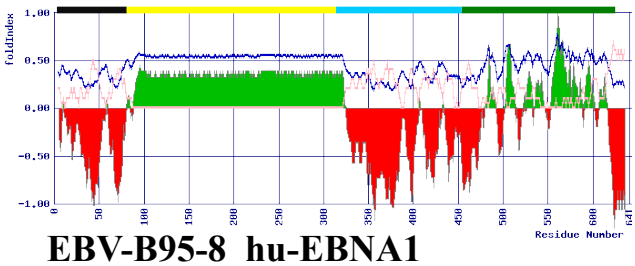
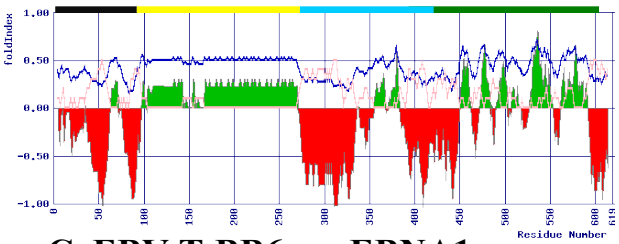
Modelling the structure of full length Epstein-Barr Virus Nuclear Antigen 1

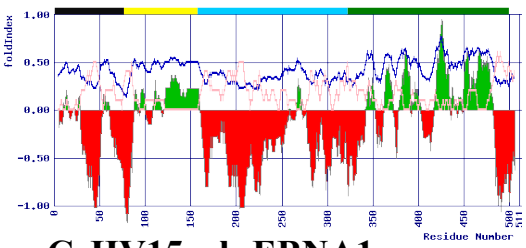
Hussain, Gatherer and Wilson **Supplemental Information**



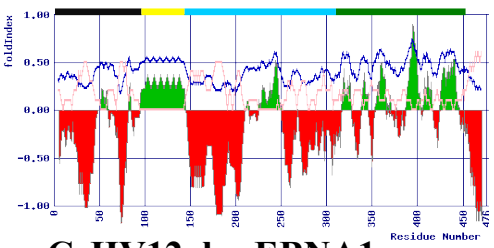
EBV-B95-8 hu-EBNA1



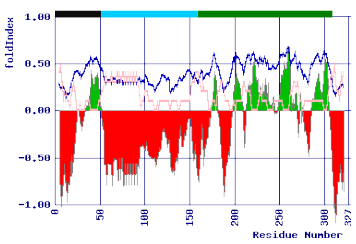
CyEBV-TsBB6 cy-EBNA1



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.

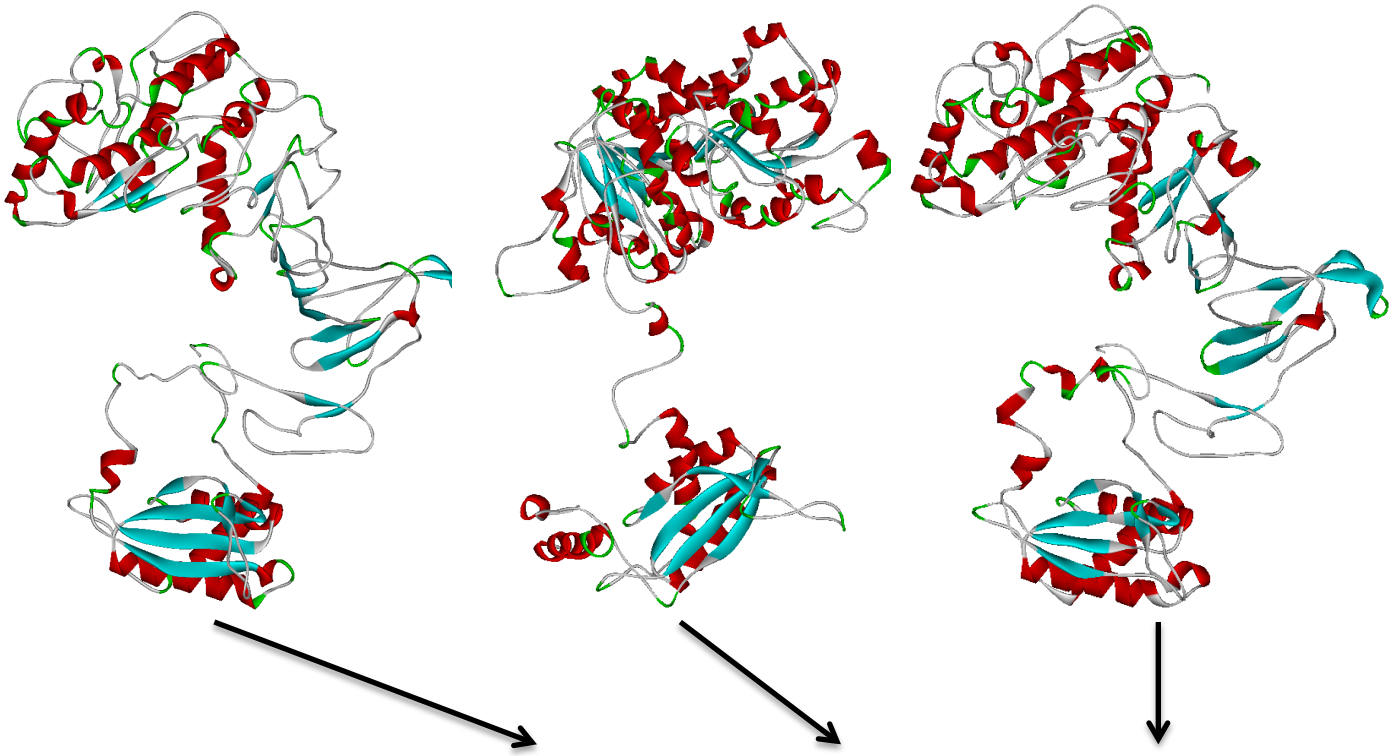
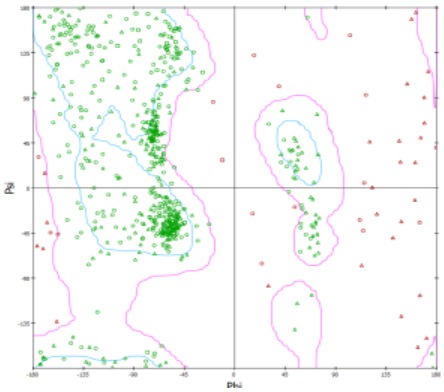
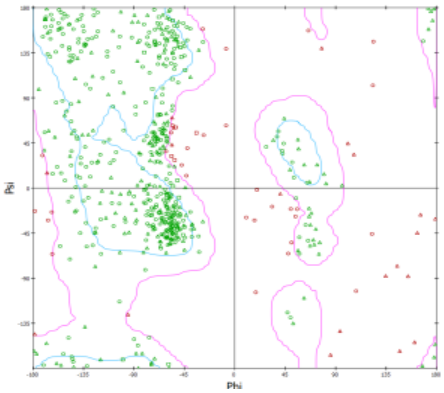


