Acta Biochimica Polonica

Vol. 54 No. 3/2007, 551–559

on-line at: www.actabp.pl

Regular paper

Fold recognition insights into function of herpes ICP4 protein

Lucjan S. Wyrwicz[1,2]✉ and Leszek Rychlewski[1]

[1]*BioInfoBank Institute, Poznań, Poland;* [2]*Department of Gastroenterology, Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Warszawa, Poland*

Received: 23 May, 2007; revised: 03 September, 2007; accepted: 10 September, 2007
available on-line: 17 September, 2007

ICP4 is an important factor regulating the life cycle of HSV1. This conserved protein has several molecular functions, including activation of expression of viral late gene transcripts and inhibition of immediate early genes. Although ICP4 and its *Alphaherpesvirinae* homologs (eg.: IE62 of VZV) have been subjects of various molecular studies, a complete view of their molecular function is lacking. Here we present the results of fold recognition and molecular modelling of ICP4 functional domains. The performed state-of-the-art bioinformatic fold recognition analysis identified a dual helix-turn-helix motif as a binding module of repressor activities (so called region 2 domain). The mapping of distant homology identified that a segment responsible for activation of late gene promoters (region 4) exhibits folding of uracil DNA glycosylase (UDG), but seems to be a non-functional homolog of UDG. Potential implications of the results are discussed.

Keywords: Herpesviridae, ICP4 transcription regulator, HSV1, HSV2, VZV, bioinformatics

INTRODUCTION

During productive infection by herpes simplex virus type 1 (HSV1), nearly 80 genes are transcribed by DNA-dependent RNA polymerase II in three phases named immediate early (IE), early (E), and late (L). ICP4 (IE175) is the major regulatory protein of HSV1 and is one of IE genes expressed at the earliest stages of virus infection (Wagner *et al.*, 1995). The gene is crucial for infection since its inactivation interrupts the progress beyond IE phase. Its product is required for the activation of transcription from the majority of viral promoters (Watson *et al.*, 1980), but the knowledge on the mechanism of its action is still incomplete. The central role of ICP4 in the progression of the viral life cycle is augmented by the fact that ICP4 protein represses transcription of three viral genes (early regulator, ICP4; latency phase product, LAT; and ORF-P) *via* interaction with high affinity binding sites at their transcription initiation sites (Faber *et al.*, 1988; Batchelor *et al.*, 1994; Gu *et al.*, 1995).

The gene with open reading frame of nearly 1300 amino acids is present in all genomes of *Alphaherpesviridae,* including also two other human pathogens: herpes simplex virus type 2 (HSV2) and Varicella-Zoster Virus (VZV). The native protein forms homodimers (Metzler *et al.*, 1985) and this interaction has a fundamental role in ICP4 functionality. ICP4 protein has several molecular functions, like DNA binding associated with repression of gene expression, nuclear localization and activation of expression of late transcripts. The activities previously mapped on the ICP4 open reading frame are distributed throughout it (DeLuca & Schaffer, 1988).

The DNA binding activity is associated with the N-terminal part of the protein (Beard *et al.*, 1986; DeLuca *et al.*, 1988; Wy *et al.*, 1990). The bipartite consensus binding site of the so called "A site" is determined as ATCGTCnnnnYCGRC, where n is any nucleotide, Y is a pyrimidine (cytosine or thymine), and R is a purine (adenine or guanine) (Faber & Wilcox, 1986). Additionally, unrelated motifs with no obvious common sequence similarity pattern

✉**Corresponding author**: Lucjan S. Wyrwicz, Department of Gastroenterology, Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, W. Roentgena 5, 02-781 Warszawa, Poland; tel: (48 22) 546 2933; Fax: (48 22) 644 0209; email: lucjan@bioinfo.pl

Abbreviations: FR, fold recognition; IE, immediate-early genes; HSV1, herpes simplex virus type 1; HSV2, herpes simplex virus type 2; HTH, helix-turn-helix; PAX, paired box domain; VZV, Varicella-Zoster virus.

have been observed (B sites) (Faber & Wilcox, 1986). The binding specificity of ICP4 and its homologs from various *Alphaherpesvirinae* is similar (Wu & Wilcox, 1991) and multiple binding sites are described throughout the genome (Michael & Roizman, 1989). The protein–DNA interaction is augmented by the proper spatial distribution of the 'A site' and general transcription factor binding sites (DiDonato & Muller, 1989). A bend of DNA within A sites is observed during the formation of protein–DNA complex (Everett *et al.*, 1992).

The regulation of expression of late genes is associated with the C-terminal segment of ICP4. Earlier studies suggested an involvement of protein–protein interactions between ICP4 and cellular factors of the polymerase complex. Analysis of ICP4-related activation of expression of viral genes indicates that ICP4 performs its action in cooperation of the host proteins of the preinitiation complex, like TATA-binding protein (TBP), $TAF_{II}250$ and TFIIB (Smith *et al.*, 1993; Carrozza & DiLuca, 1996).

