Analysis of consensus sequence patterns in *Giardia* cytoskeleton gene promoters

David V. Holberton* and Jonathan Marshall

Department of Life Science, Nottingham University, Nottingham NG7 2RD, UK

Received April 13, 1995; Revised and Accepted June 19, 1995

ABSTRACT

Protein-coding genes in the ancient eukaryote *Giardia lamblia* lack typical promoter consensus elements. We have analysed the immediate 5' flanking sequences of seven genes of related function (structural cytoskeleton proteins) to identify shared DNA motifs that might have a role in transcription initiation. Transcription start sites for five genes have been determined previously. Genomic mapping and mRNA primer extension experiments demonstrate additionally that the genes for β -giardin and median body protein are (i) present as single copies in the genome, (ii) transcribed with very short 5' leader sequences. Two search algorithms designed to extract conserved motifs from either aligned or non-aligned sequences independently discovered three sites constituting a common pattern in all seven promoters. Sites were optimally aligned using weight matrix building trials to achieve the maximum 'information content'. Profiling the information content of best alignments defines the extent of the homologies as: a 9 bp box (initiator) at the start site and upstream 18 and 6 bp boxes. The initiator is the most highly conserved element and contains a universal Py–A–Pu motif at which transcription starts. We show that the best matrices can be combined in a search pattern that correctly locates transcription start sites in genomic DNA sequences.

INTRODUCTION

Giardia lamblia is a fascinating organism. It is a primitive eukaryote with a relatively simple molecular biology. No evidence has been found for introns in genomic sequences (1) and transcripts 5'-mapped by primer extension experiments show that transcription is initiated very close to the first translated codon (2–4). Both characteristics contribute to the extreme compactness of the genome, estimated to be 10.8 Mbp of DNA (5). Other features point to an affinity with prokaryotic genetic systems. *Giardia* ribosomal subunits are smaller than for other eukaryotes (6). The small subunit rRNA (16S in *Giardia*) has greater sequence homology to archaeobacteria than to other eukaryotes (7) and codon usage in expressed genes is more similar to *Halobacterium halobium* than to other protozoans (8). The analysis of small subunit rRNA led Sogin *et al.* (7) to conclude that *Giardia*

diverged early from the root of the eukaryotic radiation and retains some primitive molecular characteristics. However, despite the atypical ribosome composition and the fact that characterized mRNAs have rudimentary leader sequences, isolated *Giardia* mRNAs have been translated effectively by ribosomes from higher eukaryotes (9).

Nothing is known of regulatory mechanisms controlling gene expression in *Giardia*. The life cycle consists of two cellular phenotypes: the trophozoite and the cyst. The switch from one phenotype to the other will involve production of stage-specific proteins (for example cyst wall components), together with developmental events such as dismantling and rebuilding the cytoskeleton. In other cells expression at class II promoters uses *cis*-acting sequence elements to activate transcription by RNA polymerase II (Pol II). In most examples elements involved in basal transcription are the TATA box, initiator and proximal upstream promoter elements, which bind *trans*-acting protein recognizers, the general transcription factors (GTFs) and their activating proteins (10,11). Kirk-Mason *et al.* (2) have mapped transcription start sites for three tubulin genes in *G.lamblia* genomic DNA and Alonso and Peattie (3) and Nohria *et al.* (4) have likewise sequenced the upstream regions of genes for α - and γ -giardin, proteins which are associated with the cytoskeleton. None of these genes has an identifiable TATA box sequence in the upstream region, although they share homologies around the start site (9). TATA-less promoters have been studied in other organisms. *In vitro* reconstitution and expression experiments have established that in these cases a more prominent role is assumed by the initiator site to correctly position Pol II (12). Most of these characterized genes are concerned with housekeeping functions, where expression may be constitutive or weakly regulated and transcription proceeds at a relatively low rate.

In this paper we map the 5' flanking regions of the β -giardin and median body protein genes. Both of these gene products are coiled coil microtubule-associated proteins (13,14). The β -giardin mapping clears up the confusion over the 5' sequence of this gene (15). Thus two more genes are added to the set of *Giardia* cytoskeleton genes for which upstream genomic sequences are known. From this minimal data set we use computer procedures to extract common sequence features that might be recognized by regulatory mechanisms acting on these regions. Because in other respects the molecular genetics of *Giardia* are atypical, we conducted this search in the first instance without reference to the database of control elements from other eukaryotes. The analysis identified three different quasi-homologous signatures shared by

* To whom correspondence should be addressed

the sequences. These putative recognition boxes are also positionally synchronized. One signal is at the transcription start site, one is at a position approximately -30 bp relative to this site and the third may appear more than once in the upstream region between -40 and -70 bp.

MATERIALS AND METHODS

Cloned *G. lamblia* genes

Genomic DNA sequences corresponding to the targeted genes were screened from a library constructed in λ ZAPII (Stratagene) plated on *Escherichia coli* XL1-Blue (14). To prepare the library, DNA from *G. lamblia* (Portland-1 strain) was digested with mung bean nuclease in the presence of 40 or 45% formamide (16) to give fragments predominantly in the range 1–10 kbp. Two clones containing the β -giardin gene were expression screened with rat R2B-specific antibodies (13). Presence of the gene was confirmed by PCR amplification of a correct sized band using primers complementary to the terminal regions of the β -giardin sequence (17). The primers (24mers) were 5'-ACGCTCAC-CCAGACGATGGACAAG-3' and 5'-TTAGTGCTTTGTGAC-CATCGAGAG-3'. One positive clone (GLZ48) was used to amplify a probe from the same primers. GLZ48 contained the β -giardin gene in a 2.2 kbp genomic fragment. For the median body protein gene clone GLZ21 (14) was used for further sequence analysis and as template for probe synthesis.

