Structural bioinformatics

Predicted function of the vaccinia virus G5R protein

Melissa Da Silva, Ling Shen, Vasily Tcherepanov, Cristalle Watson and Chris Upton*

Department of Biochemistry and Microbiology, University of Victoria, Petch Building 207, Ring Road, Victoria, BC, V8P 5C2, Canada

ABSTRACT

Motivation: Of the ∼200 proteins that have been identified for the vaccinia virus (VACV) genome, many are currently listed as having an unknown function, and seven of these are also found in all other poxvirus genomes that have been sequenced. The G5R protein of VACV is included in this list, and to date, very little is known about this essential and highly conserved protein. Conventional similarity searches of protein databases do not identify significantly similar proteins, and experimental approaches have been unsuccessful at determining protein function.

Results: Using HHsearch, a hidden Markov model (HMM) comparison search tool, the G5R protein was found to be similar to both human and archaeal flap endonucleases (FEN-1) with 96% probability. The G5R protein structure was subsequently successfully modeled using the Robetta protein structure prediction server with an archaeal FEN-1 as the template. The G5R model was then compared to the human FEN-1 crystal structure and was found to be structurally similar to human FEN-1 in both active site residues and DNA substrate binding regions.

Contact: cupton@uvic.ca

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

1 INTRODUCTION

Despite having been sequenced over 15 years ago (Goebel *et al.*, 1990), the vaccinia virus (VACV) strain Copenhagen genome still contains numerous genes that have yet to be assigned a function. Of particular interest are such proteins that are also expressed by all poxviruses, one of which is the 434 amino acid VACV G5R protein. Although little is known about the function G5R plays in the poxvirus life-cycle, it is known to be essential (da Fonseca *et al.*, 2004) and is conserved in all poxvirus species sequenced to date, implying that its role in the poxvirus replication cycle is crucial. Initial characterization of the G5R protein found that G5R is expressed at early stages of poxvirus infection and becomes associated with the viral cores (pellet fraction of solubilized intracellular mature virions), a role that is predominantly associated with proteins expressed at late times of infection although a few early proteins are associated with the viral cores (Banham and Smith, 1992; Doglio *et al.*, 2002). However, two recent studies that aimed to identify each protein associated with the intracellular mature virion using mass spectrometry were unable to detect G5R within the intracellular mature virion (Yoder *et al.*, 2006; Chung *et al.*, 2006). When we employed a hidden Markov model (HMM) comparison search tool (HHsearch) to screen for the functions of the conserved poxvirus proteins that have yet to be assigned a function (Söding *et al.*, 2005; Söding, 2005), G5R produced a significant hit with the class of proteins known as the flap endonucleases (FEN-1) supporting a previous observation (Iyer *et al.*, 2006).

Essential in a variety of organisms including human, yeast, bacterial and archaeal species, the FEN-1 protein plays a key role in both RNA primer removal during DNA replication as well as long-patch base excision repair (Shen *et al.*, 2005). FEN-1 is considered a structure-specific metallonuclease that relies on the binding of two divalent cations for catalysis and acts not only as a 5′-flap endonuclease, but also as a 5′–3′ exonuclease cleaving DNA that contains nicks and gaps (Shen *et al.*, 2005). The crystal structures of human FEN-1 (1UL1) (Sakurai *et al.*, 2005), *Archaeoglobus fulgidus* (1RXW) (Chapados *et al.*, 2004), *Methanococcus jannaschii* (1A77) (Hwang *et al.*, 1998), *Pyrococcus furiosus* (1B43) (Hosfield *et al.*, 1998) and T5 phage (1XO1) (Garforth *et al.*, 1999) have been solved and have aided in the characterization of the amino acids important in FEN-1 activity.

Five regions of the FEN-1 protein are important for activity (Shen *et al.*, 2005). First, the active site that consists of 10 residues (nine of which are charged) binds two divalent cations; one aids in catalysis and the other in protein stabilization. Second, the helical clamp region consists of a helix–loop–helix (HLH) motif that contains mainly positively charged and hydrophobic residues and binds to the 5′-flap of the DNA substrate (Fig. 1). The third region is a hydrophobic wedge that binds the 3′-flap of the DNA substrate. The fourth and fifth regions contain positively charged residues that stabilize the double-stranded portions of the DNA substrate.

