3994–4006 *Nucleic Acids Research, 2005, Vol. 33, No. 13*
doi:10.1093/nar/gki709

Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains

S. Balaji, M. Madan Babu, Lakshminarayan M. Iyer and L. Aravind*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received April 8, 2005; Revised June 21, 2005; Accepted June 29, 2005

ABSTRACT

The comparative genomics of apicomplexans, such as the malarial parasite *Plasmodium*, the cattle parasite *Theileria* and the emerging human parasite *Cryptosporidium*, have suggested an unexpected paucity of specific transcription factors (TFs) with DNA binding domains that are closely related to those found in the major families of TFs from other eukaryotes. This apparent lack of specific TFs is paradoxical, given that the apicomplexans show a complex developmental cycle in one or more hosts and a reproducible pattern of differential gene expression in course of this cycle. Using sensitive sequence profile searches, we show that the apicomplexans possess a lineage-specific expansion of a novel family of proteins with a version of the AP2 (Apetala2)-integrase DNA binding domain, which is present in numerous plant TFs. About 20–27 members of this apicomplexan AP2 (ApiAP2) family are encoded in different apicomplexan genomes, with each protein containing one to four copies of the AP2 DNA binding domain. Using gene expression data from *Plasmodium falciparum*, we show that guilds of ApiAP2 genes are expressed in different stages of intraerythrocytic development. By analogy to the plant AP2 proteins and based on the expression patterns, we predict that the ApiAP2 proteins are likely to function as previously unknown specific TFs in the apicomplexans and regulate the progression of their developmental cycle. In addition to the ApiAP2 family, we also identified two other novel families of AP2 DNA binding domains in bacteria and transposons. Using structure similarity searches, we also identified divergent versions of the AP2-integrase DNA binding domain fold in the DNA binding region of the PI-SceI homing endonuclease and the C-terminal domain of the pleckstrin homology (PH) domain-like modules of eukaryotes. Integrating these findings, we present a reconstruction of the evolutionary scenario of the AP2-integrase DNA binding domain fold, which suggests that it underwent multiple independent combinations with different types of mobile endonucleases or recombinases. It appears that the eukaryotic versions have emerged from versions of the domain associated with mobile elements, followed by independent lineage-specific expansions, which accompanied their recruitment to transcription regulation functions.

INTRODUCTION

The transcription apparatus in eukaryotes shares several generic features with the functionally equivalent systems in the two prokaryotic super-kingdoms, the archaea and the bacteria. In both prokaryotes and eukaryotes, the component of the transcriptional machinery can be categorized into three major components: (i) the RNA polymerase complex and associated protein complexes required for initiation and elongation of the transcript. (ii) The basal transcription factors (TFs) that bind the core promoter of a gene and are required for the baseline expression of any gene. (iii) The specific TFs that bind various regulatory elements distinct from the core promoter element, and either activate or repress the transcription of the gene (1). In all the three super-kingdoms of life, the core RNA polymerase subunits are orthologous, although their domain architectures and accessory complexes might show considerable variability (2–4). The archaeal and eukaryotic super-kingdoms share the majority of their core basal TFs, such as TBP, TFIIB, TFIIE and MBF1, as opposed to the bacteria that possess distinctive basal TFs in the form of the sigma factors (5–10). However, in terms of specific TFs the archaea and bacteria are closer to each other (9,11).

*To whom correspondence should be addressed. Tel: +1 301 594 2445; Fax: +1 301 435 7794; Email: aravind@ncbi.nlm.nih.gov

The majority of specific TFs from bacteria and archaea possess versions of the helix–turn–helix (HTH) DNA binding domain that are more closely related to each other than to HTH domains of eukaryotic specific TFs (11,12). In contrast, the eukaryotes are very distinct in terms of the domain composition and the evolutionary affinities of their specific TFs. While certain distinctive versions of the HTH domain, such as the homeo, Forkhead (Fkh), Bright (ARID), the MYB, PSQ and paired domains, are prevalent in the eukaryotes, they possess numerous other highly expanded families of TFs with DNA binding domains unrelated to the HTH domain (12). The families include various Zinc-chelating families, such as the C2H2 Zn-finger and the fungus-specific C6 Zn-finger, various versions of the treble-clef fold, helical domains, such as the HMG, bZip and bHLH domains and other more complex folds, such as the VP1, AP2, GCM, TIG and cytochrome F fold domains (13–17).

Comparative genomic analysis of eukaryotic TFs has revealed that the major families of TFs in eukaryotic genomes have emerged principally through the process of lineage-specific expansions (15,18,19). As a result of this, the major lineages of the eukaryotic crown group may not even share a DNA binding domain in their most prevalent TF families. For example, the most prevalent TFs in fungi contain the C6 binuclear Zn-finger, whereas this Zn-finger is completely absent in the plants and animals, which instead have their own unique TFs, like those with the VP1 domain and the nuclear hormone receptor Zn-finger domains, respectively (15,18,19). However, the chromatin level regulatory apparatus comprising diverse families of chromosomal proteins is strongly conserved across the crown group eukaryotes (15,20,21).

Previous studies on eukaryotic lineages, such as the Apicomplexa and the Diplomonads, that branch outside of the crown group showed a surprising dearth of detectable specific TFs in their proteomes, despite the presence of the expected set of basal TFs (22,23). Detailed analysis of the apicomplexan genomes of *Plasmodium falciparum* and *Cryptosporidium parvum* showed that they entirely lacked conserved DNA binding domains of specific TFs found across the eukaryotic crown group, such as the homeo, bZip, bHLH and Fkh domains (22–24). Very rare representatives of certain other families, which are common in the crown group, such as the C2H2 Zn-finger and E2F domains, were detected in these apicomplexans (22,23). The ratio of the total number of genes in the genome to the total number of TFs in free-living yeasts is in the range of 25–30 (18,23). Even though the parasitic apicomplexans possessed gene counts comparable with the free-living yeasts, the ratio of the number of genes to the total number of detectable TFs was in the range of 350–800 (23).

While a part of this discrepancy could be explained on the basis of the parasitic lifestyle of the apicomplexans, which probably does not require much intricate regulation relative to the homeostatic challenges faced by free-living organisms, it is paradoxical with respect to other observations. First, the apicomplexans possess an extensive complement of structural and regulatory chromosomal proteins, and cytoplasmic signaling proteins, such as kinases and GTPases, that are found comparable in numbers to the crown group eukaryotes (22,23). Second, they show a complex developmental cycle within their hosts, which would suggest the requirement for transcriptional regulation, and consistent with this gene expression studies have revealed an intricate developmentally regulated cascade of expression (25,26). The possible solutions (which are not mutually exclusive) for this paradox are: (i) there are undetected specific TFs that are only distantly related or unrelated to previously known DNA binding domains. (ii) There are alternative regulatory mechanisms that do not depend on TFs, such as chromatin level regulation and post-transcription regulation by non-coding RNAs.

To understand these possibilities, we conducted a systematic analysis of the predicted apicomplexan nuclear proteins. As a result, we report the discovery of a novel family of apicomplexan DNA binding proteins with a specific version of the AP2-intregrase type domains. This finding provides the first serious candidates for the principal specific TFs of the apicomplexans and helps in resolving the above-stated paradox. Furthermore, the analysis of the apicomplexan members of the AP2-intregrase DNA binding domains (27–29) provides a glimpse of the complex evolutionary history of this superfamily and reinforces the general concept of repeated origins of TFs from selfish mobile elements.

MATERIALS AND METHODS

The non-redundant (NR) database of protein sequences (National Center for Biotechnology Information, NIH, Bethesda, MD) was searched using the BLASTP program (30). Iterative database searches were conducted using the PSI-BLAST program with either a single sequence or an alignment used as the query, with the PSSM inclusion expectation (*E*)-value threshold of 0.01 (unless specified otherwise); the searches were iterated until convergence. Hidden Markov models (HMMs) were built from alignments using the hmmbuild program and searches carried out using the hmmsearch program from the HMMer package (31). For all searches with compositionally biased proteins, the statistical correction for this bias was employed (32). Entropy analysis of proteins was carried out using the SEG program (33). Multiple alignments were constructed using the T_Coffee and MUSCLE programs, followed by manual correction based on the PSI-BLAST results (34,35). Similarity-based clustering of proteins was carried out using the BLASTCLUST program (ftp://ftp.ncbi.nih.gov/blast/documents/README.bcl). All large-scale sequence and structure analyses procedures were carried out using the TASS package (L. Aravind, V. Anantharaman, S. Balaji and L. M. Iyer, unpublished data), which operates similar to the SEALS package.

Protein secondary structure was predicted by using a multiple alignment to generate a HMM and PSSM, which were then used by the JPRED program to produce a final structural prediction with 72 % or great accuracy (36,37). Protein structure manipulations were performed using the Swiss-PDB viewer program (38) and the ribbon diagrams were constructed using the PYMOL program (39). For structural searches of the PDB database the DALI and SSM programs were used (40–42). The studies on clustering-based DALI *Z*-scores have suggested that *Z*-scores >10 are characteristic of obvious relationships, such as those between two closely related proteins of the same family. Between *Z*-scores 10 and 6, typically, the relationships correspond to more distant relationships that might be recovered through sequence profile analysis and searches using HMMs. *Z*-scores <3 fall in the realm of

remote structural relationships and require additional analysis, such as comparisons of topologies to make further inference regarding these relationships (40,41).

