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]*Molecular Structure Laboratories,*

Abstract

to be written

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.

Gag is the major structural protein of both Ortho and Foamy viruses and also displays both important differences and similarities. Ortho and Foamy viral Gag are required for particle assembly, budding from the cell, reverse transcription and delivery of the viral nucleic acid into the newly infected

[1]This 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".

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. 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 but in spumaviruses Gag processing does not occur.

The recent solution of the Foamy Gag protein structure has shed new light on this relationship by revealing that the capsid structures of both viral classes share a common protein fold, with the implication that their gag proteins may be evolutionarily related [1]. An intriguing aspect of this relationship was an ambiguity in the degree of relatedness between the two domains of the gag proteins, with the Spumaretroviral Gag domains appearing almost equally similar to both the amino- and carboxy-terminal domains of the orthoretroviruses. In this paper, we investigate the nature of this relationship in greater detail and discuss its evolutionary implications.

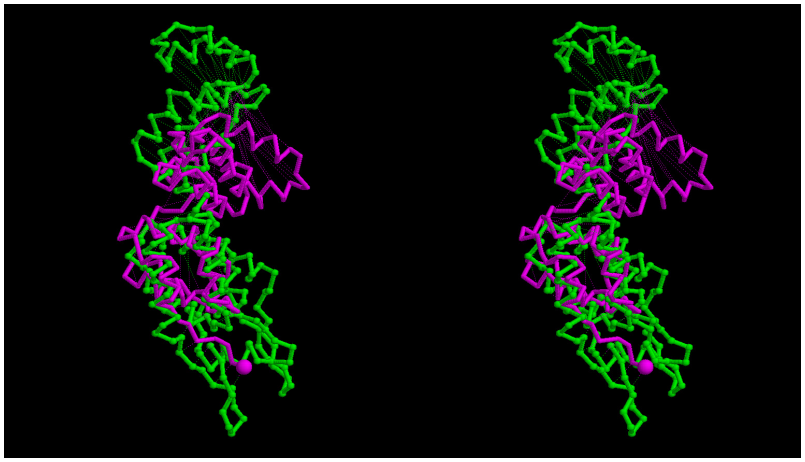
2. Results

2.1. Full-length comparison

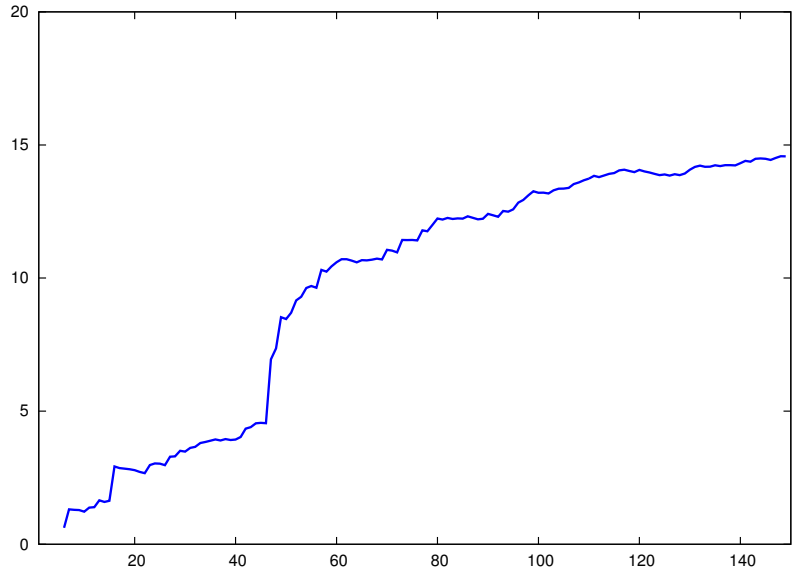
To investigate the structural relationship between the capsid structure of the ortho viruses (HIV, MLV, etc.), and the new structure of the foamy virus capsid [1] (PDB codes: 5m1g, 5m1h), the foamy virus structure was compared to one of the few full double domain ortho virus structures, the HIV capsid with PDB code: `3nte`, using the flexible superposition program `SAP` [2]. Even though this program has a tolerant approach to relative domain shifts, the comparison produced a high RMSD value of 14Å over the 100 best superposed positions. The amino (N) terminal domain positions roughly corresponded but shifts in the relative orientation of the carboxy (C) terminal domain resulted in large deviations between equivalent helices. The superposed structures are shown in Figure 1(a) and the domain divergence can be seen clearly as a jump in the cumulative RMSD plot (Figure 1(b)).

2.2. DALI searches

Although this initial superposition (Figure 1) did not appear encouraging, the foamy virus structure was scanned across the Protein DataBank (PDB), using the `DALI` program [3] to search for any similarities.



(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.

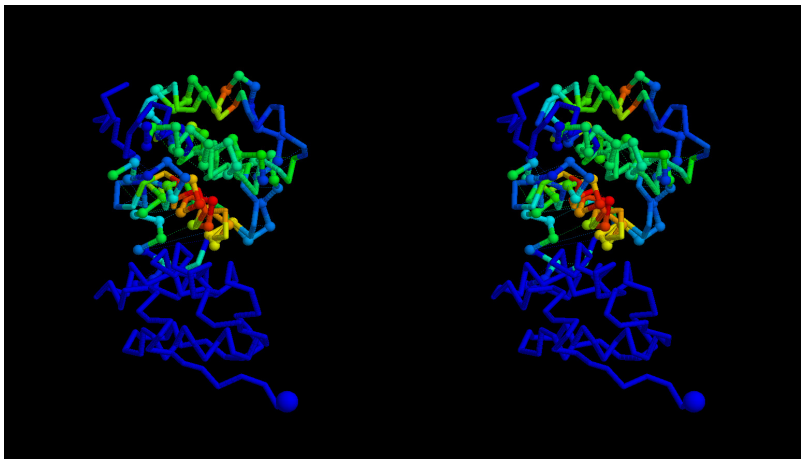
No:	Chain	Z	rmsd	lali	nres	%id	PDB	Description
1:	4x3x-A	5.0	3.1	66	82	11	PDB	MOLECULE: ACTIVITY-REGULATED CYTOSKELETON-ASSOC
2|	3g29-A	3.7	2.7	60	77	8	PDB	MOLECULE: GAG POLYPROTEIN;
3|	3g0v-A	3.7	2.9	62	76	8	PDB	MOLECULE: GAG POLYPROTEIN;
4:	2v50-D	3.6	2.2	41	998	7	PDB	MOLECULE: MULTIDRUG RESISTANCE 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.

