Genomics 100 (2012) 149–156

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

VIP DB — A viral protein domain usage and distribution database

Ting-Wen Chen [a,b,c], Richie Ruei-Chi Gan [b,d], Timothy H. Wu [c], Wen-Chang Lin [d,e,1], Petrus Tang [a,b,f,*,1]

[a] Molecular Medicine Research Center, Chang Gung University, Taoyuan, Taiwan
[b] Bioinformatics Center, Chang Gung University, Taoyuan, Taiwan
[c] Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan
[d] Department of Biological Science and Technology, National Chiao Tung University, HsinChu, Taiwan
[e] Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan
[f] Molecular Regulation & Bioinformatics Laboratory, Chang Gung University, Taoyuan, Taiwan

ARTICLE INFO

Article history:
Received 24 March 2012
Accepted 15 June 2012
Available online 24 June 2012

Keywords:
Protein domain
Domain–domain interaction
Viral protein interaction
Protein annotation

ABSTRACT

During the viral infection and replication processes, viral proteins are highly regulated and may interact with host proteins. However, the functions and interaction partners of many viral proteins have yet to be explored. Here, we compiled a VIral Protein domain DataBase (VIP DB) to associate viral proteins with putative functions and interaction partners. We systematically assign domains and infer the functions of proteins and their protein interaction partners from their domain annotations. A total of 2,322 unique domains that were identified from 2,404 viruses are used as a starting point to correlate GO classification, KEGG metabolic pathway annotation and domain–domain interactions. Of the unique domains, 42.7% have GO records, 39.6% have at least one domain–domain interaction record and 26.3% can also be found in either mammals or plants. This database provides a resource to help virologists identify potential roles for viral protein. All of the information is available at http://vipdb.cgu.edu.tw.

1. Introduction

Viruses are intracellular pathogens with minimal genome sizes and can only replicate within their host. Many viral proteins have been found to interact with their host proteins and influence the physiology of the host cell [1–4]. Some viral proteins can interact with other viral proteins. For example, several structural proteins can also interact with other viral proteins during virion assembly. These types of interactions among viral proteins and their host proteins or other viral proteins are collected in several databases, such as VirusMINT, VirHostNet, PIG and HPIDB [3,5–7]. However, these databases only contain validated information on viral protein–protein interactions from published papers, and most of the databases are limited to several families of viruses, primarily, human pathogenic viruses. There are only 952 virus–host interactions for RNA viruses collected in VirHostNet currently [3], and more than 80% of the interactions are related to proteins from either *hepatitis C virus* or *influenza A virus*. Therefore, a resource that provides a comprehensive view of all putative protein–protein interactions for viral proteins would be invaluable.

Previous studies showed that many protein–protein interactions are attributed to domain–domain interactions (DDI), and these interactions between domains are highly conserved among viruses, bacteria and humans and form cellular interaction networks [8–10]. Here, we constructed a comprehensive integrated database VIral Protein domain DataBase (VIP DB), to provide information on interaction correlations and biological functions of viral proteins by relying on their protein domains. This database includes DDI information and identifies candidate interaction partners or homologs of interaction partners from both viruses and other organisms. These possible interaction relationships may further be used to postulate the biological processes of viral proteins in their hosts. To pinpoint the types of biological processes these proteins may be involved in, pathway information from KEGG is also included when available.

This study collected 2,404 completely sequenced virus genomes with gene prediction results and the predicted protein sequences deposited in the NCBI genome database. Based on these genome annotations, we used a panel of bioinformatics tools to identify domains in these viral proteins and constructed VIP DB. Domains are the functional part of proteins and are believed to be essential for predicting protein functions [11,12]. Therefore, in addition to predicting protein interaction partners, the domain information can also be used to predict the viral protein's function. There are many established domain databases currently in existence [13–18]. These domain databases have been shown to be useful for predicting protein functions, and many of them, such as SUPERFAIMLY, InterPro and SMART, have already integrated gene ontology (GO) information into

[*] Corresponding author at: Molecular Medicine Research Center, Chang Gung University, Taoyuan, Taiwan.
E-mail addresses: g39328001@ym.edu.tw (T.-W. Chen), csardas@gmail.com (R.R.-C. Gan), g39328006@ym.edu.tw (T.H. Wu), wenlin@ibms.sinica.edu.tw (W.-C. Lin), petang@mail.cgu.edu.tw (P. Tang).
[1] These authors contributed equally.

0888-7543/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2012.06.008

the domain records in their database [19–21]. GO annotates the molecular functions, cellular components and biological processes of genes and can therefore be used to describe the function of proteins [22].

VIP DB also offers clues on other important biological features via the domains present within viruses. It provides a functional annotation for the domain and associates each domain with the biological features of the proteins. We integrated the descriptions of the domains from Pfam and the correlation between the domains and GO terms. It has been estimated that 67% of singlet-domain proteins with the same domain have a similar function. The degree of similarity is reduced for multi-domain proteins, but the degree of similarity is still up to 35% for two-domain proteins that share even one domain [23]. This result indicates that proteins sharing the same domain may have functional correlations. Therefore, domains of proteins from organisms other than viruses were also identified and included in our database to allow for these types of functional inferences.

In summary, VIP DB provides a comprehensive view of the distribution of domains in viruses and uses the information to provide suggestions for either the biological function or possible interaction partners of these viral proteins. We integrated domain–GO correlations, domain–domain interactions and the protein pathways together in a user-friendly database. By using VIP DB, users can explore the distribution of the domains that they are interested in, examine domain usages in viruses and speculate the possible biological roles of viral proteins. VIP DB is available at http://vipdb.cgu.edu.tw.

2. Results

2.1. Explore the VIP DB website

A detailed user tutorial is provided on the VIP DB website, and users can either explore VIP DB several different ways or download all of the datasets from VIP DB. Because VIP DB associates the viral proteins and other biological information correlated with domains, users can begin their query in a single virus or many viruses, a single domain or a set of domains. For users who are interested in the overall domain usage across all viruses, a comprehensive domain usage table is also provided.

2.1.1. Start with interesting viruses

VIP DB lists all viruses in a tree structure according to the taxonomy provided by NCBI. A total of 2,404 viruses were classified into 7 main categories: dsDNA, ssDNA, dsRNA, (+)ssRNA, (−)ssRNA, retrovirus, satellite and other. Users can begin searching viruses in the database via the search bar at the top right corner of the website or locate their viruses of interest directly from the hierarchical tree structure of viruses in the virus list (Figs. 2A and B). All of the proteins in a specific virus are shown with their domain information (Fig. 2C). Detailed descriptions or features of the domain are provided through links to Pfam. Pfam IDs and the standard abbreviations for the domain are displayed with the proteins. For domains with DDI (or GO annotation) information, links to the DDI (or GO) information are also provided.

