**Giardia** gene predicts a 183 kDa nucleotide-binding head-stalk protein

Jonathan Marshall and David V. Holberton*

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

*Author for correspondence

**SUMMARY**

Previously described extended proteins from the cytoskeleton of *Giardia lamblia* (β-giardin, median body protein) have been found to be segmented coiled coils with regular structural repeat patterns in their amino acid sequences. Screening a λZAPII library derived from *Giardia* genomic DNA with an antibody directed against a $34 \times 10^3 \text{ M}_r$ giardin isoform selected a gene encoding a much larger polypeptide chain (HPSR2), the sequence of which was determined by chromosome walking the open reading frame. The complete gene has been cloned and expressed as a recombinant protein of $183 \times 10^3 \text{ M}_r$. The predicted amino acid sequence of the protein has identifiable features suggesting that it might be a motor protein with an amino-terminal hydrolytic domain attached to a long coiled coil stalk. The presumed head domain is 211 residues and contains a P-loop sequence conserved in purine nucleotide-binding proteins. The remaining 1409 amino acids mainly make up a region of heptad repeats such as in myosin or the kinesin stalk, ending in a small (67 amino acids) carboxy-terminal domain. Fourier analysis of the predicted stalk shows the presence of a strong physical repeat created by regular heptad phase changes dividing the coil into segments of 25 residues. This structure most closely resembles the smaller microtubule-associated median body protein which has segments of 24 residues.

Key words: *Giardia*, coiled coil, nucleotide binding, cytoskeleton, DNA sequence

**INTRODUCTION**

Coiled coil structures are of considerable importance in the cytoskeleton. The tendency for parallel amphipathic α-helices to lock together by hydrophobic side chain docking (Crick, 1953) provides a means whereby rigid elongated molecules suitable for struts and filaments can be generated. Major coiled coil families (tropomyosins, myosins, intermediate filament (IF) proteins) were recognised early in fibrous tissues. Recent sequencing projects have discovered further notable examples in the cytoplasm (e.g. kinesin (Yang et al., 1989), desmoplakin (Green et al., 1990), tektin, (Norrander et al., 1992)), and concerned with division events in the nucleus of yeast (Alani et al., 1989; Mirzayan et al., 1992; Kolling et al., 1993) and mammalian cells (Yang et al., 1992). In the case of the best characterised motor proteins, kinesin and sarcomeric myosin heavy chains, long coiled coil segments are seen to anchor or transmit the forces produced at the hydrolytic domain to enable useful mechanical work to be done.

We are interested in the coiled coil proteins expressed in the zooflagellate *Giardia*, which may be the most primitive of known eukaryotic organisms (Sogin et al., 1989). Previously, we have sequenced two coiled coil proteins from different loci in the microtubule cytoskeleton (Holberton et al., 1988; Marshall and Holberton, 1993). Their genes were isolated by expression screening cDNA and genomic libraries with monospecific polyclonal antibodies to cytoskeleton proteins. Coiled coil amino acid sequences are recognised by characteristic heptapeptides (heptads) with alternating 3,4 residue spacing of apolar side chains. The two *Giardia* proteins are unusual in that their heptad series are regularly interrupted by extra residues which modify the apolar Fourier period. This feature has the effect of dividing the coils into successions of similar segments. Beta-giardin is one of a large number of ~30 kDa proteins (giardins) found in the cytoskeleton (Clark and Holberton, 1988; Peattie et al., 1989). It is a small coiled coil with segments of four heptads followed by a skip residue (Holberton et al., 1988). The second protein is a larger (101 kDa) three-domain protein from the median body bundle of microtubules. Its long central rod domain has segments of 24 residues created by periodically reversing the phase of the apolar repeat (Marshall and Holberton, 1993). Such structural patterns might have some role in the association of these rod proteins with microtubules. A structural homologue of β-giardin, somewhat dissimilar in sequence but with an identical heptad pattern, has now been found in fibres of the flagellar basal apparatus of the green alga *Spermatozopsis similis* (Weber et al., 1993).

We have been using rat antibodies to ~30 kDa polypeptides from the giardin isoform cluster to screen for other β-giardin-like coiled coils. One antiserum to a component more basic than β-giardin bound strongly to the immunising isoform and, at a lower level of reactivity, to β-giardin itself. Used to screen a genomic λZAPII expression library (Marshall and Holberton, 1993), the antibodies isolated two clones with the same insert. The expressed gene is an open reading frame (ORF) much longer than a ~30 kDa polypeptide, but which translates to an amino acid heptad sequence. The complete ORF has now been...
resolved as encoding a 183 kDa head-stalk type of protein. We report this sequence in the present paper. By homology, the fold of the head domain incorporates a mononucleotide phosphate binding site, as in a myosin or kinesin heavy chain. However, similarity to these two motor proteins does not extend beyond the active site in the head domain and the stalk heptads. The head domain is also smaller, about two thirds the size of the kinesin head. Like Giardia median body protein, the stalk sequence is regularly patterned by heptad phase changes, but in this case the coil segments so formed are of 25 rather than 24 residues.