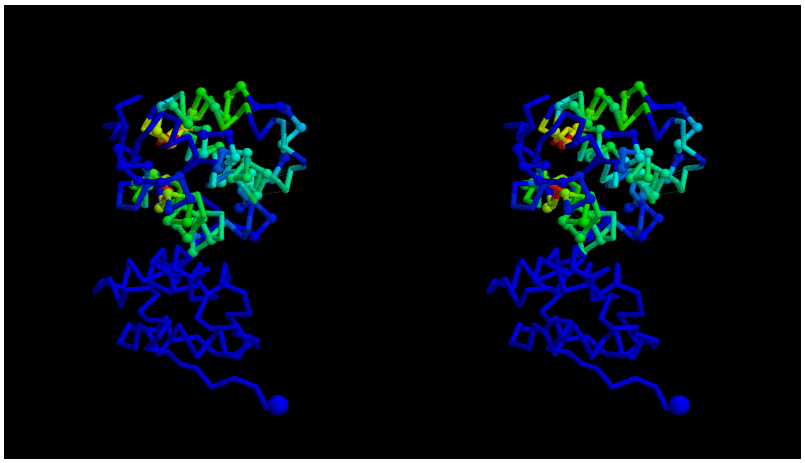
2.2.1. Full chain scan

A scan of the full-length foamy structure using the DALI server[2] over the 90% non-redundant protein structure databank identified a wide selection of retroviral capsid structures. In the ranked list of structure hits, capsids were identified from position 2 to position 550. The top hits are shown in Figure 2 (See Supplementary material for a summary of the full 550 with Z-scores over 2). Many capsids are found in the top 20 hits and although the top scoring hit is not obviously a capsid protein, it is thought to have originated from the Ty3/Gypsy retrotransposon family gag gene [4]. 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.

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

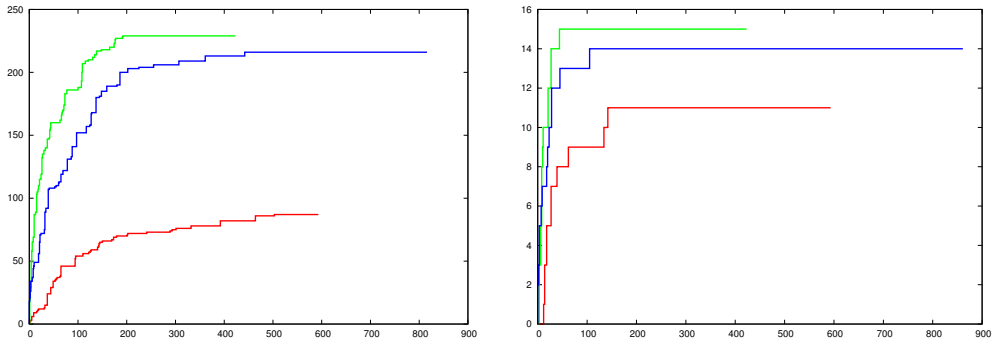


(a) `4x3x-A`



(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 scarcoma virus (PDB code: `3g29-A`).



(a) Full PDB (b) PDB-90

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.

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 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 [5]) were scanned separately

using the DALI program. The individual domains were much more specific at matching known capsid structures[3], 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 their opposed counterpart. The structure-based sequence alignments of each domain based on this equivalence are shown in Figure 5.

Although domain transposition is not impossible in viral genomes, it is sufficiently unexpected to warrant deeper investigation, especially as it is hard to imagine how an ancestral capsid protein could tolerate such a large rearrangement and still pack to form a competent shell. We therefore undertook a more thorougher evaluation using alternative methods to assess the statistical significance of these structural similarities.

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 backbone[4]). 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

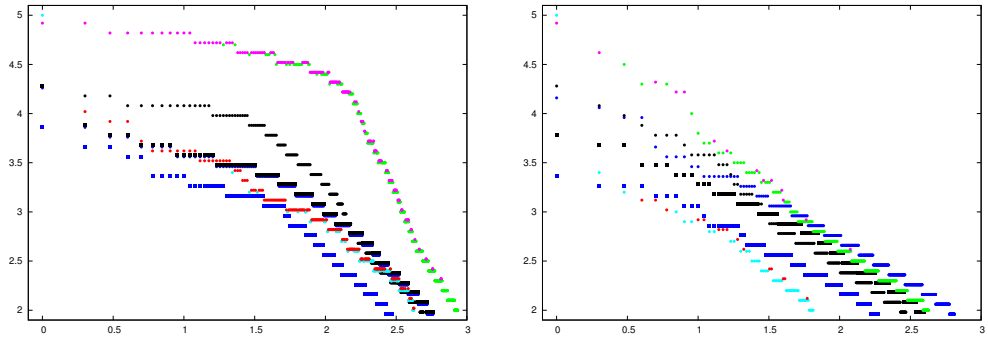
[3]True/false hits were defined by protein descriptions with the words "CAPSID", "GAG" or "P24".