Table S1	ITASSER	MOE	Composite
Ramachandran plot outliers	7.98%	4.85%	4.07%
Ramachandran plot favoured region	73.1%	80.6%	89.5%
RMSD with 1B3T	0.4Å	1.05Å	1.29Å
Bad bonds	0	0.31%	0
Bad angles	1.09%	1.72%	0.62%
QMEANnorm score	0.14	0.09	0.15

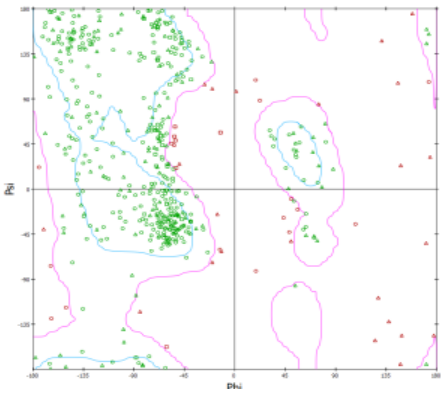
Supplementary Figure S2. EBNA1 model comparison.
The EBV B95-8 EBNA1 sequence was input to I-TASSER, which selected the EBNA1 C-terminal domain crystal structure (1B3T) and several fragment templates comprising: yeast fatty acid synthetase (2PFF), a-L-fucosidase (2Z8X), photosynthetic reaction centre (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/)