Because the functions of many viral proteins are not yet known, domains found in these proteins may provide some insight into protein functions. The most intuitive way to link functions to protein domains is through the relationships between GO terms and protein domains. VIP DB provides the GO annotations for domains if any are available. Of the 2,322 viral domains, approximately 43% have GO annotations and approximately 51% (16,129 out of 31,939) of the proteins with domains have GO information in our analysis. Therefore, the functions of about half of the domain-containing viral proteins can be inferred from their domains using this strategy.

For domains shared between viruses and mammals or viruses and plants, links to the prevalent mammal domain architecture or prevalent plant domain architecture are available through the M (for mammals) or the P (for plants) icons. Those icons link to lists of domain architecture groups, and each group contains a group of homologous proteins from either mammals or plants. The proteins listed may have further pathway information available. If a protein in the protein list is annotated in the KEGG database, an external link is provided next to the name of the protein. The function of these

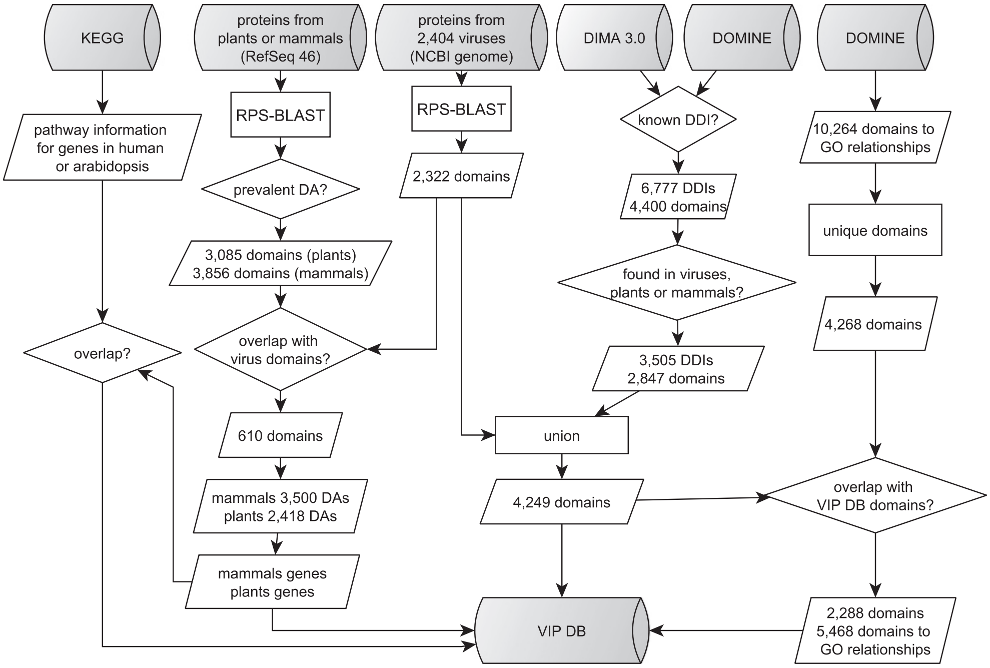
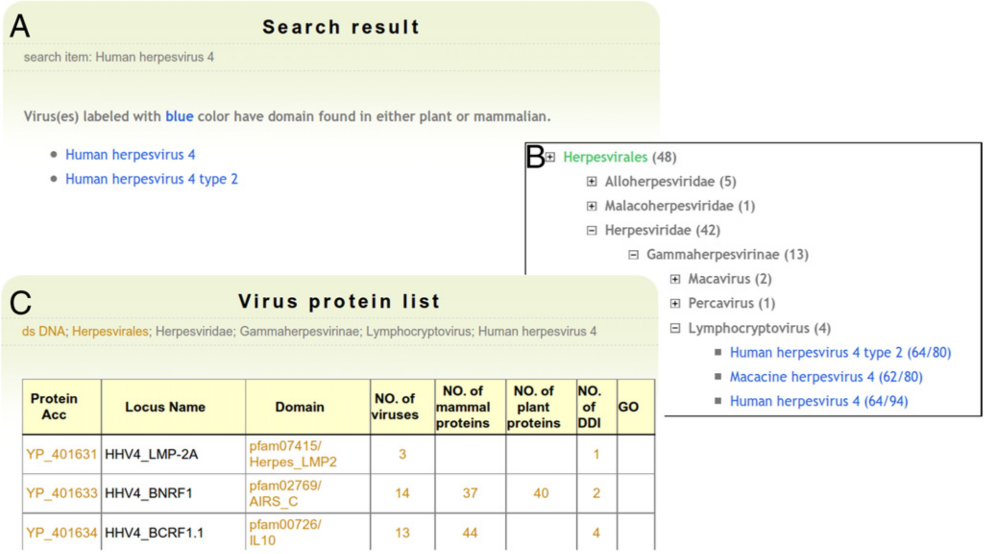

KEGG
proteins from plants or mammals (RefSeq 46)
proteins from 2,404 viruses (NCBI genome)
DIMA 3.0
DOMINE
DOMINE
pathway information for genes in human or arabidopsis
RPS-BLAST
RPS-BLAST
known DDI?
10,264 domains to GO relationships
prevalent DA?
2,322 domains
6,777 DDIs 4,400 domains
unique domains
3,085 domains (plants) 3,856 domains (mammals)
found in viruses, plants or mammals?
4,268 domains
overlap?
overlap with virus domains?
3,505 DDIs 2,847 domains
610 domains
union
mammals 3,500 DAs plants 2,418 DAs
4,249 domains
overlap with VIP DB domains?
mammals genes plants genes
VIP DB
2,288 domains 5,468 domains to GO relationships

Fig. 1. Flow chart for the construction of the viral protein domain database (VIP DB). Protein sequences from viruses, plants and mammals were downloaded from NCBI RefSeq along with the respective genomes. VIP DB suggests viral protein functions and interactive protein partners through protein domains through the integration of domain–domain interactions (shortened as DDIs in the figure), domain GO annotations, and pathway information. The DDI information was downloaded from DOMINE and DIMA 3.0, and only those known DDIs were used. The domains for the GO relationships were also downloaded from DOMINE. The pathway information was derived from KEGG. All these information were linked to the domains of viral proteins. See the main text for more details.



