Nucleic Acids Research, 1993, Vol. 21, No. 12 2823–2830

# Evolution of sequence repetition and gene duplications in the TATA-binding protein TBP (TFIID)

John M.Hancock
Molecular Evolution and Systematics Group and Bioinformatics Facility, Research School of Biological Sciences, Australian National University, Canberra, ACT 0200, Australia

Received April 6, 1993; Revised and Accepted May 27, 1993

## ABSTRACT

**Analysis of TBP gene sequences from a variety of species for clustering of short sequence motifs and for over- and underrepresentation of short sequence motifs suggests involvement of slippage in the recent evolution of the TBP N-terminal domains in metazoans, *Acanthamoeba* and wheat. AGC, GCA and CAG are overrepresented in TBP genes of other species, suggesting that *opa* arrays were amplified from motifs overrepresented in ancestral species. The phylogenetic distribution of recently slippage-derived sequences in TBP is similar to that observed in the large subunit ribosomal RNAs, suggesting a propensity for certain evolutionary lineages to incorporate slippage-generated motifs into protein-coding as well as ribosomal RNA genes. Because length increase appears to have taken place independently in lineages leading to vertebrates, insects and nematodes, TBP N-terminal domains in these lineages are not homologous. All gene duplications in the TBP gene family appear to have been recent events despite strong protein sequence similarity between TRF and *P.falciparum* TBP. The enlargement of the TBP N-terminal domain may have coincided with acquisition of new functions and may have accompanied molecular coevolution with domains of other proteins, resulting in the acquisition of new or more complex mechanisms of transcription regulation.**

## INTRODUCTION

Simple DNA sequences, defined as DNA sequences that have a clustered distribution of sequence motifs, have been shown to be common in DNA sequence databases [1] and in *Caenorhabditis elegans* genomic clones (JMH, unpublished observations). Analysis of large subunit ribosomal RNA (LSU-rRNA) genes has shown that simple sequences are not distributed uniformly in phylogenetic space, but tend to be concentrated in certain lineages [2]. In addition, the way in which such sequences have been incorporated into LSU-rRNAs during evolution suggests that selection has acted to limit both their location and sequence [3]. Such observations raise three questions about the evolution of simple sequences. Can they be incorporated into regions encoding functional proteins, if so what might the consequences of such a mode of evolution be, and does the phylogenetic distribution of such sequences in functional proteins reflect that observed in the LSU-rRNAs?

To begin to answer such questions, DNA sequences encoding the TATA-binding protein (TBP) have been subjected to phylogenetic and sequence simplicity analysis. TBP is a component of transcription factor IID that binds the TATA box [4,5]. It has been shown to comprise a conserved C-terminal domain and a variable N-terminal domain which varies length from 18 amino acids in *Arabidopsis thaliana* and *Solanum tuberosum* [6,7] to 172 amino acids in *Drosophila melanogaster* [8,9], and contains the simple sequence *opa* element ($(GCA)_n$ [10]) in vertebrates and *D.melanogaster*. Because of its bipartite structure TBP can act as a model for the analysis of the phylogenetic distribution of simple sequences. Furthermore, as it is essential not only for basal transcription by RNA polymerase II (as well as RNA polymerases I and III [11]), but also for the activation of transcription that occurs on the binding of certain protein factors (activating proteins) to control regions upstream of promoters (reviewed in Ref. 5), TBP also represents a model for the incorporation of simple sequences into genes encoding important functional molecules. The recent discovery of a developmentally regulated protein in *D.melanogaster*, known as TBP-related factor (TRF [12]), which is related to TBP, as well as the presence of of two distinct forms of TBP in *A.thaliana* and *T.aestivum*, suggests that TBP is a member of a gene family whose members may have been subject to different evolutionary pressures.

Here, DNA sequence simplicity is found to be associated not only with *opa* arrays in mammalian and Drosophila TBP, but also with the N-termini of *C.elegans*, *A.castellanii* and *Triticum aestivum* TBPs, which do not contain *opa* arrays, and parts of the mammalian and *D.melanogaster* N-termini outside their respective *opa* regions. *Opa*-like motifs, which are found associated with sequence simplicity in the *A.castellanii* N-terminal domain and are strongly overrepresented overall in *D.discoideum* and *S.tuberosum*, are suggested to have been overrepresented before the genesis of the *opa* arrays themselves, and to have served as seeds for *opa* amplification. The amplification process

may have accompanied the acquisition of new functions by the TBP N-terminal end and coevolution with proteins interacting with TBP. The pattern of phylogenetic distribution of simple sequences in TBP is similar to that in the LSU-rRNAs, and is largely confined to animals, Acanthamoeba and wheat. Two yeasts and Plasmodium show extended N-terminal domains but low overall levels of sequence simplicity. This may reflect an ancient series of slippage events that have been subsequently obscured by point mutation, as appears to have taken place in the expansion segments of *D.melanogaster* LSU-rRNAs [13].

## METHODS

### Sequences

Sources of sequences are summarized in Table 1 along with some of their characteristics. Sequences are identified by their GenBank IDs where these exist.

### Multiple alignment and phylogenetic analysis

Multiple alignment of TBP sequences was carried out using the program PILEUP implemented under GCG version 7.0 [30]. Phylogenetic trees were constructed by cladistic analysis of protein and DNA sequences using v 3.0s of the PAUP program [31] run on a Macintosh IIvx computer. 50% majority rule consensus trees were constructed from bootstrap analysis (100 replicates) carried out using the heuristic search algorithm. Analysis of protein sequences used the protpars matrix provided with PAUP to specify the number of steps required to convert between pairs of amino acids. A tree spanning all sequences was constructed using the aligned protein sequences of all C-terminal (conserved) domains. Plant sequences were resolved further using the vertebrate sequences as an outgroup (Fig. 1b). Resolution of animal species was obtained using the *P.falciparum* sequence as an outgroup (Fig. 1c).

### Sequence simplicity analysis

Regions of sequence containing clustered short (tri- and tetranucleotide) motifs were identified using a modified version of the SIMPLE computer program [1]. This looks for repeats of the trinucleotide and tetranucleotide motifs starting at each position within a sequence for repeats of themselves occurring within the 32 base pairs 5′ and 3′ to them. 1 point is awarded for each repetition of the trinucleotide motif and an additional 3 points for each repeated tetranucleotide. An overall measure of short motif clustering within the sequence is given by the Relative Simplicity Factor (RSF). This is calculated by summing scores for each position along the sequence, averaging them over the entire sequence length, and dividing by a mean score derived from ten random sequences of length 10,000 base pairs and the same base composition. RSF values are judged to be significantly greater than 1.0 if they exceed 1.0 by more than three standard deviations of the variation of the random sequences. To give an impression of the distribution of clustering of tri- and tetranucleotide motifs along the sequence, scores are averaged over blocks of 10 base pairs, divided by the mean simplicity score of ten random sequences of the same length and base composition as the whole sequence, and plotted against sequence position.