**MATERIALS AND METHODS**

**Library screening**

Primary screening antibodies were polyclonal rat antisera raised to specific Giardia cytoskeleton proteins, as previously described (Holtberton et al., 1988; Marshall and Holberton, 1993). Proteins were isolated by two-dimensional sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) of detergent-insoluble cytoskeletons dissolved in 9 M urea (Clark and Holberton, 1988). One antiserum (R1 antibodies) screened the clones described in this paper from a genomic DNA expression library in ZAPII constructed previously (Marshall and Holberton, 1993). The R1 immunogen was a prominent 34 kDa spot running in the most basic region (pH ~6.0) of the focusing gradient, well separated from the main giardin isofrom cluster (visible in Fig. 1a of Marshall and Holberton, 1993).

Library fragments were mung bean nuclease fragments of genomic DNA from Giardia lambia (Portland-1 strain). Inserts tended to coincide with coding regions because of the preference of the enzyme for A-T rich intergene sequences (McCutchan et al., 1984). For immuno-screening, amplified library samples were plated on E. coli XL1-Blue, and lacZ expression induced with isopropyl-1-thio-

c-galactoside (IPTG) (Huynh et al., 1985). Nitrocellulose filter lifts were blocked (5% Marvel dried milk powder in Tris-buffered saline: 10 mM Tris-HCl, 154 mM NaCl, pH 7.4, with 0.05%(v/v) Tween-20 (TBS/T)) before 12-16 hours incubation in primary antibodies in TBS/T with 5-10% foetal calf serum. Visualisation used secondary antibodies were polyclonal rat antisera raised to Giardia cytoskeleton proteins from SDS-PAGE. Electro-blotted transfers onto nitrocellulose were incubated 16 hours in primary antiserum diluted 1/1,000 in TBS/T with 10% foetal calf serum (Crossley et al., 1986). Preimmune serum was used as a control. Labelling was 4 hours incubation with 1/1,000 HRP-conjugated goat anti-rat IgG (ICN).

**Sequencing**

Part of the genomic sequence SR2 was found initially in two identical library clones. The DNA insert from one was sequenced on both strands. Subclones with oppositely directed serial deletions of the insert were produced using either EcoRI-mung bean nuclease (Stratagene) or EcoIII-S1 nuclease (Erase-a-base, Promega) digestion. Dideoxy chain termination sequencing (Sanger et al., 1977) used either modified T7 DNA polymerase or TAQ polymerase kits (Promega) to read into the inserts from Bluescript T7 or T3 primer sites. To cover gaps, oligonucleotide primers matching internal sites were custom synthesised (Nottingham University Biopolymer Synthesis and Analysis Unit). To extend the sequence, segments of genomic DNA were amplified from libraries of EcoRI or HindIII restricted DNA end-ligated to Vectorette 1 cassettes (Cambridge Research Biochemicals). The SR2 open reading frame was completed on three overlapping PCR (polymerase chain reaction) fragments band-purified from 1% agarose gels with QIAEX DNA extraction (Qiagen). Fragments were cycle sequenced as double stranded templates using CircumVent polymerase (New England Biolabs) and custom primers.

**Production of recombinant protein**

For expression and characterisation of the product, Giardia DNA sequences were ligated into the vector pRSET (Invitrogen Corporation) used to transform E. coli BL21(DE3)PlysS. Two inserts were prepared in the correct reading frame. A 1546 bp segment of DNA encoding the first part of the SR2 ORF was cloned via 5' BamHI and 3' HindIII sites. The DNA was amplified by PCR from BamHI restricted Giardia genomic DNA using appropriate oligonucleotide primers. The upstream primer (50 nt) was engineered to include a BamHI site. The second primer (21 nt) matched the SR2 sequence downstream of internal HindIII sites (Fig. 1). Double digested samples were gel band purified before ligating to the vector DNA. The second insert included the full SR2 ORF. To reduce the risk of introducing copying errors, it was reconstructed from the cloned pGLZ42 3.1 kb sequence, to which the missing 5' 1.8 kb was added from a longer PCR product. The two overlapping segments were restricted and ligated at a unique AvaI site (position 2505 of the coding sequence). At the 5' end, the construct incorporated a BamHI site in the PCR primer. For ligation into the vector, the 3' was restricted at a convenient KpnI site near the end of the pGLZ42 sequence.

Expression was induced with 1 mM IPTG once cultures had grown to A600 0.5 (~2 hours). After 3 hours, inclusion bodies were released by freeze-thaw lysis, and washed in 1% Nonidet NP40 (Nagai and Thogersen, 1987). Recombinant proteins solubilised in 6 M guanidine hydrochloride were collected and washed on His-bind resin by a standard procedure (Novagen, pET system manual).

Three-month rats were immunised with ~25 µg of pure protein emulsified in Freund's non-ulsurative complete adjuvant (Guilhady Antiser). Four boosting injections were given at 14 day intervals. The antibodies were characterised by immunoblotting samples of recombinant and Giardia cytoskeleton proteins from SDS-PAGE. Electro-blotted transfers onto nitrocellulose were incubated 16 hours in primary antiserum diluted 1/1,000 in TBS/T with 10% foetal calf serum (Crossley et al., 1986). Preimmune serum was used as a control. Labelling was 4 hours incubation with 1/1,000 HRP-conjugated goat anti-rat IgG (ICN).