Since the protein is essential for all *Alphaherpesviridae*, and its function is preserved, we may assume that the most critical activities are performed by the protein fragments exhibiting the highest sequence similarity throughout the protein family. McGeoch *et al.* (1986) pointed out five different functional regions of ICP4 based on patterns of conservation. In general, two conserved fragments called region 2 and 4 are mapped on independent functional regions responsible for DNA binding at 'A sites' and associated with the activation of late genes, respectively. The internal fragment of nearly 200 amino acids (region 3) located in the center of ICP4 is much more divergent. Apart from carrying a nuclear localization signal it may function as a spacer of the functional domains of different activity (McGeoch *et al.*, 1986). The N-terminal fragment (region 1) is the least conserved one. It contains multiple phosphorylation sites (Xia *et al.*, 1996) and a protein-protein interaction interface (Grondin & DeLuca, 2000). The role of region 5 localized in the C-terminal fragment is associated with late regulatory activities. The removal of its terminal 56 residues, as well as introduction of some point mutations in this region abolished its activity. It has also been suggested that region 5 acts as an enhancer of ICP4 N-terminal transactivation domain (Bruce *et al.*, 2002).

In our previous reports, we applied methods for identifying distant homology of protein families with subsequent protein structure molecular modelling to analyse the functions of divergent proteins of *Herpesvirinae*. This approach successfully identified the herpes UL24 gene product as a potential endonuclease (Knizewski *et al.*, 2006). Similar methodology allowed the identification of glycoprotein L (gL) as a protein resembling chemokine receptor ligands (Wyrwicz & Rychlewski, 2007a) and EBV BcRF1 as a late regulatory TATT-binding protein (Wyrwicz & Rychlewski, 2007b). Here we present the results of fold recognition and molecular modeling of the functional domains of the ICP4 protein.

MATERIALS AND METHODS

Fold recognition and assembly of structural alignments. Sequences of ICP4 (HSV1, HSV2) and IE62 were retrieved from the corresponding genome sequences (GenBank) and aligned using ClustalW (Thompson *et al.*, 1994) with subsequent manual corrections. The annotation of globular regions was performed using GlobPlot (Linding *et al.*, 2003). The protein sequences were divided into 300 amino acid long overlapping fragments and submitted to the Structure Prediction Meta Server (http://bioinfo.pl/meta) (Bujnicki *et al.*, 2001), which assemblies various secondary structure prediction and top-of-the-line fold recognition (FR) methods. Regions with a high propensity to create non-globular regions in GlobPlot (Linding *et al.*, 2003) and segments without consistent and confident predictions of secondary structure using PsiPred (Jones, 1999) and ProfSec (Rost *et al.*, 2004) were marked as non-globular (Fig. 1).

Since the protein structure prediction methods had been optimized for processing of globular domains, the potential globular regions were divided further into single domains according to secondary structure predictions and preliminary results of fold recognition searches. The domains with corrected boundaries were resubmitted again to the Structure Prediction Meta Server (Bujnicki *et al.*, 2001) and MetaBasic (Ginalski *et al.*, 2004). Collected models were screened with 3D-Jury, a consensus fold recognition prediction method (Ginalski *et al.*, 2003).

The protein structure prediction methods were used to identify similarity between ICP4 domains and known protein families. For both target and template sequences, close homologs were collected with PSI-BLAST and aligned by using ClustalW (Thompson *et al.*, 1994) and PCMA (Pei *et al.*, 2003) with final manual adjustments according to secondary structures (observed and predicted) and critical residues of the fold identified by literature browsing.

Identification and distant fold mapping of DNA-binding domain. Since the initial procedure of fold recognition failed to provide a confident assignment for the repressor domain (region 2), an additional algorithm was applied. All corresponding sequences of the DNA binding domain were extracted from ICP4 homologs present in the "nr" database (GenBank) clustered at 90% sequence identity (Li &