Phylogenetic analysis was carried out using the neighbor-joining and minimum evolution (least squares) methods using the MEGA package (43).

Gene expression data for the complete 48 h intraerythrocytic developmental cycle (IDC) was downloaded from http://biology.plosjournals.org/archive/1545-7885/1/1/supinfo/10.1371_journal.pbio.0000005.sd002.txt.

Missing data points, which were few in proportion compared with the experimentally measured data, were estimated using KNNImpute (44). Genes were clustered into groups based on their expression pattern using the *k*-means clustering procedure, at various *k*-values, available in the cluster program (45). The expression profile for the clustered genes was visualized using the program matrix2png (46). Correlation coefficients between the expression profiles of the ApiAP2 genes and the other genes were calculated using custom-written perl scripts.

RESULTS AND DISCUSSION

Identification of an apicomplexan family of the AP2-integrase DNA binding domains

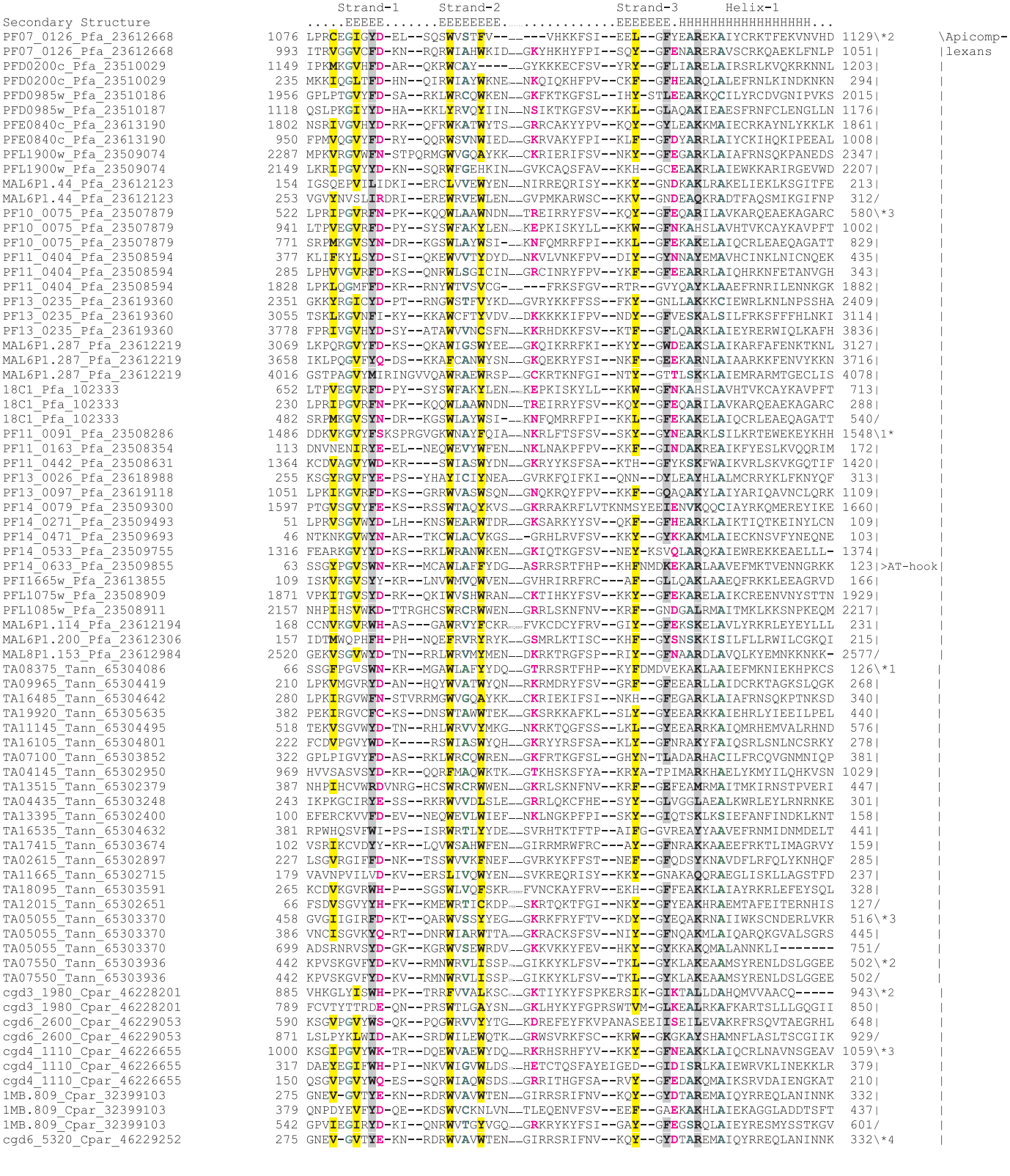
In order to gain a reliable estimate of the counts of nuclear proteins in the *P.falciparum* proteome, we systematically analyzed all the predicted proteins for known DNA binding domains and motifs, such as those found in specific TFs and components of the chromatin remodeling machinery, using previously made PSSMs and HMMs. As previously reported, most of these searches did not recover any candidates for specific TFs, such as Homeo, Forkhead, bZip, bHLH and MADS domains, which are commonly encountered in crown group eukaryotes (22,23). The protein PF14_0633 (GenBank gi: 23509855) was noted to contain a small DNA binding motif, the AT-hook (residues 36–46), which is found in numerous chromosomal proteins (47). As the AT-hook is often found linked to several other larger globular DNA binding domains in the same polypeptide, we analyzed PF14_0633 with the SEG program to identify potential globular domains (33). A prominent globular region of ~60 amino acids was predicted immediately C-terminal to the AT-hook module (region 63–123), and BLAST searches of the NR protein database (NCBI) with this segment showed that it was conserved across diverse species of the genus *Plasmodium*, and also in the other apicomplexan genera *Theileria* and *Cryptosporidium*. Further iterative PSI-BLAST searches with this globular segment as a seed recovered numerous (>15 unique proteins) statistically significant hits ($E < 0.01$) that encompassed the entire length of the query from each of the species of *Plasmodium*, *Theileria* and *Cryptosporidium*. For example, regions showing sequence similarity to PF14_0633 were recovered in *Plasmodium* MAL6P1.287 with $e = 10^{-7}$ in iteration 4, in *Theileria* TA08375 with $e = 10^{-4}$ in iteration 3 and in *Cryptosporidium* cgd6_1140/Chro.60146 with $e = 10^{-4}$ in iteration 6. This suggested that the globular segment was likely to represent a globular domain of case study ~55–65 amino acids that has undergone a lineage-specific expansion in Apicomplexa.

To further explore the evolutionary affinities of these domains, we constructed a position-specific score matrix that included all the significant hits from apicomplexans and used it to iteratively search the NR database with the PSI-BLAST program. These searches recovered significant hits to a variety of proteins from plants and mobile DNA elements, such as the floral homeotic protein Q from *Triticum* ($e = 2 \times 10^{-3}$) and 49L, an endonuclease of the EndoVII fold (also called HNH-type endonucleases) from *Xanthomonas oryzae* phage Xp10 ($e = 1.5 \times 10^{-3}$). In these proteins, the PSI-BLAST HSPs mapped completely to the AP2 DNA binding domain, which is found in plant developmental TFs, fused to several bacterial EndoVII fold endonucleases and integrases, such as the tn916 integrase (27–29). To test this potential relationship further, we initiated reciprocal searches from different AP2 domains and were able to recover members of the above-detected group of apicomplexan proteins with significant *e*-values. For example, the protein DP2593 from the bacterium, *Desulfotalea psychrophila*, with two AP2 domains, recovers the *Plasmodium* protein MAL6P1.287 with $e = 10^{-4}$ (iteration 2) and the *Cryptosporidium* protein cgd6_1140/Chro.60146 with $e = 10^{-8}$ (iteration 5). A multiple alignment of all the above-detected versions of the conserved globular domain from apicomplexans was prepared and used to predict the secondary structure of the domain using the JPRED and PHD programs. The predicted secondary structures for the apicomplexan proteins showed a conserved core of three consecutive strands and a C-terminal helix, which is congruent with the (sequence of) secondary structures of AP2-integrase DNA binding domains (AP2-IDBDs) (Figure 1). Furthermore, a HMM prepared from this multiple alignment was used to search the *Arabidopsis* proteome, and it recovered several hits to the AP2 domains ($e = 10^{-2}$–10^{-3}). Taken together, these observations suggested that the conserved globular domain found in PF14_0633 and its numerous apicomplexan homologs defines a novel family of the AP2-integrase DNA binding domain superfamily, which we hereinafter term the ApiAP2 family.