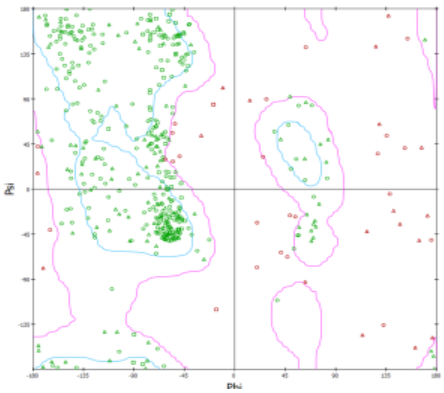
hu-EBNA1



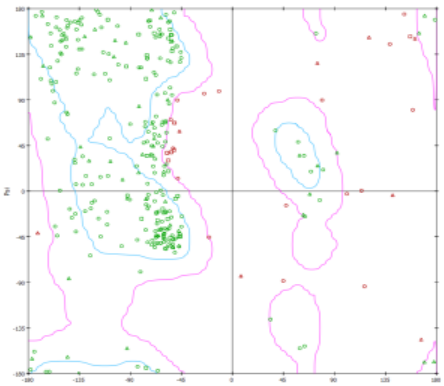
cy-EBNA1



rh-EBNA1

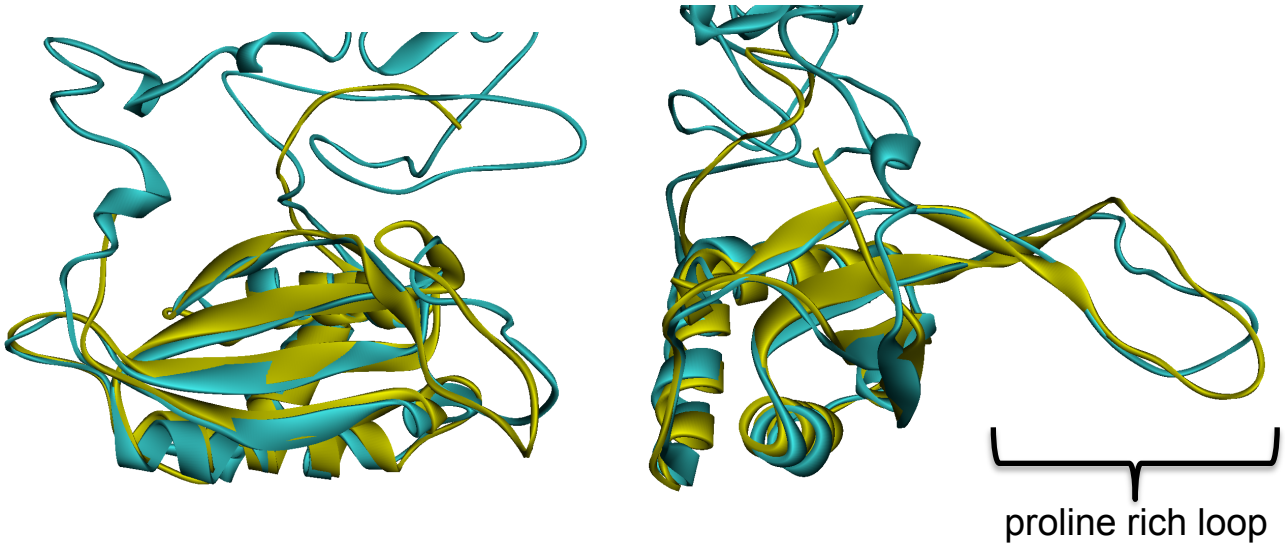


ba-EBNA1



ma-EBNA1

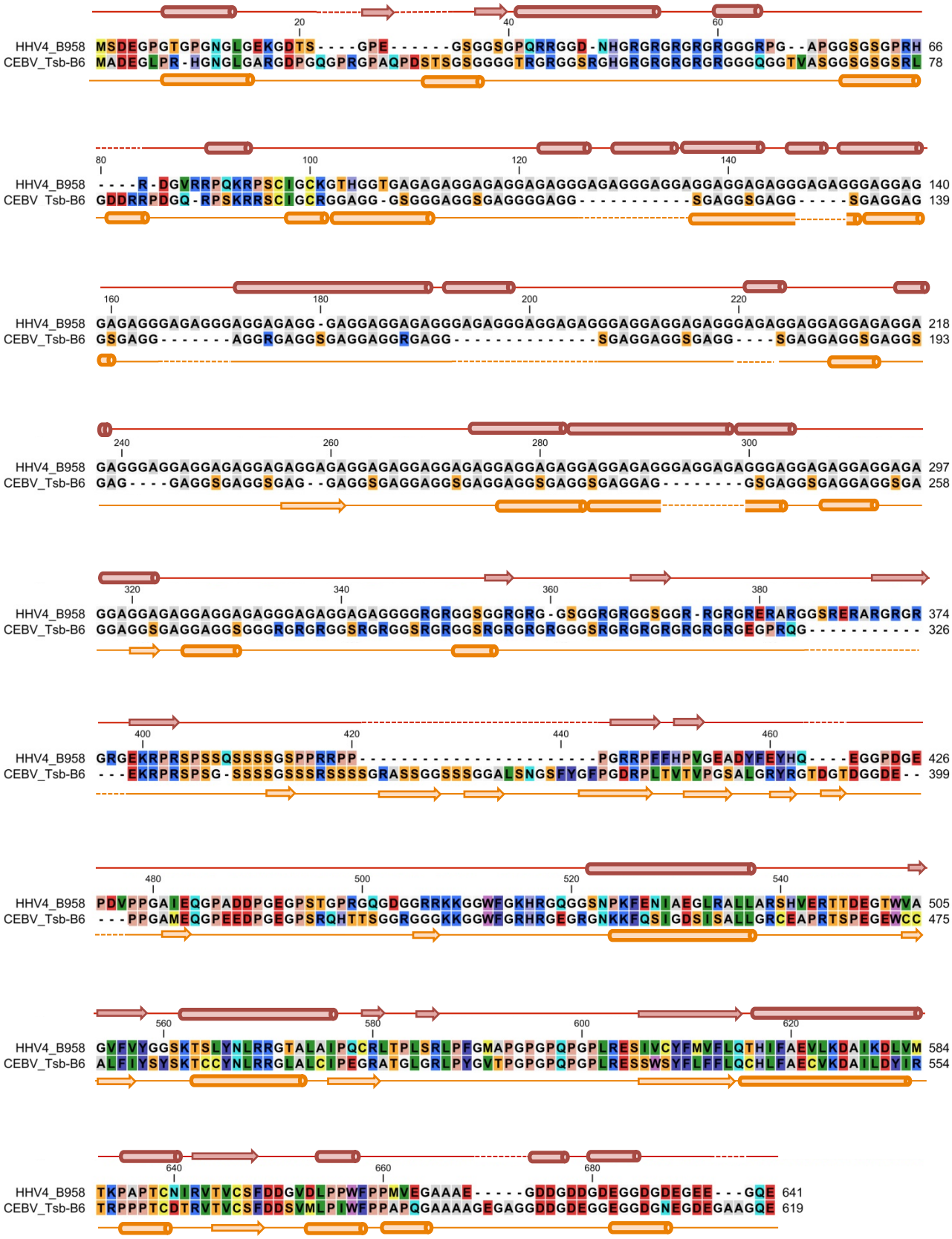
Supplementary Figure S3. Ramachandran Plots of EBNA1 modeled structures.
The human and other primate LCV EBNA1 protein structure models (as indicated) were evaluated for dihedral bond angle (Phi and Psi) distribution using Ramachandran plots. Residues in allowed and disallowed regions are represented 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.

