Parasitism as the main factor shaping peptide vocabularies in current organisms

MICHAELA ZEMKOVÁ[1], DANIEL ZAHRADNÍK[2], MARTIN MOKREJŠ[1] *and* JAROSLAV FLEGR[1]*

[1] *Faculty of Science, Department of Philosophy and History of Science, Charles University in Prague, Viničná 7, Prague, CZ-12844, Czech Republic*
[2] *Faculty of Forestry and Wood Sciences, Department of Forest Management, Czech University of Life Sciences Prague, Kamýcká 1176, Prague, CZ-165 21, Czech Republic*

(*Received 10 September 2016; revised 2 January 2017; accepted 21 January 2017; first published online 28 February 2017*)

SUMMARY

Self/non-self-discrimination by vertebrate immune systems is based on the recognition of the presence of peptides in proteins of a parasite that are not contained in the proteins of a host. Therefore, a reduction of the number of 'words' in its own peptide vocabulary could be an efficient evolutionary strategy of parasites for escaping recognition. Here, we compared peptide vocabularies of 30 endoparasitic and 17 free-living unicellular organisms and also eight multicellular parasitic and 16 multicellular free-living organisms. We found that both unicellular and multicellular parasites used a significantly lower number of different pentapeptides than free-living controls. Impoverished pentapeptide vocabularies in parasites were observed across all five clades that contain both the parasitic and free-living species. The effect of parasitism on a number of peptides used in an organism's proteins is larger than effects of all other studied factors, including the size of a proteome, the number of encoded proteins, etc. This decrease of pentapeptide diversity was partly compensated for by an increased number of hexapeptides. Our results support the hypothesis of parasitism-associated reduction of peptide vocabulary and suggest that T-cell receptors mostly recognize the five amino acids-long part of peptides that are presented in the groove of major histocompatibility complex molecules.

Key words: Peptide vocabulary usage, MHC-based recognition, immune evasion strategy, immunity, antigenic presentation, proteomics.

INTRODUCTION

Major histocompatibility complex (MHC)-based immunity recognition

The immune system recognizes the presence of proteins of foreign origin by the occurrence of peptides that are not present in a host's own proteins.

As a part of the host-parasite evolutionary arms race, a parasite could decrease the probability of its recognition by reducing the number of different peptides ('words') in its vocabulary (vocabulary reduction), and by mimetizing the peptide vocabulary of its host (vocabulary mimicry) (Flegr, 2011). It could therefore be expected that parasitic organisms will have a lower number of different peptides in their proteome than free-living organisms.

Almost all cells in vertebrate bodies (except, e.g. for sperm and trophoblast) (King *et al*. 2006) present fragments of proteins, i.e. short peptides, on their surface (Lanzavecchia, 1985). These peptides are captured in the grooves of MHC class I molecules on the surface of somatic cells and MHC class II molecules on the surface of specialized antigen-presenting cells (APC). The peptides can be recognized as non-self by the T-cells that carry molecules of a T-cell receptor with matching specificity. Each young T-cell carries one type of T-cell receptor with an affinity toward certain self or non-self-peptide. The population of T-cells is subjected to negative selection in a thymus. In this process, all T-cells carrying a receptor that recognizes any peptide presented in the thymus die or are functionally incapacitated. Therefore, when a mature T-cell recognizes a peptide outside the thymus, it is most probably a non-self-peptide that originated from the proteins of a parasitic organism. Such proteins are either synthesized in a particular host cell (typically the peptides of viral origin presented on MHC class I molecule) or originated from the proteins of parasites engulfed by APC (typically the peptides presented on MHC class II molecule) (Craiu *et al*. 1997; Trombetta and Mellman, 2005; Neefjes and Ovaa, 2013).

The mechanism of MHC-based self/non-self-discrimination suggests that the number of different peptides in a vocabulary, i.e. the number of potential targets for T-cell recognition, is a critical parameter in the host–parasite arms race and therefore also an important object of natural selection in parasitic organisms. The standardized size of any vocabulary can be expressed as the vocabulary usage, the ratio of the actual vocabulary size (the number of all

* Corresponding author: Division of Biology, Faculty of Science, Charles University in Prague, Vinicna 7, 128 44, Prague, Czech Republic. E-mail:flegr@cesnet.cz

Parasitology (2017), **144**, 975–983.
doi:10.1017/S0031182017000191

different words) to the maximal combinatorially possible vocabulary size (Popov *et al*. 1996; Orlov and Potapov, 2004).

Here we searched for indices of reduced peptide vocabulary in parasites by comparing the vocabulary usage of proteomes of 38 endoparasites (eight multicellular parasitic helminths, 30 unicellular protozoan parasites) with 33 free-living eukaryotic organisms.

METHODS

Organisms

Predicted proteomes – whole sets of proteins of given organisms – were obtained from the NCBI GenBank database and from the Sanger Institute. To provide sufficient length for a peptide vocabulary usage assay, only the organisms with proteome larger than 1 MB (size of a briefly annotated FASTA formatted file) were included in the study. The complete list of species is available in Supplementary Table S1 (Adl *et al*. 2012; Diamond and Clark, 1993; Elliott, 1973; Hamzah *et al*. 2006; Tyler and Engman, 2001).

