Are viruses a source of new protein folds for organisms? – Virosphere structure space and evolution

Aare Abroi[1)]* *and Julian Gough*[2)]

A crucially important part of the biosphere – the virosphere – is too often overlooked. Inclusion of the virosphere into the global picture of protein structure space reveals that 63 protein domain superfamilies in viruses do not have any structural and evolutionary relatives in modern cellular organisms. More than half of these have functions which are not virus-specific and thus might be a source of new folds and functions for cellular life. The number of viruses on the planet exceeds that of cells by an order of magnitude and viruses evolve up to six orders of magnitude faster. As a result, cellular species are subject to a constitutive 'flow-through' of new viral genetic material. Due to this and the relaxed evolutionary constraints in viruses, the transfer of domains between host-to-virus could be a mechanism for accelerated protein evolution. The virosphere could be an engine for the genesis of protein structures, and may even have been so before the last universal common ancestor of cellular life.

Keywords:
- evolution; protein structure; superfamily; virosphere; virus

DOI 10.1002/bies.201000126

[1)] Estonian Biocentre, Tartu, Estonia
[2)] Department of Computer Science, University of Bristol, United Kingdom

***Corresponding author:**
Aare Abroi
E-mail: aabroi@ebc.ee

Supporting information online

Introduction

Research from the last two decades has widened our understanding of viruses revealing an incredible diversity and abundance. For example there is the discovery of giant viruses like the mimivirus, isolated from amoeba, which have a genome larger than some bacteria [1]. The mimivirus' physical size is comparable to some small bacteria [1]. Mimiviruses are not alone – there are the giant close relatives of marseillevirus [2] and mamavirus [3], both infecting amoeba. The discoveries mentioned above challenge the clarity of the borders between viruses and cellular organisms (many aspects currently reviewed in ref. [4]). Another example of our widening understanding is found in the widespread existence of 'gene transfer agents' (GTA)-particles mediating genomic DNA transfer between cells – and their importance in the bacterial population. This further blurs the borders of viruses, in this case with mobile elements [5]. Also in marine environments, the abundance and diversity of viruses has been found to be extremely high (reviewed in refs. [6–8]). In different water environments the ratio of viral particles to prokaryotes varies between 5 (in lakes) and 100 (in deep ocean waters) ([9] and references therein), with 10^6-10^9 viral particles per millilitre of sea water [8]. On average, prokaryotes represent 90% of ocean biomass and viruses 94% of nucleic acid containing particles [9]. There are more than 5,000 viral genotypes or species in 100 L of sea water [7]. The sequence diversity and uniqueness of viral metagenomics data are remarkable – usually about 60% (or more) of DNA reads did not encode proteins that were significantly similar to known genes (reviewed in ref. [10]). Both the diversity and abundance of viruses and the discovery of viruses carrying proteins from photosynthesis complexes PSII and PSI have led to the acceptance of viruses as an important and integral part of the biosphere [7, 9, 11–14]. These new discoveries have led to an intensifying discussion on the positioning of the viruses (and viral genes) relative to the tree of life (TOL) and what the viruses are and what they are not ([15] and correspondence [16–21], as well as [22, 23]). In addition to being an important part of the biosphere in its own right, the virosphere is also a source of a great deal of new

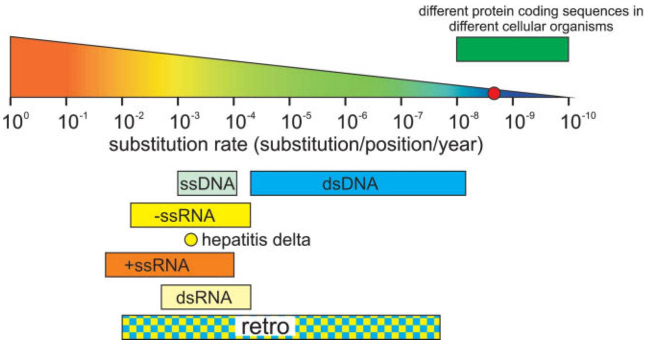

Figure 1. The ranges of substitution rates for coding sequences of viruses with different genomic architectures. Substitution rates are given as substitutions per site per year on a logarithmic scale. Data for viruses from Fields Virology [33] with updates from [34–39]. The red circle indicates the average value for mammalian nuclear coding sequences [40].

knowledge in the general field of molecular biology, as illustrated in a recent review by Enquist [24].

Since the estimated number of viral particles exceeds the number of cells by an order of magnitude, given the infectious nature of viruses, organisms are subject to a continuous ‘flow-through’ of viral genetic material [9]. Most likely every human has personally touched a remarkable abundance and diversity of viruses. Every person has been, or is, infected with viruses in more or less pathological or non-pathological ways. Due to the intensive ‘flow-through’ of genes, even the endogenisation of non-retroviral (non-reverse-transcribing) RNA virus proteins by eukaryotes has occurred; this was described for humans and yeast [25, 26], and recently for many vertebrates [27–29]. Since there are no DNA intermediates in their replication cycle this observed integration of genes from RNA viruses appears to be a very unlikely event, whereas the integration of genes from DNA viruses would appear more likely and thus is expected to be more common. Importantly, viruses have much faster evolutionary rates then their hosts and the sequence similarity may disappear very quickly (Fig. 1 and [30–32]). Depending on the particular virus, the evolution of viral coding genes is one to five orders of magnitude faster than their host, either measuring mutation per position per replication cycle or substitution per site per year (Fig. 1, also illustrated in [30–32]). Viruses have been recognised as horizontal gene transfer (HGT) vesicles in bacterial communities, however their fast evolving nature is often not taken into account in estimating their evolutionary role. Differences in evolutionary rates of viruses and their hosts of five orders of magnitude (quite a typical situation for ssRNA viruses) can be illustrated a follows: the sequence space which the average nuclear gene of vertebrates has been able to sample from the time of the Cambrian explosion until today, could have been sampled by a viral gene during the written history of humankind.

It is well known that protein structure is much more conserved than sequence ([41], quantified in ref. [42]). Thus, the interference between cellular and viral protein domains should be studied at a structural level to detect these more distant evolutionary relationships. The evolution of protein structural domains has been examined in several excellent studies, however, viruses are simply excluded from these analyses [43–45], with the exception of [46]. The exclusion of viruses is a major oversight as they are an important part of the biosphere and may contribute to the evolution of protein domains. The extent of the domain transfer or domain exchange between the virosphere and cells is an important factor in understanding how viruses have shaped the evolution of cellular organisms. To help integrate viruses into the global picture of protein evolution (and species evolution), we characterised the interference (or overlap) of cellular and virosphere structure space based on common ancestry using the SUPERFAMILY resource (Box 1) for domains of known structure in genomes (www.supfam.org, [47, 48]). The inclusion of viruses into the picture leads to interesting conclusions and many open questions, as described further in this essay.