This paper focuses on a series of bioinformatics analyses that support the hypothesis that the VACV G5R protein is structurally similar to the FEN-1 family of nucleases.

2 METHODS

The initial prediction of G5R function was carried out using a multiple alignment of 16 G5R protein sequences (Supplementary material, T1 and A1). The alignment was created using the T-coffee alignment software (Notredame *et al.*, 2000) found in Base-By-Base (Brodie *et al.*, 2004), with minor adjustments made to the alignment manually. The alignment was used as input for the HHsearch tool (Söding, 2005; Söding *et al.*, 2005).
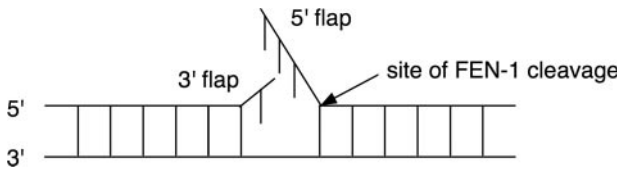
*To whom correspondence should be addressed.

Fig. 1. Optimal double-flap DNA substrate used by human FEN-1.

The VACV G5R protein sequence was used as input in the Robetta protein structure prediction server (Chivian *et al*., 2003; Rohl *et al*., 2004; Chivian *et al*., 2005; Kim *et al*., 2004). Robetta is a fully automated structure prediction server that uses comparative modeling methods to model protein structure if a related structure exists in the PDB database or *ab initio* prediction methods if no structure exists. Robetta successfully created five potential structural models of the VACV G5R protein using the crystal structure of the *A.fulgidus* FEN-1 protein (1RXV) as the template. Protein structures were superimposed using the MatchMaker feature of the Chimera visualization software (Pettersen *et al*., 2004). Electrostatic surface images were created with a PyMOL (http://www.pymol.org) plug-in (Lerner, MG, Carlson, HA. APBS plug-in for PyMOL. 2006) that utilizes the APBS electrostatic surface evaluation tool (Baker *et al*., 2001). Electrostatic potential was visualized using PyMOL with positive potential in blue and negative potential in red.

The comparative alignment between the human FEN-1 (hFEN-1) protein sequence and the VACV G5R and *A.fulgidus* protein sequences was created by performing an initial T-coffee alignment (Notredame *et al*., 2000) in Base-By-Base (Brodie *et al*., 2004) and manually editing the alignment to align active site residues that were conserved on the 3D model. The secondary structure of the proteins was determined using the ksdssp command of the Chimera program (Pettersen *et al*., 2004). This program assigns secondary structure elements based on the information found in the PDB file. Truncated versions of the two protein structures (from position 1–331 in G5R and position 1–332 in hFEN-1 and 1–326 in *A.fulgidus*) were used for both the pairwise alignment and the structure comparisons since the C-terminal domains (∼100 amino acids) of these proteins are not conserved.