This study was strongly limited by the availability of proteomes of sufficient size. Some desirable proteomes are not presently available or they are too short to be included in the analysis. Also, we tried to include only those species that are unambiguously parasitic or free-living. Similar problem with the classification of organisms arises with uncertainties concerning the cellularity of organisms. We classified colonial *Volvox carteri* or cellular slime moulds forming *Dictyostelium discoideum* and *Polysphondylium pallidum* as unicellular. We did not include prokaryotic organisms in this study due to their completely different status in evolution and different genomic structure as compared with eukaryotes. We also excluded fungi species because most of them have either short proteomes or are difficult to classify according to our criteria of parasitism (such as potato late blight *Phytophthora infestans* and other plant pathogens, and entomopathogenic ectoparasites such as genus *Metarhyzium*).

Data filtration and standardization

Groups of proteins of common origin differing from each other in only a limited number of amino acids (paralogues) are present in all eukaryotic proteomes. Only one representative of such protein family was retained and all others were excluded from analysed proteome to avoid an artificial decrease of vocabulary size in homologs and paralogues-rich proteomes during our data sampling step, see below. Similarly, comments, annotations, and special characters occurring in sequences (coding unknown amino acids, gaps, etc.) were filtered out during the loading procedure. Although they occur only rarely in the proteomes, they would cause a pronounced artificial inflation of alphabet size – the parameter that has a substantial effect on the size of vocabulary usage.

To perform a particular proteome filtration, proteins were read one by one from the input text file. Our computer program randomly selected k peptides of length n from each input protein and compared these peptides with all previously read proteins. If at least one matching peptide was found, then the protein was considered as a homolog or paralogue of a previously read protein and was excluded from the filtrated proteome. The default length of compared peptides (n) was set to 16 and the number of selected samples per protein (k) was set to 5 for most organisms. Organisms with many homologues and paralogues, such as plants, required a higher k – up to 20; otherwise the filtration was not strict enough. It was possible to directly verify the correct parameter settings by visual inspection of the graphical representation of vocabulary usage, as the vocabulary usage of 16-peptides (or longer k-mers as explained above) approached 1 in non-redundant (sufficiently filtrated) proteomes.

Data analysis

Theoretical background. Methods based on counting oligomers in nucleotide or amino acid sequences are known as 'linguistic-like' (Bolshoy, 2003). Linguistic-like tools are built on classical Shannon's technique of n-gram text decomposition, where n-gram is a word of length n. Main inputs are the length of sequence, the size of the alphabet and the length of n-grams (oligomers) (Volkovich *et al*. 2005). The sequence is split into the list of n-grams (words) of given length (called vocabulary). This method was first used by Beckmann *et al*. (1986) for searching for species-specific 'genomic signatures'. Gatherer (2007) improved this method and demonstrated the existence of such peptide vocabulary signatures. Using linguistic techniques, Pietrokovski and Trifonov (1992) identified presence of sequences of foreign origin in the yeast mitochondrial genome. Motomura *et al*. (2012) used analogy between zipf's-like distribution of words in English and that of short oligopeptides in proteins. They detected a fraction of frequently used sequences and suggested possible functional importance of certain 'short constituents amino acid sequences' (SCSs). Interestingly, these SCSs are tetra, penta and hexamers, which are oligomers of the same lengths, as we found in our study to be important for MHC recognition, see below. There are attempts to show further analogies between human texts and genetic sequences (Popov *et al*. 1996; Gimona, 2006; Eroglu, 2014; Zemkova *et al*. 2014), since both kinds of 'texts' are built from

https://doi.org/10.1017/S0031182017000191

discrete units from defined alphabets. The linguistic methods are often used to study a complexity, e.g. the richness (the diversity) of the vocabulary of particular genomic sequences. Some of such algorithms, for example the Wooton–Federhen index, are implemented and widely used bioinformatics tools, e.g. in BLAST (Wootton and Federhen, 1993; Sharon *et al*. 2005).

Data analysis. Since the peptides are trimmed randomly without any relation to the peptide function, we used concept of simple vocabulary usage (as defined further below in equation (1)) to measure the diversity of oligopeptides of particular organisms of length n ranging from 4 to 12 among the random samples of 1 000 000 n-length peptides from each proteome. Thus the final size of compared proteomes was the same. The upper bound (12 amino acids) reflects the usual length of trimmed peptides that are loaded to MHC molecules (Trombetta and Mellman, 2005).

The vocabulary usage U_n of a given organisms is defined as the ratio of the actual $U_{n,\mathrm{a}}$ vocabulary size (the number of different peptides) to the maximal combinatorially possible vocabulary size $U_{n,\max}$ for peptide length n.

$$U_n = U_{n,\mathrm{a}}/U_{n,\max}, \tag{1}$$

The theoretical number of combinatorially possible peptides of length n ($U_{n,\max}$) was computed as follows:

$$U_{n,\max} = \min(1\,000\,000, s^n) \tag{2}$$

where n is the peptide length, and s is the alphabet size, i.e. 20 for amino acid alphabet.

Computation was done first for data without filtration (containing paralogues) and then for filtered data.