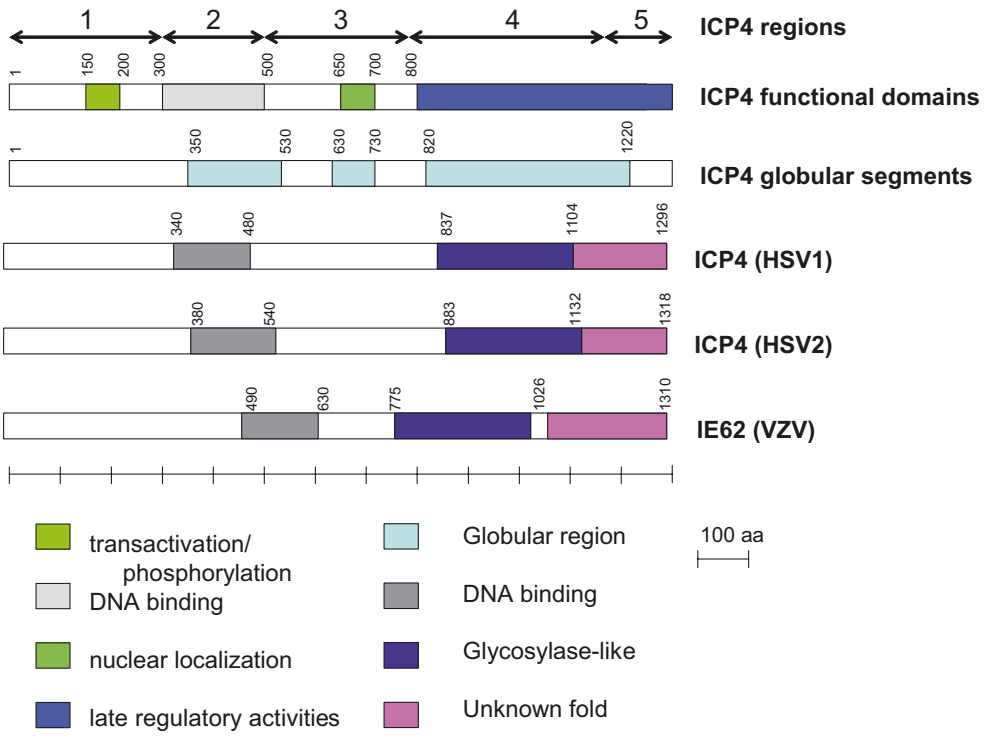

Figure 1. An overview of ICP4 organization.
ICP4 regions by McGeoch *et al.* (1986) and the functional domains are marked consistently with previous functional reports (cited in the text). Potential globular regions of HSV1 ICP4 are annotated according to GlobPlot (http://globplot.embl.de). Numbers refer to boundaries of domains and functional regions in HSV1, HSV2 and VZV (GenBank entries: gi|9629441, gi|9629330 and gi|9625936, respectively).

Godzik, 2006). The sequences were submitted again to the Protein Structure Prediction Meta Server (Bujnicki *et al.*, 2001). The results of the template selection procedure were screened for proteins involved in regulation of transcription according to functional assignments provided by the GeneOntology consortium (Gene Ontology terms: GO:0003700 — transcription factor activity, GO:0045449 — regulation of transcription, GO:0006355 — regulation of transcription, DNA-dependent, GO:0003677 — DNA binding (Harris *et al.*, 2004)). Since consistent predictions (DNA/RNA-binding three-helical bundle fold superfamily, SCOP: a.4 (Wintjens & Rooman, 1996; Aravind *et al.*, 2005)) were observed as results of several fold recognition methods (INUB (http://inub.cse.buffalo.edu/), mGenThreader (McGuffin & Jones, 2003), further analysis was performed on the proteins from this superfamily as described in the Results section. The query sequences were aligned (ClustalW; Thompson *et al.*, 1994) to template sequences from the three helical bundle superfamily with subsequent manual correction of the alignment according to secondary structure (predicted and observed). The identification of homology was confirmed by enumeration of critical residues creating the potential protein–DNA interaction interface.

RESULTS

Fold recognition for ICP4 regions

The initial screening for potential globular regions suggested that regions 1 and 3 represent a likely non-globular domains and therefore are not suitable for application of fold recognition methodology. The sequence analysis mapped the potential folding to domains from regions 2 and 4 of *Alphaherpesvirinae* ICP4 (McGeoch *et al.*, 1986). The domains and their location in the ICP4 sequences are shown in Fig. 1.

Fold recognition of repressor DNA-binding domain

The identified hits to proteins related to transcription, gene regulation and DNA binding are listed in Table 1. Among the results, we identified 102 hits to proteins associated with regulation of transcription and 83 of them represented the DNA/RNA-binding three-helical bundle fold (SCOP superfamily: a.4). Among the remaining 19 hits none of the identified folds was reported for either of the tested ICP4 homologs; the second top scoring hit — (antitermination factor NusB — SCOP a.74

Table 1. Summary of hits for the DNA-binding domain of region 2 collected with the Protein Structure Prediction MetaServer (http://bioinfo.pl/meta).

Only hits to proteins associated with transcription factor activity, regulation of transcription, and DNA binding activities are shown (SCOP folds: a.4, DNA/RNA-binding 3-helical bundle; a.22, histone-fold; a.60, SAM domain-like; a.74, cyclin-like; a.79, antitermination factor NusB; a.123, nuclear receptor ligand-binding domain; a.143, RPB6/omega subunit-like; c.47, thioredoxin fold; c.94, periplasmic binding protein-like II; d.180, conserved core of transcriptional regulatory protein vp16).