There exist protein domains in viruses which have no common ancestor in cellular organisms

In this section we are able, via protein structure, to identify domains in the virosphere which are evolutionarily distinct from anything seen in cellular life. This shows that viruses have the capability to generate new protein folds de novo.

The SUPERFAMILY database contains the genomic assignment of SCOP protein domains at the SCOP SF level for all completely sequenced genomes (Box 1). The SCOP is a hierarchical classification of protein structural domains and groups together those domains which have structural, functional and sequence evidence for a common evolutionary ancestor at the SF level [49]. Although proteins easily diverge beyond the point where there is any detectable sequence similarity, the close packing of the side-chains in the buried core of the 3D structure retains the same recognisable form. To compare proteins across the full range of evolutionary distances it is necessary to consider SF domains, the fundamental units of ancestry. Throughout this paper, we use the terms fold, SF and family as they are defined in SCOP. In SUPERFAMILY release 1.73 (based on SCOP 1.73) there are 1,304 cellular genomes: 67 archaea; 903 bacteria; 334 eukaryota containing assignments to 1,736 SFs with a significant *E*-value. To gain an initial insight we analysed the presence of different SFs in the three superkingdoms of life (archaea, bacteria and eukaryota) versus those in the virosphere. We must note that according to current knowledge the viruses do not have a common viral Last Universal Common Ancestor (LUCA) as cellular organisms do and from that point of view the virosphere is not necessarily a phylogenetically equivalent classification level (or taxonomic group) when compared to the three superkingdoms. In the virosphere we found 560 SFs. All three superkingdoms share SFs with the virosphere and 30-39% of SFs in different superkingdoms are shared with viruses (Fig. 2A)[a].

[a] The primary data used to generate Fig. 2A, B, Fig. 3A, B and the Supporting Information video are available in public SUPERFAMILY MySQL database in http://www.supfam.org/SUPERFAMILY/downloads.html. (However, not in an explicit way). Respective queries were run in MySQL.

Box 1

GTA – gene transfer agent. A virus-like element that contains random pieces of the host chromosome. They are encoded by the host genome.
HGT – horizontal gene transfer
LUCA – last universal common ancestor
SCOP – structural classification of proteins. Hierarchical classification of protein structural domains. Hierarchies from parent to child: fold class (α,β etc); fold, superfamily, family. (http://scop.mrc-lmb.cam.ac.uk/scop/index.html).
Superfamily – hierarchical level of SCOP grouping together with those structural domains which have structural, functional and sequence evidence for a common evolutionary ancestor. The highest level of SCOP with a confidence for evolutionary relationship.
SUPERFAMILY – a resource which uses a library of HMMs to assign domains of known structure to protein sequences based on the SCOP classification. The resource provides assignments to all completely sequenced genomes, viruses, plasmids, etc. (www. supfam.org).
VspSF – virosphere specific superfamily. Superfamily found in virosphere and not assigned to (not found in) any of cellular genomes.
UniProt – The Universal Protein Resource is a comprehensive resource for protein sequence and annotation data. Also called 'Protein KnowledgeBase'. The UniProt gives access to all the protein sequences which are available to the public. This is a redundant database. (www. uniprot.org).
UniProt viral – viral subset of UniProt (redundant).
NCBI viral genomes – a curated and nonredundant set of viral genomes (www.**ncbi**.nlm.nih.gov/**genomes**/VIRUSES/viruses.html).
ICTV – International Committee on Taxonomy of Viruses (http://www.ictvonline.org/).
PfamA and PfamB – a database of protein sequence families and domains. PfamA entries are high quality, manually curated families. The automatically generated entries are called PfamB. (http://pfam.sanger.ac.uk/).
SRS – integrated system for molecular biology data retrieval and applications for data analysis from multiple databases (http://srs.ebi.ac.uk/).
HMM – Hidden Markov Model. In this context statistical model representing multiple sequence alignment.
HHSearch – profile-profile comparison software which can score two HMMs against each other.
PDB – Protein Data Bank – an Information Portal to Biological Macromolecular Structures (www.pdb.org).
CASP8 – 8th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (http://predictioncenter.org/).

Although the majority of SFs in the virosphere are also found in cellular organisms, 63 SFs are not assigned to any cellular genomes and thus are virosphere-specific SF (VspSF) (Supporting Information Table 1)[b]. Therefore, according to the SCOP definition of a domain SF there is no evidence that they originate from contemporary cellular organisms. The ratio of VspSFs to the number of SFs found in the virosphere is comparable to the ratio of superkingdom specific SFs in bacteria but lower than that in eukaryotes (Fig. 2A). However, analysing overlaps between superkingdoms in more detail, a more complex picture appears, as shown in the Venn diagram (Fig. 2B). Only 9% of SFs which are specific to Archaea (within cellular life) are also found in viruses, compared to 42% of SFs found in all three superkingdoms (Fig. 2B and Supporting Information video). We see that the SFs which are specific to one or two cellular superkingdoms are far less well represented in viruses than SFs which are common to all superkingdoms.

In summary, out of the 560 SFs observed in viruses, a significant number (63) are not observed in any cellular organism, establishing the existence of virosphere specific SFs.

The observed virosphere specificity of protein domains is not an artefact of the data of our analysis

In this section, we examine historically the effect of increasing data coverage, showing that our observation will not be overturned in the future by our knowledge of protein structure and genome sequence reaching completeness. We also show that assignment errors are not critical.

We took the SFs found currently in the virosphere and tested their virosphere specificity as the number of sequenced genomes increased. Beginning with an early release of SUPERFAMILY when there were genomes from 57 species, we found 134 SFs from the used set of SFs (in SUPERFAMILY 1.75) found in the virosphere that were not found in these cellular genomes at that time (Fig. 3A). An almost ten-fold increase in the number of genomes up to 2007 reduced the number of VspSF twofold, but most of this decrease was during the year following our start point. During recent years (2007-2010) the number of genomes doubled but only resulted in an ~10% decrease in VspSFs. The increased sequencing of genomes and the number of VspSFs does not appear to lead to the complete disappearance of anything specific to the virosphere in the future. Furthermore, these data were produced on the current virosphere SFs compared to historical cellular genomic data, in reality the structural characterisation of new virosphere proteins will also increase, leading to new VspSFs over time.

