouzounis

Computational Genomics Group, Research Programme, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK

Received on June 14, 2000; revised on August 7, 2000; accepted on August 10, 2000

ABSTRACT

The mechanisms controlling gene regulation appear to be fundamentally different in eukaryotes and prokaryotes (Struhl (1999) *Cell*, **98**, 1–4). To investigate this diversity further, we have analysed the distribution of all known transcription-associated proteins (TAPs), as reflected by sequence database annotations. Our results for the primary phylogenetic domains (Archaea, Bacteria and Eukaryota) show that TAP families are mostly taxon-specific and very few transcriptional regulators are common across these domains.

Contact: ouzounis@ebi.ac.uk

The TAPs were extracted from the protein sequence databases via keyword searches (Kyrpides and Ouzounis, 1999) and clustered using an algorithm previously used to detect gene fusion events (Enright *et al.*, 1999) (Figure 1a). Of the 5894 sequences obtained, 4533 (77%) clustered into 241 families with 3 or more members. 90% of these protein families are uniquely present in one of the three primary domains. There are seven clusters that are universally present (Figure 1b). The two main RNA polymerase subunits, two observed only in Fungi (TenA and NifL) and SIR2. SIR2 is the only universal regulator that is predominantly eukaryotic as it is only present in the bacterium *Streptomyces coelicolor*. The remaining two clusters in this group contain sequences present in eukaryotes but encoded by plastid genomes of bacterial origin (Gray, 1993). The families common between the Archaea and Bacteria are all transcriptional regulators (Kyrpides and Ouzounis, 1999). Apart from the MCM (Zhang *et al.*, 1998) and SmuBP-2 (Mizuta *et al.*, 1993) clusters, the TAPs shared between the Archaea and Eukaryota are basal transcription factors and RNA polymerase subunits. Only the cold shock domain family is shared between eukaryotes and bacteria.

In Bacteria, the three major categories (Firmicutes [Gram-positive], Proteobacteria and Other [Gram-negative]) exhibit fragmentation of their transcription

components (Figure 1c) with 13% of the families unique to Gram-positive and 32% unique to Gram-negative bacteria. Only 22% of the bacterial-specific TAP families are distributed across the three categories. The eukaryote crown group (Fungi, Metazoa and Plants) exhibits a far higher level of fragmentation with only 9% of families common to all three categories (Figure 1d). 45% of the known eukaryotic TAP families are unique to metazoans, 16% unique to fungi, 8% unique to plants and 21% shared between metazoans and fungi. Given the extensive sequence information obtained for Fungi or Metazoa (including complete genome sequences), it appears that these unique TAP families are confined to these domains. Less than 1% of the TAP families are common between plants and either one of the other two categories.

The above analysis represents the most comprehensive overview of the transcriptional machinery to date and is in accordance with previous computational and experimental work. The main result is that only a minority of transcriptional components are shared between major phylogenetic taxa, which supports the hypothesis that mechanisms of gene activation are intrinsically different in Bacteria and Eukaryota. Furthermore, because of our poor understanding of archaeal-specific transcriptional control, known archaeal TAP families are mostly shared with the other two domains. This complements experimental results showing that Archaea have a basal transcriptional apparatus similar to eukaryotes but their control of gene expression is bacterial-like (Bell *et al.*, 1999). Practical applications of the above observations include the identification of taxon-specific transcription factors as drug targets (Latchman, 1997).

Note added in proof

Subsequent to the analyses described here, (Riechmann *et al.*, 2000) performed a study of *Arabidopsis thaliana* transcription factors exploiting the complete genome sequence of the plant. This involved an analysis of the distribution of transcriptional regulators across three eukaryotic king-