Principal component analysis (PCA) from a covariance matrix (unrotated) (Rencher, 2002) was used to reduce the number of nine dependent variables (vocabulary usage indexes for length of words from 4 to 12 amino acids) to four independent principal components (PCs) – each explaining more than 1% of variability. Correlations of these factors with focal binary variables were computed with analysis of covariance (ANCOVAs) (type III. sum of squares). Four proteome characteristics (length of proteome before filtration, length of filtered proteome, number of proteins – unfiltered and number of proteins – filtered) were included in the models as covariates. Because of the nested character of data (e.g., no parasitic autotroph exists in our dataset), we could not include all focal variables and all organisms into one complex model. Instead, we first computed a basic model containing only confounding variables and then subtracted the amount of variance in vocabulary usage explained by this model from the amount of variance explained by models containing the confounding variables and one focal binary variable: namely parasitism, unicellular parasitism, multicellular parasitism, multicellularity and heterotrophy, respectively. Exact variant of binomial sign test for dependent samples (Sheskin, 2003) was used to search for the overrepresentation of clades in which the parasitic species had impoverished peptide vocabulary in comparison with the free living species. For all statistical computations, standard Base-package of R software was used.

Code availability Our software: 'Complexity G' is available at figshare https://figshare.com/articles/Raw_proteomic_data_for_analysis_of_peptide_vocabulary_usage/3491057

RESULTS

Vocabulary usages for peptides of lengths from 4 to 12 amino acids (U_4–U_{12}) were computed for 71 proteomes of different organisms (see Supplementary Table S1). Individual values of U_4–U_{12} of all proteomes are listed in the Supplementary Table S2. As nine variables U_n were highly correlated, we used the method of PCA to reduce the number of variables and to obtain independent composite variables – the PCs. The first four PC had eigenvalues higher than 1 and explained 99·9% of the variability in vocabulary usage (Fig. 1). The first two principal components (PC1 and PC2) were loaded mostly by pentapeptide and hexapeptide vocabularies (U_5 and U_6). PC1 was negatively loaded by pentapeptides and positively by hexapepeptides, while PC2 was negatively loaded by all types of peptides, except tetrapeptides. PC3 was positively loaded by hexapeptides and negatively by tetrapetides, and peptides longer than seven amino acids (U_7–U_{12}). PC4 was positively loaded primarily by tetrapetides, and also partly by hexapeptides.

Vocabulary usage can be influenced by parasitism and also by various non-ecological factors, such as the complexity of an organism, genome redundancy, etc. Therefore, we used simple multivariate ANCOVAs to find out which parts of interspecies variability in vocabulary usage (independent variables PC1–PC4) could be explained by four factors that reflect the size and redundancy of proteomes (size of proteome, size of non-redundant part of proteome, number of proteins in whole proteome, and number of proteins in non-redundant part of proteome), and which parts could be explained by parasitism or other binary factors. Because of the nested character of the data, a separate multivariate analysis was performed for each binary variable of interest, namely for parasitism (parasites *vs* free-living organisms), unicellular parasitism (unicellular parasites *vs* unicellular free-living organisms),

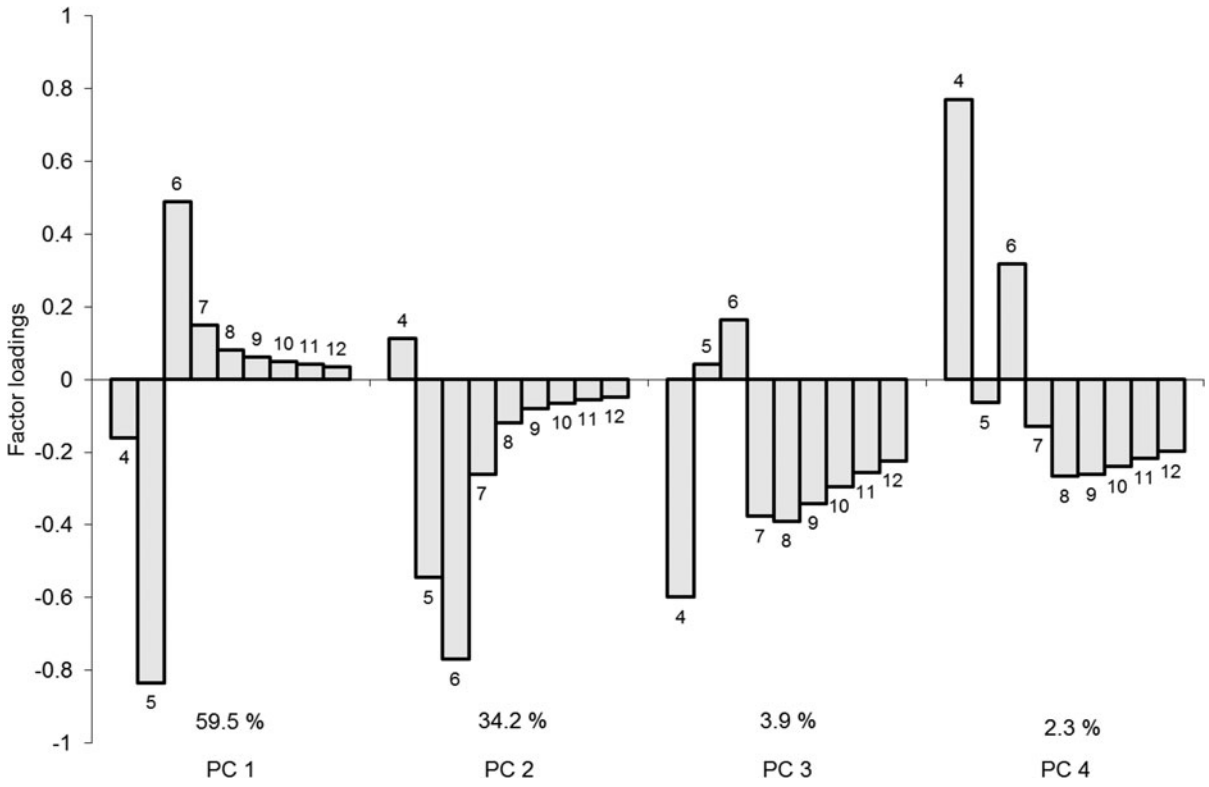

Fig. 1. Four principal components explain 99·9% of interspecies variability in peptide vocabulary usage. The figure shows particular factor loadings and corresponding percentages of explained interspecies variability in vocabulary usage. Column labels 4–12 indicate the length of the peptides, which load the particular principle component.