**Sequence analysis**

Sequence management and translations were carried out using Mount PC software (Mount and Conrad, 1986). Sequence analysis, searches and alignments used programs at SEQNET (SERC, Daresbury) in STADEN and GCG (Devereux et al., 1984) packages, accessing SWISSPROT and OWL databases. For secondary structure prediction, Fast Fourier analysis, and coiled coil heptad analysis of amino acid sequences, algorithms were programmed locally to support IBM-PC EGA and VGA graphics displays.

**RESULTS**

**Gene analysis**

Two plaques showing strong R1 antibody reactivities were found in ~2x10^5 p.f.u. from the expression library. The two isolates were amplified and rescreened for analysis. Both had inserts of ~3 kbp and proved by end sequencing to be copies of the same original clone. The Bluescript plasmid excised from one isolate (pGLZ42) was sequenced using a nested deletion strategy. Overlapping end-deleted subclones were produced to read in both directions by timed endonuclease digestion. A map of the sequencing subclones is shown in Fig. 1. The pGLZ42 insert comprised 3103 bp and was a single ORF, open from the 5' end to a stop codon 22 bp from the 3' end. Working from a 21 nt oligonucleotide primer near the 5' end, the ORF was extended by chromosome walking on three overlapping restriction fragments. This method, called 'chemical genetics' (Copley et al., 1991), makes use of ligated
adaptors for PCR screening restriction libraries to rapidly isolate matching DNA sequences. Both strands of the amplified fragments were sequenced by synthesising appropriate oligonucleotide primers.

The assembled nucleotide sequence is a 4892 bp stretch of genomic DNA spanning the single open reading frame, herein designated the \textit{SR2} gene. It will appear in the EMBL, GenBank and DDBJ Nucleotide Sequence Databases under the accession number X79815. The ORF itself is 4860 bp encoding the amino acid sequence given in Fig. 2. There is no obvious poly(A) addition signal in the short 3' non-coding sequence. The result from Southern hybridization of a 546 bp probe amplified from the 3' end of the ORF indicated a single copy of this sequence exists in the genome (result not shown). The ORF contains 32 ATG codons. Those \textit{Giardia} genes that have been sequenced at genomic level are found to code as continuous ORFs whose product is straightforwardly translated from the first ATG (Adam, 1991). The nucleotide sequence was scanned in three frames for regions of base bias consistent with coding (Staden, 1990). The ORF is the only reading frame showing positional base preferences typical of a coding region and is distinct in this respect from the immediately flanking nucleotide sequences. The codon position bias in G+C content for the 3 positions is, respectively, 56%, 31%, and 54%. Codon usage for the translated ORF is shown by the program CORRESPOND to be very similar to the median body protein gene (D^2 = 2.95). Thus, a priori arguments suggested that the ORF would encode a polypeptide chain considerably longer than, and probably different from, the immunising 34 kDa giardin subunit. Direct synthesis of the corresponding protein from the recombinant gene proved this interpretation to be correct (see below).

**Expression of the \textit{SR2} gene**

The \textit{SR2} DNA sequence was expressed in the vector pRSET (Invitrogen) which produces a version of the recombinant protein that is His-tagged at the N-terminus. The product was purified from inclusion bodies by metal chelation affinity chromatography. On gels (Fig. 3A), it ran with a mobility corresponding to 189 kDa (the extra residues added by the vector cloning site contribute 3758 to the molecular mass). This result confirmed that the ORF sequence is translated in vivo as a single full length polypeptide.

We also synthesised from a shortened DNA insert a truncated 64 kDa chain spanning the first 512 amino acids of the predicted sequence. Bacterial hosts produced this fragment at very high yields (Fig. 3A). The pure protein was used to raise rat antibodies recognising epitopes encoded at the beginning of the ORF, which were not part of the cloned sequences originally screened by the non-specific R1 antibodies. The resulting antiserum had high avidity for the \textit{SR2} recombinant proteins; furthermore it clearly recognised a single 184 kDa band corresponding to the complete gene product in \textit{Giardia} cytoskeleton proteins (Fig. 3B). This component was not detected by R1 antibody staining; conversely, there was no labelling of ~30 kDa components by the \textit{SR2}-specific antibodies.