Sequence	SCOP identifier	Number of hits
HSV1 (gi|9629441)	**a.4**	**5**
	a.79	1
HSV2 (gi|9629330)	**a.4**	**5**
	a.143	1
	a.79	1
PRV (gi|124178)	**a.4**	**4**
VZV (gi|9625936)	**a.4**	**8**
	a.22	1
CHV (gi|3551150)	**a.4**	**6**
	a.79	1
BHV5 (gi|40787932)	**a.4**	**5**
	a.79	1
FHV1 (gi|493242)	**a.4**	**4**
	c.94	1
	c.47	1
EHV1 (gi|124143)	**a.4**	**5**
	a.79	1
	a.60	1
EHV4 (gi|9629793)	**a.4**	**5**
	a.79	1
	a.60	1
MaHV2 (gi|16117372)	**a.4**	**5**
CHV7 (gi|13242468)	**a.4**	**7**
CHV1 (gi|30984487)	**a.4**	**8**
	a.79	1
	a.123	1
MeHV1 (gi|12084909)	**a.4**	**4**
	a.74	1
GHV3 (gi|10800036)	**a.4**	**3**
	d.180	1
GHV2 (gi|10180785)	**a.4**	**5**
	c.94	3
	d.180	1

superfamily) — was identified seven times and the remaining superfamilies were represented by nearly single hits.

The SCOP a.4 superfamily utilizes a major structural motif capable of binding DNA – so called helix-turn-helix motif (HTH). It is composed of two α helices joined by a short strand of amino acids and is found in many proteins that regulate gene expression. The C-terminal helix is involved in DNA binding *via* sequence-specific interaction with the major groove of double-stranded DNA (Wintjens & Rooman, 1996; Aravind *et al.*, 2005). Due to very low degree of sequence similarity no HTH protein of known structure was preferably identified as a likely homolog of the ICP4 domain. Further selection of modeling templates among the three-helical bundle superfamily was supported by the following observations:

1) the ICP4 DNA binding site consists of a bipartite motif of 11 nucleotides spaced by additional four bases (ATCGTCnnnnYCGRC);

2) the potential globular segment of ICP4 DNA-binding region (so called region 2) has six α helices (as predicted by PsiPred; Jones, 1999) and ProfSec (Rost *et al.*, 2004) accessed *via* the Protein Structure Prediction MetaServer (Bujnicki *et al.*, 2001);

3) the least conserved segment (potential inter-domain linker) divides the secondary structure elements into two blocks of three helices;

4) the third and sixth helices (last helix in each 3-helix segment) contain conserved basic residues (arginine, lysine, histidine) potentially involved in interaction with DNA elements.

Based on the above-mentioned facts we concluded that for the purposes of the modelling study we should select a protein of known structure containing two HTH domains involved in cooperative binding of DNA. We arbitrarily selected a paired box domain (PAX) (Underhill, 2000) as a domain containing two HTH domains independently interacting with closely spaced DNA motifs (Xu *et al.*, 1999). PAX proteins represent a distinct conserved class of *Eukaryota* and are critical for gene expression in the development (Lang *et al.*, 2007). PAXs contain three functional regions involved in the protein–DNA interaction: two helix-turn-helix motifs utilizing a DNA-binding module typical for this fold divided by a linker region which acts as an additional component of the DNA-binding interface (Lang *et al.*, 2007).

The structural templates of PAX domains were collected from the PDB database (6paxA, 1pdnC), additionally sequences of human PAX proteins were aligned by using ClustalW (Thomposon *et al.*, 1994). A structural alignment of ICP4 and PAX families was created manually according to the secondary structure (PsiPred predictions of ICP4 and human PAX proteins and observed structure of PAX6 from PDB entry 6paxA). The alignment of the ICP4 domain and PAX protein family is shown in Fig. 2. The basic amino acids involved in binding to phosphate moieties of double-stranded DNA were identified by inspection of the PAX6 structure and marked on Fig. 2 (1–9, A–E).