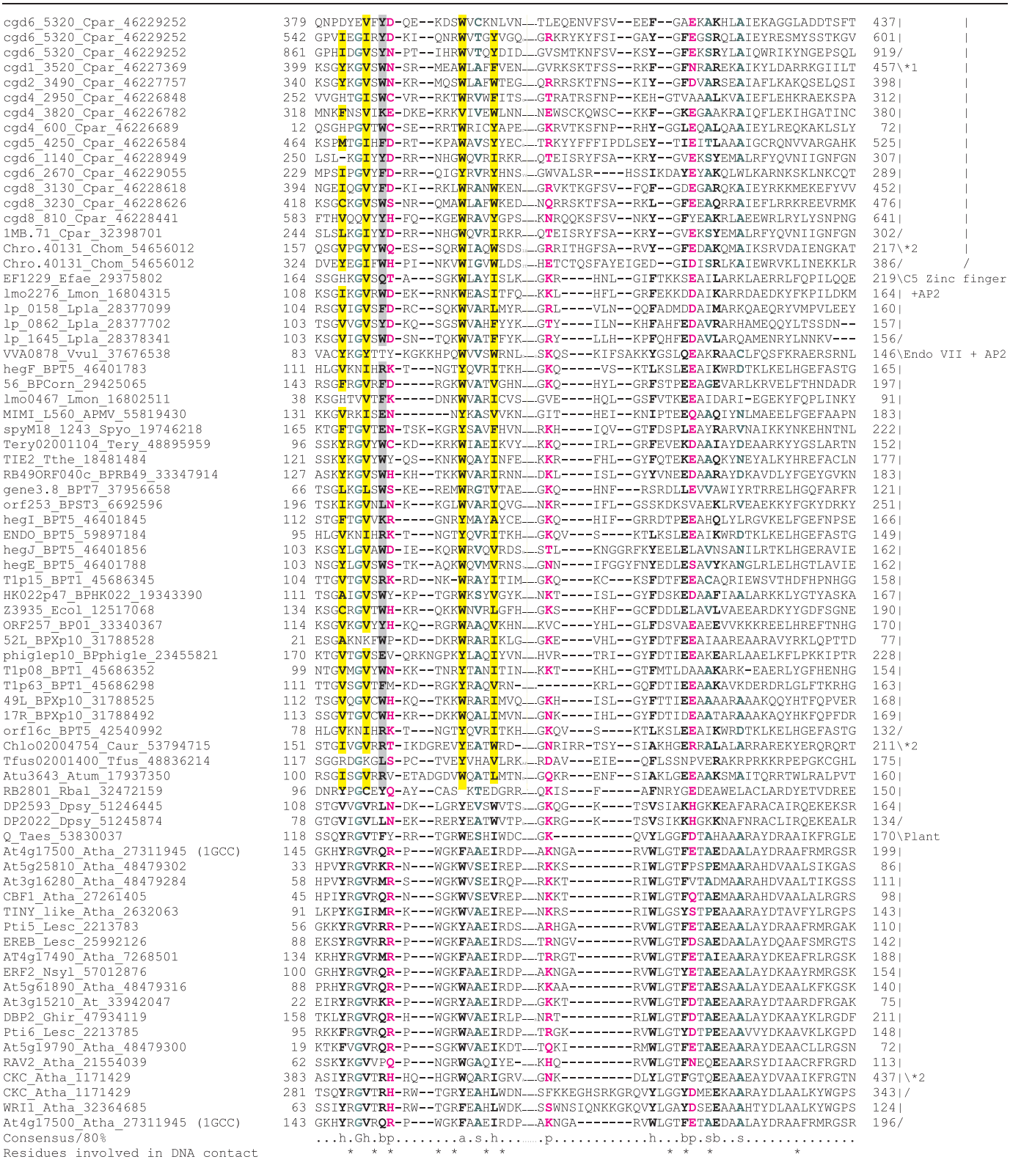
Characterization of the sequence and structure specializations of the ApiAP2 superfamily

To investigate the sequence and structure features of the ApiAP2 family, we compared the conservation patterns derived from 211 ApiAp2 domains from *Plasmodium*, *Theileria* and *Cryptosporidium* with those derived from multiple alignments of the plant AP2 proteins, those associated with EndoVII fold nucleases and other bacterial families (see below). These conservation patterns were also superimposed onto the NMR structure of the AP2 domain of the *Arabidopsis* ethylene response TF (ATERF1, PDB: 1GCC) (48) to understand their structural implications. There are 12 residues that show a strong conservation in at least 241 representatives of the 285 AP2 domains from the test-set that included diverse representatives of all the above classes, in addition to the ApiAP2 domains (Figure 1). This conservation pattern mainly corresponds to the residues that form key stabilizing hydrophobic interactions and determine the path of the backbone in the three strands and the helix of the AP2 domain (Figure 1), suggesting that the core fold of the ApiAP2 proteins would be identical to the plant, viral and bacterial AP2 domains.

However, the ApiAP2 proteins have a relatively long insert between the strands 2 and 3 (Figure 1), which is only seen in few other members of the AP2-intergrase fold, such as the second AP2 domain of the plant proteins typified by the *Arabidopsis* Wrinkled1 (49,50), and a novel member of this fold that we detected in the PI-SceI homing endonuclease (51). The conservation pattern within this insert suggests that it is likely to form a hairpin, which sticks out of the core fold in the extended conformation, similar to what is observed in the structure of the PI-SceI domain of AP2-IDBD fold (Figure 2).



Both the plant AP2 domains and those associated with the nuclease domains of the EndoVII fold have been shown to bind GC-rich sequences (28,48,52,53). In particular, forms like ATERF1 bind copies of the GCC-box motif (e.g. GCCGCC used in the ATERF1–DNA complex, whose structure was solved) (48). A total of 11 residues involved in making contacts with DNA were identified using the ATERF1 NMR structure, and the conservation and relevance of the contacts for sequence specificity were assessed with respect to the ApiAP2 domains (Table 1 and Figure 3). Of these



11 residues, 7 residues (R150, R152, W154, E160, R162, R170 and W172 in the ATERF AP2 domains structure) form contacts with the bases in the GCC boxes, while the rest of the residues are involved in backbone contacts and non-specific interactions. The average pairwise distance within the ApiAP2 family is much greater than the average pairwise distance within the plant AP2 domain family [2.6 versus 1.2; measured using the JTT score matrix (54)]. Accordingly, the majority of plant AP2 domains conserve the DNA-contacting residues seen in ATERF1, whereas there is considerably higher variability within the ApiAP2 family, suggesting a greater diversity in their binding sites.

In the ApiAP2 family, the positions corresponding to E160 and W172 are respectively occupied, most frequently, by polar amino acids with an oxygen in the side-chain and an aromatic residue (Figure 1). Thus, these positions largely retain a similar character in both the families of AP2 domains and are unlikely to contribute significantly to differential sequence specificity. However, the arginine at the end of strand 1 (position corresponding to R152 in the ATERF1 structure), which is critical for the recognition of guanine in one of the GCC boxes in the plant proteins (Table 1 and Figure 3), is replaced by a D or N in the majority of ApiAP2 proteins. This R interacts with the guanines via interactions with the oxo-groups and a D or N at this position would be more conducive for interaction with the amino groups of adenine. Similarly, the R in the middle of strand 1 (corresponding to position R150 in the ATERF1 structure), which is also critical for recognizing one of the guanines in same GCC-box as that recognized by R152 (Figure 3), is quite frequently replaced by a tyrosine or serine in the ApiAP2 family (Figure 1). A Y or S in this position is again unlikely to favor specific interactions with guanine and might actually favor an interaction with the amino group of adenine. These observations would indicate that at least a subset of the ApiAP2 proteins is likely to bind AT-containing target sequences. The ApiAP2-specific loop between strands 2 and 3 contains ~2–3 positively charged residues within 6–10 residues (Figure 1). Extrapolating on the basis of PI-SceI and ATERF1 DNA binding domains of AP2-IDBD fold, we suggest that this insert is likely to lie along the backbone of the DNA, with the positively charged residues forming multiple contacts with the phosphates. Thus, this insert is likely to play an important role in determining the affinity of the ApiAP2 domains. A search for GCC or other G and C containing oligonucleotides using Alignace (55) did not uncover any such motifs upstream of most of the genes in *P.falciparum*. Furthermore, a systematic search using a sliding window approach to identify local zones of GC richness in the intergenic regions potentially upstream of basal promoters failed to reveal any consistent patterns. The intergenic regions of the many apicomplexans, in particular the genus *Plasmodium*, are extremely AT-rich. These observations are consistent with non-GC-rich binding sites for many of the ApiAP2 proteins. In the absence of further experimental data, the extraordinary AT richness of *P.falciparum* intergenic regions makes it difficult to identify candidate binding sites for these ApiAP2 proteins by using a combination of motif searches and co-expression profiles.