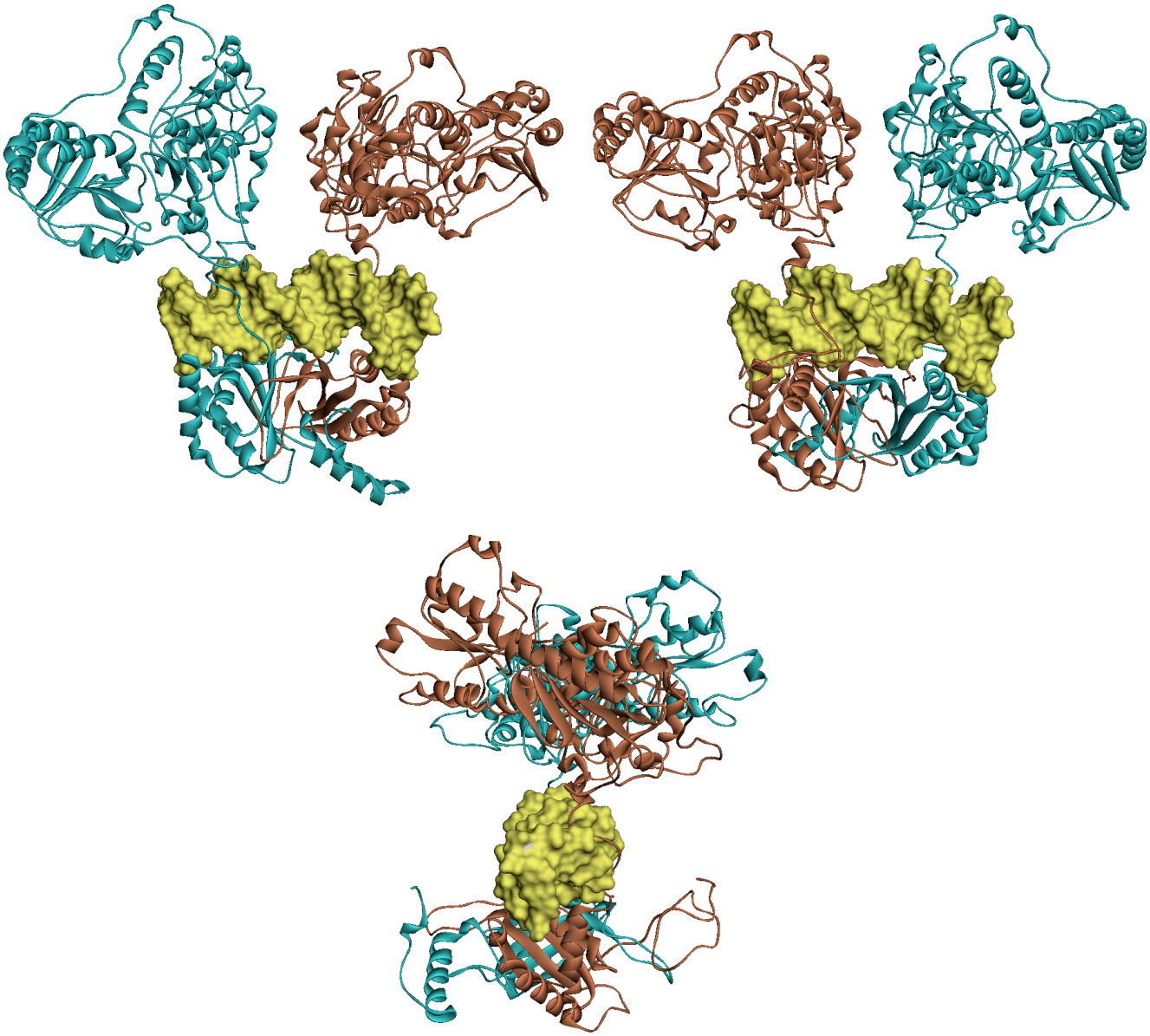


Supplementary Figure S4. Comparison of the C-terminal region of the EBNA1 composite model and the resolved structure.
The C-terminal region of the composite EBNA1 model (cyan) is shown superimposed over a monomer extracted from the resolved template 1B3T (yellow), from two angles (with a horizontal rotation 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