Modifications to the original SIMPLE program have been introduced to allow compensation for sequence length by using random sequences of the same length as the test sequence, rather than a constant 10,000 base pairs, to generate scores against which the test sequence score (Simplicity Factor, SF) was compared. This resulted in independence of RSF from sequence length and improved estimation of confidence of significance of high ($>1$) RSF values. For any sequence, SIMPLE generates a frequency distribution of the occurrence of SF scores that resembles a saw tooth pattern. To test whether a particular score occurring in a test sequence had a sufficiently low probability of occurring for it to be considered significantly high, an

**Table 1.** Characteristics of published TBP DNA sequences

| Species | Reference | Code | Accession No. | Length | 5′ | 3′ |
|---|---|---|---|---|---|---|
| *Acanthamoeba castellanii* | 14 | ACATFIID | X63133 | 777 | 234 | 543 |
| *Arabidopsis thaliana* | 6 | ATTFIIDA | X54996 | 603 | 54 | 549 |
| | | ATTFIIB | X54995 | 603 | 54 | 549 |
| *Caenorhabditis elegans* | 15 | CELTBP | L07754 | 1023 | 483 | 540 |
| *Dictyostelium discoideum* | 16 | DDITFIID | M64861 | 618 | 60 | 558 |
| *Drosophila melanogaster* | 9 | DROTFIID | M38082 | 1062 | 516 | 546 |
| | 8 | DROTATABF | M38388 | 1062 | 516 | 546 |
| *D.melanogaster* TRF | 12 | dmtrf | | 675 | 138 | 537 |
| *Homo sapiens* | 17 | HUMTFIID | M55654 | 1008 | 462 | 546 |
| | 18 | HUMTFIIDA | M34960 | 1035 | 474 | 561 |
| *Mus musculus* | 19 | MUSTFIID | D01034 | 951 | 405 | 546 |
| *Plasmodium falciparum* | 20 | PFATBP | L06060 | 687 | 135 | 552 |
| *Saccharomyces cerevisiae* | 21 | YSCBTF1Y | M29459 | 723 | 180 | 543 |
| | 22 | YSCTFIIDA | M27135 | 723 | 180 | 543 |
| | 23 | YSCTFIIDB | X16860 | 723 | 180 | 543 |
| | 24 | YSCTFIID | M26403 | 723 | 180 | 543 |
| *Schizosaccharomyces pombe* | 25 | sptfiid | X53383 | 696 | 153 | 543 |
| | 26 | YSPTFIID | X53415 | 696 | 153 | 543 |
| *Solanum tuberosum* | 7 | STTATABP | X62494 | 603 | 54 | 549 |
| *Triticum aestivum* | 27 | TATFIID | X59874 | 702 | 153 | 549 |
| | 28 | TATFIIDA | L07604 | 606 | 57 | 549 |
| *Xenopus laevis* | 29 | XLTRAF | X66033 | 891 | 348 | 543 |

NOTE: Sequences are identified by unique codes. These correspond to their GenBank IDs where printed in capitals; codes printed in lower case correspond to sequences not in the databases at the time of writing. Sequences sptfiid and YSPTFIID are identical, as are YSCTFIID, YSCTFIIDB and YSCBTF1Y. 5′ and 3′ indicate the lengths of the 5′ and 3′ parts of the gene sequence, respectively.

additional modification was introduced. The mean frequency of each SF score for the ten random sequences ($f_e$) and the corresponding frequency for the test sequence ($f_o$) were counted. The significance value $S = 1-(f_e/f_o)$ was then calculated for each score. This represents the proportion of occurrences of that score that could be expected to be due to its non-random accumulation in the test sequence. Scores with a significance value of greater than 0.9 (i.e. a ratio of non-random to background scores of 9:1 or greater) were taken to be significant and recorded. The modifications and their consequences for sequence analysis by the SIMPLE program will be described in detail elsewhere (J.M.Hancock & J.S.Armstrong, submitted).

RSFs for complete TBP sequences and their 5′ and 3′ domains, identified from the multiple alignments used to derive phylogenetic trees, are presented in Table 2. Graphic displays of the distribution of sequence simplicity are shown in Fig.2. Motifs reaching significance values of ≥ 0.9 are listed in Table 3.

The SIMPLE algorithm tests for overrepresentation of short sequence motifs within 64 base pair segments of sequences. A distinct statistical property of sequences is over- and underrepresentation of motifs within sequences as a whole, irrespective of clustering. To test for this, a modification of the triplet chi-squared method [3] was used. The frequency distribution of $\chi^2=(O-E)^2/E$ for individual motifs, where O is the observed frequency of occurrence of a particular motif and E is its expected frequency based upon base composition, was estimated using 10,000 random sequences of length 10,000 and containing equal proportions of A, C, G and T. This gave confidence limits for the probability of finding a value of $\chi^2$ greater than a particular observed value by chance. Confidence limits at the 0.01, 0.001 and 0.0001 level were then applied to triplet chi-squared analyses of complete TBP coding sequences, and motifs reaching the three levels of confidence tabulated in Table 4.

## RESULTS

### Phylogenetic analysis

Cladistic analysis of C-terminal domain protein sequences generated the tree shown in Fig.1a. This shows relatively poor resolution of sequence lineages, even within well established groups such as the animals. The only highly significant groupings, as judged by bootstrap analysis, were the vertebrates (bootstrap value=100%), plants (93%), yeasts (90%), and TRF and *P.falciparum* TBP (90%). No improvement in resolution was obtained using complete DNA sequences, or first, second or third codon position subsets. Improved resolution of plant and animal groups was obtained using subsets of DNA sequence data and well separated outgroups. Plant sequences were resolved using complete 3′ domain DNA sequences (or third or first, but not second, codon positions) and the vertebrate sequences as an outgroup (Fig. 1b), while animal sequences were resolved using third codon positions and *P.falciparum* as an outgroup (Fig. 1c). DNA sequence analysis clustered TRF with DMTFIID (bootstrap=100%), in contrast to the results of protein sequence analysis. This association was also observed when 1st codon positions of all 3′ domains were analysed, with a bootstrap value of 51%.

