

Structure

A comparative analysis of the foamy and ortho virus capsid structures reveals an ancient domain duplication --Manuscript Draft--

Manuscript Number:	
Full Title:	A comparative analysis of the foamy and ortho virus capsid structures reveals an ancient domain duplication
Article Type:	Research Article
Keywords:	Virus capsid structure; foamy virus evolution; protein structure comparison
Corresponding Author:	William R Taylor Crick Institute London, UNITED KINGDOM
First Author:	William R Taylor
Order of Authors:	William R Taylor Jonathan P Stoye Ian A Taylor
Abstract:	<p>The Spumaretrovirinae (foamy viruses) and the Orthoretrovirinae (e.g. HIV) share many similarities both in genome structure and the sequences of the core viral encoded proteins, such as the aspartyl protease and reverse transcriptase. Similarity in the Gag region of the genome is less obvious at the sequence level but has been illuminated by the recent solution of the foamy virus capsid (CA) structure. This revealed a clear structural similarity to the orthoretrovirus capsids but with marked differences that left uncertainty in the relationship between the two domains that comprise the structure. Using multiple structure comparison methods combined with statistical tests, we have shown that the relationship of the two domains conforms to a simple linear correspondence rather than a domain transposition. These similarities suggest that the origin of both viral capsids was a common ancestor with a double domain structure. In addition, we show that there is also a significant structural similarity between the amino and carboxy domains in both the foamy and ortho viruses which suggests that there may have been an even more ancient gene-duplication that preceded the double domain structure.</p>
Suggested Reviewers:	<p>Mark Gerstein Mark.Gersteini@Yale.edu structure comparison (stats)</p> <p>Mark Johnson johnson4@abo.fi structure comparison (previously on retroviruses)</p> <p>Liisa Holm liisa.holm@helsinki.fi structure comparison (DALI method)</p> <p>Peer Bork Peer.Bork@EMBL.DE retrovirus and Arc/Ty analysis</p> <p>Alexey Murzin</p>

	agm@mrc-lmb.cam.ac.uk structure comparison (SCOP database)
	David Stuart stuart-pa@strubi.ox.ac.uk virus structure
Opposed Reviewers:	

Dear Structure,

We submit the our paper on retroviral capsid comparison which shows structural evidence for early gene duplications. We believe this observation will be of considerable interest to many researchers in the retroviral field and those with a wider interest in viral evolution.

Yours,

Willie Taylor and co-authors

A comparative analysis of the foamy and ortho virus capsid structures reveals an ancient domain duplication

William R. Taylor^a, Jonathan P. Stoye^b, Ian A. Taylor^c

Francis Crick Institute, 1 Midland Rd., London NW1 1AT, UK

^a*Computational Cell and Molecular Biology,*

^b*Retrovirus-Host Interactions,*

^c*Macromolecular Structure Laboratory.*

Abstract

The *Spumaretrovirinae* (foamy viruses) and the *Orthoretrovirinae* (e.g. HIV) share many similarities both in genome structure and the sequences of the core viral encoded proteins, such as the aspartyl protease and reverse transcriptase. Similarity in the Gag region of the genome is less obvious at the sequence level but has been illuminated by the recent solution of the foamy virus capsid (CA) structure. This revealed a clear structural similarity to the orthoretrovirus capsids but with marked differences that left uncertainty in the relationship between the two domains that comprise the structure. Using multiple structure comparison methods combined with statistical tests, we have shown that the relationship of the two domains conforms to a simple linear correspondence rather than a domain transposition. These similarities suggest that the origin of both viral capsids was a common ancestor with a double domain structure. In addition, we show that there is also a significant structural similarity between the amino and carboxy domains in both the foamy and ortho viruses which suggests that there may have been an even more ancient gene-duplication that preceded the double domain structure.

Keywords: Virus capsid structure, foamy virus evolution, protein structure comparison

1. Introduction

Taxonomically, the *Orthoretrovirinae* (orthoretroviruses) and *Spumaretrovirinae*¹ (spumaviruses) make up the two subfamilies of *Retroviridae*. They share many similarities, including overall genome structures with gag, pol and env genes encoding proteins for replication and life cycles involving reverse transcription and integration into the chromosomes of infected cells. However, there are also a number of differences distinguishing these viral subfamilies, including finer details of genome organisation, the absence of a Gag-Pol fusion protein in spumaviruses and the timing of reverse transcription [1].

Gag is the major structural protein of both Ortho and Foamy viruses and is responsible for many of the differences and similarities between the viral subfamilies. Ortho and Foamy viral Gags are required for particle assembly, budding from the cell, reverse transcription and delivery of the viral nucleic acid into the newly infected cell. However, there are a number of striking differences including how the Gag precursor is targeted to the cell membrane, the absence of a Major Homology Region and Cys-His box in Foamy viruses and very different patterns of processing during viral maturation [2]. In all Ortho viruses, Gag is proteolytically cleaved to form distinct, well-studied proteins, matrix (MA), capsid (CA) and nucleocapsid (NC), found in mature virions, whilst in spumaviruses Gag processing to remove a C-terminal peptide occurs only in a fraction of the Gag molecules [3].

Structural information regarding foamy virus Gag has been limited to the crystal structure of the N-terminal Env binding region of Prototypic Foamy virus (PFV) Gag (PFV-Gag-NtD) that although maintaining some of the function of orthoretrivial MA shared no structural similarity [4]. However, more recently the solution NMR structure of the PFV Gag central CA domains has shed new light on the relationship between ortho and spumaviruses. It reveals that the CA structures of both viral subfamilies share a common protein fold, implying that their Gag proteins may be evolutionarily related [5].

However, an intriguing aspect of this relationship was an ambiguity in the degree of relatedness between the CA domains of the Gag proteins, with

¹This class is also commonly referred to as the Foamy viruses (after the morphological effect they have on infected cells) and will be referred by this name frequently below, with the term orthoretroviruses also contracted to "Ortho viruses".

34 the Spumaretroviral CA domains, NtDCEN and CtDCEN, appearing almost
35 equally similar to either the amino- (CA-NtD) or carboxy-terminal (CA-CtD)
36 domains of the orthoretroviruses. In this paper, we now investigate and clar-
37 ify the nature of this relationship and discuss its evolutionary implications.

38 2. Results

39 2.1. Full-length comparison

40 To investigate the structural relationship between the capsid structure of
41 the ortho viruses (HIV, MLV, etc.), and the new structure of the foamy virus
42 capsid [5] (PDB codes: 5m1g, 5m1h), the foamy virus structure was com-
43 pared to one of the few full double domain ortho virus structures, the HIV
44 capsid with PDB code: 3nte, using the flexible superposition program SAP
45 [6]. Even though this program has a tolerant approach to relative domain
46 shifts, the comparison produced a high RMSD value of 14Å over the 100 best
47 superposed positions. The amino (N) terminal domain positions roughly cor-
48 responded but shifts in the relative orientation of the carboxy (C) terminal
49 domain resulted in large deviations between equivalent helices. The super-
50 posed structures are shown in Figure 1(a) and the domain divergence can be
51 seen clearly as a jump in the cumulative RMSD plot (Figure 1(b)).

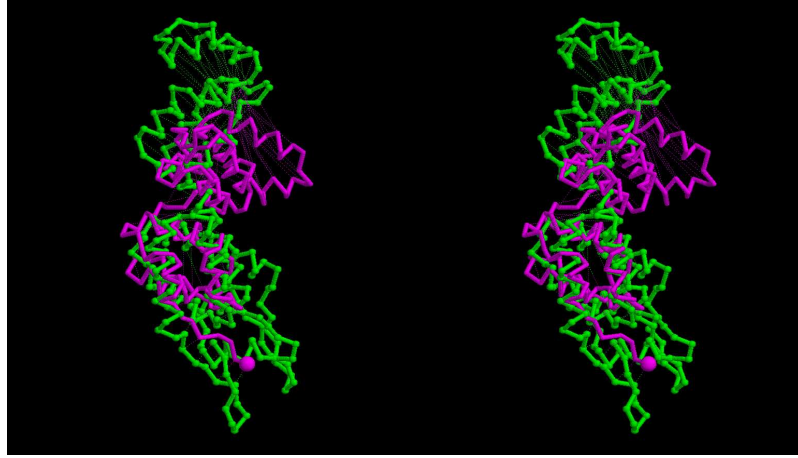
52 2.2. DALI searches

53 Although this initial superposition (Figure 1) did not appear encouraging,
54 the foamy virus structure was scanned across the Protein DataBank (PDB),
55 using the DALI program [7] to search for any similarities.

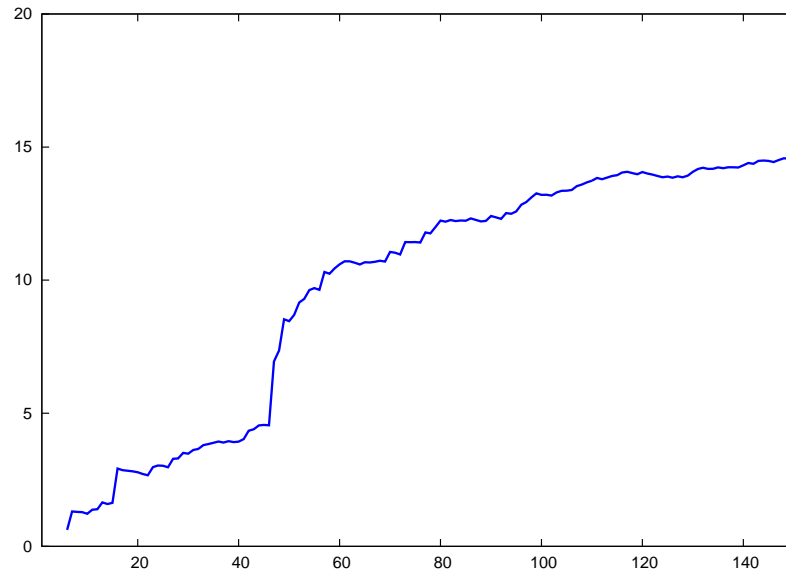
56 2.2.1. Full chain scan

57 A scan of the full-length foamy structure using the DALI server² over the
58 90% non-redundant protein structure databank identified a wide selection of
59 retroviral capsid structures. In the ranked list of structure hits, capsids were
60 identified from position 2 to position 550. The top hits are shown in Figure 2
61 (See Supplementary material for a summary of the full 550 with Z-scores over
62 2). Many capsids are found in the top 20 hits and although the top scoring
63 hit is not obviously a capsid protein, it is thought to have originated from

²http://ekhidna.biocenter.helsinki.fi/dali_server, see Methods section for details.



