Virus Genes

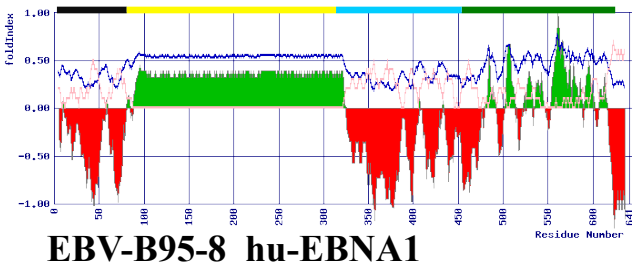
Modelling the structure of full length Epstein-Barr Virus Nuclear Antigen 1

--Manuscript Draft--

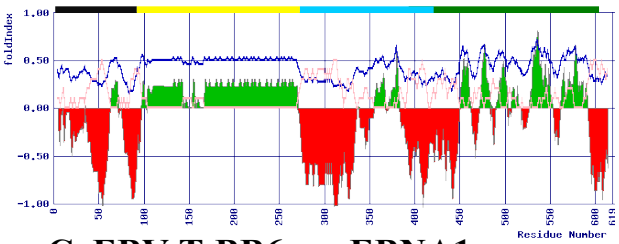
Manuscript Number:	
Full Title:	Modelling the structure of full length Epstein-Barr Virus Nuclear Antigen 1
Article Type:	Original work (Full Paper)
Section/Category:	Human Virus
Keywords:	EBV; EBNA1; structural model; I-TASSER; MOE
Corresponding Author:	Joanna Beatrice Wilson, Ph.D. University of Glasgow Glasgow, UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Glasgow
Corresponding Author's Secondary Institution:	
First Author:	Mushtaq Hussain, PhD.
First Author Secondary Information:	
Order of Authors:	Mushtaq Hussain, PhD.
	Derek Gatherer, Ph.D.
	Joanna Beatrice Wilson, Ph.D.
Order of Authors Secondary Information:	
Abstract:	Epstein-Barr virus (EBV) is a clinically important human virus associated with several cancers and is the etiologic agent of infectious mononucleosis. The viral nuclear antigen-1 (EBNA1) is central to the replication and propagation of the viral genome and likely contributes to tumourigenesis. We have compared EBNA1 homologues from other primate lymphocryptoviruses (LCV) and found that the central glycine/alanine repeat (GAr) domain, as well as predicted cellular protein (USP7 and CK2) binding sites are present in homologues in the Old World primates, but not the marmoset; suggesting that these motifs may have co-evolved. Using the resolved structure of the C-terminal one third of EBNA1 (homodimerisation and DNA binding domain), we have gone on to develop monomeric and dimeric models in silico of the full length protein. The C-terminal domain is predicted to be structurally highly similar between homologues, indicating conserved function. Zinc could be stably incorporated into the model, bonding with two N-terminal cysteines predicted to facilitate multimerisation. The GAr contains secondary structural elements in the models, while the protein binding regions are unstructured, irrespective of the prediction approach used and sequence origin. These intrinsically disordered regions may facilitate the diversity observed in partner interactions. We hypothsise that the structured GAr could mask the disordered regions, thereby protecting the protein from default degradation. In the dimer conformation, the C-terminal tails of each monomer wrap around a proline-rich protruding loop of the partner monomer, providing dimer stability, a feature which could be exploited in therapeutic design.

Modelling the structure of full length Epstein-Barr Virus Nuclear Antigen 1

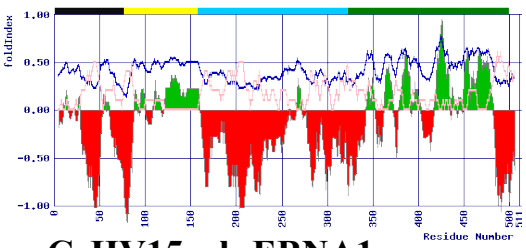
Hussain, Gatherer and Wilson **Supplemental Information**



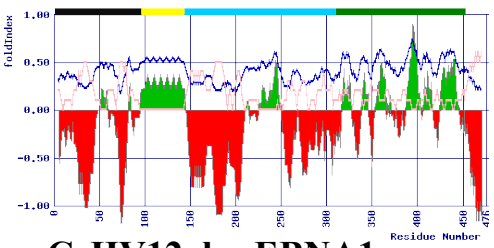
EBV-B95-8 hu-EBNA1



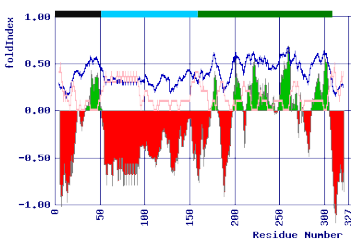
CyEBV-TsBB6 cy-EBNA1



CeHV15 rh-EBNA1



CeHV12 ba-EBNA1



CalHV3 ma EBNA1

Supplementary Figure S1. Folding propensity of EBNA1.

EBNA1 primary amino acid sequences from the primate LCVs* (as indicated) were used to predict the propensity of the proteins to fold by FoldIndex. FoldIndex uses hydrophobicity values and absolute net charge of the residues to predict structural and unstructured components of proteins. (GlobPlot2.3 employs defined propensity scales (based upon empirical protein structure data) in prediction). The distribution of hydrophobic and charged residues across the protein length are shown by blue and pink lines (respectively). Regions predicted to be disordered or structured are represented by negative (red) or positive (green) values (respectively). A simplified domain structure is indicated as a coloured bar above each plot: black: N-terminal; yellow: GAr; cyan: GR2 and protein binding sites; green: DNA binding and dimerisation. Residue number is given on the X axis.

Note: the GAr is predicted to be structured in each case. Also note, a small region within the CK2 binding domain (within the section indicated by a cyan bar) in cy-EBNA1, rh-EBNA1 and ba-EBNA1 is predicted to be structured (unlike hu-EBNA1) within the largely disordered stretch. This region maps to the extended CK2 binding site found in the Old World monkey sequences (see figure 2 of the manuscript).

*EBNA1 homologues were only identified in primate LCVs. Text mining identified several proteins from rice (*Oryza sativa*) that are annotated as EBNA1 or EBNA1-like. Similarly, some bacterial sequences (for example from *Erwinia chrysanthemi*) are also annotated as EBNA1-nuclear protein. None of these sequences show a significant BLAST score ($<10^{-50}$), or can be aligned with the LCV EBNA1 proteins (data not shown). Additionally, using reciprocal BLAST, none of these proteins show similarity with LCV EBNA1.

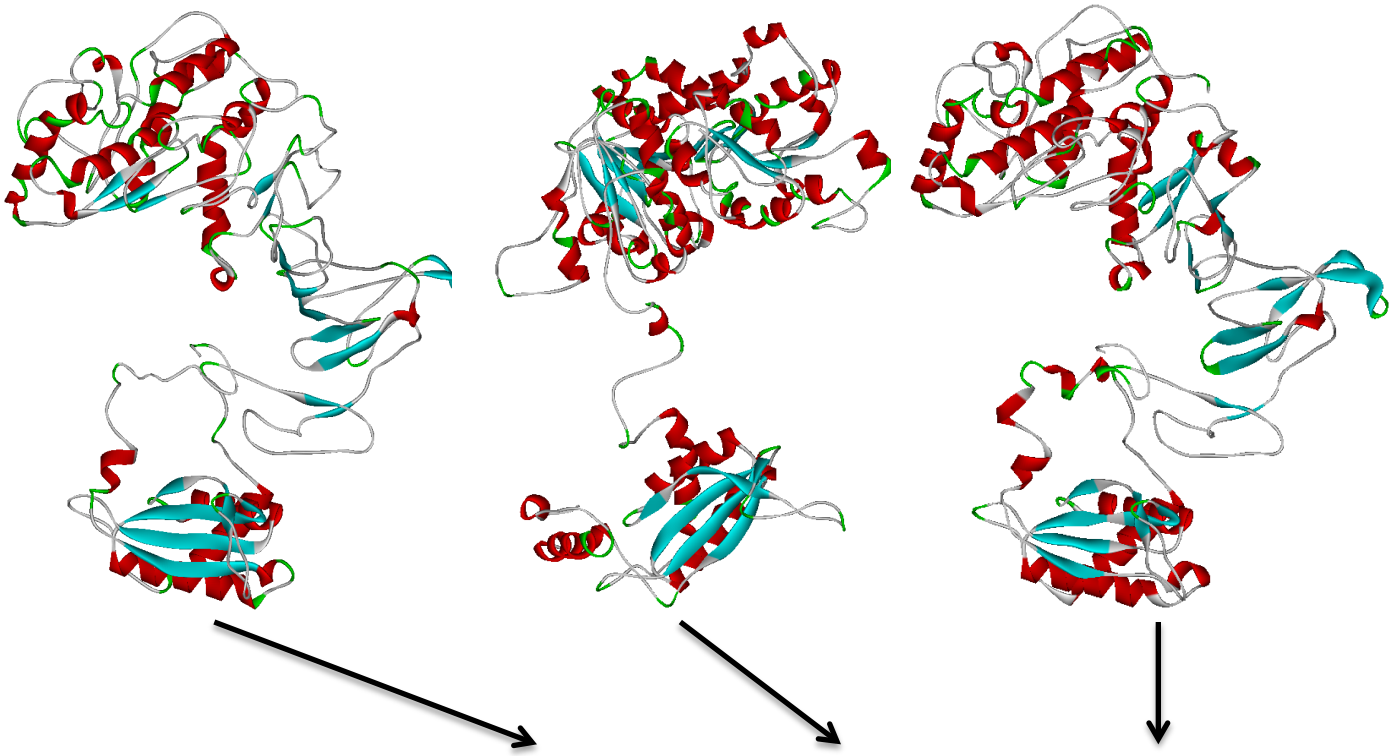
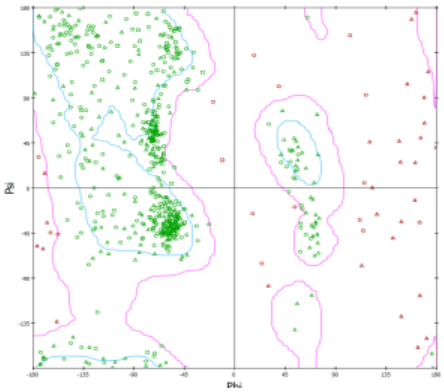
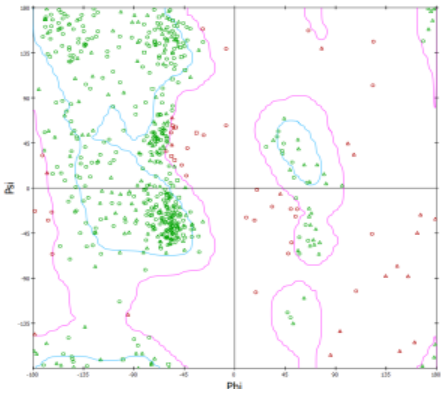


Table S1	ITASSER	MOE	Composite
Ramachandran plot outliers	7.98%	4.85%	4.07%
Ramachandran plot favoured region	73.1%	80.6%	89.5%
RMSD with 1B3T	0.4Å	1.05Å	1.29Å
Bad bonds	0	0.31%	0
Bad angles	1.09%	1.72%	0.62%
QMEANnorm score	0.14	0.09	0.15

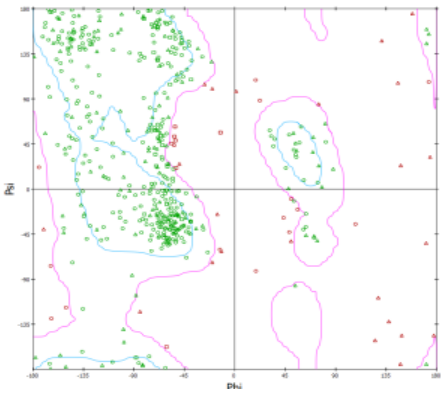
Supplementary Figure S2. EBNA1 model comparison.
The EBV B95-8 EBNA1 sequence was input to I-TASSER, which selected the EBNA1 C-terminal domain crystal structure (1B3T) and several fragment templates comprising: yeast fatty acid synthetase (2PFF), a-L-fucosidase (2Z8X), photosynthetic reaction centre (1C51), type A collagen (1YOF) and dimeric 6-phosphoglucouronate dehydrogenase (2ZYD). The 1B3T template was also used to generate models in MOE. EBNA1 models constructed using I-TASSER and MOE (and the composite of these two generated in Modeller9v8 (shown above), were assessed for structural plausibility using Molprobity and QMEAN score servers (table S1). Models were superimposed over the template (1B3T) and RMSD was estimated for each. The qualitative model energy analysis, normalised (QMEANnorm) score of a protein structural model provides a composite scoring function based on several geometrical aspects, both global (for the entire structure) and local (per residue), enabling the discrimination of good and bad models. A score in the range of 0 to 1.0 reflects a good model (optimally towards 0.5) and outside of this range (negative values or >1) reflects a poor model (for a non-membrane protein). While the composite model shows greater difference from 1B3T (RMSD), in all other respects it is better than either primary model. Overall the composite model shows more structural similarity to the I-TASSER primary model than to the MOE primary model. (I-TASSER has been ranked number 1 for several years in the CASP contest (critical assessment of protein structure prediction) http://predictioncenter.org/)