multicellular parasitism (multicellular parasites *vs* multicellular free-living organisms), multicellularity (unicellular *vs* multicellular organisms), endoparasitism (extracellular parasites *vs* intracellular parasites) and heterotrophy (heterotrophs *vs* autotrophs); see Tables 1 and 2.

The strongest and the most significant effect observed in our data was a positive effect of parasitism on the PC1, i.e. on the variable explaining the largest part of interspecies variability in the vocabulary usage. Parasites had relatively impoverished pentapeptides and tetrapeptides, and relatively enriched hexapeptide vocabularies. Parasitism explained nearly 6·5% of this variability, while, for example, all four variables describing size and redundancy of proteome together explained less than 4% of this variability (Supplementary Table S3). The main factor influencing PC2 was the unicellularity/multicellularity of organisms. Unicellular organisms had relatively impoverished vocabularies (especially the pentapeptides and hexapeptides vocabularies), except the tetrapeptides vocabulary. PC3 was negatively influenced by parasitism; the parasites had relatively impoverished pentapeptides and hexapeptides vocabularies and enriched all other vocabularies. Extracellular parasites (both unicellular and multicellular) had generally higher values of PC4, i.e. they had relatively enriched tetrapeptide and hexapeptide vocabularies and relatively impoverished other vocabularies (in comparison with intracellular parasites, not with free living organisms). The contribution of particular effects to each PC is summarized in Table 1.

The distribution of parasitic and free-living organisms within the two-dimensional (2D) space as defined by two PCs correlated with parasitism (PC1 and PC3) is shown in the Fig. 2. Parasitic organisms are clustered on the right side of the graph. Multicellular free-living and parasitic organisms are shifted left-down in comparison with unicellular free-living and parasitic organisms. The 2D space defined by PC1 (correlated with parasitism) and PC2 (correlated with multicellularity) is shown in Supplementary Fig. S4 and the 2D space defined by PC1 (correlated with parasitism) and PC4 (correlated with intracellularity) is shown in Supplementary Fig. S5.

Because of the existence of a phylogenetic relation between the analysed organisms, the results of our statistical analysis could be influenced by the effect of pseudoreplications. To eliminate this effect, we used an exact variant of binomial sign test for dependent samples (Sheskin, 2003) to search for the overrepresentation of clades in which the parasitic species had impoverished peptide vocabulary in comparison with the free-living species. Only five clades contained both parasitic and free-living organisms (Kinetoplastids, Ciliates, Nematodes, Opisthokonts and SAR). Within all five pairs, the mean value of PC1 was higher for the parasitic

https://doi.org/10.1017/S0031182017000191

Table 1. Effects of parasitism, form of parasitism, multicellularity and heterotrophy on peptide vocabulary usage

	PC1 (59·5%)			PC2 (34·2%)			PC3 (3·9%)			PC4 (2·3%)		
	$\downarrow U_5 \uparrow U_6$			$\downarrow U_5 \downarrow U_6$			$\downarrow U_{4,7-12} \uparrow U_6$			$\uparrow U_4, \uparrow U_6, \downarrow U_{7-12}$		
	Beta	%	*P*-value	Beta	%	*P*-value	Beta	%	*P*-value	Beta	%	*P*-value
Parasitism–non-parasitism	0·0661	6·5	0·000	0·0332	2·8	0·137	−0·0158	5·7	0·041	0·0077	2·3	0·196
Unicellular parasitism–Unicell. non-parasitism	0·0567	5·7	0·007	0·0345	2·4	0·220	−0·0038	0·3	0·680	0·0095	2·7	0·260
Multicell. par.–Multicell. non-parasitism	0·0627	5·3	0·001	−0·0190	1·6	0·545	−0·0296	16·4	0·057	0·0081	6·2	0·190
Unicellularity–Multicellularity	0·0409	2·6	0·008	0·0761	15·6	0·000	−0·0012	0·0	0·877	−0·0108	4·6	0·064
Unicell. Parasitism–Multicell. parasitism	0·0481	6·5	0·105	0·1237	17·7	0·002	0·0044	0·2	0·759	0·0029	0·1	0·825
Unicell. non-parasitism–Multicell. non-parasitism	0·0412	3·8	0·025	0·0709	14·5	0·022	−0·0076	1·5	0·493	−0·0125	9·5	0·076
Intracelullar parasitism–Extracell. parasitism	−0·0121	0·7	0·585	0·0207	0·8	0·495	0·0115	2·2	0·275	−0·0282	17·4	0·004
Heterotrophy–Autotrophy	0·0217	0·7	0·189	−0·0374	3·4	0·103	−0·0092	1·8	0·255	0·0165	9·8	0·006

The table summarizes the results of analyses of 32 simple multivariate ANCOVA models with five independent variables: size of proteome, size of non-redundant part of proteome, number of proteins in proteome, number of proteins in non-redundant part of proteome and one of focal binary variables listed in the first column. The columns 2–4, 5–7, 8–10 and 11–13 show results (slope beta, % of explained variability, and significance of two-sided test) for four dependent variables, namely (PC1–4). Positive beta value means that the first group of organisms of the compared pair has a higher particular PC_n value than the second group of organisms. For example, in the first row, parasites have significantly higher PC1 values than free-living organisms. Signs of correlation of PC1–4 with vocabulary usage are indicated with an arrow in the legend of each principle component, for details see Fig. 1.