(a)



(b)

Figure 1: **Full ortho/foamy virus capsid superposition.** The superposed structures are shown in part (a) as a stereo pair, coloured as green = ortho virus (HIV, PDB code: **3nte-A**) and magenta = foamy virus capsid. (The amino terminus is marked by a small sphere). Part (b) shows the cumulative RMSD plot for this superposition which plots the RMSD value (Y-axis) for increasingly larger sets of residues as ranked by their **SAP** similarity score (X-axis). The sharp rise in this trace marks the transition into subsets that include positions from the displaced domain.

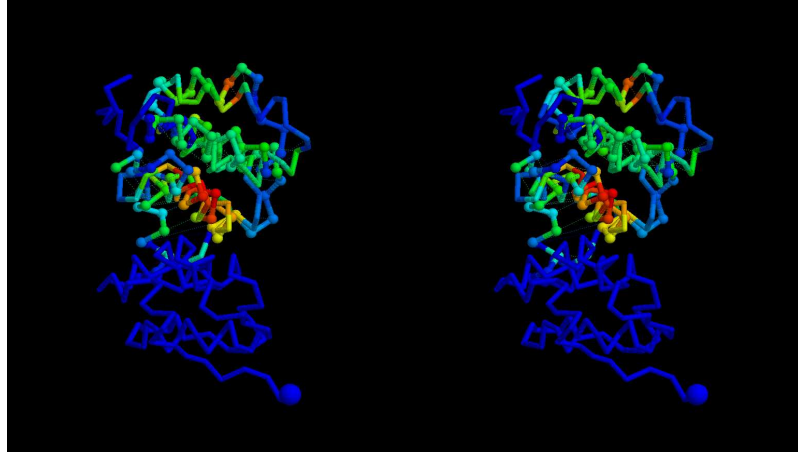
No:	Chain	Z	rmsd	lali	nres	%id	PDB	Description
1:	4x3x-A	5.0	3.1	66	82	11	PDB	MOLECULE: ACTIVITY-REGULATED CYTOSKELETON-ASSOC
2	3g29-A	3.7	2.7	60	77	8	PDB	MOLECULE: GAG POLYPROTEIN;
3	3g0v-A	3.7	2.9	62	76	8	PDB	MOLECULE: GAG POLYPROTEIN;
4:	2v50-D	3.6	2.2	41	998	7	PDB	MOLECULE: MULTIDRUG RESISTANCE PROTEIN MEXB;
5:	3j39-i	3.6	2.5	40	113	3	PDB	MOLECULE: 60S RIBOSOMAL PROTEIN L10A-2;
6	4ph2-A	3.6	3.2	69	127	7	PDB	MOLECULE: BLV CAPSID - N-TERMINAL DOMAIN;
7:	1iqp-E	3.6	3.8	69	326	7	PDB	MOLECULE: RFCS;
8:	4gco-A	3.6	3.7	55	120	11	PDB	MOLECULE: PROTEIN STI-1;
9	3g29-B	3.6	2.8	62	77	8	PDB	MOLECULE: GAG POLYPROTEIN;
10	3g1i-B	3.6	2.9	62	75	8	PDB	MOLECULE: GAG POLYPROTEIN;
11	3g21-A	3.6	2.8	60	77	8	PDB	MOLECULE: GAG POLYPROTEIN;
12:	2a0u-A	3.5	3.1	68	374	4	PDB	MOLECULE: INITIATION FACTOR 2B;
13:	1j7q-A	3.5	2.9	60	86	5	PDB	MOLECULE: CALCIUM VECTOR PROTEIN;
14:	2a0u-B	3.5	8.1	80	367	4	PDB	MOLECULE: INITIATION FACTOR 2B;
15:	1iqp-A	3.5	3.7	70	326	7	PDB	MOLECULE: RFCS;
16	4ph0-C	3.5	4.6	101	199	8	PDB	MOLECULE: BLV CAPSID;
17	4ph0-D	3.5	4.2	101	198	8	PDB	MOLECULE: BLV CAPSID;
18	4ph2-B	3.5	3.3	69	127	7	PDB	MOLECULE: BLV CAPSID - N-TERMINAL DOMAIN;
19:	1sxj-B	3.4	3.5	65	316	3	PDB	MOLECULE: ACTIVATOR 1 95 KDA SUBUNIT;
20:	2afd-A	3.4	2.7	59	88	14	PDB	MOLECULE: PROTEIN ASL1650;

Figure 2: **Top structural similarities** found by the DALI program in the 90% non-redundant PDB (PDB-90) using the full length foamy virus capsid as a query (145 residues). The columns are: the ranked number of the hit (No.), marked by a '|' for a capsid protein, otherwise ':'; the PDB entry identifier (Chain, with the chain designation after the dash); the DALI Z-score (Z) (significance estimate); the root-mean-square-deviation (rmsd) over aligned α -carbon positions; the number of aligned positions (lali); the number of residues in the matched structure (nres); the percentage sequence identity of the match (%id) followed by a description of the molecule. It can be seen from the number of matched positions (lali) that most matches are partial, covering typically less than half the query structure.

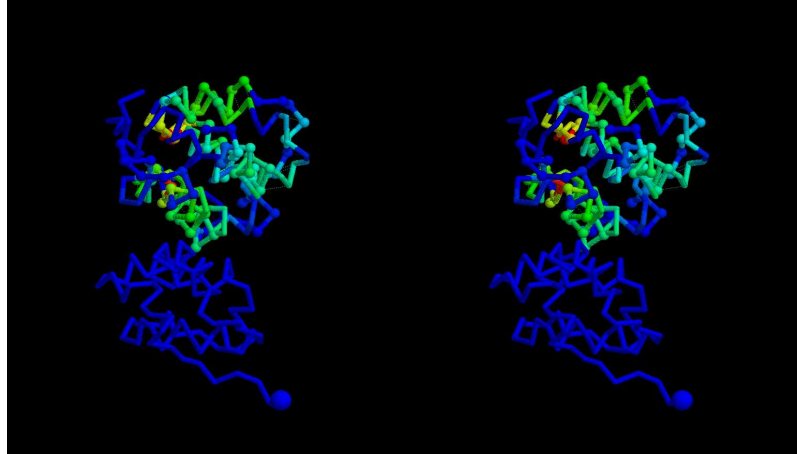
the Ty3/Gypsy retrotransposon family gag gene [8]. However, almost all of these are partial hits, covering little more than half the query structure. The structural alignment of the top two hits is shown in Figure 3 coloured to emphasise the matched regions.

The result of the DALI search indicated that the Foamy virus structure shares some similarity with the capsid structure of the ortho-viruses. However, the matches consist only of a small number of helices and appears barely more convincing than other matches to proteins that seem very unlikely to have any meaningful connection to a viral capsid. The preponderance of capsid matches throughout the list of hits might seem to add some support to the relationship but may simply be a reflection of the number of capsid structures in the structure databank.

Adding confusion to the ortho/foamy relationship is the additional observation that the distribution of matches to the ortho-virus structures between the amino (N) and carboxy (C) terminal domains are mixed. For example; taking the top 10 matches, the N-terminal domain of the Foamy structure



(a) 4x3x-A



(b) 3g29-A

Figure 3: **Top hits superposed.** The top two DALI hits to the full foamy virus capsid are shown as a α -carbon backbone (stereo pair) coloured using the residue similarity score calculated by SAP. (red = strong similarity, blue = none). The amino terminus of the foamy structure is marked by a large ball and the other structure is distinguished by small balls on its α -carbon atoms. (a) a cytoskeleton associated protein (fragment) of the arc/arg3.1 gene (PDB code: 4x3x-A), (which is thought to have originated from a Ty3/Gypsy retrotransposon family capsid) and (b) the structure of the capsid C-terminal domain of the Rous sarcoma virus (PDB code: 3g29-A).

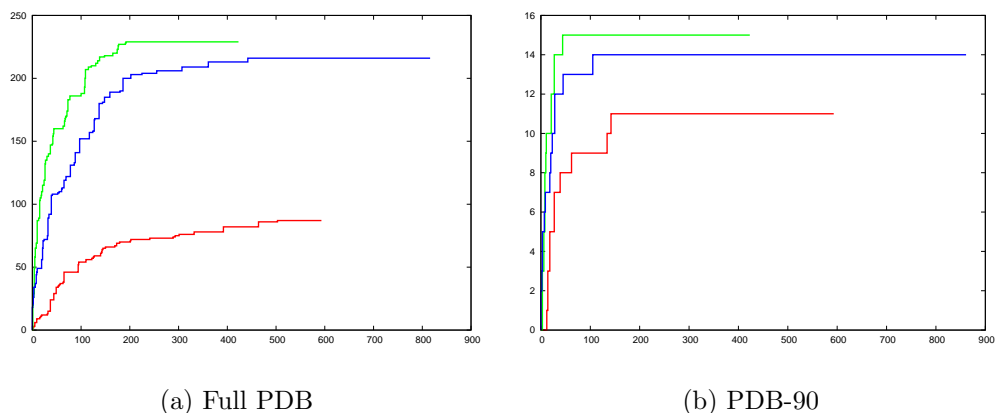


Figure 4: **PDB capsid structure matches.** The number of capsid structures identified by the DALI program in (a) the full PDB and (b) the 90% non-redundant PDB (PDB-90) is shown for queries using the full foamy capsid structure (red), the carboxy terminal domain (green) and the amino terminal domain (blue). The number of capsid hits (Y-axis) is plotted against the order of all hits ranked by Z-score down to a value of 2. A curve approaching the top left corner indicates greater specificity and the extent of a curve to the right indicates the total number of hits.

aligns with 6 C-terminal domains and 4 N-terminal domains of the ortho viruses and the best match with the corresponding Foamy C-terminal domain aligns with an ortho N-terminal domain.

2.2.2. Domain scans

To clarify the domain match specificity, the two domains of the Foamy virus (1–88 and 89–180, as defined automatically [9]) were scanned separately using the DALI program. The individual domains were much more specific at matching known capsid structures³, both in the full PDB and PDB-90 collections as can be seen from the plots in Figure 4.

