Charting Peptide Shared Sequences Between 'Diabetes-Viruses' and Human Pancreatic Proteins, Their Structural and Autoimmune Implications

Bioinformatics and Biology Insights
Volume 18: 1–23

DOI: 10.1177/11779322241289936

Stephen A James[1,2] and Istifanus A Joshua[3,4]

[1]Department of Biochemistry, Kaduna State University, Kaduna, Nigeria. [2]School of Data Sciences, Centre of Bioinformatics, Perdana University, Kuala Lumpur, Malaysia. [3]Department of Community Medicine, College of Medicine, Kaduna State University, Kaduna, Nigeria. [4]Department of Community Medicine, College of Health Sciences, Federal University Wukari, Wukari, Nigeria.

ABSTRACT: Diabetes mellitus (DM) is a metabolic syndrome characterized by hyperglycaemia, polydipsia, polyuria, and weight loss, among others. The pathophysiology for the disorders is complex and results in pancreatic abnormal function. Viruses have also been implicated in the metabolic syndrome. This study charted peptides to investigate and predict the autoimmune potential of shared sequences between 8 viral species proteins (which we termed 'diabetes-viruses') and the human pancreatic proteins. The structure and immunological relevance of shared sequences between viruses reported in DM onset and human pancreatic proteins were analysed. At nonapeptide mapping between human pancreatic protein and 'diabetic-viruses', reveal 1064 shared sequences distributed among 454 humans and 4288 viral protein sequences. The viral results showed herpesviruses, enterovirus (EV), human endogenous retrovirus, influenza A viruses, rotavirus, and rubivirus sequences are hosted by the human pancreatic protein. The most common shared nonapeptide was AAAAAAAAA, present in 30 human nonredundant sequences. Among the viral species, the shared sequence NSLEVLFQG occurred in 18 nonredundant EVs protein, while occurring merely in 1 human protein, whereas LGLDIEIAT occurred in 8 influenza A viruses overlapping to 1 human protein and KDELSEARE occurred in 2 rotaviruses. The prediction of the location of the shared sequences in the protein structures, showed most of the shared sequences are exposed and located either on the surface or cleft relative to the entire protein structure. Besides, the peptides in the viral protein shareome were predicted computationally for binding to MHC molecules. Here analyses showed that the entire 1064 shared sequences predicted 203 to be either HLA-A or B supertype-restricted epitopes. Fifty-one of the putative epitopes matched reported HLA ligands/T-cell epitopes majorly coming from EV B supertype representative allele restrictions. These data, shared sequences, and epitope charts provide important insight into the role of viruses on the onset of DM and its implications.

KEYWORDS: diabetes-virus, autoimmune, charting, shared sequence, peptide, human pancreatic proteins

RECEIVED: November 22, 2023. **ACCEPTED:** August 21, 2024.

TYPE: Research Article

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Stephen A James, Department of Biochemistry, Kaduna State University, Tafawa Balewa Way, Kabala Coastain, Kaduna 800283, Nigeria. Email: gwatiyap@kasu.edu.ng

Introduction

Diabetes mellitus (DM), a metabolic disease encompasses a number of different conditions including, long-term hyperglycaemia, ketoacidosis and ketonuria.[1-3] The underlying pathology for all these disorders is complex, which has been associated with abnormal metabolic fuel usage resulting from pancreatic dysfunction,[4] 'insulin resistance'[1] or inadequate insulin secretion, and inappropriate glucagon secretion.[3] High glucose concentration in circulation, acetoacetate formation from uncontrolled lipolysis, and high ketone bodies generation[5] has resulted to diabetic coma or even death.[6] In some cases of the advanced disease, DM syndrome has resulted to diabetic glomerulosclerosis[7] and nephropathy which is characterized by progressive loss of renal function.[8] Classical diabetic neuropathy,[9] retinopathy,[10] cardiovascular disease,[11] has been reported. Apart from genetic disorder, autoimmune reaction,[12-15] excessive growth hormone as observed in acromegaly,[16] as well as increase in glucocorticoid level in Cushing's syndrome are some of the clouded causes for the onset of DM.[17,18] In addition, infection of the pancreases have been implicated.[19]

Viral infections from cytomegalovirus (CMV), coxsackievirus B (CVB), Rotavirus A, Mumps virus, *Influenza A virus* (IAV) from Herpesviridae, Picornaviridae Reoviridae, Paramyxoviridae and Orthomyxoviridae families respectively,[20] have been implicated in the onset of DM. However, enteroviruses (EV) are the most identified viral species associated with onset of DM.[21] It is well known that viral infection is associated with autoimmune reactions,[20] possibly as a results of peptide sharing with the human host.[22] One identified outcome of such peptide sharing is cross-reactivity between autoantigen and unknown viral epitopes leading to DM.[19] The common type of DM linked to viral infection is the type 1 diabetes (T1D; insulin-dependent diabetes mellitus [IDDM]).[21] Studies have suggested that viral infection that result to T1D may arise from direct β-cell lysis, molecular mimicry, activation of autoreactive T-cells and loss of regulatory T cells.[23] One suggested mechanism of the autoimmune reaction leading to T1D arises from the recognition of $CD4^+$ and $CD8^+$ T-cells to pancreatic self-epitope such as insulin or glutamic acid decarboxylase (GAD) and can result to cross-reactivity and destruction of insulin-producing β-cells.[20]

The human leucocyte antigen (HLA) on chromosome 6 has been identified as the major genetic region for the predisposition factor to develop T1D.[24,25] The HLA class II supertype-restricted alleles DQ, DR, and DP confer high risk in the development of autoimmune reaction directed against islet-cell antigens, which then leads to T1D disease.[26]

Evidence of viral component in EVs especially in the pancreas could stimulate the proliferation of specific β-cell self-antigen autoimmune reactions and enrich the pool of T-cells.[27,28] This is followed by the induction of inflammatory cytokines,[29] thereby aggravating the destruction of the pancreas and its function. Consequently, presence of viral antigen leads to cellular response where candidate genes for T1DM, such as MDA5, PTPN2 and TYK2, regulate antiviral responses in both β-cells and the immune system by release of type I interferons.[30] The risk with MDA5 gene is associated with the development of autoantibodies targeting B cells. Viral epitopes across the infected pancreatic β-cells triggers more release of the self-epitopes and activation of T-cells and autoimmune reaction targeting self-epitopes.[31]

Pancreatic tissue of individuals with TID reveals the presence of $CD8^+$ T-cells which immensely contribute to the formation of islet lesions.[32] Thus, unravelling the role of viral pathogenicity and its association with $CD8^+$ T-cells can provide better understanding of the host-viral interactions. This knowledge is key to identifying molecular targets for vaccine strategies and drug development.

Recently, it has been reported that $CD8^+$ T-cell response is elicited by alternative reading frame (ARF) epitopes encoded by NS1 mRNA of IAV, which have been implicated in IAV-induced autoimmunity.[31] Whereas in 'diabetes-viruses' such as EVs, the viral capsid protein epitopes interaction with host pancreatic islet proteins could potentially cause β-cell destruction by autoreactive $CD8^+$ T-cell.[33] In addition, EV and CMV can trigger autoimmune reaction via molecular mimicry of host protein via presentation of viral protein as self-molecules.[34,35] Some human protein have been identified to sustain autoimmune reaction at the islet site as a result of the of EV infection which include; protein kinase R (PKR), melanoma differentiation-associated protein 5 (MDA5), retinoic acid inducible gene I (RIG-I), myxovirus resistance protein (MxA) and HLA class I.[33]

Studies have been conducted to identify the overlapping peptide epitopes specific to T-cell which may result to autoimmune reactions.[36] The approach used was based on using positional scanning synthetic combinatorial libraries (PS-SCL), which was combination of the defined amino acids in the most active mixtures found in the screening of the PS-SCL. However, Kanduc et al used an *in-silico* approach to identified massive shared sequence similarity between viral and human peptides. The peptides overlap was majorly seen with viruses such as human T-lymphotropic virus 1, Rubella virus, and hepatitis C virus, where they suggest they may incite autoimmune reactions through molecular recognition of common motifs.[22] That may serves as molecular platform for cross-reaction during immune responses following viral infections.[37] Previous studies showed Coxsackie virus, CMV and rotavirus viral protein, VP, shared peptide similarity with host GAD65 protein and tyrosine phosphatase.[38-40] This peptide sharing occurs at pentapeptide[40] or at nonapeptide and decapeptide length[19] leading to autoimmune reaction thereby precipitating T1D disease. The HLA class I and II molecules promote antigen presentation to autoreactive $CD4^+$ T helper cells and $CD8^+$ cytotoxic T lymphocytes (CTLs), thereby facilitating β-cell recognition, thus enhancing its destruction.[33,41]

Majorly, the peptide epitopes identified to date associated with T1D-specific autoreactive leading to autoimmune reactive $CD8^+$ CTL are obtained from β-cell epitopes.[39,40,42] Herein, we applied a computational approach to chart shared peptides from viral and human sequence data using 9 residues overlap, which covers pentamers key for identifying the precise antigen-antibody recognition,[43] whereas the nonamers (9-mers) will differentiate self- from non-self-epitope.[44,45] Here, we use these overlapping peptides to investigate and predict the autoimmune potential of shared sequences between 8 viral species proteins (which we termed 'diabetes-viruses') and the human pancreatic proteins. The outcome of analyses may provide new knowledge on the immune basis for the risk of T1D following viral infection, and could be useful in the development of antiviral vaccines.

Methods

A detailed outline of the methodology used in this study is summarized in Figure 1. The methodology can be divided into 3 major steps: data retrieving, data processing, and data analysis. The data retrieving involved accessing and downloading available human and viral protein sequencing relevant to our study. To remove sequences that could skew the study, such as errors and irrelevant sequences, data cleaning was done. Thus, a clean collection of human and viral sequence data for analysis was the essence of the data processing step.

Data collection and processing

The primary protein sequences of an organism are considered to be important in the identification of structure and function of the protein, which can aid in the understanding of the critical biochemical and physiological processes in the cell[46] and the molecular interaction between 2 organisms.[47] Thus, the primary protein sequence of EV (Taxonomy ID (txid): 12059), *Lymphocryptovirus* (txid: 10375), *Rhadinovirus* (txid: 10379), Roseolovirus (txid: 40272), *Herpes simplex viruses* (*Human alphaherpesvirus 1* (HSV-1) txid: 10298; and *Human alphaherpesvirus 2* (HSV-2) txid: 10310), *Human cytomegalovirus* (HCMV) also known *Human betaherpesvirus 5* (HBV-5) (txid:10359), *Human endogenous retrovirus* (txid: 11827), *IAVs* (txid: 11320), *Rotavirus* (txid: 10912), *Rubivirus* (txid: 11040), and *Varicellovirus* (txid: 10319), were downloaded from NCBI Entrez Protein (nr) database (https://www.ncbi.nlm.nih.gov/

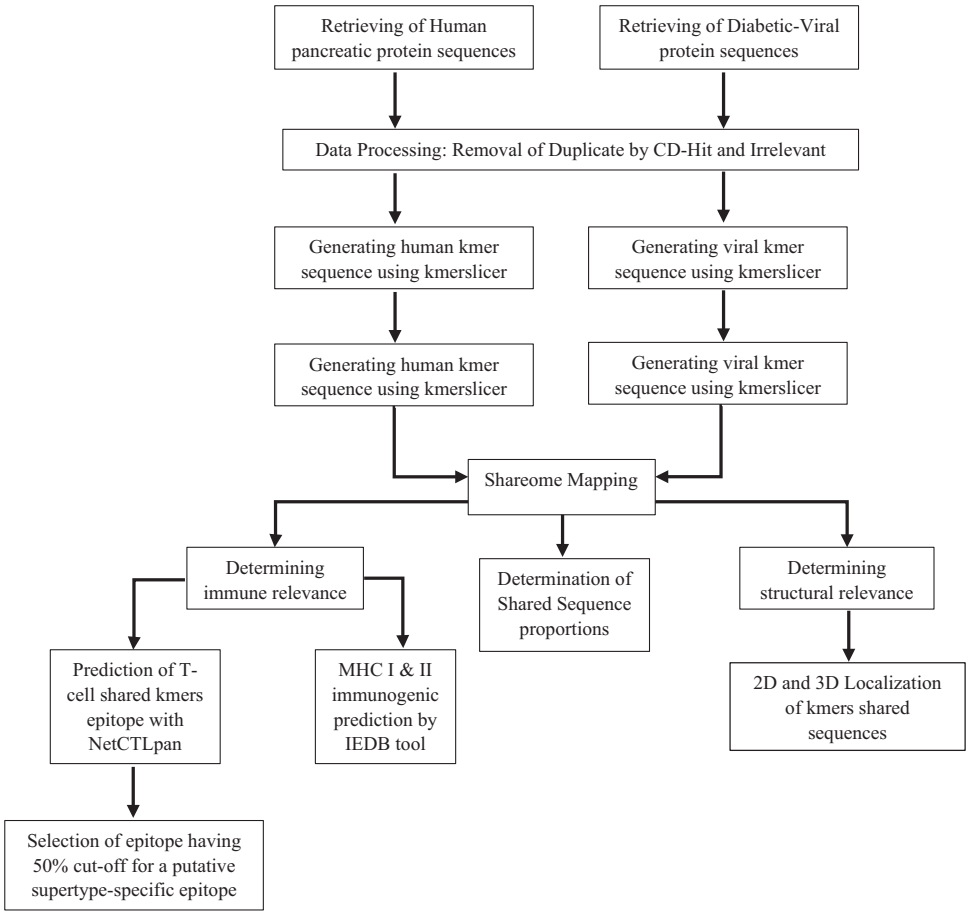

Figure 1. A Workflow for the chatting of shared sequence between diabetes-viruses and human pancreatic protein with structural and autoimmune implications.

protein/) (as of April 2020) through the NCBI Taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/). These viral protein sequences were chosen because they are reported to be associated with DM,[20,21] and the retrieval was based on either species or genus that harbor the species that infect humans, using taxonomy filter keyword.[48] Likewise, the human pancreatic protein sequences were chosen because of their specific association to carbohydrate metabolism and their presence in circulation when the pancreas is diseased.

The data collection was then followed by removal of duplicate. Each viral specie or group of species was removed of any duplicate sequences using CD-Hit software.[49] To a different file all the nonredundant viral protein of each species sequences were emerged into a single file (referred to as diabetic viral species-DVS file). Similarly, human protein sequences retrieved were removed of any duplicate sequences.