**Predicted amino acid sequence**

The ORF encodes a product, HPSR2 (= heavy chain protein from \textit{SR2} gene), of 1620 amino acids, with predicted molecular mass 183,166 Da and isoelectric point 5.14. Heptad repeats in the amino acid sequence demonstrate the protein is a member of the coiled coil superfamily. Heptad patterns arise where \(\alpha\)-helical strands are joined at hydrophobic strips on their surfaces. Apolar residues are expected once per helical turn in the amphipathic sequence \((a,b,c,d,e,f,g)_n\), where \(a\) and \(d\) are the apolar positions. Such a sequence allows two strands to twist together in a continuous rope (Crick, 1953). Algorithms for predicting secondary structure (Chou and Fasman, 1978; Garnier et al., 1978) identify the central region of HPSR2 as a highly probable \(\alpha\)-helical domain. Analysis of heptads in this region (see below) suggests that a coiled coil rod could be formed from Ser212 to Lys1553. For this stretch of 1342 residues, 1207 (90%) are predicted in the \(\alpha\)-helical conformation using a three state model (Gibrat et al., 1987). Thus this central domain could dimerize as a rod 199 nm long accounting for 83% of the molecular mass. The first 211 residues appear to form a separate head domain, and the rod ends in a small compact domain of 67 residues which is not predicted as coiled coil. The two terminal regions contain 7 of the 8 proline residues in the sequence and are prospectively folded as...
Fig. 2. Amino acid sequence of the protein chain HPS2 blocked into three domains with the central domain arranged in heptad repeats. The domains are taken from the prediction in Fig. 5. The N-terminal domain contains an NTF-binding motif (see Fig. 7). The central domain is a string of amphipathic heptad motifs, potentially a rod of 1342 residues, giving rise to the regular periodicities seen in Fig. 4. The sequence is here mostly arranged in 25-residue segments according to the Fourier period. Apolar residues (AVMILFYW) in the a and d positions are enclosed in vertical rectangles. Altered a → d phasing at the ends of segments accounts for the pattern. The heptad run is essentially continuous to Ser1486 where 2 inserted valves disrupt the pattern. Italicised residues are quantitatively weak matches to the heptad scoring matrix and might indicate local structural features, notably adjacent to Pro299 which is the only proline residue in the central domain. The alignment of segments shows strong side chain preferences in some positions. Note particularly basic charges in the segments accounts for the pattern. The heptad run is mostly arranged in 25-residue segments and phasing at the ends of

N-terminal domain:

```
1  M S R P S T A Q R Q S L E Q D S V S S L L L N T S N
57  S R Y F S N E K T K E Q N E L I K S K Y L Q S L V K L S
85  E D T E I S E Y K A I N D K Y L L K E A D T D M R K
113  M N E Q L L E S L E K A I T I N Q L S S G G S G K S G T
141  A E I I S G N A L S T A Q S R T I S T L H K L F T D L
169  D M I Q V P T E Q A N N M S E P E L V Q Q N A I N L A V
197  Y Q K R L K V E A D S R R S F
```

Coil:

```
212  S T E S E A Q    M D N L Q R Q   |   I V K I | E L E
233  L V E S R N M    V S F L E K E   C D R
250  L K N S Q E N    V V N L R T O   V Q S R D E I
271  L Q D K S M L    I S T L Q N Q   I K V M Q A E
292  K V D | A A Q A    P F V A R N D   L S L A E T
313  I K K K S V T    I E K L M E K   L S K L K T
334  S Q V A    I S Q R D Q D   A V E L A R S   V A D M R Q K
359  | L S D T    V D E K Q R I   V H Q L H I D   W D S M K K S
384  H E E S    I S Q R P R S E   I E A I K S A   F D E S T F I
409  K N A R   I S K K T T E   L T E T Q T O   L A S A L E K
434  N Q D S E S K | A | Q H A   A E V I L L A E
455  R E I A N T T I L T Q E   L Q A T K G E   L E A V R K S
480  N H E Q    V N S Y E S L   L Q K T Q Q S   Y E A L L R E
505  K E D T I L K L N E C D N A K K A K F D D H S E Q
530  S T A H V Q A L R E E   V E R V R H T   S E N L L L Q E
555  K M G V I N R L T A E   I Q A I K L D G E R A L D S
580  K D T Q I S S L M E N    V S L K T S S N
605  S S A E L S R L T L N    L E T V Q R L E S
630  K D E Q L R I A T N E   Y K E N M E K   L D A V F A K
655  K D A E L S K O T A A   L H A L R D E   L S Q V K S N
680  H L L E    V D E K L K E   V N S A Q Q N   T S M T L L S
705  N Q D E I S R L G R E    L D V V K N S   A S I N E V E
730  | A N K K I Q S N L K    I I S L Q Q E   L Q D S K N E
780  A A K A A Q N K I G    Q N T V D K Q K Q   L S S V S S D
805  K L I A S E M M R V   V S E L K H E   L S T E K Q M
830  R E A E N S R A Q L Q   I S H L E A Q   V K D A I A K A
855  K D S E    V T R L L S D   L K T V K E E   L A I V V D Q
880  K D A K   I A E R S R K   L E D T F H K   L Q R S E Q T
905  | V E V L Q A A K K E   L S V A K L N   T D Q T I A L
930  L N D R    V A N I I L A E   L N K Q K N E   T E E L L A F
955  K D L Q I Y K Q L Q C K Q   L E D S R T E   V T E A T S S
980  G R A E    I S R L Q C O I D N L G R E A   L Q N A Q I E
1005  | Y A Q K E D A L K D D   L N S A K A V   L L A E S A E
1030  K D A I    I S J K K D   L T N L R A E   L S S E E A
1055  K D V T I A R A Y Q D C   R H L Q T S   L T K S I E K
1080  K D H A I L K K Q A K Q F A   V S A A K K A Q A E
1105  | Y E A Q A I S L T G D   L I A A A K K K   Q S E L E M E
1130  I E E R E M K H A S S   A S K K A Q A E   L K V T I D Q
1155  K Q R E L E E L K R | Y L S S Q A E   A L S T N K R
1180  L L Q D E   I D K L L Q O   K G G L I N Q   I S S L K Q E
1205  | T S V S S A D
1230  R E H Q    A K K A D K D   I K T L Q D A   L R I A E E D
1273  K R V L E L G H S K   L E E I K Q Q   V Y C L I S
1262  K N K S    I T E L N Q K   V R D M Y N E   V I E A Q A A
1287  K N N E    I A R A A E K L T K S Q T H L D I V S E
1312  K D E K L A K M E A D   L A R V Q K E   I R A L E I S
1337  K K T E A L L K R E Q G E L T L R Y S A E T N M
1362  K D E R I V A L E K E   L G S F R D N I   L A R T T K L
1387  K D D E I R S L E A Q   L R D T K K   S Q T A L H E
1412  N T O L I D I T Q V K   M S D Q V Q S E A S A L K K
1462  K D A E I Q E L T R K   L G T L L E A Q   I L G Q T V S S
1487  V V    Y D A K V E Q | L D E K E R T V E L R A E
1510  | I A R L K E A
1517  S Q A T S M V S G E E N T K L A L K   | V K Q L E A E
1542  | L E E | H | R A V H D F I K
```