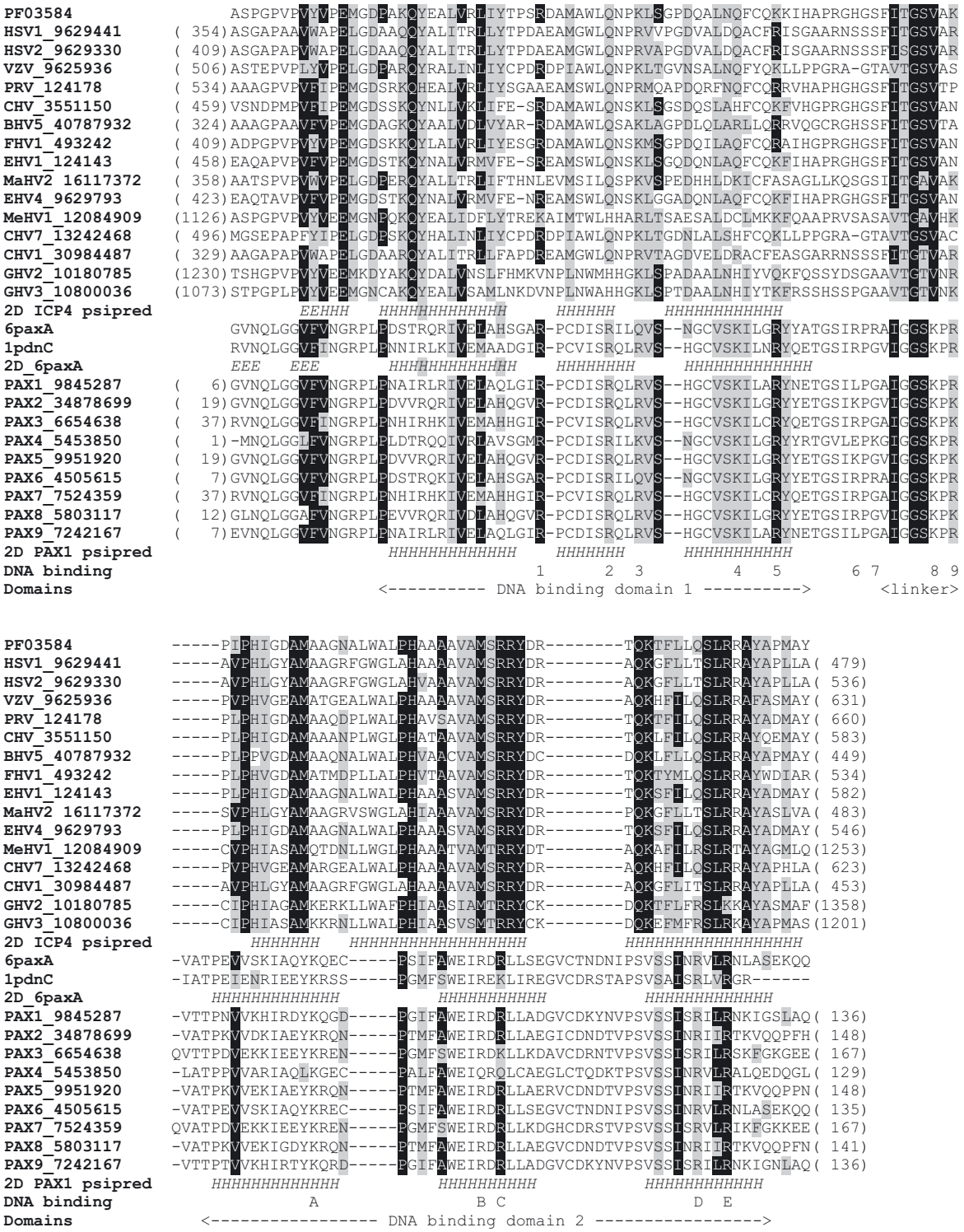

Figure 2. Alignment of ICP4 DNA-binding domain and human paired box DNA binding domains (PAX).
The corresponding sequences were extracted from the GenBank entries (coded with GenBank identifier gi; PRV, *Pseudorabies virus*; CHV, *Canine herpesvirus*; BHV5, *Bovine herpesvirus 5*; FHV, *Feline herpesvirus type 1*; EHV1, *Equine herpesvirus type 1*; MaHV1, *Macropodid herpesvirus 1*; EHV4, *Equine herpesvirus 4*; MeHV1, *Meleagrid herpesvirus 1*; CHV7, *Cercopithecine herpesvirus 7*; CHV1, *Cercopithecine herpesvirus 1*; GHV2, *Gallid herpesvirus 2*; GHV3, *Gallid herpesvirus 3*). Numbers in brackets refer to positions in the GenBank entries. The sequences of crystallographically solved proteins from PAX family, PAX6 and PAX5, are shown (6paxA, 1pdnC, respectively). The observed (PDB: 6paxA) and predicted (psipred predictions for VZV ICP4 and PAX1) secondary structure elements are coded with letters: H, α-helix, E, β-strand (extended). The distinct functional regions identified previously for PAX6 are marked below (two helix-turn-helix DNA binding domains and a linker binding to minor groove). Basic residues involved in interaction with DNA in PAX6 (PDB: 6paxA) are marked below (1–9, A–E).

Table 2. Summary of fold recognition analysis for VZV ICP4 (gi|9625936:744–1033; region 4).

The top scoring hits for each method are shown in bold font. Hits below the 3D-Jury cutoff of 50.0 (corresponding to less than 5% of prediction error) are shown in italics.