Shared sequence molecular match

Shared sequences analyses were conducted between 12 viral proteome, DVS protein sequences for nonapeptide match with human pancreatic protein sequences. Charting of the matched molecular pairs was carried out using 8 overlapping amino acid residues that originates from the same protein sequence. Each viral (or human) nonapeptide dictionary was constructed using '*kmerslicer*'[50] by window sliding method shifted by one amino acid residue. Identical nonapeptide were merged by collapsing the 'AMONG format' and joining the metadata of similar 9-*mer*, as describe by James et al. This allows for unique 9-*mers* linearization for each virus or human protein sequence sources by removing duplicates.

The unique viral dictionary was then used to map the entire human dictionary for identical match of 9-*mers* (nonamers), by employing UNIX string manipulator. Any occurrence of identical overlapping nonamers sequences between the viral-human dictionaries was considered a match.

Analyses of the shared sequences occurrence and proportion were carried out. The shareome count for the distribution of each nonapeptide was also determined. The charted nonapeptide (otherwise 'diabetes viral'- 'human pancreatic' shareome) catalogues, thus serve as conduit for the immunological studies.

2D and 3D structural and surface accessibility of shared sequences

In elucidation of all set of proteins in cells, it is important that specific function of the protein is identified. Furthermore, the

topology of protein is used in the structural characterization of protein,[51] and conservation of protein is significantly valuable than the structure; hence, structural similarity is inference to protein function, which is aligned by databases for analysis.[52] Here, we determine the structural properties of the shared sequences, such as 2D and solvent accessibility of the amino acid residue make-up of the shared sequences in the protein fold, which can be valuable information for understanding possible mechanistic role in viral-human interactions and protein functions.

The structural characterization of diabetes-viruses-human protein shared sequences was carried out by extracting both viral and human Protein Data Bank (PDB; https://www.rcsb.org/) IDs found through share-ome map. Sequences from the PDB that are from same species showed 100% identity with any of the shared nonapeptides were considered as input for the surface accessibility and 2D analyses. In addition, only viral sequences that infect the human-host were selected for this analysis. However, due to most of the crystal structures have missing atoms at the first few residues when compare to the FASTA format; therefore, the 3D localization analyses were carried by via homology modeling. Since most of the consensus sequent are not on the PDB format for visualization using visual molecular dynamics (VMD). We equally explore the non-PDB protein for the 2 D analyses. But only the human and viral proteins (that infect humans) with the highest shared sequences and one protein type were analysed.

To understand the localization of the shared sequences in the 3D structure of the proteins involved in shareome map, homology modelling was carried out using the web-based interface of SWISS MODEL (https://swissmodel.expasy.org/). FASTA sequences of Thymidylate synthase isoform 1 (NCBI: NP_001062.1), HLA class I histocompatibility antigen, A alpha chain (NCBI: P04439.2), ORF70 (NCBI: ALH45390.1) and Membrane glycoprotein UL40 (AMJ54243.1) proteins were separately submitted to the modelling query box of the SWISS MODEL. The 3D structures were built base on the principle of the homology modelling,[53] first by finding appropriate experimental templates, then aligned the chosen PDB template structure, followed by using the alignment, a 3D model of the target protein to construct a full structure. The quality of the generated model is assessed using various scoring functions and validation tools by the web SWISS MODEL. The best model was chosen and downloaded, respectively.

Identification of pathological potential of the shared sequences

To verify if the shared sequences play important role in the viral-human protein interaction that facilitates viral pathogenicity, the functions of the human protein nonapeptide shared sequences and viral epitope were analysed. Thus, to elucidate the function and interactions site of the protein of shared sequences, the catalogue of the FASTA shared nonamers sequences were submitted as query to the ScanProsite tool (https://prosite.expasy.org/scanprosite/) and NCBI Entrez conserved domain database (CDD) tool (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi; v3.17)[53] to determine domains and motifs and the conversed domain respectively. Further query against the Immune Epitope Database (www.immuneepitope.org)[54] to identify all reported immunogenic human T-cell epitopes (both class I and II) that fully overlapped with the predicted epitopes from NetCTLpan server, were carried out.

To predict the HLA class I supertype-restricted epitopes, all viral shared sequences were converted to FASTA format and submitted to a local instance (standalone) of the NetCTLpan 1.1 tool (http://www.cbs.dtu.dk/services/NetCTLpan/). To carry out the prediction the query terms received by NetCTLpan include list of alleles for a specific supertypes in a file set at the app default of 0 alongside with a file containing of each shared sequence in FASTA format. Here, the NetCTLpan 1.1 tool was run using a bash script that also extracts fields from the predicted output of the results that are pertinent to the anticipated epitopes as in the batch script below:

```
#!/bin/bash
input="/PATH /allelesfile.txt"
while IFS=read -r line
do
# echo "$line"
./netCTLpan -a "$line" -s 0 sharedSeq.fasta >> pan_output.txt
# Extract only the line specifying the supertype predicted and
epitopes without "Number. . ." line
grep '<-E' pan_output.txt| awk -F" " '{print
$1,$2,$3,$4,$5,$6,$7,$8,$9,$10}' > pan_predicted.txt
done < "$input"
```

Each shared sequence identify and used for the MHC class I binding prediction with the NetCTLpan 1.1[55] standalone scripts (command line tools) yield identical outcome as the web server version, with default parameters. The protocol involve integrated prediction of peptide MHC class I binding, proteasomal C terminal cleavage and TAP transport efficiency.[55] Where 'a' in the script above is the input 'ALLELE' (default = HLA-A02:01), 's' is 'SUPERTYPE (default = A1). Other default setting is the 'CLEAVAGE WEIGHT, ie, weight on C terminal cleavage (for netctlpan = 0.225) and TAP WEIGHT, ie, weight on TAP transport efficiency (default $t = 0.025$).

To identify all reported immunogenic, human T-cell epitopes/ligands (both class I and II) that were fully involved in the viral nonapeptide overlapped with the predicted epitopes from NetCTLpan, additional analysis via query against the Immune Epitope Database (www.immuneepitope.org)[54] was carried out. Finally, analysis of likely pathological potential of

Table 1. Viral-Human data sets used and proteome overlap at the nonapeptide level.

ORGANISM/PROTEINS/ PROTEOME	SEQUENCE RETRIEVED	NONDUPLICATE SEQUENCES	NUMBER OF NONAPEPTIDE GENERATED	NONREDUNDANT NONAPEPTIDE	HUMAN PROTEOME OVERLAP AT NONAPEPTIDE WITH VIRUSES
Human pancreatic proteins	17095	13988	4896943	788009	-
Diabetes viruses	725184	185333	92066730	4087382	1064
Herpesviruses	228692	44341	22853909	2160739	660
Enterovirus	86609	30912	17367403	637167	19
Human endogenous retrovirus	498	321	94442	16324	04
Influenza A viruses	322710	73181	35720905	499670	35
Rotavirus	83674	35885	15585042	755360	393
Rubivirus	3001	693	445029	20033	02

nonapeptides cross-reactivity by examining the functional relevance of the human proteins involved in the viral epitope overlap was determined.

Results

Viral-Human protein sequence data set

A systematic overlapping sequence mapping and analysis was carried out on 'diabetes-viruses' (pathogen) and human protein sequences. Here, we studied the possible functional relevance of viruses reported to induce diabetes versus human nonapeptide overlap by considering the protein shared sequences structure and immunological relevance, through prediction of the location of the shared sequences in the 3D structure, identification of epitopes and HLA supertype-restricted epitopes. Peptide commonality existing between the viruses and the human protein sequences retrieved was explored, using nonapeptide (9-*mers*) for its functional relevance in differentiating self and non-self-epitopes.

Available data as of May 2020 (Table 1), collected from NCBI Entrez Protein database, contained a total of 725184 sequences (FASTA format) representing search return for the 'diabetes viruses'. From these entries, all identified viruses (Herpesviruses, EV, IAVs, etc.) reported to have potential to promote cross-reactivity with the pancreatic cell were among the data retrieved (Table 1). The removal of duplicates at 100% identical protein sequences reduced these downloaded sequences to 185333 (~26% reduction–for 'diabetes viruses'). Similarly, human protein sequences (full-length and partial) were downloaded in FASTA format, with their corresponding meta-data. Around 17095 human pancreatic proteins were retrieved on removal of duplicate the initial downloaded human data reduced to 13988 protein sequences. Careful investigation of the human metadata (GenPet records) retrieved, showed some protein entries collected were non human sequences. Most (~65% of the initial download) were identified to be of viral, bacterial, and nonhuman vertebrate origin (Supplementary Material 2). The removal of the irrelevant sequences reduces the human data set significantly to 2887 (~17% reduction) protein sequences. These large variations in the data sets show that database houses duplicate sequences and may accompany with homology sequences in the metagenomics data sets and protein databases[56] during data collections.

Based on shared sequences analysis, the diabetic viral-human share-ome consist of 1064 (Supplementary Table 1 and Supplementary Material 1) peptides arising from 12708 (~0.31% of 4087382) and 2355 (~0.30% of 788009) viral and human nonredundant nonapeptides, respectively. These shared sequences originated from 4288 viral and 454 human nonredundant protein sequences, that came from 725184 and 17095 initially downloaded respectively (Table 1).

The qualitative analysis of the nonapeptide distributions of various protein shared, showed that herpesviruses had the most overlapping shared sequence of 660 (~62%), contributed from various groups of herpes retrieved; including: *HCMV*, *lymphocryptovirus*, *rhadinovirus*, *roseolovirus*, *simplexvirus*, and *varicellovirus* (Table 2). This group of viruses listed are among those identified to infect the human host and harmonious live for the rest of their lives.[57] In addition, *rotaviruses* contribute the second major nonapeptide marching accounting for ~37% (393) of the shared sequence. But *Rubivirus* only account for ~0.19% (02) of the total 'diabetes viruses'-human share-ome (Table 1).

Analysis of 'diabetes-viruses' proteome, the nonapeptide AAAAAAAVA (Figure 2; Supplementary Table 1) were the most shared sequences; this occurred in 420 nonredundant protein sequences, conversely, the same sequence overlaps with 8 human protein sequences. This shared sequence originated from herpes group of viruses (Figure 3; Supplementary Table 2). However, among the various group of viruses in the

Table 2. Herpesviruses proteome downloaded.

ORGANISM	PROTEIN SEQUENCES RETRIEVED
Human cytomegalovirus	59659
Lymphocryptovirus	53395
Rhadinovirus	11393
Roseolovirus	25891
Simplexvirus	52781
Varicellovirus	18726

'diabetes-viruses' proteome, the shared sequences NSLEVLFQG occurred in 18 nonredundant EV proteins (Figure 4), while occurring merely in 1 human protein. LGLDIEIAT occurred in 8 *IAV* overlapping to 1 in human protein sequences (Figure 5); KDELSEARE occur in 2 *rotaviruses* to 1 in human. A total of 606 'diabetes viruses' protein sequences contained a single shared sequence, mapped majorly to not more than 2 human nonredundant protein sequences (Supplementary Table 1).

The most shared nonapeptide human pancreatic protein sequence was AAAAAAAAA (Figure 2; Supplementary Table 1), present in 30 human nonredundant sequences; this nonapeptide was substantially present in 258 'diabetes viruses' sequences. This shared sequence is notably contributed by Herpesviruses (Figure 3; Supplementary Table 2). Among Rotaviruses nonapeptide with human pancreatic protein, the shared sequence HHHHHHSSG occurs 8 in human nonredundant protein, only mapped to a single viral protein (Figure 6; Supplementary Table 3). However, this same nonapeptide is shared by 6 IAV proteins and 2 EV proteins, respectively (Figures 4 and 5). Majorly, the human nonapeptide overlap was rare in viruses, where the least shared sequence was 654 with each present in a single protein (Supplementary Table 1).

A total of 22 nonapeptide were identified to be shared at least by 5 viral species. Thirteen of the shared sequences SSHHHHHHS, SSGLVPRGS, SHHHHHHSS, SGLVPRGSH, MGSSHHHHH, HSSGLVPRG, HHSSGLVPR, HHHSSGLVP, HHHHSSGLV, HHHHHSSGL, HHHHHHSSG, GSSHHHHHH, and GLVPRGSHM are shared by *Herpesviruses*, *Rotavirus*, *IAV*, and EV*s*. While nonapeptide GGGSGGGSG is shared by only 3 viral species namely *Herpesvirus*, *IAV* and *Rubivirus*. The remaining 8 nonapeptides VKGRFTISR, SVKGRFTIS, PGGSLRLSC, KGRFTISRD, HHHHHHHHH, HHHHHHHHG, GRFTISRDN, and DSVKGRFTI are shared between Herpesviruses and *IAV*. All the viral species in the identified 'diabetes viruses'- 'human pancreatic protein' shareome consist of at least a single shared sequence, of which only HERV shared sequences were not found to be shared by any other viruses in this shareome. These sequences include; EGKWSEVPY, GKWSEVPYV, KWSEVPYVQ and NKSKKRRNR nonapeptides. The human pancreatic proteins sharing nonapeptides(s) with the 'diabetes viruses' and the overlapping nonapeptides are described in detail in Supplementary Tables 1 to 3.

Shared sequence occurrence and distributions vary among both proteomes, that constitute the 'diabetes viruses'- 'human pancreatic protein' shareome. For example, the most frequent human protein involving in the shareome map is the 'Transforming growth factor-beta-induced protein ig-h3 precursor' protein, which contributes 375 nonapeptides (Table 3). Also, some diverse human proteins present in the shareome include: 'Thymidylate synthase isoform(s)', 'HLA class I histocompatibility antigen, A alpha chain', and 'Natural cytotoxicity triggering receptor 3 ligand 1 precursor', among others.

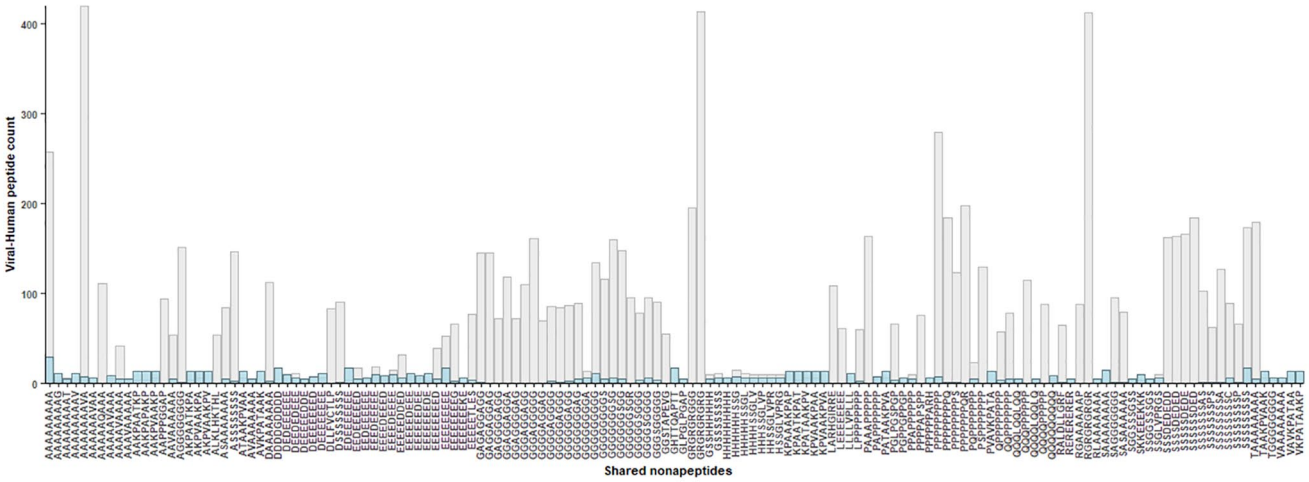

Figure 2. Nonapeptide sharing between 'diabetes viruses' proteome and human pancreatic protein sequences. The grey colour report the number of occurrences of 'diabetes viruses' nonapeptides in the shareome overlap. While the light-blue colour shows the occurrences of the human pancreatic nonapeptides in the shareome overlap.

