Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Massive peptide sharing between viral and human proteomes

Darja Kanduc a,*, Angela Stufano a, Guglielmo Lucchese a, Anthony Kusalik b

a Department of Biochemistry and Molecular Biology, University of Bari, Bari 70126, Italy
b Department of Computer Science, University of Saskatchewan, Saskatoon, Canada

ARTICLE INFO

Article history:
Received 29 February 2008
Received in revised form
28 May 2008
Accepted 30 May 2008
Published online 5 June 2008

Keywords:
Viral proteomes
Human proteome
Sequence similarity
Peptide sharing
Autoimmunity

ABSTRACT

Thirty viral proteomes were examined for amino acid sequence similarity to the human proteome, and, in parallel, a control of 30 sets of human proteins was analyzed for internal human overlapping. We find that all of the analyzed 30 viral proteomes, independently of their structural or pathogenic characteristics, present a high number of pentapeptide overlaps to the human proteome. Among the examined viruses, human T-lymphotropic virus 1, Rubella virus, and hepatitis C virus present the highest number of viral overlaps to the human proteome. The widespread and ample distribution of viral amino acid sequences through the human proteome indicates that viral and human proteins are formed of common peptide backbone units and suggests a fluid compositional chimerism in phylogenetic entities canonically classified distantly as viruses and Homo sapiens. Importantly, the massive viral to human peptide overlapping calls into question the possibility of a direct causal association between virus–host sharing of amino acid sequences and incitement to autoimmune reactions through molecular recognition of common motifs.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

Protein sequences of the human proteome as well as of a number of viral proteomes have become available in databanks, offering an opportunity for a thorough comparative analysis of their reciprocal inter-relationship(s). This area of inquiry is of special interest because of our persistent ignorance of the mechanisms at the basis of viral pathogenesis and virus–host interactions [1,19]. To define the molecular structural determinants possibly involved in viral virulence and human susceptibility, we have undertaken a systematic analysis of a large set of viral proteomes searching for amino acid sequences shared with the human proteome.

Specifically, the study was designed to answer two main questions: (i) what is the quantitative dimension of short-peptide sequence sharing between viruses and human proteomes? (ii) How many human proteins harbor viral peptide modules? The two questions are of paramount importance in the context of autoimmune diseases. Indeed, it is a common belief that an autoimmune reaction is mostly caused by a host receiving an antigen that has amino acid homology/similarity with amino acid sequences in self-antigens of the host [20]. This may result in the host immune system attacking the organs and tissues expressing the self-antigens with the shared sequences [23]. In this context and given the sustained increase in the incidence of autoimmune diseases [2,8,18,22], the mathematical quantification of peptide overlap extent between viruses and humans is essential to understand the role of structural viral similarity in the pathogenesis of autoimmunity. Here we report a virus versus human comparative screening that reveals a massive, indiscriminate, unexpected pentapeptide overlapping between viral and human proteomes.
2. Methods

The viral proteomes analyzed in this study were downloaded from www.ebi.ac.uk/genomes/virus.html and the human proteome obtained from UniProtKB (http://www.ebi.ac.uk/integr8). The human proteome originally contained 37,993 proteins at the time of download, but we filtered out (1) two viral contaminants (i.e. viral proteomes which were being expressed in human cells), and (2) duplicated sequences and fragments. After the filtering, we were left with a human proteome consisting of 36,103 unique proteins.

Similarity analyses were conducted on 30 viral proteomes for a total of 717 viral proteins, equal to 302,667 amino acids. A similarly sized control set of 30 human protein samples (for a total of 686 human proteins, equal to 302,520 amino acids) was analyzed for inter-human overlapping. More precisely, for each viral proteome size, we constructed an artificial human sub-proteome drawn from the human protein set. The proteins were selected at random without replacement, and assigned to an artificial sub-proteome at random. We kept adding human proteins to each artificial human sub-proteome until adding the next protein would have caused the total number of amino acids in the artificial sub-proteome to exceed the number of amino acids in the corresponding viral proteome. This way we eliminated possible bias in the comparison. The 30 artificial sub-proteomes are random samplings of human proteins and represent cross-sections of all human proteins. The list of the human proteins forming the 30 artificial human sub-proteomes is detailed as SwissProt entries in Table 3.

Sequence similarity analyses of each of the 30 viral proteomes (or the 30 artificial human sub-proteomes) to the human proteome were carried out using viral (or human) 5-mers sequentially overlapped by four residues. The scans were performed by custom programs written in C and utilizing suffix trees for efficiency [6]. The viral or human proteins were manipulated and analyzed as follows. The entire viral (or human) proteome was decomposed in silico to a set of 5-mers (including all duplicates). A library of unique 5-mers for each virus or human proteome was then created by removing duplicates. Next, for each 5-mer in the library, the entire human proteome was searched for instances of the same 5-mer. Any such occurrence was termed an overlap. Cursory analyses (e.g. identification of unique overlapping 5-mers, counts of unique overlapping 5-mers, counts of duplications) were performed using LINUX/UNIX shell scripts and standard LINUX/UNIX utilities. Data were plotted and linear least-squares regression was performed to determine whether any linear relationships exist between the level of overlap and the size of viral (or control) proteomes. The same procedure was applied when longer amino acid sequences (e.g., heptapeptides, octapeptides, etc.) were used as probes in the similarity analyses.

The similarity analysis for the artificial human sub-proteomes was performed as for the viral proteomes, with

| Tax Id | Virus description and abbreviation | Accession | Proteins | aa |
|--------|-----------------------------------|-----------|----------|----|
| 10407  | Hepatitis B virus (HBV)           | X51970    | 4        | 1,613 |
| 10632  | JC polyomavirus (JCV)             | J02226    | 5        | 1,629 |
| 10798  | Human parvovirus B19              | AF162273  | 3        | 2,006 |
| 12131  | Human rhinovirus 14 (HRV-14)      | K02121    | 1        | 2,179 |
| 12080  | Human poliovirus 1 (HPV-1)        | AJ132961  | 1        | 2,209 |
| 494309 | Saffold virus (SAF-V)             | EF165067  | 1        | 2,296 |
| 333760 | Human papillomavirus type 16 (HPV16) | K02718  | 8        | 2,452 |
| 11908  | Human T-cell leukemia virus 1 (HTLV-I) | U19949 | 6        | 2,589 |
| 11103  | Hepatitis C virus (HCV)           | AJ132997  | 1        | 3,010 |
| 11041  | Rubella virus                     | AF188704  | 2        | 3,179 |
| 11089  | Yellow fever virus (YFV)          | X03700    | 1        | 3,411 |
| 307044 | West Nile virus (WNV)             | AY842931  | 1        | 3,433 |
| 11676  | Human immunodeficiency virus 1 (HIV-1) | X01762 | 9        | 3,571 |
| 11292  | Rabies virus                      | M31046    | 5        | 3,600 |
| 11029  | Ross River virus (RRV)            | M20162    | 2        | 3,733 |
| 11709  | Human immunodeficiency virus 2 (HIV-2) | X05291 | 9        | 3,759 |
| 162145 | Human metapneumovirus (hMPV)      | AF371337  | 9        | 4,163 |
| 923388 | Influenza A virus (H5N1)          | AF144300  | 10       | 4,467 |
| 11250  | Human respiratory syncytial virus (HRSV) | AF013254 | 11       | 4,540 |
| 11216  | Human parainfluenza virus 3 (HPIV3) | AB012132 | 6        | 4,842 |
| 11269  | Lake Victoria marburgvirus        | Z12132    | 7        | 4,846 |
| 11161  | Mumps virus                       | A8040674  | 8        | 4,977 |
| 70149  | Measles virus                     | AY486084  | 8        | 5,205 |
| 186538 | Zaire virus                       | AF086833  | 9        | 5,493 |
| 63330  | Hendra virus                      | AF017149  | 9        | 6,056 |
| 321149 | SARS coronavirus (SARS-CoV)       | AY864806  | 13       | 14,209 |
| 10376  | Human herpesvirus 4 (HHV-4)       | AY961628  | 69       | 34,911 |
| 10368  | Human herpesvirus 6 (HHV-6)       | X83413    | 112      | 44,720 |
| 10255  | Variola virus                     | X69198    | 197      | 54,289 |
| 10359  | Human herpesvirus 5 (HHV-5)       | X17403    | 190      | 65,280 |