3DJury score	Method	#Selected Hit	Methods score	PDB identifier	Hit name DNA	SCOP family	Organism
96.38	**BasD**	**1**	**7.01**	**1lauE**	**Uracil-DNA glycosylase**	**c.18.1**	***HSV1***
93.62	BasD	5	5.71	1okbA	Uracil-DNA glycosylase	c.18.1	*G. morhua*
93.38	BasD	4	6.2	1akz_	Uracil-DNA glycosylase	c.18.1	*H. sapiens*
88.88	FFA3	5	–5.5	2booA	Uracil-DNA glycosylase	c.18.1	*D. radiodurans*
86.50	BasD	3	6.38	1eugA	Uracil-DNA glycosylase	c.18.1	*E. coli*
81.50	INUB	2	19.54	1akz_	Uracil-DNA glycosylase	c.18.1	*H. sapiens*
80.12	**3DPS**	**1**	**0.0128**	**1akz_**	**Uracil-DNA glycosylase**	**c.18.1**	***H. sapiens***
80.00	**FFA3**	**1**	**–6.3**	**1lauE**	**Uracil-DNA glycosylase**	**c.18.1**	***HSV1***
78.75	BasD	2	6.94	2booA	Uracil-DNA glycosylase	c.18.1	*D. radiodurans*
76.38	**INUB**	**1**	**25.5**	**1lauE**	**Uracil-DNA glycosylase**	**c.18.1**	***HSV1***
75.50	FFA3	3	–5.9	1eugA	Uracil-DNA glycosylase	c.18.1	*E. coli*
75.38	3DPS	3	0.0309	3eugA	Uracil-DNA glycosylase	c.18.1	*E. coli*
74.50	ORF2	8	12.12	1lauE	Uracil-DNA glycosylase	c.18.1	*HSV1*
70.75	FFA3	10	–5.0	1akz_	Uracil-DNA glycosylase	c.18.1	*H. sapiens*
63.50	3DPS	2	0.0172	1lauE	Uracil-DNA glycosylase	c.18.1	*HSV1*
61.50	ORF2	7	12.18	2booA	Uracil-DNA glycosylase	c.18.1	*D. radiodurans*
47.50	***PDBb***	**1**	***0.29***	***1lauE***	***Uracil-DNA glycosylase***	***c.18.1***	***HSV1***
29.62	*ORF2*	4	12.78	*1oatA*	*Ornithine aminotransferase*	*c.67.1*	*H. sapiens*
27.62	*ORF2*	3	*13.09*	*1z7dA*	*Ornithine aminotransferase*	*c.67.1*	*P. yoelli*
18.75	*FFA3*	6	*–5.5*	*1g9mC*	*CD4 receptor*	*b.1.1*	*H. sapiens*

Fold recognition of glycosylase-like domain

The fold recognition of region 4 was performed according to the standard procedure of fold recognition as described previously (von Grotthuss *et al.*, 2003). Among the hits collected by the Protein Structure Prediction Meta Server (Bujnicki, 2001), the uracil DNA glycosylase fold was identified by several homology modelling and threading methods (Table 2). Uracil-DNA glycosylase is a DNA repair enzyme catalyzing the reaction of cleavage of the RNA-specific base (uracil) from DNA (reviewed by Krwawicz *et al.*, 2007). As summarized in Table 3 this prediction was consistently assigned by the 3D-Jury prediction assessment system for several proteins of *Alphaherpesvirinae* ICP4 (HSV1, HSV2 – ICP4, IE62 – VZV). The structural alignment of this domain is shown in Fig. 3.

Table 3. Summary of fold recognition for ICP4 late regulator domain (region 4) according to the 3D-Jury prediction assessment method

Organism	Sequence details	3d-Jury score (PDB code; FR method)
HSV1	gi|9629441:806-1105	58.50 (1akz_; Orfeus)
HSV2	gi|9629330:868-1141	87.12 (1okbA; MetaBasic)
VZV	gi|9625936:744-1033	96.38 (1lauE; MetaBasic)

In order to test whether the ICP4 region 4 domain apart from assuming the uracil DNA glycosylase fold also retains its function an analysis of residues creating the active site was performed. The conserved residues concentrate in the internal part of the domain, while there is a weak sequence conservation on the surface and residues creating the active site are not preserved in ICP4.

DISCUSSION

The applied methodology provides an additional view into the molecular function of the ICP4 protein — a main regulator of early/late gene expression in *Alphaherpesviridae*. Previous functional studies mapped the observed activities to defined subregions of ICP4 (Fig. 1). Here we show that such regions represent distinct structural domains which can be characterized with the bioinformatic methodology of fold recognition and homology modelling.

The DNA-binding domain (region 2) — responsible for repression of early genes of HSV1 — encodes a common DNA/RNA binding fold of the three helical bundle superfamily. The consistent presence of six αhelices and preserved pattern of basic

Figure 3. Structural alignment of ICP4 glycosylase-like domain.
The corresponding *Herpesviridae* regions were extracted from the GenBank entries: HSV1, gi|9629441; HSV2, gi|9629330 and VZV, gi|9625936; numbers in brackets refer to positions in the sequences and the length of removed non-conserved fragments. The sequences of crystallographically solved uracil-DNA glycosylases from HSV1, human and Atlantic cod are shown (PDB entries: 1lauE, 1akz_, 1okbA, respectively). The observed (below; 1lauE) and predicted (above; PsiPred, ProfSec predictions for ICP4 of HSV2) secondary structure elements are coded with letter codes (compare Fig. 2).

amino acids potentially involved in binding of DNA phosphate moieties strongly supports the distant mapping of the fold. For the modelling purposes we selected a similar tandem domain from the helix-turn-helix family (PAX domain). Out of 14 basic residues located on the potential interface of the ICP4 DNA-binding domain, seven are preserved between PAX and ICP4 family (positions 1, 3, 5 in the first HTH domain, 9 in the intra-domain linker and C, D, E in the second HTH domain, compare Fig. 2). Since the conservation pattern is not accurate, we may assume some variation of the binding mode among the species of *Alphaherpesvirinae.* The highest degree of conservation was observed for the residues critical for the interaction of HTH with DNA in the PAX domain located at positions 3 and 5 in the first HTH domain, as well as C and E in the second domain.