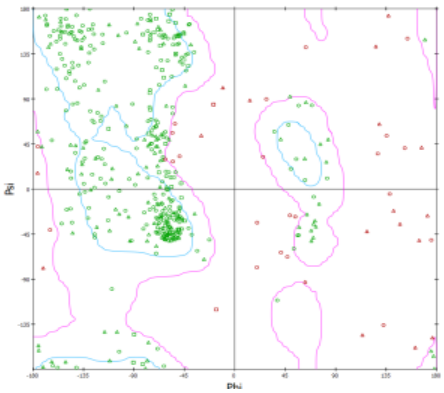
hu-EBNA1



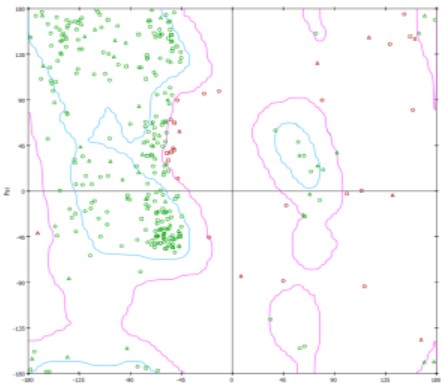
cy-EBNA1



rh-EBNA1

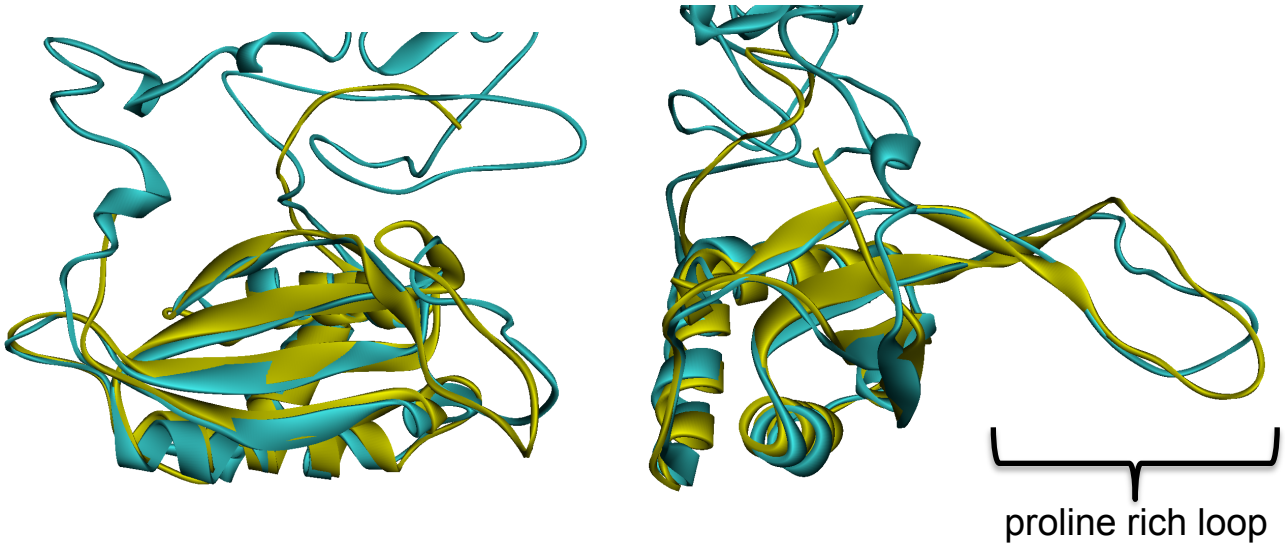


ba-EBNA1



ma-EBNA1

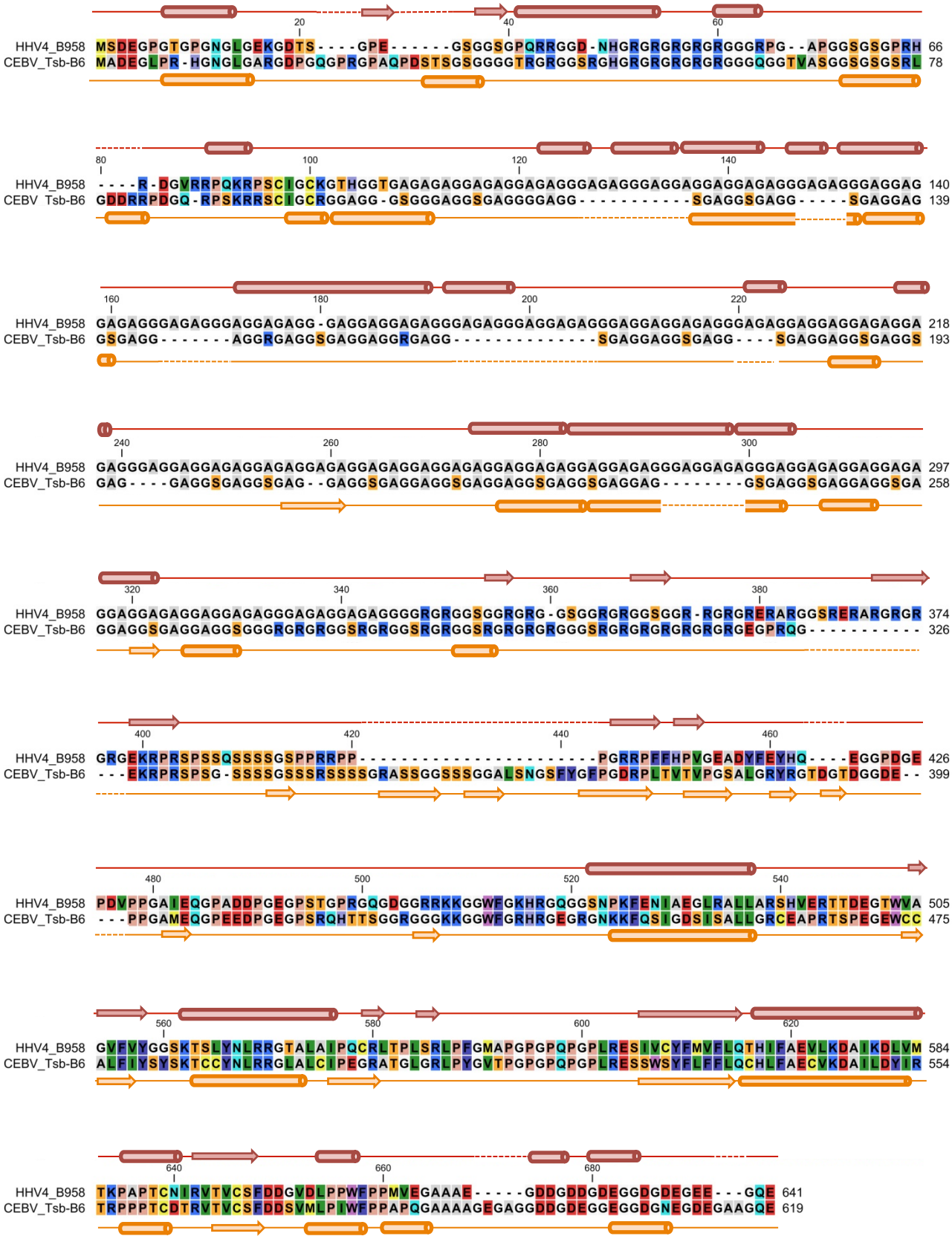
Supplementary Figure S3. Ramachandran Plots of EBNA1 modeled structures.
The human and other primate LCV EBNA1 protein structure models (as indicated) were evaluated for dihedral bond angle (Phi and Psi) distribution using Ramachandran plots. Residues in allowed and disallowed regions are represented by green and pink spots (respectively). Generously and strictly allowed regions are depicted by fuchsia and cyan contour lines (respectively). The hu-EBNA1 (B95-8 strain) composite model shows 89.5% of residues within the allowed region and an additional 5.9% in the generously allowed region. Given the proportion of Gly and Pro residues, these values are well within plausible limits.

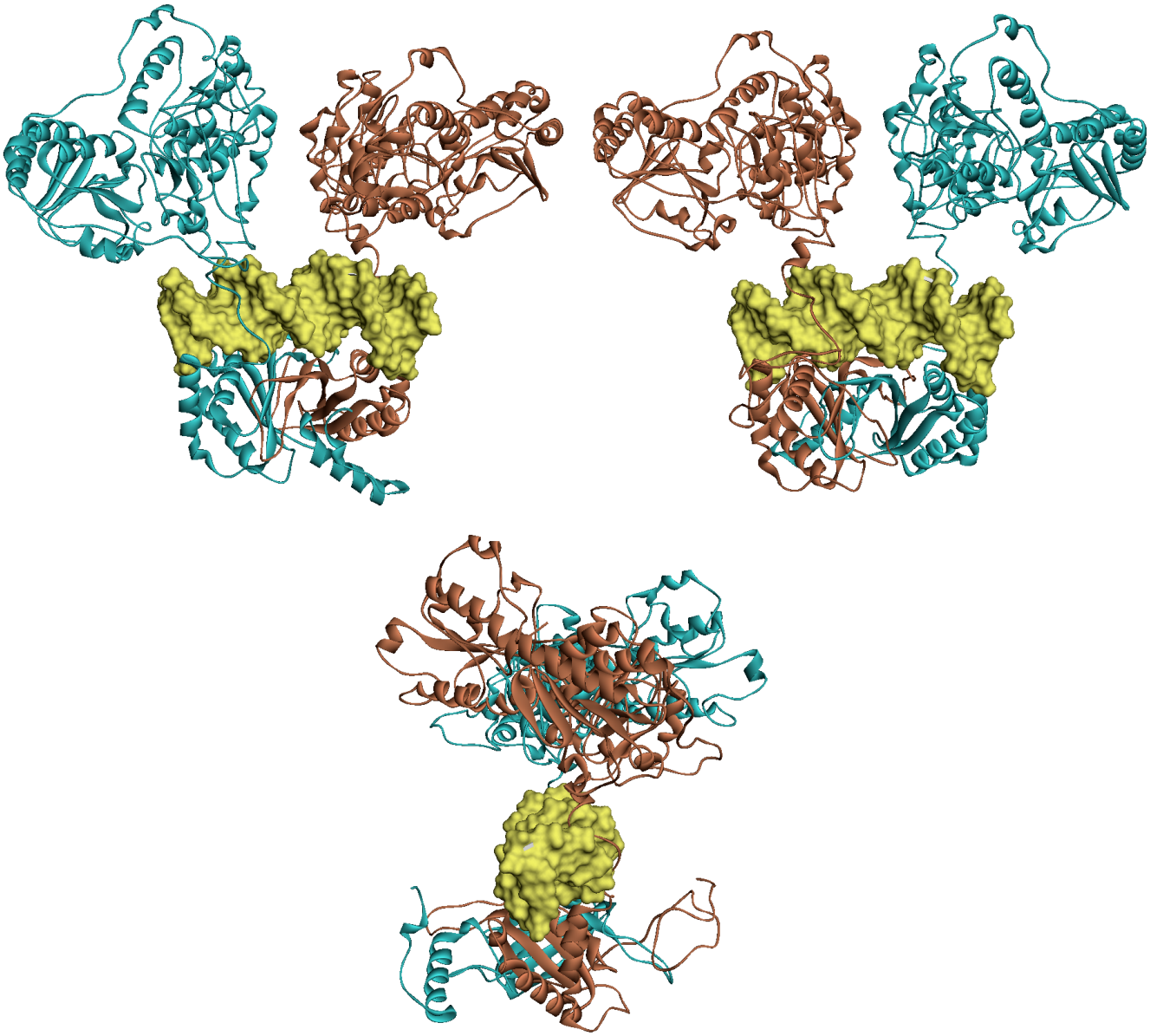


Supplementary Figure S4. Comparison of the C-terminal region of the EBNA1 composite model and the resolved structure.
The C-terminal region of the composite EBNA1 model (cyan) is shown superimposed over a monomer extracted from the resolved template 1B3T (yellow), from two angles (with a horizontal rotation of 90º). The RMSD value is 1.29Å. Note: the proline rich loop protrudes from both structures (right view).

Legend to Supplementary Figure S5. Sequence alignment between hu-EBNA1 and cy-EBNA1.
The sequences of hu-EBNA1 (EBV/HHV4 B95-8) and cy-EBNA1 (Cyno-EBV/CEBV TsB-B6) are shown aligned. The distribution of structural elements observed in the *in silico* models of each are shown as cylinders (α helices) and arrows (β sheets) in maroon (hu-EBNA1) and yellow (cy-EBNA1). Note: the modelled additional β sheets in the elongated potential CK2 binding site of cy-EBNA1 (starting at residue 348) in comparison to hu-EBNA1.

Supplementary Figure S5





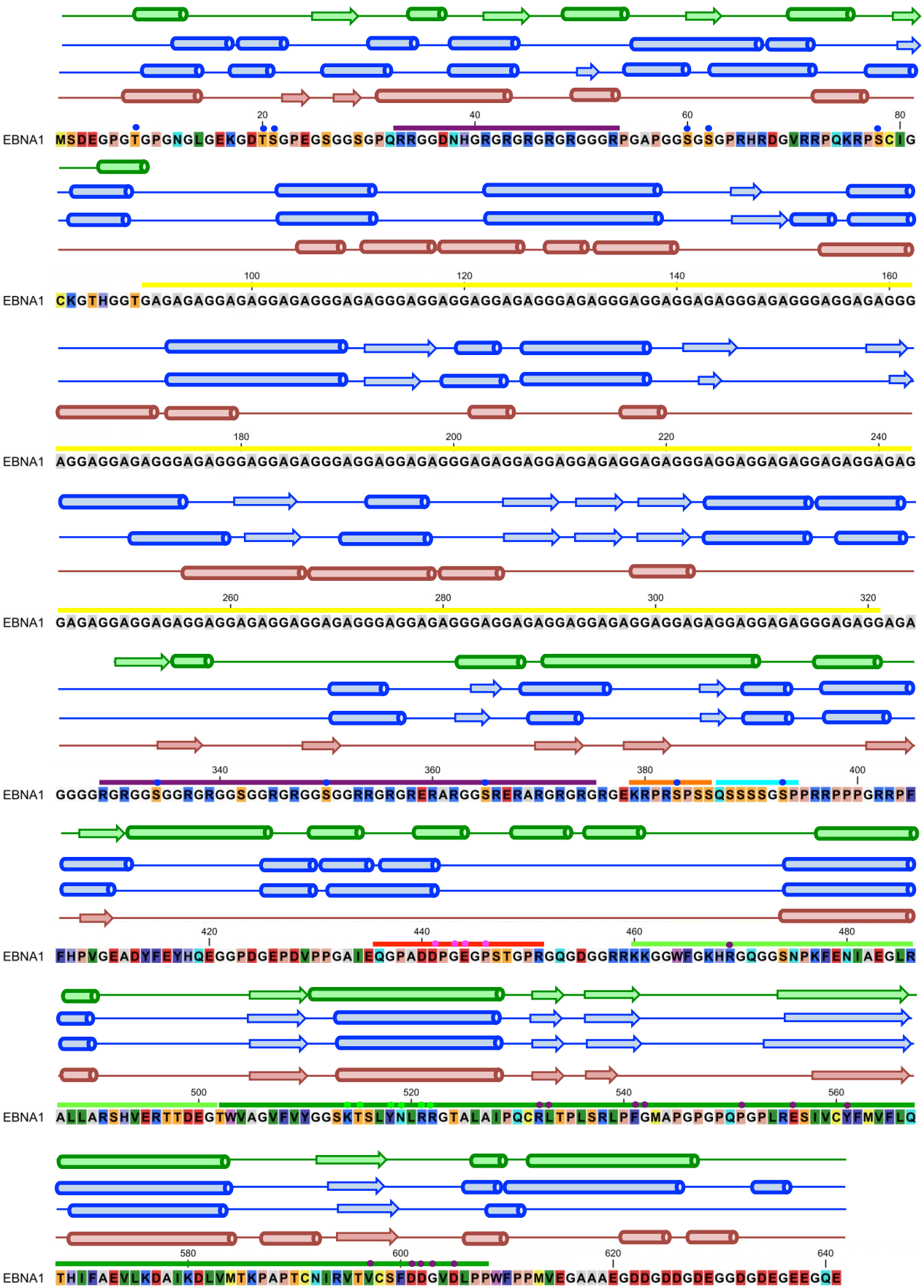
Supplementary Figure S6. EBNA1 MOE model bound with DNA.
The hu-EBNA1 dimer model developed using MOE (using spatial constraints for DNA binding) is shown. Each monomer is represented in ribbon format (brown and cyan) while DNA is shown in surface format (yellow). Since the C-terminal region of the dimer is modelled by homology, the resulting model is structurally highly similar to the 1B3T template (superimposition of the MOE-dimer model with 1B3T gives an RMSD value of 0.35 Å). Shown above: the model with 180^{o} horizontal rotation and beneath: with 90^{o} rotation. Note: the string of residues connecting the C-terminal DNA binding and dimerisation domain with the remainder of the protein lies in the major groove of the DNA. Several features of dimer stability of the full length EBNA1 SymmDock-dimer were compared with a C-terminal tail (residues 608 to 641) deleted SymmDock-dimer model (tabulated). GCS: geometrical complimentarity score (the higher, the more symmetrical the dimer); ACE: atomic contact energy; G: free energy (the lower, the more stable).

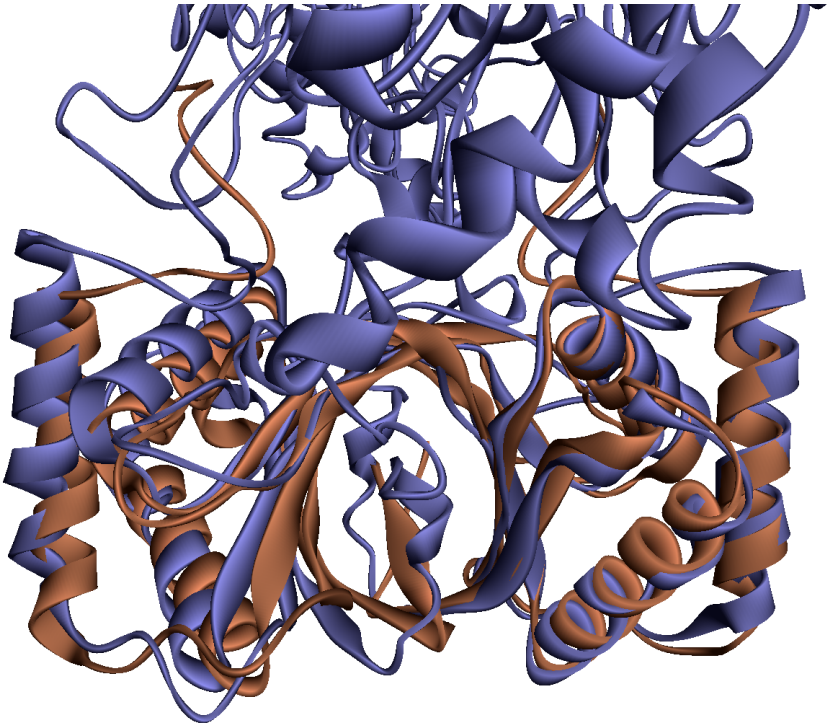
Dimer Model: feature	Full length	Tail deleted
GCS	17890	13654
ACE	-1062.7	-398
G	-11094.65	-9562

Supplementary Table S2.
Several features of dimer stability of the full length EBNA1 composite-dimer were compared with a C-terminal tail (residues 608 to 641) deleted SymmDock-dimer model (tabulated). GCS: geometrical complimentarity score (the higher, the more symmetrical the dimer); ACE: atomic contact energy; G: free energy (the lower, the more stable).