Total 717 30,2667
one modification. The modification is that all the proteins in the artificial proteome were removed from the comparison human proteome when the overlap determination was done. E.g., the artificial human sub-proteome 9 contains 12 human proteins selected from the original set of 36,103 proteins comprehensively forming the human proteome. These 12 proteins were removed from the comparison set prior to determining the overlap of artificial human sub-proteome 9 versus human; i.e. for artificial proteome 9, the overlap comparison was between 12 proteins in the artificial proteome and the remaining 36,091 human proteins. And similar for each of the other 29 artificial proteomes.

### 3. Results

The viruses were chosen based on the following criteria: (1) known to be pathogenic to human; (2) of significant health impact (e.g. HCV, rhinovirus, HIV, mumps, measles, SARS, HPV, and polio); (2) phylogenetically different; (4) proteomes established to a significant degree of completeness. In addition, the viral proteomes were chosen to span a range of proteome sizes, with the smallest viral proteome being 1613 amino acids, the largest being 65,280 amino acids, and all the viral proteomes combined amounting to 302,667 amino acids.

The viral proteomes are described in Table 1. As controls, we used 30 similarly sized protein sets drawn from the human proteome. In brief, we constructed 30 artificial sub-proteomes populated by proteins randomly selected without replacement from the human proteome. The constraints we used were that (1) the size of each artificial sub-proteome was to approximate the size of one of the viral proteomes, and (2) that the combined size of all the artificial proteomes was to approximate the combined size (in amino acids) of the viral proteomes. The 30 resultant artificial human sub-proteomes are described in Table 2. Table has four columns. The first is the artificial sub-proteome number. The second is the cumulative amino acid length of all the human proteins in that artificial sub-proteome. The third column is the cumulative amino acid length of the corresponding viral proteome. The fourth column is the number of human proteins in the artificial sub-proteome. For example, in our virus versus human analysis, information for hepatitis C virus (HCV) was used as additional probes to control the similarity interactions such as peptide–protein interaction and (auto)-immune recognition ([5,7,17,18,21] and further Refs. therein). Longer amino acid sequences (esa-, epta-, octapeptides, etc.) were used as additional probes to control the similarity pattern.

#### 3.1. Quantitative analysis of the pentapeptide overlapping of viral versus human proteomes

The numerical values that define virus-to-human similarity level are reported in Table 4. The table documents that all of the 30 viral proteomes under analysis have high and wide pentapeptide overlapping to the human proteome. Only a limited number of viral pentamers are unique to the viruses, with no counterpart in the human proteome. The overlapping extent (in terms of percentage of unique viral 5-mers which occur in the human proteome, see column 6 in Table 4) is rarely less than 90%. That means that almost 90% of the viral pentamer peptides are widely, intensively and repeatedly scattered throughout the human proteome. Numerically, the viral versus human overlap is defined by 2,907,096 total structural units critically involved in biological/pathological interactions such as peptide–protein interaction and (auto)-immune recognition ([5,7,17,18,21] and further Refs. therein). Longer amino acid sequences (esa-, epta-, octapeptides, etc.) were used as additional probes to control the similarity pattern.

| Sub-proteome number | Size of the corresponding viral proteome | Human proteins in the sub-proteome |
|---------------------|------------------------------------------|----------------------------------|
| 1                   | 1,595                                    | 7                                |
| 2                   | 1,616                                    | 4                                |
| 3                   | 1,997                                    | 9                                |
| 4                   | 2,142                                    | 9                                |
| 5                   | 2,186                                    | 6                                |
| 6                   | 2,290                                    | 7                                |
| 7                   | 2,433                                    | 4                                |
| 8                   | 2,585                                    | 7                                |
| 9                   | 2,979                                    | 12                               |
| 10                  | 3,156                                    | 19                               |
| 11                  | 3,383                                    | 8                                |
| 12                  | 3,427                                    | 11                               |
| 13                  | 3,554                                    | 11                               |
| 14                  | 3,594                                    | 10                               |
| 15                  | 3,721                                    | 12                               |
| 16                  | 3,759                                    | 11                               |
| 17                  | 4,127                                    | 8                                |
| 18                  | 4,437                                    | 13                               |
| 19                  | 4,528                                    | 13                               |
| 20                  | 4,827                                    | 8                                |
| 21                  | 4,812                                    | 11                               |
| 22                  | 4,945                                    | 11                               |
| 23                  | 5,200                                    | 8                                |
| 24                  | 5,492                                    | 16                               |
| 25                  | 6,037                                    | 15                               |
| 26                  | 14,203                                   | 35                               |
| 27                  | 34,894                                   | 72                               |
| 28                  | 44,655                                   | 98                               |
| 29                  | 54,263                                   | 111                              |
| 30                  | 65,683                                   | 130                              |
| Total               | 302,520                                  | 686                              |

a Amino acid number.
b Composed by set of human proteins as detailed under Section 2 and in Table 3.
matches. Also the viral overlapping distribution in the human proteome (as number of human proteins hosting viral 5-mers, column 5 in Table 4) is highest, being close to 100% with HHV-5 (i.e. HHV-5 pentamers are present in 35,708 human proteins out of the 36,103 ones which comprehensively form the human proteome).

Given the ample literature documentation and research data in support of the critical role exerted by pentapeptide modules in cell biology as well in antigen/antibody immunore cognition ([17], and Refs. therein), the data of Table 4 are impressive. They become even more significant, if possible, when compared to the control data obtained for inter-human

