BIOINFORMATICS DISCOVERY NOTE

Vol. 28 no. 24 2012, pages 3186–3190
doi:10.1093/bioinformatics/bts626

Sequence analysis

Advance Access publication October 24, 2012

Two distinct SSB protein families in nucleo-cytoplasmic large DNA viruses

Darius Kazlauskas and Česlovas Venclovas*

Institute of Biotechnology, Vilnius University, LT-02241 Vilnius, Lithuania

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Eukaryote-infecting nucleo-cytoplasmic large DNA viruses (NCLDVs) feature some of the largest genomes in the viral world. These viruses typically do not strongly depend on the host DNA replication systems. In line with this observation, a number of essential DNA replication proteins, such as DNA polymerases, primases, helicases and ligases, have been identified in the NCLDVs. One other ubiquitous component of DNA replisomes is the single-stranded DNA-binding (SSB) protein. Intriguingly, no NCLDV homologs of canonical OB-fold-containing SSB proteins had previously been detected. Only in poxviruses, one of seven NCLDV families, I3 was identified as the SSB protein. However, whether I3 is related to any known protein structure has not yet been established.

Results: Here, we addressed the case of 'missing' canonical SSB proteins in the NCLDVs and also probed evolutionary origins of the I3 family. Using advanced computational methods, in four NCLDV families, we detected homologs of the bacteriophage T7 SSB protein (gp2.5). We found the properties of these homologs to be consistent with the SSB function. Moreover, we implicated specific residues in single-stranded DNA binding. At the same time, we found no evolutionary link between the T7 gp2.5-like NCLDV SSB homologs and the poxviral SSB protein (I3). Instead, we identified a distant relationship between I3 and small protein B (SmpB), a bacterial RNA-binding protein. Thus, apparently, the NCLDVs have the two major distinct sets of SSB proteins having bacteriophage and bacterial origins, respectively.

Contact: venclovas@ibt.lt

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received and revised on September 19, 2012; accepted on October 10, 2012

1 INTRODUCTION

The largest currently known viruses belong to the group of eukaryotic nucleo-cytoplasmic large DNA viruses (NCLDVs) (Iyer *et al.*, 2001; Koonin and Yutin, 2010). At present, this group of double-stranded DNA viruses is classified into seven families (Boyer *et al.*, 2009). Replication of these large viruses typically does not strongly depend on the host DNA replication and transcription systems. In line with the large genome size and the relative replication independence, comparative genomics studies revealed that NCLDV genomes encode essential DNA replication proteins, including DNA polymerase, helicase and primase (Iyer *et al.*, 2006; Yutin *et al.*, 2009).

*To whom correspondence should be addressed.

However, conspicuously absent among the identified conserved NCLDV protein families (Yutin *et al.*, 2009) were canonical single-stranded DNA-binding (SSB) proteins. The SSB function in nature is strongly associated with the specific structural solution, oligonucleotide-binding (OB) fold, featuring a five-stranded β-barrel capped with an α-helix (Murzin, 1993). To our knowledge, there are only two documented exceptions, where the structure of an SSB protein is unrelated to OB-fold. The two exceptions include adenovirus (Tucker *et al.*, 1994) and Thermoproteales, a clade of hyperthermophilic Crenarchaea (Paytubi *et al.*, 2012).

Given the importance of SSB proteins in various DNA transactions, the seeming absence of canonical OB-fold SSB proteins in NCLDV genomes was all the more intriguing. Although in poxviruses, one of seven NCLDV families, the SSB protein (I3) has been identified (Gammon and Evans, 2009), whether it is based on the OB-fold or some unrelated structure has not yet been established (Greseth *et al.*, 2012).

In this study, we aimed to resolve the puzzle of 'missing' canonical OB-fold SSB proteins in the NCLDVs and to explore the evolutionary origins of the enigmatic poxviral SSB protein family.