Sequencing

For sequencing Bluescript plasmids with gene inserts were isolated on *E. coli* XL1-Blue after phagemid induction of clones. Plasmid DNA was purified as minipreps (Promega Magic minipreps) and the double-stranded template sequenced by dideoxynucleotide chain termination (18) after alkali denaturation (19). Sequencing reactions used modified T7 DNA polymerase or *Taq* polymerase (Promega) primed by appropriate oligonucleotides synthesized for chromosome walking.

Southern hybridization analysis

DNA was purified from 1.5×10^8 *G. lamblia* trophozoites by a standard method (method 1; 19). Cells were cultured at 37°C in TY1S-33 medium supplemented with bile (20) and were washed twice in TBS (10 mM Tris, 154 mM NaCl, pH 7.4) at 4°C. DNA samples were digested with a restriction endonuclease (*Eco*RI, *Bam*HI or *Hind*III) and precipitated by adding 0.1 vol 2 M NaCl and 2 vol absolute ethanol. Pellets washed in 70% ethanol were taken up in TBE buffer (0.89 M Tris, 0.89 M boric acid, 20 mM EDTA, pH 8) and run on 0.8% agarose gels (Nusieve 3:1; FMC BioProducts). For Southern blotting DNA was transferred to an immobilon S membrane (Millipore) and cross-linked with UV illumination.

Biotin-labelled probes were synthesized from PCR gene copies with the NEB Phototope kit (New England Biolabs). PCR products were first cleaned on QIAGEN tip 5, followed by QIAEX desalting, twice eluting with 20 ml TE buffer (10 mM Tris, 1 mM EDTA, pH 8). The blot was pre-incubated for 1 h at 68°C in 6 $\times$ SSC (0.9 M NaCl, 0.09 M Na citrate) containing 10 μ g/ml sonicated salmon testes DNA, before 15 h hybridization

in the same buffer containing 20 ng/ml probe DNA. Visualization used alkaline phosphatase lumigen PPD substrate as detailed in the Phototope Detection kit (New England Biolabs). The blot was first washed twice in 2 $\times$ SSC, 0.1% SDS at room temperature and twice more in 0.1 $\times$ SSC, 0.1% SDS at 68°C. For re-probing blots were stripped in 0.4 M NaOH, 0.1% SDS for 30 min at 80°C.

RNA primer extension

Total RNA was extracted from harvested trophozoites by the method of Cathaler *et al.* (21). Reverse transcription from appropriate primers used AMV reverse transcriptase (Promega) in a standard protocol (19), but without incorporating radio-labelled nucleotides. The primers were 5'-TCGTCGGGCTTGT-CCAT-3' (β -giardin) and 5'-CCATGACCTGCTCAGCA-3' (median body protein). Highest resolution results were achieved when sequence maps were post-labelled with biotinylated probes to produce chemiluminescence autographs. A sample of cDNA transcribed from the 5' mRNA was run on a sequencing gel alongside a standard DNA sequencing ladder from the same primer on a cloned gene template. The gel was blotted on to immobilon S membrane using 0.5 $\times$ TBE as transfer buffer. Labelled probes complementary to the sequence in a position 5' to the extension primer were hybridized to blots and visualized (as for Southern hybridization). For the β -giardin gene the probe was an oligonucleotide (24mer) 3'-labelled by serial addition of biotin-16-UTP using terminal deoxynucleotidyltransferase (Promega). The probe for median body protein gene was a PCR product (305 bp) synthesized upstream of the extension primer to incorporate biotin-16-UTP.

Computational analysis of sequence similarities

Seven *Giardia* promoter sequences were abstracted from the GenBank/EMBL database into a data file. The sequences were 109 bases upstream from (and including) the first translated position of the following cytoskeleton genes (with accession nos): α -tubulin (J04648); β 1-tubulin (X06748); α 1-giardin (X52485); α 2-giardin (M34550); β -giardin (X85958); γ -giardin (X55287); median body protein (X64517). Programs implementing the algorithms of Galas *et al.* (22) and Stormo and Hartzell (23) were written to run under MS-DOS on an IBM PC-compatible computer using the EGA graphics standard.

For the matrix method (23) information content (I_{seq}), measured in bits, was derived from the base frequencies at each position from the equation:

$$I_{\text{seq}} = \sum_{b=A,C,G,T} f_b \log_2(f_b/p_b) \quad (1)$$

where f_b is the observed frequency of each base and p_b is the fraction of each base in the genome. Because the genomic composition of *Giardia* is known for only a few coding sequences, which are biased by codon usage, bases were treated as equally probable ($p_A = p_C = p_G = p_T = 0.25$). To rank matrices, I_{seq} was summed across all positions. For plotting the information content profile of a site, the calculation included a statistical correction for small sample sizes (24). This correction cancels the background signal that results from oversampling bases in relation to the composition of the genome. The magnitude of the error for a collection of seven sequences is estimated to be 0.37 bits/base.