A Search result

search item: Human herpesvirus 4

Virus(es) labeled with blue color have domain found in either plant or mammalian.

- Human herpesvirus 4
- Human herpesvirus 4 type 2

B ⊞ Herpesvirales (48)
⊞ Alloherpesviridae (5)
⊞ Malacoherpesviridae (1)
⊟ Herpesviridae (42)
⊟ Gammaherpesvirinae (13)
⊞ Macavirus (2)
⊞ Percavirus (1)
⊟ Lymphocryptovirus (4)
■ Human herpesvirus 4 type 2 (64/80)
■ Macacine herpesvirus 4 (62/80)
■ Human herpesvirus 4 (64/94)

C Virus protein list

ds DNA; Herpesvirales; Herpesviridae; Gammaherpesvirinae; Lymphocryptovirus; Human herpesvirus 4

Protein Acc	Locus Name	Domain	NO. of viruses	NO. of mammal proteins	NO. of plant proteins	NO. of DDI	GO
YP_401631	HHV4_LMP-2A	pfam07415/ Herpes_LMP2	3			1	
YP_401633	HHV4_BNRF1	pfam02769/ AIRS_C	14	37	40	2	
YP_401634	HHV4_BCRF1.1	pfam00726/ IL10	13	44		4	

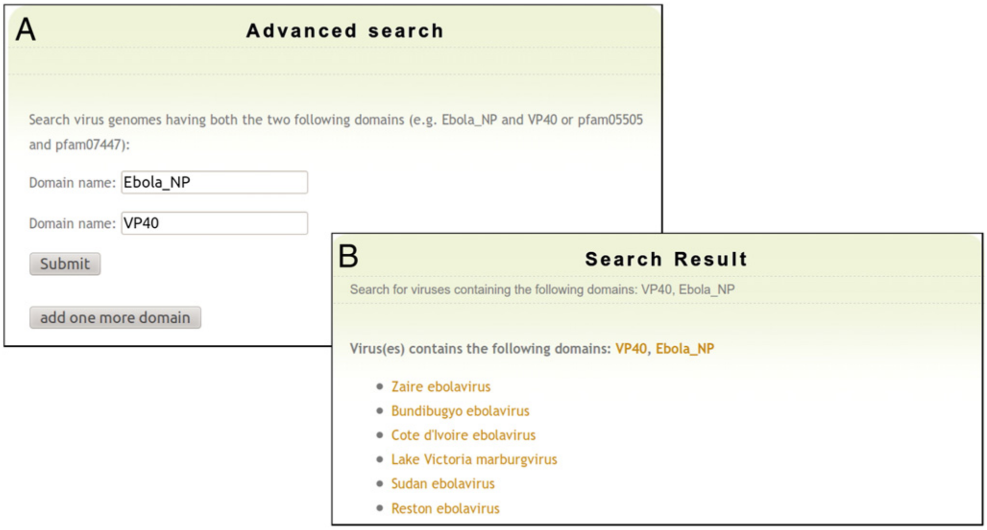
Fig. 2. Exploring the VIP DB. VIP DB contains domain and domain relevant information for 2,404 viruses. (A) Exploring can begin by searching for a virus or domain in the search box or by directly selecting the virus of interest from the virus list. (B) The virus list is organized into a taxonomic tree structure. The number next to the clade on the tree structure is the number of viruses under that clade. The numbers next to the virus name are the number of domain-containing proteins and the number of total proteins in each virus. Viruses in blue have domains that are found in either mammals or plants. (C) For each virus, the proteins with domain(s) are listed in a table. A click on the protein links the user to the main page for that protein in NCBI. Domains found in each protein are listed behind that protein, and clicking on the domain ID links to Pfam. If there are other proteins from viruses, plants or mammals that contain that domain, the user can obtain the information from the number next to the domain. Meanwhile, domain–domain interactions and domain GO information (GO icon) are provided through a link behind the domain if available.

proteins may hint at the function of the viral proteins that contain the same domain.

2.1.2. Start with interesting domains

For users with a specific interest in a single domain or a set of domains, VIP DB can be used to reveal the distribution of the domain(s) of interest in all viruses. Users can start their exploration at the domain of interest either by searching for a single domain from the search bar or by using the domain distribution table. For users interested in a set of domains, multiple domains can be searched for by using the advanced search, which can also provide users with a list of viruses containing the query domains. Users can easily find a list of viruses that have proteins that also contain the domain(s) of interest. This approach can reveal how prevalent a domain is in viruses and can be used to find functionally similar viral proteins or perhaps virus orthologs. For example, we find that the domains Ebola_NP and VP40 are restricted to, but prevalent in, five ebola viruses (*Sudan ebolavirus*, *Reston ebolavirus*, *Zaire ebolavirus*, *Bundibugyo ebolavirus*, and *Cote d'Ivoire ebolavirus*) and the *Lake Victoria marburgvirus*. These viruses are the only viruses found in the *Filoviridae* family (ssRNA negative-strand viruses; *Mononegavirales*; *Filoviridae*), as shown in Fig. 3.

For users interested in the global domain distribution of all viruses, VIP DB provides a domain distribution heat map table in which the



A Advanced search

Search virus genomes having both the two following domains (e.g. Ebola_NP and VP40 or pfam05505 and pfam07447):

Domain name: Ebola_NP

Domain name: VP40

Submit

add one more domain

B Search Result

Search for viruses containing the following domains: VP40, Ebola_NP

Virus(es) contains the following domains: VP40, Ebola_NP

- Zaire ebolavirus
- Bundibugyo ebolavirus
- Cote d'Ivoire ebolavirus
- Lake Victoria marburgvirus
- Sudan ebolavirus
- Reston ebolavirus

Fig. 3. Searching for a set of domains. Searching for multiple domains is useful for identifying viruses that use a particular set of domains. On the advanced search page, the user can query the database using two or more domains to retrieve a list of viruses that each contains all of the domains queried. In this example, we searched for the domains Ebola_NP and VP40 and found that those two domains exist in 5 ebola viruses and the *Lake Victoria marburgvirus* which also belong to the family *Filoviridae*, which contains the ebola viruses.

domain usage for each category of virus is shown (Fig. 4). In the heat map table, the color represents the frequency of the domain occurrence, and users can sort by the column they are interested in. For example, in the row pfam00589 (Phage_integrase), the dsDNA cell has a dark red background, indicating that the domain is frequently found in dsDNA viruses, while the white ssDNA cell indicates that this domain is not found in ssDNA viruses. If users are interested in the existence of this domain in other viruses, VIP DB provides a list of viruses that contains the specific domain that is accessible by clicking on that specific domain. Meanwhile, there are six separate lists of the domains found in each category of viruses that are available on the domain distribution page and the home page.