### DNA sequence repetition

RSFs of 15 TBP coding sequences plus TRF (Table 3) show a general increase of RSF with sequence length. Six sequences show RSF values significantly greater than 1.0: those from

**Table 2.** Simplicity analysis of TBP cDNA sequences

| CODE | WHOLE RSF | SIG | 5′ RSF | SIG | 3′ RSF | SIG |
|---|---|---|---|---|---|---|
| ACATFIID | 1.358 | + | 1.682 | + | 1.306 | − |
| ATTFIIDA | 1.152 | − | − | | 1.127 | − |
| ATTFIIDB | 1.184 | − | − | | 1.116 | − |
| CELTBP | 1.488 | + | 1.638 | + | 1.226 | − |
| DDITFIID | 1.145 | − | − | | 1.094 | − |
| dmtrf | 1.181 | − | 1.890 | + | 1.034 | − |
| DROTFIID | 1.476 | + | 1.776 | + | 0.969 | − |
| HUMTFIIDA | 3.686 | + | 6.107 | + | 1.182 | − |
| HUMTFIID | 3.307 | + | 5.174 | + | 0.978 | − |
| MUSTFIID | 1.915 | + | 2.747 | + | 1.030 | − |
| PFATBP | 0.937 | − | 0.821 | − | 0.796 | − |
| sptfiid | 0.982 | − | 1.407 | + | 0.900 | − |
| STTATABP | 0.857 | − | − | | 0.805 | − |
| TATFIID | 1.363 | + | 1.382 | − | 1.085 | − |
| TATFIIDA | 1.092 | − | − | | 1.064 | − |
| XLTRAF | 1.094 | − | 1.280 | − | 0.819 | − |
| YSCTFIIDA | 1.025 | − | 0.892 | − | 1.026 | − |

NOTE: RSF is the relative simplicity factor; SIG indicates whether the value of RSF was significantly greater than 1.000, indicating the action of DNA slippage on the sequence [1]. WHOLE, 5′ and 3′ refer to the region of the gene analysed. 5′ regions marked '−' were too short (<65 base pairs) to be analysed by the SIMPLE program.

**Table 3.** Frequencies of motifs associated with sequence simplicity in the TBP 5′ domain

**a)** Frequencies of the association of individual motifs with sequence simplicity

| Motif | A.c. | C.e. | D.m. | H.s. | M.m. | T.a. | X.l. | Tot |
|---|---|---|---|---|---|---|---|---|
| AAC | | | | | 3 | | | 3 |
| AAT | | 3 | | | | | | 3 |
| ACA | | | | | 3 | | | 3 |
| ACC | 4 | | | | 2 | | | 6 |
| AGC | 2 | | 10 | 31 | 12 | | | 55 |
| ATC | | 3 | | | | | | 3 |
| ATG | | 1 | | | | | | 1 |
| CAA | | 5 | | 2 | 6 | | | 13 |
| CAC | | | | 2 | | | | 2 |
| CAG | | | 14 | 42 | 15 | | 2 | 73 |
| CCA | | | 1 | 1 | 1 | | | 3 |
| CCG | | | 1 | | | | | 1 |
| GAG | | | | | | 2 | | 1 |
| GCA | | | 10 | 37 | 13 | | | 60 |
| GGA | | | | | | 6 | | 6 |

NOTE: Motifs associated with scores reaching significance levels of >0.9 were counted for each species. A.c. = *A.castellanii*; C.e. = *C.elegans*; D.m. = *D.melanogaster*; H.s. = *H.sapiens*; T.a. = *T.aestivum*; X.l. = *X.laevis*. No simple sequence motifs were detected in other 5′ regions.

**b)** Frequencies of different circularly permutable motifs in each species

| Motif | A.c. | C.e. | D.m. | H.s. | M.m. | T.a. | X.l. | Tot |
|---|---|---|---|---|---|---|---|---|
| AAC/ACA/CAA | | 5 | | 2 | 12 | | | 19 |
| AAT/ata/taa | | 3 | | | | | | 3 |
| ACC/CCA/CAC | 4 | | 1 | 3 | 3 | | | 11 |
| AGC/GCA/CAG | 2 | | 34 | 110 | 40 | | 2 | 188 |
| ATC/tca/cat | | 3 | | | | | | 3 |
| ATG/tga/gat | | | 1 | | | | | 1 |
| CCG/cgc/gcc | | | 1 | | | | | 1 |
| GAG/agg/GGA | | | | | | 8 | | 8 |

NOTE: Simple sequence motifs from Table 3a are classified into groups that are circularly permutable. Rows represent groups of motifs, with motifs in capitals those that were identified as reaching significantly high scores in sequences. Columns are the species in which they occur, identified using the same notation as Table 3a.