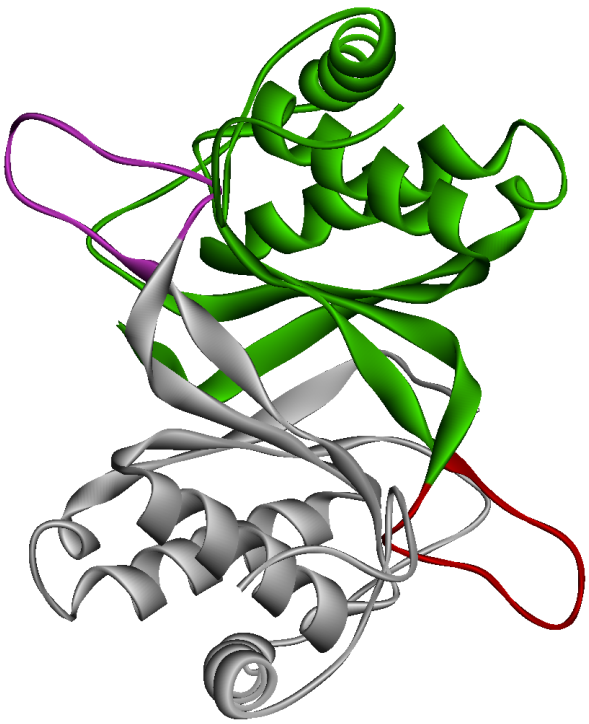
Legend to Supplementary Figure S7. Secondary structure distribution of EBNA1 models.
The distribution of the secondary structural elements of the different hu-EBNA1 models is shown above the primary sequence of EBV B95-8 EBNA1 (as assessed by HERA plot). The composite model (maroon), each MOE monomer in the dimer model (blue) and the GAr deleted composite model (green) are compared. Selected protein domains or interaction sites are indicated by coloured horizontal bars: purple: GR1 and GR2; yellow: GAr; orange: NLS; cyan: CK2 binding site; red: USP7 binding site; green: DNA binding and dimerisation domain. Coloured dots above the sequence indicate other noted residues: blue: predicted phosphorylation sites; pink: critical residues involved in USP7 binding; purple: dimerisation; green: DNA binding.

Supplementary Figure S7

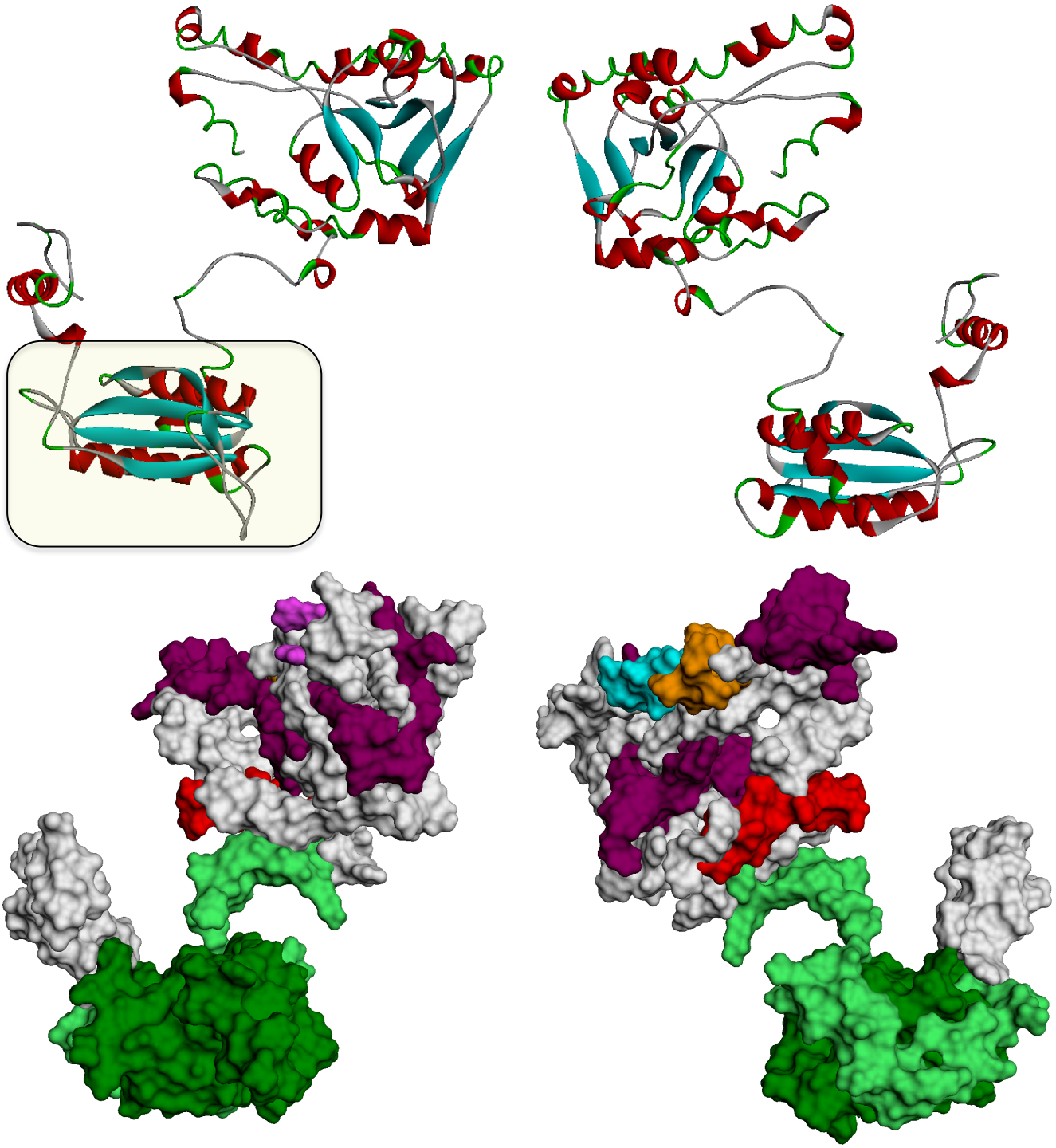




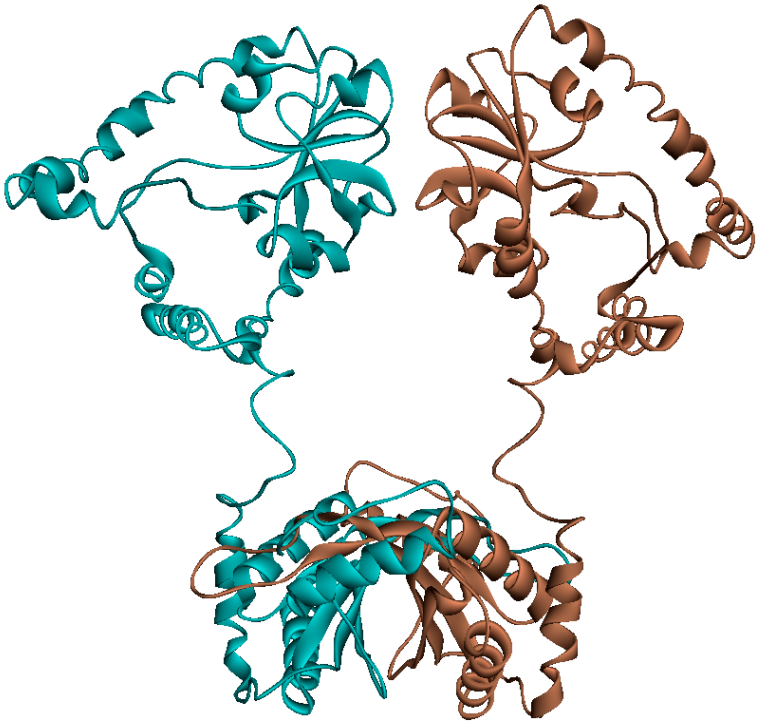
Supplementary Figure S8. Comparison of composite dimer EBNA1 model with the resolved structure.
The composite EBNA1 dimer model (blue) is shown superimposed over the resolved structure 1B3T (brown) (RMSD value 1.5 Å). One side of the β barrel (right side from the viewed angle) was used as the reference point to align the dimers. It can be seen that the modelled dimer shows an altered angle between the monomers (compared to 1B3T), such that the β barrel is slightly wider (seen here by the slight misalignment on the left side).

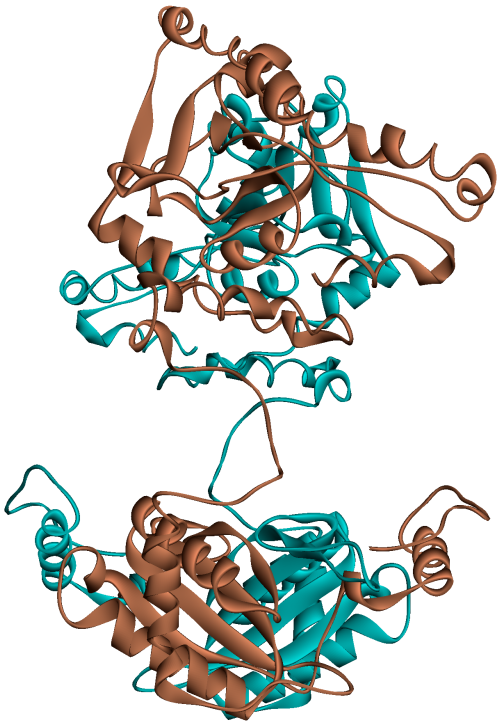


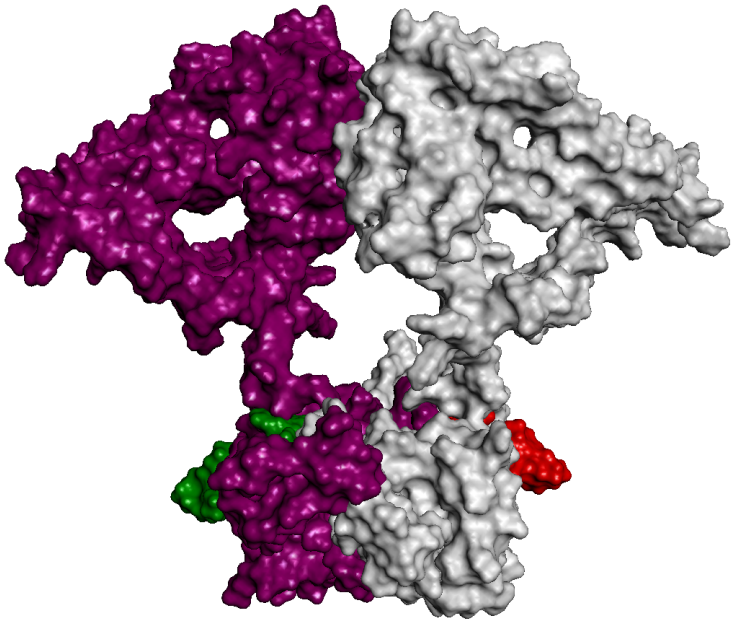
Supplementary Figure S9. The proline rich loops in the dimer.
"Top" view of the resolved EBNA1 1B3T dimer showing the protruding proline rich loops. Each monomer/loop is differently coloured green/red and silver/mauve.



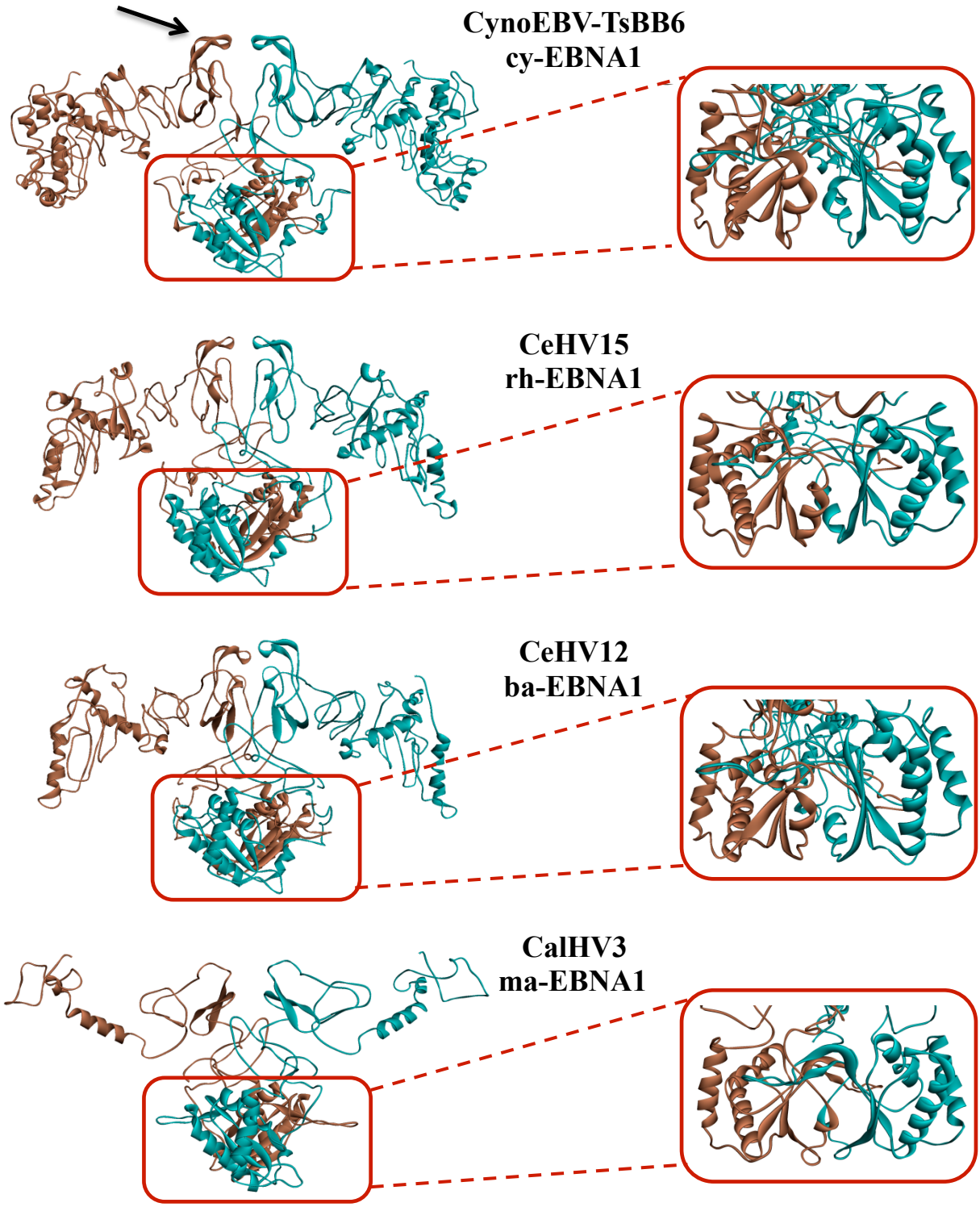
Supplementary Figure S10. GAr deleted EBNA1 Model.
Deletion of the GAr of hu-EBNA1 allows increased expression of the protein in heterologous systems and retains several of the protein's functions. A composite monomeric model of GAr-deleted hu-EBNA1 sequences was generated as for full length (using I-TASSER, MOE and Modeller). Removal of GAr impacts the predicted structure of the N terminal half, showing alpha helices in the CK2 and USP7 binding domains. Interestingly, three of the hydrogen bonds seen in 1B3T that are absent or differently paired in the full length model (such as Arg469-Glu556) are present in the dimer model of the GAr deleted EBNA1 (manuscript table 2). The composite model of GAr deleted EBNA1 is shown in both ribbon format (above) and surface view (below) with 180º rotations. The boxed region (above left) indicates the region which has been resolved (in 1B3T). The C-terminal DNA binding and dimerisation domain of the model conforms to the resolved 1B3T structure used as the original template. The surface topology images are coloured to show structural and/or functional domains (as defined in figure 2): yellow: GAr; purple: GR1 and GR2; pink: Arg71 and Arg72; cyan: CK2 interaction region; orange: NLS; red: USP7 binding site; light green and dark green: flanking region and core DNA binding and dimerisation domain.