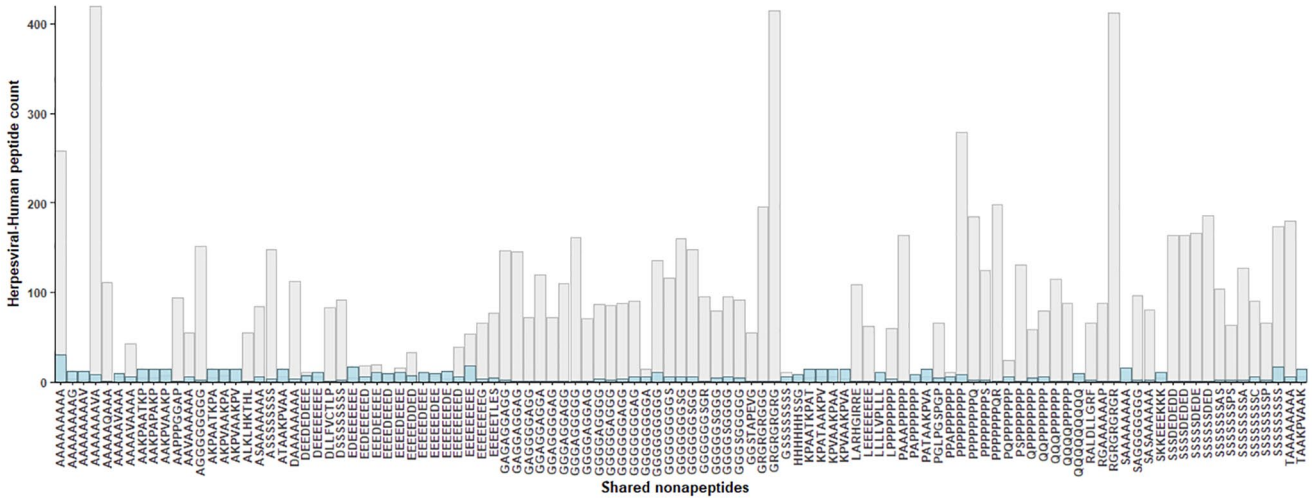

Figure 3. Herpesviruses-Human shared sequences distributions.

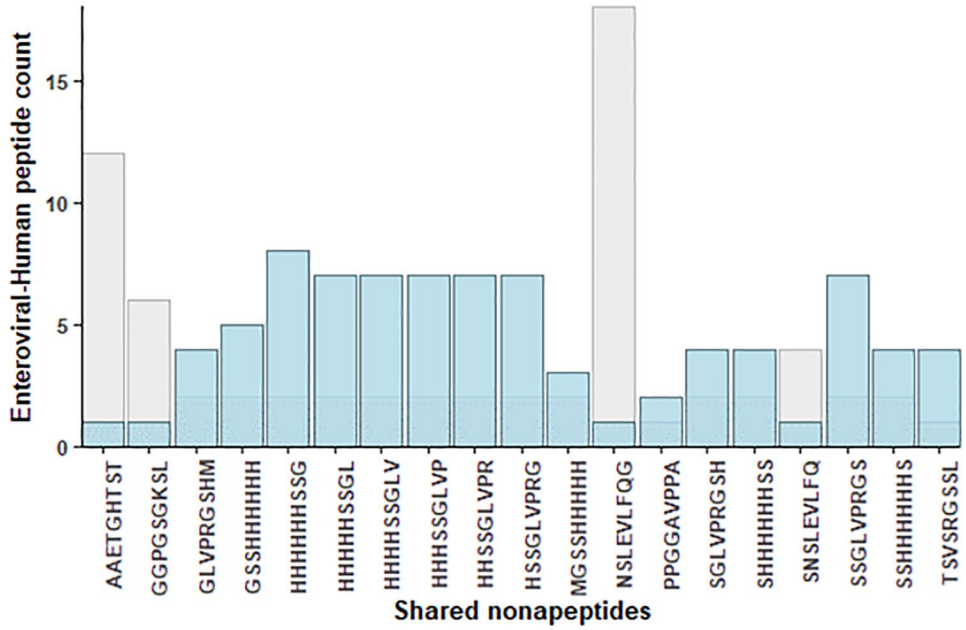

Figure 4. Enteroviruses-Human shared nonapeptides distribution.

Similarly, the 'diabetes viruses' protein with the most abundant shared sequences is 'RGD-containing collagen associated protein', consisting of 375 shared nonapeptides from *Rotavirus* (Table 3). In addition, 'Chain A, Polymerase acidic protein' from *IAV* contribute 14 shared nonapeptides, 'Thymidylate synthase' from Herpesvirus contribute 57 nonapeptide, and 'Chain A, Genome polyprotein' from EV*s* contribute 13 shared sequences, among others (Table 3).

The occurrence of short sequence in a protein has been described as a positive selection pressure behind the accumulation of amino acids repeat, which may be known functional domain, or notably a disordered region, making amino acid repeats (AARs) contribute to functionality of proteins by providing flexibility, stability and as linker elements between domains.[58] Amino acid repeats of various degrees were identified within the 'diabetes viruses'- 'human pancreatic protein' shareome (Supplementary material 1). Approximately 11% of the shared sequences were nontandem repeats (NTRs) AARs. These include: AAKPVAAKP, GHTTQAPTT, PQGPPGPPG and RALDLLGRF, among others. The shared nonapeptide GRGRGRGRG occurs 414 as simple/NTRs sequentially interspersed in viral proteins mapped to a single human pancreatic

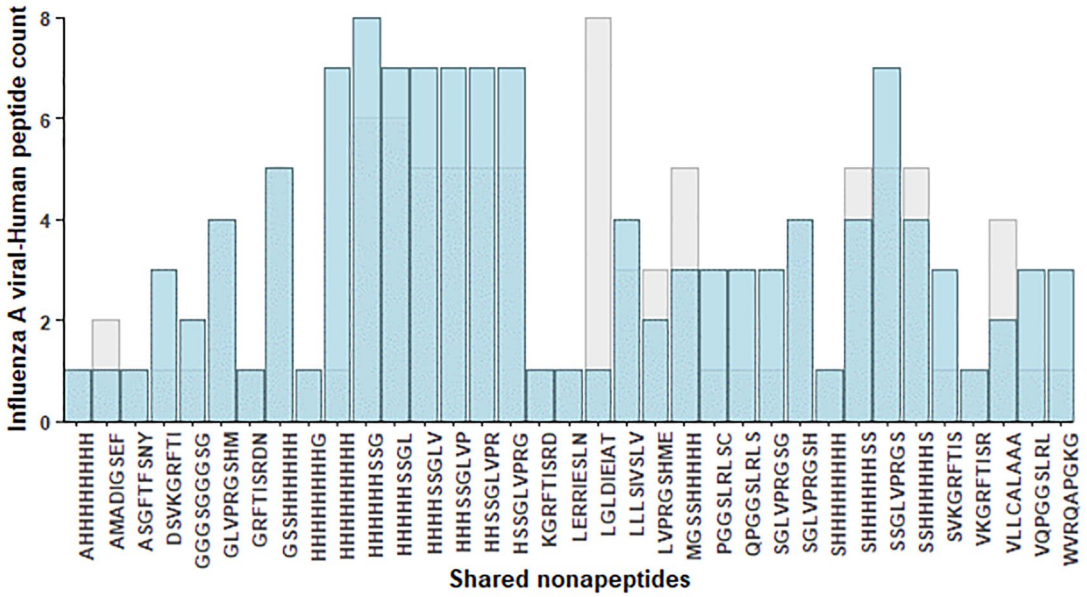

Figure 5. Influenza A Virus-Human shared nonapeptides distribution.

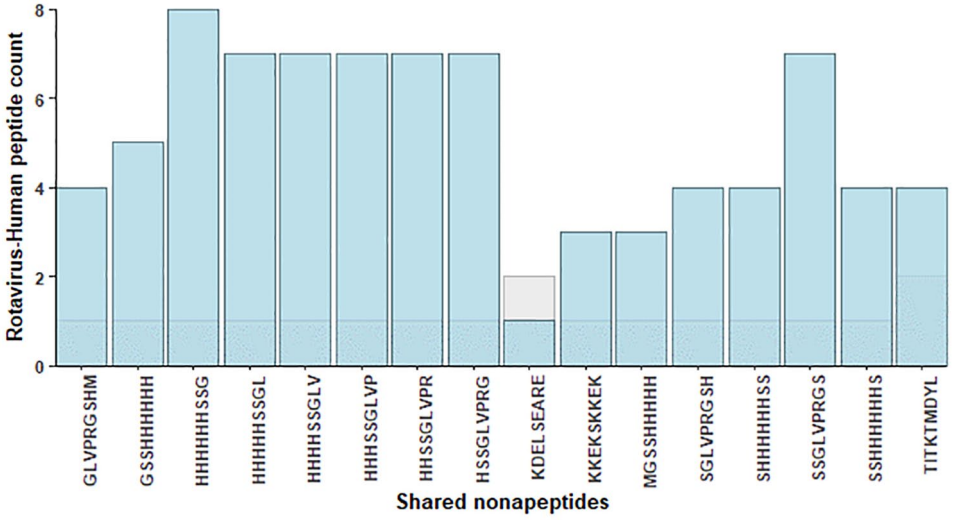

Figure 6. Rotavirus-Human shared nonapeptides distribution.

protein `Q9UMN6.1`: Histone-lysine N-methyltransferase 2B. In contrast, `QBM05694.1`: EBNA-2 protein *Human gammaherpesvirus 4* (*Epstein-Barr virus* [EBV]) protein has this same shared sequence in 13 positions. Interestingly, the nonapeptide `GAGGAGGAG` is well distribution and was found in 29 diverse positions in the *Human gammaherpesvirus 4* viral protein 'Epstein-Barr nuclear antigen 1'. Also, this shared sequence occurs in multiple position of 415 viral proteins in the share-ome overlapped with just a single human protein 'one cut domain family member 3' (NCBI ID: `NP_001073957.1`). Most of the NTRs appear in 2 positions of the viral proteins involved in the shared sequences; while 20 shared sequences appeared in 2 or more position across several human proteins in the shareome. For instance, the shared nonapeptide `KGKGKGKGK` from human 'Chain A, DNA (cytosine-5)-methyltransferase 1' (NCBI ID: `pdb|4YOC|A`) protein is shared with 4 viral proteins, `EEEEEDEEE` from human 'amyloid-like protein 2 isoform 2 precursor' (NCBI ID: `NP_001135748.1`) protein is shared with 9 viral proteins, and `AAAVAAAAA` from 'deformed epidermal autoregulatory factor 1 homolog isoform b' (`NP_001280563.1`) protein is shared with 5 viral protein sequences. In addition, 8 recognizable perfect (simple or tandem) AARs were identified in the shared sequence analysis map. These include: `AAAAAAAAA`, `EEEEEEEE`, `GGGGGGGG`,

Table 3. List of some major human pancreatic protein and 'diabetes viruses' proteins with their occurrences of shared nonapeptides.

ORGANISM	PROTEINS (NCBI ACCESSION NUMBER)	NUMBER OF PROTEINS
Homo sapiens	NP_000349.1: Transforming growth factor-beta-induced protein ig-h3 precursor	375
	NP_001062.1: Thymidylate synthase isoform 1	72
	NP_001341796.1: Thymidylate synthase isoform 2	68
	NP_001341797.1: Thymidylate synthase isoform 3	51
	NP_006159.2: Homeobox protein Nkx-6.1	24
	AAD11962.1: NK homeobox protein	24
	sp|P04439.2|HLA: HLA class I histocompatibility antigen, A alpha chain	23
	NP_054862.1: programmed cell death 1 ligand 1 isoform a precursor	22
	NP_001300958.1: programmed cell death 1 ligand 1 isoform c precursor	22
	pdb|5MQV|F: Chain F, Casein kinase I isoform delta	14
	pdb|5Y16|B: Chain B, Dual specificity phosphatase 28	13
	pdb|2QKH|A: Chain A, Glucose-dependent insulinotropic polypeptide receptor	13
	pdb|3E46|A: Chain A, Ubiquitin-conjugating enzyme E2-25 kDa	9
	pdb|5W21|B: Chain B, Fibroblast growth factor 23	8
	pdb|6O3B|G: Chain G, Antibody Fab F6, Heavy chain	6
	sp|O94875.3|SRB: Sorbin and SH3 domain-containing protein 2	3
	NP_963883.2: transcription factor MafA	3
	sp|Q92908.2|GAT: Nuclear receptor subfamily 4 group A member 3	1
	NP_001189368.1: Natural cytotoxicity triggering receptor 3 ligand 1 precursor	3
	NP_006483.2: homeobox protein aristaless-like 3	1
Influenza A virus	pdb|5VRJ|A: Chain A, Polymerase acidic protein	14
	pdb|5EG9|B: Chain B, Polymerase basic protein 2	13
	pdb|4XNM|D: Chain D, Hemagglutinin	11
	pdb|6QXE|O: Chain O, Nb8205	10
	pdb|6I3H|B: Chain B, Matrix protein 1	2
	pdb|4WSB|A: Chain A, Polymerase PA	2
	sp|P03452.2|HEM: Hemagglutinin	1
Rotavirus	BAA20089.1: RGD-containing collagen associated protein	375
	pdb|2B4I|C: Chain C, Outer capsid protein VP4	13
	sp|Q45UF6.1|NSP: nonstructural protein 1	1
	AYH64822.1: VP2	1
	AWH66514.1: nonstructural protein NSP5	1
	AIK97649.1: VP7	1
	AHZ33078.1: nonstructural protein 2	1
Retrovirus	AAD48375.1: gag polyprotein, partial	3
	CAA71420.1: pol protein, partial	1

(Continued)

Table 3. (Continued)

ORGANISM	PROTEINS (NCBI ACCESSION NUMBER)	NUMBER OF PROTEINS
Enterovirus	pdb|5GSO|E: Chain E, 3C protein	13
	pdb|3W95|A: Chain A, Genome polyprotein	13
	QFG58510.1: Polyprotein	2
	CEG62456.1: Capsid protein VP1, partial	1
	BAM36055.1: Capsid protein, partial	1
Herpesvirus	YP_009227216.1: Thymidylate synthase	57
	ALH45390.1: ORF70	43
	AEW87717.1: JM22	31
	ABX74967.1: Thymidylate synthase-like protein, partial	25
Rubivirus	sp|Q99IE7.1|POL: nonstructural polyprotein p200	1
	QED08857.1: nonstructural polyprotein, partial	1
	QED08855.1: nonstructural polyprotein	1

HHHHHHHH, PPPPPPPP, QQQQQQQQ, SSSSSSSS, and TTTTTTTT.