2 RESULTS AND DISCUSSION

2.1 Four NCLDV families have canonical OB-fold SSB proteins

We started by testing the hypothesis that NCLD viruses do possess classical OB-fold-containing SSB proteins, but with strongly diverged sequences. To this end, we performed iterative sequence searches with PSI-BLAST (Altschul *et al.*, 1997) and jackhmmer (Eddy, 2011). Searches were run against the nr70 sequence database (the NCBI non-redundant database filtered at 70% identity) using the E-value = 0.001 inclusion threshold and well-characterized SSB proteins from both cellular organisms and viruses as queries (see Supplementary Material for more details on Methods). To our surprise, when we used the T7 bacteriophage SSB protein (gp2.5) sequence (gi: 9627442) as a query, we detected statistically significant matches (E-value < 0.001) in four NCLDV families (*Phycodnaviridae*, *Mimiviridae*, *Iridoviridae* and Marseillevirus). We scrutinized this finding by testing whether these putative NCLDV SSB homologs are able to detect T7 SSB in a reverse search. For this test, we used HHsearch (Söding, 2005), a sensitive method based on comparison of sequence profiles. We generated sequence profiles for several newly identified putative NCLDV SSB proteins and queried them against profile databases. Searches against the sequence

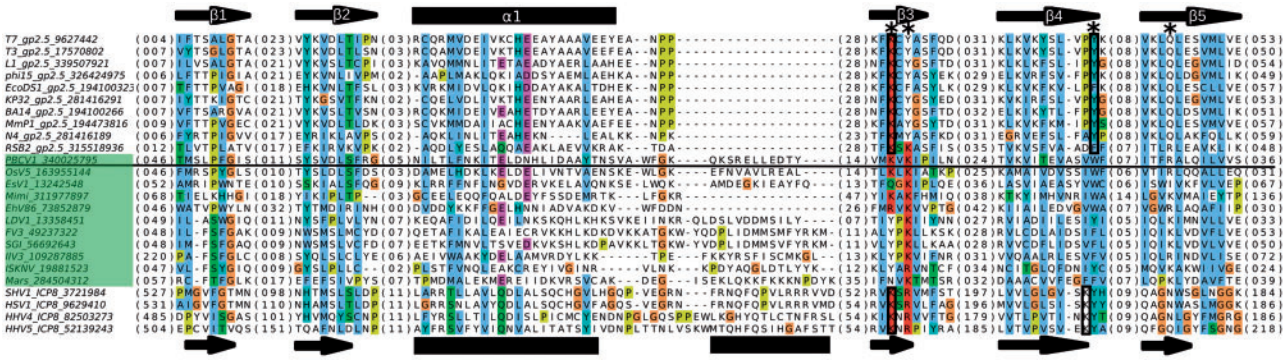

Fig. 1. Alignment of the OB-fold region of SSB proteins from T7-like phages, NCLDVs and herpesviruses. Sequence labels are constructed from the abbreviated virus name, protein name (if available) and the gi number. Secondary structures of T7 gp2.5 and herpes simplex virus 1 (HSV1) ICP8 are shown, respectively, above and below the alignment (β-strands are shown as arrows, α-helices as rectangles). NCLDV SSB residues predicted to interact with ssDNA are indicated with asterisks above the alignment. Residues, whose importance in the interaction with ssDNA was established experimentally in T7 and Suid herpes virus 1 SSBs, are indicated with the black frame. For T7 gp2.5, these are K109 and Y158 and for SHV1 ICP8 K756 and K970. Numbers of residues omitted from non-conserved regions are indicated in parentheses. NCLDV SSB protein labels have green background. *Paramecium bursaria* chlorella virus 1 (PBCV-1) SSB is underlined

profile database derived from known structures (PDB) readily identified T7 gp2.5 with highly significant HHsearch probabilities (>95%). Same queries against the Pfam protein family database detected two families annotated as domains of unknown function (DUFs). One of them (DUF2738) is a self-detection, whereas another family (DUF2815) comprises T7 gp2.5-like sequences, originating mainly from prophages (Supplementary Fig. S1). Taken together, these results provide convincing evidence that four NCLDV families encode T7-like SSB homologs. The detected relationship is not only reliable, but also specific because of distinct features of the T7 gp2.5 OB-fold domain. In T7 gp2.5, the capping α-helix is inserted between strands β2 and β3 in contrast to its typical position between β3 and β4.

In some NCLDV genomes, we detected more than one T7 SSB homolog (Supplementary Table S1). In most such cases, SSB orthologs were identified straightforwardly based on their grouping by sequence similarity using CLANS (Frickey and Lupas, 2004). The ubiquitous presence of the C-terminal tail enriched in acidic residues emerged as an additional defining feature of orthologs. It is known that the acidic C-terminal tail of T7 gp2.5 is the site of interaction for multiple proteins involved in DNA transactions (Marintcheva *et al*., 2006). The functional role of paralogs is not clear, but one of the possibilities is that they are involved in the formation of an oligomeric SSB structure. Oligomeric structures represent a functional form of SSB proteins in organisms as diverse as T7 bacteriophage, *Escherichia coli* and humans (Bochkareva *et al*., 2002; Kim and Richardson, 1994; Raghunathan *et al*., 1997).