Supplementary Figure S11. GAr deleted EBNA1 homodimer.
The GAr deleted composite EBNA1 monomer model was used to generate a dimer in SymmDock. In this model, a longer alpha helix is predicted within the C-terminal tail and as such it does not form a complete ring (see figure S10). Nevertheless, the proline-rich loop in the dimer conformation still protrudes through the space of the half ring formed.
Ribbon format views are shown with 90^{o} rotation with each monomer coloured cyan or brown (above). Monomers/proline rich loops in the surface surface topology view (left) are differently coloured: mauve/red and silver/green.



Supplementary Figure S12. Model Dimers of EBNA1 from the primate LCVs.
Homodimers were generated using SymmDock for each of the modelled (non-human) primate LCV EBNA1 homologues. Monomers are coloured cyan or brown. To the right, the C-terminal regions of each (dimerisation and DNA binding domain) are shown enlarged and rotated by 90^{o} in the horizontal plane. Note: arrow in cy-EBNA1, structure of the extended CK2 binding region. As well as differences in the N-terminal regions, some differences in the C-terminal domain structure from the hu-EBNA1 homodimer are apparent. The β barrel of the Old World monkey LCV EBNA1 structures is wider compared to hu-EBNA1. The β barrel of the ma-EBNA1 dimer is more similar to hu-EBNA1 in terms of symmetry of the interacting interface of the monomers.

Title: Modelling the structure of full length Epstein-Barr Virus Nuclear Antigen 1

Mushtaq Hussain[1], Derek Gatherer[2] and Joanna B. Wilson*

College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, G12 8QQ, UK

*Corresponding author
Joanna.wilson@glasgow.ac.uk
Tel: +44 141 3305108

[1] current address: Institute of Biological, Biochemical and Pharmaceutical Sciences, Dow University of Health Sciences, Karachi, Pakistan
[2] current address: Division of Biomedical and Life Sciences, Lancaster University, Lancaster LA1 4YG, UK

Key words

EBV, EBNA1, structural model, I-TASSER, MOE

Abstract

Epstein-Barr virus (EBV) is a clinically important human virus associated with several cancers and is the etiologic agent of infectious mononucleosis. The viral nuclear antigen-1 (EBNA1) is central to the replication and propagation of the viral genome and likely contributes to tumourigenesis. We have compared EBNA1 homologues from other primate lymphocryptoviruses (LCV) and found that the central glycine/alanine repeat (GAr) domain, as well as predicted cellular protein (USP7 and CK2) binding sites are present in homologues in the Old World primates, but not the marmoset; suggesting that these motifs may have co-evolved. Using the resolved structure of the C-terminal one third of EBNA1 (homodimerisation and DNA binding domain), we have gone on to develop monomeric and dimeric models *in silico* of the full length protein. The C-terminal domain is predicted to be structurally highly similar between homologues, indicating conserved function. Zinc could be stably incorporated into the model, bonding with two N-terminal cysteines predicted to facilitate multimerisation. The GAr contains secondary structural elements in the models, while the protein binding regions are unstructured, irrespective of the prediction approach used and sequence origin. These intrinsically disordered regions may facilitate the diversity observed in partner interactions. We hypothsise that the structured GAr could mask the disordered regions, thereby protecting the protein from default degradation. In the dimer conformation, the C-terminal tails of each monomer wrap around a proline-rich protruding loop of the partner monomer, providing dimer stability, a feature which could be exploited in therapeutic design.

Introduction

Epstein-Barr virus (EBV, or human herpesvirus 4, HHV4) is a prevalent gammaherpesvirus, infecting upwards of 90% of the world population. Primary infection in the very young is usually asymptomatic, but in adolescents and adults can cause infectious mononucleosis. Infection with EBV is also a risk factor for developing certain autoimmune disorders, including multiple sclerosis and systemic lupus erythematosus (1, 2). Importantly, EBV is associated with several cancers, including nasopharyngeal carcinoma, Burkitt's lymphoma and Hodgkin's lymphoma. EBV undergoes a lytic and a latent cycle and the virus persists for the lifetime of the host, residing in memory B-cells. During latency, propagation of the extrachromosomal viral genome requires the EBV nuclear antigen 1 (EBNA1), a

multifunctional DNA binding protein. EBNA1 plays a role in viral genome replication and is essential for its efficient mitotic segregation (3, 4). The protein also acts as a transcriptional regulator of both viral and host promoters and may promote lytic reactivation of the virus (5, 6). Furthermore, multiple lines of evidence indicate that EBNA1 contributes to the development of EBV associated tumours through increasing cell proliferation and survival and possibly by inducing oxidative stress (7-17).

EBNA1 binds to DNA as a homodimer in a site-specific manner, recognising a consensus 16bp site found multiple times in the viral genome (18-21). The site-specific DNA binding and dimerisation domain maps to the C-terminal region of the 641 amino acid protein, between residues 459 and 607 (22). EBNA1 can also attach to cellular chromosomes non-specifically, mediated through two regions termed linking regions 1 and 2 (LR1 and LR2). The N-terminal LR1 (residues 33-89) can be subdivided into two regions: a Gly and Arg repeating unit (GR1: residues 33 to 53), which allows the protein to associate with AT-rich DNA ("AT hook") and a unique region with multiple Gly then Arg residues, containing potential Ser phosphorylation sites and two highly conserved Cys residues. Both sub-regions are required for transactivation by the protein (18, 23, 24). LR2 or GR2 (residues 327-377) is also a Gly and Arg repeating unit. A host nucleolar protein, EBP2, associates with EBNA1 at GR1 and GR2, which might facilitate loading of EBNA1 onto mitotic chromosomes (25, 26). Furthermore, it has been proposed that through GR1 and GR2, EBNA1 recruits the origin replication complex (ORC) to the viral latent replication orgin (OriP) in an RNA dependent manner (27, 28).

A long Gly/Ala repeat (GAr) sequence (residues 90 to 324) regulates EBNA1 expression, renders the protein resistant to proteosomal degradation and facilitates immune evasion (29, 30). This repeat region retards translation of EBNA1 and, it has been proposed, thereby reduces production and processing of misfolded products, thus inhibiting major histocompatibility complex (MHC) class I peptide presentation of the protein (31). This idea is supported by the observation that cytotoxic T-cell recognition of EBNA1 is largely mediated by the presentation of epitopes derived from newly synthesized protein (32). However, it is not clear how this function relates to the GAr mediated property of stabilising the mature protein (30).

The structure of a C-terminal domain dimer of EBNA1 (residues 461 to 607) has been resolved, co-crystallised binding to the specific DNA recognition sequence, revealing details of this interaction and the critical contact residues (33, 34). This C-terminal region incorporates four antiparallel β strands, the eight in the dimer forming a barrel-like structure,

with two alpha helices (per monomer) situated outside the β barrel. Two sub-regions were described, a flanking region (residues 459-503) and a core domain (residues 504-607), but both contribute to sequence specific DNA binding (35).

EBNA1 associates with multiple host proteins including ubiquitin-specific protease 7 (USP7), casein kinase 2 (CK2) (36, 37), EBP2 and others (26, 37-41). The core interaction sites of EBNA1 with USP7 and CK2 have been determined and map to the central region of the protein (37, 42). Most of the protein-protein interactions of EBNA1 map within the unresolved N-terminal two thirds of the protein (7, 43), similarly, the contribution of the C-terminal tail (the last 33 residues) to the protein structure is unknown.

Given the importance of EBNA1 in the life cycle of this highly medically relevant human virus, we aimed to develop a structural understanding of the full length protein. With the difficulties in expressing the full length protein, we have generated *in silico* models of EBV EBNA1, in both monomer and homodimer conformations. The model homodimer permits prediction of the interaction surfaces and conformation of EBNA1 protein-protein interacting regions. Additionally we have generated *in silico* models of EBNA1 from the related primate lymphocryptoviruses (LCV), providing insight into the structural morphogenesis of the protein through virus and host co-evolution.

Materials and Methods

Sequence Retrieval:

EBNA1 sequences were retrieved from Uniprot (http://www.uniprot.org/) (44). Eight complete sequences were available: from human-EBV strains B95-8, AG876 and GD1 (id: P03211; Q1HVF7; Q3KSS4), from cynomolgus-EBV (Cyno-EBV or CyEBV), strains Si-IIA and TsB-B6 (id: Q9IPQ8 and Q9IPQ9), from rhesus lymphocryptovirus (CeHV15; id: O91332), from baboon Cercopithecine Herpesvirus 12 (CeHV12; id: Q80890) and from marmoset Callitrichine Herpesvirus 3, (CalHV3; id: Q993H1).

Multiple Sequence Alignment:

Sequences were aligned using Clustal X under default parameters (45). Manual adjustments were conducted using BioEdit where required. Alignments were visualized using CLC sequence viewer 5.

Phylogenetic Analysis:

Maximum likelihood trees were constructed using the Whelan and Goldman replacement model with 1000 bootstrap replicates in Mega 5.10 (46, 47).

Molecular Modelling:

Structural models were generated using I-TASSER, MOE (http://www.chemcomp.com/) and Modeller 9v8 (48, 49). The EBV B95-8 EBNA1 sequence was input to I-TASSER, which uses homology modelling where available and unaligned regions are modelled *ab initio*. Cluster centroids were generated using replica exchange Monte-Carlo simulations and by averaging all the clustered structure decoys. The structures were refined in terms of global topology (49). Models were generated in MOE using the template 1B3T (with and without DNA) and *ad hoc* outgap modelling to similar fragments from PDB for the remainder of the sequence, as follows. An initial proposed partial geometry was copied from the template chains in the solved structure of 1B3T by using all coordinates where residue identity was conserved. Otherwise, only backbone coordinates were used. Based on this initial partial geometry, Boltzmann-weighted randomized modelling (50) was employed with segment searching in PDB for regions that could not be mapped onto the initial partial geometry (51). Each of 25 models was energetically minimized in the AMBER-99 force field (52). The highest-scoring intermediate model was determined by generalized Born/volume integral (GB/VI) methodology (53). Molecular surfaces were created using the method of Connolly (54), as applied within MOE. The best models developed by both approaches were selected on the basis of the lowest RMSD deviations with the template (1B3T), lowest free energy and atomic clashes. These were used as templates to construct composite models in Modeller 9v8. The best composite model (of 50 obtained, all with identical backbone) was selected on the basis of normalized Discrete Optimized Molecule Energy (DOPE). This composite model was used to generate models for other EBNA1 homologues analysed, using MOE, repeating the process described above, by providing query and template primary structure alignment and template atomic coordinates as input.

Structure Assessment:

FoldIndex and GlobPlot2.3 were used to assess the propensity to form secondary and tertiary structures (55, 56). The default parameter for window size in FoldIndex was changed to 10 residues, to be comparable with GlobPlot2.3. The models were assessed for their structural plausibility using Molprobity (57). This includes generating Ramachandran plots (via Rampage: http://mordred.bioc.cam.ac.uk/~rapper/rampage.php), calculation of bad angles and bad bonds and atomic clash analysis. In addition models were assessed using Qualitative model energy analysis (QMEAN) normalised scoring, using the QMEAN server (58). Ubiquitination sites were predicted using CKSAAP UbSite web server (59). Pictorial

representations of structural motifs against the linear sequence were generated in the EMBL server PDBsum.

Dimer Construction:

EBNA1 dimers were constructed in SymmDock (60) using the composite monomer model as input and noted side chain interaction points in the C-terminus from the crystal structure (33). SymmDock restricts dimer models to those with a symmetric arrangement. All monomers and dimers were visualized in DSvisualizer 3.5. The top 10 dimer conformations, based on geometric scores, desolvation energies and the interface area size, as ranked in SymmDock, were analysed and superimposed over the 1B3T structure. The dimer with least RMSD to 1B3T was selected (in each case this proved to be the top ranked dimer model). A full length EBNA1 dimer was also constructed in MOE by homology modelling using 1B3T as a dimer template and with bound DNA.

Results

Phylogeny and sequence alignment of EBNA1 homologues

EBNA1 of EBV is an unusual protein; with the exception of the GAr region, it shows very limited primary sequence similarity to any other protein in the databases. However, there are several homologues of EBNA1 found in the related herpesviruses of other primates. Using EBNA1 of EBV B95-8 strain (deleting the GAr) in a BLAST search, eight complete EBNA1 sequences were retrieved, including 3 from different EBV strains and from LCVs infecting the cynomolgus monkey (*Macaca fascicularis*), rhesus macaque (*M acaca mulatta*), baboon genus (*Papio*) and marmoset family (*Callitrichidae*). The EBNA1 sequences from these viruses are termed here: hu-EBNA1 (from EBV/HHV4), cy-EBNA1 (from CyEBV), rh-EBNA1 (from CeHV15), ba-EBNA1 (from CeHV12) and ma-EBNA1 (from CalHV3). The phylogenetic history of these EBNA1 genes was inferred using the protein sequences, revealing the separation of the single New World primate viral sequence (ma-EBNA1) from the Old World primate virus sequences (figure 1). Predictably, the hu-EBNA1 sequences are closely related and distinct from the other sequences.

The sequences of the 8 EBNA1 homologues were aligned revealing that the hu-EBNA1 sequences are highly similar and are the longest amongst the homologues (figure 2 and table 1). Identity between the Old World monkey LCV EBNA1 sequences ranges between 35% and 46%, while ma-EBNA1 (the shortest protein) shows the lowest identity with the other EBNA1 sequences (table 1). The GAr of hu-EBNA1 spans 233 residues (90-324) and (in B9-58) is entirely composed of Gly and Ala residues. The GAr of the Old World monkey viruses

is shorter and intervened by other residues (primarily Ser and Val) and it is entirely absent from ma-EBNA1.

The Gly and Arg repeat regions (GR1 and GR2) of hu-EBNA1 (residues 33-53 and 327-377), which flank the GAr sequence, are present in the Old World primate virus EBNA1 homologues. However, in the absence of the GAr, ma-EBNA1 has only one GR region (aligned to GR2). A sequence adjacent to GR1 in hu-EBNA1 (KRPSCIGCKG) is highly conserved in Old World primate LCV EBNA1 but is absent from ma-EBNA1. However, two Cys residues in the N-terminal region of ma-EBNA1 (residues 38 and 43), which have been aligned to Cys79 and Cys82 of hu-EBNA1 (figure 2), might reflect a conserved function.

The hu-EBNA1 interaction sites for USP7 and CK2 are conserved in the Old World monkey virus EBNA1 homologues. In particular, the residues involved in hydrogen bonding between an EBNA1 peptide and USP7 (EBNA1 Pro442, Glu444, Gly445 and Ser447) (42) are fully conserved (figure 2). However, this USP7-binding sequence is absent from ma-EBNA1. The runs of Ser at the CK2 binding site are extended in the Old World monkey LCV EBNA1 sequences compared to hu-EBNA1. At the aligned site in the ma-EBNA1 sequence are several Ser/Pro residues, but it is not clear if these could constitute a CK2 binding site. This suggests that these sequences and their function have either been lost during the course of New World monkey viral evolution or gained during Old World monkey viral evolution. Between the CK2 and USP7 binding sites is a stretch of approximately 30 residues, similar between the Old World monkey viral EBNA1 sequences, but dissimilar to hu-EBNA1.

There is a single predicted ubiquitination site in hu-EBNA1 (Lys477), which is conserved in all the homologues. Out of the 10 proposed phosphorylation sites conserved in hu-EBNA1 sequences (61), six are also conserved in most of the primate virus homologues (figure 2). The nuclear localisation signal (NLS, 379-385) is completely conserved in EBNA1 homologues of the Old World primate viruses, but not in ma-EBNA1. A consensus NLS signal (K, K/R, x, K/R) is not present within the ma-EBNA1 sequences, however RKxRxxxK towards the N-terminus or RKRxxxR at the start of the DNA binding domain might serve as an NLS. In ba-EBNA1 and rh-EBNA1, a conserved KKRRS within the LR1 homology region could provide a second NLS.

The most conserved sequence between the EBNA1 homologues is the DNA binding and dimerisation domain (residues 459-607) (figure 2). With few exceptions, residues shown to be involved in DNA binding or important for interactions in the dimer are identical across the homologues (figure 2).

EBNA1 monomeric structure model

The propensity of hu-EBNA1 to form secondary and tertiary conformation was explored using FoldIndex, which predicts that the N-terminal and central regions of the protein are unstructured or unfolded (supplemental figure S1). By contrast, the GAr domain yielded a positive signal for likely folded conformation and the C-terminal domain shows multiple peaks of predicted structure. Similar results were obtained using GlobPlot2.3 (not shown).