[b] Supporting Information Table 1 was compiled using public data from SUPERFAMILY MySQL database, NCBI viral genomes (http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=10239) and ViralZone (www.expasy.ch/viralzone).

A)

	no of SF found	in viruses		kingdom specific
		no	%	
all	1799	560	31	
A	875	342	39	2.6
B	1316	437	33	12.1
E	1524	464	30	23.5
V	560			11.2

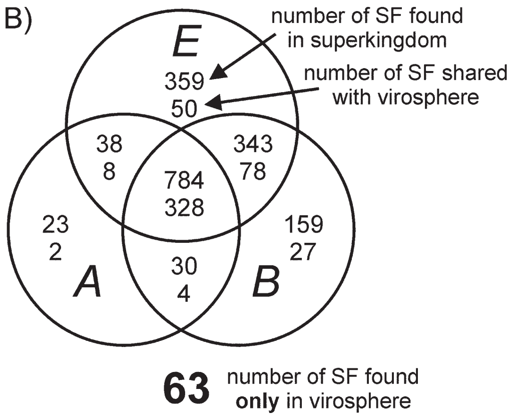
B)
E
number of SF found in superkingdom
359
50
number of SF shared with virosphere
38
8
343
78
784
328
23
2
159
27
30
4
A
B
63
number of SF found only in virosphere

Figure 2. Distribution of SFs found in the virosphere. **A:** number of SFs shared by the three superkingdoms and the virosphere (represented by 'NCBI viral genomes' – a curated and nonredundant set of viral genomes) and the fraction of superkingdom specific SFs. **B:** SFs shared by different superkingdoms and the number of SFs found in the virosphere and respective sets. The bold number outside the circles indicates the number of virosphere specific SFs. 1,736 SFs from SCOP 1.73 were found in complete genomes as of March 2010, altogether 1,304 (874) genomes: 67 (59) archaea; 903 (573) bacteria; 334 (242) eukaryota; the numbers in parentheses indicate the number of different species, ignoring redundant strains.

We also examined how sensitive the number of VspSFs is to the possibility of being found in only one, two or a few cellular genomes; the possible result of a false assignment, sequence mis-assembly or very recent integration events. The results shown in Fig. 3B show that the numbers are not greatly affected by minor assignment errors.

From the 63 VspSF, 45 SF belong to monofolds (i.e. containing one SF per fold) and thus are also virosphere specific folds. The majority therefore are not susceptible to mis-classification or over-conservative classification at the SF level in SCOP. The classification is very unlikely to be wrong even in the remaining 18-folds where it is theoretically possible.

We have thus shown that the virosphere specificity is not an artefact of the database and its limitations.

Virosphere specificity is robust under the extension to incomplete genomes and other methods

Our analysis above is based on completely sequenced genomes, so here we checked for apparently VspSFs in incomplete genomes, and also checked to see if other methods could detect them where ours had not. Our initial results were confirmed.

We ran two tests on the 63 VspSFs: a comparison to UniProt-15.15 [50] and to the PFAM protein families database (24.0 A and B) [51]. Our analysis above is based on completely sequenced genomes. UniProt, however, also contains sequences from organisms which have not been completely sequenced as well as from mobile parts of genomes (like plasmids etc.), and therefore provides a further test of virosphere specificity. Only 2 of the 63 VspSF HMMs made a significant hit against sequences in UniProt from a cellular organism. We tested these hits with alternative methods – I-TASSER and LOMETS ('Zhang-server') [52] (one of the best structure prediction servers according to CASP8 [53]) – applying, in addition to the HMM score, many different parameters (see illustrative example on http://zhanglab.ccmb.med.umich.edu/LOMETS/). However, no evidence supporting either of the two hits could be found by any of the structure prediction servers (data not shown).

To attempt to validate the 63 VspSFs using a more sensitive independent method, we used a profile-profile comparison tool HHSearch [54] (HMM versus HMM); this performs significantly better than the HMMs versus sequences (used by SUPERFAMILY in the original genome analysis). We used the HMMs of the VspSFs from SUPERFAMILY to search against 31,912 PFAM models (A and B). In most cases the PFAM families with a significant match to a VspSF in turn are not reported by PFAM to have members detected in cellular organisms. This leaves us five SFs having significant hits against five PFAM families. The I-TASSER and LOMETS analysis and an examination of sequence annotations did not verify any of them.

In summary the lack of significant hits between the SUPERFAMILY VspSF models and Uniprot or the PFAM families containing cellular sequences is strong support that the 63 VspSFs we found are likely to be genuine.

Viruses have had proportionally fewer of their structures experimentally determined than cellular organisms

To detect an SF domain there must be a representative that has had its structure determined experimentally (usually by X-ray crystallography or NMR). Below we show that the virosphere structure space is poorly characterised, so there remain more VspSFs yet to be found.

The coverage of genomes by SUPERFAMILY assignments can be shown in several ways. Here, we use two of them to illustrate that viral proteins are much less structurally characterised than cellular proteins. In Fig. 3C[c] the

[c] Figure 3C was visualized from the data at http://www.supfam.org/SUPERFAMILY/cgi-bin/gen_list.cgi.