RESULTS

Copy number of the β -giardin and median body protein genes

Southern blots of single restriction digests of *G.lamblia* genomic DNA demonstrated that only one fragment in each digest hybridized with the β -giardin probe (Fig. 1). The β -giardin probe was a preparation of biotinylated randomly primed copies of the β -giardin open reading frame (ORF). The result localizes this sequence to a single position in the genome, which was restricted by *Eco*RI to a 23 kbp fragment, by *Bam*HI to a 6.5 kbp fragment and by *Hind*III to a 3.6 kbp fragment. The *Hind*III fragment is large enough to contain more than one copy of the β -giardin gene or related sequences. We used polymerase chain reaction (PCR) amplification to map the distances between the ends of the β -giardin sequence and the ends of the fragment. Only single PCR products were obtained, indicating a single gene copy positioned in the fragment as shown in Figure 1C. Our direct sequencing of the β -giardin site in the genomic clone GLZ48 includes sufficient upstream and downstream sequence to definitely rule out the possibility of a tandem related sequence in this fragment.

This result suggests that the cDNA sequence reported previously by us (17) and the near identical genomic sequence later reported by Aggarwal *et al.* (15) are in fact readings derived from the same genomic locus. The latter genomic sequence hybridized to a *Hind*III fragment of the same size (4.2 kbp), within experimental error (15).

The probe to the median body protein gene was copied from a 1726 nt stretch of the ORF in the λ ZAPII clone GLZ21 (14). The probe spanned one internal *Bam*HI site and two internal *Hind*III sites (Fig. 1D). Southern hybridization to single restriction digests of genomic DNA produced labelled fragments of the sizes expected. One *Eco*RI fragment was labelled and was of sufficient length (8 kbp) to contain more than one gene copy. However, the fact that only two labelled *Bam*HI fragments (5 and 1.3 kbp) were produced from cutting the internal site indicates that this site exists in only one copy in the genome. In confirmation, digestion with *Hind*III resulted in two labelled fragments matching the distances between known internal sites (Fig. 1B). The smallest expected fragment (315 bp), with which the probe overlapped by 140 nt, was not detected by hybridization.

5' Sequence of β -giardin

Aggarwal *et al.* (15) determined a genomic sequence that added to the 5' nucleotide sequence of β -giardin read previously from a cDNA copy (17). They found an additional in-frame ATG that extended the ORF by 11 codons. Because of the inconsistencies in the published sequences, we re-sequenced the β -giardin gene and ~800 upstream nt from the genomic clone GLZ48. The β -giardin sequence entry in the GenBank/EMBL database has been updated with the new data (accession no. X85958). The β -giardin ORF starts two codons further upstream from the ATG read by Aggarwal *et al.* (15). Their sequence begins differently and appears to correspond to an *Eco*RI site which is not present in our genomic sequence. Thus the extended ORF translates 13 more residues at the N-terminus of β -giardin than originally reported by us (17,13). These are MSMFTSTRTLTQT. The complete deduced protein, therefore, has 272 residues with a molecular weight of 30.877 kDa and a calculated isoelectric point of 5.11.

Figure 1. Southern blot analysis of *G.lamblia* genomic DNA. Lanes are 10 μ g DNA digested with *Bam*HI (B), *Hind*III (H) or *Eco*RI (E). (A) Hybridization to a probe prepared from a 795 bp PCR copy of the β -giardin sequence in clone GLZ48. A single matching fragment is found in each digest (mapped in panel C below). (B) After stripping, the same membrane hybridized to a 1726 bp probe copied from within the median body protein ORF in genomic clone GLZ21. Labelled bands correspond to the restriction fragments predicted from the known sequence of GLZ21 (diagrammed in panel D). (C) Map of restriction enzyme sites in the vicinity of the β -giardin gene. Known sequence is shown as boxes, the β -giardin ORF as a solid box. Bars above the map represent sized PCR products amplified from single digest genomic fragments. β -Giardin internal sequences and restriction site end-ligated VECTORETTE cassettes (48) were used as primer sites. The positions of restriction sites confirm that there is only one copy of the β -giardin sequence in the 3.6 kbp *Hind*III fragment. (D) Restriction enzyme map of the sequenced ~3 kbp 5' portion (box) of the median body protein clone GLZ21. The solid box is the median body protein ORF. The bar above indicates alignment of the hybridization probe. The entire probe sequence is contained in one *Eco*RI fragment. An internal *Bam*HI site creates two hybridizing fragments. Sizes of potentially hybridizing *Hind*III fragments are shown beneath.

Figure 2. Primer extension analysis of *G.lamblia* total RNA to determine starting nucleotides of mRNAs for (A) β -giardin and (B) median body protein. Specific 17mer primers were complementary to positions +40 to +56 bp (β -giardin) and +84 to +100 bp (median body protein) from the translation start. Sequence gels run with a standard dideoxynucleotide sequencing ladder (A, C, G, T) and cDNA primer extension products (PE) prepared in the absence of dideoxynucleotides. For visualization, ladders were blotted onto nylon membranes and bands probed by hybridizing matching biotinylated 5' DNAs (see Materials and Methods). Transcribed DNA sequences are shown to the left of each panel with the first translated ATG marked. Both mRNAs have multiple starts and short leader sequences.