In order to generate a structural model for full length EBNA1 *in silico*, the hu-EBNA1 B95-8 sequence was examined for any structural similarities (using PDB Blast) with proteins of known structure in the RCSB database. None of the templates thus identified (apart from the resolved EBNA1 C-terminal domain), alone or in combination, covered EBNA1 to any extent. Therefore, a combined approach to exploit the advantages of three programmes (I-TASSER, MOE and Modeller) was conducted. I-TASSER uses homology modelling where a template is available (1B3T in this case) and for regions where no structure is available, I-TASSER searches the protein structure databases for small regions of similarity and selects several template fragments (supplemental figure S2), threading them into the model, which is then evaluated for best fit and the process is iterative. Unaligned regions are modelled *ab initio* and I-TASSER assembles thousands of possible structures which are then selected on the basis of multiple parameters (including free energy and Ramachandran plot fit). Homology modelling can also be conducted in MOE, which has the advantage of being able to restrict the model, in this case using the bound DNA in 1B3T as a required space. MOE models unresolved portions of the protein using fragments of high-resolution chains from the PDB which superpose well onto anchor residues on either side of the area in question. In order to utilise the benefits of both programmes, primary models were generated separately in I-TASSER and MOE. The two selected primary models were then used as templates employing Modeller to generate a composite structure. The best composite model was selected, which showed improved parameter values over both primary models (figure 3 and supplemental figure S2, table S1 and model 1). Plausibility of the selected composite model was supported by QMEAN score and analysis in Molprobity, which includes an evaluation of dihedral bond angles by Ramachandran plot (supplemental figure S3). As expected, the C-terminal region of the *in silico* model is almost identical to the 1B3T crystal structure which was used in model generation (supplemental figure S4).

The EBNA1 model predicts a helix towards the N-terminus of the protein (residues 31-43), within GR1 (figures 2 and 3). A conserved residue within the LR1 transactivation domain (Arg71 and also to some extent Arg72) protrudes, horn-like, from the structure. The

GAr region forms multiple helices of varying size, consistent with the FoldIndex and GlobPlot2.3 predictions that this region is structured. The remainder of the N-terminal region forms un-structured loops and turns, also consistent with the predictions. Similarly, the central region of EBNA1 (residues 325-476) incorporating the CK2 and USP7 binding sites, is modelled largely as un-structured stretches, with the inclusion of short, parallel β sheets in GR2 (residues 334-338, 348-351 and 370-374), in the NLS and in the stretch between the CK2 and USP7 core binding sites. A proline rich stretch, well conserved in all the primate sequences (residues 537-559), forms a loop (as seen in the crystal structure) which protrudes from one side of the model (supplemental figure S4). The acidic C-terminal tail of EBNA1 (residues 608-641) is predicted to be largely unstructured. In the model, this C-terminal tail curls back towards the molecule, making a ring shape (figure 3).

In order to compare the structural properties of hu-EBNA1 with homologues from the related primate viruses, models of the latter were generated using the former as a template in MOE (figure 4 and supplemental models 2 to 5). High structural similarity was observed between the dimerisation and DNA binding domain of all primate virus EBNA1 homologues. Like hu-EBNA1, the C-terminal tail of the other EBNA1 proteins curls to form a ring. However, the remainder of the proteins show several differences. With the reduction in length of the GAr, GR1 and GR2 become closer. In the absence of the GAr from ma-EBNA1, a longer N-terminal alpha helix is predicted. The beta sheets seen in GR2 of hu-EBNA1 are reduced or absent in the other modelled structures. The predicted USP7 binding sites remain as unstructured regions in the homologues. However, the longer CK2 domains, particularly of cy-EBNA1, form β sheets, as does the sequence intervening between this and the USP7 binding site (supplemental figure S5). Consistent with the models, this region was also predicted in the folding propensity plots for cy-EBNA1, rh-EBNA1 and ba-EBNA1 (but not hu-EBNA1) as a likely structured region in the middle of the unstructured stretch (supplemental figure S1). The prominent Arg71 residue noted in hu-EBNA1, also protrudes from the molecule in the modelled ba-EBNA1, but the side chain of the corresponding residue in the other homologues is buried within the structure.

EBNA1 homodimer model

Two approaches were used to generate *in silico* homodimer models of full-length EBNA1: dimer modelling in MOE and dimer generation using SymmDock. The 1B3T resolved C-terminal domain dimer with bound DNA was used as a template in MOE to generate a full

length EBNA1 dimer with DNA (figure 5 and supplemental figure S6 and model 6). Using this approach permits incorporation of the DNA molecule into the model and a string of residues connecting the C-terminal domain to the N-terminal two thirds of the protein neatly sits in the major groove of the bound DNA (figure 5). However, the N-terminal two thirds of the protein (and the C-terminal tail), modelled in MOE alone, shows several differences to the composite model described above (supplemental figure S7). In addition, this MOE-dimer model is not symmetrical, the two monomers showing a different distribution of secondary structures, most notably in the C-terminal tail (supplemental figures S6 and S7). This asymmetry may simply be a chance effect of the choice of short PDB fragments for the *ad hoc* outgroup modelling, or it may reflect the fact that the DNA structure from 1B3T is used as a spatial constraint. As DNA does not exhibit bilateral symmetry, bound dimers and models of bound dimers may also show asymmetry, indeed this has been observed for crystallised C-terminal fragments of EBNA1 (62).

In an alternative approach to generate a dimer structure, the use of SymmDock was evaluated. First a model dimer was generated using a monomer of the resolved EBNA1 C-terminal domain (a monomer from 1B3T). The dimer model thus generated in SymmDock used the correct contacts as in the resolved dimer (table 2) and superimposition of model and resolved structure showed that the two were virtually identical (with an RMSD value of 0.01Å, not shown), thus validating the approach. Consequently, the *in silico* composite full length monomer model (described above) was used to generate a homodimer in SymmDock, inputting the contact residues described for 1B3T. Of the top 20 models generated, the best was selected on the basis of lowest free energy (figures 6 and 7 and supplemental model 7). The model correctly predicts interaction between the two monomers within the C-terminal dimerisation domain. Similarly, the DNA recognition and binding sequences are appropriately positioned, relative to the resolved C-terminal domain dimer. The N-terminal two thirds of the monomers form perpendicular arms in relation to the dimerisation domain, giving a slightly different orientation compared to the MOE-dimer. Superimposition of the composite-dimer model over the atomic coordinates of the 1B3T dimer shows a close alignment (supplemental figure S8). The predicted hydrogen bonding of the SymmDock-dimer model includes bonds seen in the C-terminal domain resolved structure, as well as between residues that were not in the crystallised fragment (Table 2 and figure 7).

While the composite EBNA1 model gives improved structural reliability scores compared to either MOE-model produced in the dimer conformation, the composite model shows intermolecular clashes with the location where DNA should bind. At present, the DNA

molecule cannot be input into I-TASSER as a model constraint. However, the clashing residues in the composite-dimer model are unstructured and *in vivo* are likely to be flexible and able to move aside to allow the dimer to bind to DNA. We propose that the composite model currently provides the best prediction of the structure of full length EBNA1 (in monomer and homodimer configuration) when not bound to DNA. We hypothesize that upon binding to DNA through the C-terminal domain, the obstructing unstructured regions move aside to allow this and the angle of the N-terminal bulk of the protein, in relation to the C-terminal domain, may twist as seen in the MOE-dimer model, with the linking peptide strand sitting within the DNA major groove.

Interestingly, the proline rich loop which protrudes out from the monomeric model (noted above), in the dimer conformation penetrates into the other monomer, into the ring formed by the C-terminal tail and the core domain (figure 6). The proline rich loop was observed in the resolved structure (supplemental figure S9), however, in the absence of the C-terminal tail its significance in the dimer conformation was not previously apparent. This feature is also observed in the MOE-dimer model and a GAr deleted-hu-EBNA1 model (supplemental figures S6, S10 and S11 and model 8). The "prehensile tail" of one monomer wrapping around the proline-rich loop of the other may help to stabilise the dimer. To explore this hypothesis, we generated a dimer in SymmDock, using a monomer model deleted for the C-terminal tail (deleting residues 608 to 641). Of the top 20 best models (lowest free energy), only 8 showed the correct orientation of the two monomers to each other, compared to 15/20 for the full length sequence. Moreover, the best C-terminal tail deletion dimer model showed poorer geometry and energy value scores compared to the full length dimer (supplemental table S2), indicating that it is less stable. These data support the prediction that the C-terminal tail provides dimer stability.

EBNA1 dimers form multimers when bound to the repeated DNA recognition sites (the family of repeats, FR) at oriP (24). Two cysteine residues (Cys79 and Cys82) that are highly conserved, have been found to coordinate zinc and to facilitate these multimeric interactions. Arrayed EBNA1 dimers have been proposed to act cooperatively in transcriptional transactivation through FR (24). In order to explore this structurally, zinc ions were introduced into the full length composite dimer model, using MOE. Energy minimisation following placement of a zinc ion near Cys79 and Cys82 in each monomer predicted stable bonding between the zinc ion and the two cysteines (unlike placement of zinc near the proximal His residues, which did not allow bonding). The zinc-bound cysteines sit at the distal end of each N-terminal arm of the dimer model (figure 8 and supplemental model 9).

From this model it can be envisaged how adjacent dimers might link through zinc (arm to arm in an array), the zinc ion being coordinated by four cysteines, two provided by each of the adjacent arms of neighbouring dimers.

Dimers were also generated from each of the EBNA1 monomers of the non-human primate LCVs in SymmDock (supplemental figure S12 and models 10 to 13). In each case, dimerisation occurs between the C-terminal regions while the N-terminal regions angle away from each other.

Discussion

We have compared EBNA1 sequences from different primate LCVs to explore the evolutionary history of this viral gene. In addition, we have generated model structures of full length EBNA1, a GAr deleted version and dimers of these *in silico*.

The phylogenetic tree developed from EBNA1 sequences from related LCVs is consistent with the evolutionary history of these viruses inferred from an analysis of DNA polymerase and glycoprotein B sequences (63). From this it was proposed that LCVs have co-evolved with the host, but some evidence of interspecies virus transfer was also apparent (63, 64). The extensive differences between ma-EBNA1 and hu-EBNA1 (outside of the C-terminal domain) suggest that EBNA1 has undergone subfunctionalisation since the separation of Old and New World primates 43 MYA.

In silico protein structure modelling tools are becoming ever more sophisticated and as a consequence the structures predicted, more reliable (65). Concomitantly, the number of resolved structures is rapidly increasing, improving the resources for both template-based structure prediction and model assessment. Thus, models with good Molprobity and QMEAN scores provide a plausible structure prediction, which is the case for the EBNA1 models described here. Nevertheless, the very nature of a model is that veracity cannot be guaranteed. Despite this, consistently observed features in these models allow conclusions to be drawn and hypotheses made.

All the models generated in this study, irrespective of the prediction approach and the EBNA1 strain sequence used, model the N-terminal two thirds of the protein as a region with unstructured stretches, distinct from the C-terminal domain. This prediction is accordant with the resolved C-terminal structure and the MOE-dimer model incorporating DNA, which show that the N-terminal region is linked to the C-terminal domain via a string of residues that sit in the DNA major groove, thus separating the two protein regions by the DNA strand. The C-

terminal DNA binding and dimerisation domain is well conserved and predicted to be structurally highly similar between the EBNA1 homologues, suggesting a conservation of action in homodimerisation and sequence specific DNA binding. Introduction of zinc ions into the dimer model predicts stable binding to Cys79 and Cys82 (conserved in all homologues), which could facilitate self-association of an array of EBNA1 dimers (figure 8E). This is consistent with the observation that these residues are required for cooperative transactivation requiring zinc (24) and that deletion of LR1 from hu-EBNA1 significantly impairs the transactivation function (23).

Both GR1 and GR2 sequences of hu-EBNA1 are involved in binding to EBP2, G-rich RNA and recruitment of ORC and are required for stimulation of EBNA1 dependent viral genome replication, tethering to metaphase chromosomes and the efficient segregation of viral genomes (26, 27). Sequence and structural conservation of these regions in the EBNA1 homologues suggest a functional conservation in viral genome propagation and maintenance. The single GR region in ma-EBNA1 may reflect a more ancestral form that has not been split in two by incorporation of the GAr domain (the latter is absent from ma-EBNA1). Ma-EBNA1 also lacks the consensus NLS, however, its likely role in viral genome propagation suggests nuclear localisation might be mediated via another sequence (as described above).

In all of the EBNA1 homologues, the short proline-rich tract within the C-terminal domain is conserved. In the resolved structure this forms a loop jutting out from the molecule and this feature is recapitulated in all of the models. Interestingly, in all of the dimer models, this loop of each monomer slots into the space formed between the C-terminal tail and core domain of the other monomer, forming an interlocking structure in the manner of a “dowel pin joint”. Moreover, a dimer model generated lacking the C-terminal tail shows energy values indicative of reduced stability in comparison to the full length dimer. Although the C-terminal tail is not essential for dimer formation, the tail of each monomer (prehensile-like) curling around the dowel pin of the other, may act to stabilise the dimer. Importantly, this information could be used in designing therapeutic agents to disrupt EBNA1 dimerisation and hence action.

The USP7 core binding site is highly conserved in the Old World primate LCV EBNA1 sequences, and it seems likely that they bind to host USP7. The extended CK2 binding site in the Old World monkey sequences, could reflect differences in host CK2 or that these proteins bind to CK2 with different affinity. Alternatively, the longer, more structured site (compared to hu-EBNA1) might confer another property. Ma-EBNA1 appears to lack both sites. As

more LCV EBNA1 sequences become available, it will be interesting to determine if these domains and the GAr were acquired and evolved together.

All the models of EBNA1 predict that the protein interactions sites (with EBP2, USP7 and CK2) are intrinsically unstructured or disordered. Such regions could provide molecular flexibility, allowing for rapid association/dissociation, promiscuity in partner interactions and increased availability to modification (66-68). Proteins with signalling or transcriptional regulatory functions appear to be enriched with intrinsically disordered segments, probably because this permits a greater repertoire of specific interactions (69). Indeed, EBNA1 interacts with multiple partners and the interaction sites in some cases overlap (e.g. with EBP2 and RNA), or the core binding sites are in close proximity (e.g. CK2 and USP7). Thus while core binding sites may confer specificity in partner interactions, it is possible that EBNA1 can adopt different shapes with different partners. Thus, excluding the more structured regions, the model predicted here may reflect a "resting" shape, which in the living cell can mould to accommodate a variety of partner molecules.

Intrinsically disordered proteins (IDPs) tend to be tightly regulated, their efficient proteosomal degradation (ubiquitin-dependent and independent) being critical to the health of the cell (70). It has been proposed that IDPs are susceptible to default degradation by the 20S proteasome (ubiquitin independent) and that disordered regions of a protein must be masked in order to avoid rapid degradation; thus newly synthesized IDPs are at higher risk of degradation (70, 71). Despite the predicted disordered regions in EBNA1, the protein is highly stable in B-cells (over 30 hours) (72), suggesting that disordered regions are masked. Cytotoxic T-cell recognition of EBNA1 is largely mediated by the presentation of epitopes derived from newly synthesized protein (32). This would be consistent with EBNA1 having disordered regions susceptible to default degradation immediately post synthesis. The observed stability of EBNA1 may result from either rapid complex formation, post translational modification, or be due to some other masking activity. Based upon current knowledge and our modelling studies, we propose that the GAr might perform such a masking function.

Predictions of the propensity of EBNA1 to fold suggest that the GAr is structured; the composite model predicts that the region forms several alpha helices and the MOE-dimer model predicts a mixture of helices and β sheets. Earlier studies proposed that the GAr might form β-sheets (73). More recently it has been determined that Ala residues have a strong propensity to form α helical and break β strand conformations. Although Gly residues

examined alone tend to break both conformations, in conjunction with Ala residues, Gly shows a strong propensity to form alpha helices (74), supporting the predictions of our models.