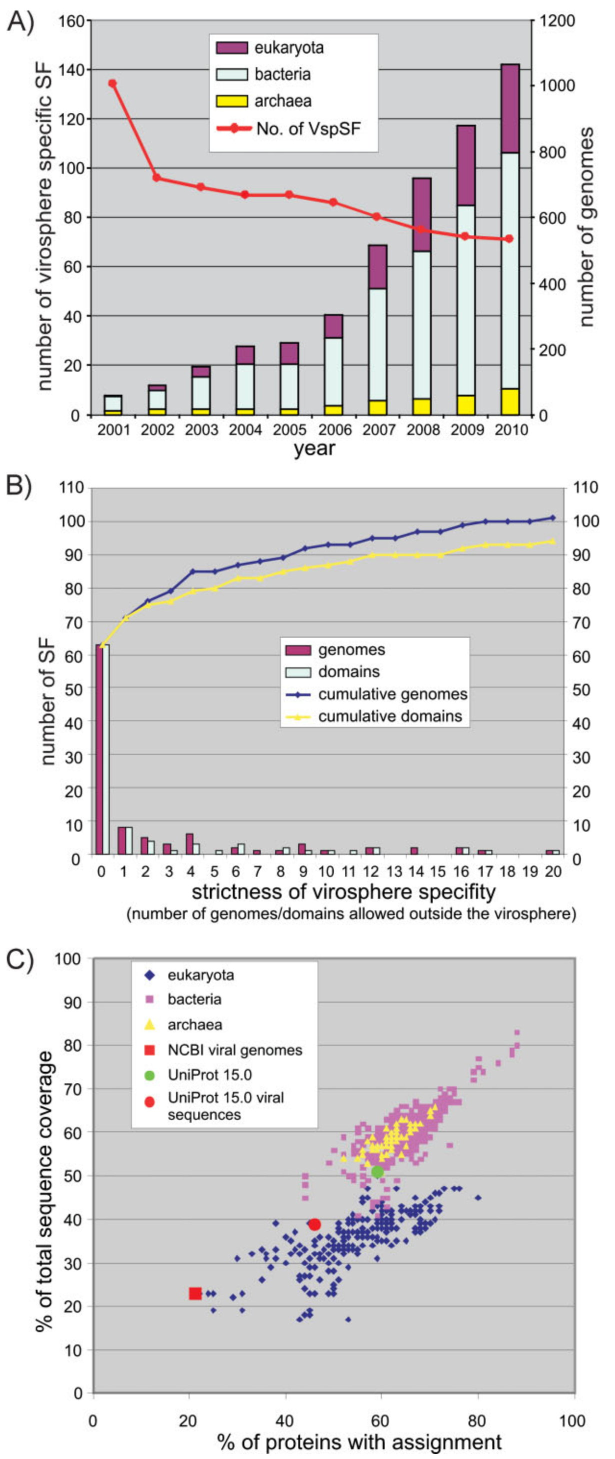

Figure 3. The number of VspSFs is not overestimated or caused by database bias. **A:** The change in the number of VspSFs over time. The 'virosphere specificity' of the SFs found in the virosphere as they would have been predicted at different time points (or against number of species in SUPERFAMILY). We calculated from historical data the number of VspSFs at any given year based on the sequences which were available at the time. SUPERFAMILY version 1.73 is used throughout this manuscript, except in this figure: the PDB entries for SCOP 1.73 were downloaded on 26 Sept 2007 (http://scop.mrc-lmb.cam.ac.uk/scop/index.html). To extend the curve with more up-to-date data it was necessary to use data from the recently updated SUPERFAMILY 1.75. The difference from 2001-2002 is due to the very limited number of eukaryote genomes available before 2002. The very large increase in the number of different species in SUPERFAMILY has a minor impact on decreasing the number of VspSFs. Solving new structures reveals new VspSFs and new cellular sequences reveal some SFs not to be virosphere specific, however we are now closer to saturation of cellular sequence space than viral structure space. **B:** The strictest requirement for virosphere specificity of a SF is that it is seen in zero cellular organisms (zero on *x*-axis, 63 SFs on *y*-axis). This graph shows what happens to the number of SFs (*y*-axis) if you relax that restriction, allowing a single (1 on the *x*-axis) cellular sequence to have the SF, or more (continuing on the *x*-axis). The criteria may be applied per genome or per domain, e.g. if one genome has two domains from the given SF, it counts as plus one for the yellow/purple and plus two for the blue/cyan. The curves (blue/yellow) are cumulative and the bars (purple/cyan) show the exact number of SFs with that number of members in cellular space. **C:** Viral sequences are strongly under-assigned in SUPERFAMILY. For each genome in the SUPERFAMILY database the '% of total sequence coverage' is plotted versus the '% of proteins with assignment' and coloured according to superkingdom. The same was done for NCBI viral genomes, Uniprot and a Uniprot viral subset.

percentage of genes (or proteins) with at least one SF assignment is plotted against the 'percentage of total sequence coverage' (% of amino acids assigned to SFs). Of all genomes the viral proteins have the lowest value in '% of genes assigned' and they also have one of the lowest values in '% of total sequence coverage'. The values for Uniprot sequences are close to the average over all genomes, however for the viral sequences in Uniprot the respective values drop significantly (more than 10%). Analysis of the distribution of structurally characterised PfamA domains shared by different superkingdoms also shows a lack of structural characterisation of virosphere specific domains. 26% of PfamA families contain a link to the known structures in Protein Data Bank (PDB) [55]. Of those families found in the genomes ~30% of those in bacteria, eukaryotes and viruses have a link to a PDB structure (Fig. 4A)[d]. Looking at the overlap, Fig. 4B shows that PfamA families present in all three superkingdoms are structurally the most characterised with more than half having a link to a PDB structure. The kingdom-specific PfamA entries are less characterised than entries shared by different kingdoms (e.g. 47% for families in archaea compared to 31% for Archaea specific families) and the virosphere-specific PfamA entries are the least well characterised having a drastic threefold drop from 30% overall to 11% for virosphere specific families. The quite high number of structurally characterised archaeal domains, 31.4%, could be the result of structural genomics initiatives as many of these are domains of unknown function (DUFs) and the fraction

[d] The data for Fig. 4 were generated via publicly available SRS search engine (srs.ebi.ac.uk). SRS integrates and combines several different databases.

Table 1. VspSFs are found in all functional classes of viruses[a]

	No. of viral families by ICTV	No. of viral families at least with one hit[b]	No. of viral genomes in dataset	Capsid/coat VspSF	Non-capsid VspSF	Total VspSF
+ssRNA	30	30	663	8	8	16
−ssRNA	8	8	131	6	4	10
dsRNA	9	9	124	4	4	8
dsDNA	27	27	748	8	14	22
ssDNA	7	5[c]	378	1		1
Retrotranscribing	4	3[d]	101	3	1	4
Hepatitis delta, Circular ssRNA	1	1	1		1	1
Satellites			129	1		1
Total	86[e]	83	2,304	31	32	63

[a] 'NCBI viral genomes' as a source of non-redundant viral genomes was downloaded on Oct. 14, 2009.

[b] From the 2304 genomes in the dataset, 2092 viral genomes have at least one significant SF assignment. No significant assignment for 212 viral genomes.

[c] None of the genomes from viral families Anelloviridae and Nanoviridae have assignment (despite of six and seven genomes in our dataset, respectively).

[d] Viral family Metaviridae had no genome sequence in database and, thus, has no assignment either.

[e] In the International Committee on Taxonomy of Viruses (ICTV) taxonomy 87 viral families, Alvernaviridae (also named Dinornaviridae) is classified as an unassigned ssRNA virus and has no hit in SF.

of Archaea specific families with a link to the PDB was much lower for the previous release (PfamA 23), 13.7%. Also, altogether 212 viral genomes (and even two viral families) do not have any significant assignment in SUPERFAMILY (Table 1)[e].

Together these data indicate that most of the virosphere proteins are not yet structurally characterised and that virosphere-specific proteins may yield many structures not yet determined and also possibly not existing in cellular proteins.