[4]Note 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.

```
Nter    PIGTVIPIQHIRSVTGEPPRNPREIPIWLGRNAPAIDGVFPVTTPDLRCRIINAILGGNIGLSLTPGDCLTWDSAVATLFIRTHGTFP
                              :  :: |   |::|        :    ::        :        |    :  :: |
3g1gA   ---------PWAD--IMQGPS--SFVDFANRLIKAVEGSDL-ARAPVIIDCFRQKSQPQQLI--PSTL-TTPGEIIKYVLDRQK----
3tirA   ---------PWAD--IMQGPS--SFVDFANRLIKAVEGSDL-ARAPVIIDCFRQKSQPQQLI-------TTPGEIIKYVLDRQ-----
3g1iA   ---------PWAD--IMQGPS--SFVDFANRLIKAVEGSDL-ARAPVIIDCFRQKSQPQQLI----TLTT-PGEIIKYVLDRQ-----
3g29A   ---------PWAD--IMQGPS--SFVDFANRLIKAVEGS---ARAPVIIDCFRQKSQPQQLI-------TTPGEIIKYVLDRQ-----
3g0vA   ---------PWAD--IMQGPS--SFVDFANRLIKAVEGSAL-ARAPVIIDCFRQKSQPQQLI-------TTPGEIIKYVLDRQ-----
3g29B   ---------PWAD--IMQGPS--SFVDFANRLIKAVEGSNL-ARAPVIIDCFRQKSQPQQLI-------TTPGEIIKYVLDRQ-----
3g1iB   ---------PWAD--IMQGPS--SFVDFANRLIKAVEGSDL-ARAPVIIDCFRQKSQPQQLI-------TTPGEIIKYVLDRQ-----
3g26A   ---------PWAD--IMQGPS--SFVDFANRLIKAVEGS---CRAPVIIDCFRQKSQPQQLI-------TTPGEIIKYVLDRQ-----
3dtjC   ---------SILD--IRQGPK--EPFRDYVDRFYKTLR--VKNW--MTATLLVQNANPD-TILKGPGA--TLEEMMTA-CQGV-----
3dtjB   ---------SILD--IRQGPK--EPFRDYVDRFYKTLR--VKNW--MTATLLVQNANPD-TILKGPGA--TLEEMMTA-CQGV-----
3dtjA   ---------SILD--IRQGPK--EPFRDYVDRFYKTLR--VKNW--MTATLLVQNANPD-TILKGPGA--TLEEMMTA-CQGV-----
3g21A   ---------PWAD--IMQGPS--SFVDFANRLIKAVEGSDL-ARAPVIIDCFRQKSQPQQLI-------TTPGEIIKYVLDRQ-----

Cter    MHQLGNVIKGIVDQEGVATAYTLGMMLSGQNYQLVSGIIRGYLPGQAVVTALQQRLDQEIDNQTRAETFIQHLNAVYEILGLNARGQSIRL
             |:   :|::   : ::::     |     :  :::     ||:: :|:: :::|:  :     :|  || :  :
1l6nA   SPRTLNAWVKVVEEKA-IPMFSALSE---GATPDLNTMLNTVGGHQAAMQMLKETINEEA--EIYKRWIILGLNKIVRMYS------PTSI
3j34U   SPRTLNAWVKVVEEKA-IPMFSALSE--GATPQDLNTMLNTVGGHQAAMQMLKETINEEA--EIYKRWIILGLNKIVRMY-------SPTS
4u0bF   SPRTLNAWVKVVEEKA-IPMFSALSC--GATPQDLNTMLNTVGGHQAAMQMLKETINEEA--EIYKRWIILGLNKIVRMY-------SPTS
4u0bG   SPRTLNAWVKVVEEK--IPMFSALSC--GATPQDLNTMLNTVGGHQAAMQMLKETINEEA--EIYKRWIILGLNKIVRMY-------SPTS
3h4eB   SPRTLNAWVKVVEEK--IPMFSALSC--GATPQDLNTMLNTVGGHQAAMQMLKETINEEA--EIYKRWIILGLNKIVRMY-------SPTS
2jprA   SPRTLNAWVKVVEEKA-IPMFSALSE--GATPQDLNTMLNTVGGHQAAMQMLKETINEEA--EIYKRWIILGLNKIVRMY-----------
1afvB   SPRTLNAWVKVVEEKAVIPMFSALSE--GATPQDLNTMLNTVGGHQAAMQMLKETINEEA--EIYKRWIILGLNKIVRMY-------SPTS
4u0bE   SPRTLNAWVKVVEEK--IPMFSALSC--GATPQDLNTMLNTV-GHQAAMQMLKETINEEA--EIYKRWIILGLNKIVRMY-------SPTS
4u0bK   SPRTLNAWVKVVEEK--IPMFSALSC--GATPQDLNTMLNTVGGHQAAMQMLKETINEEA--EIYKRWIILGLNKIVRMY-------SPTS
4u0bH   SPRTLNAWVKVVEEK--IPMFSALSC--GATPQDLNTMLNTVGGHQAAMQMLKETINEEA--EIYKRWIILGLNKIVRMY-------SPTS
2gonA   SPRTLNAWVKVVEEK-VIPXFSALSE--GATPQDLNTXLNTVGGHQAAXQXLKETINEEA--EIYKRWIILGLNKIVRXYS----------
1afvA   SPRTLNAWVKVVEEKAVIPMFSALSE--GATPQDLNTMLNTVGGHQAAMQMLKETINEEA--EIYKRWIILGLNKIVRMY-------SPTS
```

Figure 5: **Top domain similarity alignments**. The sequence alignments are shown for the top 12 capsid domain matches found by the `DALI` program using the foamy virus capsid N and C domains separately as a query over the full PDB. The sequence of the N-terminal domain (`N-ter`) is shown at the top of the first alignment block and the sequences of the C-terminal domain (`C-ter`) at the top of the second block. The sequences of the ortho-virsuses aligned below these all come from the "swapped" relationship of C and N terminal domains, respectively. These alignments, which are determined by structure not sequence, exhibit no specific similarity beyond what would be expected from aligning similar secondary structures from similar sized domains. (Amino acid identities are marked by a bar and similarities by a colon).



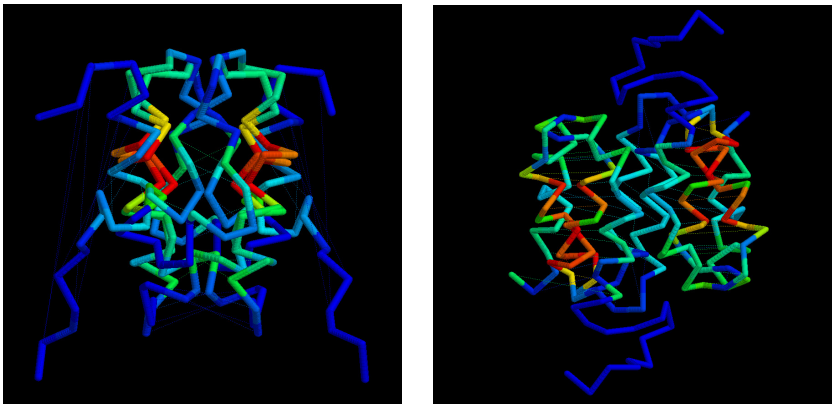
(a) full PDB (b) PDB-90

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).