Table 2. Effects of proteome size on peptide vocabulary usage

	PC1			PC2			PC3			PC4		
Length factors	Beta	%	*P*-value	Beta	%	*P*-value	Beta	%	*P*-value	Beta	%	*P*-value
Length of unfiltered proteome	$5{\cdot}29 \times 10^{-9}$	0·2	0·436	$3{\cdot}18 \times 10^{-9}$	0·1	0·737	$-4{\cdot}89 \times 10^{-9}$	3·0	0·144	$1{\cdot}02 \times 10^{-9}$	0·2	0·688
Length of filtered proteome	$-3{\cdot}52 \times 10^{-8}$	3·9	**0·002**	$1{\cdot}12 \times 10^{-8}$	0·7	0·465	$4{\cdot}11 \times 10^{-9}$	0·8	0·444	$-5{\cdot}12 \times 10^{-9}$	2·1	0·212
Number of proteins unfiltered data	$-2{\cdot}75 \times 10^{-6}$	0·3	0·367	$-1{\cdot}13 \times 10^{-6}$	0·1	0·790	$2{\cdot}34 \times 10^{-6}$	3·4	0·119	$-3{\cdot}31 \times 10^{-7}$	0·1	0·770
Number of proteins filtered data	$3{\cdot}06 \times 10^{-6}$	0·1	0·544	$-1{\cdot}92 \times 10^{-6}$	0·1	0·786	$-2{\cdot}61 \times 10^{-6}$	1·6	0·292	$2{\cdot}55 \times 10^{-6}$	2·5	0·178

Table summarizes the results of analyses of four simple multivariate ANCOVA models containing all four independent variables (column 1) and one dependent variable, i.e. the component PC1, PC2, PC3 or PC4. The beta value computed by ANCOVA indicates size and direction of the effects of four parameters characterizing the size of the proteome on particular principal components. *P*-value is a two-sided statistical significance. Significant *P*-values are printed in bold.

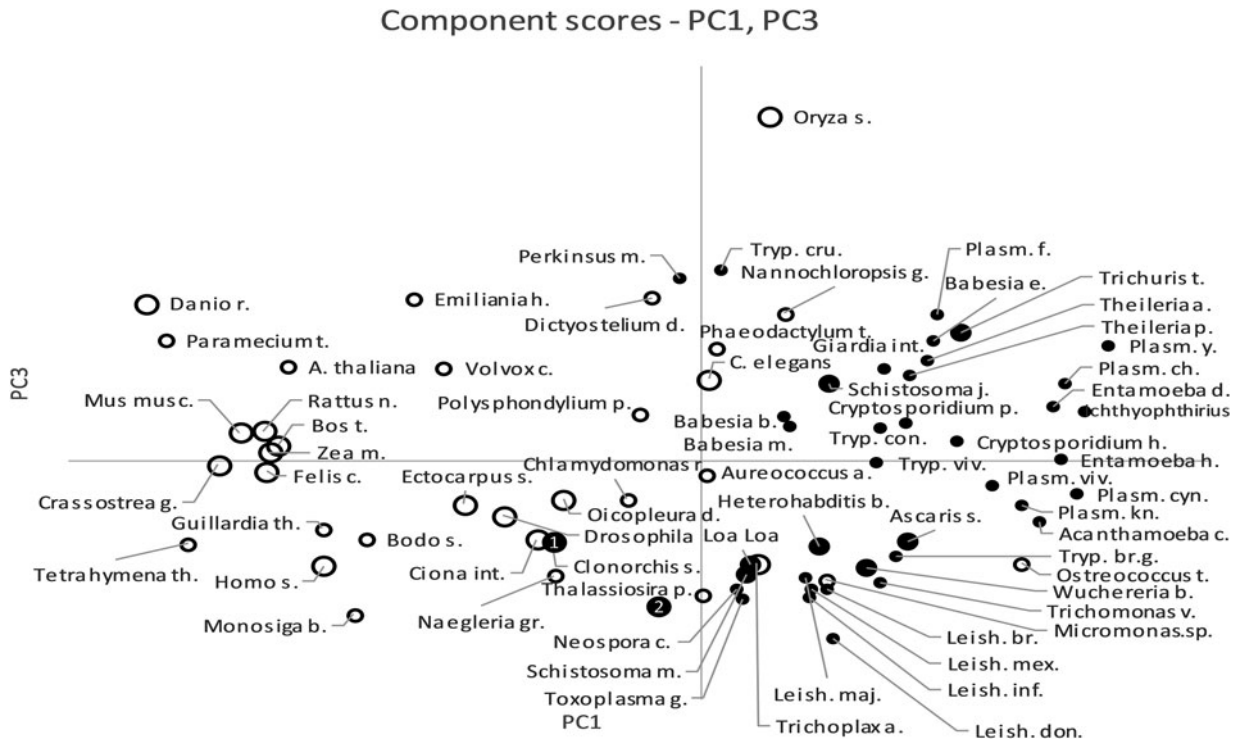

Fig. 2. Effect of parasitism on peptide vocabulary usage. Two-dimensional space defined by two principal components correlated with parasitism (PC1 - horizontal axis and PC3 - vertical axis). Dark and white circles denote the position of parasitic and free-living organisms, respectively. Larger circles indicate multicellular organisms. We used two proteome datasets for parasitic trematode *Clonorchis sinensis* – number 1 corresponds to a proteome derived from a set of genome-based proteins (assembly v2·0 from NCBI Genomes) whereas the number 2 corresponds to a transcriptome-based predicted proteome (from HelmDB).