3 RESULTS

3.1 HHsearch results

The VACV strain Copenhagen G5R gene is present in all poxviruses sequenced to date; it encodes a 434 amino acid protein. Traditional BLASTp and PSI-BLAST (Altschul *et al*., 1997; Schäffer *et al*., 2001) searches provided no clues as to the function of this protein; all significant hits were to other poxvirus G5R protein sequences. Using a diverse set of poxvirus G5R protein sequences that included each unique G5R sequence in the VOCs database (Ehlers *et al*., 2002), a multiple alignment was created (Supplementary material, A1) and used as input for the HHsearch program. HHsearch creates an HMM from the input alignment and uses it to search HMMs created from other protein databases including the PDB and SCOP databases. Profile-profile search tools, such as HHsearch are thought to be more sensitive since they compare overall profiles of protein sequences rather than comparing a single protein sequence to an alignment of potentially similar proteins (Söding, 2005; Söding *et al*., 2005). HHsearch showed, with a 96% probability, that G5R was similar to both the human (*E*-value: 1.1×10^{-5}) and *A.fulgidus* (archaeal; *E*-value: 3.2×10^{-3}) FEN-1 proteins. This is notable since the human and *A. fulgidus* protein sequences are themselves only 40% identical and the G5R and *A.fulgidus* and human protein sequences are 13 and 16% identical, respectively.
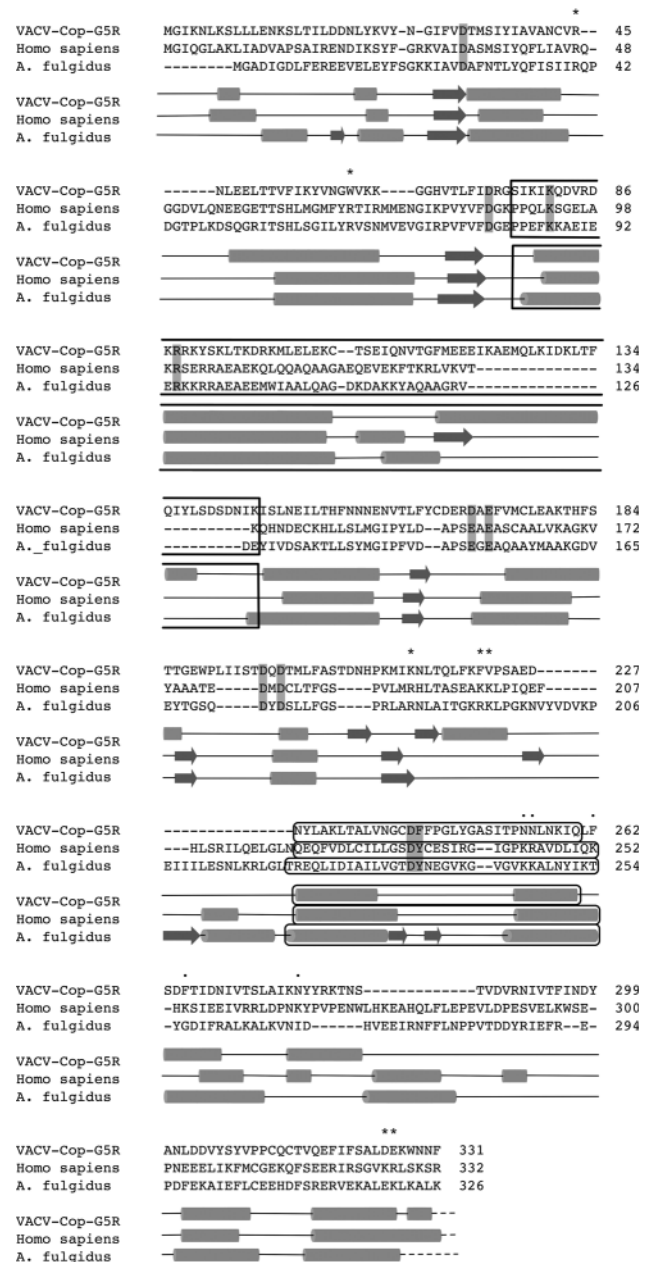

Fig. 2. Multiple alignment between VACV G5R, human FEN-1 and *A.fulgidus* protein sequences. Cylinders represent α-helices and arrows represent β-strands that were extracted by ksdssp except in the helical clamp region of G5R and hFEN-1, which was predicted using Jpred. Gray shaded boxes represent active site residues; a black box represents the helical clamp region and a rounded box represents the H3TH motif. Asterisks and dots represent key residues of the upstream and downstream substrate DNA binding regions, respectively (note: these are not contiguous in the primary amino acid sequence).