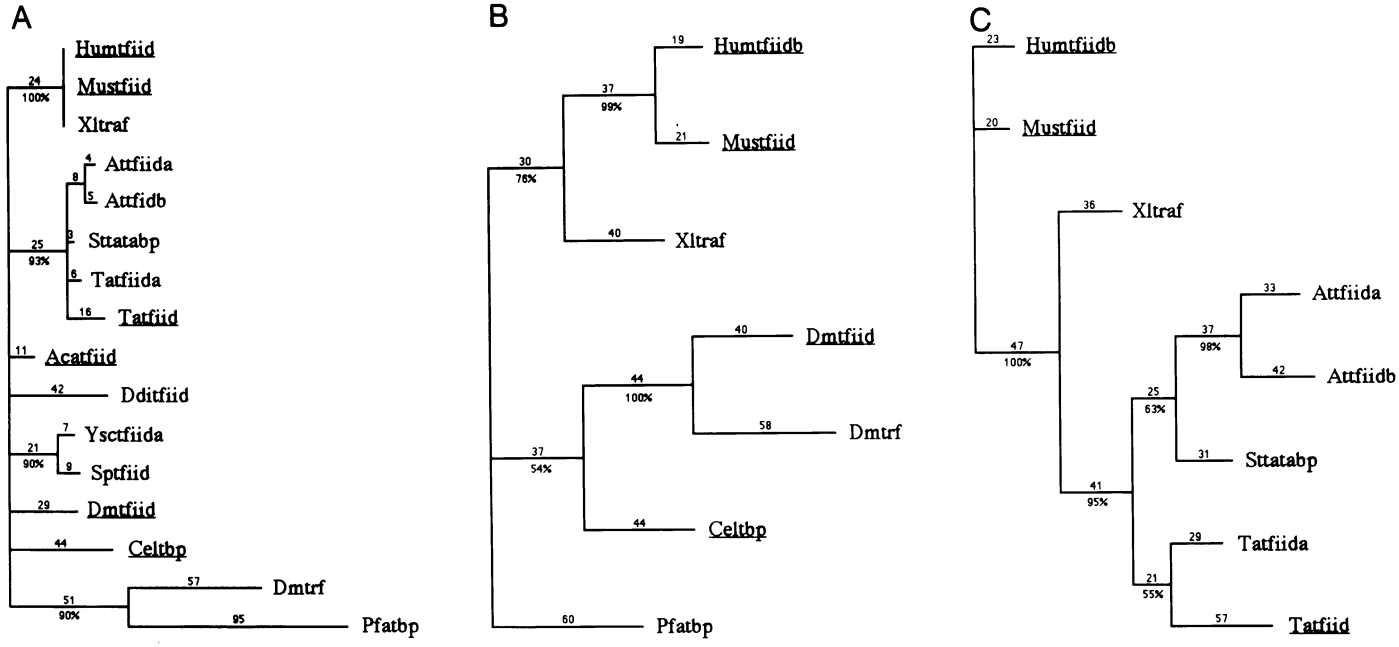


**Figure 1.** Phylograms representing relationships of TBP C-terminal domains compiled using PAUP [31]. **A**: Phylogram derived using protein sequences for all available C-terminal domain sequences; **B**: Phylogram for plant and vertebrate species derived using third codon positions of 3′ domain DNA sequences; **C**: Phylogram for animal species and *P.falciparum* TBP derived using third codon positions of 3′ domain DNA sequences. Lengths of internodes represent numbers of changes imputed to have occurred on that internode. Numbers above internodes are numbers of amino acid or base change steps and numbers below lines where present are bootstrap values (%) supporting the clade to their right. Sequences are represented by codes as in Table 1. Underlined codes represent sequences with 5′ domains having RSFs significantly greater than 1.000.

*H.sapiens*, *M.musculus*, *D.melanogaster*, *C.elegans*, *A.castellanii* and *T.aestivum* TATFIID. When sequences were divided into 5′ (variable) and 3′ (conserved) domains, only 5′ domains showed significantly high overall levels of sequence simplicity. Sequences with significantly simple 5′ domains were generally also significantly simple overall with three exceptions. The 5′ domain of TATFIID reached a similar score to the whole sequence but did not achieve significance because of its length. The 5′ domain regions of the *S.pombe* and TRF sequences reached significantly high RSF scores although their complete sequences did not do so, and no individual motif in either sequence reached a significantly high SF score.

In *H.sapiens*, *M.musculus* and *D.melanogaster*, graphical displays of the distribution of sequence simplicity within TBP sequences (Fig.2) show a high concentration of sequence simplicity in regions encoded by the *opa* repeats identified previously. Analysis of motifs associated with high sequence simplicity showed that in addition to AGC, CAG and GCA (the *opa* motifs) a further 12 motifs were associated with sequence simplicity in the 5′ domains of TBP coding sequences. The frequencies of these achieving significance were overall similar to one another, although 13 copies of CAA were detected. Some motifs were however characteristic of particular species or groups of species (see Table 3a). Grouping motifs into circularly permutable groups (Table 3b), showed that in addition to AGC/GCA/CAG three other groups of motifs made a major contribution to sequence simplicity in TBP 5′ regions: AAC/ACA/CAA, ACC/CCA/CAC and GAG/AGG/GGA. The first two of these groups of motifs were characteristic of different subsets of animal species while the third was characteristic of *T.aestivum*.

Analysis of motifs present in disproportionate quantities in TBP coding sequences by triplet chi-squared analysis (Table 4) showed that many motifs showing significant clustering in these sequences do not show overrepresentation, while many overrepresented motifs were not significantly clustered as measured by the SIMPLE algorithm. Only four motifs are represented in the same species as both overrepresented and clustered: the *opa* motifs and CAA, which was clustered and overrepresented in *C.elegans*. Six motifs that were significantly clustered in some species were also overrepresented but not clustered in others. These again included the three *opa* motifs, which were overrepresented but not clustered in *A.thaliana* (AGC), *C.elegans* and *S.tuberosum* (CAG), *D.discoideum*, and TRF (GCA). The other three motifs in this class were ATG (overrepresented in *D.melanogaster* and clustered in *C.elegans*), CAA (overrepresented and clustered in *C.elegans*, clustered in *H.sapiens* and *M.musculus* and overrepresented in *D.discoideum*) and CCA (clustered and overrepresented in *M.musculus*, clustered in *D.melanogaster* and *H.sapiens*, and overrepresented in *S.pombe*). Six motifs were found to be highly overrepresented without showing evidence of clustering in any species: AAA (*S.cerevisiae*, *S.pombe*), CTG (*X.laevis*), GCT (*A.thaliana*), TCA (*C.elegans*), TCG (*A.castellanii*) and TGG (*C.elegans*, *D.discoideum*).

As well as detecting overrepresented motifs, the triplet chi-squared method also detects underrepresented motifs. Only two motifs were underrepresented at the $p < 0.0001$ level, CTA and TAG in *C.elegans*. A number of other motifs however showed