C-terminal domain:

```
1554  P Q V M E Q K L E M N A A P T E E A S F M K Q L I K
1582  E N G Y L K Q R L N Q I A L G S D E S T R S N S S I S E
1610  G R R P G G R L R T K
```
globular domains. The very C-terminus of the chain is a strongly basic decapeptide.

Coiled coil domain

The extent of the coiled coil domain was deduced from heptad patterns. Heptad regions were identified by Fourier analysis and by database comparison with known coiled coil heptad sequences using a NEWCOIL algorithm (Lupas et al., 1991). These procedures were used previously to examine the structure of median body protein (Marshall and Holberton, 1993). In Fig. 4A the HPSR2 sequence is mapped by scanning Fourier transform of apolar residue positions. Two pronounced Fourier components characterise the presumptive coil region after residue 211; these show as contoured streaks across the map. The main streak is derived from a second order heptad repeat and the second streak is close to the third order heptad repeat. A Fourier transform of the entire rod sequence (Fig. 4B) shows the exact values of the two Fourier periods to be 3.57 and 2.27 residues, respectively. These periods are those particular orders of a fundamental 25 residue repeat that beat most closely with heptad harmonics (Table 1), i.e. 3.57 = 25/7 \equiv 7/2; 2.27 = 25/11 \equiv 7/3.

The NEWCOIL algorithm plots the probability of coiled coil sequence by scoring heptad sequences in a travelling window (Fig. 5). Like median body protein, the HPSR2 heptad pattern is modified by frequent heptad phase changes. Care is needed in interpreting the probability plot for this type of structure because out-of-phase heptads in the same window weaken the score in a periodically fluctuating manner. The plot thus appears more interrupted than for a conventional coiled coil; nonetheless, the overall region of potential coiled coil structure can be seen. Our implementation of the algorithm also records
The HPSR2 amino acid sequence analysed as a potential coiled coil sequence using the algorithm of Lupas et al. (1991). Ordinate shows probability of coil structure derived from residue frequencies in a database of amphipathic helix patterns. Scores assessed over a travelling window of 21 residues. The plot indicates a high likelihood of coiled coil structure between residues 212 and 1553. Out-of-phase effects in the window, introduced by segment boundaries, tend to disrupt the score at intervals. More extensive weaknesses in the prediction around positions 390 and 700 imply structural faults, but the heptad pattern can be read through these regions (Fig. 3). The stave above the probability plot is a diagram of the 7 alternative heptad reading frames with heptads contributing to the probability score plotted into their corresponding frame. In each frame the vertical position of a heptad is determined by its individual score. The diagram shows graphically how, in the coil region, heptad phasing regularly jumps 3 frames (a → d shift) at the end of each segment. Note that a short run of amphipathic helix is predicted also in the head domain.

Table 1. Fourier analysis of significant periodicities in the rod domain sequence

<table>
<thead>
<tr>
<th>Amino acids</th>
<th>Period</th>
<th>Rational equivalent</th>
<th>I</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apolar</td>
<td>3.57</td>
<td>25/7</td>
<td>30.67</td>
<td>4.8 x 10^{-14}</td>
</tr>
<tr>
<td>(A,V,L,I,M,F,Y,W)</td>
<td>2.27</td>
<td>25/11</td>
<td>27.17</td>
<td>1.6 x 10^{-12}</td>
</tr>
<tr>
<td>3.13</td>
<td>25/8</td>
<td>6.76</td>
<td>1.2 x 10^{-3}</td>
<td></td>
</tr>
<tr>
<td>Positive charges</td>
<td>3.13</td>
<td>25/8</td>
<td>12.83</td>
<td>2.7 x 10^{-6}</td>
</tr>
<tr>
<td>(K and R)</td>
<td>8.33</td>
<td>25/3</td>
<td>7.79</td>
<td>4.1 x 10^{-4}</td>
</tr>
<tr>
<td>2.5</td>
<td>25/10</td>
<td>6.52</td>
<td>1.5 x 10^{-3}</td>
<td></td>
</tr>
<tr>
<td>2.08</td>
<td>25/12</td>
<td>6.44</td>
<td>1.6 x 10^{-3}</td>
<td></td>
</tr>
<tr>
<td>Negative Charges</td>
<td>2.27</td>
<td>25/11</td>
<td>26.0</td>
<td>5.1 x 10^{-12}</td>
</tr>
<tr>
<td>(D and E)</td>
<td>3.57</td>
<td>25/7</td>
<td>14.88</td>
<td>3.4 x 10^{-7}</td>
</tr>
<tr>
<td>2.14</td>
<td></td>
<td></td>
<td>7.01</td>
<td>9.0 x 10^{-4}</td>
</tr>
<tr>
<td>2.66</td>
<td></td>
<td></td>
<td>5.96</td>
<td>2.6 x 10^{-3}</td>
</tr>
<tr>
<td>8.35</td>
<td>25/3</td>
<td>5.3</td>
<td>5.0 x 10^{-3}</td>
<td></td>
</tr>
</tbody>
</table>