A multiple alignment of the VACV G5R, hFEN-1 and *A.fulgidus* proteins (Fig. 2) show conservation of essential components of the hFEN-1 active site. The poxvirus G5R orthologs are themselves very diverse; the VACV G5R protein shares only 17% amino acid identity with the entomopoxvirus proteins (Supplementary material, T2), and 33–47% amino acid identity when compared to lepori, sui,
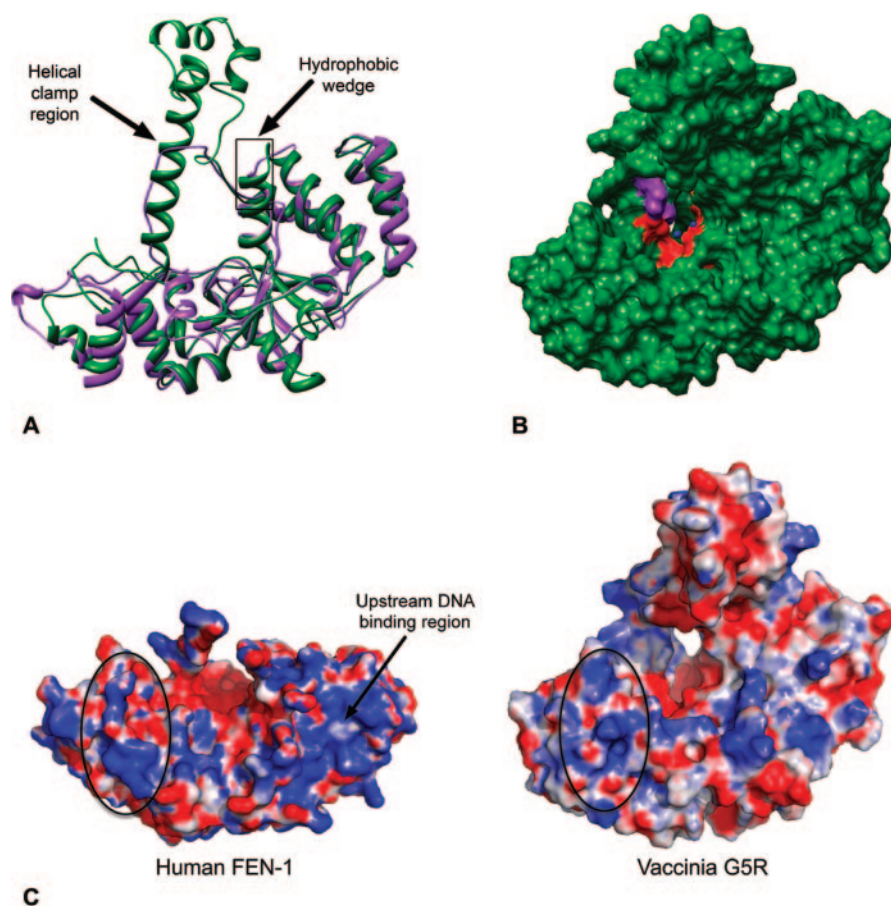
Fig. 3. (**A**) Superimposed structures of human FEN-1 (1UL1, chain X) (purple) and vaccinia G5R (green). (**B**) Surface diagram of VACV G5R with negatively charged active site residues in red, positively charged and hydrophobic amino acid residues in purple and two magnesium ions in blue. (**C**) Surface diagrams of hFEN-1 and VACV G5R colored by electrostatic properties as calculated by APBS. Positively charged regions are colored in blue, negatively charged regions in red. Ovals over both diagrams represent the helix-3 turn-helix motif (downstream DNA binding region) that is conserved in both.

capri, yata or avipoxvirus orthologs. It is therefore significant that the hFEN-1 active site residues are also completely conserved in all the poxvirus proteins suggesting strong functional conservation.

3.2 Structural modeling of the VACV G5R protein

The significant sequence conservation and availability of FEN-I crystal structures prompted us to create a structural model of the VACV G5R protein (Fig. 3). The five models created with Robetta were very similar; model 4 (Supplementary material, M1) was chosen as the representative model since it had the most compact protein structure. Since the *A.fulgidus* and human proteins are distantly related we chose to model the VACV protein on the *A.fulgidus* FEN-1 structure and carry out subsequent analysis with the human structure (1UL1) to permit more rigorous comparisons.