Comparative genomics and domain architectures of ApiAP2 proteins

We systematically searched for copies of the ApiAP2 domains in the previously published apicomplexan genomes using PSI-BLAST PSSMs and HMMs. We detected between 35 and 43 copies of domain in *P.falciparum*, *Plasmodium chabaudi*, *Plasmodium yoelii* and *Plasmodium berghei*, 25 copies in *Theileria annulata*, and 25–30 copies in *C.parvum* and *Cryptosporidium hominis*. We used single linkage clustering with the BLASTCLUST program and neighbor-joining with the MEGA program (43) to classify the ApiAP2 domains into orthologous groups. As a result, we obtained 40 reliable orthologous groups of ApiAP2 domains with at least one representative from any three of the four species in the *Plasmodium* genus (data not shown; see Supplementary Material). This observation taken together with the presence of 43 copies of the ApiAP2 domain in *P.falciparum* (the best annotated of the *Plasmodium* species) suggests that the discrepancy in counts in the different species is likely to be consequence of lower quality of sequence data and assembly in the other three species. Similarly, the difference between two *Cryptosporidium* species appears to be a consequence of the lower quality of the *C.hominis* genome sequence. The copies of the ApiAP2 domains were traced to ~27 different proteins in *Plasmodium*, 21 in *Theileria* and 19 in *Cryptosporidium*, each containing one to four repeats of the ApiAP2 domain (Figure 4). An analysis of chromosomal distribution of the ApiAP2 genes shows that they are not clustered on any particular chromosome or chromosomal regions, unlike genes for several cell surface protein families in Apicomplexa (22). An examination of the orthologous clusters of the ApiAP2 domains shows that 16 of them from at least 15 distinct proteins are shared by *Theileria* and *Plasmodium*, whereas 11 of them from at least 9 distinct proteins are shared by *Cryptosporidium* and *Plasmodium*. Given that *Cryptosporidium* and *Plasmodium* represent a very early divergence

Figure 1. Alignment of AP2 domains. Proteins are denoted by their gene names, species abbreviations and GenBank identifier (gi) numbers. The number of AP2 domains in a polypeptide is shown to the right of the alignment. Residues involved in contacting DNA in the solution structure of the AP2 domain (pdb id: 1GCC) are shown below the alignment. The secondary structure was derived from the solution structure of the AP2 domain (PDB ID: 1GCC). E represents a β strand; H, helix. The coloring reflects the conservation profile at 80% consensus. The coloring scheme and consensus abbreviations are as follows: h, hydrophobic (h: ACFILMVWY) and a, aromatic (a: FWY) residues shaded yellow; b, big (LIYERFQKMW) residues shaded gray, s, small (AGSVCDN) residues colored green; and p, polar (STEDKRNQHC) residues colored magenta. Species abbreviations are as follows: APMV: Acanthamoeba polyphaga mimivirus; Atha: *A.thaliana*; Atum: *Agrobacterium tumefaciens*; BP01: Bacteriophage Felix 01; BPCorn: Mycobacteriophage Corndog; BPHK022: Enterobacteria phage HK022; BPRB49: Enterobacteria phage RB49; BPST3: *Streptococcus thermophilus* bacteriophage ST3; BPT1: Enterobacteria phage T1; BPT5: Bacteriophage T5; BPT7: Enterobacteria phage T7; BPXp10: *X.oryzae* bacteriophage Xp10; BPphig1e: Bacteriophage phig1e; Caur: *Chloroflexus aurantiacus*; Chom: *Cryptosporidium hominis*; Cpar: *C.parvum*; Dpsy: *D.psychrophila*; Ecol: *Escherichia coli*; Efae: *Enterococcus faecalis*; Ghir: *Gossypium hirsutum*; Lesc: *Lycopersicon esculentum*; Lmon: *Listeria monocytogenes*; Lpla: *Lactobacillus plantarum*; Nsyl: *Nicotiana sylvestris*; Pfa: *Plasmodium falciparum*; Rbal: *Rhodopirellula baltica*; Spyo: *Streptococcus pyogenes*; Taes: *Triticum aestivum*; *Theileria annulata*; Tery: *Trichodesmium erythraeum*; Tfus: *Thermobifida fusca*; Tthe: *Tetrahymena thermophila*; Vvul: *Vibrio vulnificus*.

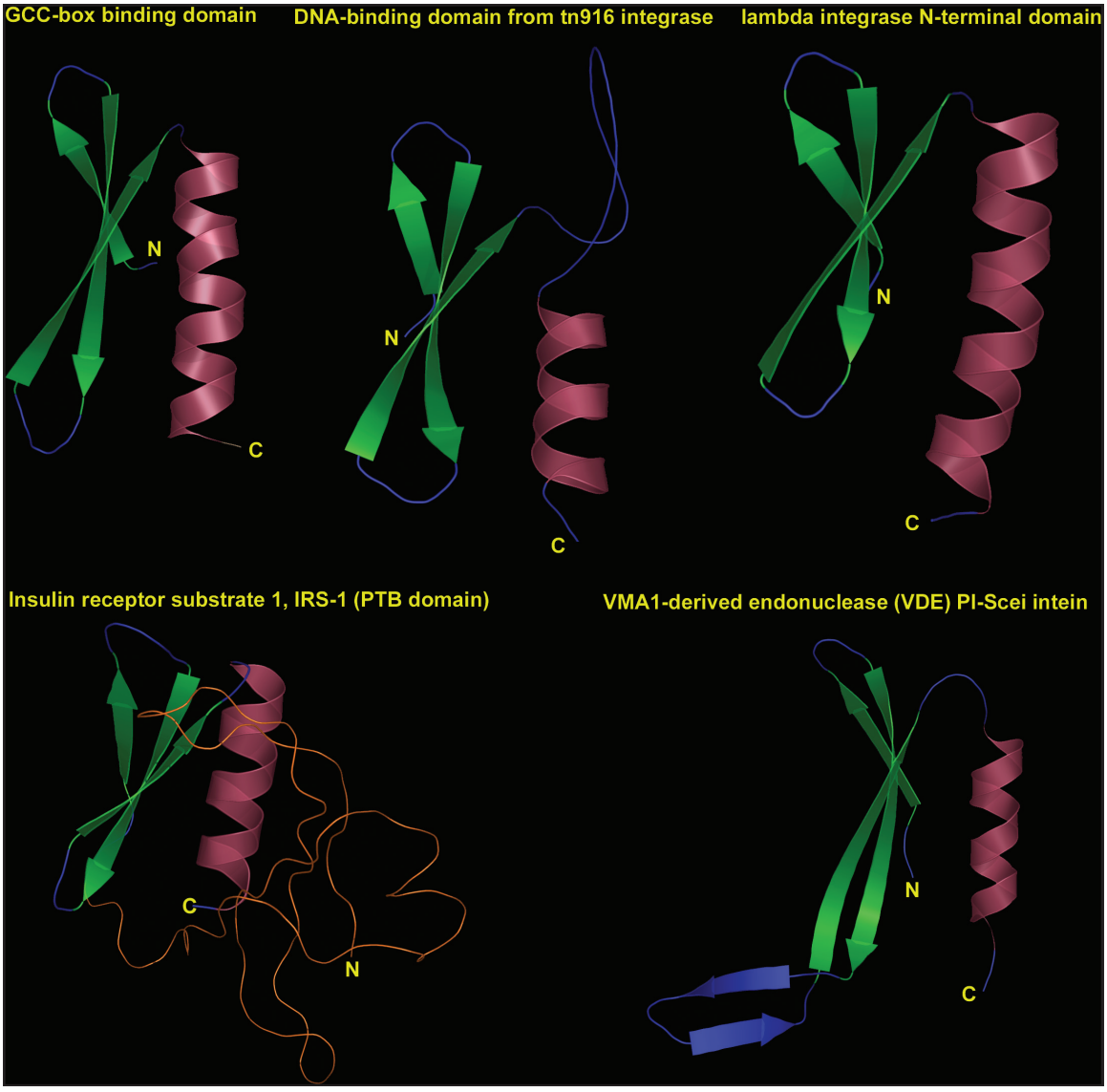


Figure 2. Structures of different domains of the AP2-IDBD fold. Strands and helices of the AP2-IDBD fold are colored green and pink, respectively. PDB ids for the displayed structures as follows; 1gcc: GCC-box binding domain; 1bb8: tn916 integrase DNA binding domain; 1kjk: lambda integrase N-terminal domain; 1qqg: Insulin receptor substrate 1 (IRS-1); 1lwt: PI-SceI homing endonuclease DNA binding domain.

Table 1. Most frequent amino acids at DNA-contacting positions (numbered 1–11 and labeled according to 1gcc), according to their order of occurrence, deduced from the solution structure of GCC-box binding domain of ATERF1 (PDB ID 1gcc) and the comprehensive multiple alignment are shown

Family	1 R147	2 R150	3 R152	4 W154	5 K156	6 E160	7 R162	8 R170	9 W172	10 T175	11 Y186
ApiAP2	K 27	Y, R, S 24,15,13	D, N 53, 14	Q, K, R 18,17,17	R, S, A 23,18,17	Q, E, Y, T 14,10,10,9	Y, K 20,17	K 45	Y, F 52,22	G 88	F, C 28,19
Plant AP2 family	R {B} 80	R {G 20} 70	R {G 5,G 20,G 21} 65	W {T 3,A 4} 60	K, R {B} 55, 35	E {C 7} 65	R {G 17,G 18} 65	R {G 8} 75	W {G 5,C 6} 70	T {B} 70	Y {B} 80
Bacterial/Viral AP2 domains	K, T, R 27,20,14	S, T, R, Y 27,13,10,10	N, D, H 17,15,15	T, R, S 23,19,13	K, R 44,23	R, Q, T 35,13,13	T, R 21,19	G, K 50,22	F 65	K, L, I 23,19,15	R, A, E 23,19,15

The protein–DNA contacts were identified by using the hydrogen-bond length (3.5 Å) and apolar interaction distance cut-off of 5 Å. The frequencies of occurrence of the amino acids, approximately to nearest integer percentage values, are given directly below the corresponding amino acid codes in each of the columns. Interacting bases and their nucleotide sequence numbers (as in 1gcc) are shown within curly brackets in each of the DNA-contacting positions for the plant protein sequences. The symbol 'B' within the brackets denotes interactions only to backbone phosphate groups.