Viral species distribution analysis showed that most of the viral proteins in the 'diabetes viruses'- 'human pancreatic protein' share-ome comes from HHV-5 (formerly known as *HCMV*). This HHV-5 virus contributes ~33% (of the 4288) viral nonredundant protein sequences that contained the shared nonapeptides (Figure 7). *Human gammaherpesvirus 4*—EPV (EBV) account for ~26% of the shared 'diabetes viruses'-'human pancreatic protein' shareome. Some notable viruses that appeared had a significant drop in the shared sequences such as *Human alphaherpesvirus 1*-HSV1 (*Herpes simplex virus type 1*), *Human alphaherpesvirus 2*—HSV2 (*Herpes simplex virus 2*) contributes ~14% and ~11% of the shared nonapeptides respectively. Other viral species that contributed less less than 2% shared nonapeptide include: *Rubella virus*, *Human gammaherpesvirus 8*, *HBV 6*, *IAV*, *HBV 6B*, *Rhinovirus B*, *Rhinovirus A* and *HBV 6A*, among others (Figure 7). Interestedly, nonhuman viral species where also involved in the shared nonapeptide overlap. These include, *Macacine alphaherpesvirus 1* (~2%), *Macacine betaherpesvirus* 3 (~1%), *Papiine alphaherpesvirus 2* (~1%), and *Cercopithecine betaherpesvirus 5* (~1%), among others (Figure 7).

Structural visualization of shared sequences across human and viral proteins

A total of 33 PDB crystal structures were identified in the share-ome map and 20 where present for the human (see Supplementary Table 4). Using the selection criteria as earlier discussed in method section, only 3 human PDB proteins were used against 5 viral PDB proteins. The consensus shared sequences for human and the viral proteins are shown in Table 4. Here, the human protein 5Y16_B: 'Chain B, Dual specificity phosphatase 28' has a single consensus shared sequence, while 5 viral proteins (see Table 4) consensus shared sequence are mapped to this human protein.

The secondary structure prediction of the human protein 5Y16_B shows 21 residues of the shared consensus sequences '$_{1}$MGSSHHHHHHSSGLVPRGSHM$_{21}$' in this protein are exposed and the residues are disordered (see Figure 8). Similarly, this consensus shared sequence occurs in 4 viral proteins namely (see Table 4): *Human alphaherpesvirus 1* (*Herpes simplex virus 1*–HSV-1) 'Chain C, Tripartite terminase subunit UL15 (4IOX_C)', *IAV (A/California/04/2009(H1N1)* 'Chain A, Polymerase acidic protein (5VRJ_A)', *Enterovirus A71* (EV-A71) 'Chain E, 3C protein (5GSO_E)' and *Human enterovirus 71 (strain 7423/MS/87)* 'Chain A, Genome polyprotein (3W95_A)'. On the contrary, *IAV* Chain D, Hemagglutinin (4XNM_D) consists of 19 residues '$_{1}$MGSSHHHH HHSSGLVPRGS$_{19}$' in the consensus shared sequences that are in the human protein. These viral consensus sequences are equally exposed (Figure 8). Similarly, the prediction of surface accessibility of other human and viral protein is shown in Figures 9 and 10.

Another observation is between human 'Chain A, Pancreatic lipase-related protein 1' (PDB: 2PPL_A) and Human herpesvirus 5 strain AD169 'Chain D, UL89 HCMV' (PDB ID: 6EY7_D) protein. Their consensus shared sequence is HHHHHHDYDIPTTENLYFQG spanning a long position 4-23 in human and 3-22 in HHV-5 strain strain AD169 proteins. This sequence in human 'Chain A, Pancreatic lipase-related protein 1' spans through coil and disordered region of the protein (Figure 9A). The concensus amino acids shared sequences

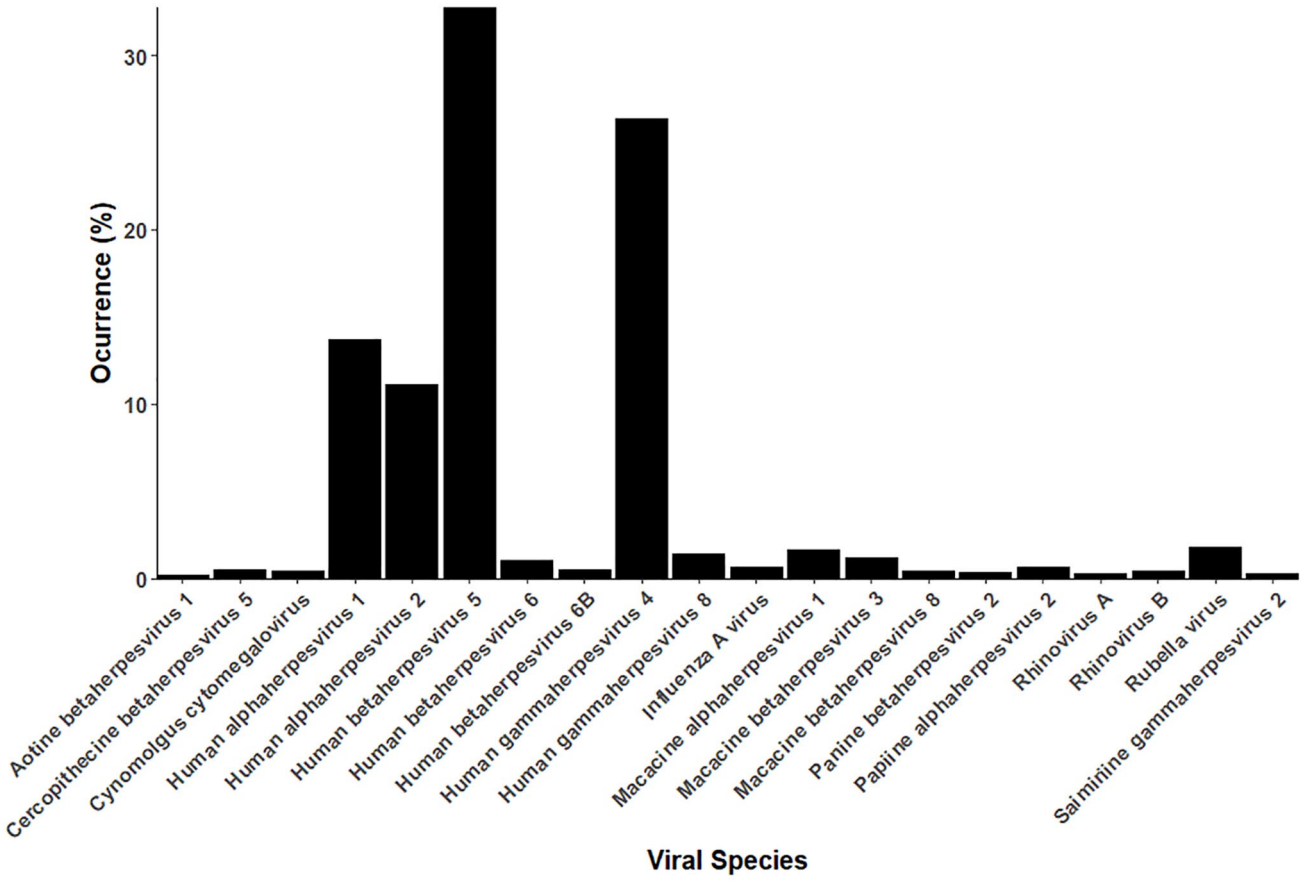

Figure 7. Major viral species distribution in 'diabetes viral'-human shared sequences.

are well exposed relative to other part of the protein. Similarly, the viral HHV-5 strain strain AD169 protein 'Chain D, UL89 HCMV' has its concensus amino acids covering the coil and diosorded region of the protein and they are also well exposed just as that on the human 'Chain A Pancreatic lipase-related protein 1'.

Structural characterization of human and 3 'diabetics viruses' shareome screening shows that human 'Chain F, Casein kinase I isoform delta' (PDB ID: 5MQV_F) shared similar consensus shared sequences $_{1}$MGSSHHHHHHSSGLVPRGSHME$_{22}$ with *Human enterovirus 71 (strain 7423/MS/87)* 'Chain A, Genome polyprotein' (or '2A proteinase (C110A)') (PDB ID: 3W95_A), *Enterovirus A71* (EV-A71) 'Chain E, 3C protein' (PDB ID: 5GSO_E) and *IAV (A/California/04/2009(H1N1))*" Chain A, Polymerase acidic protein" (PDB ID: 5VRJ_A). However, in the consensus sequence for *Human enterovirus 71 (strain 7423/MS/87)* and *Enterovirus A71* the 'E' residue is missing making the consequences to be only 21 residues. The NetSurfP analysis of these proteins shows the amino acids residues of the shared sequences are 100% exposed and the consensus sequences spans through coil regions of the proteins (Figure 10). Although, the human Casein kinase I isoform delta the protein (Chain F) and *IAV (A/California/04/2009(H1N1))* Polymerase acidic protein (Chain A) have few of their amino acid residues of the shared sequences in the strand (Figure 10A) and helix (Figure 10D) region of the proteins, respectively.

Another involving charting is the human protein NP_001062.1: 'Thymidylate synthase isoform 1' that has consensus shared sequences discontinued into 7 parts, while in the viral *Human gammaherpesvirus 8* (HHV8) protein ALH45390.1: ORF70 has 2 consensus shared sequences (see Table 4). On the 2D structure, the human Thymidylate synthase isoform 1 have region that are well buried and exposed. For example, the first consensus shared sequence '$_{49}$DRTGTGTLSVFGMQARYSLRDEFPLLTTKRVFW$_{81}$' has ~61% of its residue buried and ~39% exposed, while they cover the coil and strand regions (Figure 11). Similarly, the remaining consensus sequences lie within the region of expose, buried and/or helix, strand, or coil (Figure 11). For instance, the consensus shared sequence "$_{207}$ELSCQLYQRSGDMGLGVPFNIASYALLTYMIAH$_{239}$' on the 2D shows that ~6% of the residues are exposed, while ~94% of the residues are buried. Of these buried residues, most of them are in helix region and few lies within the strand and coil segments. Comparably, the HHV8 viral protein 'ORF70' consensus consequences '96PLLTTKRVFW105' and '231ELSCQLYQRSGDMGLGVPFNIASYALLTYM260' residues are well buried and occupying the strand, helix and coil regions. The remaining

Table 4. Some identified consensus sequences present in the 'diabetes viral'-human shared protein sequences.

HUMAN PROTEIN/CONSENSUS SHARED SEQUENCES	MAPPED VIRAL PROTEIN/CONSENSUS SHARED SEQUENCES
5Y16_B: Chain B, Dual specificity phosphatase 28 $_{1}$MGSSHHHHHHSSGLVPRGSHM$_{21}$	4IOX_C: Chain C, Tripartite terminase subunit UL15 $_{1}$MGSSHHHHHHSSGLVPRGSHM$_{21}$ 4XNM_D: Chain D, Hemagglutinin $_{1}$MGSSHHHHHHSSGLVPRGS$_{19}$ 3W95_A: Chain A, Genome polyprotein $_{1}$MGSSHHHHHHSSGLVPRGSHM$_{21}$ 5VRJ_A Chain A, Polymerase acidic protein $_{1}$MGSSHHHHHHSSGLVPRGSHM$_{21}$ 5GSO_E Chain E, 3C protein $_{1}$MGSSHHHHHHSSGLVPRGSHM$_{21}$
5MQV_F: Chain F, Casein kinase I isoform delta $_{1}$MGSSHHHHHHSSGLVPRGSHME$_{22}$	3W95_A: Chain A, Genome polyprotein $_{1}$MGSSHHHHHHSSGLVPRGSHM$_{21}$ 5GSO_E: Chain E, 3C protein $_{1}$MGSSHHHHHHSSGLVPRGSHM$_{21}$ 3W95_A: Chain A, Genome polyprotein $_{1}$MGSSHHHHHHSSGLVPRGSHM$_{21}$ 5VRJ_A: Chain A, Polymerase acidic protein $_{1}$MGSSHHHHHHSSGLVPRGSHME$_{22}$
2PPL_A: Chain A, Pancreatic lipase-related protein 1 $_{4}$HHHHHHDYDIPTTENLYFQG$_{23}$	6EY7_D Chain D, UL89 HCMV $_{3}$HHHHHHDYDIPTTENLYFQG$_{22}$
NP_001062.1: Thymidylate synthase isoform 1 $_{49}$DRTGTGTLSVFGMQARYSLRDEFPLLTTKRVFW$_{81}$; $_{126}$REEGDLGPVYGFQWRHFGAEYR$_{147}$; $_{164}$VIDTIKTNP$_{172}$; $_{174}$DRRIIMCAWNP$_{184}$; $_{186}$DLPLMALPPCH$_{196}$; $_{207}$ELSCQLYQRSGDMGLGVPFNIASYALLTYMIAH$_{239}$; $_{248}$FIHTLGDAHIYLNHIE$_{263}$	ALH45390.1: ORF70 $_{96}$PLLTTKRVFW$_{105;\ 231}$ELSCQLYQRSGDMGLGVPFNIASYALLTYM$_{260}$
P04439.2: HLA class I histocompatibility antigen, A alpha chain $_{2}$AVMAPRTLLLL$_{12}$; $_{150}$LNEDLRSWTA$_{159}$; $_{161}$DMAAQITKRKWEAA$_{174}$; $_{224}$TLRCWALGFYPA$_{235}$; $_{255}$VETRPAGDGTFQKWAA$_{270}$	AMJ54243.1: Membrane glycoprotein UL40 $_{14}$AVMAPRTLLL$_{23}$

portions of the consensus shared sequences are mapped to non-human viral sequences, thus not included in the results analysis.