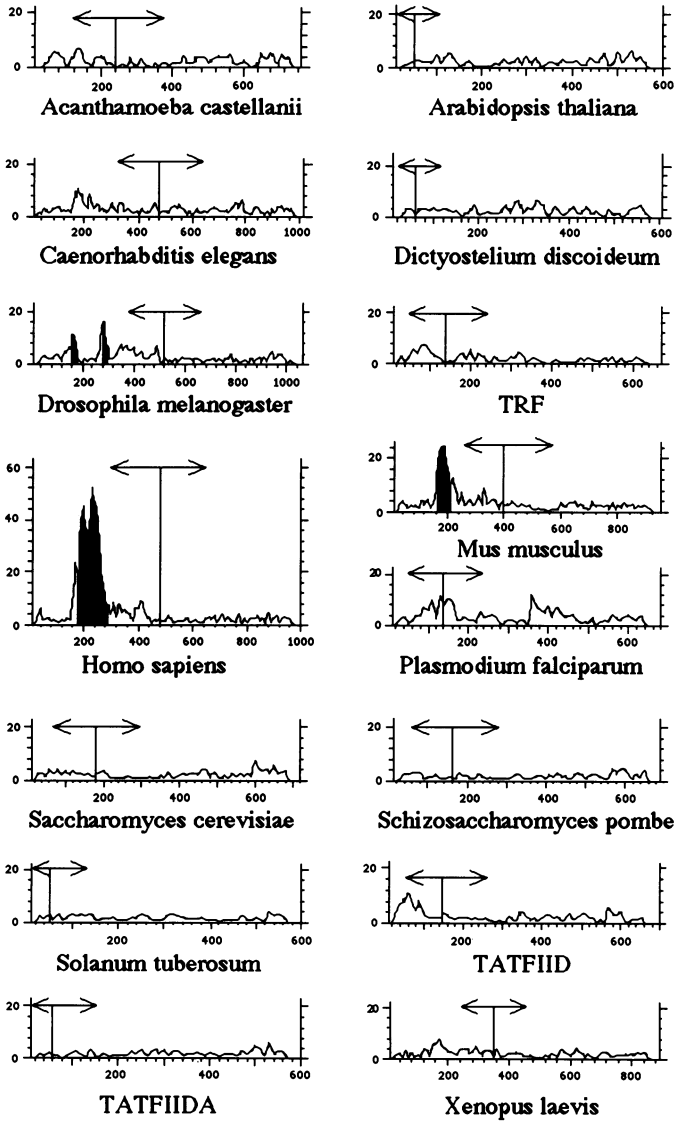


**Figure 2.** DNA sequence simplicity profiles of TBP cDNA sequences generated by the SIMPLE program [1] with modifications as described in Methods. In each case the horizontal axis represents the length of the sequence, numbered at 200 bp intervals, and the vertical axis the simplicity score. The boundary between the 5′ and 3′ domains is marked by a vertical line, the 5′ domain being indicated by a left-pointing arrow (←) and the 3′ domain by a right-pointing arrow (→). Species from which sequences derived, or in the cases of the *T.aestivum* sequences the sequence codes, are below their respective profiles. Regions of simplicity corresponding to tandem *opa* repeats are shaded in the *D.melanogaster*, *H.sapiens* and *M.musculus* profiles.

slight underrepresentation in a number of species, in particular TAG (5 species), ATA (4 species), CCG and CGT (3 species).

## DISCUSSION

### Origins of DNA repetition in the TBP N-terminal domain

The overall length variation of the N-terminus of TBP, indicative of insertion/deletion events during its evolution, is suggestive of a role for slippage-like processes in its evolution. There is also circumstantial evidence to suggest that DNA slippage has contributed to the evolution of the N-terminal *opa* regions, particularly in vertebrates. Firstly there is a change in the length of these regions between the Xenopus, mouse and human genes, which otherwise show high (90–95%) protein sequence similarity in their N-terminal domains [19,29]. Secondly, a copy number difference of *opa* subrepeats is seen in this region between the two sequenced human genes [17,18]. A number of lines of evidence link high levels of sequence simplicity (significantly high RSF scores) with the occurrence of DNA slippage within sequences. Simple sequence arrays show *in vivo* length polymorphism in large subunit rRNA genes [32], in introns of Drosophila protein coding genes (T.Kidd, D.Ish-Horowicz & JMH, unpublished observations), and make up the large number of highly polymorphic microsatellite markers in wide use in gene mapping. In addition, *in vitro* studies show that such sequences undergo slippage spontaneously in the presence of DNA polymerase [33]. Taking these observations together with the significant level of sequence simplicity found within TBP 5′ regions, it therefore seems likely that DNA slippage has played a major role in the evolution of this domain. In contrast, the C-terminal domain shows uniformly low RSF scores although a small number of insertion/deletion events have taken place there during its evolution, resulting in some length variation (see Table 1).

High RSF scores in TBP 5′ domains do not in all cases reflect the presence of tandem repetition within them, instead reflecting a pattern of distribution of repeated motifs wherein clustered motifs are interspersed with each other and with non-clustered motifs. Such a pattern of clustered interspersion is referred to as cryptic simplicity [1]. Cryptic simplicity may reflect the remnants of earlier slippage events and may also provide the raw material for future slippage. The observation of cryptic simplicity in *A.castellanii* TBP associated with *opa*-like motifs, but in the absence of their tandem repetition, prompted further analysis of overrepresented motifs in other species. This identified the most commonly overrepresented motifs to be AGC, CAG and GCA, as in SIMPLE analysis, but the distribution of the over-representation of these motifs among species was wider than their association with sequence simplicity. As well as Drosophila, human and mouse, *S.tuberosum* (CAG), and *D.discoideum* (GCA) showed very high overrepresentation of this class of motif, while *A.thaliana* (AGC) and *C.elegans* (CAG), as well as TRF (GCA), showed less marked overrepresentation. Over-representation of *opa*-like motifs is therefore widespread in TBP encoding sequences even where *opa* arrays are not present. These observations suggest that *opa*-like motifs were overrepresented in TBP genes before the amplification of *opa* arrays was initiated in the animal lineage. As *opa* repeats in TBP genes always encode polyglutamine regions, this may reflect selection for glutamine-rich domains in TBP, resulting in a propensity for the fixation of such slippage products and their subsequent exploitation in transcription factors and other *opa*-containing proteins.

Although it contains a short *opa* array, the sequence encoding *X.laevis* TBP did not achieve a significantly high score in SIMPLE analysis. The 5′ region of the sequence did however achieve a score above one, although not significantly so for its length, and contained CAG motifs that reached significantly high simplicity scores. Alignments of the *X.laevis* sequence with the mouse and human sequences show high levels of sequence conservation (90% at the protein level, excluding gaps), the main difference being in the length of the *opa*-encoded region. This is consistent with the hypothesis that the ancestral sequence for