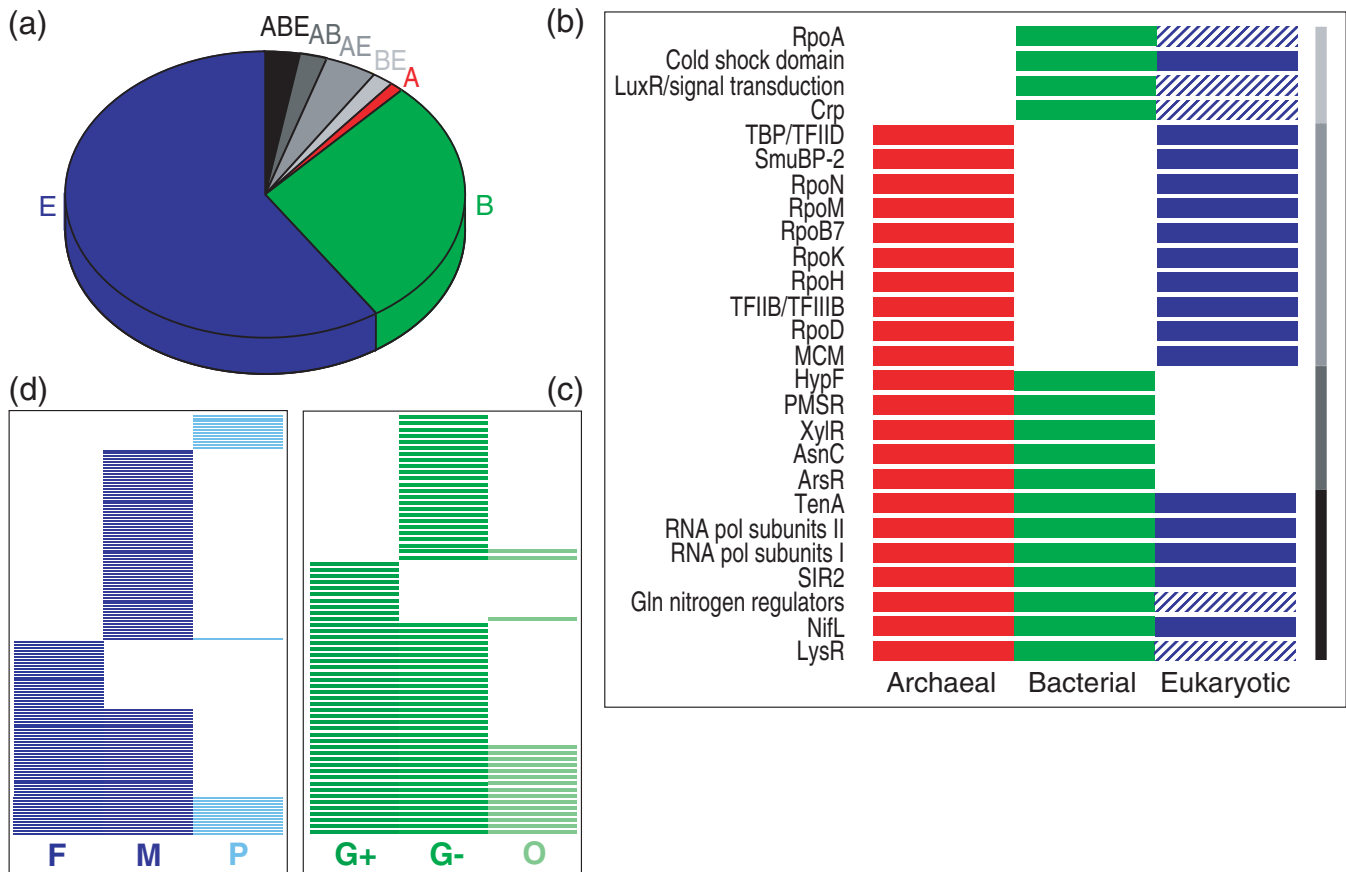


Fig. 1. TAP family distribution across the three domains of life. Sequences annotated with the word ‘transcription’ in the keyword and description fields of SWISS-PROT and SWISS-NEW and ‘transcription’ in the description field of SP-TrEMBL were extracted using SRS (<http://srs6.ebi.ac.uk>). 5894 transcription-associated proteins (TAPs) were obtained and filtered to remove regions of low complexity. The TAPs were clustered by sequence similarity using BLAST and subsequently Smith–Waterman analysis for non-symmetric BLAST hits—involving sequence randomisation, symmetrification of the similarity matrix and multi-domain detection (Enright and Ouzounis, 2000). A total of 2.3 million individual Smith–Waterman sequence comparisons were performed for non-symmetric BLAST hits. The ‘organism classification’ string for each TAP was extracted from its database entry and associated with the appropriate sequence cluster. Clusters containing three or more TAPs were annotated to provide functional descriptions of the whole family. Viral-only sequences and clusters were excluded from this analysis. All data are available on the WWW at: <http://www.ebi.ac.uk/research/transcription/clusters/>. (a) *Results of TAP clustering.* **A**, Archaea; **B**, Bacteria; **E**, Eukaryota. Numerical breakdown of chart (figures: number of families, in *italic*: number of TAPs in a particular grouping): ABE = 2.9% (7, 452); AB = 2.1% (5, 72); AE = 4.2% (10, 175); BE = 1.7% (4, 395); A = 1.2% (3, 20); B = 28.6% (69, 812); E = 59.3% (143, 2607). (b) *TAP families present in two or more primary domains.* Hashed rectangles indicate eukaryotic TAPs encoded by chloroplast or cyanelle genomes. The grey-scale bar on the right-hand side indicates the four non-unique domain groupings in (a). (c), *TAP family distribution in Gram-positive and Gram-negative bacteria.* **G+**, Firmicutes; **G-**, Proteobacteria; **O**, All other Gram-negative bacteria that are not Proteobacteria. Numerical breakdown: G+/G-/O = 21.7% (15, 325); G+/G- = 29.0% (20, 316); G+/O = 1.5% (1, 4); G-/O = 2.9% (2, 6); G+ = 13.0% (9, 41); G- = 31.9% (22, 120); O = 0.0% (0, 0). (d), *TAP family distribution in the eukaryote crown group.* **F**, Fungi; **M**, Metazoa; **P**, Plants. Numerical breakdown: F/M/P = 9.1% (13, 413); F/M = 20.9% (30, 631); F/P = 0.0% (0, 0); M/P = 0.7% (1, 316); F = 16.1% (23, 106); M = 44.8% (64, 1058); P = 8.4% (12, 83).