than for the free-living organisms ($P = 0{\cdot}031$). No correlation existed between PC1 and PC3. Therefore, it was also possible to compare the mean PC3 of parasitic and free-living organisms within the same five clades. Again, the mean value of PC3 was lower for parasitic organisms in all five clades. Global tests for all ten pairs (Table 3) showed highly significant ($P < 0{\cdot}001$) support for our hypothesis of lower peptide vocabulary diversity (namely higher PC1 and lower PC3) in parasitic organisms.

DISCUSSION

Our results suggest that the number of different peptides per proteome could reflect an ecological strategy of these species, namely the difference between parasitic and non-parasitic organisms. Nearly all variability (99·9%) of vocabulary usage in eukaryotic organisms was explained by four PCs. PC1, the factor explaining nearly 60% of interspecies variability (59·5%), was influenced most strongly by the

https://doi.org/10.1017/S0031182017000191

Table 3. Comparison of five clades containing both-parasitic and free-living organisms for values of PC1 and PC3

	Opisthokonts		Kinetoplastids		Ciliates		Nematods		Sar	
Clade	Parasites	Free-living	Parasites	Free-living	Parasites	Free-living	Parasites	Free-living	Parasites	Free-living
PC1	0·0404	−0·1190	0·0501	−0·1324	0·1508	−0·2075	0·0193	0·0027	0·0864	0·0271
PC3	−0·0114	−0·0015	−0·0156	−0·0197	0·0126	0·0050	−0·0263	0·0209	0·0098	−0·0051

Average values of PC1 and PC3; *P*-value for exact binomial test are 0·031 for 5 pairs and 0·0009 for ten compared pairs.

effect of parasitism and less strongly by the length of a filtered proteome, i.e. size of non-redundant part of the proteome. PC2 was influenced by the unicellularity/multicellularity of the organism, PC3 was influenced by parasitism in multicellular organisms and PC4 reflected differences between extracellular and intracellular parasites.

The final richness of any peptide vocabulary is the result of several independent forces which differ in strength (reflected by amount of explained variability) and direction. Some of them, like the strong primary effect of parasitism (PC1) had positive influence on the size of the hexapeptide vocabulary, while another, like weaker effect unicellularity (PC2), and much weaker secondary effect of parasitism (PC3) had negative effect on the size of the hexapeptide vocabulary. It is possible, that the PC3 could reflect an existence of some mechanism which, in parasites, could increase size of the longer peptides vocabularies in order to partly compensate the reduction of shorter peptides vocabularies.

From all factors studied, including the lengths of proteomes, parasitism had the strongest effect on peptide vocabulary usage. We detected this effect independently in two sets of organisms, multicellular organisms and unicellular organisms. The results suggest that parasites have lower diversity of pentapeptides, which is partly compensated for by higher diversity of hexapeptides. It can be hypothesized that T-cells recognize peptides of five aminoacid residues in length when attached in the groove of MHC I protein. The length of trimmed peptides, which are loaded onto MHC I protein, is about 8–10 amino acids and the length of those loaded on MHC II is even higher. However, it was experimentally shown that only the residues at the top of the binding groove are recognized by T-cell receptors while those at bottom of the groove are used to bind the peptide to the groove of MHC protein (Vyas *et al*. 2008). Peptides usually contain only 2 or 3 amino acids that are critical for T-cell recognition; however, to trigger the response of the T-cell receptor the peptides must be longer by at least one or two additional amino acid residues (Vyas *et al*. 2008). This agrees with our observation that parasites have the most strongly impoverished pentapeptide and partly impoverished tetrapeptide vocabulary. It is highly probable that for the preservation of functionality of proteins some minimal 'linguistic' complexity is required. Therefore, reduction at the level of pentapeptides and tetrapeptides should probably be compensated for by a richer hexapeptide vocabulary.

The length of actually recognized parts of peptides on MHC I is lower than on MHC II (Vyas *et al*. 2008). The MHC I and II present mostly peptides from intracellular and extracellular parasites, respectively. Therefore, we could expect that the type of parasitism (intracellular *vs* extracellular) should affect the vocabulary usage – the intracellular parasites should have more impoverished shorter peptides-vocabularies than longer peptides-vocabularies. Indeed, we observed the effect of this type of parasitism on PC4, i.e. on the factor loaded mostly by high values of tetrapeptides (Supplementary Fig. S5). The effect of intracellularity/extracellularity on vocabulary usage was relatively weak. It must be noted, however, that some extracellular antigens can also be presented through the MHC I pathway (in a process known as cross-presentation) (Paz *et al*. 1999; Trombetta and Mellman, 2005) and that many seemingly intracellular parasites (such as representatives of the phylum Apicomplexa) in fact occupy an interior of parasitophorous vacuole – the organelle in some respects homologous to an extracellular, rather than intracellular, compartment (Lingelbach and Joiner, 1998).