VspSFs are found in all major functional classes of viruses and most fold classes

Inspection of the distribution of VspSFs across viruses and across structural fold classes shows that they are widespread, and not restricted by viral or structural class.

The host range of different viruses varies greatly depending on the class (i.e. nature and strandedness) of their genome [56]. Conversely, the major groupings of cellular hosts are infected by different viral classes, e.g. there are no dsDNA viruses with a plant host and only a few RNA viruses infect bacterial hosts (illustrated in www.expasy.org/viralzone). The evolutionary rates are also different depending on the class (Fig. 1). Thus, according to these very deep and principal differences, functional classes of viruses have to be handled separately. We analysed the distribution of VspSFs between major functional classes of viruses. As shown in Table 1, the VspSFs is found in all major functional classes of viruses. Thus, the VspSFs are not restricted to specific kinds of genetic material. Not one of the VspSFs is found in more than one functional class of viruses, however, 5 VspSFs are found in more than one viral family (Supporting Information Table 1). The host range of viruses coding VspSFs is also very broad – prokaryotes as well as different eukaryotic taxons (yeast, metazoa, plants) are present (Supporting Information Table 1). The VspSFs are found in all SCOP structural classes (α, β, etc.) except the α/β-fold class (Supporting Information Table 1). The virosphere is not exceptional here as this fold class is also very rare in other superkingdom-specific SF (Abroi and Gough, unpublished results). The fold class distribution of SFs found in the virosphere is essentially the same as in SCOP or in cellular genomes. The presence of VspSFs is widespread and general.

Only half of VspSFs have functions related to viral coat/capsid

Surprisingly, looking at the functions of VspSFs we find that more than half of the protein domains which are specific to viruses do not have functions specific to viruses.

The default assumption is that, of course, there must be VspSFs because they are required for the viral capsid/coat function that in general is not found in cellular organisms (if we exclude the retroviral ones [57]). However, only up to half of VspSFs are capsid/coat proteins. We were conservative in classifying proteins as non-capsid; we are also counting proteins related to activities specific to viruses, such as attachment and

[e] Distribution of SUPERFAMILY hits in viral genomes (represented by 'NCBI viral genomes' http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=10239) in different viral families and functional classes of viruses according to combined data from SUPERFAMILY and NCBI Viral Genomes.

A)

	PfamA families	% with link to PDB	% of families shared by virosphere
all	11912	26	21
A	3016	47	22
B	7141	31	19
E	7880	31	15
V	2495	30	100

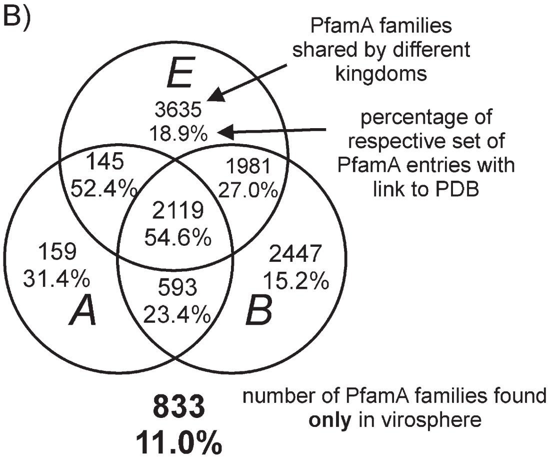
B)
PfamA families shared by different kingdoms
percentage of respective set of PfamA entries with link to PDB
E
3635
18.9%
145
52.4%
1981
27.0%
2119
54.6%
159
31.4%
2447
15.2%
593
23.4%
A
B
833
11.0%
number of PfamA families found only in virosphere

Figure 4. Virosphere-specific PfamA entries are under-represented in the PDB. **A:** Number of PfamA families in superkingdoms and percentage of PfamA families having a link to the PDB or percentage of families shared with the virosphere. **B:** Number of PfamA families in the respective sets and percentage of PfamA families with a link to the PDB. A – archaea; B – bacteria; E – eukaryota; V - viruses. Queries were run via SRS interface (srs.ebi.ac.uk) in March 2010 using PfamA 24.0. If the PfamA family has a link to the PDB it was counted as a structurally characterised family without taking into account 'alignment length'.

entry, encapsidation, etc. as 'capsid/coat' in this context (Table 1). The other half of VspSFs is related to activities which are not exclusively specific to viruses. Most of them are related to nucleic acid metabolism (DNA/RNA binding, transcription, replication etc.), with only 5 of the 32 VspSFs not being related to nucleic acid metabolism (Supporting Information Table 1). Thus, about half of VspSFs have functions related to the cell, and specifically to nucleic acid metabolism.

Virus-to-host (V2H) and host-to-virus (H2V) transfers of protein domains

In this section and the following three, we begin to speculate that the virosphere could be a source of new SFs for cellular organisms (via V2H transfer), as also hypothesised for viruses and plasmids [58]. For this to be possible there must be HGT between viruses and hosts.

Here, by V2H transfer, we mean where the new domain is fixed and has spread in the population (selected for) rather than just the integration of viral genetic material into the host chromosome. V2H transfers have not been very widely described to date for several reasons: only very recent transfers from V2H, where DNA sequence similarity is very high, can easily be detected; the viral-like sequences might be filtered out from genomic sequences as supposed contamination; the viral genes or viral protein domains are usually excluded from the phylogenetic analysis of domains or genes; viruses have much faster evolutionary rates, so the sequence similarity between homologous domains in viruses and their host may disappear very quickly. Nevertheless, there are still several examples that illustrate the V2H possibility for dsRNA, ssDNA and both classes of ssRNA viruses [25–28, 59, 60]. The extent of the transfer however is hard to estimate for the reasons described above.

The extensive overlap between the virosphere and cellular structure space (~500 SF out of 1,736) gives rise to the question of the direction of domain transfer, and indeed of the ancient history of viruses and cellular organisms. The abundance of SFs in the virosphere vs. cellular genomes suggests the possibility that transfers from V2H can be found in this dataset. Good candidates for this are the SFs found in many viral genomes and only a tiny fraction of cellular genomes (Supporting Information video). Work is in progress to analyse and find evidence for this. It is believed (but not proven on a large scale) that there is also much of 'from host-to-virus (H2V)' transfer [15]. It is possible that a continuous exchange exists in both directions.

When discussing domain transfer from V2H, we must bear in mind some virology. The host range and tissue tropism for viruses is defined as a species or tissues where productive infection takes place, i.e. where new infectious viral particles are produced. However, viruses can transfer genetic material in a non-productive way into a much larger number of species or tissues. As an example of this, the widespread use of viral vectors for gene targeting based on adenoviruses, lentiviruses, baculoviruses, vaccinia viruses, etc., should be noted.