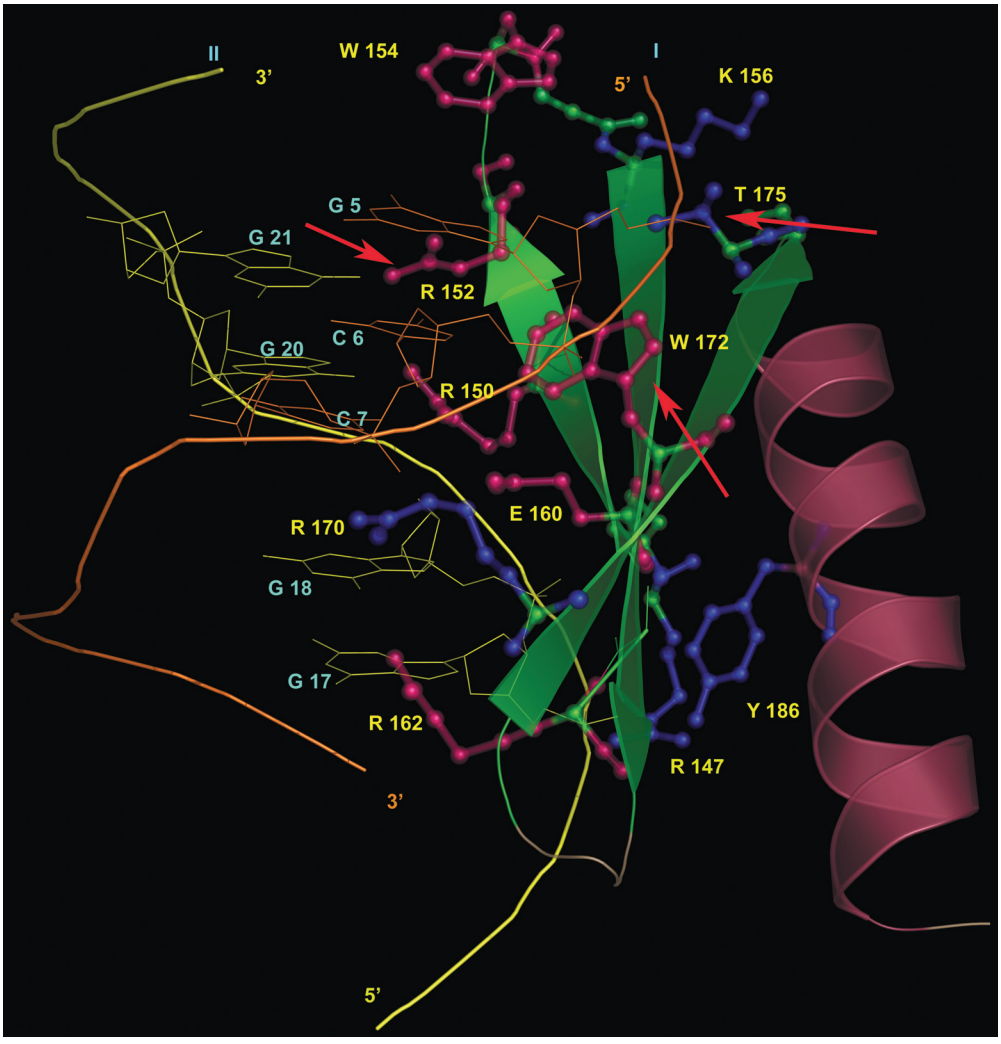


Figure 3. DNA interactions of the AP2 domain. The solution structure of the *A.thaliana* GCC-box binding domain in complex with DNA (PDB Id: 1gcc) is shown. Strands are colored green and the helix is colored pink. Complementary DNA strands are labeled I and II and colored orange and yellow, respectively. The side-chains of DNA-contacting residues are displayed in the ball and stick format. Residues that interact with DNA bases are colored pink and those that predominantly interact with the DNA backbone are colored blue. Red arrows indicate positions that are well conserved in the ApiAP2 family (see Figure 1 and Table 1 for the equivalent residues in the ApiAP2 proteins).

within Apicomplexa (56), it is likely that the common ancestor of Apicomplexa already possessed at least nine members of the ApiAP2 family. The higher number of orthologous ApiAP2 domains shared by *Plasmodium* and *Theileria* supports a closer relationship between these two lineages within Apicomplexa. This is consistent with phylogenetic studies, which have suggested an apicomplexan crown group that includes the piroplasms (*Theileria*) and hemosporidians (*Plasmodium*) to the exclusion of the basal lineages, the gregarines and *Cryptosporidium* (56). Thus, starting from a core set of at least nine proteins inherited from the ancestral form, the ApiAP2 family appears to have proliferated further through independent duplications as the different apicomplexan lineages emerged.

In terms of domain architectures, the majority of members of the ApiAP2 family contain a single AP2 domain, which is often the only globular domain in the entire protein. Furthermore, ApiAP2 proteins with 2–4 AP2 domains are also encountered in all the apicomplexan genomes (Figure 4). The AT-hook is the only other DNA binding motif that is found in association with the AP2 domain in a few apicomplexan proteins (Figure 4) and is consistent with the similar combination of the AT-hooks with other globular DNA binding domains (47). In this respect, the ApiAP2 family is similar to the plant TFs of the AP2 family, which also always contains single or duplicated AP2 domains as the principal globular domain/s in the protein. Outside of apicomplexans and plants, similar duplicate AP2 domain proteins are encountered only in a small family of bacterial proteins typified by the DP2593 from *D.psychrophila* (Figure 4). All the other AP2 domains from bacteria, viruses and mobile DNA elements contain a fusion of the AP2 domain with the EndoVII nuclease, at least two distinct members of lambda integrase superfamily (namely the phage lambda integrase proper and the tn916-type integrases, which are closer to the XerC/D recombinases) and one to two copies of a novel cysteine-rich domain with five conserved cysteines (e.g. lmo2276) (Figure 4). These distinctive domain architectural themes lend support to the idea that

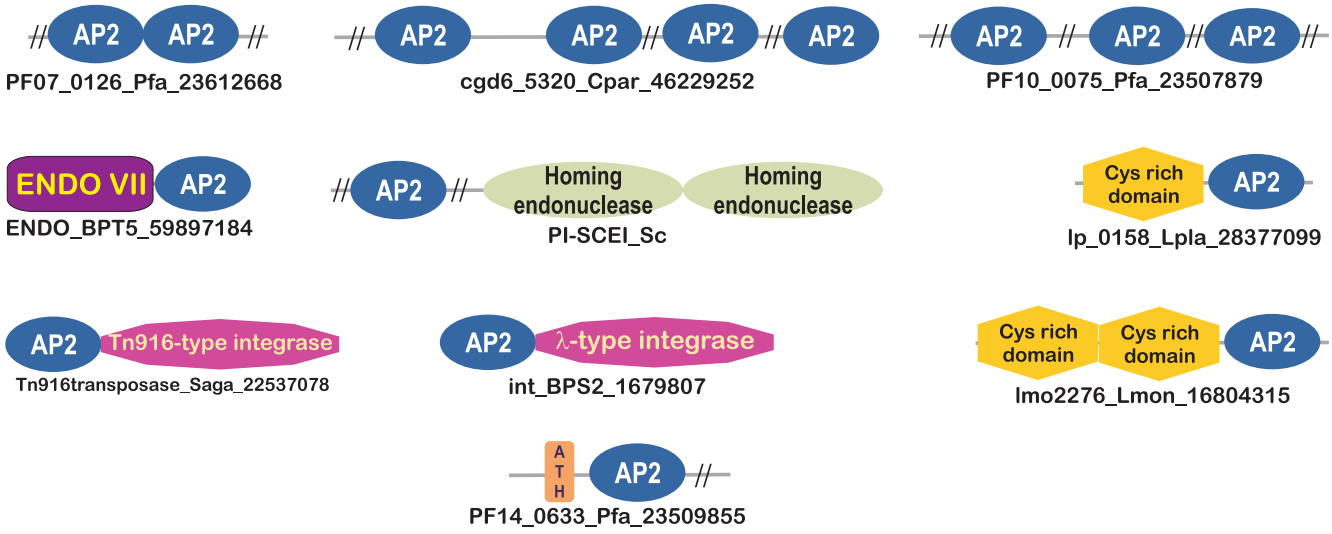


Figure 4. Domain architectures of AP2 domain proteins. Domains are represented by their standard notations. ATH represents the AT-hook. The protein naming scheme and species abbreviations are as in Figure 1.

members of the ApiAP2 family are specific TFs of the apicomplexans, rather than integrases of mobile DNA elements.