| Sub-proteome | Proteins (Swiss-Prot accession number) |
|--------------|----------------------------------------|
| 1 | Q8NAW6; Q6IBU4; ARFP1_HUMAN; SODM_HUMAN; AIF1_HUMAN; MAGB1_HUMAN; Q6ZNA3 |
| 2 | Q6Z2V9; Q6ETW5; Q6UY40; Q9U773 |
| 3 | Q7UT18; Q5TC13; GALT_HUMAN; LV1H_HUMAN; Q6ZUC0; Q9H0S3; TRXR2_HUMAN; Q5TG59; Q16367 |
| 4 | Q6NT40; Q5HY76; Q6KX19; Q8BMX4; Q57270; Q9UH93; Q6ZT88; Q6ZSJ1; Q52127 |
| 5 | Q6IPM5; GPR39_HUMAN; Q6ZVY8; Q6YQH2; Q4IVX4; S10AF_HUMAN |
| 6 | Q7CVQY9; Q6FSQ5; Q6VSD3; Q8VVE6; Q6Z2N3; Q9H882; Q4IVT5 |
| 7 | Q6S5G0; Q3YL75; Q8BXX2; Q5G9Q5 |
| 8 | DSC8_HUMAN; PKSCA_HUMAN; Q6PD83; Q8TB28; Q8TBV1; PHP1H_HUMAN; Q9N6R9 |
| 9 | LAP2A_HUMAN; Q7S100; Q9I8AY; Q9NWL9; Q6ZQN3; ZNF73_HUMAN; VEGFC_HUMAN; Q96FS0; Q5TBJ1; Q8I1X3; Q6A909; Q6S591 |
| 10 | Q5IV79; ATPK_HUMAN; Q6ZT66; Q6S3Q3; SYT9_HUMAN; Q5J8P; Q715S7; Q5YF53; Q6P80 |
| 11 | Q53E87; Q12771; DXD1_HUMAN; Q66X2; GNA12_HUMAN; Q8TCS7; DAX1_HUMAN; Q9BFY6 |
| 12 | Q6G7M5; P5R23_HUMAN; Q5SF8; Q6NK0; Q4IVT5; Q96G74; Q6F9N7; Q6Q3S7; Q52127 |
| 13 | Q5T1I_HUMAN; Q565L8; Q5T0J9; Q6B08; Q6ZMV2; Q86T62; Q4G161; Q6H0Q4; TSH3_HUMAN; Q6ZNA9; Q5TG33 |
| 14 | ORS1U_HUMAN; Q8RBW6; Q5SAE7; Q5001_HUMAN; Q6NC13; Q6F9N7; Q6Q3S7; Q52127 |
| 15 | Q6Z2V3; ADH4_HUMAN; Q86TC3; Q5ZME7; Q5M2O; Q6F9N7; Q6Q3S7; Q52127 |
| 16 | AT1I_HUMAN; Q565L8; Q5T0J9; Q6B08; Q6ZMV2; Q86T62; Q4G161; Q6H0Q4; TSH3_HUMAN; Q6ZNA9; Q5TG33 |
| 17 | OR5U1_HUMAN; Q9BRW6; FBX2_HUMAN; Q5TAE7; Q5001_HUMAN; Q6NC13; Q6F9N7; Q6Q3S7; Q52127 |
| 18 | Q6ZRR8; Q5T435; Q6NUS8; CX033_HUMAN; ALG1_HUMAN; Q6ZUT0; Q86UH7; NCKX5_HUMAN; Q9BRY8; Q6F9N7; Q6Q3S7; Q52127 |
| 19 | Q6ZTL0; Q6H9L7; Q5F6S3; Q5NK0; Q4IVT5; Q96G74; Q6F9N7; Q6Q3S7; Q52127 |
| 20 | VPS16_HUMAN; GLUC_HUMAN; Q5T49_HUMAN; Q5H100; Q6K6S6; Q5X17; Q9HF66 |
| 21 | Q5ZS5N; DNAL4_HUMAN; Q59F6; Q5T3G3; OR211_HUMAN; Q9NWV6; DPF3_HUMAN; Q96F50; Q5T884; Q6V3P3 |
| 22 | GAK14_HUMAN; Q69612; IL17D_HUMAN; Q6ZWP8; Q5JXU3; Q8N40; Q8NEC4; Q9NL7; Q5X5N7; Q6ZQV9 |
| 23 | Q5TS4V; Q4VHES; Q5SD3_HUMAN; Q6Z5R4; Q8X67; BCLW_HUMAN; SPR2B_HUMAN; Q5S5U1 |
| 24 | Q5STC6; Q6H066; SG11_HUMAN; MLF2_HUMAN; Q6ZJS6; Q5R6G0; Q8WBU4; Q5W65; Q8N966; Q5YH77 |
| 25 | GP1A4_HUMAN; Q6Z6R8; Q5X18; G1O10_HUMAN; Q6A43; Q5ZV28 |
| 26 | Q6L2Z1; Q7RA6; CF060_HUMAN; Q5YV0; NEF2_HUMAN; APEX1_HUMAN; Q9R6E8; Q8T3B3; KLF11_HUMAN; Q9NH68; Q5T71A3; Q5NV17; Q5T870; Q6N36; Q6ZK9; Q5Z69 |
| 27 | GP1L00_HUMAN; Q6T7Q0; KPSH1_HUMAN; DAZ2_HUMAN; Q5U5D4; Q5T585; ZNF6_HUMAN; Q9N6V7; QKG1B_HUMAN; AMYV_HUMAN; KRA44_HUMAN; Q6P78; Q86B3; TR10_HUMAN; MAGC1_HUMAN; Q96F50; Q5T884; Q6V3P3 |
| 28 | VPS16_HUMAN; GLUC_HUMAN; Q5T49_HUMAN; Q5H100; Q6K6S6; Q5X17; Q9HF66 |
pentapeptide overlapping illustrated in Table 5. When 30 human artificial sub-proteomes were analyzed for pentapeptide overlapping to the entire human proteome, we obtained clear numerical evidence that the inter-human level of perfect 5-mer matching is of the same order of magnitude as the viral overlapping to human proteome detailed in Table 4, i.e. there are 3,713,010 inter-human versus 2,907,096 viral perfect 5-mer matches to the human proteome (see Table 5). Likewise, the number of human proteins involved in viral and inter-human overlaps is highest and practically identical in the numbers: human proteins are involved in the inter-human peptide overlapping for 697,373 times and in the viral versus human

| Sub-proteome | Proteins (Swiss-Prot accession number) |
|--------------|---------------------------------------|
| 29           | Q6P2D1; Q9P1E4; RHPN2_HUMAN; Q8N657; Q5SJU4; DOCK3_HUMAN; PLOD1_HUMAN; SEL1L_HUMAN; O43379; Q6S382; STX4_HUMAN; Q15281; CP2A6_HUMAN; Q9H2Q4; Q5TF13; Q6UX27; Q5JR60; Q59FA5; Q6L6T2; K2CG6_HUMAN; GBRA2_HUMAN; CEBPG_HUMAN; DCD2B_HUMAN; Q6A2Z_HUMAN; IGF1R_HUMAN; Q8N139; Q6W6W7; JIP2_HUMAN; Q86SE5; Q6IF87; Q9H4T6; FBLN3_HUMAN; 1C06_HUMAN; MOB2_HUMAN; Q6RL15; Q9HK50; NEUR4_HUMAN; Q8NY1Y0; SPA6G_HUMAN; CRB81_HUMAN; THN5L_HUMAN; Q6NW61; 9KD_HUMAN; Q92638; RCBT1_HUMAN; IMDH2_HUMAN; Q5VYN8; Q6UXJ7; Q5QGD7; Q9657; FHL3_HUMAN; Q59F8X; 000318; Q13584; ATRN_HUMAN; Q5VXJ0; 000172; Q7TB16; FBX47_HUMAN; Q59EY9; Q6ED5; ABC9C_HUMAN; Q86T11; Q655S4; Q5PC80; Q7ZSY7; Q6IA40; RB33A_HUMAN; TAAR2_HUMAN; ZNH1I_HUMAN; Q6MWZ7; Q5VSZ2; CI074_HUMAN; TMM28_HUMAN; O95082; Q5ST44; PTPRU_HUMAN; SIGI5_HUMAN; Q6K25; Q5QWX7; Q8IU00; CABCA_HUMAN; Q56IG1; TIMD1_HUMAN; Q7S799; Q5Z772; Q8NN98; Q8TA0J; Q5P041; UBR1_HUMAN; Q5BY14; Q3V6T2; PART1_HUMAN; BTLN3_HUMAN; Q566T1; Q5WYS5; Q6AQ7A; Q5WXV1; Q58D36; RNAS9_HUMAN; Q5QXV0; ORSP3_HUMAN; DCK4_HUMAN; Q5NT50; Q5S376; 086K2_HUMAN; Q58DF6; EPC1_HUMAN; Q96BJ2; Q6M6Q; Q5Z6V4; Q6WY5G; ETV2_HUMAN; Q6UXY4; GPCE5_HUMAN; Q6UA4A; Q6ZM6B; Q5Q99; Q8Z2U3; Q59U7D; QBAT7V; Q6V6P5; Q54F15_HUMAN; RB3GR_HUMAN; Q6ZSQ0; Q5P073; UEB2_T_HUMAN; Q5JUP7; Q50P35; Q562W8; STAC3_HUMAN; Q57TF9; Q5W6PW; SALL2_HUMAN; Q9H3F1; Q6M24; Q6R3X3; Q5M311; Q9HBS1; Q5S5X9; Q5NE1I; Q5BP2A_HUMAN; Q52W4A; F102B_HUMAN; HES5_HUMAN; Q5JUUE; UTP20_HUMAN; TLR4_HUMAN; Q5Q5L8; H2V8G6; H4VCA_HUMAN; Q56V9; Q587V4; HX9A_HUMAN; Q5SC55; QTBE17; Q58NW8; TRPV1_HUMAN; Q5N3Q9; Q5P99; Q5FQ02; FA47B_HUMAN; Q5055_HUMAN; BAD_HUMAN; Q57755; Q5LTW9; Q5S6F2; Q5WV52; PCD16_HUMAN; 3BHS7_HUMAN; CB013_HUMAN; Q5613_HUMAN; Q56A6; Q5HH91; Q5N11Y8; Q5B4V2; Q525ID; MYOTI_HUMAN; Q56P0L; Q58IC2; Q58WSJ3; SPDYA_HUMAN; Q59569; Q5Q2D9; Q557XR2; Q5ST3L7; Q58ND3; Q5JR95; Q5TYY8