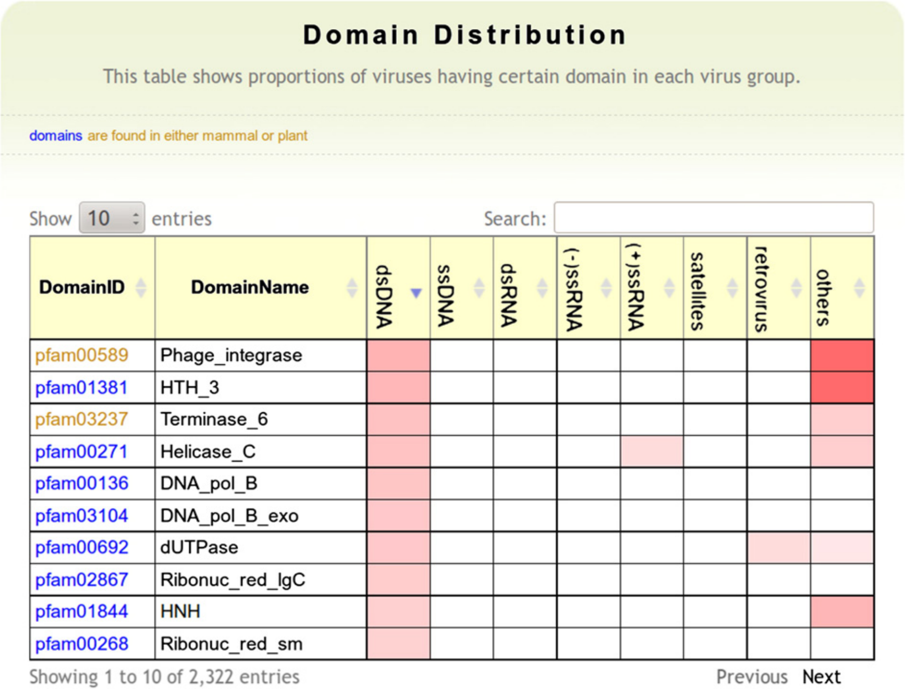
2.1.3. Investigate potential protein–protein interactions

Putative interactions between viral proteins and other proteins are also provided in VIP DB. DDI information can be used to infer candidate interacting proteins or homologs of candidate interacting proteins for viral proteins. As many interactions between proteins are derived from the interaction between domain pairs and many viral proteins had been shown to interact with viral protein or their host proteins, this allows us to link potential interaction partners. VIP DB includes DDI records and provides links to indicate DDI relationships between virus domains or between virus domains and domains found in prevalent domain architectures in mammals or plants. Overall, there are 920 viral domains with at least a DDI relationship with either itself or other domains. For those domains with DDI information, the interacting domain partner(s) is/are indicated by the DDI icon links.

2.2. A case study

The Epstein–Barr virus (EBV, also known as *human herpesvirus 4*) causes many human diseases and is prevalent among humans. Approximately 90% of the adult human population has been infected by EBV. Searching *human herpesvirus 4* on VIP DB resulted in two viral hits: *human herpesvirus 4* and *human herpesvirus 4 type 2*. Clicking on *human herpesvirus 4* indicated that there are 94 proteins in EBV, and 64 of the viral proteins contained at least one domain. These domain-containing proteins are listed in a table along with additional information, such as the number of viruses that share each domain, the number of proteins that share a domain, and the DDI and GO terms. For example, as shown in Fig. 5, HHV4_BORF2 (YP_401655) contains Ribonuc_red_lgN and Ribonuc_red_lgC. Clicking on the domain name leads the users to the domains' functional descriptions on Pfam. These two domains make up a ribonucleotide reductase (RNR) that can catalyze the formation of deoxyribonucleotides from ribonucleotides. Also shown in the table (Fig. 5), there are 18 mammalian proteins that also have this domain. We found that the same two domains can be found in the large subunit of the human ribonucleoside-diphosphate reductase, RRM1 (NP_001024.1), which contains three domains (ATP-cone, Ribonuc_red_lgN, and Ribonuc_red_lgC). This human protein is involved in purine metabolism, pyrimidine metabolism, glutathione metabolism and metabolic pathways. Although BORF2 does not contain the ATP-cone domain, it is likely to be a ribonucleotide reductase and therefore is involved in purine and pyrimidine metabolism, two pathways that are crucial for DNA synthesis. Therefore, BORF2 may be able to produce more deoxyribonucleotides during the reproduction of EBV as the concentration of deoxyribonucleotides in the host decrease due to the replication of the EBV DNA.

By using the GO information provided in the table, we found that the two domains found in BORF2 are consistent with the previously suggested function of BORF2. The GO annotations for Ribonuc_red_lgN are ribonucleoside-diphosphate reductase activity (Molecular function GO:004748), ATP binding (Molecular function GO:0005524), DNA replication (Biological process GO:0006260) and oxidation reduction (Biological process GO:0055114). The GO annotations for Ribonuc_red_lgC are ribonucleoside-diphosphate reductase activity (Molecular function GO:0004748), DNA replication (Biological process GO:0006260), and



Domain Distribution

This table shows proportions of viruses having certain domain in each virus group.