Gene expression patterns and the potential role for the ApiAp2 family in regulating life-cycle progression in apicomplexans

In order to further understand the biological functions of the ApiAp2 proteins, especially in the context of the complex life-cycles of the apicomplexan parasites, we exploited the high-throughput gene expression data obtained for the asexual IDC of *P.falciparum* (26). This and other studies have shown that the gene expression in the IDC of *P.falciparum* occurs in a continuous cascade with the induction of most genes occurring just once in the cycle, only at the time when their products are required (25,26). We found that 22 of the 26 genes encoding ApiAP2 proteins in *P.falciparum* were expressed in different stages of the IDC. Many genes were represented by more than one sequence tag that showed temporally consistent expression patterns, suggesting that the underlying expression data were sufficiently robust to make conclusions about stage-specific gene expression. In order to get a better picture of the stage-specific expression of the ApiAP2 genes, we clustered the genes based on their expression patterns using *K*-means clustering with pairwise Euclidean distance metric at various values of *K*. At $K = 5$, this procedure gave rise to four major clusters, each with 4–6 distinct ApiAP2 genes, that approximately corresponded to four major developmental stages, such as the ring stage, the trophozoite, early schizonts and the late schizont–merozoite stage (Figure 5). This indicates that different ApiAP2 genes may indeed function in specific developmental stages, and this observation is again consistent with their being specific TFs. The fifth cluster, however, contains only two genes that show anomalous expression patterns. These two genes showed apparent elevated expression in two discontinuous developmental stages. However, it is currently not clear if this biphasic expression is a genuine signal or not. The remaining ApiAP2 genes of *P.falciparum*, which were not detected in the IDC expression profiles, are probably uniquely utilized for other stages, such as intra-hepatocytic and sexual development, or in the insect vector. However, due to the absence of comparable expression data for these stages, we were unable to verify this possibility.

The striking differential expression of the ApiAP2 genes in specific developmental stages strongly suggests that they could mediate transcriptional regulation of stage specific genes. Within each stage-specific guild, individual ApiAP2 genes show further slight temporal differences in their expression patterns. This suggests that even within a given developmental stage a more complex combinatorial interplay between different specific TFs of the ApiAp2 family could set up expression patterns of particular genes. A comparison of the expression patterns of members of the ApiAP2 family with that of the rest of the genes might provide hints regarding the genes whose expression they might regulate. In particular, those genes showing strongly correlated expression (either positive or negative) with a particular guild of ApiAP2 genes might be regulated and maintained in that expression state by the products of that guild. The *K*-means clustering of all other genes (excluding the ApiAP2 genes) with $K = 5$ resulted in the detection of four major clusters correlating well with the four major expression classes of the ApiAP2 genes and the four developmental stages. A comparison of these expression profiles with the ApiAP2 genes might help in narrowing the potential target genes for the *P.falciparum* ApiAP2 genes (see Supplementary Material). Interestingly, nine of ApiAP2 genes expressed in the IDC of *P.falciparum* have potential orthologs in *Cryptosporidium*, and they are found in expression guilds corresponding to each of the IDC stages. Although the intracellular life-cycles of the two parasites show many specific differences, they follow an overall similar pattern of developmental progression that is also observed in other apicomplexans. If the orthologous ApiAP2 genes shared by *P.falciparum* and *Cryptosporidium* show generally similar expression patterns, then it is likely that their products regulate some of the common aspects of apicomplexan development. In contrast, the lineage-specific members

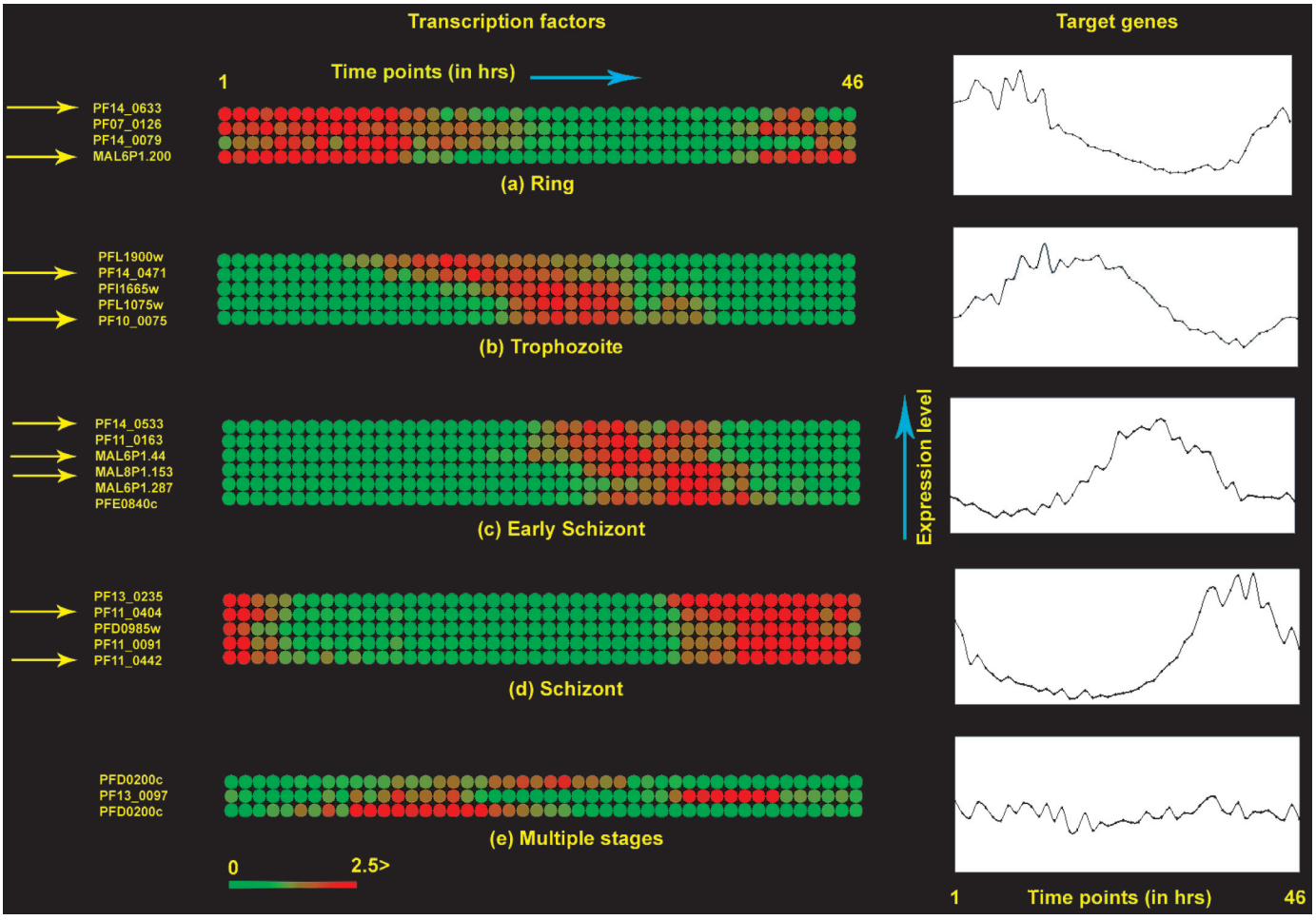


Figure 5. Expression patterns of AP2 proteins. Stage-specific expression of the ApiAp2 TFs and their potential target genes during the IDC. Microarray gene expression data were available for 46 timepoints as shown (26). Using *K*-means clustering, the predicted ApiAp2 TFs were grouped into five clusters. The first four clusters correspond to the four major developmental stages: (a) ring (b) trophozoite (c) early schizont and (d) schizont, whereas the fifth cluster (e) consists of genes that show the expression at two discontinuous developmental stages. Gene names for the ApiAp2 domain containing proteins are given by the sides, and an arrow next to the gene name indicates the presence of an ortholog in *Cryposporidium*. Note that there is at least one TF from each stage that has an ortholog in *Cryptosporidium*. The graphs on the right represent the average expression profile of non-ApiAp2 genes that show a high correlation in their expression profile with the ApiAp2 genes. The expression of such genes in a stage-specific manner suggests that these genes could be the potential targets for the predicted TFs.

would be expected to contribute to the taxon-specific diversity in gene expression.

The ApiAP2 domains and the evolutionary radiation of the AP2-IDBDs in prokaryotes and eukaryotes

Previous studies based on sequence analysis had identified two major families of AP2 domains, such as the plant TF family and the EndoVII fold (HNH) homing endonuclease-associated family (27–29). Structure comparisons had also identified two more closely related families, such as those associated with the catalytic domains of the lambda-type integrases and the tn916-type integrases (see SCOP database). Our sequence profile searches identified three new families namely the ApiAP2 family, and two bacterial families typified by the DP2593 (with two AP2 domains) and lmo2276 (fused to a Zn-chelating domain with five conserved cysteines) proteins. To identify other more divergent representatives, we conducted structural similarity searches using the DALI and SSM program. The AP2-IDBDs have a simple topology, and topologically equivalent units are found as sub-structures in larger domains (e.g. the RNAse H domain). Hence, we filtered our hits by performing reciprocal searches with each of the hits, and only considering those cases where the three-strand-helix unit formed a self-contained module or a distinct domain. One previously un-recognized AP2-IDBD identified in these searches was the DNA binding domain of the intein-associated homing endonuclease PI-SceI (Figure 2). This domain (corresponding to region 82–150 in PDB: 1lwt) occurs at the N-terminus of the two tandem homing endonuclease domains, which are unrelated to the EndoVII fold (HNH) endonucleases, and contacts DNA in manner very similar to the plant AP2 domains. Thus, the PI-SceI DNA binding domain represents the fourth independent instance in which AP2 domains are associated with a distinct endonuclease domain.