Differences between parasites and free-living organisms are clearly visible from the figures of component scores. Here the parasites are aggregated in a region of positive PC1 values, while the free-living organisms are clustered in the region of negative PC1 values. There are some interesting exceptions to this trend. A rich peptide vocabulary (including pentapeptides) of *Perkinsus marinus* can be explained by the fact that its host (oyster) does not possess an MHC-based immune system. Although free-living, the *Ostreococcus tauri* has a highly reduced, parasite-like, peptide vocabulary. This tiny green alga is the smallest and the most reduced autotrophic eukaryote in our dataset, so the reduction of its vocabulary could be related to its extreme simplicity.

https://doi.org/10.1017/S0031182017000191

We have no explanation for the non-parasite-like (overly rich) vocabulary of the trematode *Clonorchis sinensis*, except possible undetected contamination by genes from cat liver tissue or from any sample processed in the respective sequencing laboratory (Wang *et al*. 2011). When purely transcriptome-based predicted proteins from this organism (Yang and Wang, 2013) were analysed (downloaded from HelmDB), the position of *C. sinensis* in the 2D space of PC1 and PC3 moved towards the cluster of parasitic organisms (Fig. 2).

Though it was not a subject of the present study, we detected the effect of multicellularity on the second strongest PC (Supplementary Fig. S4). Multicellular organisms, both parasitic and free-living, have relatively richer hexapeptide and pentapeptide vocabularies, which could be an effect of the higher complexity of multicellular organisms. It must be noted, however, that only representatives of three phyla of multicellular organisms (Metazoa, Metaphyta and Charophyta) were included in this analysis. Therefore, this result may be biased by the effect of pseudoreplications.

Four independent lines of evidence, namely impoverished vocabulary in unicellular parasites, multicellular parasites, results of a phylogenetic contrast test performed on five pairs of sister taxa, and the fact that *Perkinsus* (one of the two parasites of hosts lacking MHC in the analysed dataset) has unreduced peptide dictionary are in an agreement with our *a priori* hypothesis about the reduced peptide vocabulary of parasitic organisms. Other explanations of the observed pattern, for example the theoretical possibility of the reduction of non-housekeeping proteins in parasites, are of course also legitimate and should be tested when necessary proteomes become available. The results also suggest that T-cells recognize MHC-attached peptides of lengths 4–5 amino acids, which could possibly be of importance in vaccine construction. Most of between-species variability of vocabulary usage is among 4–6 amino acids long peptides. This corresponds to length of basic functional units, 'words', described by Motomura's hypothesis of SCSs (Motomura *et al*. 2012). Our analysis included all proteomes larger than 1·2 MB, which were available by May 2015. It would be possible to reproduce our findings in future, with newly appearing proteomes as additional independent datasets. Similarly, it will be possible to use the developed software for testing the related peptide vocabulary mimicry hypothesis by studying similarities of peptide vocabularies between parasites and their specific hosts.

SUPPLEMENTARY MATERIAL

The supplementary material for this article can be found at https://doi.org/10.1017/S00311820170 00191.

ACKNOWLEDGMENTS

We thank to Fatima Cvrčková, Ivan Čepička, Vojtěch Žárský, Jaroslav Kulda, Julie Novakova, and Charlie Lotterman for their advices, suggestions and help.

FINANCIAL SUPPORT

The work was supported by Project UNCE 204004 (Charles University in Prague).

REFERENCES

Adl, S. M., Simpson, A. G. B., Lane, C. E., Lukes, J., Bass, D., Bowser, S. S., Brown, M. W., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., le Gall, L., Lynn, D. H., McManus, H., Mitchell, E. A. D., Mozley-Stanridge, S. E., Parfrey, L. W., Pawlowski, J., Rueckert, S., Shadwick, L., Schoch, C. L., Smirnov, A. and Spiegel, F. W. (2012). The revised classification of eukaryotes. *Journal of Eukaryotic Microbiology*, **59**, 429–493. doi: 10.1111/j.1550-7408.2012.00644.x.

Beckmann, J. S., Brendel, V. and Trifonov, E. N. (1986). Intervening sequences exhibit distinct vocabulary. *Journal of Biomolecular Structure and Dynamics* **4**, 391–400.

Bolshoy, A. (2003). DNA sequence analysis linguistic tools: contrast vocabularies, compositional spectra and linguistic complexity. *Applied Bioinformatics* **2**, 103–112.

Craiu, A., Akoplan, T., Goldberg, A. and Rock, K. L. (1997). Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 10850–10855. doi: 10.1073/pnas.94.20.10850.

Diamond, L. S. and Clark, C. G. (1993). A redescription of *Entamoeba histolytica* Schaudinn, 1903 (Emended Walker, 1911) Separating it from *Entamoeba dispar* Brumpt, 1925. *Journal of Eukaryotic Microbiology* **40**, 340–344. doi: 10.1111/j.1550-7408.1993.tb04926.x.

Elliott, A. M. (1973). *Biology of Tetrahymena*. Hutchinson & Ross, Dowden.

Eroglu, S. (2014). Language-like behavior of protein length distribution in proteomes. *Complexity* **20**, 12–21. doi: 10.1002/cplx.21498.

Flegr, J. (2011). *Pozor, Toxo! Tajná učebnice praktické metodologie vědy (Watch out for Toxo! The secret guide to practical science)*. Academia, Prague.