3.2.1 Secondary and tertiary structure comparison In addition to the high conservation of active site residues, the secondary structure of the G5R and hFEN-1 proteins is also well conserved. Of the 18 α-helices found in hFEN-1, all but four align with similar secondary structure in G5R (Fig. 2). Likewise, four of the seven β-strands found in hFEN-1, align very well with G5R. The relatively short length of the β-strands in these proteins may explain this lower

modeling success rate. Due to the inability to observe residues in the helical clamp region in the crystal structure of chain X in hFEN-1 (Sakurai *et al.*, 2005), the secondary structure in this region for both G5R and hFEN-1 was predicted using Jpred (Cuff *et al.*, 1998). Jpred assigns a confidence to its predictions at each amino acid in the protein sequence based on a scale ranging from 0 to 9, where nine is confidently predicted. The secondary structure of this missing region in hFEN-1 as well as the corresponding region in G5R was predicted with a confidence level that ranged from 7 to 9 and consists of two α-helices and a short β-strand at the end of the clamp region for hFEN-1 and just two α-helices for this region in G5R (Fig. 2, black box). Archaeal FEN-1 proteins also have helical secondary structure in this region, further supporting the α-helices predicted for the model.

The tertiary structure of hFEN-1 and G5R is also very well conserved. With the exception of the unmapped helical clamp region of hFEN-1, the positioning of the majority of alpha helices and beta-strands in G5R is conserved and the RMSD of the superposition was found to be 1.4 Å over 251 alpha-carbon pairs (Fig. 3A). For comparison, the RMSD of the superposition of G5R and the *A.fulgidus* structures is 0.81 Å over 281 alpha-carbon pairs and for *A.fulgidus* and hFEN-1 structures is 1.2 Å over 261 alpha-carbon pairs. Thus, each of the structures are relatively equally well conserved.

Conservation of the surface electrostatic properties between VACV G5R and hFEN-1 is shown in Figure 3C; blue areas of hFEN-1 are positively charged and represent regions responsible for contacting substrate DNA (Fig. 3C).

3.2.2 Important residues in the hFEN-1 active site

Ten amino acids are believed to comprise the active site of hFEN-1, seven of which are negatively charged and coordinate the binding of two divalent cations, two that are positively charged and are thought to contribute to the nucleophilic attack that breaks the phosphate backbone of the substrate DNA and one that is hydrophobic and may be involved in hydrogen bonding with other active site residues (Shen *et al.*, 2005). Six of the seven negatively charged residues are fully conserved in G5R (Figs 2 and 3B); glutamate 158 and tyrosine 234 in hFEN-1 are replaced by aspartate and phenylalanine in the corresponding positions of G5R. Both substitutions maintain the biochemical properties of the amino acids found in FEN-1 and therefore are likely to be functional in G5R. The two positively charged residues that are keys for FEN-1 activity are lysine 93 and arginine 100 (Shen *et al.*, 2005). These two residues are absolutely conserved in all poxvirus G5R protein sequences (Supplementary material, A1) and their positions are conserved on the structural model of G5R when compared to the hFEN-1 structure.

3.2.3 Regions important in substrate binding

There are four regions that are important in binding the substrate DNA in hFEN-1. The first is the region important in binding the 5′-flap of substrate DNA, which is located at the helical clamp region and contains the HLH motif that binds ssDNA (Shen *et al.*, 2005). The G5R structural model has four α-helices in this region that when combined into two sets of two helices separated by a 6 amino acid loop, closely resembles an HLH motif (Fig. 3A). When the secondary structure of G5R was predicted for this region using Jpred, the HLH motif was found to be conserved and consists of two α-helices separated by a 9 amino acid loop (Fig. 2). The helical clamp region of hFEN-1 consists of bulky and positively charged amino acid residues on the inner side of each of the helices of the HLH motif that are responsible for contacting the 5′-flap DNA (Shen *et al.*, 2005). Of the 70 amino acids comprising the helical clamp region in G5R, 20 of them (28.6%) are positively charged or hydrophobic compared to hFEN-1, which has 12 of 47 amino acids (25.5%) being positively charged or hydrophobic in the helical clamp region.

The binding of the 3′-flap of the substrate DNA occurs at the hydrophobic wedge in hFEN-1 (Shen *et al.*, 2005). This region begins at position 42 of hFEN-1 and consists of the sequence: FLIAV (Fig. 2). The corresponding region in G5R, although not identical to that of hFEN-1 begins at position 41, on both the structural model (Fig. 3A) and the alignment (Fig. 2), and consists of the sequence: VANCV. This sequence in G5R is primarily hydrophobic suggesting that it may also be capable of binding to a 3′-flap DNA region.