Another more intriguing hit, which was consistently recovered, was to C-terminal module of domains with the PH-like fold. The PH-like fold includes a variety of eukaryote-specific domains, such as the PTB, PH, Ran-binding and EVH1 domains that are involved in a very diverse range of biochemical roles, such as protein–protein interactions, DNA repair, mRNA de-capping and lipid-binding (57–61). The PH-like

fold is a composite fold (62) with an N-terminal four-strand module closely related to the monomeric four-stranded units of β-propeller proteins (63) and a C-terminal three-strand-helix unit, which we found to be specifically related to the AP2-IDBD fold (Figure 2). The PH-fold is widely utilized across the eukaryotes, but is currently not observed in the bacteria or archaea (58). As the PH-like fold is absent in the prokaryotes, unlike the β-propeller and the stand-alone AP2-IDBD folds, it appears to be a late innovation in evolution that occurred only after the primal eukaryote had emerged. Given that the PH-like fold is a fairly complex fold with no equivalents elsewhere, its innovation in the eukaryotes is likely to have occurred via the combination of two pre-existing modules, such as a monomeric unit of the β-propeller and a domain of the AP2-IDBD fold.

Other than the C-terminal module of the PH-like fold, most other members of the AP2-IDBD fold show rather sporadic phyletic distributions. Their multiple associations with mobile DNA elements suggests that they originally emerged as a DNA binding domain of an integrase/homing endonuclease and appear to have combined on multiple occasions with evolutionarily distinct classes of endonuclease modules in different mobile elements. From such a precursor they appear to have invaded the nuclear genome of eukaryotes, where the AP2-IDBD acquired new functions. At least two distinct invasions appear to have occurred—an early one, which probably gave rise to the C-terminal module of the PH-like fold and a late one, which gave rise to the plant TF family. On a number of occasions the HTH domains of transposases appear to have given rise to TFs, such as those with Paired, PSQ, and CENBP DNA binding domains (12,64–66). The BED finger domain, which is found in certain animal and plant TFs has been shown to be derived from the DNA recognition modules of activator-element type transposons (67). Similarly, the β-barrel DNA binding domain of the other major group of plant specific transcription factors, the VP1 TFs, appears to have been derived from the DNA binding domain of certain mobile restriction endonucleases (68). This suggests that the recruitment of DNA binding domains of transposases or integrases as TFs appears to be a recurrent theme in evolution.

This leads to a question as to whether the ApiAP2 family of AP2-IDBDs represents an independent acquisition from a mobile DNA element. The small size of the AP2 domain does not provide sufficient information to address this problem by means of conventional phylogenetic analysis. Although, as mentioned above, the domain architectures of the ApiAP2 family are reminiscent of the plant AP2 proteins, there are no specific sequence or predicted structure features that link these two groups to exclusion of other families. Moreover, the sequence conservation patterns make it clear that the expansions of the AP2 domains in the plant and apicomplexan clades are independent lineage-specific events. However, it is known that the apicomplexans are a chimeric lineage that has acquired a number of genes from a secondary endosymbiont of the primary plant lineage (including chlorophytes, rhodophytes and glaucocystophytes) (69), which gave rise to their apicoplast organelle (22,23,70). Hence, the most parsimonious explanation would be that the ApiAP2 family was derived from a plant AP2-like protein transferred from the rhodophyte, which was the apicoplast progenitor. This hypothesis can be tested with the availability of more sequence information from other sisters groups of the apicomplexans, such as the dinoflagellates, and early branching plant lineages, such as rhodophytes. Alternatively, given the distinctness of the ApiAP2 family it is possible that they were independently acquired from a bacterium or transposable element, similar to the TIE elements observed in ciliates (29).

CONCLUSIONS

Previous comparative genomics analyses had suggested an unexpected dearth of specific TFs in the apicomplexans, despite the presence of comparable number of genes and other signaling pathways as in unicellular free-living eukaryotes. Using sensitive sequence profile analysis methods, we show that the apicomplexans possess a large lineage-specific family of DNA binding proteins, the ApiAP2 family, with one or more copies of AP2 domain. By analogy to the plant TFs with the AP2 domain, we propose that these apicomplexan proteins are likely to function as the specific TFs in this lineage. This finding considerably reduces the ratio of genes to specific TFs in Apicomplexa compared with previous reports. While it is still lower than those seen in yeasts and other free-living eukaryotes with comparable genome sizes, it provides major candidates for understanding conventional specific transcription in Apicomplexa. It is possible that some other TFs specific to Apicomplexa remain undetected. Our searches of the proteome with sensitive profiles for DNA binding domains, and an examination of the multigene families fails to reveal any additional candidates. An analysis of the expression patterns of ApiAP2 genes during *P.falciparum* intraerythrocytic development suggests that different guilds of these TFs are specifically expressed in four major temporal phases corresponding to the ring, trophozoite, early schizont and late schizont–merozoite stages of development. This suggests that they might specifically regulate the expression of developmental stage specific target genes and maintain the progression of the development procession. We show that the domains of the AP2-IDBD have associated with different endonucleases domains on multiple occasions in evolution and appear to have contributed the primary TFs in both plants and apicomplexans through lineage-specific expansions. We also provide evidence that the PH-like fold appears to have emerged early in the eukaryotic lineage through the fusion of a β-propeller-like monomeric unit with a domain of the AP2-IDBD fold.

We hope that the findings presented here will spur future experimental investigations of the ApiAP2 family, which are likely to provide leads into hitherto unexpected aspects of apicomplexan transcriptional regulation.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by National Institutes of Health, National Library of Medicine Intramural Research Program.

Conflict of interest statement. None declared.