The human 'HLA class I histocompatibility antigen, A alpha chain (HLA-A) (P04439.2)' protein has 5 consensus shared (Table 4) mapped to several viral protein (Supplementary Material 1). However, here in Table 4 we showed only one consensus shared sequence $_{14}$AVMAPRTLLL$_{23}$ of *Human betaherpesvirus* 5 (HHV-5) 'AMJ54243.1: Membrane glycoprotein UL40' protein that is shared with the human HLA-A. The other viral shared sequences mapped are nonhuman viruses (see Supplementary Material 1), thus not satisfy our selection criteria. In addition, the coverage of each consensus shared sequence is a function of the overlapping extension of the kmer length of the peptide generated between the host and the pathogen protein sequences. Besides, higher or longer *kmers* mapping encourages higher specificity and reduces smaller repertoire of the shareome charting.

In addition to the 2D analysis we carried out 3D analysis which was achieve by modelling protein structure of 2 shared human and viral protein sequences to gain insight localization shared sequences in 3D. This was necessitated because the localisation of shared sequences on the 3D structure using the 33 PDB proteins was not possible since most of the crystal structure had missing residues, thus the need to modelled the 3D structures of the protein. Herein, the 3D visualization showed that some shared sequences on the human Thymidylate synthase isoform 1 (NP_001062.1) are distributed on the surface and exposed (Figure 12-1A and 1B and Figure 11). Few residues were noted to occupied cleft/pocket of the protein, which may also be catalytic site for substrate interaction because the protein is an enzyme. Similarly, the virus HHV-8 ORF70 (ALH45390.1) identified to share some consensus peptide sequences with the human protein have their shared sequences distributed on the surface, exposed and respectively (Figure 12-2A and 2B). HHV-8 ORF70 proteins show few residue surrounding surfaces of the cleft, which may be a binding site for interaction with human host.

Similarly, modelling 3D structure of HLA class I histocompatibility antigen, A (NCBI: sp|P04439.2|HLAA_HUMAN) reveal the shared sequences are prominent at the peptide-binding cleft, covering some identified pocket/groove of the HLA (Figure 13). The distribution of the shared peptide at this region might have immunological relevance in terms of the

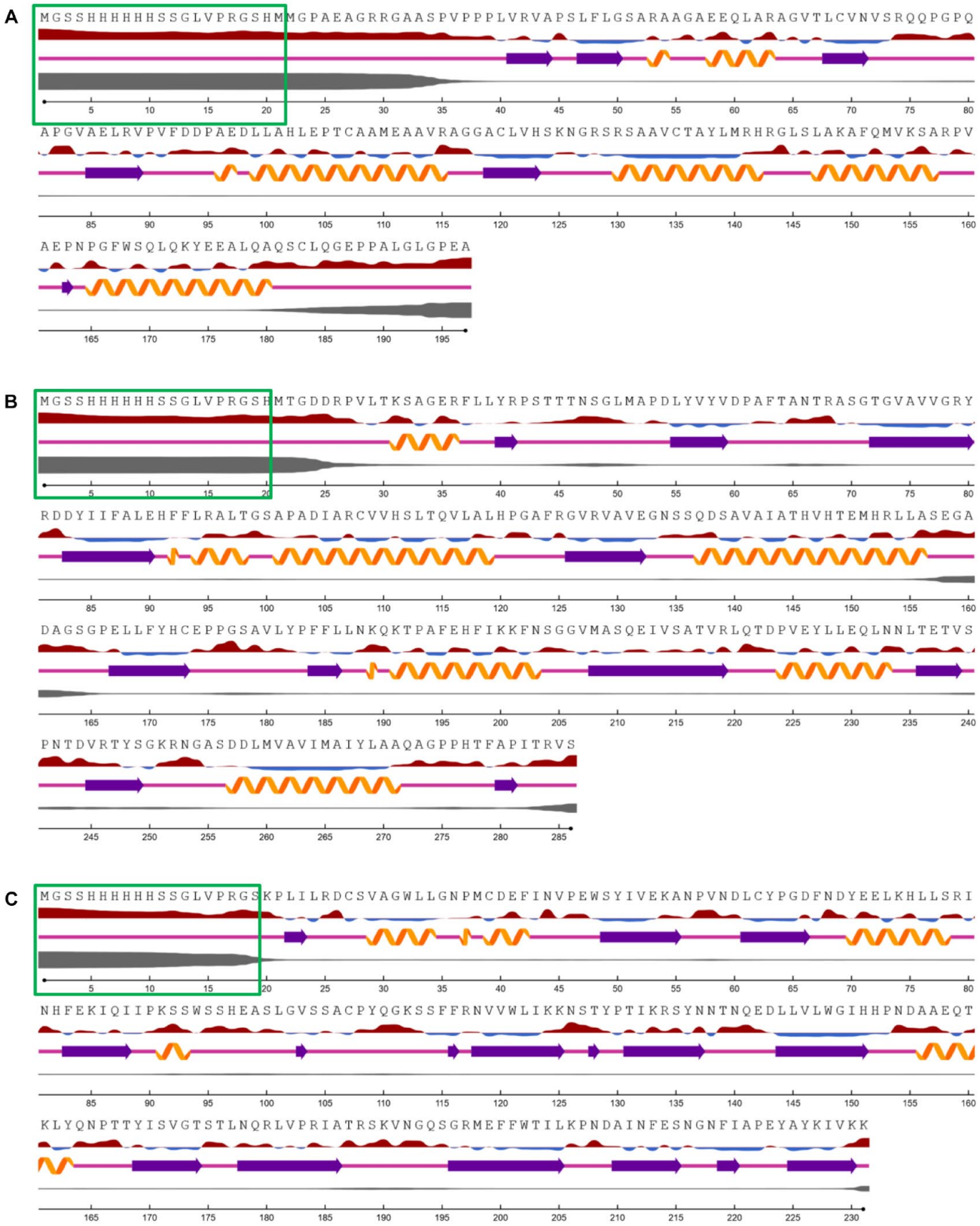

Figure 8. Secondary structure showing relative surface accessibility (RSA) of human and HSV-1 and IAV proteins. Marked area in Green Square shows the shared sequences in the protein. Graphical colours show predictions where ( ) red signifies expose residue and blue represent buried residue. Orange spiral ( ) indicates region in the 2D that are helix, why purple ( ) thick arrow stands for region that are strand, while region of single pink lines ( ) represent coils; in addition, region of disordered residue ( ) is indicated by thick ash lines. (A) Human 'Chain B, Dual specificity phosphatase 28', RSA analysis shows all of the residue are expose. Similarly, (B) *Human alphaherpesvirus 1* (*Herpes simplex virus 1* – HSV-1) 'Chain C, Tripartite terminase subunit UL15', and (C) *Influenza A virus* 'Chain D, Hemagglutinin' shared sequences are well exposed and are region of disordered.

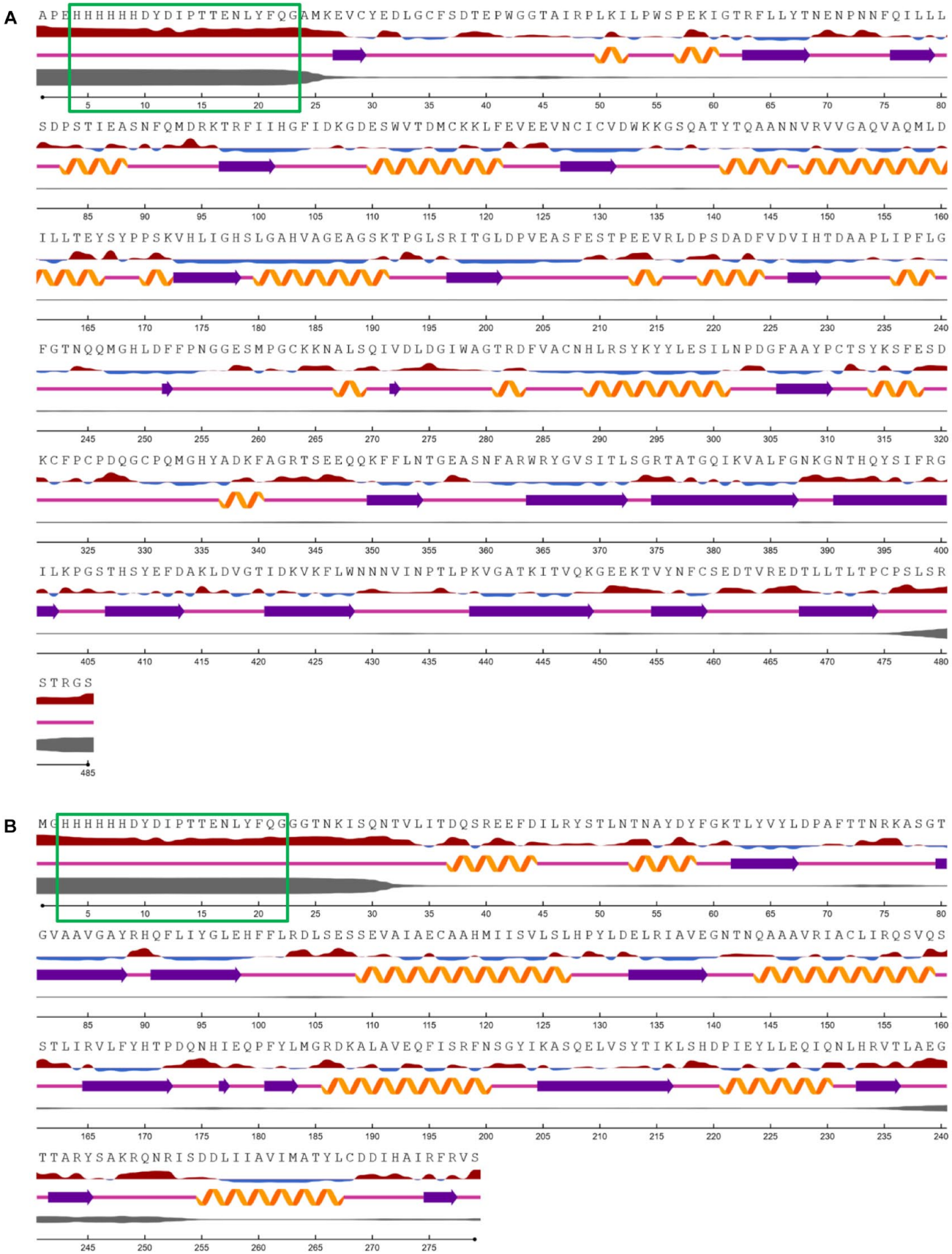

Figure 9. Secondary structure of the human 'Chain A, Pancreatic lipase-related protein 1' (A) and Human herpesvirus 5 strain AD169 'Chain D, UL89 HCMV' protein (B). Marked area in Green Square shows the shared sequences in the protein.

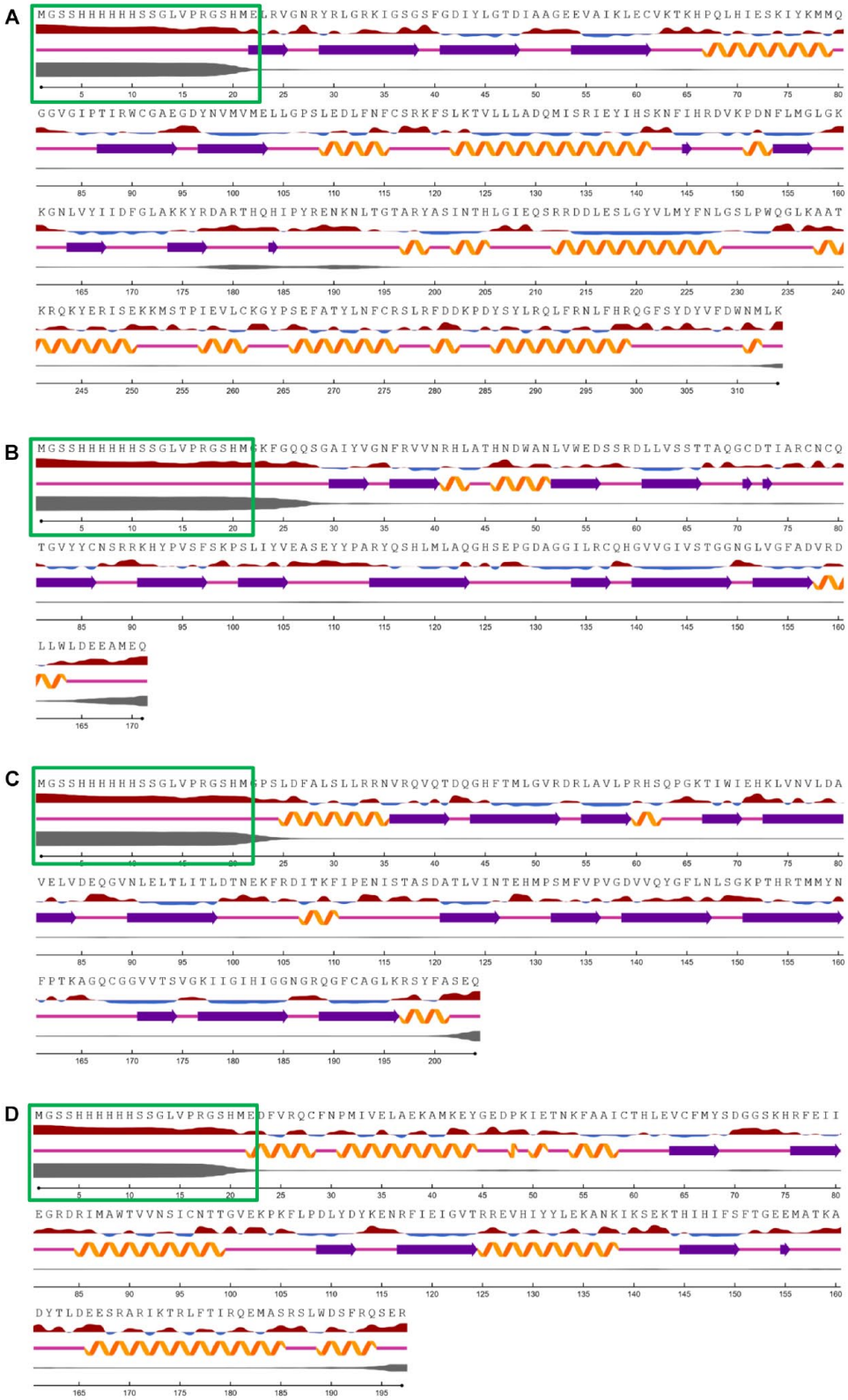

Figure 10. Secondary structures of (A). human Casein kinase I isoform delta (Chain F) protein; (B). *Human enterovirus 71 (strain 7423/MS/87)* Genome polyprotein (Chain A) (also known as 2A proteinase (C110A)) protein; (C). *Enterovirus A71* 3C protein (Chain E) and D). *Influenza A virus* (A/California/04/2009(H1N1)) Polymerase acidic protein (Chain A) (also known as PA Endonuclease). Marked area in Green Square shows the shared sequences are well exposed and lies in the coil region of the protein. In addition, some few residues are found in strand (A) and slightly helix (D) region of the proteins. Graphical colours show predictions where ( ) red signifies expose residue and blue represent buried residue. Orange spiral ( ) indicates region in the 2D that are helix, why purple ( ) thick arrow stands for region that are strand, while region of single pink lines ( ) represent coils; in addition, region of disordered residue ( ) is indicated by thick ash lines.