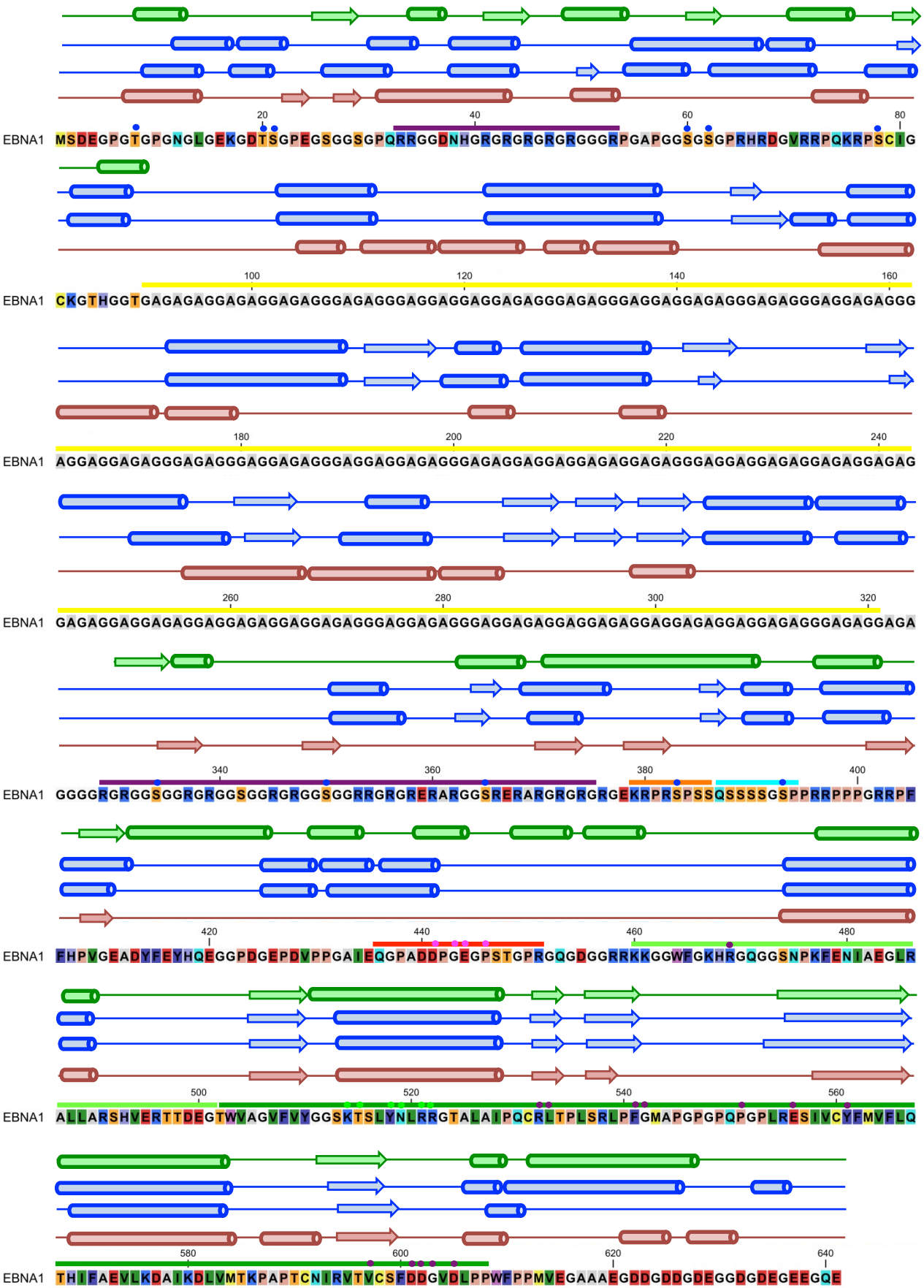
Supplementary Figure S6. EBNA1 MOE model bound with DNA.
The hu-EBNA1 dimer model developed using MOE (using spatial constraints for DNA binding) is shown. Each monomer is represented in ribbon format (brown and cyan) while DNA is shown in surface format (yellow). Since the C-terminal region of the dimer 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).

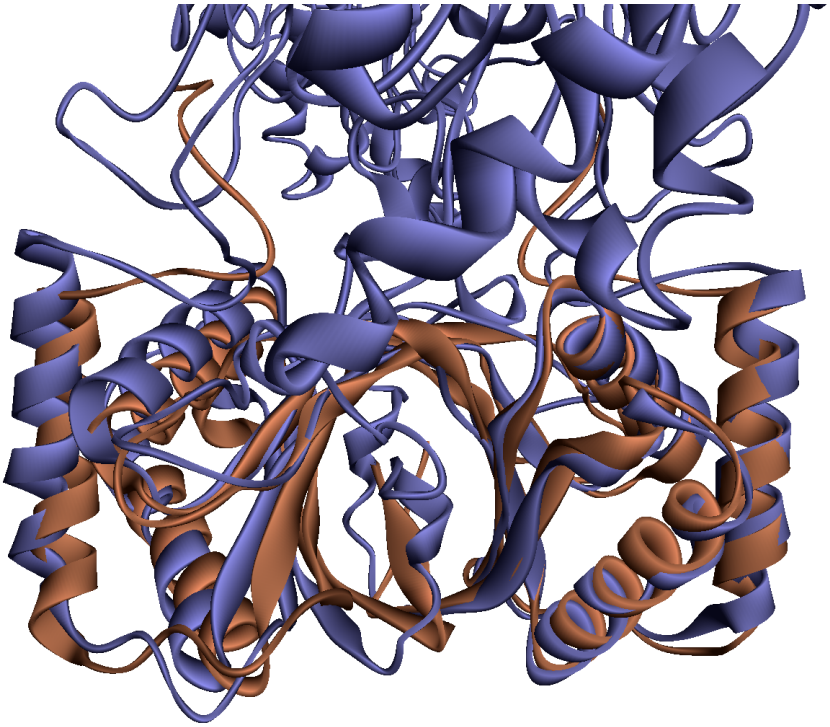
Dimer Model: feature	Full length	Tail deleted
GCS	17890	13654
ACE	-1062.7	-398
G	-11094.65	-9562

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).