doms (Fungi, Metazoa and Plants) and showed that about 45% of *A.thaliana* transcription factor families are confined to plants. Hence, these data further strengthen our observation (based on TAP family distribution across the three domains of life) that the majority of TAP families appear to be taxon-specific.

REFERENCES

- Bell,S.D., Cairns,S.S., Robson,R.L. and Jackson,S.P. (1999) Transcriptional regulation of an archaeal operon *in vivo* and *in vitro*. *Mol. Cell*, **4**, 971–982.
- Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based

- on gene fusion events. *Nature*, **402**, 86–90.
- Enright,A.J. and Ouzounis,C.A. (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, **16**, 451–457.
- Gray,M.W. (1993) Origin and evolution of organelle genomes. *Curr. Opin. Genet. Dev.*, **3**, 884–890.
- Kyrpides,N.C. and Ouzounis,C.A. (1999) Transcription in Archaea. *Proc. Natl. Acad. Sci. USA*, **96**, 8545–8550.
- Latchman,D.S. (1997) How can we use our growing understanding of gene transcription to discover effective new medicines? *Curr. Opin. Biotechnol.*, **8**, 713–717.
- Mizuta,T.R., Fukita,Y., Miyoshi,T., Shimizu,A. and Honjo,T. (1993) Isolation of cDNA encoding a binding protein specific to 5'-phosphorylated single-stranded DNA with G-rich sequences. *Nucleic. Acids Res.*, **21**, 1761–1766.
- Riechmann,J.L., Heard,J., Martin,G., Reuber,L., Jiang,C.-Z., Keddie,J., Adam,L., Pineda,O., Ratcliffe,O.J., Samaha,R.R., Creelman,R., Pilgrim,M., Broun,P., Zhang,J.Z., Ghandehari,D., Sherman,B.K. and Yu,G.-L. (2000) *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.
- Struhl,K. (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, **98**, 1–4.
- Zhang,J.J. *et al.* (1998) Ser727-dependent recruitment of MCM5 by Stat1 α in IFN- γ -induced transcriptional activation. *EMBO J.*, **17**, 6963–6971.