Table 3 (Continued) | Proteins (Swiss-Prot accession number)
| 30 | PIAS3_HUMAN; MP2KS_HUMAN; Q6E2A4; Q6E6L3; SIA7A_HUMAN; Q6W6W9; Q6Z2W4; ARSG_HUMAN; Q56FVS; Q5T09B; OLR1_HUMAN; Q7T352; Q6PN1; A1M1_HUMAN; Q5NGL5; CDD37_HUMAN; Q95082; Q5ST44; PTRPU_HUMAN; SIGL5_HUMAN; Q6K25; Q5QWX7; Q8IU00; CABCA_HUMAN; Q56IG1; TIMD1_HUMAN; Q7S799; Q5Z772; Q8NN98; Q8TA0J; Q5P041; UBR1_HUMAN; Q5BY14; Q3V6T2; PART1_HUMAN; BTLN3_HUMAN; Q566T1; Q5WYS5; Q6AQ7A; Q5WXV1; Q58D36; RNAS9_HUMAN; Q5QXV0; ORSP3_HUMAN; DCK4_HUMAN; Q5NT50; Q5S376; 086K2_HUMAN; Q58DF6; EPC1_HUMAN; Q96BJ2; Q6M6Q; Q5Z6V4; Q6WY5G; ETV2_HUMAN; Q6UXY4; GPCE5_HUMAN; Q6UA4A; Q6ZM6B; Q5Q99; Q8Z2U3; Q59U7D; QBAT7V; Q6V6P5; Q54F15_HUMAN; RB3GR_HUMAN; Q6ZSQ0; Q5P073; UEB2_T_HUMAN; Q5JUP7; Q50P35; Q562W8; STAC3_HUMAN; Q57TF9; Q5W6PW; SALL2_HUMAN; Q9H3F1; Q6M24; Q6R3X3; Q5M311; Q9HBS1; Q5S5X9; Q5NE1I; Q5BP2A_HUMAN; Q52W4A; F102B_HUMAN; HES5_HUMAN; Q5JUUE; UTP20_HUMAN; TLR4_HUMAN; Q5Q5L8; H2V8G6; H4VCA_HUMAN; Q56V9; Q587V4; HX9A_HUMAN; Q5SC55; QTBE17; Q58NW8; TRPV1_HUMAN; Q5N3Q9; Q5P99; Q5FQ02; FA47B_HUMAN; Q5055_HUMAN; BAD_HUMAN; Q57755; Q5LTW9; Q5S6F2; Q5WV52; PCD16_HUMAN; 3BHS7_HUMAN; CB013_HUMAN; Q5613_HUMAN; Q56A6; Q5HH91; Q5N11Y8; Q5B4V2; Q525ID; MYOTI_HUMAN; Q56P0L; Q58IC2; Q58WSJ3; SPDYA_HUMAN; Q59569; Q5Q2D9; Q557XR2; Q5ST3L7; Q58ND3; Q5JR95; Q5TYY8
| Virus          | 1     | 2     | 3     | 4     | 5     | 6   |
|---------------|-------|-------|-------|-------|-------|-----|
| Hendra virus  | 5,210 | 6,020 | 4,810 | 57,646| 22,644| 92.3|
| SARS-CoV      | 9,739 | 14,157| 8,853 | 108,632| 27,957| 90.9|
| HHV-4         | 32,009| 34,635| 29,955| 531,946| 35,092| 93.5|
| HHV-6         | 41,834| 44,272| 38,277| 467,206| 35,057| 91.4|
| Smallpox virus| 52,017| 53,501| 47,180| 498,970| 35,035| 90.7|
| HHV-5         | 61,001| 64,520| 56,300| 883,952| 35,708| 92.2|
| All           | 257,035| 299,701| 234,691| 2,907,096|       |     |

Human proteome formed by 36,103 proteins and 15,771,565 occurrences of 2,388,563 unique 5-mers. Column number refers to: (1) unique 5-mers in the viral proteome; (2) total number of 5-mers in the viral proteome (including multiple occurrences); (3) unique viral 5-mers occurring in the proteome; (4) viral overlap occurrences in the human proteome (including multiple occurrences); (5) number of human proteins involved in overlap; (6) % of unique viral 5-mers which occur in the human proteome (i.e. 100 \ times column 3/column 1).

`a` Abbreviations as in Table 1.

`b` The results of linear regression analysis between columns 1 and 3, and 1 and 4 are: column 3 = 0.91811 \ times column 1 + 1.2272 (r = 0.99993). Column 4 = 12.636 \ times column 1 + 269.01 (r = 0.97452).

c Obtained by combining all 30 viral proteomes into one viral proteome, and then computing the overlap with the human proteome.