The results of these scans strengthened the identification of the relationship to the ortho capsids and supported the swapped specificity for the N-terminal match of the Foamy structure with the C-terminal match of the ortho virus and *vica versa*, with all top 12 hits of each domain matching

³True/false hits were defined by protein descriptions with the words "CAPSID", "GAG" or "P24".

2.3. Structural alignment significance

2.3.1. Reversed-structure searches

For each comparison, the DALI program calculates an empirical Z-score, combining an estimation of significance with protein length normalisation. The program reports all matches over $Z=2$, however, when the proteins are small and especially when the structures being compared are both predominantly alpha-helical in nature, then matches over this cutoff include many functionally unrelated hits where the similarity has arisen through the fortuitous alignment of a few helices.

Therefore, to calculate a stricter cutoff on score, we created a decoy probe by reversing the alpha-carbon backbone then reconstructing the full atomic structure, using a simple algorithm to regenerate a full backbone⁴. Figure 6 plots the ranked DALI Z-scores for the separate (native) foamy domains. As would be expected, the larger C-terminal domain has hits with a higher significance than the smaller N-terminal domain: the former covers the range $Z=2.5$ to $Z=5$ over the true hits (magenta dots) whereas the latter tracks a similar profile running one Z-value unit lower (2–4 over true red dots). Plotting the Z-scores against the log of their rank produces almost linear traces for the hits from the PDB-90, making it easy to compare N-domain (red/cyan dots) with C-domain (magenta/green dots) (for T/F hits) in Figure 6.

The equivalent scans with the reversed domain structures, using both the foamy and ortho (HIV) structures (neither of which should have any particular relationship to the capsid or any other natural protein) also found hits with high Z-scores (black and blue points in Figure 6, respectively). When compared with the native domains (Figure 6), these decoys had a profile that tracked mostly above the N-terminal native domain but below the C-terminal domain. However, with the latter domain, this was only distinct in the hits to the full PDB whereas with the PDB-90, the native domain was only clearly better over the top 10 matches, half of which were to non-capsid structures.

The results with the simple reversed decoy using DALI suggested that the match of the foamy virus domains to the ortho virus capsid N-terminal domain may be due to chance and that the match to the C-terminal domain

⁴Note that reversing the α -carbon backbone does not change the chirality of the α -helices but as DALI requires a full atomic backbone, this must be restored on the reversed chain.

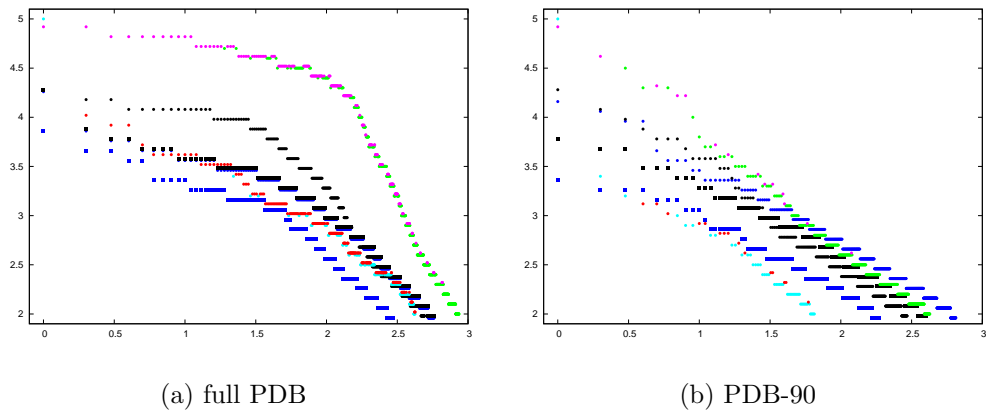


Figure 6: **Ranked DALI scores with decoys.** The DALI Z-scores (Y-axis) are plotted against the \log_{10} of their ranked position in the list of hits (X-axis) with the amino-terminal domain (N) as T=red, F=cyan dots and the carboxy-terminal domain (C) as T=magenta and F=green dots, where T is a true capsid hit and F is a false hit to a non-capsid protein. Four sets of decoys are compared to these, consisting of the reversed foamy capsid domains in black and the reversed HIV capsid domains in dark-blue (with a circle = N and a square = C domains in both). The DALI score for each set of hits has been slightly displaced to prevent coincident dots from being obscured. (This happens because of the integral number of residues and the DALI score being specified to only one decimal place).

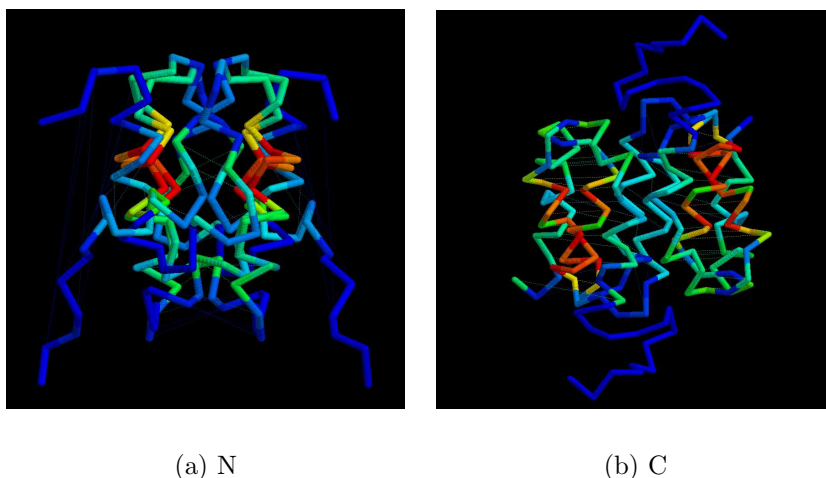


Figure 7: **Native/decoy similarity.** When superposed using the program **SAP**, both N-terminal (left) and C-terminal (right) domains have some degree of similarity to their reversed decoy 'doppleganger', which is more marked for the N domain. The superposed structures are coloured by the **SAP** residue-level score as red = high similarity, blue = low. The N domain has roughly 60 equivalent α -carbon positions compared to only 24 in the larger C domain.

134 looks meaningful if based on the hits to the full PDB but may be only
 135 marginal based on the PDB-90 hits.

136 However, both the N and C terminal domains pocess a degree of internal
 137 symmetry which gives rise to a partial match with their reversed 'dopple-
 138 ganger' decoys. The N-terminal domain superposed on its decoy had an
 139 RMSD of 5.4/60 ($\text{\AA}/\alpha$ -carbons) and 5.5/24 for the C-terminal domain. The
 140 higher symmetry of the smaller domain may be sufficient to explain its poor
 141 level of specificity seen in Figure 6 and to try and resolve this ambiguity, a
 142 more diverse set of decoys were generated based on cyclic permutation and
 143 segment swapping combined with chain reversal [10].

144 2.3.2. Customised decoy comparisons

145 To improve the statistical analysis of the foamy/ortho capsid similarity,
 146 we employed a method based on the generation of a population of customised
 147 'decoy' models to provide a background distribution of unrelated protein
 148 scores [10]. This method retains the advantage of the simple reversed struc-
 149 tures where every comparison that constitutes the random pool is between

two models of the same size and secondary structure composition as the pair of native structures being compared. For this study we collected 12 capsid N-terminal domains and 7 C-terminal domains, each of which were compared with the foamy N-terminal domain and the foamy C-terminal domain. (The structures are identified in Table 1 with full details in the Methods section).

For each domain pair to be compared, decoys were created using cyclic permutation and segment swapping with chain reversal to generate a family of customised decoys for each comparison [10]. All pairs of forward/reversed decoys were then compared, with each pair being drawn from a pool of models generated from the two native structures. This ensures that the native domains (which may have different lengths) are always evaluated against a decoy pair with the same length combination. (See Methods section for details). All the decoy comparisons, of which there are typically 150–300 for each comparison, can then be compared to the native pair on a plot of RMSD against the number of matched residues (α -carbon atoms). An example is shown in Figure 8 for the comparison of the HIV1 structure (PDB codes: 1ak4 (N) and 1a43 (C)) domains against the foamy virus gag domains.

2.3.3. Statistical analysis of the decoy comparisons

The quality of the comparisons in Figure 8 can be quantified as a combination of their RMSD (R) and the number of matched (superposed) positions (N). However, as explained in the Methods section, for statistical analysis, it is easier to combine this pair of numbers as a single number, called the a -value ($\text{Equ}^n. 1$), which is the scaling factor that causes a theoretical curve to pass through the point (R, N).

When expressed by a single a -value all the data points in a comparison, such as Figure 8(c), can be plotted as a frequency histogram and examined to see if they approximate a Normal distribution. The distributions were found to be a good fit to unskewed Gaussians and so were treated as normal distributions (rather than extreme value distributions that have also been considered previously as a model for random structure comparison scores [11, 10]). The frequency data from the comparison of the orthoN domain from HIV1 and the foamyC domain (Figure 8(c)) is shown in Figure 9(a) along with a Normal distribution that has the same mean (μ) and standard deviation (σ) as the data. On this plot, the value of a ($\text{Equ}^n. 1$) for the comparison of the native pair of domains is also plotted (blue triangle) and from its position, a Z-score can be calculated.