REFERENCES

1. Lodish,H., Berk,A., Zipursky,S.L., Matsudaira,P., Baltimore,D. and Darnell,J.E. (1999) *Molecular Cell Biology*. W.H. Freeman & Co., NY.
2. Cramer,P. (2002) Common structural features of nucleic acid polymerases. *Bioessays*, **24**, 724–729.
3. Iyer,L.M., Koonin,E.V. and Aravind,L. (2003) Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. *BMC Struct. Biol.*, **3**, 1.
4. Borukhov,S. and Nudler,E. (2003) RNA polymerase holoenzyme: structure, function and biological implications. *Curr. Opin. Microbiol.*, **6**, 93–100.
5. Langer,D., Hain,J., Thuriaux,P. and Zillig,W. (1995) Transcription in archaea: similarity to that in eucarya. *Proc. Natl Acad. Sci. USA*, **92**, 5768–5772.
6. Losick,R. (1998) Summary: three decades after sigma. *Cold Spring Harb. Symp. Quant. Biol.*, **63**, 653–666.
7. Gross,C.A., Chan,C., Dombroski,A., Gruber,T., Sharp,M., Tupy,J. and Young,B. (1998) The functional and regulatory roles of sigma factors in transcription. *Cold Spring Harb. Symp. Quant. Biol.*, **63**, 141–155.
8. Fassler,J.S. and Gussin,G.N. (1996) Promoters and basal transcription machinery in eubacteria and eukaryotes: concepts, definitions, and analogies. *Methods Enzymol.*, **273**, 3–29.
9. Bell,S.D. and Jackson,S.P. (1998) Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features. *Trends Microbiol.*, **6**, 222–228.
10. Bell,S.D., Jaxel,C., Nadal,M., Kosa,P.F. and Jackson,S.P. (1998) Temperature, template topology, and factor requirements of archaeal transcription. *Proc. Natl Acad. Sci. USA*, **95**, 15218–15222.
11. Aravind,L. and Koonin,E.V. (1999) DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res.*, **27**, 4658–4670.
12. Aravind,L., Anantharaman,V., Balaji,S., Babu,M.M. and Iyer,L.M. (2005) The many faces of the helix–turn–helix domain: transcription regulation and beyond. *FEMS Microbiol. Rev.*, **29**, 231–262.
13. Englbrecht,C.C., Schoof,H. and Bohm,S. (2004) Conservation, diversification and expansion of C2H2 zinc finger proteins in the *Arabidopsis thaliana* genome. *BMC Genomics*, **5**, 39.
14. Grishin,N.V. (2001) Treble clef finger—a functionally diverse zinc-binding structural motif. *Nucleic Acids Res.*, **29**, 1703–1714.
15. Riechmann,J.L., Heard,J., Martin,G., Reuber,L., Jiang,C., Keddie,J., Adam,L., Pineda,O., Ratcliffe,O.J., Samaha,R.R. *et al.* (2000) *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.
16. Schjerling,P. and Holmberg,S. (1996) Comparative amino acid sequence analysis of the C6 zinc cluster family of transcriptional regulators. *Nucleic Acids Res.*, **24**, 4599–4607.
17. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
18. Chervitz,S.A., Aravind,L., Sherlock,G., Ball,C.A., Koonin,E.V., Dwight,S.S., Harris,M.A., Dolinski,K., Mohr,S., Smith,T. *et al.* (1998) Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, **282**, 2022–2028.
19. Lespinet,O., Wolf,Y.I., Koonin,E.V. and Aravind,L. (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.*, **12**, 1048–1059.
20. 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.
21. Aravind,L. and Subramanian,G. (1999) Origin of multicellular eukaryotes—insights from proteome comparisons. *Curr. Opin. Genet Dev.*, **9**, 688–694.
22. Aravind,L., Iyer,L.M., Wellems,T.E. and Miller,L.H. (2003) Plasmodium biology: genomic gleanings. *Cell*, **115**, 771–785.
23. Templeton,T.J., Iyer,L.M., Anantharaman,V., Enomoto,S., Abrahante,J.E., Subramanian,G.M., Hoffman,S.L., Abrahamsen,M.S. and Aravind,L. (2004) Comparative analysis of apicomplexa and genomic diversity in eukaryotes. *Genome Res.*, **14**, 1686–1695.
24. Coulson,R.M., Hall,N. and Ouzounis,C.A. (2004) Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*. *Genome Res.*, **14**, 1548–1554.
25. Le Roch,K.G., Zhou,Y., Blair,P.L., Grainger,M., Moch,J.K., Haynes,J.D., De La Vega,P., Holder,A.A., Batalov,S., Carucci,D.J. and Winzeler,E.A. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, **301**, 1503–1508.
26. Bozdech,Z., Llinas,M., Pulliam,B.L., Wong,E.D., Zhu,J. and DeRisi,J.L. (2003) The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.*, **1**, E5.
27. Wessler,S.R. (2005) Homing into the origin of the AP2 DNA binding domain. *Trends Plant Sci.*, **10**, 54–56.
28. Magnani,E., Sjolander,K. and Hake,S. (2004) From endonucleases to transcription factors: evolution of the AP2 DNA binding domain in plants. *Plant Cell*, **16**, 2265–2277.
29. Wuitschick,J.D., Lindstrom,P.R., Meyer,A.E. and Karrer,K.M. (2004) Homing endonucleases encoded by germ line-limited genes in Tetrahymena thermophila have APETELA2 DNA binding domains. *Eukaryot. Cell*, **3**, 685–694.
30. Altschul,S.F., Madden,T.L., Schaffer,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.
31. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
32. Schaffer,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.
33. Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.*, **18**, 269–285.
34. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
35. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
36. Cuff,J.A., Clamp,M.E., Siddiqui,A.S., Finlay,M. and Barton,G.J. (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892–893.
37. Cuff,J.A. and Barton,G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.
38. Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
39. Delano,W.L. (2002) *The PyMOL Molecular Graphics System.* DeLano Scientific, San Carlos, CA, USA.
40. Holm,L. and Sander,C. (1995) Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.*, **20**, 478–480.
41. Holm,L. and Sander,C. (1996) The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res.*, **24**, 206–209.
42. Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta. Crystallogr. D. Biol. Crystallogr.*, **60**, 2256–2268.
43. Kumar,S., Tamura,K. and Nei,M. (2004) MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform.*, **5**, 150–163.
44. Troyanskaya,O., Cantor,M., Sherlock,G., Brown,P., Hastie,T., Tibshirani,R., Botstein,D. and Altman,R.B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.
45. de Hoon,M.J., Imoto,S., Nolan,J. and Miyano,S. (2004) Open source clustering software. *Bioinformatics*, **20**, 1453–1454.
46. Pavlidis,P. and Noble,W.S. (2003) Matrix2png: a utility for visualizing matrix data. *Bioinformatics*, **19**, 295–296.
47. Aravind,L. and Landsman,D. (1998) AT-hook motifs identified in a wide variety of DNA-binding proteins. *Nucleic Acids Res.*, **26**, 4413–4421.
48. Allen,M.D., Yamasaki,K., Ohme-Takagi,M., Tateno,M. and Suzuki,M. (1998) A novel mode of DNA recognition by a beta-sheet revealed by the solution structure of the GCC-box binding domain in complex with DNA. *EMBO J.*, **17**, 5484–5496.

49. Cernac,A. and Benning,C. (2004) WRINKLED1 encodes an AP2/EREB domain protein involved in the control of storage compound biosynthesis in *Arabidopsis*. *Plant J.*, **40**, 575–585.
50. Krizek,B.A. (2003) AINTEGUMENTA utilizes a mode of DNA recognition distinct from that used by proteins containing a single AP2 domain. *Nucleic Acids Res.*, **31**, 1859–1868.
51. Moure,C.M., Gimble,F.S. and Quiocho,F.A. (2002) Crystal structure of the intein homing endonuclease PI-SceI bound to its recognition sequence. *Nature Struct. Biol.*, **9**, 764–770.
52. Buttner,M. and Singh,K.B. (1997) *Arabidopsis thaliana* ethylene-responsive element binding protein (AtEBP), an ethylene-inducible, GCC box DNA-binding protein interacts with an ocs element binding protein. *Proc. Natl Acad. Sci. USA*, **94**, 5961–5966.
53. Ohme-Takagi,M. and Shinshi,H. (1995) Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element. *Plant Cell*, **7**, 173–182.
54. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.
55. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.
56. Leander,B.S., Clopton,R.E. and Keeling,P.J. (2003) Phylogeny of gregarines (Apicomplexa) as inferred from small-subunit rDNA and beta-tubulin. *Int. J. Syst. Evol. Microbiol.*, **53**, 345–354.
57. Musacchio,A., Gibson,T., Rice,P., Thompson,J. and Saraste,M. (1993) The PH domain: a common piece in the structural patchwork of signalling proteins. *Trends Biochem. Sci.*, **18**, 343–348.
58. Blomberg,N., Baraldi,E., Nilges,M. and Saraste,M. (1999) The PH superfold: a structural scaffold for multiple functions. *Trends. Biochem. Sci.*, **24**, 441–445.
59. Gervais,V., Lamour,V., Jawhari,A., Frindel,F., Wasielewski,E., Dubaele,S., Egly,J.-M., Thierry,J.-C., Kieffer,B. and Poterszman,A. (2004) TFIIH contains a PH domain involved in DNA nucleotide excision repair. *Nature Struct. Mol. Biol.*, **11**, 616–622.
60. She,M., Decker,C.J., Sundramurthy,K., Liu,Y., Chen,N., Parker,R. and Song,H. (2004) Crystal structure of Dcp1p and its functional implications in mRNA decapping. *Nature Struct. Mol. Biol.*, **11**, 249–256.
61. Vetter,I.R., Nowak,C., Nishimoto,T., Kuhlmann,J. and Wittinghofer,A. (1999) Structure of a Ran-binding domain complexed with Ran bound to a GTP analogue: implications for nuclear transport. *Nature*, **398**, 39–46.
62. Yoon,H.S., Hajduk,P.J., Petros,A.M., Olejniczak,E.T., Meadows,R.P. and Fesik,S.W. (1994) Solution structure of a pleckstrin-homology domain. *Nature*, **369**, 672–675.
63. Fulop,V. and Jones,D.T. (1999) Beta propellers: structural rigidity and functional diversity. *Curr. Opin. Struct. Biol.*, **9**, 715–721.
64. Smit,A.F. and Riggs,A.D. (1996) Tiggers and DNA transposon fossils in the human genome. *Proc. Natl Acad. Sci. USA*, **93**, 1443–1448.
65. Izsvak,Z., Khare,D., Behlke,J., Heinemann,U., Plasterk,R.H. and Ivics,Z. (2002) Involvement of a bifunctional, paired-like DNA-binding domain and a transpositional enhancer in Sleeping Beauty transposition. *J. Biol. Chem.*, **277**, 34581–34588.
66. Sitbon,E. and Pietrokovski,S. (2003) New types of conserved sequence domains in DNA-binding regions of homing endonucleases. *Trends Biochem. Sci.*, **28**, 473–477.
67. Aravind,L. (2000) The BED finger, a novel DNA-binding domain in chromatin-boundary-element-binding proteins and transposases. *Trends Biochem. Sci.*, **25**, 421–423.
68. Yamasaki,K., Kigawa,T., Inoue,M., Tateno,M., Yamasaki,T., Yabuki,T., Aoki,M., Seki,E., Matsuda,T., Tomo,Y. *et al.* (2004) Solution structure of the B3 DNA binding domain of the *Arabidopsis* cold-responsive transcription factor RAV1. *Plant Cell*, **16**, 3448–3459.
69. Foth,B.J. and McFadden,G.I. (2003) The apicoplast: a plastid in *Plasmodium falciparum* and other Apicomplexan parasites. *Int. Rev. Cytol.*, **224**, 57–110.
70. Martin,W., Stoebe,B., Goremykin,V., Hapsmann,S., Hasegawa,M. and Kowallik,K.V. (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature*, **393**, 162–165.