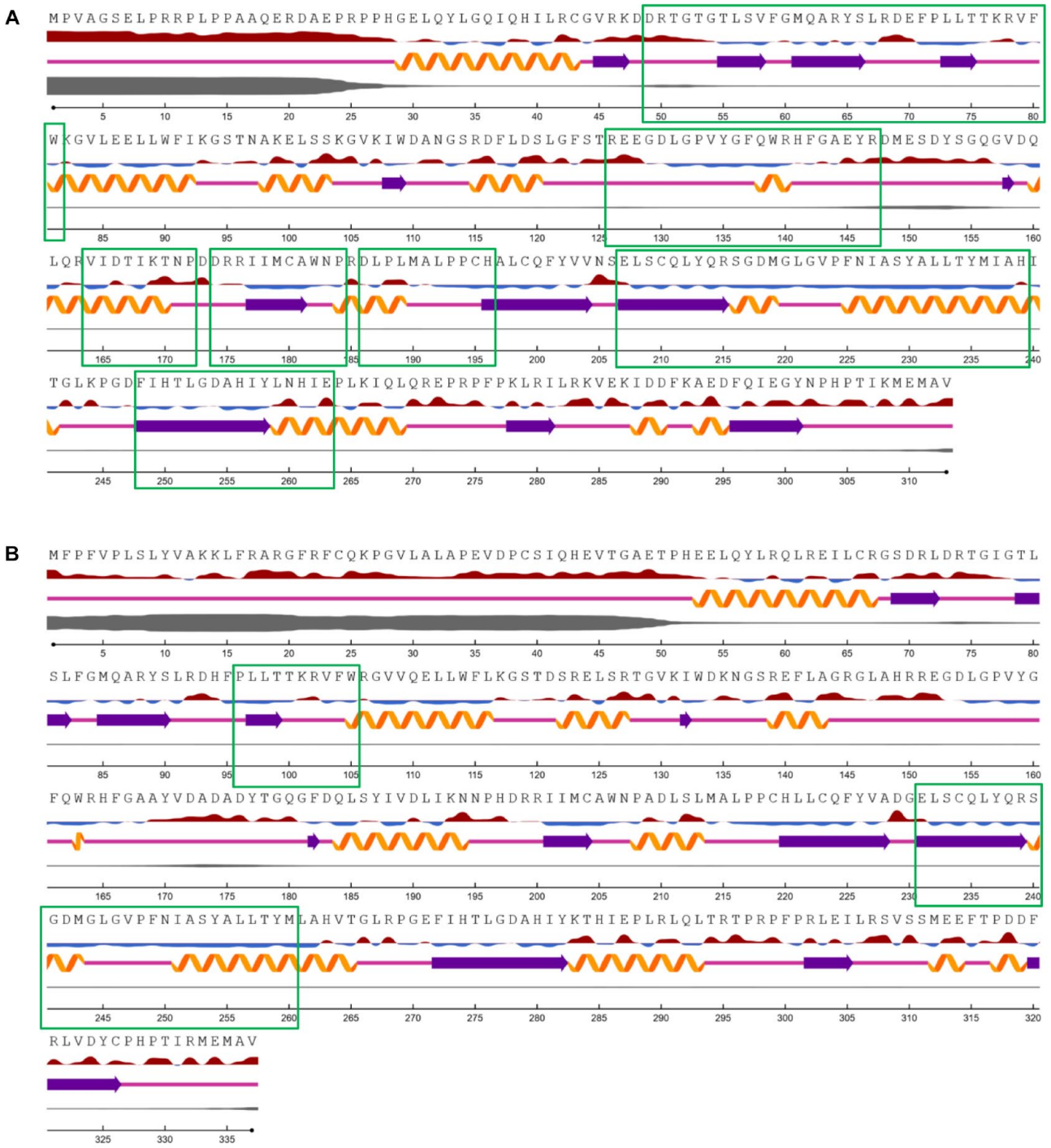

Figure 11. Secondary structure showing relative surface accessibility (RSA) of human Thymidylate synthase isoform 1 and HHV8 ORF70 proteins. Marked area in Green Square shows the shared sequences in the protein. Graphical colours show predictions where ( ) red signifies expose residue and blue represent buried residue. Orange spiral ( ) indicates region in the 2D that are helix, why purple ( ) thick arrow stands for region that are strand, while region of single pink lines ( ) represent coils; in addition, region of disordered residue ( ) is indicated by thick ash lines. A) Human 'Thymidylate synthase isoform 1', RSA analysis shows 61% of the consensus residues are buried, while 31% are expose, their coverage falls within helix, strand and coil regions. B) *Human gammaherpesvirus 8* (HHV-8) 'ORF70' with consensus residues majorly buried and covers both strand, helix and coil regions.

HHV-8 viral infection which equally has its $_{14}$AVMAPRTLLL$_{23}$ sequences in the human HLA class 1 $_{2}$AVMAPRTLLLL$_{12}$ sequences (Table 4). However, the modelling of the HHV-8 membrane glycoprotein UL40 protein had missing residue in the region of the shared sequences; thus, the 3D structure is not presented. Nevertheless, sharing similar sequences between the

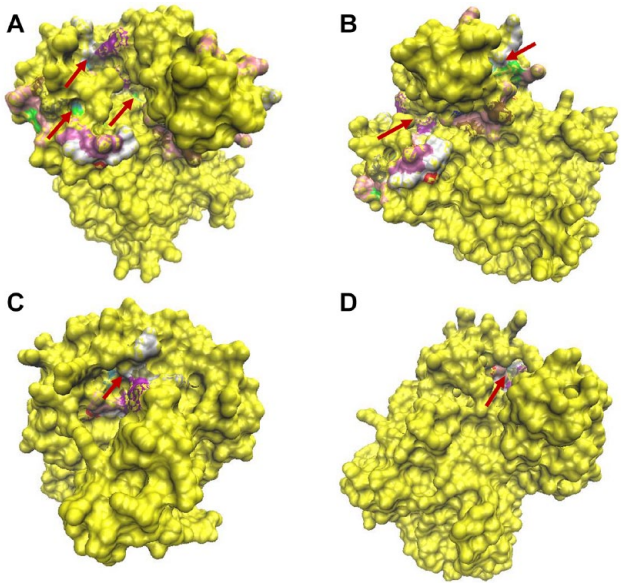

Figure 12. Structure model of human (1) and viral (2) protein involved in shareome charting. (1). Localization of Thymidylate synthase isoform 1 (NCBI: NP_001062.1) shared sequences on the 3D structure. 1A: Human protein x-y-z oriented (showing the shared sequences being exposed at the pocket of the protein (arrows indicate pocket on the modelled protein). 1B: Orientation along y-z-x-axis position displaying more of the shared sequences surrounding the pocket of the human protein. 2). Localization of *Human gammaherpesvirus 8* (HHV-8) ORF70 protein (NCBI: ALH45390.1). 2A and 2B showing different orientation (x-z-y- and x-y-z-axis, respectively) with shared sequences being exposed at the pocket of the viral protein (arrows indicate pocket on the modelled protein). The colours on the 3D represent residue names as described in VMD colour categories.

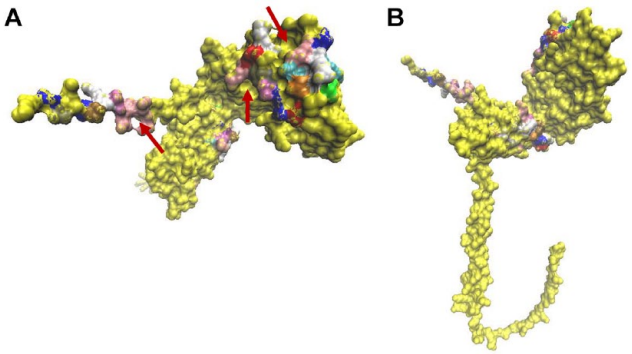

Figure 13. Structural model of HLA class I histocompatibility antigen, A alpha chain (P04439.2) protein involved in shareome charting. 1). Localization of shared sequences on the HLA class 1 3D structure at the peptide-binding cleft. 1A: The HLA class 1 protein y-z-x-axis oriented (showing the shared sequences being exposed at and surrounding the cleft of the protein (arrows indicate pocket/grooves on the modelled protein). 1B: Orientation along z-x-y-axis position displaying more of the shared sequences surrounding the peptide-binding cleft of the HLA class 1.

viruses that is ordinary to be presented by HLA class I molecules can have several implications for the immune response.

Immune epitope studies

T-cell epitopes of viral protein in diabetes viruses-human share-ome were predicted computationally for peptide binding to MHC molecules. Nonapeptides predicted as supertypes-restricted HLA-restricted epitopes using the combined score of 0.8 for at least half of the representative alleles, recognizes peptides presented by HLA involve in multiple overlaps spanning individual supertypes (Table 5). The entire 1064 shared sequences were predicted HLA-A and B supertype-restricted epitopes. Only 30 A1 Supertype epitopes were returned as shared nonapeptide predicted, to be binders of A1 supertype-restricted representative alleles: HLA-A*26:01, HLA-A*26:02, HLA-A*26:03, HLA-A*30:02, HLA-A*30:03, and HLA-A*32:01 (Table 5). Similarly, other predicted A and B supertype-restricted representatives are shown in Supplementary table 5. Also, predicted are HLA supertype-restricted epitopes with promiscuity against many alleles and supertypes (Supplementary Table 5). Validation of the predicted epitope base on comparison to all reported human T-cell epitopes of viral sequences available in the IEDB resource (Table 6; Supplementary Table 6), showed that most of the epitope predicted originate from *EV B*. Other viral epitope predicted come from *EV C*, *HBV 6A*, *HBV 6B*, *Human endogenous retrovirus K*, *Human herpesvirus 4*—(EBV), *IAV*, *Rotavirus A*, and *Rubella virus*. Specific viral epitopes that were validated to be restricted to HLA include: *IAV* NP, HA and matrix protein 1; *Rotavirus A* Outer capsid glycoprotein VP7; *Rubella virus* "Structural polyprotein"; *HBV 6B* "Protein U2"; and *Human herpesvirus 4*—(EBV) "Latent membrane protein 2" and "mRNA export factor ICP27 homolog", among others (Supplementary Table 6). However, records of the *EV B* and *EV C* returned did not show an explicit protein of the hits predicted, but show that the whole genome polyprotein are involved.

Discussion

In this study, we charted, analysed the structures and immune relevance of protein shared sequences between viruses reported in DM onset and human pancreatic proteins. The distribution analysis of the 'diabetes viruses'-human shared sequences, reveals that nonapeptide present in the *Herpesviruses*, *Enterovirus*, *Human endogenous retrovirus*, *IAVs*, *Rotavirus* and *Rubivirus* are equally hosted by the human pancreatic proteins. The identified sets of shared sequences obtained from the initial total removal of duplicated sequences at 100% could have led to loss of additional details on the natural divergence within the sequences. Even the filtering of the nonamers we observed greater reduction of redundant nonapeptides for both the viral and human nonapeptides (Table 1); this can only implied that despite distinct variations that may be present in viral species of viruses, there exist limited changes in the amino acid residues of their protein sequences. When considering the protein shared nonamers function between 'diabetes-viruses' and

Table 5. Number of shareome nonapeptides predicted as supertypes-restricted HLA-restricted epitopes.

HLA SUPERTYPE	REPRESENTATIVE ALLELES	NO. OF PUTATIVE HLA SUPERTYPE-RESTRICTED EPITOPES[A,B]
A1	HLA-A*26:01, HLA-A*26:02, HLA-A*26:03, HLA-A*30:02, HLA-A*30:03, HLA-A*32:01	30(18)
A01-A03	HLA-A*30:01	8
A01-A24	HLA-A*29:02	6
A2	HLA-A*02:01, HLA-A*02:02, HLA-A*02:03, HLA-A*02:04, HLA-A*02:05, HLA-A*02:06, HLA-A*02:14, HLA-A*68:02, HLA-A*69:01	51 (21)
A24	HLA-A*23:01, HLA-A*24:02	9 (5)
A3	HLA-A*03:01, HLA-A*11:01, HLA-A*31:01, HLA-A*33:01, HLA-A*33:03, HLA-A*68:01, HLA-A*03:13, HLA-A*03:14	24 (13)
B7	NA	NA
B8	NA	NA
B27	HLA-B*15:03, HLA-B*15:18, HLA-B*27:05, HLA-B*27:06, HLA-B*27:07, HLA-B*27:09, HLA-B*39:02	24 (13)
B44	HLA-B*18:01, HLA-B*40:01, HLA-B*40:02, HLA-B*40:06, HLA-B*44:02, HLA-B*44:03	9 (4)
B58	HLA-B*15:16, HLA-B*15:17, HLA-B*58:01	21 (11)
B62	HLA-B*15:01, HLA-B*15:02, HLA-B*15:12, HLA-B*15:13	21 (8)

[a]Only included the predictions that satisfied the combined score of 0.8 for at least half of the representative alleles.
[b]A total (in nonparenthesis with duplicate) and unique numbers (in parenthesis) of HLA supertype-restricted epitopes predicted including promiscuity against many alleles and supertypes.

human sequences, it become necessary to consider the significance of specific interacting residues. Since related commonality of sequences can present similar cell functions in terms of signalling transduction, growth, transport, binding, enzymes activities and immune regulations. Thus, proteins and specific shared nonapeptides are ways for viruses to interact with and enter the human-host cells or the host immune system.