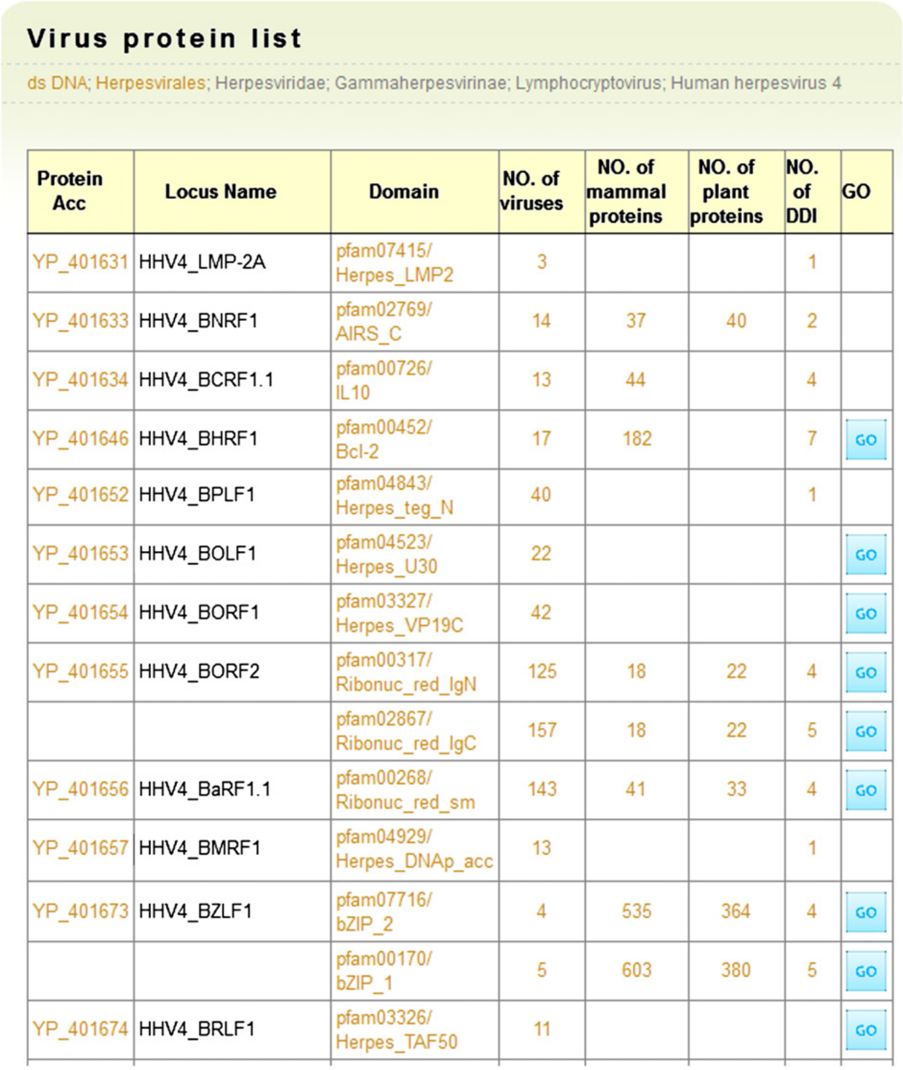
domains are found in either mammal or plant

Show 10 entries Search:

DomainID	DomainName	dsDNA	ssDNA	dsRNA	(-)ssRNA	(+)ssRNA	satellites	retrovirus	others
pfam00589	Phage_integrase	🟥							🟥
pfam01381	HTH_3	🟥							🟥
pfam03237	Terminase_6	🟥							🟥
pfam00271	Helicase_C	🟥				🟥			🟥
pfam00136	DNA_pol_B	🟥							
pfam03104	DNA_pol_B_exo	🟥							
pfam00692	dUTPase	🟥						🟥	🟥
pfam02867	Ribonuc_red_lgC	🟥							
pfam01844	HNH	🟥							🟥
pfam00268	Ribonuc_red_sm	🟥							

Showing 1 to 10 of 2,322 entries Previous Next

Fig. 4. Domain usage in viruses. All domains used in VIP DB are listed in this table, which provides the user with a comprehensive understanding of the domain distribution among the eight categories of viruses. The color of the cells represents the prevalence of the domains in viruses. A darker red color indicates a higher prevalence, while a lighter red indicates a lower prevalence. Clicking on the Pfam ID gives the user a list of viruses that contain the indicated domain. This table has sort functions for each column and provides a search function. Users can search for a domain using either its domain name or its Pfam ID.



Virus protein list

ds DNA; Herpesvirales; Herpesviridae; Gammaherpesvirinae; Lymphocryptovirus; Human herpesvirus 4

Protein Acc	Locus Name	Domain	NO. of viruses	NO. of mammal proteins	NO. of plant proteins	NO. of DDI	GO
YP_401631	HHV4_LMP-2A	pfam07415/ Herpes_LMP2	3			1	
YP_401633	HHV4_BNRF1	pfam02769/ AIRS_C	14	37	40	2	
YP_401634	HHV4_BCRF1.1	pfam00726/ IL10	13	44		4	
YP_401646	HHV4_BHRF1	pfam00452/ Bcl-2	17	182		7	GO
YP_401652	HHV4_BPLF1	pfam04843/ Herpes_teg_N	40			1	
YP_401653	HHV4_BOLF1	pfam04523/ Herpes_U30	22				GO
YP_401654	HHV4_BORF1	pfam03327/ Herpes_VP19C	42				GO
YP_401655	HHV4_BORF2	pfam00317/ Ribonuc_red_lgN	125	18	22	4	GO
		pfam02867/ Ribonuc_red_lgC	157	18	22	5	GO
YP_401656	HHV4_BaRF1.1	pfam00268/ Ribonuc_red_sm	143	41	33	4	GO
YP_401657	HHV4_BMRF1	pfam04929/ Herpes_DNAp_acc	13			1	
YP_401673	HHV4_BZLF1	pfam07716/ bZIP_2	4	535	364	4	GO
		pfam00170/ bZIP_1	5	603	380	5	GO
YP_401674	HHV4_BRLF1	pfam03326/ Herpes_TAF50	11				GO

Fig. 5. Partial table of proteins from EBV. *Human herpesvirus 4* proteins with domains are shown in this table. Information such as the number of viruses, the number of mammalian proteins, and the number of plant proteins that contain the specific domain are listed next to each domain and clicking on the numbers link to a list of viruses or a list of mammalian or plant proteins. The last two columns are the number of known domain–domain interactions and icon links to the relevant GO annotation.

andribonucleoside-diphosphate reductase complex (GO:0005971 of cellular component). In addition, based on the DDI records, the Ribonuc_red_lgN and Ribonuc_red_lgC domains are reported to interact with each other. We further examined whether these two domains are found in other viruses. There were 125 viruses that contain at least one protein with the Ribonuc_red_lgN domain, and 157 viruses contain at least one protein with the Ribonuc_red_lgC domain. All of the identified viruses with these domains are dsDNA viruses. Using the advanced search, we were further able to identify that there are a total of 125 viruses that contain both of these domains, and all of these viruses are dsDNA viruses, which again supports their suggested biological function.