| Human sub-proteome* | 1     | 2     | 3     | 4     | 5     | 6   |
|---------------------|-------|-------|-------|-------|-------|-----|
| 1                   | 1,566 | 1,567 | 1,496 | 27,014| 14,381| 95.5|
| 2                   | 1,598 | 1,600 | 1,576 | 26,352| 13,850| 98.6|
| 3                   | 1,937 | 1,956 | 1,885 | 44,087| 17,222| 97.3|
| 4                   | 2,100 | 2,106 | 2,064 | 30,025| 14,400| 98.2|
| 5                   | 2,058 | 2,162 | 1,986 | 40,374| 17,653| 96.5|
| 6                   | 2,131 | 2,262 | 2,098 | 62,564| 16,698| 98.4|
| 7                   | 2,289 | 2,417 | 2,227 | 37,796| 17,303| 97.2|
| 8                   | 2,556 | 2,557 | 2,475 | 33,387| 16,439| 96.8|
| 9                   | 2,858 | 2,931 | 2,721 | 81,650| 20,105| 95.2|
| 10                  | 3,067 | 3,120 | 3,042 | 48,918| 19,069| 99.1|
| 11                  | 3,306 | 3,351 | 3,198 | 59,666| 21,568| 96.7|
| 12                  | 3,355 | 3,383 | 3,271 | 68,549| 22,697| 97.4|
| 13                  | 3,478 | 3,510 | 3,377 | 56,469| 21,221| 97.0|
| 14                  | 3,529 | 3,554 | 3,417 | 60,138| 21,490| 96.8|
| 15                  | 3,663 | 3,673 | 3,596 | 88,229| 22,454| 98.1|
| 16                  | 3,626 | 3,715 | 3,545 | 96,827| 22,061| 97.7|
| 17                  | 4,084 | 4,095 | 3,958 | 72,451| 23,645| 96.9|
| 18                  | 4,374 | 4,385 | 4,112 | 68,579| 24,055| 94.0|
| 19                  | 4,423 | 4,476 | 4,295 | 79,283| 24,472| 97.1|
| 20                  | 4,751 | 4,795 | 4,735 | 70,296| 22,686| 99.6|
| 21                  | 4,646 | 4,768 | 4,555 | 115,916| 26,284| 98.0|
| 22                  | 4,886 | 4,901 | 4,850 | 95,959| 25,348| 99.2|
| 23                  | 5,048 | 5,168 | 4,740 | 72,400| 23,453| 93.8|
| 24                  | 5,408 | 5,428 | 5,144 | 101,686| 26,713| 95.1|
| 25                  | 5,941 | 5,977 | 5,841 | 128,904| 27,751| 98.3|
| 26                  | 13,143| 14,063| 12,700| 290,022| 32,681| 96.6|
| 27                  | 33,281| 34,606| 32,460| 664,539| 35,181| 97.5|
| 28                  | 42,683| 44,263| 41,535| 787,778| 35,757| 97.3|
| 29                  | 51,865| 53,804| 50,075| 895,672| 35,489| 96.5|
| 30                  | 62,219| 65,156| 60,506| 1,103,962| 35,629| 97.2|
| All                | 254,808| 299,749| 246,322| 3,713,010|      |     |

All 30 artificial human sub-proteomes constitute 686 proteins and are numbered from 1 to 30. The comparison human proteome contained 36,103 proteins and 15,771,565 occurrences of 2,388,563 unique 5-mers. Column number refers to: (1) unique 5-mers in the artificial sub-proteome; (2) total number of 5-mers in the artificial sub-proteome (including multiple occurrences); (3) unique 5-mers from the artificial sub-proteome occurring in the human proteome; (4) occurrences in the human proteome of 5-mers from artificial sub-proteome (including multiple occurrences); (5) number of human proteins in the human proteome involved in overlap; (6) % of unique 5-mers from the artificial sub-proteome which occur in the human proteome (i.e. 100 \ times column 3/column 1).

`a` Analogous to viral proteomes in size (see Table 1), and composed by set of human proteins as detailed in Table 3.

`b` The results of linear regression analysis between columns 1 and 3, and 1 and 4 are: column 3 = 0.97083 \ times column 1 + 0.76628 (r = 0.99998). Column 4 = 17.921 \ times column 1 + 62.786 (r = 0.97452).

`c` Obtained by combining all 30 artificial sub-proteomes into one artificial sub-proteome, and then computing the overlap with the entire human proteome.
overlap for 633,229 times (see data from Tables 4 and 5, respectively). This indicates a massive, repeated, acritical usage of the same pentapeptide blocks in viral and human proteomes.

Plotting the viral 5-mer occurrences in the human proteome as a function of viral proteome length (see data under Table 4, columns 1 and 4). The symbols refer to: (●), HTLV-1; (■), Rubella virus; (□), HCV; (+), other viral data point. The regression line (—) has an equation of $y = 12.636x - 269.01$, with a Pearson correlation coefficient ($r$ value) of 0.97452. Both $x$- and $y$-axis are log scale.

![Graph showing viral 5-mer occurrences in the human proteome](image)

### 3.2. Qualitative analysis of the pentapeptide overlapping of viral versus human proteomes

To understand the overlap usage in entities so distant in the evolutionary time and so enormously different in the evolutionary history such as viruses and humans, we investigated the profile of viral motif distribution along the human proteome. Representative histograms are reported in Fig. 2. It can be seen that, independently of the virus size, the density of the viral 5-mer motif matches along the human proteome presents constant behavior, with virus portions endowed with high similarity alternating with portions scarcely represented in the human proteome. The same alternating behavior is shown by the histograms of inter-human overlapping by analyzing sub-proteomic sets of low, medium and large size (Fig. 3).

Figs. 2 and 3 confirm the commonality of pentapeptide block usage in viruses and humans and, more in general, document the existence of a basic structural platform in the protein world. As shown in Tables 6 and 7, the viral versus human overlapping is still remarkable by using viral esa- or eptapeptide sequences as probes.

### 3.3. Non-stochastic nature of the peptide overlapping between the viral and human proteomes under analysis

The results illustrated in Tables 4, 6 and 7 are indicative of a widespread peptide overlapping between viral proteins and the *Homo sapiens* proteome. In order to understand how the above reported data are mathematically governed, we explored the 30 viral proteomes and the 30 human sub-proteomes under analysis for the degree of internal redundancy and the viral versus human proteome overlapping at n-mer level (with $n$ from 5 to 16 amino acids). The quantitative data we found are reported in Table 8 listing three orders of data: (a) the number of unique occurrences of $n$-peptides in the 30 viral (or the 30 human) proteome samples; (b) the total number of occurrences (i.e. including multiple occurrences) of $n$-peptides in the 30 viral (or the 30 human) proteome samples; and (c) the relative viral versus human $n$-peptide overlaps.

The first set of data, column 1 through 6 in Table 8, provides evidence that the numerical values for uniquely expressed peptides are always lower than the actual total values for peptide occurrences. This is remarkable in light of the enormously high number of potential $n$-peptides which theoretically are available. As an example, in face of the possible 3,200,000 pentamers, both the 30 viral proteomes and the 30 human sub-proteomes present a high degree of repetitiveness in their 5-mer composition. In fact, the 30 viral proteomes and the 30 human sub-proteomes are respectively formed by a total of 299,701 and 299,749 5-mers and, although there is the possibility of forming 3,200,000 different pentapeptides, these total 5-mers present 42,666 and 44,941 repeated pentapeptides, in the viral and human proteins, respectively. The intra-viral (or intra-human) peptide redundancy strikingly persists at higher $n$-mer level, despite the exponentially increasing numbers of theoretically possible $n$-mers. The intra-repetitiveness is mathematically quantified in Table 8, columns 3 and 6, for the 30 viral proteomes and the 30 human sub-proteomes, respectively.
The second set of data illustrates the overlapping at n-peptide level (with n from 5 to 16 amino acids) from the 30 viral proteomes versus the 30 human sub-proteomes (Table 8, columns 7 and 8). Again, we note that, although there is a potential pentapeptide reservoir amounting to 3,200,000, still 42,647 of unique viral pentapeptides (16.6% of the total) appear in the human proteins forming the 30 human sub-proteomes. When considering the total viral pentapeptides including multiple occurrences (Table 8, column 8), the viral contribution to the 30 human sub-proteomes raises to 56,230 pentamers. These repeated viral overlaps in the human proteins appear to further support the existence of a defined peptide repertoire which is utilized (even repeatedly) in protein composition. A similar observation holds in considering the viral peptide overlapping to the 30 human sub-proteomes at higher n-mer level (Table 8, columns 7 and 8).