The hu-EBNA1 GAr sequence renders EBNA1 resistant to proteosomal degradation and inhibits self-synthesis, resulting in impaired immune presentation (30, 31). However, the length and purity of the repeat has a profound effect upon is action. The smaller, impure GAr of rh-EBNA1 and ba-EBNA1 confer little or no self-synthesis inhibition, but nevertheless these proteins are stable, with a half-life similar to hu-EBNA1 (72). In addition, rh-EBNA1 expressing cells are more efficiently recognised by cytotoxic T-cells and the epitopes are derived from newly synthesized protein (72). It has been proposed that by retarding translation, the hu-GAr sequence reduces the generation of misfolded products which would otherwise be processed and presented by MHC (31). In an extension to this hypothesis, we propose that the partially structured GAr sequence masks the disordered regions of EBNA1, thereby inhibiting default degradation by the 20S proteasome. Several observations are consistent with this hypothesis; 1) retarded translation might allow the GAr to fold in advance of translation of the disordered central region inhibiting early default degradation; 2) GAr masking of the disordered regions would continue to protect the mature protein from degradation; 3) GAr masking would not override ubiquitin mediated degradation, consistent with observations (73); 4) the hypothesis allows for separation of two functions of GAr, retarding translation and conferring protein stability, which has been observed experimentally for hu-EBNA (31) and would explain the observed differences in activity between hu-EBNA1 and the rh-EBNA1 and ba-EBNA1 proteins. This hypothesis is also consistent with the possibility that evolutionary acquisition of disordered regions (such as the USP7 and CK2 binding regions), necessitated co-acquisition of a masking function.

EBNA1 is the only viral protein consistently expressed in all proliferating EBV infected cells and thus represents a key therapeutic target. As such, knowledge of its structure and interactions will contribute to the understanding of viral biology and also aid in the design of potential anti-viral drugs.

Statement of Author contributions

MH conducted the majority of the analyses described herein and contributed to drafting the manuscript. DG generated all models in MOE and contributed to project progression.

JBW conceived, directed and coordinated the project and wrote the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

M.H. is a recipient of a Higher Education Commission of Pakistan and Dow University of Health Sciences Scholarship.

Legends to Figures

Figure 1

Phylogenetic tree of EBNA1. The evolutionary history of the LCV EBNA1 genes was inferred using the encoded amino acid sequences, employing the maximum likelihood method with the Whelan and Goldman replacement model. The consensus phylogenetic tree compiled from 1000 bootstrap replicates shows the support percentages at each node. One clade comprises three EBV homologues of EBNA1 (hu-EBNA1). Other EBNA1 sequences from non-human primate LCVs form a separate clade (herpesviruses from cynomolgous monkey (CyEBV), rhesus macaque (CeHV15) and baboon (CeHV12)), while the marmoset CalHV3 EBNA1 homologue outgroups both clades. A schematic representation of the domain structure of selected EBNA1 molecules is shown on the right (with amino acid length indicated). The predicted protein domains or sites are colour coded: purple: Gly, Arg repeat regions (GR1 and GR2), yellow: GAr, orange: NLS, cyan: CK2 binding site, red: USP7 core binding site, green: DNA binding domain (core in dark green).

Figure 2

Multiple sequence alignment of EBNA1. The amino acid alignment of EBNA1 homologues of human and other primate LCVs is shown using EBNA1 of EBV B95-8 strain as the reference sequence, with Rasmol colour coding of residues. The secondary structural elements (based on the composite model of EBV-B95-8) are shown above as arrows (β sheets) and cylinders (α helices). Selected protein domains or interaction sites are indicated by coloured horizontal bars: purple: GR1 and GR2, yellow: GAr, orange: NLS (379-385, KRPRSPS), cyan: CK2 binding site, red: USP7 binding site, green: DNA binding domain (core in dark green). Coloured dots above the sequence indicate other noted residues: blue: predicted phosphorylation sites (note conservation of Ser60, Ser62, Ser78, Ser365, Ser383, Ser393); pink: critical residues involved in USP7 binding; purple: dimerisation; green: DNA binding (note conservation of residues involved in DNA binding: Lys514, Thr515, Tyr518,

Asn519, Arg521 and Arg522 and dimer interactions: Arg469, Tyr510, Arg532, Leu533, Phe541, Gly542, Pro553, Glu556, Tyr561, Val597, Ser599, Asp601, Asp602, and Asp605) .

Figure 3

EBV EBNA1 composite model structure. A composite model structure of EBV B95-8 EBNA1 was generated using I-TASSER, MOE and Modeller and is presented in A-C, with 180^{o} rotation in D-F. (A, D) Ribbon diagram, indicating the region which has been resolved from the crystal structure (1B3T) boxed. (B, E) Electrostatic surface diagrams. (C) Surface topology with highlighted structural and/or functional domains (as defined in figure 2): yellow: GAr; purple: GR1 and GR2; pink: Arg71 and Arg72; cyan: CK2 interaction region; orange: NLS; red: USP7 binding site; light green and dark green: flanking region and core DNA binding and dimerisation domain. Note, the C-terminal acidic tail (arrow in A) loops round to form a "ring" with the C-terminal region of the molecule.

Figure 4

Primate LCV EBNA1 structures. The modelled structures of EBNA1 monomers from different primate LCVs (as indicated) are shown from 2 angles (180^{o} rotation) in ribbon format. Structural and/or functional domains as they relate to EBV EBNA1 are colour coded as in figure 3.

Figure 5

EBV EBNA1 MOE dimer model structure. A ribbon diagram of the EBV B95-8 EBNA1 MOE dimer model is depicted with bound DNA (yellow) shown in surface topology. Note the string of residues connecting the C-terminal domain with the remainder of the protein sitting in the DNA major groove.

Figure 6

EBV EBNA1 composite model dimer structure. (A): Ribbon diagram of the EBV B95-8 EBNA1 composite dimer model is shown with the monomers coloured cyan or brown. The C-terminal region involved in homodimerisation, DNA recognition and binding is enlarged and rotated by 90^{o} in the horizontal plane in (B) and viewed from the "top" (C). The central barrel can be clearly viewed in (B). In (C), the penetrating loop and DNA binding core domain are indicated for each monomer. (D and E): Surface topology diagrams of the monomer and

dimer (respectively), showing the penetrating loop, residues Phe541-Lys555 coloured red or green.

Figure 7

Contacts in the EBNA1 composite dimer model. The hydrogen bond pattern between the two EBNA1 monomers (cyan and brown) is shown. Black dashed lines represent hydrogen bonds and the interacting amino acids are labelled.

Figure 8

Zinc in the EBNA1 composite dimer model. Stable bonding between zinc (green) and Cys79 and Cys82 (red ball and stick) is predicted in the EBNA1 full length dimer model, shown in ribbon format with 90^{o} rotation in the vertical plane between (A) and (B). (C and D): Enlarged image of interaction as seen in (B). E: A cartoon of zinc linked EBNA1 dimers bound to DNA.

Table 1

	EBV GD1	EBV AG876	CEBV TsbB6	CEBV SiIIA	CeHV15	CeHV12	CalHV3
EBV B958	97	88	46	44	38	36	22
EBV GD1		87	46	43	38	36	22
EBV AG876			44	42	37	35	21
CEBV TsbB6				90	54	47	21
CEBV SiIIA					56	50	23
CeHV15						59	25
CeHV12							26

Table 1. Percentage sequence identity matrix of EBNA1 homologues

Table 2

No.	1B3T	SymmDock-dimer model	Gly/Ala deleted model
1	NS	NF	Lys313-Arg314 (x2)
2	NS	Glu367-Arg368 (x2)	NF
3	NS	Arg370-Ser386	NF
4	NS	Arg382-Asp455	NF
5	NS	Ser386-Asp455	NF
6	Arg469-Glu556 (x2)	NF	Arg469-Glu556
7	Tyr510-Asp605	Tyr510-Trp609 *	Tyr510-Asp605
8	NS	Tyr510-Phe610 *	NF
9	NF	NF	Arg521-Pro553 (x2)
10	Arg532-Gly542 (x2)	Arg532-Gly542 (x2)	Arg532-Met543 (x3) *
11	Arg532-Phe541	NF	Arg532-Gln550 *

12	Leu533-Pro553	Leu533-Pro553	NF
13	Gly542-Pro607	NF	NF
14	NF	NF	Ala544- Glu641
15	NF	NF	Arg555-Gly470 (x2)
16	NF	Ile558-Glu629 (x2)	NF
17	Tyr561-Tyr561	Tyr561-Trp609 *	Tyr561-Tyr561
18	NF	NF	Ala588-Asp625
19	NF	NF	Cys591-Asp625
20	Arg594-Asp605	Arg594-Pro608 *	NF
21	NS	Arg594-Phe610	NF
22	Thr596-Asp602	NF	NF
23	Val597-Asp601	NF	NF
24	Ser599-Ser599	NF	NF

Table 2. Comparison of intermolecular hydrogen bond pattern between the EBNA1 composite dimer model, the GAr deleted model and the resolved C-terminal domain structure. Intermolecular hydrogen bonds observed in 1B3T (residues 461-607), the full length composite EBNA1 SymmDock-dimer model (residues 1-641) and the GAr deleted (missing residues 91 to 327 inclusive) EBNA1 dimer model are listed. In the latter, residues are numbered according to the full length protein. Residues sharing more than one bond are indicated (x2 or x3). NS: not structured; NF: not found; * denotes a difference in bond partner compared to 1B3T.

References

1. Ascherio A., and Munger K.L., Journal of neuroimmune pharmacology : the official journal of the Society on NeuroImmune Pharmacology 5, 271-277, 2010.
2. Niller H.H., Wolf H., and Minarovits J., Autoimmunity 41, 298-328, 2008.
3. Lindner S.E., and Sugden B., Plasmid 58, 1-12, 2007.
4. Sivachandran N., Thawe N.N., and Frappier L., Journal of virology 85, 10425-10430, 2011.
5. Canaan A., Haviv I., Urban A.E., Schulz V.P., Hartman S., Zhang Z., Palejev D., Deisseroth A.B., Lacy J., Snyder M., Gerstein M., and Weissman S.M., Proceedings of the National Academy of Sciences of the United States of America 106, 22421-22426, 2009.
6. Sivachandran N., Wang X., and Frappier L., Journal of virology 86, 6146-6158, 2012.
7. Frappier L., Viruses 4, 1537-1547, 2012.
8. Hong M., Murai Y., Kutsuna T., Takahashi H., Nomoto K., Cheng C.M., Ishizawa S., Zhao Q.L., Ogawa R., Harmon B.V., Tsuneyama K., and Takano Y., Journal of cancer research and clinical oncology 132, 1-8, 2006.
9. Kaul R., Murakami M., Choudhuri T., and Robertson E.S., Journal of virology 81, 10352-10361, 2007.
10. Kennedy G., Komano J., and Sugden B., Proceedings of the National Academy of Sciences of the United States of America 100, 14269-14274, 2003.
11. Kube D., Vockerodt M., Weber O., Hell K., Wolf J., Haier B., Grasser F.A., Muller-Lantzsch N., Kieff E., Diehl V., and Tesch H., Journal of virology 73, 1630-1636, 1999.
12. Sheu L.F., Chen A., Meng C.L., Ho K.C., Lee W.H., Leu F.J., and Chao C.F., The Journal of pathology 180, 243-248, 1996.
13. Tsimbouri P., Drotar M.E., Coy J.L., and Wilson J.B., Oncogene 21, 5182-5187, 2002.
14. Wilson J.B., Bell J.L., and Levine A.J., EMBO J 15, 3117-3126, 1996.

15. Yin Q., and Flemington E.K., Virology 346, 385-393, 2006.
16. Gruhne B., Sompallae R., Marescotti D., Kamranvar S.A., Gastaldello S., and Masucci M.G., Proceedings of the National Academy of Sciences of the United States of America 106, 2313-2318, 2009.
17. Kamranvar S.A., and Masucci M.G., Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, UK 25, 1017-1025, 2011.
18. Sears J., Ujihara M., Wong S., Ott C., Middeldorp J., and Aiyar A., Journal of virology 78, 11487-11505, 2004.
19. Wang J., Lindner S.E., Leight E.R., and Sugden B., Molecular and cellular biology 26, 1124-1134, 2006.
20. Dresang L.R., Vereide D.T., and Sugden B., Journal of virology, 2009.
21. Yates J., Warren N., Reisman D., and Sugden B., Proceedings of the National Academy of Sciences of the United States of America 81, 3806-3810, 1984.
22. Ambinder R.F., Mullen M.A., Chang Y.N., Hayward G.S., and Hayward S.D., Journal of virology 65, 1466-1478, 1991.
23. Singh G., Aras S., Zea A.H., Koochekpour S., and Aiyar A., Journal of virology 83, 4227-4235, 2009.
24. Aras S., Singh G., Johnston K., Foster T., and Aiyar A., PLoS pathogens 5, e1000469, 2009.
25. Nayyar V.K., Shire K., and Frappier L., Journal of cell science 122, 4341-4350, 2009.
26. Shire K., Ceccarelli D.F., Avolio-Hunter T.M., and Frappier L., Journal of virology 73, 2587-2595, 1999.
27. Norseen J., Thomae A., Sridharan V., Aiyar A., Schepers A., and Lieberman P.M., The EMBO journal 27, 3024-3035, 2008.
28. Snudden D.K., Hearing J., Smith P.R., Grasser F.A., and Griffin B.E., EMBO J 13, 4840-4847, 1994.
29. Apcher S., Komarova A., Daskalogianni C., Yin Y., Malbert-Colas L., and Fahraeus R., Journal of virology 83, 1289-1298, 2009.
30. Levitskaya J., Sharipo A., Leonchiks A., Ciechanover A., and Masucci M.G., Proceedings of the National Academy of Sciences of the United States of America 94, 12616-12621, 1997.
31. Yin Y., Manoury B., and Fahraeus R., Science 301, 1371-1374, 2003.
32. Voo K.S., Fu T., Wang H.Y., Tellam J., Heslop H.E., Brenner M.K., Rooney C.M., and Wang R.F., The Journal of experimental medicine 199, 459-470, 2004.
33. Bochkarev A., Barwell J.A., Pfuetzner R.A., Bochkareva E., Frappier L., and Edwards A.M., Cell 84, 791-800, 1996.
34. Bochkarev A., Bochkareva E., Frappier L., and Edwards A.M., Journal of molecular biology 284, 1273-1278, 1998.
35. Cruickshank J., Shire K., Davidson A.R., Edwards A.M., and Frappier L., The Journal of biological chemistry 275, 22273-22277, 2000.
36. Holowaty M.N., and Frappier L., Biochemical Society transactions 32, 731-732, 2004.
37. Sivachandran N., Cao J.Y., and Frappier L., Journal of virology 84, 11113-11123, 2010.
38. Wang Y., Finan J.E., Middeldorp J.M., and Hayward S.D., Virology 236, 18-29, 1997.
39. Holowaty M.N., Sheng Y., Nguyen T., Arrowsmith C., and Frappier L., The Journal of biological chemistry 278, 47753-47761, 2003.
40. Malik-Soni N., and Frappier L., Journal of virology 86, 6999-7002, 2012.
41. Wang S., and Frappier L., Journal of virology 83, 11704-11714, 2009.