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



(a) N (b) C

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.

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

However, both the N and C terminal domains pocess a degree of internal symmetry which gives rise to a partial match with their reversed 'doppleganger' decoys. The N-terminal domain superposed on its decoy had an RMSD of 5.4/60 (Å/α-carbons) and 5.5/24 for the C-terminal domain. The higher symmetry of the smaller domain may be sufficient to explain its poor level of specificity seen in Figure 6 and to try and resolve this ambiguity, a more diverse set of decoys were generated based on cyclic permutation and segment swapping combined with chain reversal [6].

2.3.2. Customised decoy comparisons

To improve the statistical analysis of the foamy/ortho capsid similarity, we employed a method based on the generation of a population of customised 'decoy' models to provide a background distribution of unrelated protein scores [6]. This method retains the advantage of the simple reversed structures 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).

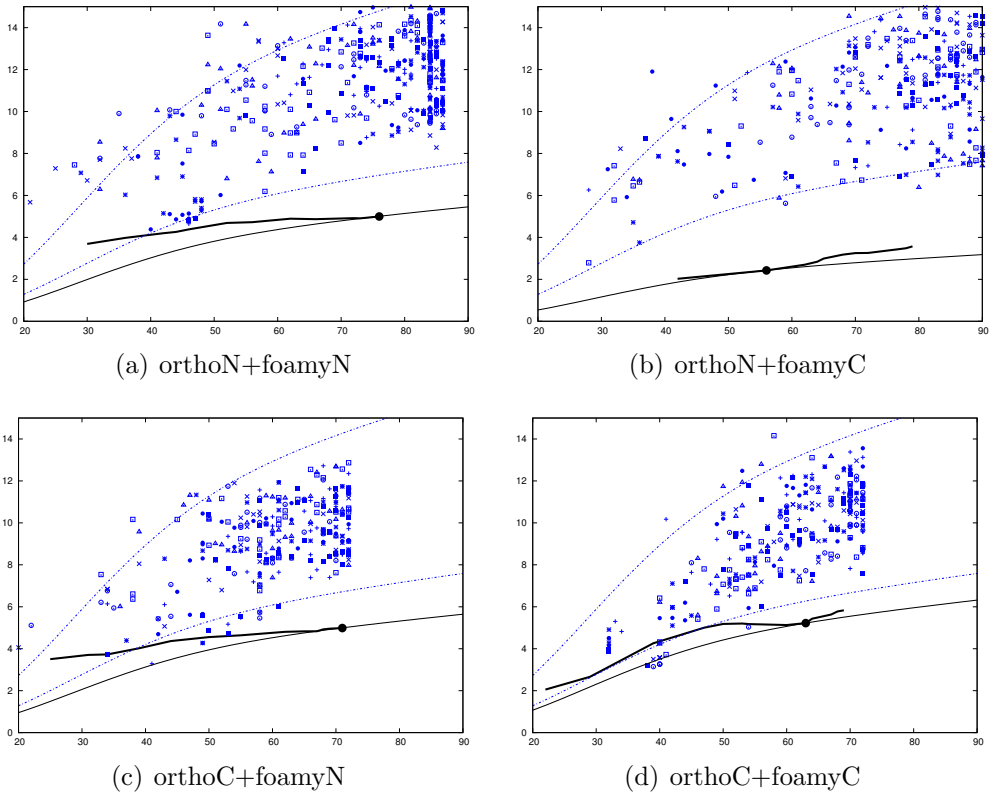
For each domain pair to be compared, decoys were created using cyclic permutation and segment swapping with chain reversal to generate a family of customised decoys for each comparison [6]. All pairs of forward/reversed decoys were then compared, with each pair being drawn from a pool of models generated 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 (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 [7, 6]). 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 (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.

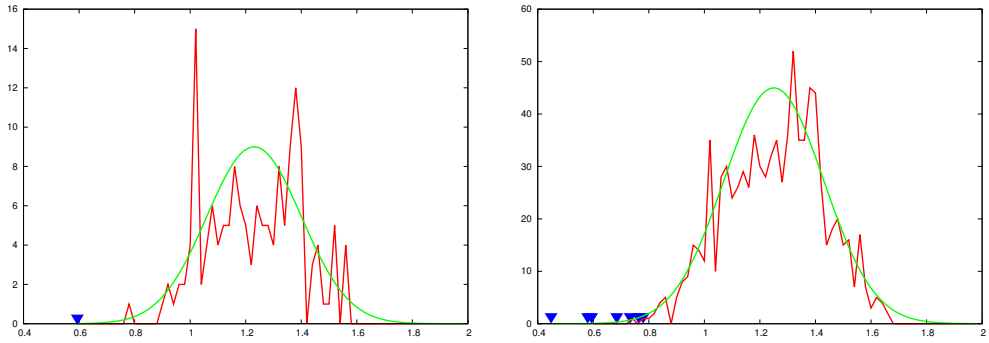
In this way, the significance of all combinations of the native ortho and foamy domain superpositions were calculated, using the background distribution of 'customised' decoy comparisons based on each individual native



(a) orthoN+foamyN (b) orthoN+foamyC

(c) orthoC+foamyN (d) orthoC+foamyC

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 (Equ^n. 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 10^{th} percentile boundaries of the decoy generated distributions, with $\mathtt{a} = 1.7$ (upper) and $\mathtt{a} = 0.8$ (lower). (See Methods section for details).



(a) (b)

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.

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 caluclated 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)).