One of the phycodnaviruses [*Emiliania huxleyi* virus 86 (EhV86)], in addition to strongly diverged T7 SSB-like protein (Supplementary Fig. S1 and Table S1), has two homologs of eukaryotic/archaeal SSB (replication protein A). The T7 gp2.5-like protein in EhV86 is most distant among the corresponding homologs in other phycodnaviruses, suggesting that there may be a certain functional interplay between the gp2.5-like protein and replication protein A homologs.

The EhV86 case notwithstanding, the identified relationship between putative NCLDV SSB proteins and T7 gp2.5 suggests their similar role in DNA replication and recombination. This notion is strengthened by the genomic context. We noticed that frequently putative NCLDV SSB proteins are encoded in the vicinity of proteins involved in DNA replication and recombination, such as DNA polymerases, helicases, topoisomerases, clamp and clamp loader subunits. Moreover, transcriptomics data indicate that putative SSB genes in Frog virus 3 (*Iridoviridae*), PBCV-1 (*Phycodnaviridae*) and Mimivirus are actively expressed early in the infection, together with DNA replication components (Legendre *et al*., 2010; Majji *et al*., 2009; Yanai-Balser *et al*., 2010).

2.2 Structure of NCLDV SSB proteins is consistent with the single-stranded DNA (ssDNA) -binding function

To get a further insight regarding structure and function of putative NCLDV SSB proteins, we constructed a homology model for a representative (gi: 340025795) from the phycodnavirus PBCV-1. The model was built using iterative optimization of both the sequence–structure alignment and the set of protein structures used as modeling templates (Venclovas and Margelevičius, 2009). The initial PBCV-1 SSB model was constructed using the crystal structure of T7 gp2.5 (PDB: 1je5) as the single template. However, our analysis suggested that the capping α-helix (α1 in T7 gp2.5; Fig. 1) and the adjacent region of PBCV-1 SSB might be more similar to the corresponding OB-domain substructure of the herpesviral SSB protein (ICP8, PDB: 1urj; Supplementary Fig. S2). Indeed, the combination of T7 and herpesviral SSB structures led to a substantial improvement. According to the energy estimation with ProSA-web (Wiederstein and Sippl, 2007), the final PBCV-1 SSB model (available at: http://www.ibt.lt/bioinformatics/models/pbcv1_ssb/) fared better than the T7 gp2.5 structure, for which missing loops were modeled in before the evaluation. Prosa *Z*-score values were −7.34 for the PBCV-1 SSB model and −6.96 for T7 gp2.5 (more negative values imply a more energetically favorable structure). Thus, evaluation results suggested that the model is likely to be sufficiently accurate for the exploration of functional properties.

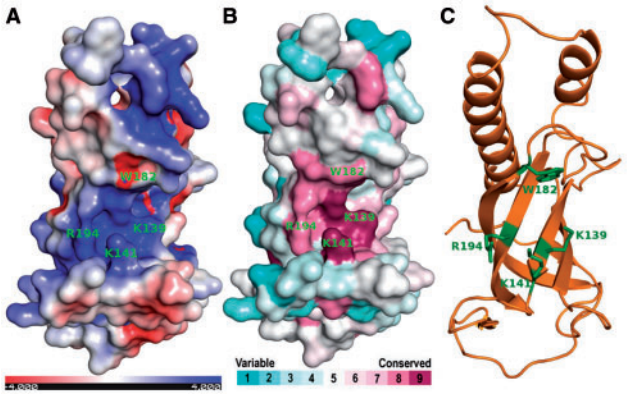

Fig. 2. PBCV-1 SSB structural model and its putative ssDNA-binding site. Residues predicted to be important for ssDNA binding are labeled. (**A**) The surface electrostatic potential map (red color, negative potential; blue, positive; scale units, K_bT/e_c). (**B**) The position-specific conservation of NCLDV SSB proteins mapped onto the surface of the PBCV-1 SSB model (dark red indicates strong conservation and cyan variable regions). (**C**) Cartoon representation of the model with labeled residues shown as sticks

To see whether the modeled PBCV-1 SSB structure indeed is suggestive of the single-stranded DNA (ssDNA) binding, we examined its surface properties (Fig. 2). Consistent with the predicted ssDNA-binding function, we observed an increased positive electrostatic potential in the region, which is thought to bind ssDNA in T7 gp2.5 (Hyland *et al.*, 2003). Moreover, we established that the same PBCV-1 SSB surface region features some of the highest evolutionary conservation (Fig. 2B). Based on the conservation and/or electrostatic properties, we identified four residues that are most likely to participate in the ssDNA binding (Figs 1 and 2).