42. Saridakis V., Sheng Y., Sarkari F., Holowaty M.N., Shire K., Nguyen T., Zhang R.G., Liao J., Lee W., Edwards A.M., Arrowsmith C.H., and Frappier L., Molecular cell 18, 25-36, 2005.
43. Frappier L., Current opinion in virology 2, 733-739, 2012.
44. Magrane M., and Consortium U., Database : the journal of biological databases and curation 2011, bar009, 2011.
45. Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., and Higgins D.G., Nucleic acids research 25, 4876-4882, 1997.
46. Whelan S., and Goldman N., Molecular biology and evolution 18, 691-699, 2001.
47. Tamura K., Peterson D., Peterson N., Stecher G., Nei M., and Kumar S., Molecular biology and evolution, 2011.
48. Eswar N., Webb B., Marti-Renom M.A., Madhusudhan M.S., Eramian D., Shen M.Y., Pieper U., and Sali A., Current protocols in protein science / editorial board, John E Coligan [et al] Chapter 2, Unit 2 9, 2007.
49. Roy A., Kucukural A., and Zhang Y., Nature protocols 5, 725-738, 2010.
50. Levitt M., Journal of molecular biology 226, 507-533, 1992.
51. Fechteler T., Dengler U., and Schomburg D., Journal of molecular biology 253, 114-131, 1995.
52. Wang J., Cieplak, P., Kollman, P.A., J ComputChem 21, 1049-1074, 2000.
53. Labute P., Journal of computational chemistry 29, 1693-1698, 2008.
54. Connolly M.L., Science 221, 709-713, 1983.
55. Linding R., Russell R.B., Neduva V., and Gibson T.J., Nucleic acids research 31, 3701-3708, 2003.
56. Prilusky J., Felder C.E., Zeev-Ben-Mordehai T., Rydberg E.H., Man O., Beckmann J.S., Silman I., and Sussman J.L., Bioinformatics 21, 3435-3438, 2005.
57. Chen V.B., Arendall W.B., 3rd, Headd J.J., Keedy D.A., Immormino R.M., Kapral G.J., Murray L.W., Richardson J.S., and Richardson D.C., Acta crystallographica Section D, Biological crystallography 66, 12-21, 2010.
58. Benkert P., Tosatto S.C., and Schomburg D., Proteins 71, 261-277, 2008.
59. Chen Z., Chen Y.Z., Wang X.F., Wang C., Yan R.X., and Zhang Z., PloS one 6, e22930, 2011.
60. Schneidman-Duhovny D., Inbar Y., Nussinov R., and Wolfson H.J., Nucleic acids research 33, W363-367, 2005.
61. Duellman S.J., Thompson K.L., Coon J.J., and Burgess R.R., The Journal of general virology 90, 2251-2259, 2009.
62. Barwell J.A., Bochkarev A., Pfuetzner R.A., Tong H., Yang D.S., Frappier L., and Edwards A.M., The Journal of biological chemistry 270, 20556-20559, 1995.
63. Ehlers B., Spiess K., Leendertz F., Peeters M., Boesch C., Gatherer D., and McGeoch D.J., The Journal of general virology 91, 630-642, 2010.
64. Perelman P., Johnson W.E., Roos C., Seuanez H.N., Horvath J.E., Moreira M.A., Kessing B., Pontius J., Roelke M., Rumpler Y., Schneider M.P., Silva A., O'Brien S.J., and Pecon-Slattery J., PLoS genetics 7, e1001342, 2011.
65. Huang Y.J., Mao B., Aramini J.M., and Montelione G.T., Proteins 82 Suppl 2, 43-56, 2014.
66. Shire K., Kapoor P., Jiang K., Hing M.N., Sivachandran N., Nguyen T., and Frappier L., Journal of virology 80, 5261-5272, 2006.
67. Liu J., Faeder J.R., and Camacho C.J., Proceedings of the National Academy of Sciences of the United States of America 106, 19819-19823, 2009.
68. Tompa P., Current opinion in structural biology 21, 419-425, 2011.

69. Babu M.M., van der Lee R., de Groot N.S., and Gsponer J., Current opinion in structural biology 21, 432-440, 2011.
70. Tsvetkov P., Reuven N., and Shaul Y., Nature chemical biology 5, 778-781, 2009.
71. Gsponer J., Futschik M.E., Teichmann S.A., and Babu M.M., Science 322, 1365-1368, 2008.
72. Tellam J., Rist M., Connolly G., Webb N., Fazou C., Wang F., and Khanna R., European journal of immunology 37, 328-337, 2007.
73. Tellam J., Sherritt M., Thomson S., Tellam R., Moss D.J., Burrows S.R., Wiertz E., and Khanna R., The Journal of biological chemistry 276, 33353-33360, 2001.
74. Fujiwara K., Toda H., and Ikeguchi M., BMC structural biology 12, 18, 2012.