The fold recognition methodology in region 4 suggested that the globular domain located between residues 837 and 1104 of HSV1 ICP4 (Fig. 3) encodes a uracil-DNA glycosylase-like fold. With the consistent predictions of various distant similarity and threading methods (Table 2) the mapping of the fold is highly likely to be correct. Also the match in the pattern of secondary structure (predicted for ICP4 and observed for UDG) strongly supports the confident scores of the fold recognition protocols used. Since the residues creating the active site are not preserved (not shown), we conclude that this protein is unable to perform the enzymatic function of uracil-DNA glycosylase. Although retaining of a fold without preserving its enzymatic function is an important mechanism of evolution of protein families, such situations are observed relatively infrequently. We may expect that ICP4 and its homologs utilize the uracil-DNA glycosylase fold to interact with DNA of herpes gene promoters and either specifically recognize modified nucleotides (e.g.: methylated cytosines) or perform a chemical modification of promoter DNA. Further experimental work is needed to answer the above question and to investigate the possibility of utilization of potential inhibitors of this domain in repression of ICP4 function (Speina *et al*., 2005).

Notably, the glycosylase-like fold is smaller than region 4 described by McGeoch *et al.* (1986). Potentially, this region of ICP4 may create an additional structurally independent domain. With the presented methodology we were unable to get an insight into the structure of the C-terminal segment of region 4 and the whole region 5, but other protein modelling algorithms may prove to be successful with those regions (Kolinski *et al*., 2004; Ekonomiuk *et al*., 2005).

The fold recognition approach provides an additional view into the function of the complex *Alphaherpesvirinae* gene expression regulator by providing evidence for a distant homology between ICP4 and proteins of established function. Although the presented structural assignments should be treated with rather limited confidence, this analysis clearly shows that bioinformatics has an important role in annotation of divergent genomes, like those of *Herpesviridae.* Application of profile-profile methodology in detection of distant similarity and various methods of fold recognition supported by homology modelling allows the identification of unexpected structural assignments. Our analysis, apart from providing an insight into the action of te ICP4 protein, allows further exploitation of the presented data in a rational design of experimental studies (e.g. mutation studies or binding assays).

Acknowledgements

This work was supported by the European Commission (LSHG-CT-2003-503265, LSHG-CT-2004-512035) and MNiSW (2P05A00130). LSW is supported by the Focus program from The Foundation for Polish Science.

REFERENCES

Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM (2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev* **29**: 231–262.

Batchelor AH, Wilcox KW, O'Hare P (1994) Binding and repression of the latency-associated promoter of herpes simplex virus by the immediate early 175K protein. *J Gen Virol* **75**: 753–767.

Beard P Faber S, Wilcox KW, Pizer LI (1986) Herpes simplex virus immediate early infected-cell polypeptide 4 binds to DNA and promotes transcription. *Proc Natl Acad Sci USA* **83**: 4016–4020.

Bruce JW, Wilcox KW (2002) Identification of a motif in the C terminus of herpes simplex virus regulatory protein ICP4 that contributes to activation of transcription. *J Virol* **76**: 195–207.

Bujnicki J, Elofsson A, Fischer D, Rychlewski L (2001) Structure prediction meta server. *Bioinformatics* **17**: 750–751.

Carrozza MJ, DeLuca NA (1996) Interaction of the viral activator protein ICP4 with TFIID through TAF250. *Mol Cell Biol* **16**: 3085–3093.

DeLuca N, Schaffer P (1988) Physical and functional domains of the herpes simplex virus transcriptional regulatory protein ICP4. *J Virol* **62**: 732–743.

DiDonato J, Muller M (1989) DNA binding and gene regulation by the herpes simplex virus type 1 protein ICP4 and involvement of the TATA element. *J Virol* **63**: 3737–3747.

Ekonomiuk D, Kielbasinski M, Kolinski A (2005) Protein modeling with reduced representation: statistical potentials and protein folding mechanism. *Acta Biochim Polon* **52**: 741–748.

Everett R, DiDonato J, Elliott M, Muller M (1992) Herpes simplex virus type 1 polypeptide ICP4 bends DNA. *Nucleic Acids Res* **20**: 1229–1233.

Faber S, Wilcox K (1986) Association of the herpes simplex virus regulatory protein ICP4 with specific nucleotide sequences in DNA. *Nucleic Acids Res* **14**: 6067–6083.

Faber SW, Wilcox KW (1988) Association of herpes simplex virus regulatory protein ICP4 with sequences spanning the ICP4 gene transcription initiation site. *Nucleic Acids Res* **16**: 555–570.

Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**: 1015–1018.

Ginalski K, von Grotthuss M, Grishin NV, Rychlewski L (2004) Detecting distant homology with Meta-BASIC. *Nucleic Acids Res* **32**: W576–W581.

Grondin B, DeLuca N (2000) Herpes simplex virus type 1 ICP4 promotes transcription preinitiation complex formation by enhancing the binding of TFIID to DNA. *J Virol* **74**: 11504–11510.

