Sequence analysis paradigm shift reveals unsuspected semantic properties of species proteomes

Authors: Antonio Starcevic[1, *], Ena Melvan[1, 2, †], Toni Cvrljak[1, †], Janko Diminic[1], Jurica Zucko[1], Paul F. Long[3]

Affiliations:

[1]Laboratory for bioinformatics, Faculty of Food Technology and Biotechnology, University of Zagreb, 10000 Zagreb, Croatia

[2]Department of Biological Science, Faculty of Science, Macquarie University, Sydney, Australia

[3]King's College London, Faculty of Life Sciences & Medicine, London, United Kingdom

*Correspondence to: Antonio Starcevic, antonio.starcevic@gmail.com

† These authors contributed equally to this work.

Abstract

Alignment-based methods dominate molecular biology. However, by primarily allowing one-to-one comparisons, they are focused on gene-centered viewpoint and lack the broad aperture needed for complex biological systems analysis. We hypothesized existence of contextual information related to gene's inclusion in a molecular network of the cell being distributed among more than one sequence. The need for conservation of established interactions, which is arguably more important to the evolutionary success of species than conservation of individual function was the rationale behind this. To test whether this information exists, we applied distributional semantics method - Latent Semantic Analysis (LSA) to thousands of species proteomes. Using natural language processing we identified Latent Taxonomic Signatures (LTSs), a novel proteome distributed feature supporting the argument that protein-coding genes do not evolve as taxonomy independent variables. LTSs reflect constraint imposed to individual gene/protein evolution by their genome/proteome context. In summary, discovery of LTSs indicates that genes had to trade some of their "selfishness" by becoming parts of genome conglomerates.

Introduction

Major role of DNA is information storage. This information enables function, development, growth and reproduction of all known organisms and is organized in the form of genes, gene clusters, chromosomes and ultimately genomes. If DNA represents the code of life, the way it is organized within cells - genome can be compared to the organization of a computer database or a structured document. Even the challenges modern databases face, such as data redundancy, archiving, versioning, replication etc. can be compared to genome related ones, having evolution constantly performing unscheduled updates. In this analogy, the atomic information unit is the gene. A gene is defined as the basic physical and functional unit of heredity acting on its stored information to make proteins, which in return perform actual work ensuring that the gene will multiply and persist through replication. However, genomes are composed of thousands of genes, participating in complex interaction networks. The fact that these networks also need to be conserved through course of evolution, poses a legitimate question on how individual genes and related proteins belonging to different families retain information related to inclusion in already established networks, through course of their evolution. Modern evolutionary studies rely heavily on alignment-based homology to infer relatedness, which has trouble accounting for sequences sharing common biological function or interacting, but lacking sufficient homology to be aligned. We hypothesize existence of latent information common to sequences sharing a network or a context, and herein describe a Latent Semantic Analysis (LSA) (Deerwester *et al*, 1990) vector space model based on species proteomes, which we used to extract biological information from sequences with or without presence of remote homology in an unsupervised manner. The concept behind LSA is the Distributional hypothesis (Harris, 1954), whereby linguistic items with similar distributions have similar meanings. LSA has been extensively used in text processing applications e.g. analyzing relationships between sets of documents (Chen *et al*, 2013) and the same principle can be applied to studying evolution by analyzing the distribution of the 4 letter nucleotide code of genes, or the 20 letter amino acid one in proteins. For example, LSA has been successfully used to predict enzyme substrate specificity (Baranašić *et al*, 2014) and a variant of LSA called probabilistic latent semantic analysis (PLSA) was used to analyze proteomes of closely related pathogens (Rusakovica *et al*, 2014). Others have gone further and proposed bio-vectors referring to biological sequences in general and more specifically, protein-vectors for proteins and gene-vectors for DNA sequences. These vectors achieved >90% family classification accuracy, outperforming existing classification methods (Asgari & Mofrad, 2015). Herein, all available protein sequences from NCBI “nr” database were grouped according to source organism. Resulting species proteomes were used to create species LSA model that was benchmarked against the NCBI Taxonomy database in order to test whether reductionist approach favored by alignment-based methods producing a gene-centered view can be complemented by more integral, natural language processing method putting emphasis on