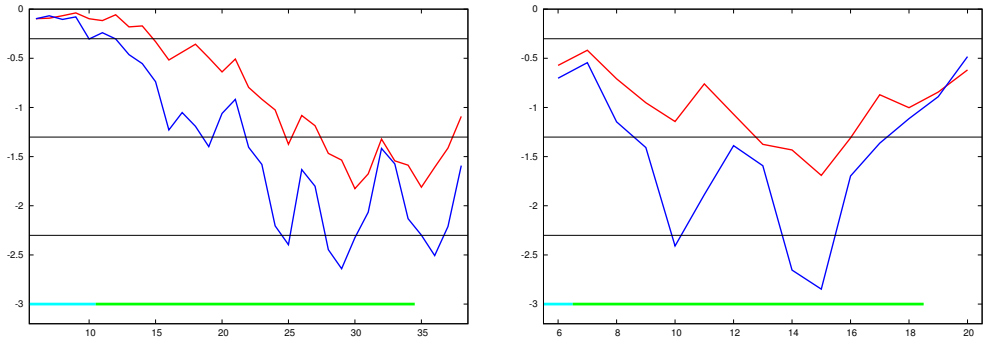
The majority of values in Figure 10(a) lie below the 0.05 probability level for the larger sample sizes, with those for the top-half bias statistic (blue line) being more significant than the moment-based statistic (red line). While confirming the visual trend towards a bias of higher scoring like-type 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 towards a preferred like-domain equivalance, 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 the 0.005 level (Figure 10(b)).

T-test statistic:.

a	**ortho-N**					
	foamy-N			**foamy-C**		
virus	pool	*a*-value	Z-score	pool	*a*-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

b	**ortho-C**					
	foamy-N			**foamy-C**		
virus	pool	*a*-value	Z-score	pool	*a*-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** (Equ^n. 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.



(a) Full data set (b) 5N+5C data set

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 combinfation 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.

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 `DALI` results, now seems less likely. The preferred, and biologically more reasonable, result is that the ortho virus domain are related to the foamy virus

domains as a result of genetic divergence from a common, double domain ancestor.

2.4. Internal duplication

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

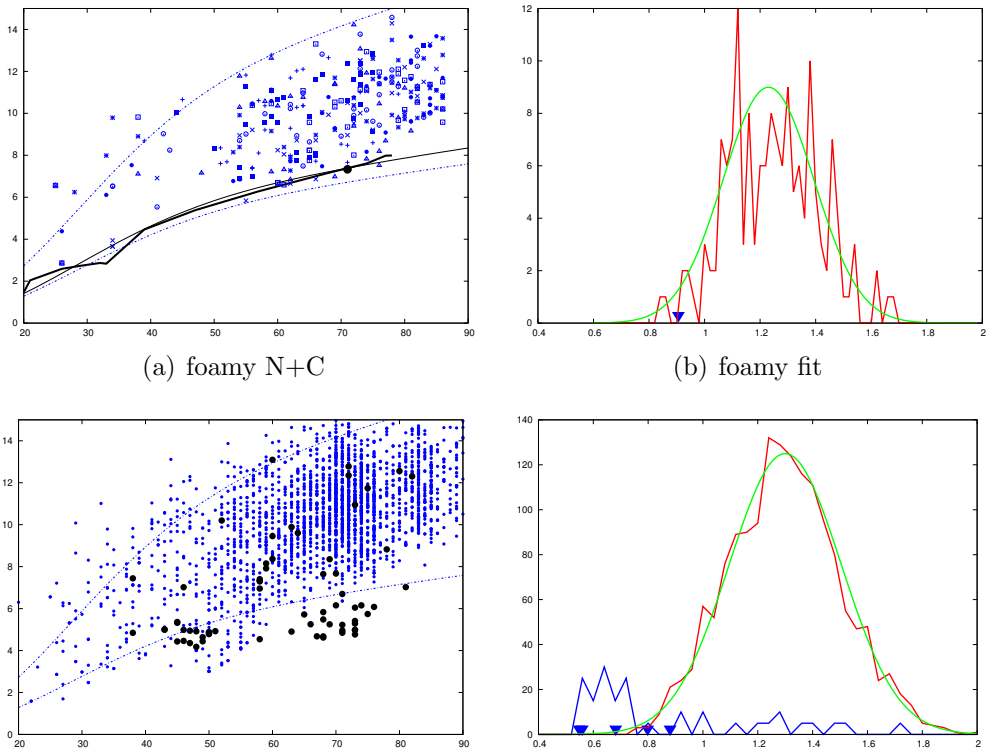
Such a relationship between the foamy domains implies an equivalent relationship in the ortho viruses and a similar comparison in structures of their N and C domains finds matches with Z-scores ranging from 2 to 4. As with the comparison of the ortho and foamy structures, these can be pooled to allow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

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

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

2.5. Fold-space representation

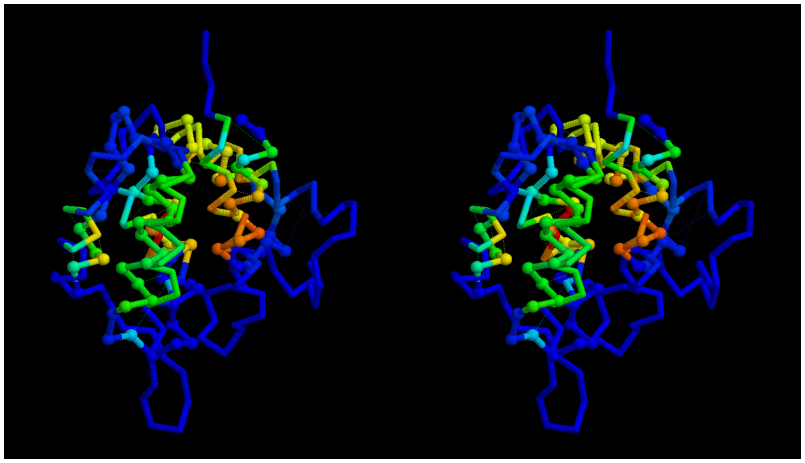
To summarise the structural relationships among the ortho and foamy domains, the matrix of pairwise comparisons was projected into a three-



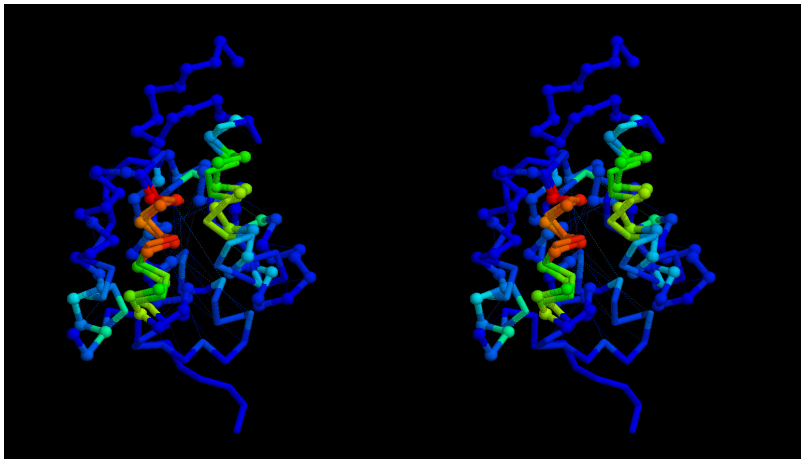
(a) foamy N+C (b) foamy fit

(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.