2.3. Summary of domain usage in viruses

The first step in constructing VIP DB was to identify all of the domains in viruses, which provided us with a chance to analyze the overall domain usage in viruses. We found that 40.6% of the viral proteins contained at least one known domain. This low rate of domain detection suggests that viral proteins have diverged from the domains recorded in the Pfam database. Further investigation showed that the low domain detection rate was mainly due to dsDNA viruses and dsRNA viruses, as shown in Table 1. We found that the percentages of proteins with known domains (domain coverage) among the different virus categories varied widely, which may be explained by the biological nature of the viruses. The sizes of viral genomes were highly divergent as were the number of proteins in the different virus categories. Typically, only dsDNA viruses are capable of having large genomes and encoding many proteins, which is in stark contrast to most RNA viruses, which have small genomes and encode for relatively few proteins. In addition,

Table 1
Protein number and proportion of domain-containing protein in viruses.

Virus groups	# of viruses	Average # of proteins per virus genome	% of domain-containing proteins	Correlation coefficient between # of proteins and # of domain-containing proteins
dsDNA	783	89.7	36.8%	R=0.60
dsRNA	132	5.3	40.9%	R=0.61
ssDNA	426	5.9	81.1%	R=0.57
(+)ssRNA	683	3.7	81.5%	R=0.87
(−)ssRNA	135	6.4	77.7%	R=0.62
Retrovirus	102	4.6	72.5%	R=0.73
Satellites	112	1.0	89.5%	R=0.04
Others	31	41.1	48.7%	R=0.90

some viral genomes encode for polyproteins, and these polyproteins can be further cleaved into functional proteins. These polyproteins actually contain more than one functional protein and consequently are more likely to contain a known domain. In summary, many viral proteins do not contain known domains, and the domain coverage differs significantly between the different virus categories.

We also investigated the correlation coefficient between the number of proteins and the number of protein-containing domain(s). The overall correlation coefficient was 0.74, and this coefficient was relatively low compared to those found in eukaryotes (correlation coefficients of 0.99 and 0.94 for mammals and plants, respectively), but the number of proteins and the number of domain-containing proteins are still highly correlated. The correlation coefficient may be misleading due to the diverse number of proteins in the different virus categories. After further study on the correlation between the different categories, we found that although the correlation coefficient varied among the different categories, most of the categories have shown a highly correlated trend (Table 1). This finding suggests that the proportions of proteins that contain a known domain are more or less consistent across the viruses within each category.

In addition to the overall domain usage of viruses, we are also interested in how many of the domains identified in the present study were virus specific. By comparing the 2,322 virus domains with domains found in vertebrates, invertebrates, plants, fungi, and bacteria, we found that 360 of the identified domains were virus-specific domains (Supplementary file 1). Of those domains, only 31 did not have 'viral', 'virus', or 'phage' in their domain function descriptions. Manual examination of those 31 domains found that many of them have functions related to the virus-specific aspects of their life cycle (e.g., peptidase, packaging protein, structure protein that may be involved in receptor binding) or domains of unknown function. In short, of all the domains found in viruses, only a small fraction (15.5%) of them were virus specific and could not be found in other sequenced organisms. Therefore, it is likely that the functions of many viral proteins may be inferred from their domains that are shared with other organisms, either from the domain annotations themselves or from proteins containing these domains.

3. Discussion

Although many viral proteins have been reported to interact with host proteins, there are few high-throughput experiments that have been conducted to investigate the host–virus or virus–virus interactions, and few virus–host protein interactions have been recorded [2]. VIP DB suggests interaction patterns derived from the DDI information. In VIP DB, 39.6% of the domains identified in viruses have at least one domain interaction partner. Of the 2,404 viruses, less than one fourth (562) of the viruses have no DDI information for their proteins. Moreover, those 562 viruses are mostly restricted to two categories of viruses, ssDNA and satellite viruses. Most of the viruses from the other six categories have at least one DDI partner found for their proteins. Recently, several reports focused on virus–host interaction networks and found that viral proteins usually interact with the hub proteins in human interaction networks [10,24]. The candidate protein–protein interaction information provided by VIP DB is a useful resource to further examine these interaction networks.

VIP DB provides domain co-occurrence between viral proteins and mammalian or plant proteins, which may indicate functional correlations. This type of co-occurrence may also be used to infer evolutionary history because co-occurrence of protein domains can result from lateral gene transfer (LGT) between viruses and their host. Previous reports have suggested that several DNA viruses may have captured proteins from their host [25–27]. Detecting LGT using primary protein sequence comparisons may be difficult because viruses usually have a higher mutation rate, and as a consequence, their protein sequences may not be well conserved [28]. However, the domains are the functional regions of the proteins and are more likely to be conserved due to evolutionary constraints. Therefore, domains shared between viruses and hosts may be a more sensitive marker for detecting LGT. That is, the co-occurrence of domains revealed by VIP DB could be used to investigate this type of evolutionary event. As described in the Results section, only 15.5% of the domains found in viruses were unique to viruses. Therefore, VIP DB provides many domain co-occurrence data, which may be useful for host–virus LGT studies. Detailed histories of the gene transfer between viruses and their hosts may be revealed by further phylogenetic analysis of the proteins containing those domains.

The domain distribution information in VIP DB also provides a resource for systematic studies of viral genomes. The collection of overall domain distribution in viruses could be used to investigate a clade-specific domain usage in viruses. For example, as shown in Fig. 3, the domain Ebola_NP was found in all six viruses in the *Filoviridae* family and was only found in this family. This type of clade-specific domain usage may correlate with the specific traits of that clade. We further systematically searched for family-specific domains and found 868 domains that are restricted to a single family. Of those, 172 domains were found in all of the viruses in that particular family. We list all of the family-specific domains in Supplementary Table 1. Previous reports have suggested that the phylogeny could be successfully reconstructed from gene or domain contents within genomes [29–31]. The domain usage distributions provided by VIP DB across the viruses may also be used to determine phylogeny between viruses.

In brief, we constructed a viral protein domain database, VIP DB, which allows the user to investigate protein domain distribution in viruses and get information regarding the possible biological roles for viral proteins based on the included domain information. The functions of viral proteins may be inferred from the domain's GO annotations, DDI information and other proteins that share the same domains with the viral protein. This database provides integrated data that is useful for both transferring protein-function annotations and identifying pathway(s) in which the protein may be involved. In addition, as VIP DB contains DDI relationships, viral protein interaction partners may be inferred using that data. VIP DB also provides an overview of domain usage in viruses, which could be used to examine viral phylogenies or to identify clade-specific domains that may have crucial functions.