In this way, the significance of all combinations of the native ortho and

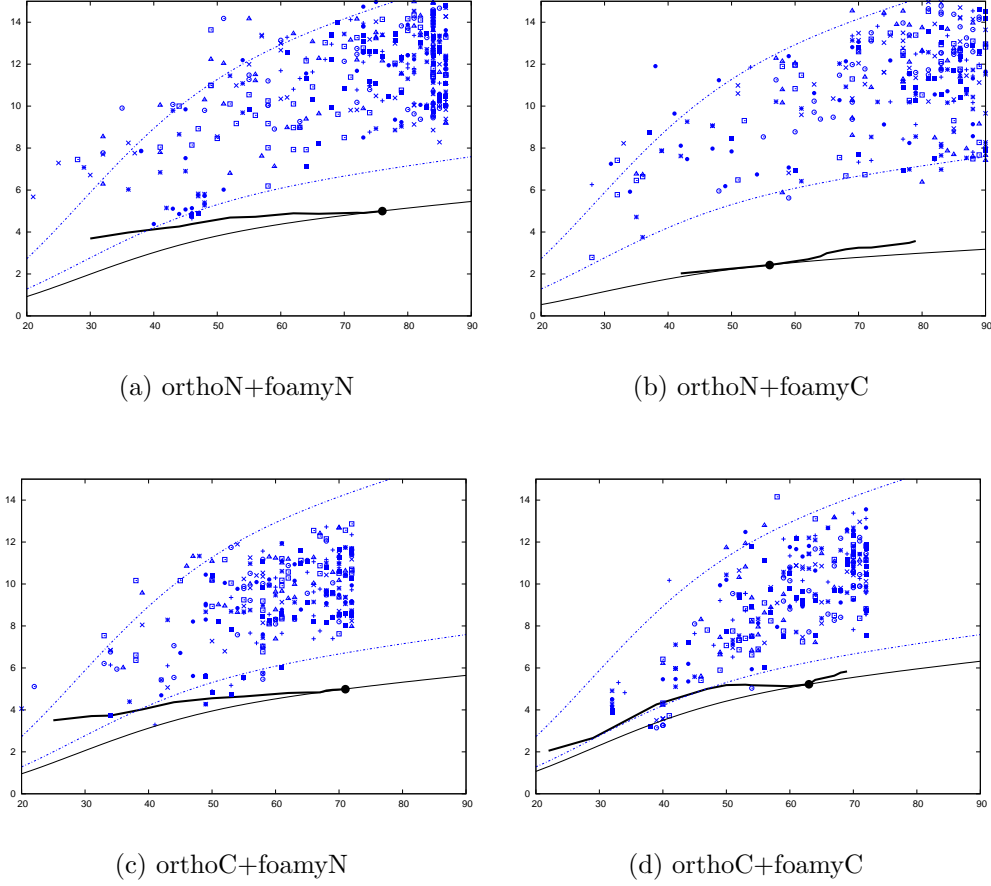


Figure 8: **ortho/foamy domains compared with customised decoys**. Each amino (N), carboxy (C) domain combination between the ortho retrovirus capsid structure (HIV1) and the foamy virus capsid structure is plotted as a line for increasingly large subsets of matched positions against their RMSD (Y-axis), as in Figure 2. The point on this line marks the lowest a -value (Equⁿ. 1), however, to be consistent with the decoy data, the full alignment length was used. The decoy comparison data (blue) is plotted in a variety of symbols with each representing a different combination of decoy construction. The dashed blue lines (which are the same in all plots) mark the approximate 10th percentile boundaries of the decoy generated distributions, with $a = 1.7$ (upper) and $a = 0.8$ (lower). (See Methods section for details).

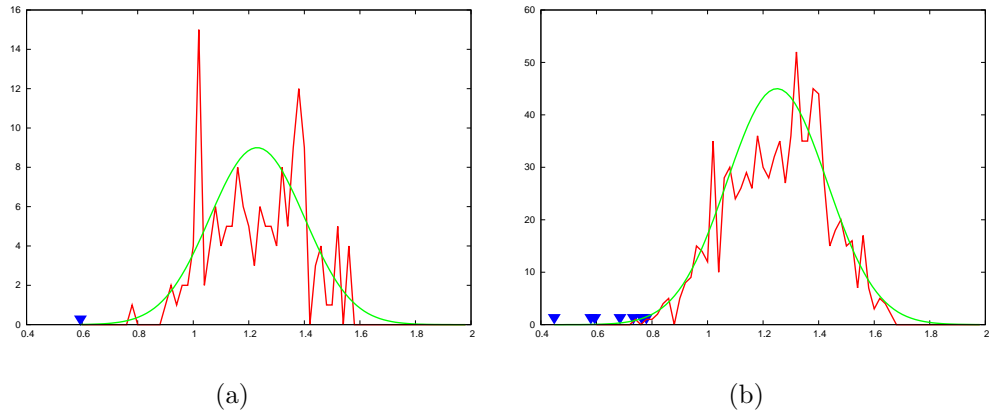


Figure 9: **ortho-C and foamy-N domain comparison statistics.** The a -value (normalised RMSD) for the comparison of the ortho-C and foamy-N decoy domains (Figure 8(b)) are plotted as a frequency distribution (red) along with a bell-shaped Normal distribution curve (green) with matching mean (μ) and spread (σ). Part (a) shows the distribution for the HIV1 C-terminal domain ($\mu = 1.23$) and spread ($\sigma = 0.17$) with the position of the native structure comparison plotted as a blue (inverted) triangle. Its position lies 0.64 units below the mean giving a Z-score of $0.64/0.17 = 3.76$. Part (b) shows the combined data from seven representative viruses (in Table 1). These data comprise two distributions, that of the combined decoys and also the much smaller distribution of native scores (blue triangles). This allows a T-test to be made on the significance of their separation.

foamy domain superpositions were calculated, using the background distribution of 'customised' decoy comparisons based on each individual native pair. The resulting Z-scores (σ units) are collected in Table 1. The degree of similarity between the domains ranged from less than 1σ to over 5σ , with the latter (highly significant) result being obtained for both a swapped (NC) and forward (CC) combination. However, of the top 12 scores, only three now came from swapped pairings.

Asymmetry statistics:

To quantify the degree of bias for domains of like-type (NN, CC) to be more similar than those of mixed-type (NC, CN), the observed ranking of like and mixed pairs, based on their Z-value (Table 1), was compared to that expected by chance. The positions of all pairs in the list were shuffled a million times and the asymmetry of each arrangement was quantified as the number of like-pairs in the top half and also by their second moment: $\sqrt{((\sum r_i^2)/N)}$, where r is the rank of the like-pair i in a list of N pairs. The chance of obtaining a distribution with more like-pairs being ranked higher can be calculated by summing the area of the tail of each empirical distribution that lies beyond the observed value. However, these values were calculated over all pairs and neglects the principle that emphasis should be given to the more significant similarities. Rather than rely on a single significance cutoff (like 3σ) or an arbitrary cutoff (like the "3-out-of-12" mentioned above), we calculated statistics for all such cutoffs (Figure 10(a)).

The majority of values in Figure 10(a) lie below the 0.05 probability level for the larger sample sizes, with those for the top-half bias statistic (blue line) being more significant than the moment-based statistic (red line). While confirming the visual trend towards a bias of higher scoring like-type domain similarities, the analysis summarised in Figure 10(a) is complicated by having unequal numbers of amino and carboxy domain comparisons and also by including some closely related structures. To produce a more balanced dataset, one of each pair of the two most similar carboxy domain structures was discarded leaving five structures and for each of these, their matching amino terminal domain was also retained, leaving: BLV-1, HIV-1, HML2, HTLV-1 and RSV. Despite having a smaller set of comparisons (5N + 5C domains giving 20 rather than 38 Z-scores), the results for this reduced set indicated an equally clear bias towards a preferred like-domain equivalence, especially as measured by their occurrence in the upper half of the ranked list, with several having a probability below the 0.05 level and a few below

<i>a</i>	ortho-N					
	foamy-N			foamy-C		
virus	pool	<i>a</i> -value	Z-score	pool	<i>a</i> -value	Z-score
BLV6	300	0.552	4.073	244	0.542	3.692
BLV	251	0.550	4.494	184	0.400	3.669
HIV6	312	0.551	3.781	220	0.405	3.579
HIV1	312	0.573	3.703	213	0.402	3.692
HML2	264	0.777	2.166	196	0.438	4.594
HTLV	400	0.592	4.030	328	0.457	4.013
JSRV	225	1.063	0.896	190	0.601	3.237
MLV	326	0.751	3.044	188	0.508	3.151
MPMV	269	0.565	3.902	185	0.523	2.918
PSIV	285	0.621	3.731	235	0.369	5.019
RELIK	234	0.639	3.688	237	0.700	3.297
RSV	204	0.543	3.123	239	0.526	3.542

<i>b</i>	ortho-C					
	foamy-N			foamy-C		
virus	pool	<i>a</i> -value	Z-score	pool	<i>a</i> -value	Z-score
BLV6	144	0.763	3.019	212	0.709	4.046
BLV	154	0.578	3.400	204	0.556	4.047
HIV1	157	0.593	3.760	174	0.705	3.362
HIV6	179	0.780	3.175	177	0.640	4.380
HML2	185	0.732	3.027	184	0.676	3.900
HTLV	156	0.685	3.847	163	0.694	2.807
RSV	155	0.448	3.754	235	0.403	5.009

Table 1: **Ortho and foamy domain comparison Z-score statistics.** For each amino (N) and carboxy (C) domain pair between an ortho virus structure and the foamy virus capsid structure, a **Z-score** is calculated based on the **a-value** (Equⁿ. 1) derived from the comparison RMSD and length, relative to the **pool** of background decoy comparisons. The ortho **virus** identity is indicated by the code to the left, full details of which can be found in the Methods section. The top 12 Z-scores are high-lighted in bold, only three of which support a swapped domain match.

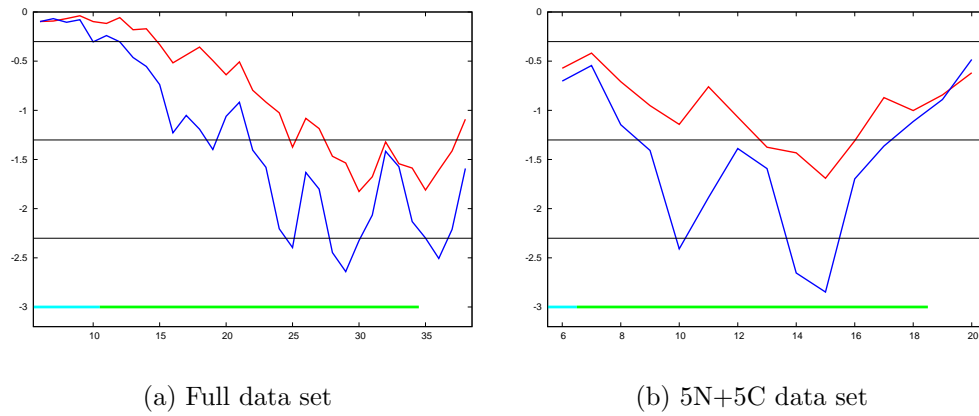


Figure 10: **Asymmetry statistics for like/mixed domain pairs.** Given the ranked list of domain pairings, the chance for more domain pairs of like-type to be found higher than the observed order was evaluated from empirical distributions measured by two statistics: the second moment of the rank value (red) and the number of like-type pairs in the top half (blue). These statistics were calculated for all subsets from the 6 top pairs up to the full set of comparisons (X-axis) and for each, the chance of a better score is plotted as the \log_{10} of the probability (Y-axis). The horizontal lines mark the 0.5, 0.05 and 0.005 levels. The line at the 0.001 level is coloured by the Z-score for each pair as: green = over 3 and cyan = over 4 sigma. Part (a) shows the probabilities calculated from the full set of 7 carboxy and 12 amino domains and part (b) shows the same values calculated on a more balanced set of 5 non-redundant carboxy domains and their matching amino domains.

	orthoN	orthoC
foamyN	Avg: 6.67e-01 < 1.32e+00 Tprob = 4.62e-21 ** StD: 1.61e-01 = 2.12e-01 Fprob = 1.84e-01	Avg: 6.51e-01 < 1.25e+00 Tprob = 2.35e-16 ** StD: 1.17e-01 = 1.89e-01 Fprob = 1.12e-01
foamyC	Avg: 4.92e-01 < 1.29e+00 Tprob = 4.09e-10 ** StD: 1.02e-01 < 2.21e-01 Fprob = 7.37e-03 **	Avg: 6.22e-01 < 1.30e+00 Tprob = 3.81e-23 ** StD: 1.12e-01 = 1.77e-01 Fprob = 1.20e-01

Table 2: **ortho and foamy capsid domain comparison T-test significance.** For each combination of domains between the ortho and foamy viruses, the probability is given that the two means from each distribution (Avg values) were sampled from the same distribution. (i.e., that the native and decoy comparisons are not distinct). All domain pairings are extremely significant. An F-test was used to test if the standard deviations (Std) of each sample were distinct and if not, the a T-test was made on the assumption of equal standard deviations.

the 0.005 level (Figure 10(b)).