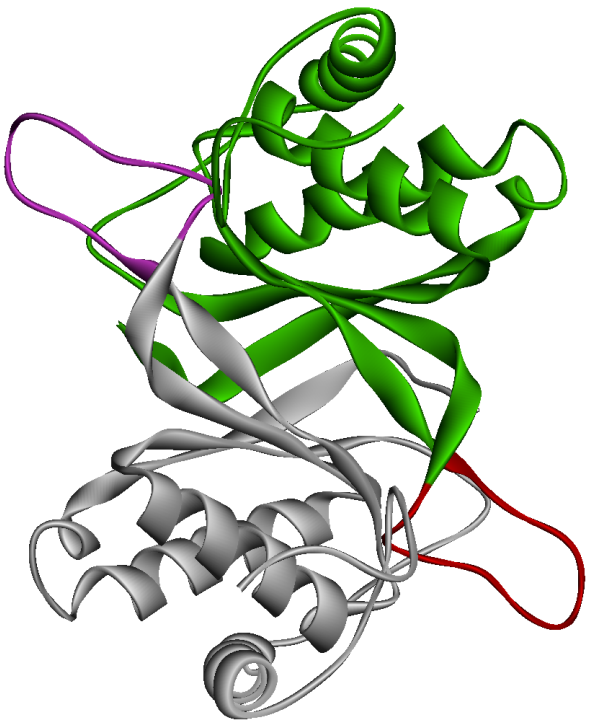
Legend to Supplementary Figure S7. Secondary structure distribution of EBNA1 models.
The distribution of the secondary structural elements of the different hu-EBNA1 models is shown above the primary sequence of EBV B95-8 EBNA1 (as assessed by HERA plot). The composite model (maroon), each MOE monomer 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