Transcription start sites

Reverse transcription of the 5' leader sequences on mRNAs for β -giardin and median body protein mapped the initiation sites for transcription. The presence of transcripts in total RNA extracted from *G.lamblia* trophozoites and in a poly(A)⁺ fraction was first demonstrated by Northern blotting. In each case the probe hybridized to a single band, corresponding to a 1.06 kbp β -giardin mRNA and a 2.9 kbp median body protein mRNA (result not shown). Maps of the 5'-ends of these strands are shown in Figure 2. The cDNAs produced by primer extension were labelled by hybridizing specific biotinylated probes. This method gave clearly defined multiple bands identifying the termini of the extension products against a sequence ladder. In relation to the first codon of the ORF, β -giardin transcripts start between nucleotide positions -5 and -10 and median body protein transcripts start at positions -3, -4 and -5. Thus both genes are transcribed with very short leader sequences, in accordance with results obtained previously from other *Giardia* genes (9).

Sequence patterns in promoters

Promoters of seven *Giardia* cytoskeleton genes were examined for conserved sequence patterns that might be involved in a common regulatory mechanism. In other promoters transcription signals mostly conform to motifs at characteristic locations (25). We wished to analyse the *Giardia* sequences without prior assumptions about consensus strings, i.e. at the most basic level

Figure 3. Aligned *Giardia* promoter sequences scanned for common subsequences by the method of Galas *et al.* (22). A weighted frequency score is plotted against the position of the right-most base in the score window. Parameters are: word length six bases, window width eight bases, allowed mismatches one. Consensus words are shown above the main peaks. The horizontal line is a confidence level; ~1% of aligned random sequences are expected to achieve this score. (A) Sequences aligned on the first translated base. (B) Sequences aligned on the initiator consensus.

of finding unknown patterns imperfectly represented in the set. A number of methods formally addressing this problem have been published. We found that two procedures in particular, the Galas-Waterman method (22,26) and the method of Stormo and Hartzell (23), gave consistent and interesting results.

The first method scans pre-aligned sequences and determines the most frequent complete 'word' at each position. Positional information is thus intrinsic to the definition of a consensus pattern. Some positional flexibility is incorporated by allowing the common words to be found in a wider search window. Figure 3 plots score profiles, where the score is the weighted frequency of the consensus word. In this example the words are 6mers, with one allowed mismatched base occurring in a window of eight positions, which we found to be the most discriminating parameters. The weighting is 1 for each perfect match and 0.83 for imperfect matches (22). The *Giardia* sequences were aligned initially on the first translated codon (as in Fig. 4). Regions of similarity were: adjacent to the coding region (coinciding with the transcription initiator), a broad zone of ~20 nt between positions -39 and -17 and a weaker upstream site centred on position -49. The consensus words at these peak positions are indicated. The initiator consensus is strongly conserved. At the other sites the score indicated that three or four of the seven sequences conformed to the local consensus under the constraint of pre-alignment. The sequence correspondence found at these key sites has a very low probability (<0.06%) of occurring by chance in aligned random sequences (27). Sequences were re-aligned on the T/C-A step in the initiator consensus to produce the second profile in Figure 3. This refinement aligned further motifs in about half of the sequences at an additional upstream site centred on position -41.

The common motifs that contribute to the consensus matches in these results were extracted independently by the second algorithm (23,28). This method finds the most strongly related

Figure 4. *Giardia* cytoskeleton gene promoters. Part of the data set closest to the coding region, covering all of the sites identified by the analysis. The full sequences extended 108 bases 5' to the first translated codon (italicized on right). Note that numbering is from the translation start. Sequence order: α 1-giardin, α 2-giardin, β -giardin, γ -giardin, median body protein, α -tubulin, β 1-tubulin. Arrows with multiple roots mark transcription start sites mapped by primer extension experiments; a heavy root indicates the most prominent band, where known. Motifs (capitals) extracted by the search matrices are: nine wide initiator sites (overscored); 18 wide '-30 box' sequences (boxed); upstream six wide sites (underlined), including the highest scoring site per sequence (double underlined). For the γ -giardin sequence the '-30 box' and initiator sites overlap by two bases. For median body protein an alternative '-30 box', also selected by the 18 wide matrix, starts seven positions closer to the initiator. This potential site is better aligned with the other promoters, but has a lower score and is thus not the objective choice.

motifs regardless of position, though, in practice, the matched sequences were generally found at equivalent positions. The algorithm proceeds in stages. Briefly, for a given word size a large set of nucleotide frequency matrices is built iteratively by aligning the many possible word combinations obtainable from the sequences being compared. Generations of matrices are developed by adding sequences in turn from the data set and the outcome may be sensitive to the order in which this is done. The comparison is not exhaustive, since only the highest scoring subset is kept from one generation to the next. A score is attributed to each matrix from its information content (see Materials and Methods) and is used to rank the set. The information content is measured in bits/base, totalling to bits/site. It expresses the reduced choice brought about by conformity; for example, if at a given position only one of the four alternative bases is observed, the implied constraint is a gain of 2 bits of information. In the next stage the top ranking matrices are used to search the data set to locate the mostly closely conforming sequence positions. Finally, the extracted patterns can be aligned and the information content of the alignment calculated on a position-by-position basis to interpret the important features of the site (24).