T-test statistic:

An alternative to the above analysis, which still remains marginally significant, is to pool the raw comparison data for all the domain comparisons and their background distributions giving now not just a single value compared to a distribution but two distributions (Figure 9(b)). For these data, a significance was calculated using Student’s T-test, the values of which are given in Table 2.

From these results, it can be seen that all the four possible pairings are highly significant with probabilities ranging from 10^{-10} to over 10^{-20} . It is also clear that the two swapped pairings (NC and CN) have higher probabilities than the forward pairings (NN and CC). Combining the probabilities (P) as: $\Delta P = \log_{10}(P_{NN}P_{CC}) - \log_{10}(P_{NC}P_{CN})$, gives a value of 17.7 (42.7 - 25.0) which means that the swapped pairing is almost 18 orders of magnitude less likely than the forward pairing. Calculating the same statistic on the reduced 5N+5C domain data set gave a similar result but with a difference reduced 1000-fold to 15 orders of magnitude.

The unexpected swapped pairing, which was indicated originally by the

242 DALI results, now seems less likely. The preferred, and biologically more
 243 reasonable, result is that the ortho virus domain are related to the foamy virus
 244 domains as a result of genetic divergence from a common, double domain
 245 ancestor.

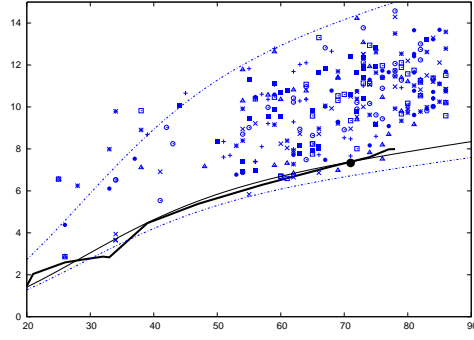
246 *2.4. Internal duplication*

247 The transposed pairings of N/C and C/N (ortho/foamy) domains still
 248 retain a high structural significance and this suggests that the two domains
 249 are derived from a common ancestral structure, probably as the result of a
 250 prior gene-duplication event that has been retained more clearly in the less
 251 embellished foamy virus structures. Comparing the two foamy domains gives
 252 a Z-score of 2.077 sigma which, although of marginal significance, supports
 253 this model. (Figure 11(a, b)).

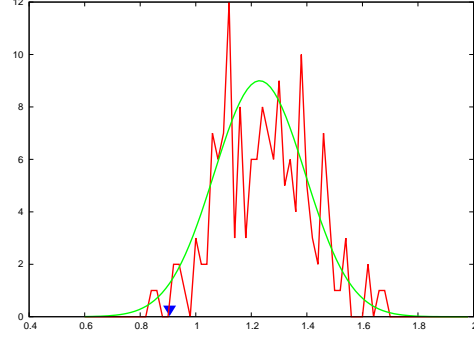
254 Such a relationship between the foamy domains implies an equivalent
 255 relationship in the ortho viruses and a similar comparison in structures of
 256 their N and C domains finds matches with Z-scores ranging from 2 to 4.
 257 As with the comparison of the ortho and foamy structures, these can be
 258 pooled to allow a joint T-test to be applied. This gave a probability of
 259 10^{-8} that the true N/C domain comparisons were drawn from the decoy
 260 distribution, adding strong support to the hypothesis of an ancient gene
 261 duplication occurring before the split of the ortho and foamy virus families.
 262 (Figure 11(c, d), blue triangles). Supporting this relationship, earlier studies
 263 also suggested an internal duplication in the ortho viruses but were based
 264 largely on very distant sequence similarity [12].

265 This test was applied only to the comparison of domains between viruses
 266 with known structures for both domains, however, it is not unreasonable to
 267 compare amino and carboxy domains across all viruses. The longer loops
 268 in the ortho virus domains gives greater scope of structural variation and a
 269 wide range of variation was seen ranging from RMSD values under 4 to over
 270 12. When normalised for length (a -value from Equⁿ. 1) and partial matches
 271 under 60 positions excluded, a distinct cluster remains between $a = 0.5 \dots 0.8$
 272 (4...6Å RMSD) but still with a long tail to higher values. Despite this tail,
 273 the T-test on the distributions is highly significant at 2.7×10^{-17} .

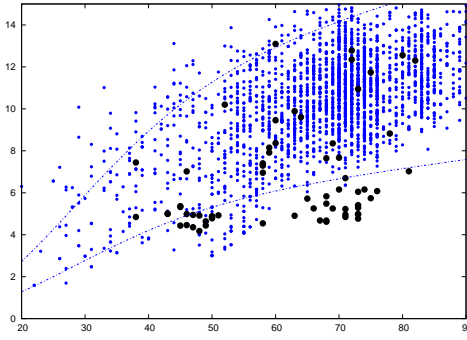
274 One of the better N/C ortho similarities is shown in Figure 12(a), along
 275 with the N/C ortho domain superposition in Figure 12(b).



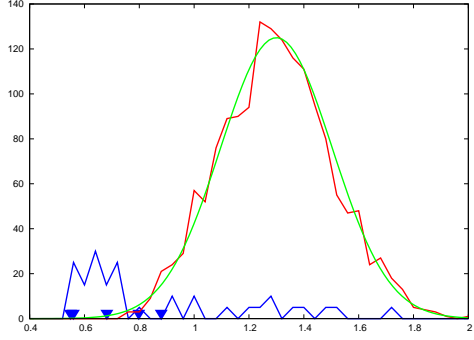
(a) foamy N+C



(b) foamy fit

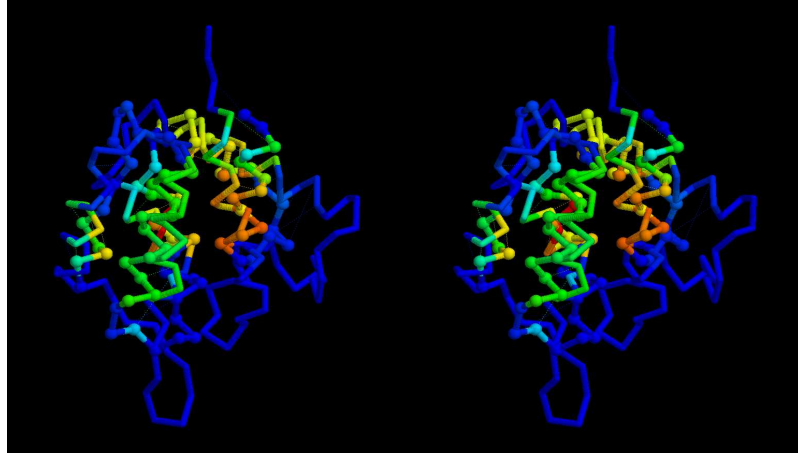


(c) ortho N+C

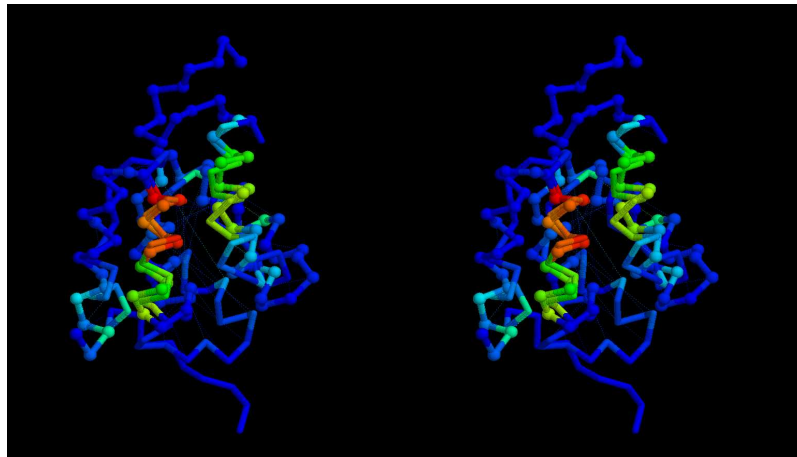


(d) ortho fit

Figure 11: **N and C domains compared with customised decoys.** *a)* The N and C domains of the foamy virus (black) compared to decoys (blue) with *(b)* the derived frequency plot with the native comparison marked by a blue triangle. (See legend to Figure 8 for details). *c)* The N and C domains of the ortho virus combinations (black) with *(d)* the derived frequency plot showing the native comparison for pairs from the same virus (blue triangles) with the distribution of all native pairs shown as a scattered frequency plot (blue line). (See Methods section for details).



(a) ortho



(b) foamy

Figure 12: **Amino and carboxy domains superposed.** *a* ortho virus domains and *b* foamy virus domains are shown as a stereo pair with their α -carbon backbones coloured by the residue similarity score calculated by **SAP**. (red = strong similarity, blue = none). The amino terminal domain is distinguished by small balls on its α -carbon positions and the amino terminus lies to the top in both panels.