Finally, the third set of data documents the overlaps at n-peptide level (with n from 5 to 16 amino acids) from the 30 viral proteomes versus the entire human proteome (Table 8, columns 9 and 10). It can be seen that, independently of the vastness of the human protein sample, the extent of the viral overlap is relevant, especially when considering that the human proteome has the possibility of drawing from an enormous number of potential theoretical n-peptides. For example, notwithstanding the astonishingly great repertoire of theoretical possible undecapeptides (that is: 204,800,000 millions), nonetheless eight viral overlaps (one of which even repeated) are present in the sub-proteomic human samples. The numbers raise to 352 and 661 (including multiple occurrences of the same 11-mer) when considering the viral 11-mer overlaps in the entire human proteome.

Our conclusion is that the mathematical redundancy present in the protein world is not stochastic (i.e. is not pure random chance), but rather reflects strong peptide usage bias since certain peptides are repeatedly used (and shared) in (and among) viral and human proteins. It seems that peptide usage in the protein world (and, consequently, the degree of similarity, repetition, overlapping) is not dictated by pure mathematical laws of distribution. Rather, powerful constraints appear to be at work by forcing the use of restricted platforms of n-mers and privileging the repeated usage of the chosen ones. Possibly, as anticipated by preliminary data from our labs (Kusakil et al., manuscript in preparation), these powerful constraints able to influence and favor specific n-mer synthesis might be found among physico-chemical factors such as hydrophobicity, residue bulkiness, $\Delta G$ of peptide bond formation, plus biological factors such as codon bias.
Fig. 3 – Human versus human pentapeptide overlapping: similarity profile of low-, medium- and large-sized artificial human sub-proteomes to the human proteome. Human sub-proteome set (see also Table 2): (A) 4 (total aa: 2142); (B) 26 (total aa: 14,203); (C) 29 (total aa: 54,263).

Table 6 – Viral versus human proteome overlap at the 6-mer level

| Virus                  | 1     | 2     | 3     | 4     | 5     | 6     |
|------------------------|-------|-------|-------|-------|-------|-------|
| HBV                    | 1,589 | 1,593 | 496   | 1,394 | 1,224 | 31.2  |
| JCV                    | 1,527 | 1,604 | 461   | 1,468 | 1,301 | 30.1  |
| Human parvovirus B19  | 1,442 | 1,991 | 431   | 1,341 | 1,150 | 29.8  |
| HRV-15                 | 2,174 | 2,174 | 579   | 1,456 | 1,270 | 26.6  |
| HPV-1                  | 2,204 | 2,204 | 619   | 1,483 | 1,288 | 28.0  |
| SAF-V                  | 2,285 | 2,285 | 622   | 1,392 | 1,280 | 27.2  |
| HPV16                  | 2,412 | 2,412 | 739   | 2,010 | 1,703 | 30.6  |
| HTLV-I                 | 2,559 | 2,559 | 967   | 3,261 | 2,658 | 37.7  |
| HCV                    | 3,005 | 3,005 | 1,025 | 4,195 | 3,099 | 34.1  |
| Rubella virus          | 3,169 | 3,169 | 1,133 | 3,918 | 3,119 | 35.7  |
| YFV                    | 3,406 | 3,406 | 1,005 | 2,680 | 2,322 | 29.5  |
| WNIV                   | 3,428 | 3,428 | 996   | 2,744 | 2,365 | 29.0  |
| HIV-1                  | 3,084 | 3,526 | 904   | 2,135 | 1,852 | 29.3  |
| Rabies virus           | 3,575 | 3,575 | 1,162 | 2,635 | 2,195 | 32.5  |
| RRV                    | 3,613 | 3,613 | 1,069 | 2,708 | 2,301 | 29.5  |
| HIV-2                  | 3,283 | 3,714 | 1,006 | 4,653 | 2,780 | 30.6  |
| hMPV                   | 4,117 | 4,118 | 1,245 | 3,263 | 2,720 | 30.2  |
| H5N1                   | 4,408 | 4,417 | 1,191 | 2,674 | 2,332 | 27.0  |
| HRSV                   | 4,483 | 4,485 | 1,258 | 2,847 | 2,455 | 28.0  |
| HPV3                   | 4,812 | 4,812 | 1,350 | 3,194 | 2,721 | 28.0  |
| Lake Victoria marburgvirus | 4,811 | 4,811 | 1,497 | 5,631 | 3,695 | 31.1  |
### Table 6 (Continued)

| Virus | 1 | 2 | 3<sup>a</sup> | 4<sup>b</sup> | 5 | 6 |
|-------|---|---|-------------|-------------|---|---|
| Mumps virus | 4,787 | 4,937 | 1,463 | 3,734 | 3,017 | 30.5 |
| Measles virus | 4,937 | 5,165 | 1,585 | 3,823 | 3,208 | 32.1 |
| Zaire virus | 4,868 | 5,448 | 1,448 | 3,488 | 2,929 | 29.7 |
| Hendra virus | 5,210 | 6,011 | 1,516 | 3,538 | 3,035 | 29.0 |
| SARS-CoV | 9,767 | 14,144 | 2,704 | 9,438 | 5,770 | 27.6 |
| HHV-4 | 32,634 | 34,566 | 12,358 | 50,013 | 18,445 | 37.8 |
| HHV-6 | 42,432 | 44,160 | 12,126 | 35,793 | 16,847 | 28.5 |
| Variola virus | 53,136 | 53,304 | 13,870 | 33,911 | 16,167 | 26.1 |
| HHV-5 | 63,115 | 64,330 | 21,873 | 87,931 | 23,629 | 34.6 |
| All<sup>c</sup> | 282,932 | 298,966 | 87,224 | 259,655 | |

Human proteome formed by 36,103 proteins and 15,734,725 occurrences of 8,247,275 unique 6-mers. Column number refers to: (1) unique 6-mers in the viral proteome; (2) total number of 6-mers in the viral proteome (including multiple occurrences); (3) unique viral 6-mers occurring in the human proteome; (4) viral overlap occurrences in the human proteome (including multiple occurrences); (5) number of human proteins involved in overlap; (6) % of unique viral 6-mers which occur in the human proteome (i.e. 100 \times \text{column 3}/\text{column 1}).

<sup>a</sup> Abbreviations as in Table 1.

<sup>b</sup> The results of linear regression analysis between columns 1 and 3, and 1 and 4 are: column 3 = 0.31426/\text{column 1}/42.237 (r = 0.98762).

<sup>c</sup> Column 4 = 1.0972/\text{column 1}/877.68 (r = 0.93373).

<sup>c</sup> Obtained by combining all 30 viral proteomes into one viral proteome, and then computing the overlap with the entire human proteome.