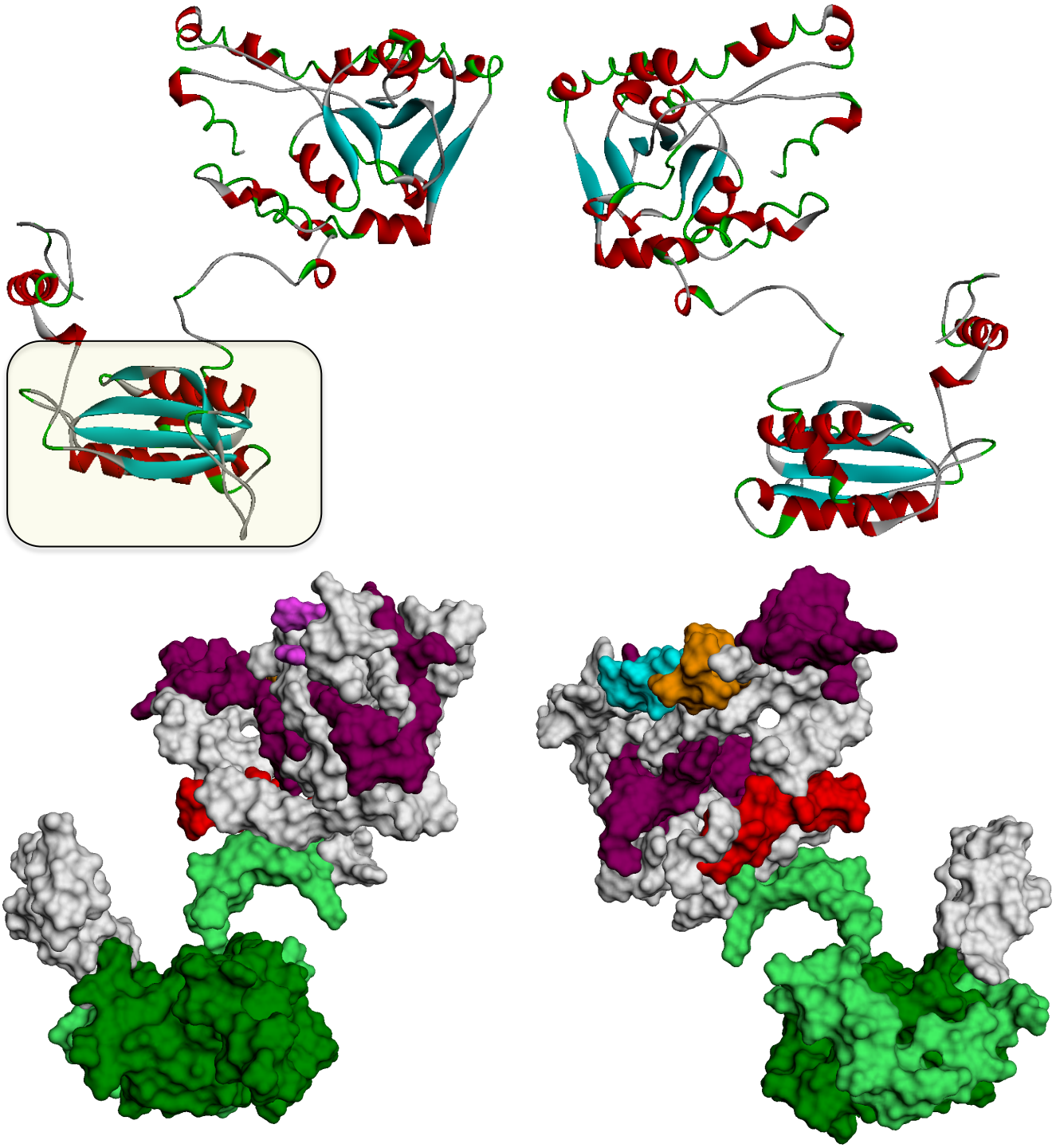




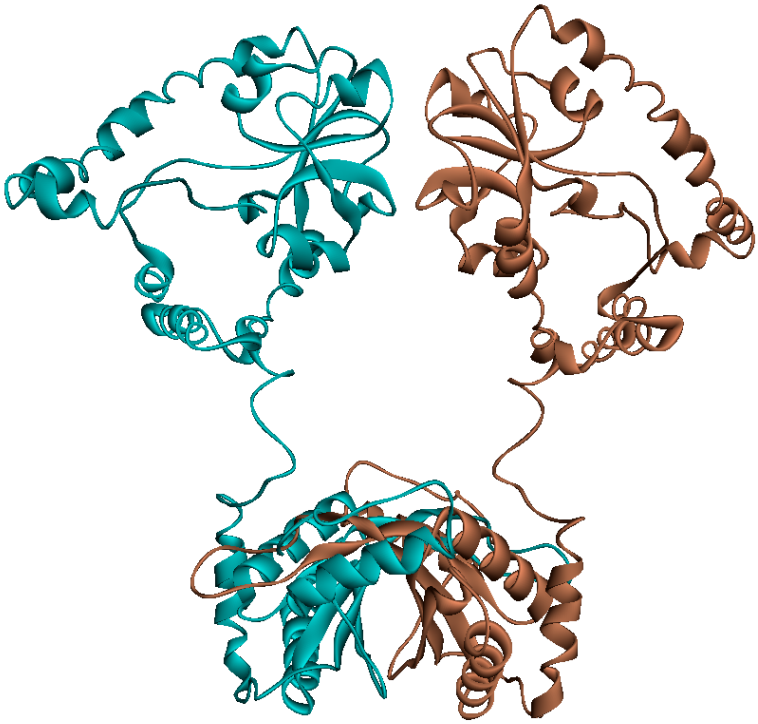
Supplementary Figure S8. Comparison of composite dimer EBNA1 model with the resolved structure.
The composite EBNA1 dimer model (blue) is shown superimposed over the resolved structure 1B3T (brown) (RMSD value 1.5 Å). One side of the β barrel (right side from the viewed angle) was used as the reference 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).