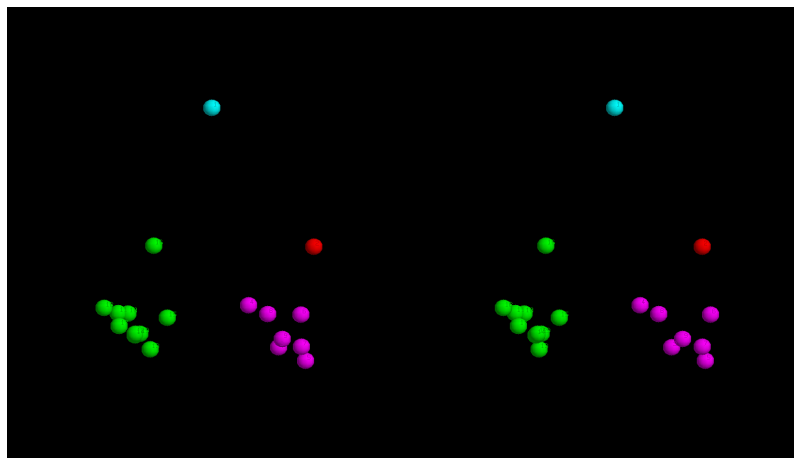


Figure 13: **Fold-space representation of all domains.** All the viral domains considered in the paper were projected into a 3D fold-space representing the relationship of their **SAP** weighted RMSD values. The domains are coloured as: foamyN = cyan, foamyC = red, orthoN = green and ortho C = magenta.

276 2.5. Fold-space representation

277 To summarise the structural relationships among the ortho and foamy
278 domains, the matrix of pairwise comparisons was projected into a three-
279 dimensional fold-space. (See methods for details). This produces a best
280 visual representation of the RMSD values between domains.

281 As can be seen from Figure 13, the N and C domains of the ortho viruses
282 form distinct clusters with the foamy C domain lying closer to the ortho
283 C-domain cluster. The foamy N-domain, however, maintains a fairly equal
284 distance from both ortho domain clusters but lies closer to its C-terminal
285 partner.

286 3. Conclusions

287 3.1. Structure comparison

288 3.1.1. Pairwise significance

289 The comparison of small domains that are largely composed of α -helices
290 presents a challenging problem in how to interpret the significance of the
291 RMSD values. As the individual helical secondary structure elements (SSEs)
292 constitute a sizeable fraction of the domain, it takes only the chance align-
293 ment of a few helices to result in a low RMSD over a large proportion of the
294 structure, giving an apparently meaningful result.

295 The use of the customised decoy-model sets, as illustrated here, attempts
296 to avoid this problem by recreating a large number of possible folds that
297 were generated using the same (reconnected) SSEs. Moreover, to avoid any
298 chance recreation of native fragments, each comparison always involved the
299 comparison of a native (forward) chain direction with a reversed chain. Using
300 these models, a background distribution of decoy/decoy comparisons allowed
301 us to calculate Z-scores for each native/native comparison between the dif-
302 ferent Gag proteins. This has the advantage that every comparison in the
303 background distribution involved two models with the same length, residue
304 packing density and secondary composition as the native pair. These values
305 indicated a clearly significant relationship between the foamy and ortho CA
306 structures.

307 3.1.2. *Direct or transposed domain order?*

308 Although the decoy model alignment strategy did confirm the relationship
309 between the foamy and ortho CA structures, the Z-scores did not point to a
310 clear resolution of whether the domains should have a direct correspondence
311 (NN and CC match) or a transposed relationship (NC and CN) as significant
312 individual matches were found across all pairings. Testing for a bias towards
313 more significant like-domain pairings (NN, CC) in the list of similarities
314 ranked by Z-score confirmed the visual bias towards a natural correspondence
315 but only at a marginal level of significance (around 0.05). By contrast, the
316 application of a T-test on the combined raw comparison data returned a very
317 clear distinction between the direct and the transposed relationships, clearly
318 favouring the more natural forward order.

319 However, although the astronomic probabilities calculated by the T-test
320 seem very convincing, they must be viewed in the light of the much lower
321 probabilities calculated from the asymmetry statistics. Both calculations
322 involve assumptions and are limited by the small number of known structures
323 so neither can be taken as definitive. Nevertheless it would seem likely that
324 the true level of significance may lie somewhere between the two results and as
325 both of these objective assessments point in the direction of the NN and CC
326 domain order, there is no reason to adopt the more unexpected transposed
327 domain order.

328 3.2. *Evolutionary implications*

329 On the basis of these structural comparisons, and a variety of recently
330 described functional assays [5], we can conclude that the central region of the

331 spumavirus gag gene encodes a polypeptide sequence related to that of the
 332 corresponding region of orthoretroviral, CA. It therefore seems reasonable to
 333 suppose that the last common ancestor of orthoretroviruses and spumaviruses
 334 possessed such a sequence. Moreover this region appears to be made up from
 335 two related all helical subdomains suggesting a gene duplication event in a
 336 common precursor.

337 In our initial search employing foamy virus CA using the DALI program,
 338 we made the observation that the strongest similarity of the foamy virus
 339 CA domains was actually with a cellular protein, Arc (Activity-Regulated
 340 Cytoskeleton-associated protein). Arc is required for neural synaptic growth
 341 and activity [13, 14, 15, 16] and mis-regulation and/or deletion contributes
 342 to diseases of cognition [17, 14, 16]. Arc has widespread and clear sequence
 343 homologues as far back as insects and probably deeper, giving it a very
 344 ancient origin somewhere close to the metazoan root [12, 18] and based on
 345 sequence homology Arc is considered to be a relic of an ancient Ty3/Gypsy
 346 retrotransposon [8], preserved as a living fossil in metazoan genomes. Given
 347 the structural relatedness of foamy virus CA and Arc, this might suggest
 348 an equally ancient origin for foamy virus CA. As it is believed that the
 349 Ty3/Gypsy family of retrotransposons gave rise to retroviruses [19], it will
 350 therefore be of considerable interest to determine whether the Gag of Ty
 351 elements also comprise CA proteins with a two-domain structure.

352 It is also noteworthy that Ty3 Gag is significantly smaller than that
 353 of the foamy and orthoretroviruses and although it contains CA related se-
 354 quences there is no equivalent of either orthoretroviral MA or PFV Gag-NtD,
 355 regions of Gag necessary for membrane targeting, budding and extracellu-
 356 lar release of virions. Therefore, given the very different structures of MA
 357 [20, 21, 22, 23] and Gag-NtD [4], this raises the possibility that the MA and
 358 Gag-NtD domains of the orthoretroviruses and foamy viruses were co-opted
 359 by independent events that has resulted in the viruses employing different
 360 mechanisms to facilitate budding from the cell. Notably, Gag from Gypsy, an
 361 Errantivirus capable of extracellular replication [24] and Arc contain addi-
 362 tional N-terminal domains. In Gypsy-Gag this domain is distantly sequence-
 363 related to orthoretroviral MA [12]. By contrast, in Arc it contains a coiled
 364 coil region [8] reminiscent of spumavirus Gag-NtD [4, 25] further supporting
 365 the notion of a shared origin for Arc and foamy virus Gag that is distinguish-
 366 able from an alternative acquisition pathway giving rise to Gypsy and the
 367 orthoretroviruses.

368 4. Methods

369 4.1. Structural data

370 The foamy virus structures were obtained from the Protein Structure
371 Databank (PDB code:5M1G) [5].

372 The ortho virus structures used, with their shorthand code in bold and
373 PDB code in teletype, were:

- 374 • **BLV**: bovine leukemia virus (deltaretrovirus) **4PH1** (N-ter.dom) and
375 **4PH2** (C-ter.dom) [26],
- 376 • **BLV6**: bovine leukemia virus (hexameric) **4PH0** (both dom.s) [26],
- 377 • **HIV1**: human immunodeficiency virus 1 (lentivirus) **1AK4** (N-ter.dom)
378 [27] and **1A43** (C-ter.dom) [28],
- 379 • **HIV6**: human immunodeficiency virus 1 **3H47** (both dom.s) [29],
- 380 • **HML2**: human endogenous retrovirus type-K (betaretrovirus) [30],
- 381 • **HTLV**: human T-cell leukemia virus (deltaretrovirus) **1QRJ** (both dom.s)
382 [31],
- 383 • **JSRV**: jaagsiekte sheep Retrovirus (betaretrovirus) **2V4X** (N-ter.dom)
384 [32],
- 385 • **MLV**: murine leukemia virus (gammaretrovirus) **1U7K** (N-ter.dom) [33],
- 386 • **MPMV**: Mason-Pfizer monkey virus (betaretrovirus) **2KGF** (N-ter.dom)
387 [34],
- 388 • **PSIV**: prosimian immunodeficiency virus (ancient lentivirus) **2XGV** (N-
389 ter.dom) [35],
- 390 • **RELIK**: rabbit endogenous lentivirus type-K (ancient lentivirus) **2XGU**
391 (N-ter.dom) [35],
- 392 • **RSV**: Rous sarcoma virus (alpharetrovirus) **3G1I** (both dom.s) [36].

393 4.2. Structure comparison

394 4.2.1. DALI

395 The DALI method for searching the PDB with a structural query [7] was
396 accessed via the server at: http://ekhidna.biocenter.helsinki.fi/dali_server.
397 The DALI method reports the significance of each match with an estimated Z-
398 score which is the raw comparison score, normalised by the combined length
399 of the proteins. Z-scores down to a value of 2 are reported by the program.

400 The list of DALI hits (ranked by Z-score) were assessed by how many high-
401 scoring capsid structures had been identified. These true/false (T/F) hits
402 were defined simply by protein descriptions that contained the words "CAP-
403 SID", "GAG" or "P24". This may have misclassified a few (low scoring) hits
404 to the matrix protein and missed some hits where the primary description
405 refers to a cyclophilin structure solved in complex with the capsid.

406 DALI reports structural hits in both the full PDB and a reduced collection
407 of structures that have no pair of proteins with over 90% sequence identity,
408 referred to as the 90% non-redundant or PDB-90 collection. It was found,
409 however, that some hits, seen in the full PDB were not found in the PDB-90,
410 for example in Figure 6, all of the top 31 hits of the N-domain against the
411 full PDB are missing in the PDB-90 hits. The most likely explanation is
412 that the PDB-90 secection has not been updated at the same time as the full
413 collection. For this reason, hits to both databases were monitored.