The evolutionary history of viral protein domains can be elucidated

Here, we discuss the ways in which we can elucidate the phylogeny, diversity and environmental distribution of viruses.

As explained in the introduction, structural information is crucial to understanding V2H and H2V transfers. This may also be used in the same way to examine the evolutionary history of viral protein domains within the virosphere, i.e. between different viral genomes (similarly proposed for structures of viral capsid/coat proteins by Krupovic et al. [61]). The presence of domain SFs and their combinations in proteins has sufficient information content to, in the case of some medium and large viral genomes, determine the existence of certain branches in the (largely unresolved) phylogeny of viruses, as demonstrated in cellular organisms [62]. Examining the structural relatives of viral protein domains not only in the chromosomes of cellular and viral genomes, but also in the mobile parts of the genomes (like plasmids) is important when considering the transfer of viral proteins. The representation of VspSFs and unique (to viruses) combinations of domains in environmental sampling sequence sets (meta-genomes) can be used to describe the viral diversity and ubiquity in these samples, and may suggest to what extent we have explored the virosphere via sequencing so far. Answers to all these questions can be extracted from the assignments in the SUPERFAMILY database.

What is the origin of virosphere specific superfamilies?

Above we established the existence of VspSFs. In this section we examine the possible origins of them, crucial to the question of the genetic interaction between the virosphere and cellular life.

If VspSFs have no cellular structural homologues then they have no visible cellular ancestry, so where did they come from? There are multiple potential origins of VspSFs: they may be ancestral viral SFs (for example some viral capsid proteins); ancestral cellular proteins borrowed by viruses but later lost by cells, or the respective taxon died out; or they could be a de novo viral SF evolved by viruses. The origin of VspSFs may give us important clues about the evolutionary interaction between the virosphere and the cellular protein world, with implications for the non-VspSFs, and possibly some hints about evolution pre-LUCA.

It has been hypothesised that some type of viral capsid proteins, like the double-jellyroll of adenoviruses Prd1 and STIV, are very ancient according to their abundance in different types of viruses with respect to the range of their hosts [63]. Also other types of viral capsids are found in viruses infecting hosts from different domains [64]. Thus, the hypothesis that some viral lineages and viral capsids precede the LUCA have some strong support (argued in ref. [56]) and could be classified as ancient viral SFs.

Since VspSFs may originally be from ancestral cellular proteins but subsequently have been lost via gene loss or extinction, the virosphere may be able to work as a buffer or reservoir for SFs the cells have lost. In this way they could provide a shortcut in time from ancestral cellular to contemporary cellular proteins. The host-to-virus-to-host transfer loop was also proposed [62] to explain the evolutionary history of MCM proteins in archaea. The virosphere could be much more than only a store, as it may also work as a workshop to evolve the domains (see below).

Could the virosphere be a source for de novo superfamilies?

Below we propose the hypothesis that due to the vast differences between viral evolution and cellular evolution since the LUCA, viruses could be responsible for some of the rare de novo creations of cellular SFs.

The probability of evolving a new SF de novo is extremely low, but this probability is much higher in viruses than in cells. The most important aspect – very fast evolutionary rates of viruses – has already been mentioned. Viruses, as parasites, are not concerned with the fate of the cell after the new virus generation has been produced (especially the lytic viruses). Viruses can express (even at a high level) genes which are energetically unfavoured by cells (moderately mis-folded proteins, or loosely packed proteins) or proteins which are toxic to the cells. These would be selected against in cellular evolution (avoiding misfolding is also one of the driving forces for cellular protein domain evolution [65, 66]). Further evolution of these domains may lead to stable and/or nontoxic domains. For this reason, in the pathway of domain evolution, viruses can overcome some barriers that cells cannot. The long branches on the phylogenetic trees (fast evolution) and accumulation of insertion/deletion and domain swapping was also observed in the case of mobile element (including viruses) encoded MCM proteins [61]. Thus, there is the hypothetical possibility that this gives viruses the ability to develop new protein SFs and families, acting as an evolutionary engine and store. Is it possible that viruses are actually beneficial to cellular life in the long term and that they are the medium through which life can still make use of the ancient pre-LUCA mechanisms in which the majority of existing SFs were originally forged?

Why uncharacterised regions of the virosphere are crucial for the understanding of structure space

In this section, we consider the regions of the virosphere that are yet to have a representative protein structure experimentally determined. These regions are excellent targets for structural genomics since they would help us to understand the extent to which nature has explored structure space, and (if viruses are a source of new protein folds) elucidate the nature of de novo protein evolution itself.

We showed above that the structural characterisation of viral sequences has been relatively neglected so far (Fig. 4) and in the future may lead to the discovery of many new SFs. They would therefore make good targets for structural genomics initiatives. A good example here is the structure of the protein from the archaeal plasmid pT26-2, where all three domains seem to be new folds [58] (classification from SCOP and CATH is yet to confirm the statement made by the authors). Viral metagenomics data and studies of archaeal viruses show that many genes in the virosphere are not yet found in databases (reviewed in ref. [10, 67]). The currently sequenced viruses (viral genomes in the database) are strongly biased to medically and economically important viruses. The host genomes that have been sequenced have also been subject to more or less the same sorts of selection bias. The more commensal (*commensalism* is a class of relationship between two organisms where one organism benefits but the other is unaffected) or symbiotic viruses have just started to be sequenced and characterised.

The virosphere structure space may contain structurally stable but evolutionary not-yet-observed protein folds, also called the 'dark matter' of protein structure space [68]. There is a discrete-continuous duality of protein structure space (recently reviewed by Grishin and coworkers [69]), which is based on knowledge that the space is largely discrete in the evolutionary sense, but continuous geometrically. The viruses help to bridge this duality in at least two ways. First, the viral structure space may broaden the populated structure space (e.g. the VspSFs are mostly located in the periphery on the 'galaxy of folds' [70], http://toolkit.tuebingen.mpg.de/hhcluster) and also generate 'tunnels' on the 'valleys' between folds. Second, if the folds are islands of stability in an ocean of an overwhelming majority of unstable conformations, using the analogy of Lupas and Koretke [71], viruses as fast evolving parasites help to swim (to evolve) the domains from one island to another (yet unpopulated) island. Since viruses evolve many times faster than cellular organisms, their structural characterisation may give us crucial evidence for helping to quantify the extent to which nature has explored structure space, one of the important outstanding questions in molecular biology.