Fast Fourier transformation was performed on the HPSR2 putative coil sequence, residues 212 to 1553, embedded in zeros to a linear series of 2048 elements. The most significant peaks for each type of residue are shown. The transforms were scaled by the method of McLachlan and Stewart (1976). The probability (P) of a peak of intensity I occurring by chance is exp(-I).

Fig. 5. The HPSR2 amino acid sequence analysed as a potential coiled coil sequence using the algorithm of Lupas et al. (1991). Ordinate shows probability of coil structure derived from residue frequencies in a database of amphipathic helix patterns. Scores assessed over a travelling window of 21 residues. The plot indicates a high likelihood of coiled coil structure between residues 212 and 1553. Out-of-phase effects in the window, introduced by segment boundaries, tend to disrupt the score at intervals. More extensive weaknesses in the prediction around positions 390 and 700 imply structural faults, but the heptad pattern can be read through these regions (Fig. 3). The stave above the probability plot is a diagram of the 7 alternative heptad reading frames with heptads contributing to the probability score plotted into their corresponding frame. In each frame the vertical position of a heptad is determined by its individual score. The diagram shows graphically how, in the coil region, heptad phasing regularly jumps 3 frames (a → d shift) at the end of each segment. Note that a short run of amphipathic helix is predicted also in the head domain.

High scoring heptad matches in their corresponding reading frame (Fig. 5). This plot helps to locate heptad phasing. For example, used with a tropomyosin sequence, the scored heptads all appear in the same frame, whereas for the nematode myosin II rod (McLachlan and Karn, 1983) the trace changes frame four times at the skip residue positions, as expected. For HPSR2, a pattern emerges of regular segments repeatedly frame-shifting throughout most of the coil domain. These are the segments of 25 residues detected as a Fourier periodicity. They are aligned in Fig. 2 which identifies all of the heptads in the HPSR2 sequence between Ser212 and Lys1553. The heptad sequence is nearly continuous but some irregular pattern changes or weakly predicted heptad might interrupt the coil (see legend to Fig. 2). The section from Ser334 to Ser1486 has 45 repeats of 25 residues and is almost completely regular in this respect, save for one missed phase change at position 434 and an extra heptad from Ile1205 to Asp1211. The frame shifts in this section are all heptad a → d position phase changes. The top end of the coil from Ser212 to Thr333 has no segments, but according to the heptad frame plot there is a single d → a heptad phase change at Leu250. It also contains a potential proline helix break in a weak heptad. Possibly this region is a neck providing a hinge between the head domain and more regular stalk. The C-terminal section of the coil is punctuated by two additional residues breaking the heptad series near Val1487. The heptad phasing is not easy to assign after this point because of irregular apolar positions. In Fig. 2 we maintain a reasonably continuous pattern through this region while indicating the weakly matched heptads. Despite these weaknesses, scoring windows between 17 and 28 residues consistently predict a coil extending to Lys1553.

Like other α-fibrous proteins, the rod domain has a high proportion of apolar and charged residues defining the two surfaces of the amphipathic strand (Cohen and Parry, 1986). In this domain there are 349 apolar residues (26%), almost half being leucine, and 466 charged residues (35%) with an excess of 48 acidic charges. Charged residues also contribute strongly to the segment boundaries, being more concentrated in the 4-residue extra turns breaking the heptad pattern. Here, 50% of the residues have charged side chains, with a similar number of acidic and basic groups. Typically, the cluster involves a charge pair placed in the centre of the apolar seam (Fig. 6A). Plotting net charge at segment positions shows strong ± net pairs also at the f/g positions in the first full heptad, and at the e/g positions in heptad 2 (Fig. 6B). Thus the distribution of acidic residues mostly conforms to the heptad repeat positions whilst less typical positions for basic side chains create significant new Fourier intensities for the periods 8.3 (25/3) and 3.125 (25/8) residues (Table 1).