The data set of *Giardia* promoter sequences was processed repeatedly by the algorithm, varying the matrix width (between six and 24 elements) and sequence order. Results tended to converge on several best scoring matrices distinguishing alternative sequence patterns. These were maximally differentiated by choosing the most appropriate site width (23). Distributions of matrix scores for three widths are shown in Figure 5. The most discriminating matrices were those of highest information content, in the right-most columns.

The first site is best described by a nine wide matrix (Fig. 5A). Two matrices out of 215 stand well above the distribution. Both

result from aligning sequences at the initiator site (overscored in Fig. 4). Used to scan the promoter sequences, the top ranking matrix returned six of the seven initiator sequences as producing the highest scores. The exception was the γ -giardin initiator, which scored slightly lower than a matching sequence 19 positions further upstream in the promoter. The complete set of initiator sites, including the γ -giardin initiator, was the alignment represented by the second ranking matrix. This alignment is plotted in Figure 6A, to show the profile of the initiator region in terms of information content. In accordance with the analysis, the site spans nine base positions where the information content is significantly enhanced. Note that by maximising the information content, objective mathematical support is brought to an alignment that otherwise might be made intuitively from the runs of A at these sites (9). A key feature is the universal A in the fifth position of the site, preceded by a T (or C in the less conserved γ -giardin initiator). The Py-A step, or the nucleotide following it, maps as the strongest transcription start site for these genes.

The second example (Fig. 5B) relates sequences found 5' to the initiator sites. This region of the promoters tended to be sampled when wider matrices of 15–20 positions were constructed and was optimally aligned by the highest scoring 18 wide matrix. The set of best matches is boxed in Figure 4. Although each promoter sequence was scanned over 108 positions, the conserved motifs are all located in the region proximal to the start site. Individually their positions vary with respect to the initiator; the two sites are not strongly spatially coupled and may be separated by 6–16 other bases or, in the case of the γ -giardin promoter, overlap by two positions. Thus, referenced in each case to the transcription start (taken to be the Py-A dinucleotide), these boxes begin at nucleotide positions -35, -29, -28, -20, -38, -29 and -29.

Figure 5. Histograms summarizing the results of matrix building trials from *Giardia* promoter sequences. Matrices ranked and counted according to their information content (23). The 'best matrices' used to extract sets of related sequence patterns (capitalized in Fig. 4) are arrowed. (A and B) Analysis of 109 base promoter regions: (A) 215 nine wide matrices; the arrowed matrix combines sequences at the initiator; (B) 214 18 wide matrices; the arrowed matrix selects motifs comprising the '-30 box'. (C) 150 six wide matrices processed from 50 base segments upstream of the '-30 box' sequences.

The information content profile of the '-30 box' is shown in Figure 6B. Two peaks corresponding to conserved base choices are separated by positions where there is little or no conservation. Sequences are most closely matched at the 5' peak. The separation of the two peaks is 10.5 bases, which is equal to the pitch of the DNA helix. This observation suggests that the function of this site depends primarily on contacts made with one side of the helix. In the tubulin promoters this site has been identified as a potential 'CCAAT box' (2). We tested all of the *Giardia* sequences against

a weighted CCAAT box matrix derived from aligning 502 promoters from other eukaryotes (29) and found that none matched this standardized definition, even at relaxed stringencies.

The conserved motifs at the two major sites described above dominated the output from the algorithm at most matrix widths. To examine similarity in the upstream region, the analysis was repeated, taking from each promoter the next 50 bases 5' to the '-30 box'. After trials at different matrix widths, conserved motifs were extracted with a short six wide matrix (Fig. 5C). In addition to the strongest matches, a second conforming site was found in all of the promoters except for the β -giardin sequence. These sites are marked in Figure 4. The information content profile is shown for the seven primary sites in Figure 6C and for the set of 13 sites in Figure 6D. From the profile, the trailing TT dinucleotide is the most conserved feature.

Prediction of transcription start sites

The above analysis provides matrix descriptions of the conserved initiator and two other motifs found in all of the promoters. Together they form an ordered sequence of sites adjacent to the point in these genes at which transcription starts; therefore, it was possible to devise a search algorithm to locate start sites in *Giardia* genomic sequences by linking the three matrices by rules specifying the spacing permitted between them (49). The resulting pattern description (available on request) encodes all of the statistical evidence of similarity available from the seven promoters.

The full genomic sequences of the cytoskeleton genes were examined with the search pattern. In total, 11 545 bases of genomic DNA were scanned for matches. The longest continuous sequence was the 3078 bases containing the median body protein gene. Table 1 gives the results when matrix cut-offs were set at a relatively low stringency close to 1 bit/base. The algorithm correctly located the start sites of all seven genes, as intended. In addition, the β -giardin sequence predicted a second start position eight bases upstream of the correct start which met the search criteria at a lower score. No potential start sites other than these were predicted in the *Giardia* DNA.