4. Materials and methods

4.1. Data resources

Viral genomes were downloaded from the NCBI genome database. A total of 2,404 viral genomes encoding for a total of 78,630 proteins were analyzed in the present study. The protein sequences from mammals, plants, invertebrates and fungi were downloaded from the NCBI RefSeq database [32]. In total, 313,435 mammalian proteins and 217,694 plant proteins were included in VIP DB. The DDI relationships were downloaded from DIMA 3.0 and DOMINE [33,34]. The relationships between Pfam and GO terms were also obtained from the DOMINE database. Pathway information for genes in human (*Homo sapiens*) and Arabidopsis (*Arabidopsis thaliana*) were downloaded from the KEGG FTP website [35].

4.2. Implementation

The structure of VIP DB is shown as a flow chart in Fig. 1. The goal of VIP DB is to identify domains in viruses and integrate information regarding the domains and GO terms, pathway information, DDI and domain co-occurrence. We identified 2,322 distinct domains that occur in all of the proteins from 2,404 viral genomes using RPS-BLAST [14]. After identifying the domains, linkages between the domains and GO terms or DDI with experimental validation were added. We also searched for viral domains that can also be found in other organisms, specifically in mammals and plants.

4.2.1. Domains in viral proteins

VIP DB aimed to reveal domain-related information by deducing the function of domains/motifs found in viral proteins. After the viral protein sequences were downloaded from NCBI, the first step was to identify a potential domain. Protein domains were identified by RPS BLAST with Pfam v24 [14,18] using the default parameters except that the expected value was set to less than 0.01. Domain hits with coverage greater than 50% of the length of the position-specific score matrix (PSSM) were used to construct the database.

4.2.2. Domains in other eukaryotes

To identify virus domains that are also present in either mammals or plants, all mammalian and plant proteins with accession number starting with NP_ or XP_, which indicate curated proteins or proteins under curation, were used in this analysis. The identification of mammalian and plant domains was performed with the same criteria as was used in viral protein domain detection except that the domain hit ranges were required to be greater than 70% of the length of the PSSMs. Domain architectures, including the orders and compositions of domains, were then discovered from the domain hit results. Proteins with the same domain architecture are usually homologous proteins, and proteins conserved through many species are more likely to have crucial biological functions [36,37]. Therefore, we were primarily interested in proteins with domain architectures that were prevalent in many organisms. Proteins were filtered based on the prevalence of their domain architectures, and the top 35% of the domain architectures were kept for further analysis. This step generated two sets of proteins that included 3856 and 3085 distinct domains in mammals and plants, respectively.

4.2.3. Pathway information and GO

To provide information on the biological functions of the identified domains, we also included known pathway information in VIP DB. For the pathway information, we chose the relatively well-studied human and *Arabidopsis* as representatives for mammals and plants. The relationships between the pathways and proteins from human and *Arabidopsis* were downloaded from KEGG [35], and the protein IDs were mapped to their pathway information through the GI numbers.

Domains are the functional regions of proteins. Mapping domains and GO annotations provides the most straightforward method for linking functions and domains. GO annotations for domains were downloaded from DOMINE [33]. Connecting the protein domain and GO information allows VIP DB to provide the GO descriptions and hyperlinks to AmiGO [38] on the website.

4.2.4. Domain–domain interactions

We obtained the DDI relationships from two structure-based databases, DIMA [34] and DOMINE [33], which contain both known and predicted DDIs. Those known DDIs were inferred from high-resolution 3D structures from iPfam and 3did [36,39]. We kept only the known interactions, which resulted in a total of 6777 DDIs after taking the union of the known DDIs from the two databases. These 6777 unique DDIs are provided on VIP DB, and the user can easily access this DDI information by linking out from one domain to its interacting domain(s).

Acknowledgments

This work was supported in part by a grant from Academia Sinica and the National Science Council.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2012.06.008.

References

[1] K. Van Vliet, M.R. Mohamed, L. Zhang, N.Y. Villa, S.J. Werden, J. Liu, G. McFadden, Poxvirus proteomics and virus–host protein interactions, Microbiol. Mol. Biol. Rev. 73 (2009) 730–749.

[2] P.O. Vidalain, F. Tangy, Virus–host protein interactions in RNA viruses, Microbes Infect. 12 (2010) 1134–1143.

[3] V. Navratil, B. de Chassey, L. Meyniel, S. Delmotte, C. Gautier, P. Andre, V. Lotteau, C. Rabourdin-Combe, VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus–host interaction networks, Nucleic Acids Res. 37 (2009) D661–D668.

[4] F.T. Vreede, E. Fodor, The role of the influenza virus RNA polymerase in host shut-off, Virulence 1 (2010) 436–439.

[5] R. Kumar, B. Nanduri, HPIDB—a unified resource for host–pathogen interactions, BMC Bioinformatics 11 (Suppl. 6) (2010) S16.

[6] T. Driscoll, M.D. Dyer, T.M. Murali, B.W. Sobral, PIG—the pathogen interaction gateway, Nucleic Acids Res. 37 (2009) D647–D650.

[7] A. Chatr-aryamontri, A. Ceol, D. Peluso, A. Nardozza, S. Panni, F. Sacco, M. Tinti, A. Smolyar, L. Castagnoli, M. Vidal, M.E. Cusick, G. Cesareni, VirusMINT: a viral protein interaction database, Nucleic Acids Res. 37 (2009) D669–D673.

[8] B. Schuster-Bockler, A. Bateman, Reuse of structural domain–domain interactions in protein networks, BMC Bioinformatics 8 (2007) 259.

[9] Z. Itzhaki, E. Akiva, Y. Altuvia, H. Margalit, Evolutionary conservation of domain–domain interactions, Genome Biol. 7 (2006) R125.

[10] Z. Itzhaki, Domain–domain interactions underlying herpesvirus–human protein–protein interaction networks, PLoS One 6 (2011) e21724.

[11] I. Shah, L. Hunter, Identification of divergent functions in homologous proteins by induction over conserved modules, Proc. Int. Conf. Intell. Syst. Mol. Biol. 6 (1998) 157–164.