---

### Table 7 – Viral versus human proteome overlap at the 7-mer level

| Virus<sup>a</sup> | 1 | 2 | 3<sup>b</sup> | 4<sup>b</sup> | 5 | 6 |
|-------------------|---|---|-------------|-------------|---|---|
| HBV | 1,586 | 1,589 | 66 | 102 | 93 | 4.1 |
| JCV | 1,523 | 1,599 | 61 | 108 | 101 | 4.0 |
| Human parvovirus B19 | 1,440 | 1,988 | 49 | 120 | 104 | 3.4 |
| HRV-14 | 2,173 | 2,173 | 60 | 118 | 100 | 2.7 |
| SAf-V | 2,283 | 2,283 | 35 | 51 | 51 | 1.5 |
| HPV16 | 2,404 | 2,404 | 102 | 163 | 121 | 4.2 |
| HTLV-I | 2,553 | 2,553 | 143 | 283 | 215 | 5.6 |
| HCV | 3,004 | 3,004 | 129 | 403 | 369 | 4.2 |
| Rubella virus | 3,167 | 3,167 | 171 | 309 | 282 | 5.3 |
| YFV | 3,405 | 3,405 | 96 | 164 | 153 | 2.8 |
| WNv | 3,427 | 3,427 | 96 | 177 | 170 | 2.8 |
| HIV-1 | 3,079 | 3,517 | 96 | 163 | 148 | 3.1 |
| Rabies virus | 3,570 | 3,570 | 132 | 208 | 192 | 3.6 |
| RRV | 3,595 | 3,595 | 119 | 184 | 169 | 3.3 |
| HIV-2 | 3,278 | 3,705 | 121 | 1,347 | 535 | 3.6 |
| hMPV | 4,109 | 4,109 | 123 | 234 | 212 | 2.9 |
| HSN1 | 4,400 | 4,407 | 96 | 147 | 139 | 2.1 |
| HRSV | 4,474 | 4,474 | 106 | 174 | 162 | 2.3 |
| HPV3 | 4,806 | 4,806 | 126 | 193 | 170 | 2.6 |
| Lake Victoria marburgvirus | 4,804 | 4,804 | 151 | 281 | 253 | 3.1 |
| Mumps virus | 4,780 | 4,929 | 142 | 214 | 204 | 2.9 |
| Measles virus | 4,931 | 5,157 | 158 | 257 | 244 | 3.2 |
| Zaire virus | 4,861 | 5,439 | 141 | 212 | 196 | 2.9 |
| Hendra virus | 5,204 | 6,002 | 126 | 219 | 203 | 2.4 |
| SARS-CoV | 9,759 | 14,131 | 299 | 790 | 699 | 3.0 |
| HHV-4 | 32,670 | 34,497 | 1,899 | 7,050 | 3433 | 5.8 |
| HHV-6 | 42,426 | 44,048 | 1,242 | 3,954 | 2236 | 2.9 |
| Variola virus | 53,044 | 53,107 | 1,638 | 2,760 | 1916 | 3.0 |
| HHV-5 | 63,216 | 64,140 | 3,176 | 16,745 | 6030 | 5.0 |
| All<sup>c</sup> | 284,918 | 298,966 | 10,863 | 259,655 | |

Human proteome formed by 36,103 proteins and 15,697,964 occurrences of 10,431,975 unique 7-mers. Column number refers to: (1) unique 7-mers in the viral proteome; (2) total number of 7-mers in the viral proteome (including multiple occurrences); (3) unique viral 7-mers occurring in the human proteome; (4) viral overlap occurrences in the human proteome (including multiple occurrences); (5) number of human proteins involved in overlap; (6) % of unique viral 7-mers which occur in the human proteome (i.e. 100 \times \text{column 3}/\text{column 1}).

<sup>a</sup> Abbreviations as in Table 1.

<sup>b</sup> The results of linear regression analysis between columns 1 and 3, and 1 and 4 are: column 3 = 0.042176/\text{column 1}/36.723 (r = 0.95637); column 4 = 0.17329/\text{column 1}/410.74 (r = 0.84397).

<sup>c</sup> Obtained by combining all 30 viral proteomes into one viral proteome, and then computing the overlap with the entire human proteome.
Table 8—Actual versus theoretical n-peptide occurrences in viral and human proteins

| Peptidea (n-mer) | Theoretical numberb (×10^3) | Actual occurrences |
|------------------|-----------------------------|--------------------|
|                  | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
| 5                | 3.2 | 257,035 | 299,701 | 42,666 | 254,808 | 299,749 | 44,941 | 42,647 | 56,230 | 234,691 | 2,907,096 |
| 6                | 64  | 282,932 | 298,966 | 16,034 | 286,515 | 299,058 | 12,543 | 4,394 | 5,089 | 87,224 | 259,655 |
| 7                | 1,280 | 284,918 | 298,232 | 13,314 | 289,592 | 298,367 | 8,775 | 455 | 617 | 10,863 | 31,737 |
| 8                | 25,600 | 284,857 | 297,498 | 12,641 | 289,744 | 297,676 | 7,932 | 80 | 163 | 1,668 | 8,805 |
| 9                | 512,000 | 284,497 | 296,764 | 12,267 | 289,486 | 296,985 | 7,499 | 30 | 81 | 656 | 4,035 |
| 10               | 10,240,000 | 284,022 | 296,030 | 12,008 | 289,096 | 296,294 | 7,198 | 14 | 27 | 437 | 1,311 |
| 11               | 204,800,000 | 283,494 | 295,296 | 11,802 | 288,629 | 295,603 | 6,974 | 8 | 9 | 352 | 661 |
| 12               | 4,096,000,000 | 282,926 | 294,562 | 11,636 | 288,121 | 294,912 | 6,701 | 5 | 5 | 292 | 449 |
| 13               | 81,920,000,000 | 282,341 | 293,829 | 11,488 | 287,578 | 293,223 | 6,445 | 3 | 3 | 240 | 316 |
| 14               | 1,638,400,000,000 | 281,735 | 293,096 | 11,361 | 287,017 | 293,535 | 6,518 | 1 | 1 | 204 | 254 |
| 15               | 32,768,000,000,000 | 281,116 | 292,363 | 11,247 | 286,441 | 292,847 | 6,406 | 0 | 0 | 171 | 207 |
| 16               | 655,360,000,000,000 | 280,478 | 291,630 | 11,152 | 285,842 | 292,160 | 6,318 | 0 | 0 | 148 | 176 |

Column number: (1) actual unique n-mers in the 30 viral proteomes; (2) actual n-mers in the 30 viral proteomes (including multiple occurrences); (3) number of repeated n-mers in the 30 viral proteomes; (4) actual unique n-mers in the 30 human sub-proteomes; (5) actual n-mers in the 30 human sub-proteomes (including multiple occurrences); (6) number of repeated n-mers in the 30 human sub-proteomes; (7) unique viral n-mers overlaps in the 30 human sub-proteomes; (8) total viral n-mers overlaps in the 30 human sub-proteomes (including multiple occurrences); (9) unique viral n-mers overlaps in the human proteome; (10) total viral n-mers overlaps in the human proteome (including multiple occurrences).

a Peptide length, with n from 5 to 16 amino acids.
b The number of possible amino acid combinations is given by 20^n.

4. Discussion

The difference between biological entities such as viruses and cellular organisms is quite obvious when the functional roles of the proteins found in viruses and in cellular organisms are compared [24]. On the other hand, according to the data we present here, this obvious difference disappears at the peptide phenetic level. The mathematical quantification of the widespread and high pentapeptide similarity between viral and human proteomes is surprising, raising fundamental questions. In particular, the ample viral-human peptide sharing is significant to (auto)immune phenomena as well as to evolutionary pathways.