414 4.2.2. SAP

415 The SAP method for structure comparison [6] was run as a local copy
416 which can be accessed at: <https://github.com/WillieTaylor/util>. As
417 part of determining the alignment between two structures, the SAP program
418 calculates a similarity score for each pair of matched positions which is how
419 similar the rest of the structure looks from the viewing-frame of the su-
420 perposed residues. This value can be used both to weight the importance
421 of positions when calculating the (rigid-body) RMSD superposition and to
422 colour positions in the superposed structures. [37]. (As in Figure 3).

423 If the matched positions are ranked by this value, then RMSD values can
424 be calculated over increasingly larger subsets to high-light the extent of a
425 well matched core before the contribution of variable loops, or domain shifts,
426 leads to higher RMSD values. (As in Figure 1(b)).

4.3. Decoy structure construction

4.3.1. Reversed structure decoys

Simple structural decoys were generated from native PDB structures by reversing the order of the α -carbon atoms in the PDB file using the Unix command line:

```
cat native.pdb | grep ' CA ' | sort -nr -k2 > reverse.pdb
```

The reversal of a protein chain does not alter the chirality of the alpha helix and these decoys can be used directly in SAP. However, DALI requires all main-chain atoms and these must be regenerated for the reversed decoys. This was done using the simple `ca2main` program which can also be found at: <https://github.com/WillieTaylor/util>.

4.3.2. Customised decoys

Customised structural decoys were generated for each comparison using each of the pair of structures being compared to create two pools of decoys then comparing all decoys in the first pool against all decoys from the second but with their chain reversed as described in the previous section.

The decoys were created as described in Ref.[10]: starting by cyclising the chain then introducing new termini in each surface loop to create cyclic permutations. In addition, when three loop regions lie in close proximity, their ends are also reconnected in such a way that if a chain, comprising four segments (1...4) runs from amino (N) to carboxy (C) termini through three adjacent loop regions a-b, c-d and e-f (i.e.: N,1,a-b,2,c-d,3,e-f,4,C) then the reconnected chain runs: N,1,a-d,3,e-b,2,c-f,4,C with each switch being made at the least disruptive point between a pair of loops. This chain switching does not create any reversed segments which would otherwise form regions of local matching when the whole chain is reversed.

In a pair of structures, if each have four surface loops where breaks can be made, then including the native termini, this gives five cyclic permutations and if two groups of loops can be reconnected then a total of 15 distinct decoys can be made from each native starting structure. As these can be compared pairwise, a pool of 225 decoy derived data points is generated that constitutes the random background against which the native/native comparison can be assessed.

For example, in Figure 8, the 36 data points marked by a solid circle come from the comparison of six cyclic permutations of a native ortho domain

462 compared with six permutations of a reversed foamy domain that includes a
 463 single loop reconnection.

464 Every pair drawn from this pool will have the same lengths as the two
 465 native structures as well as the same secondary structure composition, surface
 466 exposure, residue packing density and inertial properties but each decoy will
 467 have a different chain fold.

468 4.4. Statistical tests

469 4.4.1. RMSD length normalisation

470 The quality of structure comparisons can be characterised by a combina-
 471 tion of their RMSD value and the number of matched (superposed) positions.
 472 How to combine these values has been the subject of much discussion over
 473 the years and central to this is the expected random RMSD value for two
 474 proteins of a given length [38, 39, 40]. However, when reviewed [10], all these
 475 measures were approximations of a simple square-root function of the protein
 476 length (as originally proposed by McLachlan on theoretical grounds [38]) but
 477 with an added term to depress the RMSD values obtained with small units
 478 or structure that are dominated by secondary structure elements (and super-
 479 secondary structure motifs) giving a lower than expected RMSD value. The
 480 formula that best captures this is: $R = \sqrt{N(1 - \exp(-N^2/s^2))}$, where, R is
 481 the expected random RMSD for N matched positions and s is the damping
 482 factor in the inverted Gaussian term (equivalent to the standard deviation
 483 in the Normal distribution).

484 Any point that lies on this line can be considered "exactly" random with
 485 those above it being "more" random and those below it "less" random. This
 486 can be quantified as a single number which is the value of a scaling factor
 487 (a), which when applied to the curve, makes it pass through any given point.
 488 If a comparison has an RMSD of R over N positions, then $R = a\sqrt{N(1 - \exp(-N^2/s^2))}$
 489 and when

$$a = R/(\sqrt{N(1 - \exp(-N^2/s^2))}), \quad (1)$$

490 the line will pass through the data point. This reduces the pair of values
 491 (R, N) to a single value a that is a simpler quantity for statistical analysis.

492 The best value for s is slightly dependent on the nature of the proteins
 493 being compared. For artificial (random-walk) models with no secondary struc-
 494 ture, no modification will be needed but the proteins considered here have
 495 segments of packed alpha helices that can be locally similar over two to three

496 helices. To correct for this, a value of $s = 30$ was used (or $1/s^2 = 0.11$)
 497 which is higher than the value of $1/s^2 = 0.03$ used previously. That this is
 498 a reasonable fit to the data can be seen in the way the dashed blue lines
 499 in Figure 8 track the upper and lower boundary of the decoy comparison
 500 results.

501 When $a = 1$, the point lies on the random line and when $a = 0$, the RMSD
 502 is zero, so values of a that approach this lower bound will be of interest when
 503 evaluating similarity.

504 4.4.2. Frequency plots

505 The a -values obtained using Equⁿ. 1 were plotted as frequency histograms
 506 using only data points that had a length of $N \pm 10$, where N is the
 507 maximum number of matched positions in the comparison of the two native
 508 structures. Previously, a cumulative plot of RMSD was used to select an
 509 optimal value for N (giving the minimum a -value). This can be important
 510 if the full set of matched positions is dominated by a high deviations from
 511 variable loop regions. However, in the current application, the small length
 512 of the foamy virus loops meant that this was not an important aspect and
 513 the full number of matched positions was taken. Otherwise, the same cor-
 514 rection would have to be applied to all decoy comparisons to maintain a fair
 515 comparison. (See Figure 8, where the black dot marks the minimum a -value
 516 length).

517 The mean and standard deviation of the a -values in the $N \pm 10$ region
 518 were calculated and the corresponding Normal distribution used to calcu-
 519 late Z-scores for the associated native comparison. (See Figure 9(a), for an
 520 example).

521 4.4.3. T-tests

522 Data from separate native/native comparisons, with their customised de-
 523 coy data, were combined giving not only a much larger background popula-
 524 tion of decoy derived scores but also a small population of native comparison
 525 scores that can be tested to calculate the probability that they were drawn
 526 from the same population as the decoy data. To do this, a T-test was used
 527 which takes the size, mean, and standard deviation of each distribution and
 528 calculates a probability. The implementaion of this test was taken from the
 529 Numerical Reicpies collection [41] which implements one of two variants of
 530 the test depending on whether the distributions have statistically distinct

531 standard deviations. (Routines `ttest()` and `tutest()`). The choice of rou-
532 tine is based on a preapplication of an F-test on the standard-deviations.
533 (Using the routine `ftest()`).

534 The values quoted in the Results section are for a two-tailed T-test, how-
535 ever, as it is expected that the native comparisons should always be more
536 similar than comparisons between random models, then a one-tailed T-test
537 would be valid, which gives half the probability. As the values in the Tables
538 are so significant and only the relative relationships are of interest, then the
539 choice is unimportant.

540 4.5. Fold-space clustering

541 The results of the pairwise similarity within a set of structures can be
542 visualised by treating the RMSD values as Euclidean distances⁵ and reducing
543 their dimensionality to sufficiently few dimensions to be visualised: usually
544 2D or, better 3D, to visualise the space with less distortion. Rather than use a
545 simple multi-dimensional scaling (MDS) method ([42]), the more complicated
546 method of multi-dimensional projection was used ([43], see [44] for a simpler
547 exposition).

548 This method reduces the dimensionality of the projection in gradual
549 stages with each step employing triangle-inequality balancing and hyper-
550 dimensional real-space refinement. In the real-space refinement stages, a
551 weight can be applied to pairwise distances. (This cannot be done in direct
552 MDS projection, which can only assign a mass to each point). Weights were
553 assigned to distances as a function of their inverse RMSD, up to a maximum
554 value of 1.

555 The method is robust and has been widely applied to rough models ([45])
556 and predicted inter-residue distances that constitute highly non-metric data
557 sets ([46]).

558 References

- 559 [1] D. Lindemann, A. Rethwilm, Foamy virus biology and its application
560 for vector development, *Viruses* 3 (2011) 561–585.

⁵In theory, pairwise RMSD values are guaranteed to constitute a consistent Euclidean metric, but only in N-1 dimensions (where N is the number of structures compared).

- 561 [2] E. Mllers, The foamy virus Gag proteins: what makes them different?,
562 Viruses 5 (2013) 1023–1041.
- 563 [3] R. Flgel, K. Pfrepper, Proteolytic processing of foamy virus Gag and
564 Pol proteins, Curr. Top. Microbiol. Immunol 277 (2003) 63–88.
- 565 [4] D. Goldstone, T. Flower, N. Ball, M. Sanz-Ramos, M. Yap, R. Ogrodow-
566 icz, N. Stanke, J. Reh, D. Lindemann, J. Stoye, I. Taylor, A unique
567 spumavirus Gag N-terminal domain with functional properties of or-
568 thoretroviral matrix and capsid, PLoS pathogens 9 (2013) e1003376.
- 569 [5] N. Ball, G. Nicastro, M. Dutta, D. Pollard, D. Goldstone, M. Sanz-
570 Ramos, A. Ramos, E. Mllers, K. Stirnnagel, N. Stanke, D. Lindemann,
571 J. Stoye, W. Taylor, P. Rosenthal, I. Taylor, Structure of a spumaretro-
572 virus gag central domain reveals an ancient retroviral capsid, PLoS
573 Path. 12 (2016) e1005981. Doi:10.1371/journal.ppat.1005981.
- 574 [6] W. R. Taylor, Protein structure alignment using iterated double dy-
575 namic programming, Prot. Sci 8 (1999) 654–665.
- 576 [7] L. Holm, C. Sander, Protein-structure comparison by alignment of dis-
577 tance matrices, J. Molec. Biol. 233 (1993) 123–138.
- 578 [8] W. Zhang, J. Wu, M. Ward, S. Yang, Y. Chuang, M. Xiao, R. Li,
579 D. Leahy, P. Worley, Structural basis of arc binding to synaptic proteins:
580 implications for cognitive disease., Neuron 86 (2015) 490–500.
- 581 [9] W. R. Taylor, Protein structure domain identification, Prot. Engng. 12
582 (1999) 203–216.
- 583 [10] W. R. Taylor, Decoy models for protein structure score normalisation,
584 J. Molec. Biol. 357 (2006) 676–699.
- 585 [11] M. Levitt, M. Gerstein, A unified statistical framework for sequence
586 comparison and structure comparison, Proc Natl Acad Sci USA 95
587 (1998) 5913–5920.
- 588 [12] M. Campillos, T. Doerks, P. Shah, P. Bork, Computational characteri-
589 zation of multiple gag-like human proteins, Trends in Genetics 22 (2006)
590 285–589.