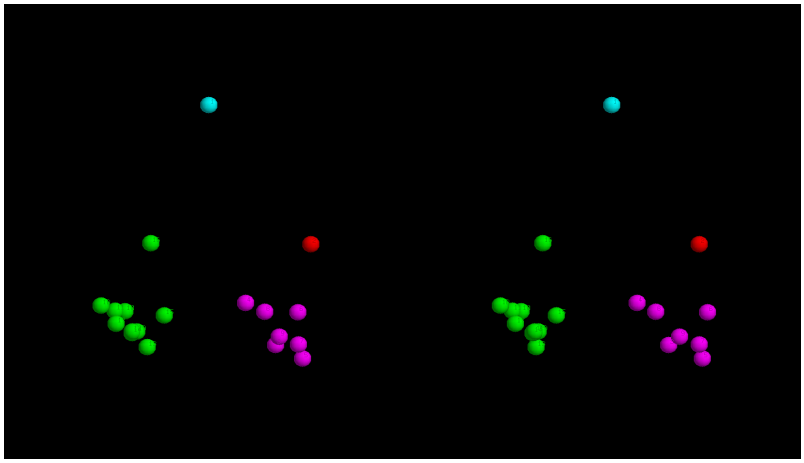


Figure 13: **Fold-space representation of all domains**. All the viral domains considered in the paper were projected into a 3D fold-space representing the relationship of their `SAP` weighted RMSD values. The domains are coloured as: foamyN = cyan, foamyC = red, orthoN = green and ortho C = magenta.

dimensional fold-space. (See methods for details). This produces a best visual representation of the RMSD values between domains.

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

3. Conclusions

3.1. Structure comparison

3.1.1. Pairwise significance

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

The use of the customised decoy sets attempts to avoided this problem by recreating a large number of possible folds that were generated using

the same (reconnected) SSEs. Moreover, to avoid any chance recreation of native fragments, each comparison always involved the comparison of a native (forward) chain direction with a reversed chain. Using these models, a background distribution of decoy/decoy comparisons allowed us to calculate Z-scores for each native/native comparison with the advantage that every comparison in the background distribution involved two models with the same length, density and secondary composition as the native pair. These values indicated a clearly significant relationship between the foamy and ortho structures.

3.1.2. Direct or transposed domain order?

However, the Z-scores did not point to a clear resolution of whether the domains should have a direct correspondance (NN and CC match) or a transposed relationship (NC and CN) with significant individual matches found across all pairings. Testing for a bias towards more significant like-domain pairings (NN,CC) in the list of similarities ranked by Z-score confirmed the visual bias towards a direct correspondance but only at a marginal level of significance (around 0.05). By contrast, the application of a T-test on the combined raw comparison data returned a very clear distinction between the direct and the transposed relationships, clearly favouring the more natural forward order.

Although the 'astronomic' probabilities calculated by the T-test seem very convincing, they must be viewed in the light of the much lower probabilities calculated from the asymmetry statistics. Both calculations involve assumptions and are limited by the small number of known structures so neither can be taken as definitive. It would seem likely that the "true" level of significance may lie somewhere between the two results but as both point in the direction of the NN and CC domain order, there is no reason to adopt the more unexpected transposed domain order.

3.2. Evolutionary implications

On the basis of these structural comparisons, and a variety of functional assays described elsewhere, we can conclude that the central domain of the spumavirus Gag gene encodes a polypeptide sequence related to that of the corresponding region of orthoretroviruses, CA. It therefore seems reasonable to suppose that the last common ancestor of orthoretroviruses and spumaviruses possessed such a sequence. Moreover this region appears to be

made up from two related subdomains suggesting a gene duplication event in a common precursor.

In our initial search of the foamy virus capsid using the DALI program, we made the curious observation that the strongest similarity of the foamy virus capsid was with the ARC protein (Activity-Regulated Cytoskeleton-associated protein) that is active in neural synaptic growth and activity (and several other developmental associated functions). The ARC protein has widespread and clear (non-gag) sequence homologues as far back as insects and probably deeper, giving it a very ancient origin somewhere close to the metazoan root [8]. If ARC is considered to be a relic of an ancient Ty3/Gypsy retrotransposon [4], preserved as a 'living fossil' in the genomes of metazoa, this relationship would suggest an equally ancient origin for the foamy virus.

Alternatively, the foamy viruses may have co-opted an ARC protein to facilitate budding and their escape from the cell. As it is believed that the Ty3/Gypsy family of intracellular retrotransposons gave rise to retroviruses [9], it will therefore be of considerable interest to determine whether such elements possess CA proteins with a two-domain structure. Finally, it is worth noting that the Gag protein of Ty3 is significantly shorter that that of the retroviruses and it is possible that the N-terminal domains of the orthoretroviruses and spumaviruses were co-opted at different times to facilitate budding from the cell surface. If so, the very different structures of this region in the orthoretroviruses and spumaviruses might suggest independent acquisition events.

4. Methods

4.1. Structural data

The foamy virus structures were obtained from the Protein Structure Databank (PDB code:`5M1G`) [1].