**Table 4.** Over- and underrepresented trinucleotide motifs in TBP coding regions

| Ac | Ata | Atb | Ce | Dd | Dm | TRF | Hsa | Mm | Pf | Sc | Sp | St | Ta | Taa | Xl | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AAA | | | | | | | | | | | +++ | +++ | | | | |
| AAG | | + | | | | | | | | | | | | | | |
| ACG | | | | | | | | | −− | | | | | | | |
| AGA | + | | | | | | | | | | | | | | | |
| AGC | | + | | | | ++ | | +++ | +++ | | | | | | | |
| ATA | | | | −− | | | | − | − | | | − | | | | |
| ATG | | | | | | + | | | | | | | | | | |
| ATT | | | | + | | | | | | | | | | | | |
| CAA | | | | +++ | ++ | | | | | | | | | | | |
| CAG | | | | + | | +++ | | +++ | +++ | | | | +++ | | | +++ |
| CAT | + | | | | | | | | | | | | | | | |
| CCA | | | | | | | | | + | | | + | | | | |
| CCG | | | | | | | | − | − | | | | | | | − |
| CGA | | | | | | | | −− | −− | | | | | | | |
| CGC | | | | | | | | − | − | | | | | | | |
| CGG | | | | | | | | | − | | | | | | | − |
| CGT | | | | | | − | | − | − | | | | | | | |
| CTA | | | | −−− | | | | | | | | | | | | |
| CTG | | | | | | | | | | | | | | | | +++ |
| CTT | | + | | | | | | | | | | | | | | |
| GCA | | | | | +++ | +++ | ++ | +++ | ++ | | | | | | | |
| GCG | | | | | | | | −(−) | | | | | | | | |
| GCT | | +++ | + | | | | | | | | | | | | | |
| GTA | | | | | | | | | | | | | − | | | |
| GTC | | | | | | | | | − | | | | | | | |
| TAA | | | | −− | | | | | | | | | − | | | |
| TAC | | | | − | | | | − | | | | | | | | |
| TAG | − | | | −−− | | − | | | | | | − | | − | − | |
| TCA | | | | +++ | | | | | | | | | | | | |
| TCC | | | | | | + | | | | | | | | | | |
| TCG | +++ | | | | | | | − | − | | | | | | | − |
| TGC | | | | | + | | | | | + | + | | | | | |
| TGG | | | | +++ | ++ | | | | | | | | | | | |
| TTC | | | | + | | | | | | | | | | | | |
| TTG | | | | + | | | | | | | | | | | | |
| TTT | | | | | | | | (+) | | | | | | | | + |

Over- and underrepresentation of trinucleotides as identified by triplets chi-squared analysis is represented by different numbers of '+' or '−' symbols. '+++' and '−−−' represent over- or underrepresentation at the $p<0.0001$ level, '++' and '−−' at $p<0.001$ and '+' and '−' at $p<0.01$. Over- and underrepresentations are arranged by species (columns): Ac = *A.castellanii*; Ata = ATTFIIDA; Atb = ATTFIIDB; Ce = *C.elegans*; Dd = *D.discoideum*; Dm = *D.melanogaster* TBP; TRF = *D.melanogaster* TRF; Hs = *H.sapiens*; Mm = *M.musculus*; Pf = *P.falciparum*; Sc = *S.cerevisiae*; Sp = *S.pombe*; Ta = TATFIID; Taa = TATFIIDA; Xl = *X.laevis*. Brackets in the Hs column indicate levels of significance achieved in HUMTFIID but not HUMTFIIDA.

the vertebrates had no extended *opa* array and that expansion of the TBP *opa* array has taken place much more rapidly in the vertebrate lineage leading to mammals than in that leading to Xenopus. Given the lack of sequence similarity and an apparent lack of homology of the *C.elegans* and *D.melanogaster* N-terminal domains to those of vertebrates, compared to the high level of conservation shown by the vertebrate sequences over more than 500 million years of divergence, the expansions seen in the lineages leading to Drosophila and Caenorhabditis seem likely to have taken place independently since the divergence of these lineages from each other and from vertebrates. The independent preferential fixation of polyglutamine encoding domains in insect (*D.melanogaster*) and vertebrate TBP may reflect common features of their transcription machinery not shared with nematodes (*C.elegans*), or it may reflect convergent evolution between the lineages.

### Phylogenetic distribution of simple sequences in TBP genes

Comparison of the pattern of distribution of sequence simplicity between species with the phylogenetic tree derived from the C-terminal domain of TBP (Fig.1) shows a strong clustering of sequence simplicity within the phylogenetic tree, the only lineages other than animals showing elevated sequence simplicity being Acanthamoeba and a monocotyledenous plant (wheat). This distribution of sequence simplicity and N-terminal domain length on the phylogenetic tree is suggestive of a number of independent slippage events leading to length increase in TBP in different lineages. The significant levels of TBP sequence simplicity and long TBP N-terminal ends shown by most animal sequences and their sharing of many of the same repetitive motifs is consistent with the action of slippage within animal TBP genes at least since the divergence of protists from metazoans, which pre-dates the Ediacaran fauna of 640 Ma ago [34]. Amongst the protists included in this analysis only Acanthamoeba shows high levels of sequence simplicity, whereas Dictyostelium and Plasmodium show no detectable sequence simplicity. According to 18S rRNA data the Acanthamoeba lineage originated relatively recently compared to Dictyostelium and Plasmodium, which are representatives of more ancient evolutionary lineages [35,36].

By contrast to the animal lineage, the plant and yeast lineages are relatively poor in sequence simplicity, with the exception of the wheat gene TATFIID. As the *A.thaliana*, *S.tuberosum* and wheat TATFIIDA TBP genes show low sequence simplicity and a N-terminal domain shorter than that of *D.discoideum*, the longer

N-terminal domain in TATFIID probably arose after the separation of the monocotyledenous and dicotyledenous plants, suggesting more recent onset of fixation of slippage products in the lineage leading to the monocotyledonous plants than was the case for the animals (monocotyledenous and dicotyledenous plants diverged approximately 230–350 million years ago [37]).

The yeasts and Plasmodium are exceptions to the relationship between high sequence simplicity and long TBP N-terminal ends as they have long TBP N-terminal ends but low levels of sequence simplicity. There are two possible explanations for this: their N-terminal domains may have been derived by slippage in the distant evolutionary past and subsequently ceased to undergo slippage, resulting in the subsequent erosion of patterns of sequence simplicity by point mutation (see Ref. 13), or their N-terminal ends may have originated in a different manner to those of other species. There is some circumstantial evidence in favour of the former explanation. Both yeasts show highly significant overrepresentation of AAA, which might be a remnant of early slippage events, and the 5′ region of the *S.pombe* gene reaches a significant RSF when analysed separately.

Many of the features of the pattern of evolution of the TBP N-terminal domain are similar to patterns observed in the LSU-rRNAs, which have also been suggested to expand by the incorporation of slippage-generated sequences [2,3] and which also show relationships between RSF and sequence length, phylogenetic distribution and intraspecific length polymorphism [2,3,32]. These common features of TBP and LSU-rRNA evolution may reflect a general susceptibility of the animal and monocotyledenous plant lineages to the fixation of slippage-generated sequences in the recent evolutionary past.

## Phylogenetic relationships and gene duplications in the TBP gene family

Phylogenetic analysis of both protein and DNA sequences of the TBP C-terminal (3′) domain shows association of the pairs of TBP genes detected in Arabidopsis and Triticum, suggesting that these duplications are relatively recent events. However, the phylogenetic tree based on protein sequences (Fig.1a) locates the *D.melanogaster* gene TRF closest to *P.falciparum* TBP with a high degree of confidence (bootstrap=90%), while trees derived from third codon positions of DNA sequences cluster TRF with DMTFIID (Fig.1c). The most conservative interpretation of these apparently contradictory results is that the duplication giving rise to TRF and TBP took place after insects separated from nematodes, and that the similarity in protein sequence observed between TRF and *P.falciparum* TBP reflects some common aspect of their function.