At present the characterised viral structure space is populated very sparsely and is very biased. Hopefully, a more complete view of viral structure space will help us to understand the evolution of viral as well as cellular proteins, and how they affect each other.

Conclusions

To our knowledge this is the first comprehensive characterisation of the overlap of the virosphere and cellular structure space. We have shown conclusively that there are a significant number of virosphere specific SFs which have no evolutionary relative in cellular organisms. We have shown that these VspSFs are genuine and that there are most likely many more yet to be discovered; they are present in all major classes of viruses. These SFs are only present in viruses, which due to their extreme rate of evolution and lack of constraint, could be a source of novel protein folds not only for themselves but for cellular organisms via horizontal transfer. Moreover, only half of the VspSFs have viral capsid/coat functions with the remaining half having functions relevant to cellular life (mostly nucleic acid metabolism). A more complete characterisation of viral structure space should be a priority for experimental determination to understand de novo fold evolution, very early evolution and virus to host genetic transfer. Most studies of molecular evolution deliberately focus exclusively on cellular life, yet the inclusion of viruses and their protein domain content is crucial to our understanding of the genesis of nature.

Acknowledgments

We thank Cyrus Chothia, Erkki Truve and Maido Remm for critical reading of this paper. The work of AA was supported in part by the target financial project SF0180175B, by the European Regional Development Fund through Centre of Excellence in Chemical Biology 3.2.0101.08-0017 and by ETF8812 to AA from the Estonian Science Foundation. AA likes to thank the EMBO Short term fellowship ASTF63-2010 and EC DGR for FP7 Ecogene grant 205419 for supporting collaboration with JG.

References

1. **Raoult D**, **Audic S**, **Robert C**, **Abergel C**, et al. 2004. The 1.2-megabase genome sequence of Mimivirus. *Science* **306**: 1344–50.
2. **Boyer M**, **Yutin N**, **Pagnier I**, **Barrassi L**, et al. 2009. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc Natl Acad Sci USA* **106**: 21848–53.
3. **La Scola B**, **Desnues C**, **Pagnier I**, **Robert C**, et al. The virophage as a unique parasite of the giant mimivirus. *Nature* 2008. **455**: 100–4.
4. **Claverie JM**, **Abergel C.** 2010. Mimivirus: the emerging paradox of quasi-autonomous viruses. *Trends Genet* **26**: 431–7.
5. **Lang AS**, **Beatty JT.** 2007. Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol* **15**: 54–62.
6. **Breitbart M**, **Rohwer F.** 2005. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* **13**: 278–84.
7. **Rohwer F**, **Thurber RV.** 2009. Viruses manipulate the marine environment. *Nature* **459**: 207–12.
8. **Suttle CA.** 2005. Viruses in the sea. *Nature* **437**: 356–61.
9. **Suttle CA.** 2007. Marine viruses–major players in the global ecosystem. *Nat Rev Microbiol* **5**: 801–12.
10. **Kristensen DM**, **Mushegian AR**, **Dolja VV**, **Koonin EV.** 2010. New dimensions of the virus world discovered through metagenomics. *Trends Microbiol* **18**: 11–9.
11. **Lindell D**, **Jaffe JD**, **Johnson ZI**, **Church GM**, et al. 2005. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**: 86–9.
12. **Mann NH**, **Clokie MR**, **Millard A**, **Cook A**, et al. 2005. The genome of S-PM2, a ''photosynthetic'' T4-type bacteriophage that infects marine Synechococcus strains. *J Bacteriol* **187**: 3188–200.
13. **Sharon I**, **Alperovitch A**, **Rohwer F**, **Haynes M**, et al. 2009. Photosystem I gene cassettes are present in marine virus genomes. *Nature* **461**: 258–62.
14. **Sullivan MB**, **Lindell D**, **Lee JA**, **Thompson LR**, et al. 2006. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol* **4**: e234.
15. **Moreira D**, **Lopez-Garcia P.** 2009. Ten reasons to exclude viruses from the tree of life. *Nat Rev Microbiol* **7**: 306–11.
16. **Claverie JM**, **Ogata H.** 2009. Ten good reasons not to exclude giruses from the evolutionary picture. *Nat Rev Microbiol* **7**: 615.
17. **Hegde NR**, **Maddur MS**, **Kaveri SV**, **Bayry J.** 2009. Reasons to include viruses in the tree of life. *Nat Rev Microbiol* **7**: 615.
18. **Koonin EV**, **Senkevich TG**, **Dolja VV.** 2009. Compelling reasons why viruses are relevant for the origin of cells. *Nat Rev Microbiol* **7**: 615.
19. **Ludmir EB**, **Enquist LW.** 2009. Viral genomes are part of the phylogenetic tree of life. *Nat Rev Microbiol* **7**: 615.
20. **Navas-Castillo J.** 2009. Six comments on the ten reasons for the demotion of viruses. *Nat Rev Microbiol* **7**: 615.
21. **Raoult D.** 2009. There is no such thing as a tree of life (and of course viruses are out!). *Nat Rev Microbiol* **7**: 615.
22. **Forterre P**, **Prangishvili D.** 2009. The origin of viruses. *Res Microbiol* **160**: 466–72.
23. **Villarreal LP**, **Witzany G.** 2010. Viruses are essential agents within the roots and stem of the tree of life. *J Theor Biol* **262**: 698–710.
24. **Enquist LW.** 2009. Virology in the 21st century. *J Virol* **83**: 5296–308.
25. **Horie M**, **Honda T**, **Suzuki Y**, **Kobayashi Y**, et al. 2010. Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* **463**: 84–7.
26. **Taylor DJ**, **Bruenn J.** 2009. The evolution of novel fungal genes from non-retroviral RNA viruses. *BMC Biol* **7**: 88.
27. **Belyi VA**, **Levine AJ**, **Skalka AM.** 2010. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. *PLoS Pathog* **6**: e1001030.
28. **Liu H**, **Fu Y**, **Jiang D**, **Li G**, et al. 2010. Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. *J Virol* **84**: 11876–87.
29. **Katzourakis A**, **Gifford RJ.** Endogenous viral elements in animal genomes. *PLoS Genet* 2010. **6**: e1001191.
30. **Duffy S**, **Shackelton LA**, **Holmes EC.** 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* **9**: 267–76.
31. **Gago S**, **Elena SF**, **Flores R**, **Sanjuan R.** 2009. Extremely high mutation rate of a hammerhead viroid. *Science* **323**: 1308.
32. **Sanjuan R**, **Nebot MR**, **Chirico N**, **Mansky LM**, et al. 2010. Viral mutation rates. *J Virol* **84**: 9733–48.
33. **Domingo E.** 2007. Virus evolution. In Fields BN, Knipe DM, Howley PM, Griffin DE, eds; *Fields Virology*, 5th edn. Philadelphia, PA, USA: Lippincott Williams & Wilkins. p. 3177.