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

- **BLV**: bovine leukemia virus (deltaretrovirus) `4PH1` (N-ter.dom) and `4PH2` (C-ter.dom) [10],
- **BLV6**: bovine leukemia virus (hexameric) `4PH0` (both dom.s) [10],
- **HIV1**: human immunodeficiency virus 1 (lentivirus) `1AK4` (N-ter.dom) [11] and `1A43` (C-ter.dom) [12],

- **HIV6**: human immunodeficiency virus 1 `3H47` (both dom.s) [13],
- **HML2**: human endogenous retrovirus type-K (betaretrovirus) [14],
- **HTLV**: human T-cell leukemia virus (deltaretrovirus) `1QRJ` (both dom.s) [15],
- **JSRV**: jaagsiekte sheep Retrovirus (betaretrovirus) `2V4X` (N-ter.dom) [16],
- **MLV**: murine leukemia virus (gammaretrovirus) `1U7K` (N-ter.dom) [17],
- **MPMV**: Mason-Pfizer monkey virus (betaretrovirus) `2KGF` (N-ter.dom) [18],
- **PSIV**: prosimian immunodefficiency virus (ancient lentivirus) `2XGV` (N-ter.dom) [19],
- **RELIK**: rabbit endogenous lentivirus type-K (ancient lentivirus) `2XGU` (N-ter.dom) [19],
- **RSV**: Rous sarcoma virus (alpharetrovirus) `3G1I` (both dom.s) [20].

4.2. Structure comparison

4.2.1. `DALI`

The `DALI` method for searching the PDB with a structural query [3] was accesed via the server at: `http://ekhidna.biocenter.helsinki.fi/dali_server`. The `DALI` method reports the significance of each match with an estimated Z-score which is the raw comparison score, normalised by the combined length of the proteins. Z-scores down to a value of 2 are reported by the program.

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

`DALI` reports structural hits in both the full PDB and a reduced collection of structures that have no pair of proteins with over 90% sequence identity, referred to as the 90% non-redundant or PDB-90 collection. It was found, however, that some hits, seen in the full PDB were not found in the PDB-90,

for example in Figure 6, all of the top 31 hits of the N-domain against the full PDB are missing in the PDB-90 hits. The most likely explanation is that the PDB-90 secection has not been updated at the same time as the full collection. For this reason, hits to both databases were monitored.

4.2.2. `SAP`

The `SAP` method for structure comparison [2] was run as local copy which can be accessed at: `mathbio`. As part of determining the alignment between two structures, the `SAP` program calculates a similarity score for each pair of matched positions which is how similar the rest of the structure looks from the viewing-frame of the superposed residues. This value can be used both to weight the importance of positions when calculating the (rigid-body) RMSD superposition and to colour positions in the superposed structures. [21]. (As in Figure 3).

If the matched positions are ranked by this value, then RMSD values can be calculated over increasingly larger subsets to high-light the extent of a well matched core before the contribution of variable loops, or domain shifts, 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: `mathbio`.

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 by Taylor [6], 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 swapped. That is: if a chain runs from amino (N) to carboxy (C) termini through three adjacent loop regions `a-b`, `c-d` and `e-f` as: N,`a-b,c-d,e-f`,C then the swapped chain runs: N,`a-d,e-b,c-f`,C with each switch being made at the least disruptive point. This swap 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 compared with six permutations of a reversed foamy domain that includes a single loop reconnection.

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

4.4. Statistical tests

4.4.1. RMSD length normalisation

The quality of structure comparisons can be characterised by a combination of their RMSD value and the number of matched (superposed) positions. How to combine these values has been the subject of much discussion over the years and central to this is the expected random RMSD value for two proteins of a given length [22, 23, 24]. However, when reviewed [6], all these measures were approximations of a simple square-root function of the protein length (as originally proposed by McLachlan on theoretical grounds [22]) but with an added term to depress the RMSD values obtained with small units or structure that are dominated by secondary structure elements (and super-secondary structure motifs) giving a lower than expected RMSD value. The formula that best captures this is: $R = \sqrt{N}(1 - \exp(-N^2/s^2))$, where, R is

the expected random RMSD for N matched positions and s is the damping factor in the inverted Gaussian term (equivalent to the standard deviation in the Normal distribution).

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

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

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

The best value for s is slightly dependent on the nature of the proteins being compared. For artifical (random-walk) models with no secondary structure, no modification will be needed but the proteins considered here have segments of packed alpha helices that can be locally similar over two to three helices. To correct for this, a value of $s = 30$ was used (or $1/s^2 = 0.11$) which is higher than the value of $1/s^2 = 0.03$ used previously. That this is a reasonable fit to the data can be seen in the way the dashed blue lines in Figure 8 track the upper and lower boundary of the decoy comparison results.

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

4.4.2. Frequency plots

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

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

4.4.3. T-tests

Data from separate native/native comparisons, with their customised decoy data, were combined giving not only a much larger background decoy derived population of scores but also a smaller distribution of native comparison scores that can be tested to calculate the probability that they were drawn from the same population as the decoy data. To do this, a T-test was used which takes the size, mean, and standard deviation of each distribution and calculates a probability. The implementaion of this test was taken from the Numerical Reicpies collection [25] which implements one of two variants of the test depending on whether the distributions have statistically distinct standard deviations. (Routines `ttest()` and `tutest()`). The choice of routine is based on a preapplication of an F-test on the standard-deviations. (Using the routine `ftest()`).

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

4.5. Fold-space clustering

The results of the pairwise similarity within a set of structures can be visualised by treating the RMSD values as Euclidean distances[5] and reducing their dimensionality to sufficiently few dimensions to be visualised: usually 2D or, better 3D, to visualise the space with less distortion. Rather than use a simple multi-dimensional scaling (MDS) method ([26]), the more complicated method of multi-dimensional projection was used ([27], see [28] for a simpler exposition).

This method reduces the dimensionality of the projection in gradual stages with each step employing triangle-inequality balancing and hyper-

[5]In 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).

dimensional real-space refinement. In the real-space refinement stages, a weight can be applied to pairwise distances. (This cannot be done in direct MDS projection, which can only assign a mass to each point). Weights were assigned to distances as a function of their inverse RMSD, up to a maximum value of 1.

The method is robust and has been widely applied to rough models ([29]) and predicted inter-residue distances that constitute highly non-metric data sets ([30]).

References

[1] N. Ball, G. Nicastro, M. Dutta, D. Pollard, D. Goldstone, M. Sanz-Ramos, A. Ramos, E. Mllers, K. Stirnnagel, N. Stanke, D. Lindemann, J. Stoye, W. Taylor, P. Rosenthal, I. Taylor, Structure of a spumaretrovirus gag central domain reveals an ancient retroviral capsid, PLoS Path. ? (2016) ? Submitted.