[12] C. Chothia, A.M. Lesk, The relation between the divergence of sequence and structure in proteins, EMBO J. 5 (1986) 823–826.

[13] I. Letunic, L. Goodstadt, N.J. Dickens, T. Doerks, J. Schultz, R. Mott, F. Ciccarelli, R.R. Copley, C.P. Ponting, P. Bork, Recent improvements to the SMART domain-based sequence annotation resource, Nucleic Acids Res. 30 (2002) 242–244.

[14] A. Marchler-Bauer, J.B. Anderson, F. Chitsaz, M.K. Derbyshire, C. DeWeese-Scott, J.H. Fong, L.Y. Geer, R.C. Geer, N.R. Gonzales, M. Gwadz, S. He, D.I. Hurwitz, J.D. Jackson, Z. Ke, C.J. Lanczycki, C.A. Liebert, C. Liu, F. Lu, S. Lu, G.H. Marchler, M. Mullokandov, J.S. Song, A. Tasneem, N. Thanki, R.A. Yamashita, D. Zhang, N. Zhang, S.H. Bryant, CDD: specific functional annotation with the Conserved Domain Database, Nucleic Acids Res. 37 (2009) D205–D210.

[15] D. Wilson, M. Madera, C. Vogel, C. Chothia, J. Gough, The SUPERFAMILY database in 2007: families and functions, Nucleic Acids Res. 35 (2007) D308–D313.

[16] S. Hunter, R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R.D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A.F. Quinn, J.D. Selengut, C.J. Sigrist, M. Thimma, P.D. Thomas, F. Valentin, D. Wilson, C.H. Wu, C. Yeats, InterPro: the integrative protein signature database, Nucleic Acids Res. 37 (2009) D211–D215.

[17] C.J. Sigrist, L. Cerutti, E. de Castro, P.S. Langendijk-Genevaux, V. Bulliard, A. Bairoch, N. Hulo, PROSITE, a protein domain database for functional characterization and annotation, Nucleic Acids Res. 38 (2010) D161–D166.

[18] R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy, A. Bateman, The Pfam protein families database, Nucleic Acids Res. 38 (2010) D211–D222.

[19] D.A. de Lima Morais, H. Fang, O.J. Rackham, D. Wilson, R. Pethica, C. Chothia, J. Gough, SUPERFAMILY 1.75 including a domain-centric gene ontology method, Nucleic Acids Res. 39 (2011) D427–D434.

[20] N.J. Mulder, R. Apweiler, T.K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R.R. Copley, E. Courcelle, U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. Griffiths-Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S.E. Orchard, M. Pagni, D. Peyruc, C.P. Ponting, J.D. Selengut, F. Servant, C.J. Sigrist, R. Vaughan, E.M. Zdobnov, The InterPro Database, 2003 brings increased coverage and new features, Nucleic Acids Res. 31 (2003) 315–318.

[21] I. Letunic, R.R. Copley, S. Schmidt, F.D. Ciccarelli, T. Doerks, J. Schultz, C.P. Ponting, P. Bork, SMART 4.0: towards genomic data integration, Nucleic Acids Res. 32 (2004) D142–D144.

[22] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat. Genet. 25 (2000) 25–29.

[23] H. Hegyi, M. Gerstein, Annotation transfer for genomics: measuring functional divergence in multi-domain proteins, Genome Res. 11 (2001) 1632–1640.

[24] M.D. Dyer, T.M. Murali, B.W. Sobral, The landscape of human proteins interacting with viruses and other pathogens, PLoS Pathog. 4 (2008) e32.

[25] L.A. Shackelton, E.C. Holmes, The evolution of large DNA viruses: combining genomic information of viruses and their hosts, Trends Microbiol. 12 (2004) 458–465.

[26] J. Filee, N. Pouget, M. Chandler, Phylogenetic evidence for extensive lateral acquisition of cellular genes by nucleocytoplasmic large DNA viruses, BMC Evol. Biol. 8 (2008) 320.

[27] D.J. McGeoch, Molecular evolution of the gamma-Herpesvirinae, Philos. Trans. R. Soc. Lond. B Biol. Sci. 356 (2001) 421–435.

[28] H. Sakaoka, K. Kurita, Y. Iida, S. Takada, K. Umene, Y.T. Kim, C.S. Ren, A.J. Nahmias, Quantitative analysis of genomic polymorphism of herpes simplex virus type 1 strains from six countries: studies of molecular evolution and molecular epidemiology of the virus, J. Gen. Virol. 75 (Pt 3) (1994) 513–527.

[29] M. Gerstein, Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census, Proteins 33 (1998) 518–534.

[30] B. Snel, P. Bork, M.A. Huynen, Genome phylogeny based on gene content, Nat. Genet. 21 (1999) 108–110.

[31] S. Yang, R.F. Doolittle, P.E. Bourne, Phylogeny determined by protein domain content, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 373–378.

[32] K.D. Pruitt, T. Tatusova, W. Klimke, D.R. Maglott, NCBI reference sequences: current status, policy and new initiatives, Nucleic Acids Res. 37 (2009) D32–D36.

[33] S. Yellaboina, A. Tasneem, D.V. Zaykin, B. Raghavachari, R. Jothi, DOMINE: a comprehensive collection of known and predicted domain–domain interactions, Nucleic Acids Res. 39 (2011) D730–D735.

[34] Q. Luo, P. Pagel, B. Vilne, D. Frishman, DIMA 3.0: domain interaction map, Nucleic Acids Res. 39 (2011) D724–D729.

[35] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, M. Hirakawa, KEGG for representation and analysis of molecular networks involving diseases and drugs, Nucleic Acids Res. 38 (2010) D355–D360.

[36] A. Stein, R.B. Russell, P. Aloy, 3did: interacting protein domains of known three-dimensional structure, Nucleic Acids Res. 33 (2005) D413–D417.

[37] T.W. Chen, T.H. Wu, W.V. Ng, W.C. Lin, DODO: an efficient orthologous genes assignment tool based on domain architectures. Domain based ortholog detection, BMC Bioinformatics 11 (Suppl. 7) (2010) S6.

[38] S. Carbon, A. Ireland, C.J. Mungall, S. Shu, B. Marshall, S. Lewis, AmiGO: online access to ontology and annotation data, Bioinformatics 25 (2009) 288–289.

[39] R.D. Finn, M. Marshall, A. Bateman, iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions, Bioinformatics 21 (2005) 410–412.