Gu BH, Kuddus R, DeLuca NA (1995) Repression of activator-mediated transcription by herpes simplex virus ICP4 *via* a mechanism involving interactions with the basal transcription factors TATA-binding protein and TFIIB. *Mol Cell Biol* **15**: 3618–3626.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**: D258–D261.

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**: 195–202.

Knizewski L, Kinch L, Grishin NV, Rychlewski L, Ginalski K (2006) Human herpesvirus 1 UL24 gene encodes a potential PD-(D/E)XK endonuclease. *J Virol* **80**: 2575–2577.

Kolinski A (2004) Protein modeling and structure prediction with a reduced representation. *Acta Biochim Polon* **51**: 349–371.

Krwawicz J, Arczewska K, Speina E, Maciejewska A, Grzesiuk E (2007) Bacterial DNA repair genes and their eukaryotic homologues: 1. Mutations in genes involved in base excision repair (BER) and DNA-end processors and their implication in mutagenesis and human disease. *Acta Biochim Polon* **54**: 413–434.

Lang D, Powell SK, Plummer RS, Young KP, Ruggeri BA (2007) PAX genes: roles in development pathophysiology, cancer. *Biochem Pharmacol* **73**: 1–14.

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

Linding R, Russell RB, Neduva V, Gibson TJ (2003) GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res* **31**: 3701–3708.

McGeoch D, Dolan A, Donald S, Brauer DH (1986) Complete DNA sequence of the short repeat region in the genome of herpes simplex virus type 1. *Nucleic Acids Res* **14**: 1727–1745.

McGuffin LJ, Jones DT (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* **19**: 874–881.

Metzler D, Wilcox K (1985) Isolation of herpes simplex virus regulatory protein ICP4 as a homodimeric complex. *J Virol* **55**: 329–337.

Michael N, Roizman B (1989) Binding of the herpes simplex virus major regulatory protein to viral DNA. *Proc Natl Acad Sci USA* **86**: 9808–9812.

Pei J, Sadreyev R, Grishin NV (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics* **19**: 427–428.

Rost B, Yachdav G, Liu J (2004) The PredictProtein server. *Nucleic Acids Res* **32**: W321–W326.

Smith C, Bates P, Rivera-Gonzalez R, Gu B, DeLuca NA (1993) ICP4 the major transcriptional regulatory protein of herpes simplex virus type 1 forms a tripartite complex with TATA-binding protein and TFIIB. *J Virol* **67**: 4676–4687.

Speina E, Cieśla JM, Graziewicz MA, Laval J, Kazimierczuk Z, Tudek B (2005) Inhibition of DNA repair glycosylases by base analogs and tryptophan pyrolysate Trp-P-1. *Acta Biochim Polon* **52**: 167–178.

Thompson J, Higgins D, Gibson T (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673–4680.

Underhill DA (2000) Genetic and biochemical diversity in the Pax gene family. *Biochem Cell Biol* **78:** 629–638.

von Grotthuss M, Pas J, Wyrwicz L, Ginalski K, Rychlewski L (2003) Application of 3D-Jury GRDB, Verify3D in fold recognition. *Proteins* **53** (Suppl 6): 418–423.

Wagner EK, Guzowski JF, Singh J (1995) Transcription of the herpes simplex virus genome during productive and latent infection. In *Progress in Nucleic Acid Research and Molecular Biology* (Cohn WH, Moldave K, ed) pp 123–165. San Diego CA: Academic Press.

Watson RJ, Clements JB (1980) A herpes simplex virus type 1 function continuously required for early and late virus RNA synthesis. *Nature* **285**: 329–30.

Wintjens R, Rooman M (1996) Structural classification of HTH DNA-binding domains and protein-DNA interaction modes. *J Mol Biol* **262**: 294–313.

Wu C-L, Wilcox KW (1990) Codons 262 to 490 from the herpes simplex virus ICP4 gene are sufficient to encode a sequence-specific DNA binding protein. *Nucleic Acids Res* **18**: 531–538.

Wu C-L, Wilcox KW (1991) The conserved DNA-binding domains encoded by the herpes simplex virus type 1 ICP4 pseudorabies virus IE180, varicella-zoster virus ORF62 genes recognize similar sites in the corresponding promoters. *J Virol* **65**: 1149–1159.

Wyrwicz LS, Rychlewski L (2007a) Herpes glycoprotein gL is distantly related to chemokine receptor ligands. *Antiviral Res* **75**: 83–86.

Wyrwicz LS, Rychlewski L (2007b) Identification of Herpes TATT-binding protein. *Antiviral Res* **75**: 167–172.

Xia K, DeLuca N, Knipe D (1996) Analysis of phosphorylation sites of herpes simplex virus type 1 ICP4. *J Virol* **70**: 1061–1071.

Xu HE, Rould MA, Xu W, Epstein JA, Maas RL, Pabo CO (1999) Crystal structure of the human Pax6 paired domain-DNA complex reveals specific roles for the linker region and carboxy-terminal subdomain in DNA binding. *Genes Dev* **13**: 1263–1275.