Gatherer, D. (2007). Peptide vocabulary analysis reveals ultraconservation and homonymity in protein sequences. *Bioinformatics and Biology Insights* **1**, 129–137.

Gimona, M. (2006). Protein linguistics – a grammar for modular protein assembly? *Nature Reviews Molecular Cell Biology* **7**, 68–73. doi: 10.1038/nrm1785.

Hamzah, Z., Petmitr, S., Mungthin, M., Leelayoova, S. and Chavalitshewinkoon-Petmitr, P. (2006). Differential detection of *Entamoeba histolytica, Entamoeba dispar*, and *Entamoeba moshkovskii* by a single-round PCR assay. *Journal of Clinical Microbiology* **44**, 3196–3200. doi: 10.1128/Jcm.00778-06.

King, R. C., Stansfield, W. D. and Mulligan, P. K. (2006). *A Dictionary of Genetics*. Oxford University Press, New York.

Lanzavecchia, A. (1985). Antigen-specific interaction between T-cells and B-cells. *Nature* **314**, 537–539.

Lingelbach, K. and Joiner, K. A. (1998). The parasitophorous vacuole membrane surrounding *Plasmodium* and *Toxoplasma*: an unusual compartment in infected cells. *Journal of Cell Science* **111**, 1467–1475.

Motomura, K., Fujita, T., Tsutsumi, M., Kikuzato, S., Nakamura, M. and Otaki, J. M. (2012). Word decoding of protein amino acid sequences with availability analysis: a linguistic approach. *PLoS ONE* **7**, 1–15. doi: ARTN e5003910.1371/journal.pone.0050039.

Neefjes, J. and Ovaa, H. (2013). A peptide's perspective on antigen presentation to the immune system. *Nature Chemical Biology* **9**, 769–775. doi: 10.1038/Nchembio.1391.

Orlov, Y. L. and Potapov, V. N. (2004). Complexity: an internet resource for analysis of DNA sequence complexity. *Nucleic Acids Research* **32**, W628–W633. doi: 10.1093/nar/gkh466.

Paz, P., Brouwenstijn, N., Perry, R. and Shastri, N. (1999). Discrete proteolytic intermediates in the MHC class I antigen processing pathway

and MHC I-dependent peptide trimming in the ER. *Immunity* **11**, 241–251. doi: 10.1016/S1074-7613(00)80099-0.

Pietrokovski, S. and Trifonov, E. N. (1992). Imported sequences in the mitochondrial yeast genome identified by nucleotide linguistics. *Gene* **122**, 129–137. doi: 10.1016/0378-1119(92)90040-V.

Popov, O., Segal, D. M. and Trifonov, E. N. (1996). Linguistic complexity of protein sequences as compared to texts of human languages. *BioSystens* **38**, 65–74.

Rencher, A. C. (2002). *Methods of Multivariate Analysis*, pp. 380–408. Wiley, New York.

Sharon, I., Birkland, A., Chang, K., El-Yaniv, R. and Yona, G. (2005). Correcting BLAST e-values for low-complexity segments. *Journal of Computational Biology* **12**, 980–1003. doi: 10.1089/cmb.2005.12.980.

Sheskin, D. J. (2003). *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC Press, Boca Raton, FL, USA.

Trombetta, E. S. and Mellman, I. (2005). Cell biology of antigen processing *in vitro* and *in vivo*. *Annual Review of Immunology* **23**, 975–1028. doi: 10.1146/annurev.immunol.22.012703.104538.

Tyler, K. M. and Engman, D. M. (2001). The life cycle of *Trypanosoma cruzi* revisited. *International Journal for Parasitology* **31**, 472–481. doi: 10.1016/S0020-7519(01)00153-9.

Volkovich, Z., Kirzhner, V., Bolshoy, A., Nevo, E. and Korol, A. (2005). The method of N-grams in large-scale clustering of DNA texts. *Pattern Recognition* **38**, 1902–1912. doi: 10.1016/j.patcog.2005.05.002.

Vyas, J. M., Van der Veen, A. G. and Ploegh, H. L. (2008). The known unknowns of antigen processing and presentation. *Nature Reviews Immunology* **8**, 607–618. doi: 10.1038/nri2368.

Wang, X. Y., Chen, W. J., Huang, Y., Sun, J. F., Men, J. T., Liu, H. L., Luo, F., Guo, L., Lv, X. L., Deng, C. H., Zhou, C. H., Fan, Y. X., Li, X. R., Huang, L. S., Hu, Y., Liang, C., Hu, X. C., Xu, J. and Yu, X. B. (2011). The draft genome of the carcinogenic human liver fluke *Clonorchis sinensis*. *Genome Biology* **12**, 1–14. doi: Artn R10710.1186/Gb-2011-12-10-R107.

Wootton, J. C. and Federhen, S. (1993). Statistics of local complexity in amino-acid-sequences and sequence databases. *Computers & Chemistry* **17**, 149–163. doi: 10.1016/0097-8485(93)85006-X.

Yang, X. W. and Wang, T. M. (2013). A novel statistical measure for sequence comparison on the basis of k-word counts. *Journal of Theoretical Biology* **318**, 91–100. doi: 10.1016/j.jtbi.2012.10.035.

Zemkova, M., Trifonov, E. and Zahradnik, D. (2014). One common structural feature of 'words' in protein sequences and human texts. *Journal of Biomolecular Structure and Dynamics* **32**, 1085–1091. doi: 10.1080/07391102.2013.809317.