[2] W. R. Taylor, Protein structure alignment using iterated double dynamic programming, Prot. Sci 8 (1999) 654–665.

[3] L. Holm, C. Sander, Protein-structure comparison by alignment of distance matrices, J. Molec. Biol. 233 (1993) 123–138.

[4] W. Zhang, J. Wu, M. Ward, S. Yang, Y. Chuang, M. Xiao, R. Li, D. Leahy, P. Worley, Structural basis of arc binding to synaptic proteins: implications for cognitive disease., Neuron 86 (2015) 490–500.

[5] W. R. Taylor, Protein structure domain identification, Prot. Engng. 12 (1999) 203–216.

[6] W. R. Taylor, Decoy models for protein structure score normalisation, J. Molec. Biol. 357 (2006) 676–699.

[7] M. Levitt, M. Gerstein, A unified statistical framework for sequence comparison and structure comparison, Proc Natl Acad Sci USA 95 (1998) 5913–5920.

[8] M. Campillos, T. Doerks, P. Shah, P. Bork, Computational characterization of multiple gag-like human proteins, Trends in Genetics 22 (2006) 285–589.

[9] C. Llorens, M. Fares, A. Moya, Relationships of gag-pol diversity between Ty3/Gypsy and retroviridae LTR retroelements and the three kings hypothesis., BMC Evol. Biol. 8 (2008) e276.

[10] G. Obal, F. Trajtenberg, F. Carrion, L. Tome, N. Larrieux, X. Zhang, O. Pritsch, A. Buschiazzo, Conformational plasticity of a native retroviral capsid revealed by X-ray crystallography., Science 349 (2015) 95–98. DOI: 10.1126/science.aaa5182.

[11] T. Gamble, F. Vajdos, S. Yoo, D. Worthylake, M. Houseweart, W. Sundquist, C. Hill, Crystal structure of human cyclophilin A bound to the amino-terminal domain of HIV-1 capsid., Cell 87 (1996) 1285–1294.

[12] D. Worthylake, H. Wang, S. Yoo, W. Sundquist, C. Hill, Structures of the HIV-1 capsid protein dimerization domain at 2.6å resolution., Acta Crystallogr.,Sect.D 55 (1999) 85–92. DOI: 10.1107/S0907444998007689.

[13] O. Pornillos, B. Ganser-Pornillos, B. Kelly, Y. Hua, F. Whitby, C. Stout, W. Sundquist, C. Hill, M. Yeager, X-ray structures of the hexameric building block of the HIV capsid., Cell 137 (2009) 1282–1292. DOI: 10.1016/j.cell.2009.04.063.

[14] G. Mortuza, M. Dodding, D. Goldstone, L. Haire, J. Stoye, I. Taylor, Structure of B-tropic MLV capsid N-terminal domain., J. Mol. Biol. 376 (2008) 1493–1508.

[15] S. Khorasanizadeh, R. Campos-Olivas, C. Clark, M. Summers, Sequence-specific 1H, 13C and 15N chemical shift assignment and secondary structure of the HTLV-I capsid protein., J. Biomol. NMR 14 (1999) 199–200.

[16] G. B. Mortuza, D. C. Goldstone, C. Pashley, L. F. Haire, M. Palmarini, W. R. Taylor, J. P. Stoye, I. A. Taylor, Structure of the capsid amino terminal domain from the betaretrovirus, Jaagsiekte sheep retrovirus., J. Molec. Biol. 386 (2009) 1179–1192.

[17] G. B. Mortuza, L. F. Haire, A. Stevens, S. J. Smerdon, J. P. Stoye, I. A. Taylor, High-resolution structure of a retroviral capsid hexameric amino-terminal domain., Nature 431 (2004) 481–485.

[18] P. Macek, J. Chmelik, I. Krizova, P. Kaderavek, P. Padrta, L. Zidek, M. Wildova, R. Hadravova, R. Chaloupkova, I. Pichova, T. Ruml, M. Rumlova, V. Sklenar, NMR structure of the N-terminal domain of capsid protein from the mason-pfizer monkey virus, J. Mol. Biol. 392 (2009) 100–114. DOI: 10.1016/j.jmb.2009.06.029.

[19] D. C. Goldstone, M. W. Yap, L. E. Robertson, L. F. Haire, W. R. Taylor, A. Katzourakis, J. P. Stoye, I. A. Taylor, Structural and functional analysis of prehistoric lentiviruses uncovers an ancient molecular interface., Cell Host Microbe 8 (2010) 248–259.

[20] G. Bailey, J. Hyun, A. Mitra, R. Kingston, Proton-linked dimerization of a retroviral capsid protein initiates capsid assembly, Structure 17 (2009) 737–748. DOI: 10.1016/j.str.2009.03.010.

[21] F. Rippmann, W. R. Taylor, Visualization of structural similarity in proteins, J. Molec. Graph. 9 (1991) 3–16.

[22] A. D. McLachlan, How alike are the shapes of two random chains?, Biopolymers 23 (1984) 1325–1331.

[23] F. E. Cohen, M. J. E. Sternberg, On the prediction of protein structure: the significance of the root-mean-square deviation, J. Molec. Biol. 138 (1980) 321–333.

[24] V. N. Maiorov, G. M. Crippen, Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins, J. Mol. Biol. 235 (1994) 625–634.

[25] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, Numerical Recipes: The Art of Scientific Computing, Cambridge Univ. Press (Cambridge, UK), 1986.

[26] N. P. Brown, C. A. Orengo, W. R. Taylor, A protein structure comparison methodology, Computers Chem. 20 (1996) 359–380.

[27] A. Aszódi, W. R. Taylor, Hierarchical inertial projection: a fast distance matrix embedding algorithm., Computers Chem. 21 (1997) 13–23.

[28] W. R. Taylor, A. C. W. May, N. P. Brown, A. Aszódi, Protein structure: Geometry, topology and classification, Rep. Prog. Phys. 64 (2001) 517–590.

[29] W. R. Taylor, V. Chelliah, S. M. Hollup, J. T. MacDonald, I. Jonassen, Probing the "dark matter" of protein fold-space, Structure 17 (2009) 1244–1252.

[30] A. Aszódi, W. R. Taylor, Folding polypeptide α-carbon backbones by distance geometry methods, Biopolymers 34 (1994) 489–506.

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