Indeed, in the context of the physio(patho)logical relationships between viruses and humans, it is mandatory to observe that short-peptide motifs can exert a central role in cell adhesion, signal transduction, hormone activity, regulation of transcript expression, and enzyme activity [3,9,10,14,15,17]. Likewise, the antigen-antibody recognition process can reductionistically be circumscribed to interacting modules of five amino acid residues which, therefore, appear to be sufficient structural immunological determinants ([11–13,17] and Refs. therein). Consequently, we notice that the data shown here call into question the possibility of a direct causal association between virus–host sharing of amino acid motifs and incitement of autoimmune reactions [20]. Indeed, the molecular mimicry hypothesis suggests that, when bacterial/viral agents share epitopes with a host’s proteins, an immune response against the infectious agent may result in formation of cross-reacting antibodies that bind the shared epitopes on the normal cell and result in the auto-destruction of the cell. In the present case, the molecular mimicry hypothesis implies that viral infections should be a practically infinite source of autoimmune diseases since this study demonstrates that viral 5-mer matches are disseminated throughout practically all the human proteome and each viral match is repeated almost more than 10 times (see Table 4). Consequently, autoimmune diseases should theoretically approach a 100% real incidence, since the 30 viruses we examined practically are more or less disseminated throughout the entire human species. Taking HCV as an example, we note that the virus versus human overlap analysis in Table 4 produces the mathematical evidence that HCV (as well as HTLV-1 and Rubella virus) are even “more similar” to the human proteins than that expected by the linear regression relationship reported in Fig. 3. The data reinforce and add to our previous report on the sharing of numerous perfect exact matches between HCV and human proteomes [16]. Since 2760 out of 3003 HCV 5-mers occur in the human proteome, the 46,731 occurrences include multiple/repeated matches in the human proteins. More exactly, each HCV 5-mer that overlaps with the human proteome are repeated an average of 16.5 times. Moreover, the 46,731 occurrences occur in 20,269 different human proteins (Table 4). That means that almost 60% of the 36,103 human proteins forming the H. sapiens proteome contain viral HCV pentamers. This datum is of cogent interest in the present worldwide advance of HCV infection, since it further excludes sequence similarity as a possible mechanism in HCV-associated autoimmunity phenomena and, at the same time, might address research towards new unexplored scenarios. These observations, of course, apply to almost all the viruses listed in Table 1.

When analyzed in the evolutionary context, the viral and human protein sets studied here yield a closely congruent overlapping pattern to the human proteome at the 5-mer level as well as at higher n-mer levels, indicating a common compositional mosaicism in viral and human proteins. The absence of specific phenetic n-mer clusterings in the viral versus human proteome (and vice versa) suggests a synergetic platform of development of non-equilibrium systems which can be traced back to the Prigoginian model rather than to the...
Darwinian evolutionary scheme, where harmful protein sequences are “selectively removed” whereas neutral or useful peptides are left alone. De facto, it seems that viruses and mammalians have drawn and draw from a common ancient melting pot of short-peptide modules in building up their proteomes, either simple or complex. Indeed, it seems that human evolution has paid very little attention to a peptide motif being viral or not. The reconstruction of the informational network relating all living organisms might be helped by an integration with the present data which recall a time when the totality of life on Earth was a simple heterogeneous community that shared only the ability to synthesize proteins with a ribosomal machine and that freely exchanged genetic material [4].

REFERENCES

[1] Alter HJ. To C or not to C: these are the questions. Blood 1995;85:1681–95.
[2] Aragones JM, Bolibr I, Bonfill X, Bufill E, Mummany A, Alonso F, et al. Myasthenia gravis: a higher than expected incidence in the elderly. Neurology 2003;60:1024–6.
[3] Deroo S, El Kasmi KC, Fournier P, Theisen D, Brons NH, Herrmann M, et al. Enhanced antigenicity of a four-contact-residue epitope of the measles virus hemagglutinin protein by phage display libraries: evidence of a helical structure in the putative active site. Mol Immunol 1998;35:435–43.
[4] Doolittle RF. Evolutionary aspects of whole-genome biology. Curr Opin Struct Biol 2005;15:248–53.
[5] Gautam AM, Pearson CI, Smiley DE, Steinman L, McDevitt HO. A polyalanine peptide with only five native myelin basic protein residues induces autoimmune encephalomyelitis. J Exp Med 1992;176:605–9.
[6] Gusfield D. Algorithms on strings, trees, and sequences: computer science and computational biology. Cambridge University Press; 1997.
[7] Hemmer B, Kondo T, Gran B, Pinilla C, Cortese I, Pascal J, et al. Minimal peptide length requirements for (CD4+) T cell clones-implications for molecular mimicry and T cell survival. Int Immunol 2000;12:375–83.
[8] Jacobson DL, Gange SJ, Rose NR, Graham NM. Epidemiology and estimated population burden of selected autoimmune diseases in the United States. Clin Immunol Immunopathol 1997;84:223–43.
[9] Johansson J, Ekberg K, Shafqat J, Henriksson M, Chibalin A, Wahren J, et al. Molecular effects of proinsulin C-peptide. Biochem Biophys Res Commun 2002;295:1035–40.
[10] Kanduc D. Translational regulation of human papillomavirus type 16 E7 mRNA by the peptide SEQIKA, shared by rabbit (alpha1)-globin and human cytokeratin 7. J Virol 2002;76:7040–8.
[11] Kanduc D. Correlating low-similarity peptide sequences and allergic epitopes. Curr Pharm Des 2008;14:289–95.
[12] Kanduc D, Lucchese A, Mittelman A. Non-redundant peptidomes from DAPs: towards “the vaccine”? Autoimmun Rev 2007;6:290–4.
[13] Kanduc D. Immunogenicity in peptide-immunotherapy: from self/nonself to similar/dissimilar sequences. In: Sigalov A, editor. Multichain immune recognition receptor signaling: from spatiotemporal organization to human disease. Austin, TX: Landes Biosciences; 2008.
[14] Kapica M, Laubitz D, Puzio J, Jankowska A, Zabielski R. The ghrelin pentapeptide inhibits the secretion of pancreatic juice in rats. J Physiol Pharmacol 2006;57:691–700.
[15] Kudirka JC, Panupinthu N, Tesseym MA, Dixon SJ, Bernier SM. P2Y nucleotide receptor signaling through MAPK/ERK is regulated by extracellular matrix: involvement of beta3 integrins. J Cell Physiol 2007;213:54–64.
[16] Kusualik A, Bickis M, Lewis C, Li Y, Lucchese G, Marincola FM, et al. Widespread and ample peptide overlapping between HCV and Homo sapiens proteomes. Peptides 2007;28:1260–7.
[17] Lucchese G, Stufano A, Trost B, Kusualik A, Kanduc D. Peptidology: short amino acid modules in cell biology and immunology. Amino Acids 2007;33:703–7.
[18] Nimmi HL, Houghten RA, Walker LE, Reisfeld RA, Wilson IA, Hoge JM, et al. Generation of protein-reactive antibodies by short peptides is an event of high frequency: implications for the structural basis of immune recognition. Proc Natl Acad Sci USA 1983;80:4949–53.
[19] Oldstone MB. A suspenseful game of ‘hide and seek’ between virus and host. Nat Immunol 2007;8:325–7.
[20] Oldstone MB. Molecular mimicry and immune-mediated diseases. FASEB J 1998;12:1255–65.
[21] Reddehase MJ, Rothbard JB, Koszinowski UH. A pentapeptide as minimal antigenic determinant for MHC class I-restricted T lymphocytes. Nature 1989;337:651–3.
[22] Redelings MD, McCoy L, Sorvillo F. Multiple sclerosis mortality and patterns of comorbidity in the United States from 1990 to 2001. Neuroepidemiology 2006;26:102–7.
[23] Stevenson F, Natvig J. Autoantibodies revealed: the role of B-cells in autoimmune diseases. Immunol Today 1999;20:296–8.
[24] Van Regenmortel MH. Virus species and virus identification: past and current controversies. Infect Genet Evol 2007;7:133–44.