Head domain

From secondary structure predictions, the first 211 residues fold as a compact domain of amphipathic helices, β-strands and loops. One short region is homologous with the defining glycine-rich P-loop of purine nucleotide phosphate (NTP)-binding proteins (Saraste et al., 1990). Consequently, database searches with the N-terminal domain sequence produced a series of part matches to proteins with NTP binding sites. Fig. 7 shows alignment of the HPSR2 motif with examples of P-loop sequences, including myosin and kinesin. The Giardia protein sequence is not a match to the P-loop definition in the PROSITE directory (Bairoch, 1991) because of the substitution of Leu for Gly in the first position of the consensus. However, the same substitution occurs in some known ATP-binding proteins, causing Gorbalenya and Koonin (1989) to introduce a separate category of ‘short’ motif, to which the Giardia sequence conforms. The role of the canonical first Gly of the P-loop motif is in any case obscure. Crystallographic
studies have demonstrated that its side chain is not obviously involved in binding substrate analogues (Dreusicke et al., 1988; Tong et al., 1991).

Adenylate kinase (AK1) usefully represents NTP-binding proteins because the atomic coordinates are known for variants of the enzyme from the pig (Dreusicke et al., 1988), E. coli (Muller and Schulz, 1992) and yeast (Egner et al., 1987). We examined HPSR2 and mammalian AK1 sequences for similarity by dot-plotting at different stringencies. Short runs of limited homology were found to coincide with two important structures at the active centre of AK1 (Fig. 8). The first corre-

sponds to the P-loop and includes the immediately upstream β-strand and downstream α-helix involved in binding the phosphate chain. When the sequences are optimally aligned, this site, originally called the ‘A-site’ (Walker et al., 1982) has about 42% identity, 58% similarity over 40 residues, which is a significant homology (Sander and Schneider, 1991). The similarity is apparent also in the local plot of Chou-Fasman secondary structure propensities which predict a strand-loop-helix structure for this site in the HPSR2 head domain (Fig. 9). The second homology in the dot plot is a weaker match to an α-helix immediately following the classical ‘B-site’ motif (Walker et al., 1982) in AK1. The B-site includes 1 or 2 aspartic acid side chains at the end of a β-strand thought to be involved in chelating Mg$^{2+}$. The B-site sequence has not been found in all NTP-binding proteins (Fry et al., 1986). A short sequence of hydrophobic residues in HPSR2, matching a β-strand run and ending with two aspartic acids, is found 11 residues N-terminally from the dot plot homology. Thus a plausible B-site structure embracing these components in HPSR2 can be aligned with the AK1 site by incorporating a longer loop between β-strand and α-helix. The best gapped alignment (Fig. 8B) shows that the overall homology is modest (25%) and remains only a tentative match to the functional site. It is also the case that two residues of the classical B-site consensus located 9 and 4 positions N-terminally from the β-strand motif are missing from the Giardia HPSR2 sequence.
DISCUSSION

We have found a long open reading frame in the genome of Giardia lamblia which can be expressed in full in a recombinant bacterial host, and which produces a 183 kDa component of the cytoskeleton. The protein is potentially a motor protein heavy chain with a large coiled coil stalk domain. The relationship, if any, of the heavy chain to the smaller giardin antigen is unknown. The giardin subunit is unlikely to be cleaved directly from the longer chain. Attempts at Edman peptide microsequencing showed it to be blocked by N-terminal modification indicating that it is not a cleavage product. Antigenic overlap of the two proteins may be due to similar heptad-based sequence epitopes recognised by R1 antibodies.

The HPSR2 sequence is interesting because it is structurally explicit with interpretable homologies to known functional domains. Moreover, patterns in heptad repeats relate it to other Giardia coiled coils. These proteins share the property that the apolar seam uniting two strands is regularly displaced by heptad faults. HPSR2 is most similar to median body protein where the relevant faults are phase changes, and which is also a three domain protein. In general, however, HPSR2 resembles a myosin or kinesin heavy chain where an active mononucleotide-binding domain is tethered to a long coiled coil stalk, but lacks specific homology to either of these two protein families outside of the functionally convergent sites.

The head domain contains recognisable homologues of the A and B substrate contact sites of nucleotide-binding proteins. In three dimensions, the two sites are adjacent structural units of a Rossmann folding pattern at the core of NTP-binding proteins. Nucleotide binding is in a cleft between the loops emerging C-terminally from successive strands of the central $\beta$-sheet (Fry et al., 1986). The P-loop of the A-site binds the phosphate chain of ATP or GTP. Contact is mediated by proximity of the $\alpha$-phosphate to the Gly-Lys-Thr/Ser run marking the transition into the following $\alpha$-helix (Bradley et al., 1987). At the same time, the $\beta$- and $\gamma$-phosphates are complexed with Mg$^{2+}$ to one or more negative charges at the end of the $\beta$-strand identified with the B-site. For GTP-binding proteins like ras p21, an additional consensus motif in this region, Asp-X-Gly, couples the negative charge and a glycine involved in $\gamma$-phosphate contact (Tong et al., 1991). This motif is seen also in the HPSR2 homology in Fig. 8B. In