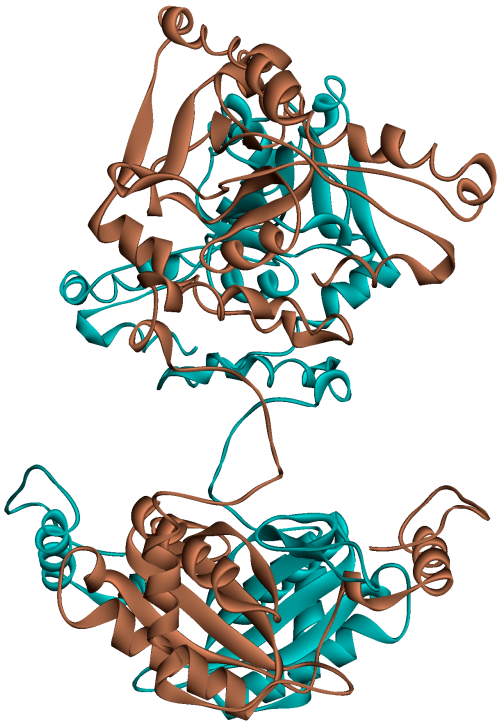


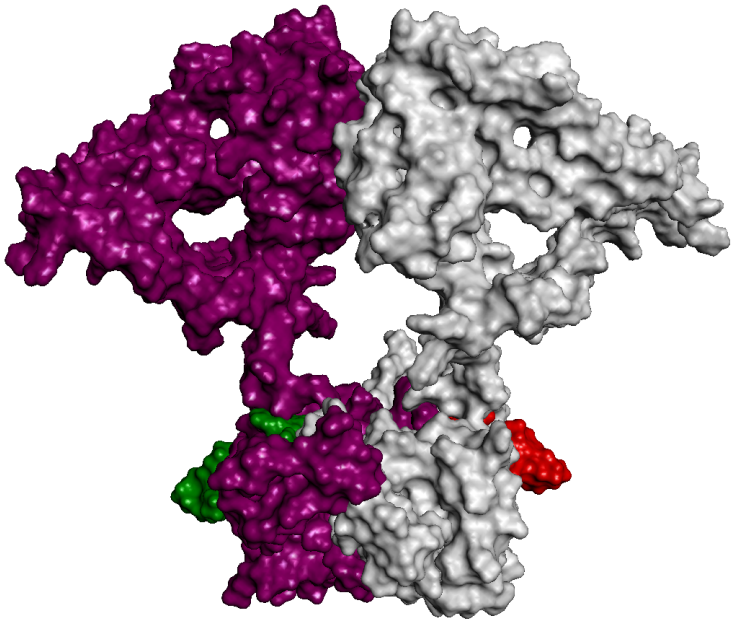
Supplementary Figure S9. The proline rich loops in the dimer.
"Top" view of the resolved EBNA1 1B3T dimer showing the protruding proline rich loops. Each monomer/loop is differently coloured green/red and silver/mauve.



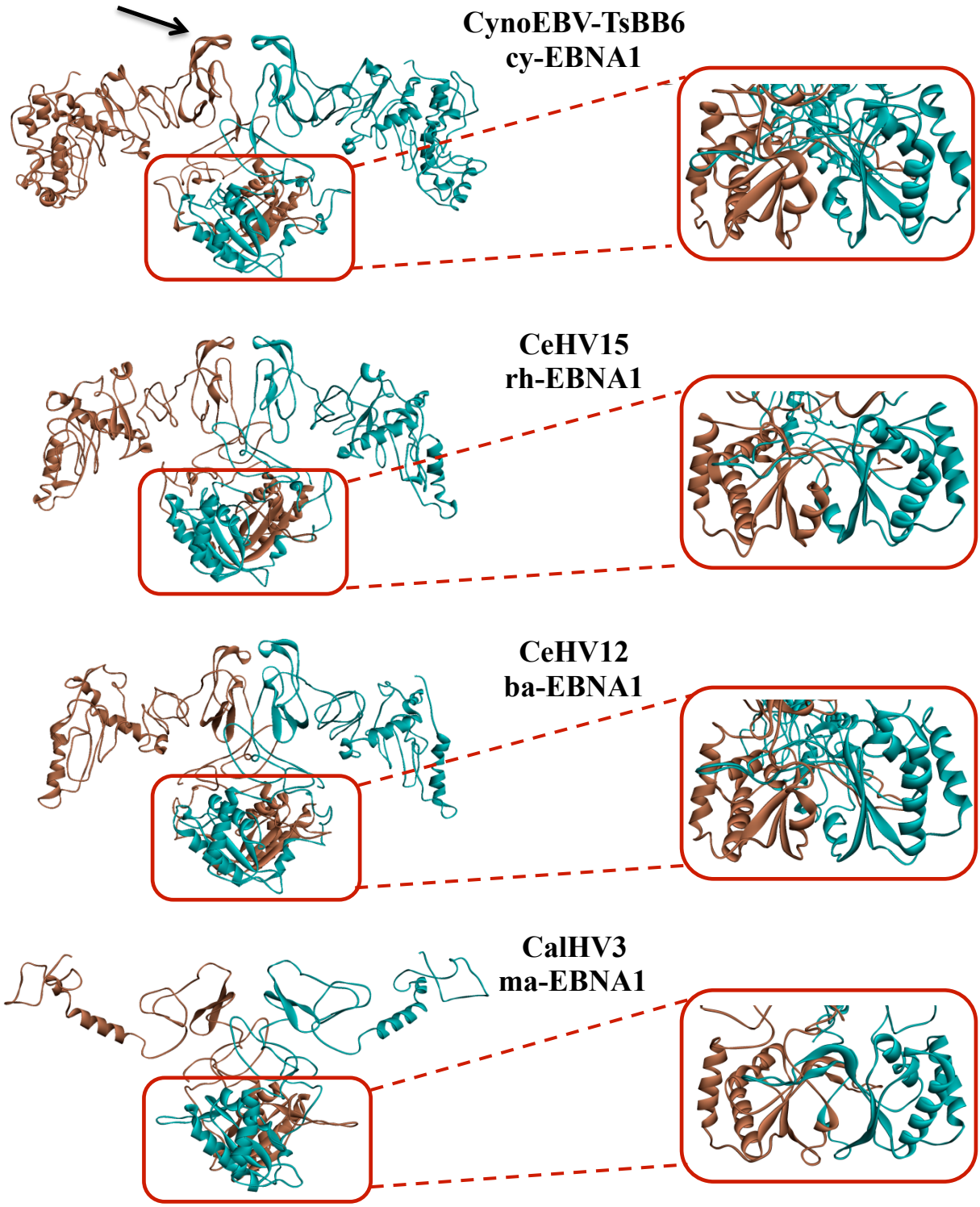
Supplementary Figure S10. GAr deleted EBNA1 Model.
Deletion of the GAr of hu-EBNA1 allows increased expression of the protein in heterologous systems and retains several of the protein's functions. A composite monomeric model of GAr-deleted hu-EBNA1 sequences was generated as for full length (using I-TASSER, 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.