The remaining two regions are responsible for binding the upstream and downstream double-stranded substrate DNA. Each of these regions contains key positively charged residues that are responsible for making contact with the substrate DNA. The upstream DNA binding region includes the following residues of hFEN-1: R47, R70, R192, K200, K201, K326 and R327. Of these seven residues, one (R47) is conserved in the G5R protein (R46 in

G5R). Although the remaining residues may match similar amino acids in the sequence alignment (Fig. 2) the structure appears to be poorly conserved in this region; this, however, may be the result of poor modeling due to poor sequence conservation or real structural differences. The electrostatic surface diagram of the corresponding upstream DNA binding region of G5R (Fig. 3C) shows some positively charged residues that may contribute to DNA binding by virtue of their position in the structure despite the apparent lack of conservation at the level of primary sequence that is seen in Fig. 2.

Analysis of the downstream DNA substrate binding regions of the two structures revealed some regions of conservation and some differences between the proteins. This region on hFEN-1 consists of five important positively charged amino acids as well as a Helix-3–Turn–Helix (H3TH) motif. The five residues that are involved in hFEN-1 downstream DNA binding are: K244, R245, K252, K254 and K267. These residues appear to correspond to N254, N255, I259, L261 and V271, respectively, on the G5R model suggesting a loss of positive charge. However, the H3TH region of hFEN-1 is highly conserved in the G5R structure (Fig. 2, rounded boxes; Fig. 3C). On hFEN-1, the first helix of the H3TH motif is located at Q220 and ends at G231, which corresponds to the G5R helix that begins at N228 and ends at V237. The 3T region of hFEN-1 is slightly shorter than the G5R 3T region beginning at S232 and ending at G242 and corresponds to positions N238 through T252 in G5R, respectively. The second helix of G5R is slightly shorter than that seen for hFEN-1 with the hFEN-1 helix beginning at P243 and ending at K252 and corresponding to P253–Q260, respectively for G5R.

3.2.4 Other features that may play a role in G5R activity

There are two other features that have been observed in hFEN-1 that may play a role in G5R activity. The first is the post-translational phosphorylation of S187 in hFEN-1, which has been shown to play a role in the regulation of exo/endonuclease activity in hFEN-1 (Shen *et al.*, 2005). The corresponding residue in G5R is S204 and is conserved in both the alignment and the structural model.

Finally, we note that not only are the N-terminal sequences, ending at position 10, of the G5R and FEN-1 proteins very highly conserved among their respective groups, but also that there is significant conservation between the two sets of proteins (Fig. 2). Although the exact function of this short sequence is not known, the crystal structure of hFEN-1 revealed that it is in close proximity to the active site and H3TH residues and thus may stabilize interactions with the substrate (Sakurai *et al.*, 2005).

3.3 Functional considerations

Given the multiple lines of evidence presented here, we conclude that the poxvirus G5R proteins are likely to share significant structural similarity to the FEN-1 family. Since related proteins are also encoded by herpes (Oroskar and Read, 1989; Taddeo *et al.*, 2002) and iridoviruses (Iyer *et al.*, 2006) these may represent very ancient viral genes that have evolved from a FEN-1-like ancestor. However, it is not clear whether endonuclease activity has been preserved, and given the wide range of activities described for FEN-1 proteins, including its participation in the base excision repair pathway, genome replication and probable participation in apoptosis (Shen *et al.*, 2005), it may be difficult to ascertain the natural substrates for the viral proteins.

Since the traditional role that hFEN-1 plays in RNA primer removal during DNA replication is thought not to be a part of the poxvirus DNA replication cycle, the G5R protein may have evolved novel functions in poxviruses in a manner similar to the uracil DNA glycosylase (De Silva and Moss, 2003). It is likely that the C-terminal domain, which appears to be unique to the poxvirus proteins, plays a role in its specific function.

ACKNOWLEDGEMENTS

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.

Baker,N.A. *et al.* (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA*, **98**, 10037–10041.