Table 1. Matches to promoter motifs and the start site pattern found in *Giardia* genes

Gene	bp	Observed start ^a	Number of matches ^b		Initiator	Predicted start sites	
			CAATTT	-30		Location	Score
α 1-Giardin	1239	262-264 ⁽³⁾	5 (4)	1 (1)	1 (2)	262	45.4
α 2-Giardin	1870	211-213 ⁽³⁾	13 (17)	1 (1)	1 (1)	210	45
β -Giardin	1961	778-783	11 (32 ^c)	2 (2)	3 (3)	782	44.1
						774	42.1
γ -Giardin	1807	266 ⁽⁴⁾	13 (15)	1 (2)	3 (8 ^d)	266	42.1
MBP	3078	329-331	28 (27)	1 (1)	3 (3)	330	48
						329	44.7
α -Tubulin	123	103-108 ⁽²⁾	3 (6)	1 (1)	1 (2)	105	48.5
β 1-Tubulin	1467	103-108 ⁽²⁾	7 (6)	1 (1)	1 (1)	105	45.7
Controls							
pBR322	4363		26	0	6	-	-
sv40	5244		104	4	23	-	-
m13	6408		108	3	20	-	-
Adenovirus 2	35937		269	10	72	-	-

^aObserved starts refer to positions in genomic sequence files corresponding to 5'-mapped messengers in the indicated references or in this paper.

^bNumber of individual motifs found in genomic sequences. Numbers in parentheses are matches found by a matrix re-calculated to exclude the target sequence. Cut-off values were: CAATTT box = 7 bits; -30 box = 17 bits; initiator = 11 bits; except for ^ccut-off = 6.6 bits and ^dcut-off = 9.6 bits.

Figure 6. Information content plotted across aligned consensus sites in *Giardia* promoters. Matched motifs identified by best matrices (Fig. 5) were the basis for alignment and calculation of information content in bits/base. Data is corrected for small sample size by an error estimate (24). The variance of the estimate is indicated by bars extending 1 SD above and below the curve at each base position. With the correction the genome information content outside the significant site should sum to zero. Consensus base choices are given beneath. (A) The nine wide initiator site. Total information content = $9.73 (\pm 2.5)$ bits. (B) The 18 wide '-30 box' site. Total information content = $11.87 (\pm 5.0)$ bits. Note that the valley between the two halves of the site has information content no higher than some of the background positions. (C and D) The six wide upstream boxes: (C) the highest scoring boxes selected one per sequence, total information content = $6.34 (\pm 1.7)$ bits; (D) all 13 potential matches selected by the top-ranking search matrix, total information content = $6.24 (\pm 0.92)$ bits.

Given the small number of sequences in the data set and, therefore, the large effect each has in determining the matrix values, it is not surprising that all of the homologous sites were matched. Consequently, we tested the usefulness of the algorithm as a predictor in the following way. Each sequence in turn was removed from the data set and the matrices re-calculated from the remaining six promoters; the gene left out was then scanned with the modified pattern. In every case the correct start site was located using data from only the other six genes. To predict the β -giardin and γ -giardin gene starts the cut-off values were reduced slightly for the 'CAATTT box' site and initiator respectively (Table 1). The other five gene starts were predicted at the original cut-off values. These results show that common information in the promoter sequences is satisfactorily abstracted in the matrix model. Although the proximity of the three different motifs provides the formal description, juxtaposition of the '-30 box' and the initiator suffices to specify unique locations in the *Giardia* DNA, because very few stringent matches to the latter signals appear in the 11.5 kbp examined. As a further control, four complete plasmid and viral genomes providing long stretches of DNA sequence were searched at the same stringency, without finding any matches to the *Giardia* promoter pattern (Table 1). The longest single sequence (36 kbp) from adenovirus-2 has 72

positions matching the initiator signal, none of which is sited in a context reproducing the start site pattern.

DISCUSSION

The *Giardia* genes examined in this paper produce cytoskeleton proteins that are constitutive elements of cell structure. Mapping their respective mRNAs has shown that transcripts are initiated predominantly from an adenine which is one of several following a pyrimidine (usually a thymine) close to the first translated codon (2-4, this paper). This process is not exact, as indicated by the presence of mRNA species mapping to adjacent positions spanning 3 (α -giardins and median body protein) or 6 bp (β -giardin and α and β tubulins). Attention has been drawn previously to the similarity of *Giardia* start site sequences (9) and the common motifs in tubulin genes (2). However, this is the first systematic study of *Giardia* genes where homologies have been extracted by numerical methods in order to construct a matrix representation of the promoter cores. A weight matrix combines sequences that are supposedly alternative expressions of the same genetic signal. It assumes that the response of the potential recognizer can be quantified independently at each nucleotide position. In fact, the dissociation constants of complexes formed

between DNA binding proteins and alternative DNA probes correlate well with scores attributed to the sequences by weight matrix models (30,31). The matrix building approach has demonstrated the fundamental position-dependent similarity of many eukaryotic promoters, at the same time revealing differences between major groups of organisms (invertebrates versus vertebrates, for instance; 29).

Accurate transcription by Pol II depends on a minimal initiation complex of >20 proteins being assembled at the start of a gene (10). In higher eukaryotes the multiprotein factor TFIID normally binds stably to the promoter core through the component TBP (TATA box binding protein). TBP attaches specifically to an ~20 bp stretch of DNA spanning the TATA sequence by sequence-dependent contacts made with the minor groove (11). In the absence of the TATA motif, entry of the GTFs and Pol II depends on interactions with the start site, although *in vitro* transcription still requires the TFIID complex (32–34), including, apparently, TBP (35). Some TATA-less promoters continue to bind TBP to the –30 region, albeit weakly (36). Thus the dispositions of TBP and other TFIID components may be the same for complexes formed on TATA-containing and TATA-less promoters (35).