The involvement of at least two of these residues in ssDNA binding is supported by experimental studies of T7 and herpesviral SSB proteins. Suid herpes virus SSB K756 corresponding to PBCV-1 SSB K139 is critical for ssDNA binding (Wu *et al.*, 2009), while the point mutation of respective T7 2.5 residue (K109I) alters the ssDNA-binding mode (Hyland *et al.*, 2003). The Y158C point mutation renders T7 SSB defective in ssDNA binding (Hyland *et al.*, 2003), presumably by disrupting the stacking interaction with a nucleotide base. Consistent with such a role, the corresponding position in NCLDV SSB proteins (W182 in PBCV-1 SSB) is occupied exclusively by aromatic residues. The SSB protein of Suid herpes virus also has an aromatic residue (Y971) in this position. Although the importance of Y971 has not been addressed directly, the mutation of adjacent lysine (K970) significantly affected ssDNA binding (Wu *et al.*, 2009). The importance of the remaining two positions in NCLDV SSB proteins (K141 and R194 in PBCV-1) has no experimental support coming from either T7 or herpesviral SSBs. However, their location within the putative ssDNA-binding cleft and a fairly strong conservation suggest a likely role in ssDNA binding.

2.3 Poxviral SSB proteins (the I3 family) are evolutionary distinct

In poxviruses, we did not detect either T7 gp2.5-like or any other OB-fold-containing SSB homolog. On the other hand, vaccinia virus, the prototypic poxvirus, has long been known to encode the ssDNA-binding protein, named I3 (Rochester and Traktman, 1998). Several reports have provided convincing evidence that I3 is the replicative SSB protein and a major player in viral DNA recombination (Gammon and Evans, 2009; Greseth *et al.*, 2012). However, despite extensive efforts, no relationship between I3 and any other protein family could be established (Greseth *et al.*, 2012).

To search for homologous proteins related to the I3 family, we first performed standard PSI-BLAST and jackhmmer runs (up to five iterations, inclusion *E*-value = 0.001) against the nr70 sequence database using representative I3 sequences as queries. However, these searches revealed no statistically significant matches outside of the I3 family. To increase sensitivity, we performed the same searches with the more permissive *E*-value inclusion threshold (0.01). This time, jackhmmer (but not PSI-BLAST) readily detected small protein B (SmpB) sequences as significant matches. To test whether the I3 alignment with SmpB is indicative of true homology or is just a spurious match, we performed additional analyses. We noticed that only the central region of I3 sequences was aligned to SmpB. It is commonly known that sequence segments of low complexity or unrelated domains may hinder homology detection. Thus, we reasoned that if I3 and SmpB are homologous, the removal of unaligned regions of I3 should improve the detection of SmpB. This indeed turned out to be the case. One of the N- and C-terminally truncated I3 sequences (Crocodilepox virus I3; gi: 115531731) in the jackhmmer search was now able to detect SmpB using a more stringent inclusion threshold (*E*-value = 0.001).

Next, we performed homology searches using HHsearch. Initially, we generated profiles based on representative full-length I3 sequences and queried them against the PDB-derived profile database. HHsearch detected SmpB from *Aquifex aeolicus* (PDB: 1p6v) as the best match. The most significant results (HHsearch probability = 89%) were obtained using the Yoka poxvirus I3 (gi: 345107239) profile. This result alone would not be considered entirely reliable. However, HHsearch and jackhmmer returned essentially identical alignments for the same central region of I3 (Fig. 3). Moreover, when we ran queries with profiles based on truncated I3 sequences, HHsearch results in detecting SmpB have improved further, reaching the 95% probability (e.g. I3; gi: 41057468). As an additional test, we performed searches in the reverse direction using SmpB profiles as queries against the Pfam profile database. HHsearch did identify poxviral I3 profile as the best match (disregarding self-match to SmpB), but with the insufficient confidence (probability = 90%). Again, if only the aligned region of SmpB was used, HHsearch was able to detect the I3 family with the highly significant 95% probability (SmpB; gi: 240047684). Collectively, all these homology search results strongly suggest common ancestry for I3 and SmpB families.