Banham,A.H. and Smith,G.L. (1992) Vaccinia virus gene-b1r encodes a 34-kda serine threonine protein kinase that localizes in cytoplasmic factories and is packaged into virions. *Virol.*, **191**, 803–812.

Brodie,R. *et al.* (2004) Base-By-Base: single nucleotide-level analysis of whole viral genome alignments. *BMC Bioinformatics*, **5**, 96.

Chapados,B.R. *et al.* (2004) Structural basis for FEN-1 substrate specificity and PCNA-mediated activation in DNA replication and repair. *Cell*, **116**, 39–50.

Chivian,D. *et al.* (2003) Automated prediction of CASP-5 structures using the Robetta server. *Proteins*, **53**, 524–533.

Chivian,D. *et al.* (2005) Prediction of CASP6 structures using automated Robetta protocols. *Proteins*, **61**, 57–166.

Chung,C.S. *et al.* (2006) Vaccinia virus proteome: identification of proteins in vaccinia virus intracellular mature virion particles. *J. Virol.*, **80**, 2127–2140.

Cuff,J.A. *et al.* (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892–893.

da Fonseca,F.G. *et al.* (2004) Vaccinia virus mutants with alanine substitutions in the conserved G5R gene fail to initiate morphogenesis at the nonpermissive temperature. *J. Virol.*, **78**, 10238–10248.

De Silva,F.S. and Moss,B. (2003) Vaccinia virus uracil DNA glycosylase has an essential role in DNA synthesis that is independent of its glycosylase activity: catalytic site mutations reduce virulence but not virus replication in cultured cells. *J. Virol.*, **77**, 159–166.

Doglio,L. *et al.* (2002) The Vaccinia virus E8R gene product: a viral membrane protein that is made early in infection and packaged into the virions' core. *J. Virol.*, **76**, 9773–9786.

Ehlers,A. *et al.* (2002) Poxvirus Orthologous Clusters (POCS). *Bioinformatics*, **18**, 1544–1545.

Garforth,S.J. *et al.* (1999) Mutagenesis of conserved lysine residues in bacteriophage T5 5′-3′ exonuclease suggests separate mechanisms of endo-and exonucleolytic cleavage. *Proc. Natl Acad. Sci. USA*, **96**, 38–43.

Goebel,S.J. *et al.* (1990) The complete DNA sequence of vaccinia virus. *Virology*, **179**, 247–266.

Hosfield,D.J. *et al.* (1998) Structure of the DNA repair and replication endonuclease and exonuclease FEN-1: coupling DNA and PCNA binding to FEN-1 activity. *Cell*, **95**, 135–146.

Hwang,K.Y. *et al.* (1998) The crystal structure of flap endonuclease-1 from *Methanococcus jannaschii*. *Nat. Struct. Biol.*, **5**, 707–713.

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

Kim,D.E. *et al.* (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.*, **32**, W526–W531.

Notredame,C. *et al.* (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

Oroskar,A.A. and Read,G.S. (1989) Control of mRNA stability by the virion host shutoff function of herpes simplex virus. *J. Virol.*, **63**, 1897–1906.

Pettersen,E.F. *et al.* (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.

Rohl,C.A. *et al.* (2004) Modeling structurally variable regions in homologous proteins with rosetta. *Proteins*, **55**, 656–677.

Sakurai,S. *et al.* (2005) Structural basis for recruitment of human flap endonuclease 1 to PCNA. *EMBO J.*, **24**, 683–693.

Schäffer,A.A. *et al.* (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.

Shen,B. *et al.* (2005) Multiple but dissectible functions of FEN-1 nucleases in nucleic acid processing, genome stability and diseases. *Bioessays*, **27**, 717–729.

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

Söding,J. *et al.* (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244–W248.

Taddeo,B. *et al.* (2002) The patterns of accumulation of cellular RNAs in cells infected with a wild-type and a mutant herpes simplex virus 1 lacking the virion host shutoff gene. *Proc. Natl Acad. Sci. USA*, **99**, 17031–17036.

Yoder,J.D. *et al.* (2006) Pox proteomics: mass spectrometry analysis and identification of Vaccinia virion proteins. *Virol J.*, **3**, 10.