The common elements in *Giardia* promoters need to be probed experimentally to determine how they contribute to normal transcription. With Pol II recruited to the mapped start sites, consistent elements are placed in appropriate positions (–30 and at the initiator) to interact with GTFs of a committed complex. Yet the sequences found in these positions are unlike their counterparts in other organisms, including the related protistan *Trichomonas vaginalis* (37), which represents a primitive eukaryotic lineage almost as ancient as that leading to *Giardia*.

The most conserved leading positions of the –30 box include a segment of A+T nucleotides. Nonetheless, none of the individual sequences scores above the threshold for a TATA box (29), despite being appropriately spaced from the start site. Interestingly, the stretch with the highest information content covering the first half of the consensus (CAAAA/TA/CT) partially matches an active –30 element in the TATA-less promoter governing transcription of mouse ribosomal protein (rpS16) genes (32). Mutation studies of the latter show that transcription competence depends on the integrity of the hexamer segment AAAAAT. This element apparently binds a specific nuclear factor that is not TFIID, but can be replaced by a conventional TATA box binding a second factor (presumed to be TFIID) without altering promoter efficiency.

DNAs with poly(A) segments are often either intrinsically curved or prone to bend in response to binding proteins (reviewed in 38). The potential curvature vectors (39) of the *Giardia* A+T-rich promoter sequences indicate that the ~50 bp proximal to the start site is a region capable of sustaining a pronounced bend (unpublished calculations). This tendency is greatest over the –30 region, suggesting that the DNA may be locally adapted to wrap over a transcription complex.

The *Giardia* start site sequence is fairly strongly conserved. It can be regarded as a polypurine tract [essentially poly(A)] interrupted by a run of one to four pyrimidines (usually T). This sequence differs from the characterized functional initiator sequences studied in other TATA-less promoters, which take several non-homologous forms (40). In contrast, the seven known *T.vaginalis* promoter elements broadly conform to the Inr or polypyrimidine initiator (consensus PyPyA₄NT/APyPy) from

mammalian TATA-less promoters (41,37), which is also closely related to the eukaryotic CAP site consensus (29). In only one respect can the *Giardia* initiator be related to the Inr element: that the initiating base is an A immediately or closely following a pyrimidine. Mutagenesis studies of mammalian and viral sequences have shown that the +1 adenosine is crucial for promoter efficiency (33,42). Comparing a broad range of organisms, it is the only strictly conserved nucleotide (43).

Because the *Giardia* initiator sequences feature a TA step embedded in a run of compatible Ts and As, they come close to matching the TATA box consensus. The Bucher (29) TATA matrix produces scores above cut-off for the start site sequences of α 1-giardin, β -giardin and the tubulins. In the absence of direct studies it is not possible to judge whether these matches might be significant in terms of binding TFIID; various A+T-rich sequences have been shown to support a functional interaction with the protein complex (44). High A+T content is a general feature of the *Giardia* promoters and other sites match the TATA box motif at a higher score, presumably fortuitously. Altogether, 26 examples can be found in the genomic sequences of the seven genes. Recognition of the initiator sequences in other TATA-less promoters does not rely on TBP (45); more probably, alternative subunits of TFIID are involved in making contact (32,34, 46). There is also evidence that recognition by Pol II itself might be sufficient to recruit the other GTFs (47). An alternative explanation of why the *Giardia* initiators are A+T-rich might be to facilitate DNA strand separation at the start point.

So far, genes with TATA boxes have not been demonstrated in the very earliest protistan lineages leading to *T.vaginalis* and *Entamoeba histolytica*, as well as the even more primitive *G.lambli*a sequences examined here. Probably, at this level of evolution the region around the start site, demonstrated to be functional in *T.vaginalis* (37), is the most important element in determining both the location of transcription and promoter strength. Dependence on the initiator is retained in higher eukaryotes in those genes not subject to strong regulation that remain TATA-less. Without the strong second anchorage provided by TBP–TATA DNA recognition, initiation is less precise. In these circumstances it is thought that the transcription complex might slide on the DNA, producing multiple start sites from adjacent nucleotides (40), as seen in mRNAs of *Giardia* and, to a lesser extent, *T.vaginalis*. Given that *G.lambli*a is the most ancient of the amitochondrial protists, it appears that the primitive start site is a simple repetitive sequence containing a single Py–A step and that this pattern pre-dates the more structured Inr element that unites *T.vaginalis* and higher eukaryotes.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the Wellcome Trust through a project grant. DVH is a Wellcome Senior Research Fellow. We would like to thank John Keyte for synthesizing oligonucleotides.

REFERENCES

- 1 Logsdon, J.M. and Palmer, J.D. (1994) *Nature (Lond.)*, **369**, 526.
- 2 Kirk-Mason, K.E., Turner, M.J. and Chakraborty, P.R. (1989) *Mol. Biochem. Parasitol.*, **36**, 87–100.
- 3 Alonso, R.A. and Peattie, D.A. (1992) *Mol. Biochem. Parasitol.*, **50**, 95–104.