SmpB is a bacterial protein that binds tmRNA, a hybrid RNA molecule having functions of both transfer RNA and messenger RNA. SmpB adopts a distinct structural fold with no significant similarity to any other structures (Dong *et al.*, 2002). Despite that, SmpB was proposed to contain an 'embedded' OB-fold (Dong *et al.*, 2002), a proposition that appears to be misleading. Our analysis revealed that the SmpB structure is a *bona fide*

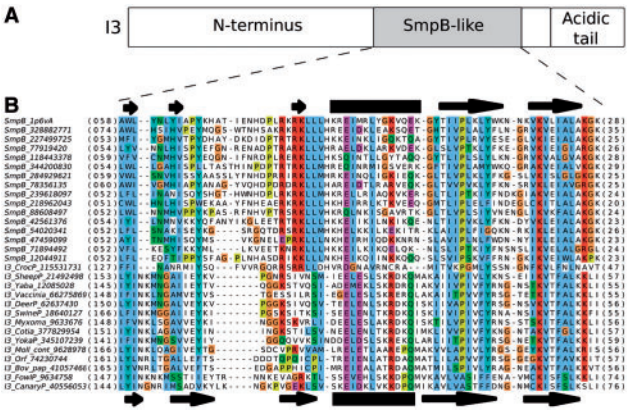

Fig. 3. Similarity between poxviral SSB (I3) and SmpB families. (**A**) Distinct sequence regions of I3. (**B**) Multiple alignment of SmpB and I3 representatives. The structure-derived secondary structure of *A.aeolicus* SmpB (PDB: 1p6v) and the predicted secondary structure of Vaccinia I3 are shown above and below the alignment, respectively

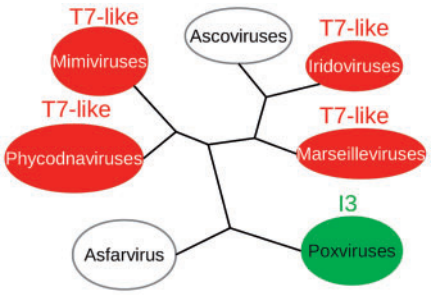

Fig. 4. Distribution of T7-like and I3 SSB families mapped onto the consensus NCLDV evolutionary tree (adopted from Koonin and Yutin, 2010)

duplication of the $\beta\alpha\beta\beta$ structural motif (Supplementary Fig. S3). Thus, the pseudosymmetric SmpB structure, made of two consecutive repeats, is in stark contrast to the repeat-free asymmetric OB-fold.

Interestingly, only the second SmpB repeat is aligned with I3 in its entirety (Fig. 3), suggesting its stronger evolutionary conservation. Consistent with this notion, the aligned regions of both SmpB and I3 harbor a number of functionally important residues (Greseth *et al.*, 2012; Gutmann *et al.*, 2003). The central SmpB-like region of I3 is followed by the C-terminus, enriched in aspartates and glutamates and predicted to feature a significant intrinsic disorder. In this regard, the I3 C-terminus is reminiscent of unstructured acidic tails involved in protein–protein interactions in canonical bacterial and phage SSB proteins.

2.4 Taxonomic distribution of NCLDV SSBs and its evolutionary implications

We mapped the resulting distribution of NCLDV SSB proteins onto the consensus phylogenetic tree of NCLD viruses (Fig. 4). As NCLDVs are thought to be monophyletic (Iyer *et al.*, 2001), the distribution is suggestive of T7-like SSB being already present in the ancestral virus. This is consistent with the idea that bacteriophages played an important role in the early NCLDV evolution (Iyer *et al.*, 2001, 2006; Koonin and Yutin, 2010). Indeed, many of NCLDV genes essential for genome replication (DNA primase-helicase, NAD-dependent ligase and Holliday junction resolvase) show bacteriophage origins (Koonin and Yutin, 2010). Our newly detected T7 gp2.5-like protein family is yet another addition to this list. On the other hand, the unrelated I3 family, confined to poxviruses, is likely the result of a more recent non-orthologous replacement event.