34. **Carpi G**, **Holmes EC**, **Kitchen A.** 2010. The evolutionary dynamics of bluetongue virus. *J Mol Evol* **70**: 583–92.
35. **Duffy S**, **Holmes EC.** 2009. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *J Gen Virol* **90**: 1539–47.
36. **Gibbs AJ**, **Ohshima K**, **Phillips MJ**, **Gibbs MJ.** 2008. The prehistory of potyviruses: their initial radiation was during the dawn of agriculture. *PLoS One* **3**: e2523.
37. **Hughes AL**, **Irausquin S**, **Friedman R.** 2010. The evolutionary biology of poxviruses. *Infect Genet Evol* **10**: 50–9.
38. **Kowalik TF**, **Li JK.** 1991. Bluetongue virus evolution: sequence analyses of the genomic S1 segments and major core protein VP7. *Virology* **181**: 749–55.
39. **Shah SD**, **Doorbar J**, **Goldstein RA.** 2010. Analysis of host-parasite incongruence in papillomavirus evolution using importance sampling. *Mol Biol Evol* **27**: 1301–14.
40. **Kumar S**, **Subramanian S.** 2002. Mutation rates in mammalian genomes. *Proc Natl Acad Sci USA* **99**: 803–8.
41. **Chothia C**, **Lesk AM.** 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J* **5**: 823–6.
42. **Illergard K**, **Ardell DH**, **Elofsson A.** 2009. Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. *Proteins* **77**: 499–508.
43. **Caetano-Anolles G**, **Sun FJ**, **Wang M**, **Yafremava LS**, et al. 2008. Origins and evolution of modern biochemistry: insights from genomes and molecular structure. *Front Biosci* **13**: 5212–40.
44. **Caetano-Anolles G**, **Wang M**, **Caetano-Anolles D**, **Mittenthal JE.** 2009. The origin, evolution and structure of the protein world. *Biochem J* **417**: 621–37.
45. **Chothia C**, **Gough J.** 2009. Genomic and structural aspects of protein evolution. *Biochem J* **419**: 15–28.
46. **Levitt M.** 2009. Nature of the protein universe. *Proc Natl Acad Sci USA* **106**: 11079–84.
47. **Gough J**, **Chothia C.** 2002. SUPERFAMILY: HMMs representing all proteins of known structure. COP sequence searches, alignments and genome assignments. *Nucleic Acids Res* **30**: 268–72.
48. **Gough J**, **Karplus K**, **Hughey R**, **Chothia C.** 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* **313**: 903–19.
49. **Murzin AG**, **Brenner SE**, **Hubbard T**, **Chothia C.** 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**: 536–40.
50. **The UniProt Consortium.** 2010. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* **38**: D142–8.
51. **Finn RD**, **Mistry J**, **Tate J**, **Coggill P**, et al. 2010. The Pfam protein families database. *Nucleic Acids Res* **38**: D211–22.
52. **Zhang Y.** 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinf* **9**: 40.
53. **Cozzetto D**, **Kryshtafovych A**, **Fidelis K**, **Moult J**, et al. 2009. Evaluation of template-based models in CASP8 with standard measures. *Proteins* **77**: 18–28.
54. **Hildebrand A**, **Remmert M**, **Biegert A**, **Soding J.** 2009. Fast and accurate automatic structure prediction with HHpred. *Proteins* **77**: 128–32.
55. **Berman HM**, **Westbrook J**, **Feng Z**, **Gilliland G**, et al. 2000. The Protein Data Bank. *Nucleic Acids Res* **28**: 235–42.
56. **Koonin EV**, **Senkevich TG**, **Dolja VV.** 2006. The ancient Virus World and evolution of cells. *Biol Direct* **1**: 29.
57. **Varela M**, **Spencer TE**, **Palmarini M**, **Arnaud F.** 2009. Friendly viruses: the special relationship between endogenous retroviruses and their host. *Ann N Y Acad Sci* **1178**: 157–72.
58. **Keller J**, **Leulliot N**, **Soler N**, **Collinet B**, et al. 2009. A protein encoded by a new family of mobile elements from Euryarchaea exhibits three domains with novel folds. *Protein Sci* **18**: 850–5.
59. **Belyi VA**, **Levine AJ**, **Skalka AM.** 2010. Sequences from ancestral single stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more than 40-50 million years old. *J Virol* **84**: 12458–62.
60. **Kapoor A**, **Simmonds P**, **Lipkin IW.** 2010. Discovery and characterization of mammalian endogenous parvoviruses. *J Virol* **84**: 12628–35.
61. **Krupovic M**, **Gribaldo S**, **Bamford DH**, **Forterre P.** 2010. The evolutionary history of archaeal MCM helicases: a case study of vertical evolution combined with hitch-hiking of mobile genetic elements. *Mol Biol Evol* **27**: 2716–32.
62. **Yang S**, **Doolittle RF**, **Bourne PE.** 2005. Phylogeny determined by protein domain content. *Proc Natl Acad Sci USA* **102**: 373–8.
63. **Bamford DH.** 2003. Do viruses form lineages across different domains of life? *Res Microbiol* **154**: 231–6.
64. **Bamford DH**, **Grimes JM**, **Stuart DI.** 2005. What does structure tell us about virus evolution? *Curr Opin Struct Biol* **15**: 655–63.
65. **Drummond DA**, **Wilke CO.** 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**: 341–52.
66. **Zhou T**, **Weems M**, **Wilke CO.** 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol* **26**: 1571–80.
67. **Prangishvili D**, **Garrett RA**, **Koonin EV.** 2006. Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. *Virus Res* **117**: 52–67.
68. **Taylor WR**, **Chelliah V**, **Hollup SM**, **MacDonald JT**, et al. 2009. Probing the ‘‘dark matter’’ of protein fold space. *Structure* **17**: 1244–52.
69. **Sadreyev RI**, **Kim BH**, **Grishin NV.** 2009. Discrete-continuous duality of protein structure space. *Curr Opin Struct Biol* **19**: 321–8.
70. **Alva V**, **Remmert M**, **Biegert A**, **Lupas AN**, et al. 2010. A galaxy of folds. *Protein Sci* **19**: 124–30.
71. **Lupas AN**, **Koretke KK.** 2008. Evolution of protein folds. In Pitsch M, Schwede T, eds; *Computational Structural Biology: Methods and Applications*. Hackensack, NJ: World Scientific. p. 792.