While no functional differentiation has been reported between the pairs of TBP genes in Arabidopsis and wheat, *D.melanogaster* TRF shows a developmentally regulated pattern of expression [12]. It therefore appears that Drosophila has made use of a pre-existing DNA-binding motif as the starting material for the evolution of a novel developmentally regulated gene. Such a process is similar to the process of co-option described for the insect pair-rule developmental gene *even-skipped* [38], and raises the possibility that a family of developmentally regulated genes involved in the control of transcription may have arisen from TBP over time. Although in the case of *D.melanogaster* TRF and TBP, TBP has a more enlarged N-terminal domain than TRF, TRF nevertheless shows overrepresentation of GCA which may be a remnant of ancient slippage events. The combination of gene co-option with the evolution of new functions by the incorporation of slippage-derived sequences would provide a powerful mechanism for the evolution of new developmental strategies. As TBP contains read-made DNA-binding and regulatory domains, it may have served as raw material for a larger gene family than has so far been detected.

## Molecular coevolution in the transcription apparatus?

The N-terminal domain of TBP has been shown to be involved in transcription activation in *D.melanogaster* [39], and in *S.cerevisiae* [40] and has also been proposed to play such a role in *A.castellanii* [14]. The N-terminal domain of *D.melanogaster* TBP cannot however be functionally replaced by that of *S.cerevisiae* [39]. This suggests that taxon-specific functions, in particular interactions with putative coactivators [39,41,42], have been acquired during the evolution of this domain. The coincidence of this with the enlargement of the N-terminal domain of TBP by the incorporation of slippage-generated sequences suggests that at least some of these taxon-specific functions may have been acquired as a result of its enlargement. Such a process is likely to have been a gradual one, reflecting the successive addition of short stretches of sequence corresponding to one or two amino acids by slippage, their screening by selection, and gradual functional incorporation into the transcriptional machinery. It would also have necessitated alterations in proteins interacting with the N-terminal domain, by point mutation and/or slippage. This would represent an example of molecular coevolution [43] between proteins involved in transcription regulation.

This situation would be analogous to that observed in the RNA polymerase I transcription system, where species-incompatibility is observed between different species of Drosophila (reviewed in Ref. 43). In this system, taxon-specificity has also arisen as a result of molecular coevolution, both between RNA polymerase I transcription factors, and between the RNA polymerase I complex and its promoter, which evolves rapidly [43–45]. The genes encoding the coevolving protein components of the RNA polymerase I transcription machinery also contain simple sequences (JMH, unpublished observations).

Molecular coevolution has been possible in the RNA polymerase I system because the rDNA is a multigene family which can accommodate a proportion of promoter variants without serious selective disadvantage [43]. In the case of interactions of TBP with other proteins, as in the case of mammalian RNA polymerase I cofactors [45], it is necessary to look elsewhere for the redundancy that must exist in the interactions to allow variant proteins to arise without being eliminated by selection. In such cases, molecular coevolution may reflect a degree of non-specificity in the interactions between the proteins concerned. For example, such interactions may involve generalized interactions between charged surfaces of proteins, or a number of individual interactions between pairs of amino acids, any one of which may be disrupted without loss of the interaction itself. This latter kind of redundancy is analogous to that existing in individual stems of ribosomal RNAs undergoing compensatory mutation [46]. An example of such redundancy may be the *opa*-encoded regions of vertebrate and *D.melanogaster* TBPs, which encode polyglutamine regions and appear to have been particularly susceptible to slippage, as glutamine-rich domains of transcriptional regulation proteins have been suggested to be involved in protein-protein interactions between transcription factors [47]. Although *opa*-like motifs were predominant in the TBP-coding sequences of the metazoan species

with high levels of sequence simplicity (*H.sapiens*, *M.musculus* and *D.melanogaster*), a number of other motifs were also found associated with sequence simplicity in these and other species (see Table 3), reflecting the action of slippage on a variety of sequence substrates. As these motifs encode part of a functional protein, the fixation of different products of slippage in different species and in different parts of the N-terminal domain is likely to reflect different evolutionary pressures acting on TBP during its evolution in different species.

The possibility that species-specific functions of the TBP N-terminal domain have arisen by a process of adoption of new functions as it has undergone DNA-slippage has potentially far-reaching implications for the evolution of other molecular systems. The best documented evolutionary systems in which slippage has been shown to act involve sequences that are either functionless or whose functions are indeterminate (such as the expansion segments of the large-subunit rRNA [2,3], minisatellite sequences [48] and mitochondrial D-loops [49]). There is however increasing evidence of involvement of slippage in a variety of protein systems as disparate as involucrins [50], chorion proteins [51], Balbiani ring proteins [52], the Drosophila clock gene *period* [53], and developmental genes such as the gap segmentation gene *hunchback* [54], and the neurogenic gene *mastermind* [55]. The data presented here raise the possibility that slippage may be involved in the coevolutionary change of sets of interacting proteins in some of these systems.

## ACKNOWLEDGEMENTS

I thank Gabriel Dover and Michael Ashburner for their support during the early stages of this work.