- 4 Nohria, A., Alonso, R.A. and Peattie, D.A. (1992) *Mol. Biochem. Parasitol.*, **56**, 27–38.
- 5 Fan, B., Korman, S.H., Cantor, C.R. and Smith, C.L. (1991) *Nucleic Acids Res.*, **19**, 1905–1908.
- 6 Boothroyd, J.C., Wang, A., Campbell, D.A. and Wang, C.C. (1987) *Nucleic Acids Res.*, **15**, 4065–4084.
- 7 Sogin, M.L., Gunderson, J.H., Elwood, H.J., Alonso, R.A. and Peattie, D.A. (1989) *Science*, **243**, 75–77.
- 8 Char, S. and Farthing, M.J.G. (1992) *J. Protozool.*, **39**, 642–644.
- 9 Adam, R.D. (1991) *Microbiol. Rev.*, **55**, 706–732.
- 10 Buratowski, S. (1994) *Cell*, **77**, 1–3.
- 11 Tjian, R. and Maniatis, T. (1994) *Cell*, **77**, 5–8.
- 12 Weis, L. and Reinberg, D. (1992) *FASEB J.*, **6**, 3300–3309.
- 13 Holberton, D.V., Baker, D.A. and Marshall, J. (1988) *J. Mol. Biol.*, **204**, 789–795.
- 14 Marshall, J. and Holberton, D.V. (1993) *J. Mol. Biol.*, **231**, 521–530.
- 15 Aggarwal, A., Adam, R.D. and Nash, T.E. (1989) *Infect. Immun.*, **57**, 1305–1310, correction p. 3950.
- 16 McCutchan, T.F., Hansen, J.L., Dame, J.B. and Mullins, J.A. (1984) *Science*, **225**, 625–628.
- 17 Baker, D.A., Holberton, D.V. and Marshall, J. (1988) *Nucleic Acids Res.*, **16**, 7177.
- 18 Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
- 19 Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd Edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- 20 Keister, D.B. (1983) *Trans. R. Soc. Trop. Med. Hyg.*, **77**, 487–488.
- 21 Cathala, G., Savouret, J.-F., Mendez, B., West, B.L., Karin, M., Martial, J.A. and Baxter, J.D. (1983) *DNA*, **2**, 329–335.
- 22 Galas, D.J., Eggert, M. and Waterman, M.S. (1985) *J. Mol. Biol.*, **186**, 117–128.
- 23 Stormo, G.D. and Hartzell, G.W. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 1183–1187.
- 24 Schneider, T.D., Stormo, G.D., Gold, L. and Ehrenfeucht, A. (1986) *J. Mol. Biol.*, **188**, 415–431.
- 25 Breathnach, R. and Chambon, P. (1981) *Annu. Rev. Biochem.*, **50**, 349–383.
- 26 Waterman, M.S. and Jones, R. (1990) *Methods Enzymol.*, **183**, 221–237.
- 27 Waterman, M.S., Arratia, R. and Galas, D.J. (1984) *Bull. Math. Biol.*, **46**, 515–527.
- 28 Hertz, G.Z., Hartzell, G.W. and Stormo, G.D. (1990) *CABIOS*, **6**, 81–92.
- 29 Bucher, P. (1990) *J. Mol. Biol.*, **212**, 563–578.
- 30 Berg, O.G. and von Hippel, P.H. (1987) *J. Mol. Biol.*, **193**, 723–750.
- 31 O'Neill, M.C. (1989) *J. Mol. Biol.*, **207**, 301–310.
- 32 Hariharan, N. and Perry, R.P. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 1526–1530.
- 33 Smale, S.T. and Baltimore, D. (1989) *Cell*, **57**, 103–113.
- 34 Pugh, B.F. and Tjian, R. (1991) *Genes Dev.*, **5**, 1935–1945.
- 35 Hernandez, N. (1993) *Genes Dev.*, **7**, 1291–1308.
- 36 Wiley, S.R., Kraus, R.J. and Mertz, J.E. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 5814–5818.
- 37 Quon, D.V.K., Delgadillo, M.G., Khachi, A., Smale, S.T. and Johnson, P.J. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 4579–4583.
- 38 Harrington, R.E. (1992) *Mol. Microbiol.*, **6**, 2549–2555.
- 39 De Santis, P., Palleschi, A., Savino, M. and Scipioni, A. (1990) *Biochemistry*, **29**, 9269–9273.
- 40 Zawel, L. and Reinberg, D. (1993) *Prog. Nucleic Acids Res. Mol. Biol.*, **44**, 67–108.
- 41 Javahery, R., Khachi, A., Lo, K., Zenzie-Gregory, B. and Smale, S.T. (1994) *Mol. Cell. Biol.*, **14**, 116–127.
- 42 Lee, R., Concino, M. and Weinmann, R. (1988) *Virology*, **165**, 51–56.
- 43 O'Shea-Greenfield, A. and Smale, S.T. (1992) *J. Biol. Chem.*, **267**, 1391–1402.
- 44 Hahn, S., Buratowski, S., Sharp, P.A. and Guarente, L. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 5718–5722.
- 45 Pugh, B.F. and Tjian, R. (1990) *Cell*, **61**, 1187–1197.
- 46 Kaufmann, J. and Smale, S.T. (1994) *Genes Dev.*, **8**, 821–829.
- 47 Carcamo, J., Buckbinder, L. and Reinberg, D. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 8052–8056.
- 48 Copley, C.G., Boot, C., Bundell, K. and McPheat, W.L. (1991) *Biotechnology*, **9**, 74–79.
- 49 Staden, R. (1988) *CABIOS*, **4**, 53–60.