Fig. 8. (A) Dotplot showing similarity in the amino acid sequences of adenylate kinase (vertical) and the head domain of HPSR2, residues 1 to 211 (horizontal). Similarity was examined with the program DIAGON taking amino acid exchange values from the MDM78 frequency matrix (Staden, 1982). The plotted points are matched positions in windows of 25 residues where the overall score has a probability <5 x 10$^{-5}$ of occurring by chance. Two segments of the HPSR2 head domain match regions in adenylate kinase adjacent to the ATP-binding site. Segment 1 (S1) is an alignment coinciding with the kinase A-site (P-loop). Segment 2 (S2) is aligned downstream of the B-site. (B) Optimum alignments of the homologies in (A). Gapped alignments of bovine adenylate kinase and HPSR2 were examined using BESTFIT with different penalties to find the highest scores for various segment lengths. Dashes indicate identical residues, double and single dots conservative substitutions. Upper alignment: segment 1 of HPSR2 head domain aligned with the adenylate kinase sequence (AK1) around the A-site (overscored). Maximum homology of 42%, similarity of 58%. Lower alignment: best alignment of segment 2 with the adenylate kinase sequence following the B-site (overscored). The position of the long gap was adjusted manually to produce an identity of 25% and similarity of 52.5%.
adenylate kinase (Dreusicke et al., 1988) is aligned in the top panel. The beta-loop-alpha X-ray structure of the A-site in porcine sheet, and turn propensities as 4 residue averages (Chou and Fasman, 1978). The helix surface is 100° for the former, 50° for the latter. In the predicted HPSR2 structure, the disrupted turns are marked by pairs of opposite charges in the contacting surfaces (Fig. 6). Charge pair formation across the two strands involving the $e$ and $g$ inner positions is a usual means of stabilising coiled coils (Cohen and Parry, 1986). Similar pairs also involving the $d$ positions at the phase changes in HPSR2 might help hold strands together at the vulnerable points where interleaved apolar side chain packing cannot be maintained. The long side chains of the lysine residues in these positions might allow the positional flexibility required to accommodate the reorientation of the seam.

It is likely that structural proteins similar to the Giardia coiled coils are present in other eukaryotic organisms. The beta-giardin conformation is found also in SF-assemblin from the green flagellate Spermatozopsis similis (Weber et al., 1993). SF-assemblin is isolated from striated fibres attached to flagellar root microtubules that are similar to the striated ribbons of beta-giardin in the Giardia cytoskeleton. We suggested in earlier discussions of these patterns that segmentation of coil structures might enable fitting to microtubule lattices for microtubule-associated proteins (Holberton et al., 1988; Marshall and Holberton, 1993). We do not know if this is likely to be relevant for HPSR2. Another possibility raised earlier was that because of the ancestral position of the organism, we might be examining a set of prototype coiled coil sequences. These genes were perhaps assembled originally from short heptad-encoding segments to match the dimensions of polymeric combining surfaces in the cytoskeleton. However, we now know that these structures are found in other protists less primitive than Giardia. Also, we have partly sequenced a Giardia coiled coil gene screened by anti-muscle tropomyosin which has a conventional continuous apolar seam (unpublished). Giardia, therefore, appears to express both conventional and segmented coiled coils, which implies that there are strong functional reasons for the segmented patterns.

HPSR2 the putative B-site occurs N-terminal to the A-site, which is atypical. In principle, because the polypeptide chain is recurrently threaded through the beta-sheet, juxtaposition of the sites can be achieved whatever their order in the sequence (Taylor and Green, 1989). In AK1, ras 21p, and other well characterised small cytosolic NTP-binding proteins (Moller and Amons, 1985), the A-site is the first full loop in the fold, within the first 20–30 residues of the N-terminal. In contrast, in the more complex myosin and kinesin head domains, as in the HPSR2 sequence, the A-site position is well downstream from the N-terminus (myosin residue 178; kinesin residue 92; HPSR2 residue 131).

There are no further homologies in the head or tail terminal domains to suggest combining sites for other proteins in an active complex, equivalent to the myosin actin-binding or kinesin tubulin-binding sites, or indeed other PROSITE (release 8.10) catalogued binding sites for cytoskeleton proteins (Bairoch, 1991).

The 25 residue segments in HPSR2 which impose structural stutters on the continuity of strand coiling are seen to arise from a series of heptad $a \to d$ phase substitutions. By contrast, the altered phasing in median body protein is a series of $d \to a$ changes giving 24 residue repeats (Marshall and Holberton, 1993). The beta-giardin skip residue pattern is a third variation. One consequence of displacing the seam in the segmented coils is to change their pitch. Comparing the alternative patterns, in median body protein the reduced pitch will cause the coil to be overwound, whereas the extra residues in beta-giardin and the phase changes in HPSR2 lengthen the pitch and lead to partial unwinding. McLachlan and Karn (1983) estimated from model building that the distortion around a skip position in myosin is spread over a stretch of 14–21 residues. The displacement introduced by a phase change is less severe. The rotation on the helix surface is 100° for the former, 50° for the latter. In the predicted HPSR2 structure, the disrupted turns are marked by pairs of opposite charges in the contacting surfaces (Fig. 6).

We gratefully acknowledge the support of the Wellcome Trust through a project grant. D.V.H. is a Wellcome Senior Research Fellow. We would like to thank John Keyte for synthesising oligonucleotides. This work has benefitted from the use of the SEQNET facility.

REFERENCES


Plasmodium