## REFERENCES

1. Tautz, D., Trick, M., & Dover, G.A. (1986) *Nature* **322,** 652–656.
2. Hancock, J.M., & Dover, G.A. (1988) *Mol Biol Evol* **5,** 377–391.
3. Hancock, J.M., & Dover, G.A. (1990) *Nuc Acids Res* **18,** 5949–5954.
4. Dynlacht, B.D., Hoey,T., & Tjian, R. (1991) *Cell* **66,** 563–576.
5. Greenblatt, J. (1991) *Cell* 66, 1067–1070.
6. Gasch, A., Hoffmann, A., Horikoshi, M., Roeder, R.G., & Chua, N-H. (1990) *Nature* **346,** 390–394.
7. Holdsworth, M.J., Grierson, C., Schuch, W. & Bevan, M. (1992) *Plant Mol Biol* **19,** 455–464.
8. Muhich, M., Iida, C.T., Horikoshi, M., Roeder, R.G., & Parker, C.S. (1990) *Proc Natl Acad Sci USA* **87,** 9148–9152.
9. Hoey, T., Dynlacht, B.D., Peterson, M.G., Pugh, B.F., & Tjian, R. (1990) *Cell* **61,** 1179–1186.
10. Wharton, K.A., Yedvobnick, B., Finnerty, V.G., & Artavanis-Tsakonas, S. (1985) *Cell* **40,** 55–62.
11. Cormack, B.P. & Struhl, K. (1992) *Cell* **69,** 685–696.
12. Crowley, T.E., Hoey, T., Liu, J.-K., Jan, Y.N., Jan L.Y. & Tjian, R. (1993) *Nature* **361,** 557–561.
13. Ruiz Linares, A., Hancock, J.M., & Dover, G.A. (1991) *J Mol Biol* **219,** 381–390.
14. Wong, J.M., Liu, F. & Bateman, E. (1992) *Gene* **117,** 91–97.
15. ichtsteiner, S. (unpublished). L07754.
16. Blume, J.E., Shaw, D.R., & Ennis, H.L. (unpublished). X53415.
17. Hoffmann, A., Sinn, E., Yamamoto, T., Wang, J., Roy, A., Horikoshi, M. & Roeder, R.G. (1990) *Nature* **346,** 387–390.
18. Kao, C.C., Lieberman, P.M., Schmidt, M.C., Zhou, Q., Pei, R., & Berk, A.J. (1990) *Science* **248,** 1646–1650.
19. Tamura, T., Sumita, K., Fujino, I., Aoyama, A., Horikoshi, M., Hoffmann, A., Roeder, R.G., Muramatsu, M., & Mikoshiba, K. (1991) *Nuc Acids Res* **19,** 3861–3865.
20. McAndrew, M.B., Read, M., Sims, P.F. & Hyde, J.E. (1993) *Gene* **124,** 165–171.
21. Cavallini, B., Faus, I., Matthes, H., Chipoulet, J.M., Winsor, B., Egly, J.M. & Chambon, P. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 9803–9807.
22. Hahn, S., Buratowski, S., Sharp, P.A., & Guarente, L. (1989) *Cell* **58,** 1173–1181.
23. Horikoshi, M., Wang, C.K., Fujii, H., Cromlish, J.A., Weil, P.A., & Roeder, R.G. (1989) *Nature* **341,** 299–303.
24. Schmidt, M.C., Kao, C.C., Pei, R., & Berk, A.J. (1989) *Proc Natl Acad Sci USA* **86,** 7785–7789.
25. Fikes, J.D., Becker, D.M., Winston, F., & Guarente, L. (1990) *Nature* **346,** 291–294.
26. Hoffmann, A., Horikoshi, M., Wang, C.K., Schroeder, S., Weil, P.A., & Roeder, R.G. (1990) *Genes Dev* **4,** 1141–1148.
27. Kawata, T., Minami, M., Tamura, T., Sumita, K., & Iwabuchi, M. (1992) *Plant Mol Biol* **19,** 867–872.
28. Apsit, V., Freberg, J.A., Chase, M.R., Davis, E.A. & Ackerman, S. (unpublished). L07604.
29. Hashimoto, S., Fujita, H., Hasegawa, S., Roeder, R.G. & Horikoshi, M. (1992) *Nuc Acids Res* **20,** 3788.
30. Devereux, J.P., Haeberli, P., & Smithies, O. (1984) *Nuc Acids Res* **12,** 387–395.
31. Swofford, D.L. (1991) PAUP: Phylogenetic Analysis Using Parsimony, Version 3.0s. Computer program distributed by the Illinois Natural History Survey, Champaign, Illinois..
32. Gonzalez, I.L., Gorski, J.L., Campen, T.J., Dorney, D.J., Erickson, J.M., Sylvester, J.E., & Schmickel, R.D. (1985) *Proc Natl Acad Sci USA* **82,** 7666–7670.
33. Schlötterer, C. & Tautz, D. (1992) *Nuc Acids Res* **20,** 211–215.
34. Cloud, P & Glaessner, M.F. (1982) *Science* **217,** 783–792..
35. Sogin, M.L., Elwood, H.J. & Gunderson, J.H. (1986) *Proc Natl Acad Sci USA* **83,** 1383–1387.
36. Gunderson, J.H., McCutchan, T.F. & Sogin, M.L. (1986) *J Protozool* **33,** 525–529.
37. Brandl, R., Mann, W. & Sprinzl, M. (1992) *Proc R Soc Lond B* **249,** 13–17.
38. Patel, N.H., Ball, E.E. & Goodman, C.S. (1992) *Nature* **357,** 339–342.
39. Pugh, B.F., & Tjian, R. (1990) *Cell* **61,** 1187–1197.
40. Zhou, Q., Schmidt, M.C., & Berk, A.J. (1991) *EMBO J* **10,** 1843–1852.
41. Berger, S.L., Cress, W.D., Cress, A., Triezenberg, S.J., & Guarente, L. (1990) *Cell* **61,** 1199–1208.
42. Kelleher, R.J., III, Flanagan, P.M., & Kornberg, R.D. (1990) *Cell* **61,** 1209–1215.
43. Dover, G.A., & Flavell, R.B. (1983) *Cell* **38,** 622–623.
44. Tautz, D., Tautz, C., Webb, D. & Dover, G.A. (1987) *J Mol Biol* **195,** 525–542.
45. Bell, S.P., Pikaard, C.S., Reeder, R.H., & Tjian, R. (1989) *Cell* **59,** 489–497.
46. Hancock, J.M., Tautz, D., & Dover, G.A. (1988) *Mol Biol Evol* **5,** 393–414.
47. Mitchell, P.J., & Tjian, R. (1989) *Science* **245,** 371–378.
48. Kelly, R., Gibbs, M., Collick, A., & Jeffreys, A.J. (1991) *Proc R Soc Lond B* **245,** 235–245.
49. Hoelzel, A.R., Hancock, J.M., & Dover, G.A. (1991) *Mol Biol Evol* **8,** 475–493.
50. Djian, P., & Green, H. (1989) *Proc Natl Acad Sci USA* **86,** 8447–8451.
51. Burke, W.D., & Eickbush, T.H. (1986) *J Mol Biol* **190,** 357–366.
52. Paulsson, G., Lendahl, U., Galli, J., Ericsson, C., & Wieslander, L. (1990) *J Mol Biol* **211,** 331–349.
53. Costa, A.R., Peixoto, A.A., Thackeray, J.R., Dalgleish, R., & Kyriacou, C.P. (1991) *J Mol Evol* **32,** 238–246.
54. Treier, M., Pfeifle, C., & Tautz, D. (1989) *EMBO J.* **8,** 1517–1525.
55. Newfeld, S.J., Smoller, D.A., & Yedvobnick, B. (1991) *J Mol Evol* **32,** 415–420.