The fundamental idea of deciphering the structural function of viral proteins is to shed light on the nature of the viral architect, which involves protein-protein interaction and is a prime target for therapeutic research. According to analysis, most of the shared nonapeptide sequences were exposed, while the remaining ones had some of their incomplete residues exposed. Thus, it can be inferred that the shared sequences are typically buried or exposed and solvent-accessible (Figures 8 to 11). The exposed residues (M, R, N, D, C, S, Q, K, E, K, T, G, P, L, A, and H) interact with the molecules in the immediate vicinity and are frequently hydrophilic in nature. These residues may also be engaged in important metabolic pathways. The exposed shared residues in groups may serve as binding sites for the viral protein. Furthermore, secondary structure analysis demonstrated that the common sequences were not limited to any of the particular structural units, removing rigidity in localization. Despite the vast distribution of expose residue in the shared sequence region of the protein, few shared sequences were buried, suggesting stability of the protein structure as contributed by their unaccessible residues' nature, which often face stricter evolutionary constraints than accessible residues.[59] When amino residues are buried in polypeptide chain they are frequently hydrophobic a driving force in protein folding, 3D stability and functions,[60] These buried residues often contribute to the formation of hydrophobic core of a protein.[61] Here, we identify A, C, D, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y residues as buried that may contribute as molecular chaperones during folding, preventing aggregation and misfolding. It is well established that where protein binding or catalytic sites are at found in protein 3D structure, buried residue can be identified is such environment.[62] This implied that the shared sequences in the 'diabetes viruses'-human shareome mapped could be crucial in facilitating substrate binding and catalysis. Sharing these similar sequences may promote cross-reactivity during viral infections due host immune (T-cells) surveillance targeting viral antigens that are shade with human proteins or due to nonspecific immune activation. In some cases, these activities may provide cross-protection against different viruses, or results to detrimental autoimmune reactions of vital cells and tissues.[63]

Table 6. Reported human T-cell epitopes/HLA ligands of Enteroviruses, Herpesvirus, Influenza A virus, Rotavirus and Rubella virus that matched the predicted epitopes.

DETAILS	EPITOPE	ANTIGEN	ORGANISM
11081	EAIPALTAVETGHTSQV	Genome polyprotein	Enterovirus B
18615	GAHETSLSASGNSIIHYTNI	Genome polyprotein	Enterovirus B
21518	GNSIIHYTNINYYKDAASNS	Genome polyprotein	Enterovirus B
31330	KILPEVKEKHEF	Genome polyprotein	Enterovirus B
41554	MGAQVSTQKTGAHETSLSAS	Genome polyprotein	Enterovirus B
46733	NYYKDAASNSANRQDFTQDPSKFTEPVKDV	Genome polyprotein	Enterovirus B
58829	SKFTEPVKDVMIKSLPALN	Genome polyprotein	Enterovirus B
117846	EWLKVKILPEVK	Genome polyprotein	Enterovirus B
118222	WLKVKILPEVKEKHEFLNRL	Genome polyprotein	Enterovirus B
118883	LKVKILPEVKEKHEFLSRLKQLP	Genome polyprotein	Enterovirus B
120682	KILPEVKEKHEFLSRL	Genome polyprotein	Enterovirus B
135900	ATCRFYTLDSIK	Genome polyprotein	Enterovirus B
135908	DINTVTTVAQSR	Genome polyprotein	Enterovirus B
135911	DRVRSITLGNST	Genome polyprotein	Enterovirus B
135920	ETSLSAAGNSII	Genome polyprotein	Enterovirus B
135951	LLESQIATIEQT	Genome polyprotein	Enterovirus B
135958	LTAVETGHTSQV	Genome polyprotein	Enterovirus B
135974	PSNSASVPALTA	Genome polyprotein	Enterovirus B
135976	PVKDVMIKSLPAPALNSPTVEEC	Genome polyprotein	Enterovirus B
135986	RSGPSNSASVPA	Genome polyprotein	Enterovirus B
135987	RSITLGNSTITT	Genome polyprotein	Enterovirus B
135990	SASVPALTAVET	Genome polyprotein	Enterovirus B
135997	SLEKKMSNYIQF	Genome polyprotein	Enterovirus B
135999	SQIATIEQTAPS	Genome polyprotein	Enterovirus B
136002	TKQMVQMRRKLE	Genome polyprotein	Enterovirus B
136003	TRKDINTVTTVA	Genome polyprotein	Enterovirus B
136006	VETGHTSQVTPS	Genome polyprotein	Enterovirus B
136010	VMIKSLPALNSPTVEEC	Genome polyprotein	Enterovirus B
136011	VPALTAVETGHT	Genome polyprotein	Enterovirus B
226838	FKPKHVKAYVRP	Genome polyprotein	Enterovirus B
549266	WLKNKLIPEVKEKHEFLSRL	Genome polyprotein	Enterovirus B
30661	KEVPALTAVETGAT	Genome polyprotein	Enterovirus C
129061	KELLQSYVSKNNN	Uncharacterized protein U95	Human betaherpesvirus 6A
122774	GGVAVVIGRFFG	Protein U2	Human betaherpesvirus 6B
118381	GKTCPKEIPKGSKNT	putative env	Human endogenous retrovirus K
6568	CLGGLLTMV	Latent membrane protein 2	Human herpesvirus 4 (Epstein-Barr virus)

(Continued)

Table 6. (Continued).

DETAILS	EPITOPE	ANTIGEN	ORGANISM
20788	GLCTLVAML	mRNA export factor ICP27 homolog	Human herpesvirus 4 (Epstein-Barr virus)
20354	GILGFVFTL	Matrix protein 1	Influenza A virus
13263	ELRSRYWAI	Nucleoprotein	Influenza A virus
7136	CTELKLSDY	Nucleoprotein	Influenza A virus
27283	ILRGSVAHK	Nucleoprotein	Influenza A virus
48237	PKYVKQNTLKLAT	Hemagglutinin	Influenza A virus
69270	VKLYRKLKREITFHGAKEIS	Matrix protein 1	Influenza A virus
133607	IIVILSPLLNAQN	Outer capsid glycoprotein VP7	Rotavirus A
133729	VILLNYVLKSLTR	Outer capsid glycoprotein VP7	Rotavirus A
79506	AFGHSDAACWGFPTDTVMSV	Structural polyprotein	Rubella virus
79628	PTDVSCEGLGAWVPTAPCARI	Structural polyprotein	Rubella virus
119822	EACVTSWLWSEGEGAVFYRVDLHFINLGT	Structural polyprotein	Rubella virus
120093	MDFWCVEHDRPPPATPTSLTT	Structural polyprotein	Rubella virus
120127	PFLGHDGHHGGTLRVGQHHRNASDV	Structural polyprotein	Rubella virus
120187	RVKFHTETRTVWQLSVAGVSC	Structural polyprotein	Rubella virus

Interestingly, considering structural and functional relatedness between viral and human shared sequences in this study, HHV8 ORF70 proteins like that of human-host thymidylate synthase are involved in nucleotide synthesis. The ORF70 encodes a viral thymidylate synthase, an enzyme essential for nucleotide synthesis, which is crucial for viral DNA replication. This enzyme helps the virus synthesize the building blocks of DNA, supporting its replication and proliferation within the host.[64] In a close related virus *Equid alphaherpesvirus 3* (formerly known: *Equid herpesvirus type 3* (EHV-3)), ORF70 differently encodes glycoprotein G (gG), which is involved in modulating the host immune response by interacting with chemokines, thereby helping the virus evade the immune system. While in human the protein 'thymidylate synthase isoform 1' play a key role in the initial steps of thymidylate (dTMP) biosynthesis essential DNA synthesis, repairs and cell division.[65] The sharing of similar nonapeptides sequences (Table 4 and Figure 11) between HHV-8 ORF70 and human thymidylate synthase isoform 1 proteins can have various implications, including the potential for autoimmune reactions or broader immune dysfunction that may result to cellular function alteration or biochemical derangement.

Likewise, the 3D crystal structure via modelling provides detailed insights into molecular architecture and possible sites for interactions with other molecules, the region of shared sequences that can specifically bind to and modulate as targets regions. Here, the structure in Figures 12 and 13 depict this ascension. Having residues on the surface of the grooves/clefts especially on the human thymidylate synthase and HLA Class 1 of the shared sequences have implications. Presence of similar peptide sequences of viruses on the host protein have long been establishes to results cross-reactivity, immune escape and autoimmunity.[66] Consequently, these sharing of peptide sequences between the viral sequences to human pancreatic proteins may contribute to the destruction of the cells associate to glucose metabolism and its regulations. Thus, identifying and understanding the sequences that HLA class 1 molecules can present is crucial for the designing effective vaccine.

Nearly one-tenth of the shared nonapeptides showed immune relevance (Table 5), these include 111 HLA class I supertype-restricted epitopes predicted and a reasonable number (51) matched to reported T-cell epitopes/ligands, giving the predictions a degree of credibility. However, these putative epitopes would likely apply to a significant section of the human population due to promiscuity across various supertypes. This indicates that there may be a cross-ethnic risk of an autoimmune reaction. Besides, the charted epitopes give us a list of potential epitopes that might initiate cross-reactivity in the human-host pancreas leading to Type I diabetes due to mimicking of sequences by the viruses. This assumption is on

the basis of molecular mimicry theory which stated that an immune reaction against an infectious agent may result in the development of cross-reacting antibodies that bind the shared epitopes on a normal cell and cause the cell autodestruction of the cell when viral agents share epitopes with a host's proteins.[22] In addition, they could also be potential epitopes that might be excluded from virus-specific candidate vaccine formulations, thereby preventing autoimmunity.

Human gammaherpesvirus 4—(EBV), among our defined 'diabetes viruses' in this study and other known human viruses happen to be ubiquitous in the human population.[67] Here, we identified EBV nonapetides that are repeats (such as GRGRGRGRG and GAGGAGGAG) which covers a subset or full-length of the shared sequences. Their role may be associated with molecular recognition and molecular assembly as it is common to amino acids repeats.[68] For instance, the EBV protein 'Epstein-Barr virus (EBV) nuclear antigen 2 (EBNA-2)' involved in the protein shared sequences, plays a crucial role in facilitating the B-cell proliferation in EBV infection,[69] thus promoting B-cell transformation. This protein has nontandem AAR GRGRGRGRG shared with human protein 'Histone-lysine N-methyltransferase 2B'. This human protein has been associated with the pancreatic immune-pathophysiology,[70] development and regulation of the pancreas metabolic activities relating to carbohydrate.[71] Thus, the shared sequences or the AARs can provide a platform for molecular mimicry leading to human hyperimmune responses to *Human gammaherpesvirus 4*—(EBV) followed by autoimmune cross-reactions. These actions can promote the onset of type I DM in the human host.

Another herpes virus, *Human betaherpesvirus 5*—HHV-5 (*HCMV*) is known for its wide spread among the human populace with lifetime latent infection and can occasionally reactivate[72] and capable of evading the natural killer cells by encoding 'decoy' molecules that imitate MHC molecule.[73] Their presence in this shareome overlap (Figure 7) especially with its 'Chain D, UL89 HCMV' known for assisting viral DNA cleavage and packaging.[74] The amino acids residues are well exposed and their protein shared sequences are mapped to that of the human protein 'Chain A Pancreatic lipase-related protein 1' (Table 4). Since HHV-5 is capable of periodical reactivation in the host, its presence in the pancreas may initiate cross-reactivity due to the shared sequences they have covering even more residues beyond nonamers (Table 4). These viral proteins can further be explored as a promising target for vaccine development and drug target.

In our shareome analysis the virus with major T-cell epitopes/HLA ligand predicted is the *EVs* (Table 6) from the *Picornaviridae* family. Their protein shared sequences are also well exposed and its sequences are mapped to various human proteins such as 'Chain F, Casein kinase I isoform delta' among others (Tables 3 and 4; Figure 10). Both shared sequence residues of the viral (*Human enterovirus 71 (strain 7423/MS/87)* '2A proteinase (C110A)') and the human proteins are expose as shown in Figure 10, this may signify that the viral protein being an enzyme can modify other human protein, in such process leading to cross-reactivity in the pancreatic cells. Moreover, *EVs* have been shown to have a strong pancreotropism; severe islet damage has been demonstrated in fatal CVB infection cases,[75] human islets show strong expression of the coxsackie virus receptor (CAR)[76] and beta cells are permissive for EV in vitro.[77] It has recently been shown that the inflammatory state of the pancreas can be explained by direct or indirect viral effects.[78,79]

Conclusion

This study identified shared nonapeptides and the proteins encoding between 'diabetes-viruses' and human pancreatic proteins characterizing them in terms of their structure and immunological relevance. The 'diabetes-viruses' charted to the human nonapeptide matches indicated possible cross-reactivity with the pancreatic tissues that may result to destruction of the β-cells, due to potential interactions between viral and human nonapeptide that could compete. Data from this presence study can be used to design antiviral immunotherapeutic approaches targeting onset of type I DM thereby inhibiting the cross-reactivity with human antigens.

Acknowledgements

The authors appreciate the Management of Perdana University for the various resources provided to SAJ to achieve this research after his PhD studies especially during the COVID-19 lockdown.

Author contributions

SAJ designed and carried out the study. SAJ and IAJ draft the manuscript and contributed to its further improvement. Both authors read and approved the final manuscript.

ORCID iD

Stephen A. James https://orcid.org/0000-0003-4788-9778

SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

REFERENCES

1. Baynest HW. Classification, pathophysiology, diagnosis and management of diabetes mellitus. *J Diabetes Metab*. 2015;6. doi:10.4172/2155-6156.1000541
2. Gosmanov AR, Gosmanova EO, Kitabchi AE. Hyperglycemic crises: Diabetic Ketoacidosis (DKA) and Hyperglycemic Hyperosmolar State (HHS). In: Feingold, KR, Anawalt, B, Boyce, A, et al., ed. *Acute Endocrinology*. Humana Press; 2018:119-147. doi:10.1007/978-1-60327-177-6_6
3. Kitabchi AE, Umpierrez GE, Miles JM, Fisher JN. Hyperglycemic crises in adult patients with diabetes. *Diabetes Care*. 2009;32:1335-1343. doi:10.2337/dc09-9032
4. Hu F, Qiu X, Bu S. Pancreatic islet dysfunction in type 2 diabetes mellitus. *Arch Physiol Biochem*. 2020;126:235-241. doi:10.1080/13813455.2018.1510967