Figure 1
Click here to download high resolution image

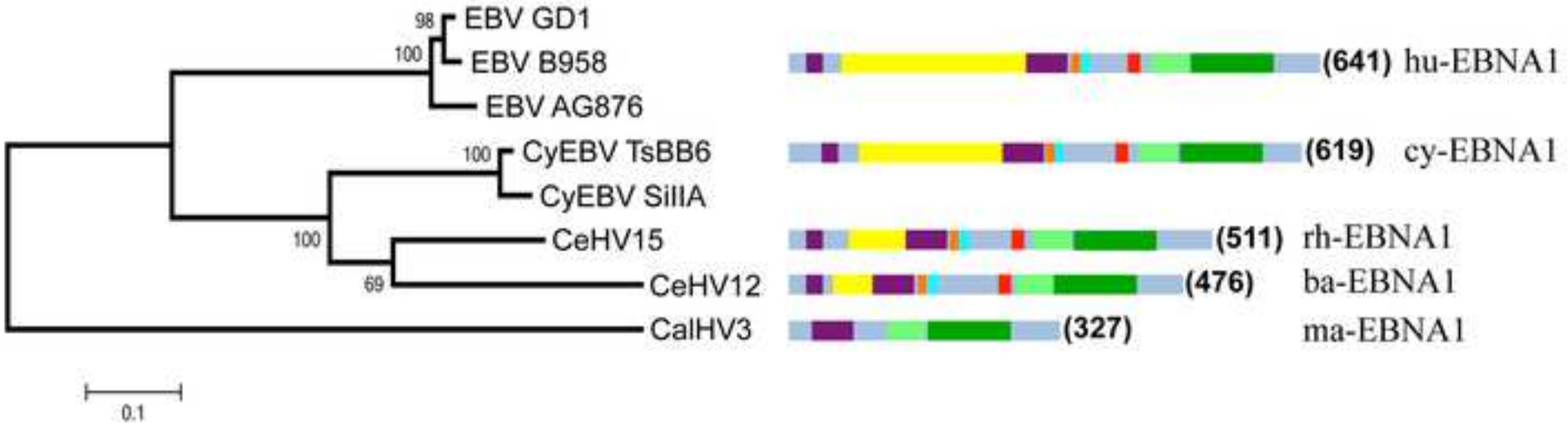


Figure 2

Click here to download high resolution image

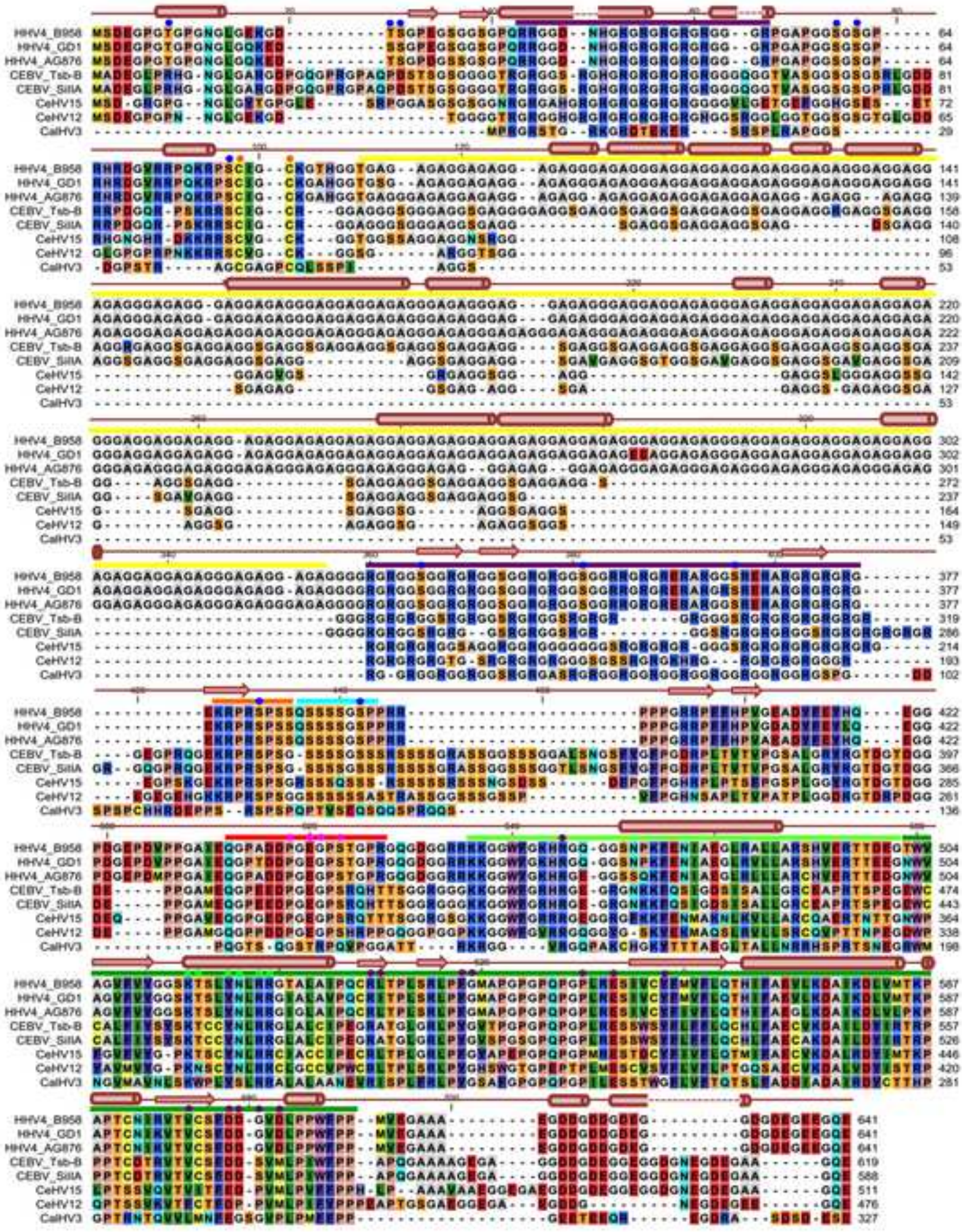


Figure 3
Click here to download high resolution image

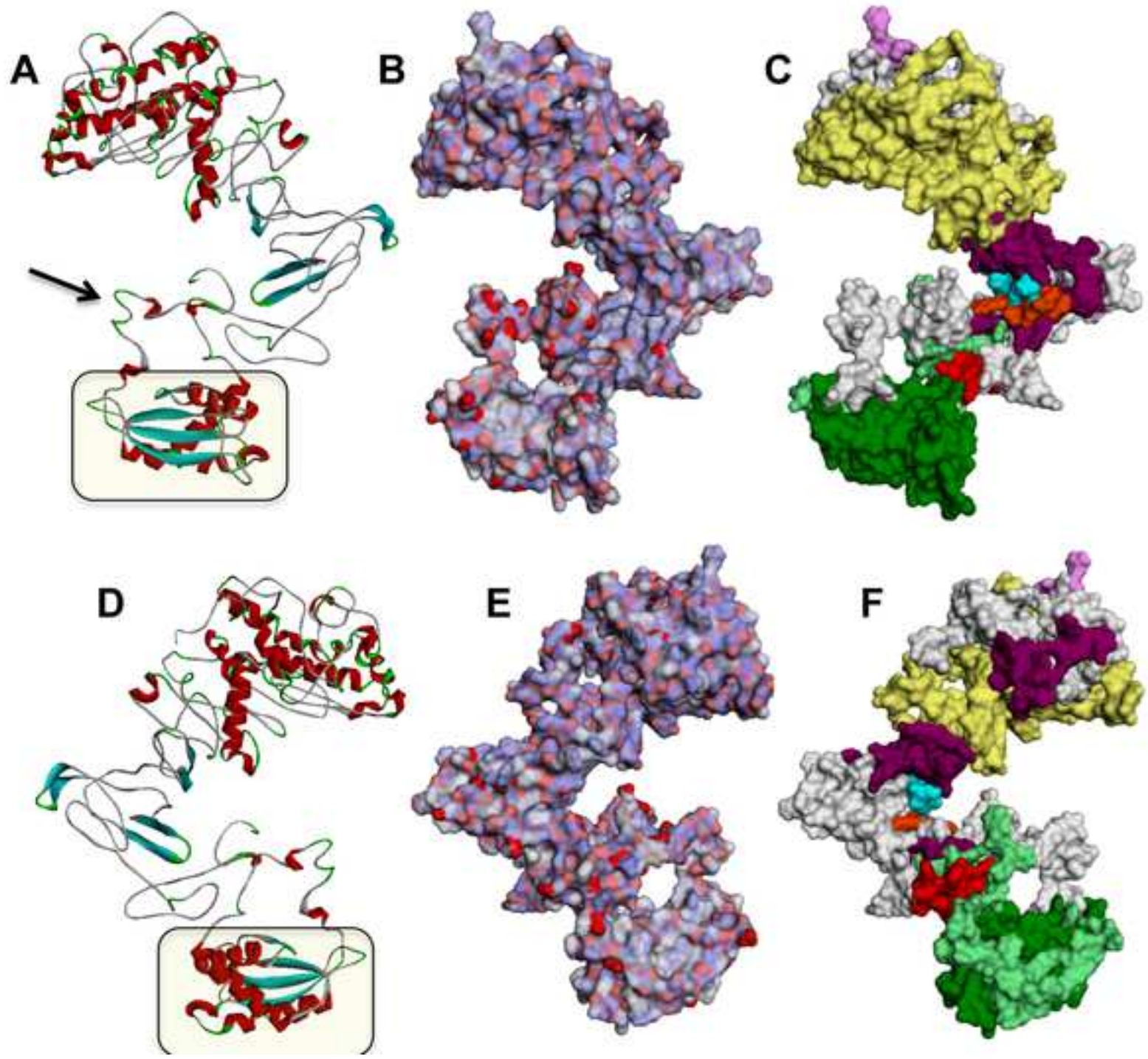


Figure 4
Click here to download high resolution image

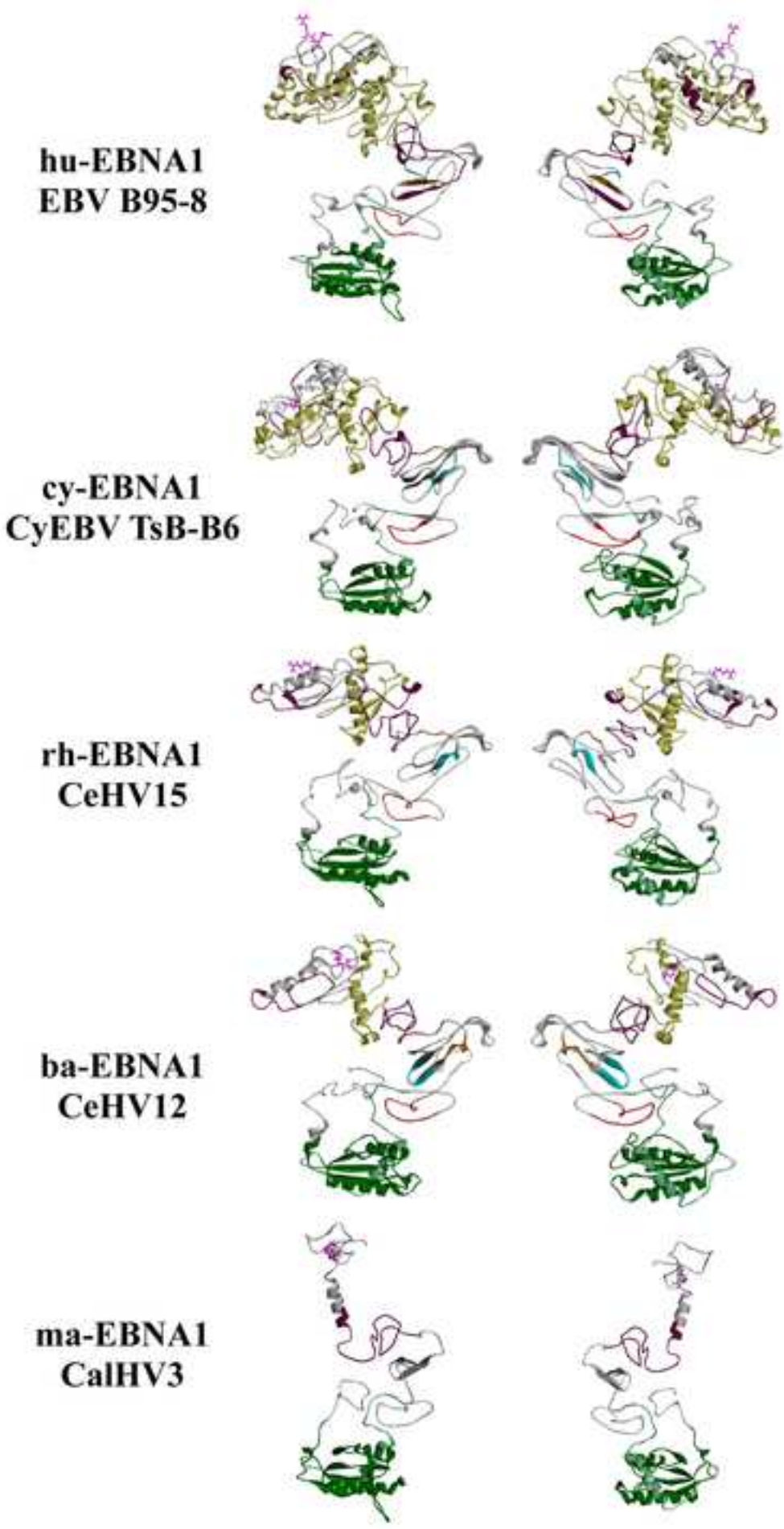


Figure 5
Click here to download high resolution image

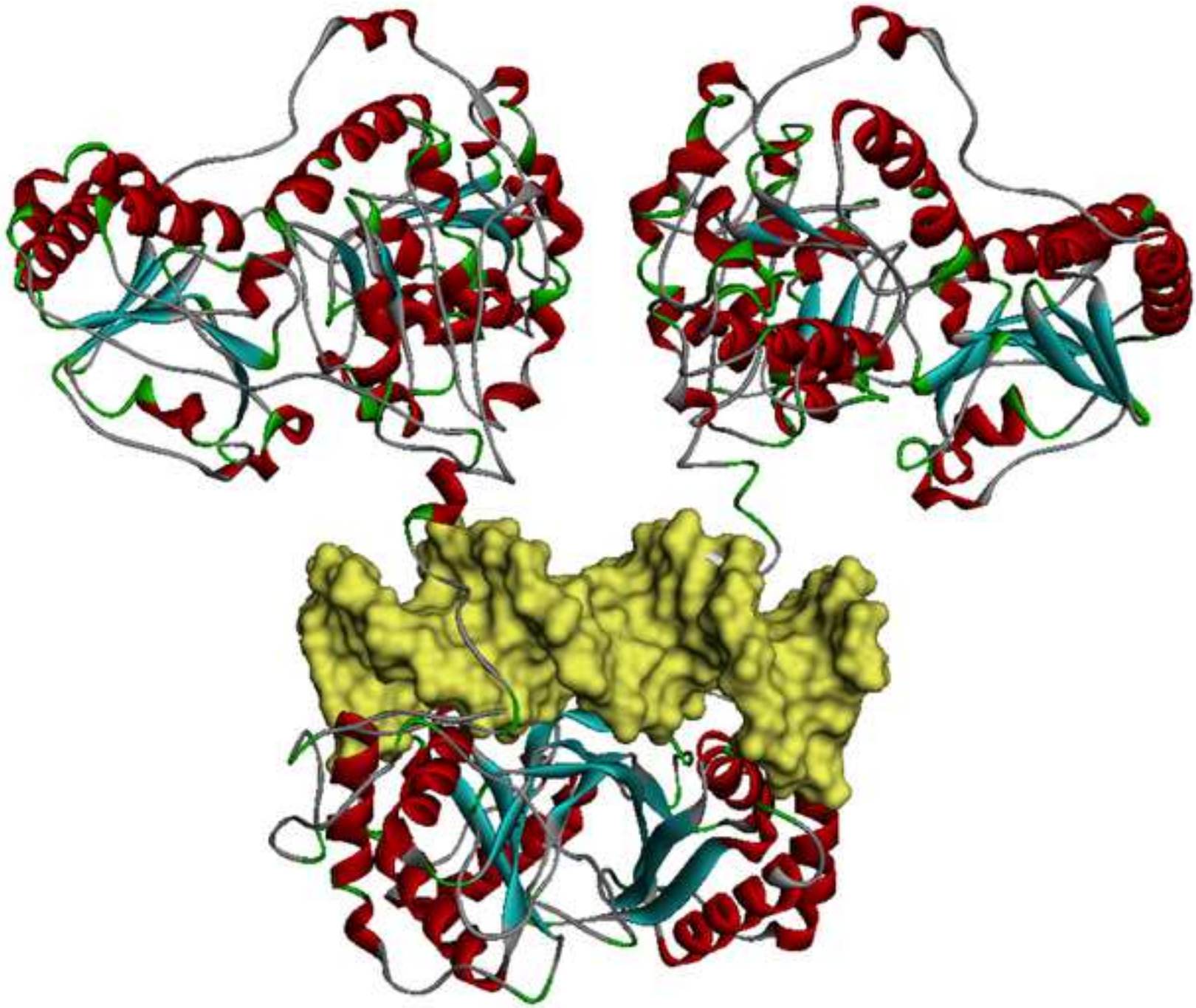


Figure 6
Click here to download high resolution image

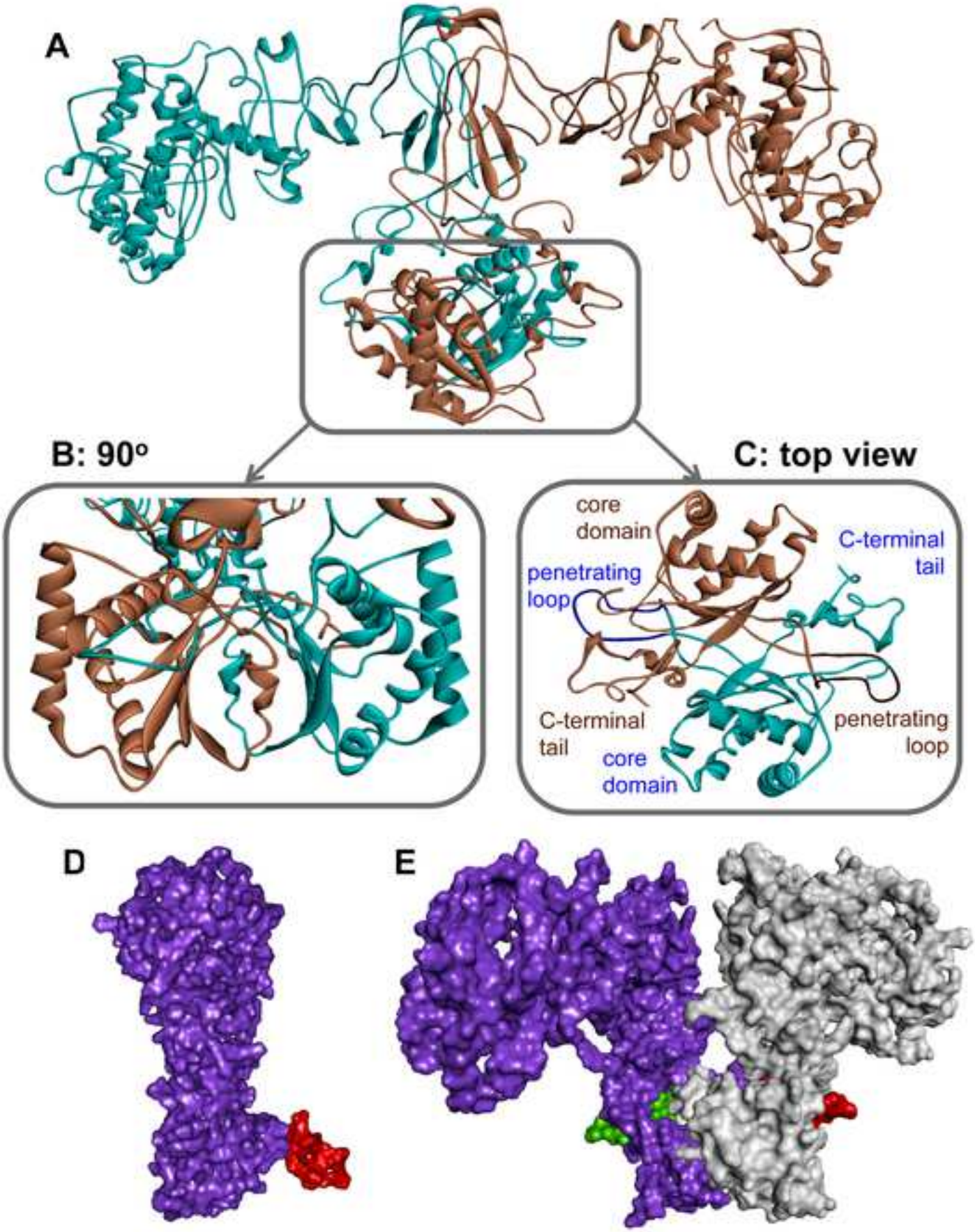


Figure 7
Click here to download high resolution image

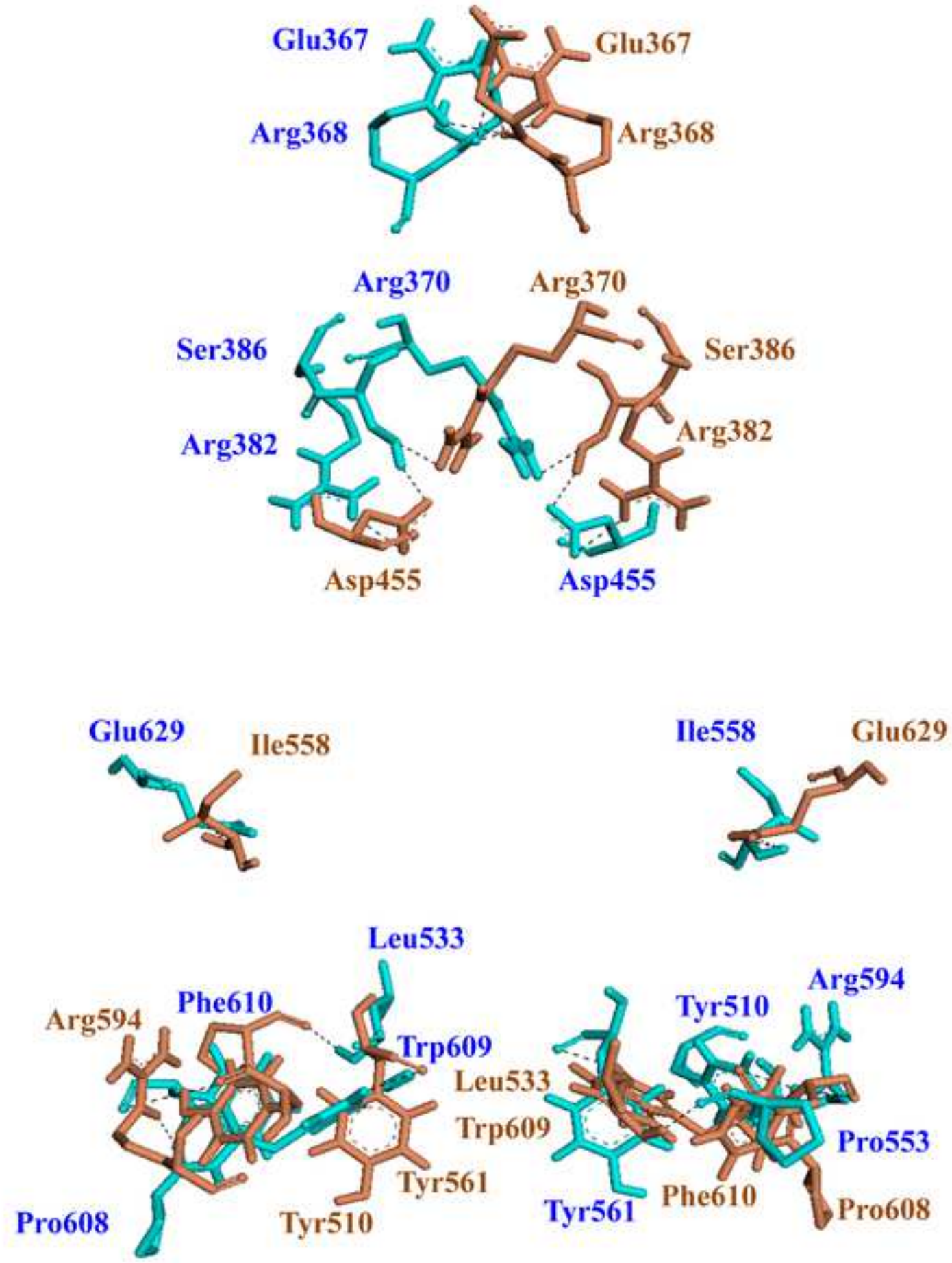
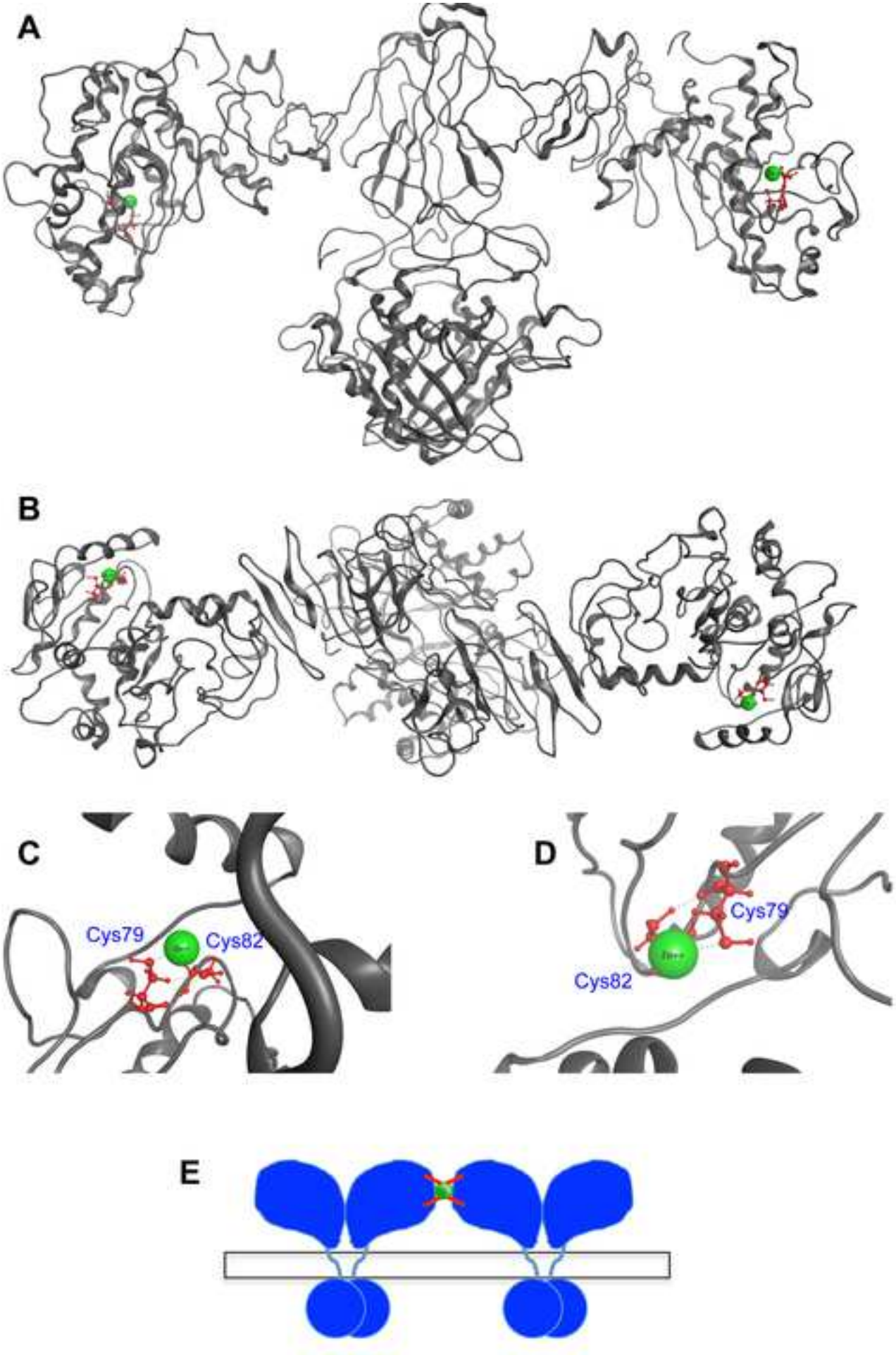


Figure 8
Click here to download high resolution image



Supplementary Material

Click here to download Supplementary Material: Supplementary Figures.pdf

model 1
Click here to download Supplementary Material: Model 1.pdb

model 2

Click here to download Supplementary Material: Model 2.pdb

model 3
Click here to download Supplementary Material: Model 3.pdb

model 4
Click here to download Supplementary Material: Model 4.pdb

model 5
Click here to download Supplementary Material: Model 5.pdb

model 6
Click here to download Supplementary Material: Model 6.pdb

model 7
Click here to download Supplementary Material: Model 7.pdb

model 8
Click here to download Supplementary Material: Model 8.pdb

model 9
Click here to download Supplementary Material: Model 9.pdb

model 10
Click here to download Supplementary Material: Model 10.pdb

model 11
Click here to download Supplementary Material: Model 11.pdb

model 12
Click here to download Supplementary Material: Model 12.pdb

model 13
Click here to download Supplementary Material: Model 13.pdb