Funding: Howard Hughes Medical Institute (grant number 55005627) and Ministry of Education and Science of Lithuania.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bochkareva,E. *et al.* (2002) Structure of the RPA trimerization core and its role in the multistep DNA-binding mechanism of RPA. *EMBO J.*, **21**, 1855–1863.

Boyer,M. *et al.* (2009) Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc. Natl Acad. Sci. USA*, **106**, 21848–21853.

Dong,G. *et al.* (2002) Structure of small protein B: the protein component of the tmRNA-SmpB system for ribosome rescue. *EMBO J.*, **21**, 1845–1854.

Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.

Frickey,T. and Lupas,A. (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**, 3702–3704.

Gammon,D.B. and Evans,D.H. (2009) The 3′-to-5′ exonuclease activity of vaccinia virus DNA polymerase is essential and plays a role in promoting virus genetic recombination. *J. Virol.*, **83**, 4236–4250.

Greseth,M.D. *et al.* (2012) Molecular genetic and biochemical characterization of the vaccinia virus I3 protein, the replicative single-stranded DNA binding protein. *J. Virol.*, **86**, 6197–6209.

Gutmann,S. *et al.* (2003) Crystal structure of the transfer-RNA domain of transfer-messenger RNA in complex with SmpB. *Nature*, **424**, 699–703.

Hyland,E.M. *et al.* (2003) The DNA binding domain of the gene 2.5 single-stranded DNA-binding protein of bacteriophage T7. *J. Biol. Chem.*, **278**, 7247–7256.

Iyer,L.M. *et al.* (2001) Common origin of four diverse families of large eukaryotic DNA viruses. *J. Virol.*, **75**, 11720–11734.

Iyer,L.M. *et al.* (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res.*, **117**, 156–184.

Kim,Y.T. and Richardson,C.C. (1994) Acidic carboxyl-terminal domain of gene 2.5 protein of bacteriophage T7 is essential for protein-protein interactions. *J. Biol. Chem.*, **269**, 5270–5278.

Koonin,E.V. and Yutin,N. (2010) Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. *Intervirology*, **53**, 284–292.

Legendre,M. *et al.* (2010) mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. *Genome Res.*, **20**, 664–674.

Majji,S. *et al.* (2009) Transcriptome analysis of Frog virus 3, the type species of the genus Ranavirus, family Iridoviridae. *Virology*, **391**, 293–303.

Marintcheva,B. *et al.* (2006) Essential residues in the C terminus of the bacteriophage T7 gene 2.5 single-stranded DNA-binding protein. *J. Biol. Chem.*, **281**, 25831–25840.

Murzin,A.G. (1993) OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *EMBO J.*, **12**, 861–867.

Paytubi,S. *et al.* (2012) Displacement of the canonical single-stranded DNA-binding protein in the Thermoproteales. *Proc. Natl Acad. Sci. USA*, **109**, E398–E405.

Raghunathan,S. *et al.* (1997) Crystal structure of the homo-tetrameric DNA binding domain of Escherichia coli single-stranded DNA-binding protein determined by multiwavelength X-ray diffraction on the selenomethionyl protein at 2.9-A resolution. *Proc. Natl Acad. Sci. USA*, **94**, 6652–6657.

Rochester,S.C. and Traktman,P. (1998) Characterization of the single-stranded DNA binding protein encoded by the vaccinia virus I3 gene. *J. Virol.*, **72**, 2917–2926.

Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Tucker,P.A. *et al.* (1994) Crystal structure of the adenovirus DNA binding protein reveals a hook-on model for cooperative DNA binding. *EMBO J.*, **13**, 2994–3002.

Venclovas,Č. and Margelevičius,M. (2009) The use of automatic tools and human expertise in template-based modeling of CASP8 target proteins. *Proteins*, **77** (**Suppl. 9**), 81–88.

Wiederstein,M. and Sippl,M.J. (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.*, **35**, W407–W410.

Wu,S.L. *et al.* (2009) Mutagenesis identifies the critical regions and amino acid residues of suid herpesvirus 1 DNA-binding protein required for DNA binding and strand invasion. *Virus Res.*, **140**, 147–154.

Yanai-Balser,G.M. *et al.* (2010) Microarray analysis of Paramecium bursaria chlorella virus 1 transcription. *J. Virol.*, **84**, 532–542.

Yutin,N. *et al.* (2009) Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. *Virol. J.*, **6**, 223.