- 591 [13] S. Chowdhury, J. Shepherd, H. Okuno, G. Lyford, R. Petralia, N. Plath,
592 D. Kuhl, R. Huganir, P. Worley, Arc/Arg3.1 interacts with the endocytic
593 machinery to regulate AMPA receptor trafficking, *Neuron* 52 (2006)
594 445–459.
- 595 [14] S. Park, J. Park, S. Kim, J. Kim, J. Shepherd, C. Smith-Hicks,
596 S. Chowdhury, W. Kaufmann, D. Kuhl, A. Ryazanov, et al., Elongation
597 factor 2 and fragile X mental retardation protein control the dynamic
598 translation of Arc/Arg3.1 essential for mGluR-LTD, *Neuron* 59 (2008)
599 70–83.
- 600 [15] J. Shepherd, G. Rumbaugh, J. Wu, S. Chowdhury, N. Plath, D. Kuhl,
601 R. Huganir, P. Worley, Arc/arg3.1 mediates homeostatic synaptic scal-
602 ing of AMPA receptors, *Neuron* 52, 475–484 (2006).
- 603 [16] M. Waung, B. Pfeiffer, E. Nosyreva, J. Ronesi, K. Huber, Rapid transla-
604 tion of Arc/Arg3.1 selectively mediates mGluR-dependent LTD through
605 persistent increases in AMPAR endocytosis rate, *Neuron* 59 (2008) 84–
606 97.
- 607 [17] F. Niere, J. Wilkerson, K. Huber, Evidence for a fragile x mental retar-
608 dation protein-mediated translational switch in metabotropic glutamate
609 receptor-triggered arc translation and long-term depression, *J. Neurosci.*
610 32 (2012) 5924–5936.
- 611 [18] J. Volff, Cellular genes derived from Gypsy/Ty3 retrotransposons in
612 mammalian genomes, *Annals New York Acad. Sci.* 1178 (2009) 233–
613 243.
- 614 [19] C. Llorens, M. Fares, A. Moya, Relationships of gag-pol diversity be-
615 tween Ty3/Gypsy and retroviridae LTR retroelements and the three
616 kings hypothesis., *BMC Evol. Biol.* 8 (2008) e276.
- 617 [20] C. Hill, D. Worthylake, D. Bancroft, A. Christensen, W. Sundquist,
618 Crystal structures of the trimeric human immunodeficiency virus type
619 1 matrix protein: implications for membrane association and assembly,
620 *Proc. Natl. Acad. Sci. U.S.A.* 93 (1996) 3099–3104.
- 621 [21] J. Prchal, P. Srb, E. Hunter, T. Ruml, R. Hrabal, The structure of
622 myristoylated mason-pfizer monkey virus matrix protein and the role

- 623 of phosphatidylinositol-(4,5)-bispophate in its membrane binding, J.
624 Mol. Biol. 423 (2012) 427–438.
- 625 [22] Z. Rao, A. Belyaev, E. Fry, P. Roy, I. Jones, D. Stuart, Crystal structure
626 of SIV matrix antigen and implications for virus assembly, Nature 378
627 (1995) 743–747.
- 628 [23] N. Riffel, K. Harlos, O. Iourin, Z. Rao, A. Kingsman, D. Stuart, E. Fry,
629 Atomic resolution structure of moloney murine leukemia virus matrix
630 protein and its relationship to other retroviral matrix proteins, Structure
631 10 (2002) 1627–1636.
- 632 [24] S. Song, T. Gerasimova, M. Kurkulos, J. Boeke, V. Corces, An env-like
633 protein encoded by a drosophila retroelement: evidence that gypsy is
634 an infectious retrovirus, Gene Dev 8, 2046-2057 (1994).
- 635 [25] J. Tobaly-Tapiero, P. Bittoun, M. Giron, M. Neves, M. Koken, A. Saib,
636 H. de The, Human foamy virus capsid formation requires an interaction
637 domain in the N-terminus of Gag, J Virol 75, 4367-4375 (2001).
- 638 [26] G. Obal, F. Trajtenberg, F. Carrion, L. Tome, N. Larrieux, X. Zhang,
639 O. Pritsch, A. Buschiazzi, Conformational plasticity of a native retrovi-
640 ral capsid revealed by X-ray crystallography., Science 349 (2015) 95–98.
641 DOI: 10.1126/science.aaa5182.
- 642 [27] T. Gamble, F. Vajdos, S. Yoo, D. Worthylake, M. Houseweart,
643 W. Sundquist, C. Hill, Crystal structure of human cyclophilin A bound
644 to the amino-terminal domain of HIV-1 capsid., Cell 87 (1996) 1285–
645 1294.
- 646 [28] D. Worthylake, H. Wang, S. Yoo, W. Sundquist, C. Hill, Structures of
647 the HIV-1 capsid protein dimerization domain at 2.6Å resolution., Acta
648 Crystallogr.,Sect.D 55 (1999) 85–92. DOI: 10.1107/S0907444998007689.
- 649 [29] O. Pornillos, B. Ganser-Pornillos, B. Kelly, Y. Hua, F. Whitby, C. Stout,
650 W. Sundquist, C. Hill, M. Yeager, X-ray structures of the hexameric
651 building block of the HIV capsid., Cell 137 (2009) 1282–1292. DOI:
652 10.1016/j.cell.2009.04.063.

- 653 [30] G. Mortuza, M. Dodding, D. Goldstone, L. Haire, J. Stoye, I. Taylor,
654 Structure of B-tropic MLV capsid N-terminal domain., *J. Mol. Biol.* 376
655 (2008) 1493–1508.
- 656 [31] S. Khorasanizadeh, R. Campos-Olivas, C. Clark, M. Summers,
657 Sequence-specific ^1H , ^{13}C and ^{15}N chemical shift assignment and sec-
658 ondary structure of the HTLV-I capsid protein., *J. Biomol. NMR* 14
659 (1999) 199–200.
- 660 [32] G. B. Mortuza, D. C. Goldstone, C. Pashley, L. F. Haire, M. Palmarini,
661 W. R. Taylor, J. P. Stoye, I. A. Taylor, Structure of the capsid amino
662 terminal domain from the betaretrovirus, Jaagsiekte sheep retrovirus.,
663 *J. Molec. Biol.* 386 (2009) 1179–1192.
- 664 [33] G. B. Mortuza, L. F. Haire, A. Stevens, S. J. Smerdon, J. P. Stoye,
665 I. A. Taylor, High-resolution structure of a retroviral capsid hexameric
666 amino-terminal domain., *Nature* 431 (2004) 481–485.
- 667 [34] P. Macek, J. Chmelik, I. Krizova, P. Kaderavek, P. Padrta, L. Zidek,
668 M. Wildova, R. Hadravova, R. Chaloupkova, I. Pichova, T. Ruml,
669 M. Rumlova, V. Sklenar, NMR structure of the N-terminal domain
670 of capsid protein from the mason-pfizer monkey virus, *J. Mol. Biol.* 392
671 (2009) 100–114. DOI: 10.1016/j.jmb.2009.06.029.
- 672 [35] D. C. Goldstone, M. W. Yap, L. E. Robertson, L. F. Haire, W. R. Taylor,
673 A. Katzourakis, J. P. Stoye, I. A. Taylor, Structural and functional anal-
674 ysis of prehistoric lentiviruses uncovers an ancient molecular interface.,
675 *Cell Host Microbe* 8 (2010) 248–259.
- 676 [36] G. Bailey, J. Hyun, A. Mitra, R. Kingston, Proton-linked dimerization
677 of a retroviral capsid protein initiates capsid assembly, *Structure* 17
678 (2009) 737–748. DOI: 10.1016/j.str.2009.03.010.
- 679 [37] F. Rippmann, W. R. Taylor, Visualization of structural similarity in
680 proteins, *J. Molec. Graph.* 9 (1991) 3–16.
- 681 [38] A. D. McLachlan, How alike are the shapes of two random chains?,
682 *Biopolymers* 23 (1984) 1325–1331.

- 683 [39] F. E. Cohen, M. J. E. Sternberg, On the prediction of protein structure:
684 the significance of the root-mean-square deviation, J. Molec. Biol. 138
685 (1980) 321–333.
- 686 [40] V. N. Maiorov, G. M. Crippen, Significance of root-mean-square devia-
687 tion in comparing three-dimensional structures of globular proteins, J.
688 Mol. Biol. 235 (1994) 625–634.
- 689 [41] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, Numer-
690 ical Recipes: The Art of Scientific Computing, Cambridge Univ. Press
691 (Cambridge, UK), 1986.
- 692 [42] N. P. Brown, C. A. Orengo, W. R. Taylor, A protein structure compar-
693 ison methodology, Computers Chem. 20 (1996) 359–380.
- 694 [43] A. Aszódi, W. R. Taylor, Hierarchical inertial projection: a fast distance
695 matrix embedding algorithm., Computers Chem. 21 (1997) 13–23.
- 696 [44] W. R. Taylor, A. C. W. May, N. P. Brown, A. Aszódi, Protein structure:
697 Geometry, topology and classification, Rep. Prog. Phys. 64 (2001) 517–
698 590.
- 699 [45] W. R. Taylor, V. Chelliah, S. M. Hollup, J. T. MacDonald, I. Jonassen,
700 Probing the “dark matter” of protein fold-space, Structure 17 (2009)
701 1244–1252.
- 702 [46] A. Aszódi, W. R. Taylor, Folding polypeptide α -carbon backbones by
703 distance geometry methods, Biopolymers 34 (1994) 489–506.

704 *Acknowledgements:*. The work was supported by the Francis Crick Institute
705 under awards: FC001179 (WRT), FC001162 (JPS) and FC001178 (IAT). The
706 Crick receives its core funding from Cancer Research UK, the UK Medical
707 Research Council, and the Wellcome Trust.