5. Misra S, Oliver NS. Diabetic ketoacidosis in adults. *BMJ*. 2015;351:h5660. doi:10.1136/bmj.h5660
6. Kim H, Kim W, Choi JE, Kim C, Sohn J. Short-term effect of ambient air pollution on emergency department visits for diabetic coma in Seoul, Korea. *J Prev Med Public Health*. 2018;51:265-274. doi:10.3961/jpmph.18.153
7. Alicic RZ, Rooney MT, Tuttle KR. Diabetic kidney disease: challenges, progress, and possibilities. *Clin J Am Soc Nephrol*. 2017;12:2032-2045. doi:10.2215/CJN.11491116
8. Kelly KJ, Dominguez JH. Rapid progression of diabetic nephropathy is linked to inflammation and episodes of acute renal failure. *Am J Nephrol*. 2010;32:469-475. doi:10.1159/000320749
9. Ramdharry G. Peripheral nerve disease. *Handb Clin Neurol*. 2018;159:403-415. doi:10.1016/B978-0-444-63916-5.00026-4
10. Prakash G, Agrawal R, Natung T. Role of lipids in retinal vascular and macular disorders. *Indian J Clin Biochem*. 2017;32:3-8. doi:10.1007/s12291-016-0560-2
11. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care*. 2009;31. doi:10.2337/dc09-S062
12. Antosik K, Borowiec M. Genetic factors of diabetes. *Arch Immunol Ther Exp (Warsz)*. 2016;64:157-160. doi:10.1007/s00005-016-0432-8
13. Mannering SI, Pathiraja V, Kay TW. The case for an autoimmune aetiology of type 1 diabetes. *Clin Exp Immunol*. 2016;183:8-15. doi:10.1111/cei.12699
14. de Candia P, Prattichizzo F, Garavelli S, et al. Type 2 diabetes: how much of an autoimmune disease. *Front Endocrinol (Lausanne)*. 2019;10:451. doi:10.3389/fendo.2019.00451
15. Krzewska A, Ben-Skowronek I. Effect of associated autoimmune diseases on type 1 diabetes mellitus incidence and metabolic control in children and adolescents. *Biomed Res Int*. 2016;2016:6219730-6219712. doi:10.1155/2016/6219730
16. Ferraù F, Albani A, Ciresi A, Giordano C, Cannavò S. Diabetes secondary to acromegaly: physiopathology, clinical features and effects of treatment. *Front Endocrinol (Lausanne)*. 2018;9:358. doi:10.3389/fendo.2018.00358
17. Bauerle KT, Harris C. Glucocorticoids and diabetes. *Mo Med*. 2016;113:378-383. doi:10.1007/s11428-016-0070-0
18. Barbot M, Ceccato F, Scaroni C. Diabetes mellitus secondary to cushing's disease. *Front Endocrinol (Lausanne)*. 2018;9:284. doi:10.3389/fendo.2018.00284
19. Judkowski VA, Allicotti GM, Sarvetnick N, Pinilla C. Peptides from common viral and bacterial pathogens can efficiently activate diabetogenic T-cells. *Diabetes*. 2004;53:2301-2309. doi:10.2337/diabetes.53.9.2301
20. Filippi CM, von Herrath MG. Viral trigger for type 1 diabetes: pros and cons. *Diabetes*. 2008;57:2863-2871. doi:10.2337/db07-1023
21. Coppieters KT, Boettler T, von Herrath M. Virus infections in type 1 diabetes. *Cold Spring Harb Perspect Med*. 2012;2:a007682. doi:10.1101/cshperspect.a007682
22. Kanduc D, Stufano A, Lucchese G, Kusalik A. Massive peptide sharing between viral and human proteomes. *Peptides*. 2008;29:1755-1766. doi:10.1016/j.peptides.2008.05.022
23. Principi N, Berioli MG, Bianchini S, Esposito S. Type 1 diabetes and viral infections: what is the relationship? *J Clin Virol*. 2017;96:26-31. doi:10.1016/j.jcv.2017.09.003
24. Raffel LJ, Noble JA, Rotter JI. HLA on chromosome 6: the story gets longer and longer. *Diabetes*. 2008;57:527-528. doi:10.2337/db07-1756
25. Lie BA, Todd JA, Pociot F, et al. The predisposition to type 1 diabetes linked to the human leukocyte antigen complex includes at least one non–class II gene. *Am J Hum Genet*. 1999;64:793-800. doi:10.1086/302283
26. Noble JA, Valdes AM. Genetics of the HLA region in the prediction of type 1 diabetes. *Curr Diab Rep*. 2011;11:533-542. doi:10.1007/s11892-011-0223-x
27. Hober D, Alidjinou EK. Enteroviral pathogenesis of type 1 diabetes: queries and answers. *Curr Opin Infect Dis*. 2013;26:263-269. doi:10.1097/QCO.0b013e3283608300
28. Hober D, Sane F. Enteroviral pathogenesis of type 1 diabetes. *Discov Med*. 2010;10:151-160.
29. Dorman JS, Bunker CH. HLA-DQ Locus of the human leukocyte antigen complex and type 1 diabetes mellitus: a HuGE review. *Epidemiol Rev*. 2000;22:218-227. doi:10.1093/oxfordjournals.epirev.a018034
30. Op de Beeck A, Eizirik DL. Viral infections in type 1 diabetes mellitus – why the β cells? *Nat Rev Endocrinol*. 2016;12:263-273. doi:10.1038/nrendo.2016.30
31. Smatti MK, Cyprian FS, Nasrallah GK, Al Thani AA, Almishal RO, Yassine HM. Viruses and autoimmunity: a review on the potential interaction and molecular mechanisms. *Viruses*. 2019;11:762. doi:10.3390/v11080762
32. Coppieters KT, von Herrath MG. Viruses and cytotoxic T lymphocytes in type 1 diabetes. *Clin Rev Allergy Immunol*. 2011;41:169-178. doi:10.1007/s12016-010-8220-4
33. Rodriguez-Calvo T. Enterovirus infection and type 1 diabetes: unraveling the crime scene. *Clin Exp Immunol*. 2019;195:15-24. doi:10.1111/cei.13223
34. Roer BO, Hiemstra HS, Schloot NC, et al. Molecular mimicry in type 1 diabetes. *Ann N Y Acad Sci*. 2006;958:163-165. doi:10.1111/j.1749-6632.2002.tb02961.x
35. Pugliese A. Autoreactive T cells in type 1 diabetes. *J Clin Invest*. 2017;127:2881-2891. doi:10.1172/JCI94549
36. Borràs E, Martin R, Judkowski V, et al. Findings on T cell specificity revealed by synthetic combinatorial libraries. *J Immunol Methods*. 2002;267:79-97. doi:10.1016/S0022-1759(02)00142-4
37. Kanduc D. The comparative biochemistry of viruses and humans: an evolutionary path towards autoimmunity. *Biol Chem*. 2019;400:629-638. doi:10.1515/hsz-2018-0271
38. Hyöty H, Taylor KW. The role of viruses in human diabetes. *Diabetologia*. 2002;45:1353-1361. doi:10.1007/s00125-002-0852-3
39. Schloot NC, Willemen SJM, Duinkerken G, Drijfhout JW, de Vries RR, Roep BO. Molecular mimicry in type 1 diabetes mellitus revisited: T-cell clones to GAD65 peptides with sequence homology to Coxsackie or proinsulin peptides do not crossreact with homologous counterpart. *Hum Immunol*. 2001;62:299-309. doi:10.1016/s0198-8859(01)00223-3
40. Härkönen T, Lankinen H, Davydova B, Hovi T, Roivainen M. Enterovirus infection can induce immune responses that cross-react with beta-cell autoantigen tyrosine phosphatase IA-2/IAR. *J Med Virol*. 2002;66:340-350. doi:10.1002/jmv.2151
41. Panagiotopoulos C, Trudeau JD, Tan R. T-cell epitopes in type 1 diabetes. *Curr Diab Rep*. 2004;4:87-94. doi:10.1007/s11892-004-0062-0
42. Roep BO, Peakman M. Antigen targets of type 1 diabetes autoimmunity. *Cold Spring Harb Perspect Med*. 2012;2:a007781. doi:10.1101/cshperspect.a007781
43. Kanduc D, Shoenfeld Y. From HBV to HPV: designing vaccines for extensive and intensive vaccination campaigns worldwide. *Autoimmun Rev*. 2016;15:1054-1061. doi:10.1016/j.autrev.2016.07.030
44. Stauss HJ. Peptides feeling groovy. *Curr Biol*. 1991;1:328-330. doi:10.1016/0960-9822(91)90102-3
45. Parham P. Oh to be twenty seven again. *Nature*. 1991;351:523. doi:10.1038/351523a0
46. You ZH, Chan KC, Hu P. Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. *PLoS ONE*. 2015;10:e0125811. doi:10.1371/journal.pone.0125811
47. Schulze S, Schleicher J, Guthke R, Linde J. How to predict molecular interactions between species. *Front Microbiol*. 2016;7:442-413. doi:10.3389/fmicb.2016.00442
48. Coordinators NR, Acland A, Agarwala R, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2013;41:D8-D20. doi:10.1093/nar/gks1189
49. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28(23):3150-3152. doi:10.1093/bioinformatics/bts565
50. James SA, Ong HS, Hari R, Khan AM. A systematic bioinformatics approach for large-scale identification and characterization of host-pathogen shared sequences. *BMC Genomics*. 2021;22:700. doi:10.1186/s12864-021-07657-4
51. Nguyen MN, Madhusudhan MS. Biological insights from topology independent comparison of protein 3D structures. *Nucleic Acids Res*. 2011;39. doi:10.1093/nar/gkr348
52. Sharma A, Manolakos ES. Multi-criteria protein structure comparison and structural similarities analysis using pyMCPSC. *PLoS ONE*. 2018;13:e0204587. doi:10.1371/journal.pone.0204587
53. Marchler-Bauer A, Bo Y, Han L, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res*. 2017;45:D200-D203. doi:10.1093/nar/gkw1129
54. Vita R, Mahajan S, Overton JA, et al. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res*. 2019;47:D339-D343. doi:10.1093/nar/gky1006
55. Stranzl T, Larsen MV, Lundegaard C, Nielsen M. NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics*. 2010;62:357-368. doi:10.1007/s00251-010-0441-4
56. Barrientos-Somarribas M, Messina DN, Pou C, et al. Discovering viral genomes in human metagenomic data by predicting unknown protein families. *Sci Rep*. 2018;8:28. doi:10.1038/s41598-017-18341-7
57. Sehrawat S, Kumar D, Rouse BT. Herpesviruses: harmonious pathogens but relevant cofactors in other diseases. *Front Cell Infect Microbiol*. 2018;8:177. doi:10.3389/fcimb.2018.00177
58. Kumar AS, Sowpati DT, Mishra RK. Single amino acid repeats in the proteome world: structural, functional, and evolutionary insights. *PLoS ONE*. 2016;11:1-19. doi:10.1371/journal.pone.0166854
59. Lin YS, Hsu WL, Hwang JK, Li WH. Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Mol Biol Evol*. 2007;24:1005-1011. doi:10.1093/molbev/msm019
60. van Dijk E, Hoogeveen A, Abeln S. The hydrophobic temperature dependence of amino acids directly calculated from protein structures. *PLoS Comput Biol*. 2015;11:e1004277. doi:10.1371/journal.pcbi.1004277
61. Banach M, Fabian P, Stapor K, Konieczny L, Roterman AI. Structure of the hydrophobic core determines the 3D protein structure-verification by single mutation proteins. *Biomolecules*. 2020;10. doi:10.3390/biom10050767
62. Ribeiro AJM, Tyzack JD, Holliday GL, Borkakoti N, Thornton JM. A global analysis of function and conservation of catalytic residues in enzymes. *J Biol Chem*. 2020;295:314-324. doi:10.1074/jbc.REV119.006289

63. Rouse BT, Sehrawat S. Immunity and immunopathology to viruses: what decides the outcome. *Nat Rev Immunol*. 2010;10:514-526. doi:10.1038/nri2802
64. Majerciak V, Alvarado-Hernandez B, Lobanov A, Cam M, Zheng Z-M. Genome-wide regulation of KSHV RNA splicing by viral RNA-binding protein ORF57. *PLoS Pathog*. 2022;18:e1010311. doi:10.1371/journal.ppat.1010311
65. Hu Frisk J, Pejler G, Eriksson S, Wang L. Structural and functional analysis of human thymidylate kinase isoforms. *Nucleosides Nucleotides Nucleic Acids*. 2022;41:321-332. doi:10.1080/15257770.2021.2023748
66. Martins YC, Jurberg AD, Daniel-Ribeiro CT. Visiting molecular mimicry once more: pathogenicity, virulence, and autoimmunity. *Microorganisms*. 2023;11. doi:10.3390/microorganisms11061472
67. CDC. Epstein-Barr and Infectious Mononucleosis (Mono). Published 2023. Accessed April 3, 2023. https://www.cdc.gov/epstein-barr/index.html
68. Luo H, Nijveen H. Understanding and identifying amino acid repeats. *Brief Bioinform*. 2014;15:582-591. doi:10.1093/bib/bbt003
69. Wang X, Wang Y, Wu G, Chao Y, Sun Z, Luo B. Sequence analysis of Epstein-Barr virus EBNA-2 gene coding amino acid 148-487 in nasopharyngeal and gastric carcinomas. *Virol J*. 2012;9:49. doi:10.1186/1743-422X-9-49
70. Urrutia G, de Assuncao TM, Mathison AJ, et al. Inactivation of the euchromatic histone-lysine n-methyltransferase 2 pathway in pancreatic epithelial cells antagonizes cancer initiation and pancreatitis-associated promotion by altering growth and immune gene expression networks. *Front Cell Dev Biol*. 2021;9:681153. doi:10.3389/fcell.2021.681153
71. Koutsioumpa M, Hatziapostolou M, Polytarchou C, et al. Lysine methyltransferase 2D regulates pancreatic carcinogenesis through metabolic reprogramming. *Gut*. 2019;68:1271-1286. doi:10.1136/gutjnl-2017-315690
72. Cannon MJ, Schmid DS, Hyde TB. Review of cytomegalovirus seroprevalence and demographic characteristics associated with infection. *Rev Med Virol*. 2010;20:202-213. doi:10.1002/rmv.655
73. Carrillo-Bustamante P, Keşmir C, de Boer RJ. Virus encoded MHC-like decoys diversify the inhibitory KIR repertoire. *Plos Comput Biol*. 2013;9:e1003264. doi:10.1371/journal.pcbi.1003264
74. Krosky PM, Underwood MR, Turk SR, et al. Resistance of human cytomegalovirus to benzimidazole ribonucleosides maps to two open reading frames: UL89 and UL56. *J Virol*. 1998;72:4721-4728. doi:10.1128/JVI.72.6.4721-4728.1998
75. Jenson AB, Rosenberg HS, Notkins AL. Pancreatic islet-cell damage in children with fatal viral infections. *Lancet (London, England)*. 1980;2:354-358. doi:10.1016/S0140-6736(80)90349-9
76. Oikarinen M, Tauriainen S, Honkanen T, et al. Analysis of pancreas tissue in a child positive for islet cell antibodies. *Diabetologia*. 2008;51:1796-1802. doi:10.1007/s00125-008-1107-8
77. Skog O, Korsgren O, Frisk G. Modulation of innate immunity in human pancreatic islets infected with enterovirus in vitro. *J Med Virol*. 2011;83:658-664. doi:10.1002/jmv.21924
78. Koma T, Huang C, Kolokoltsova OA, Brasier AR, Paessler S. Innate immune response to arenaviral infection: a focus on the highly pathogenic New World hemorrhagic arenaviruses. *J Mol Biol*. 2013;425:4893-4903. doi:10.1016/j.jmb.2013.09.028
79. Parks GD, Alexander-Miller MA. Paramyxovirus activation and inhibition of innate immune responses. *J Mol Biol*. 2013;425:4872-4892. doi:10.1016/j.jmb.2013.09.015