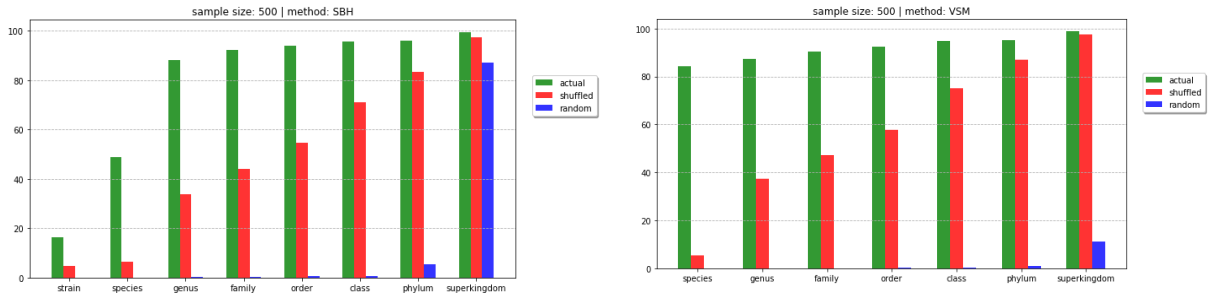
context based information and entire system. The model can easily be accessed on: http://matrix.pbf.hr/.

Results

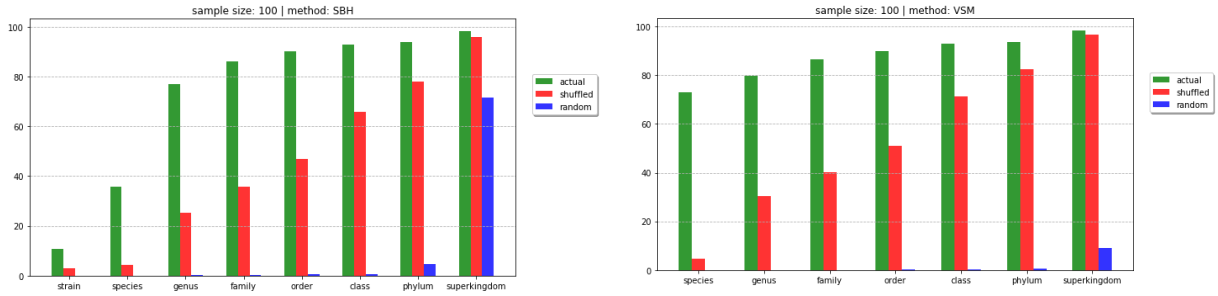
Alignment-free taxonomy benchmarking

In order to assess whether different proteins making up a species proteome store context-based information beside alignment-based one important for the primary function, taxonomy benchmarking test was performed. Latent Semantic Analysis is able to utilize entire protein sets for both query and model creation. For this purpose, three protein query sets of different sizes were selected and deliberately removed from the remaining proteomes used for LSA model. In order to ensure same benchmarking conditions, initial removal of 500 proteins from each of the 54,526 proteomes provided the largest training set from which two smaller ones were extracted (100 and 50 proteins each) and the remaining proteome was used to build the LSA model used in all three benchmarking tests. The model was built by transforming protein sequences into lists of "words", in this case 3-peptide motifs based on a sliding window approach (Supplementary Figure 1A). These 3-peptide frequencies were then weighted, embedded in form of species vectors and utilized for pairwise comparison based on cosine similarity (Supplementary Figure 1B). The taxonomy related prediction capacity of the LSA model was assessed by challenging the model with three query sets of different size for each of the 54,526 species, using single best-hit approach (SBH) and a voting scheme method (VSM) (Supplementary Figure 1B). Briefly, for each taxon query sequence set, the SBH approach assigned taxonomy of the most similar subject taxa vector (from LSA training set), whilst the VSM method took the five most similar taxa vectors from the training set and used these in a voting scenario (Supplementary Figure 1C). The results using both SBH and VSM approaches gave significant agreement with the NCBI taxonomy, ranging from 66.9% - 78.9% for SBH and from 82.6% - 92% for VSM, average taxonomy correlation calculated across all taxonomic ranks from the smallest (50 proteins) to largest (500 proteins) sample used in benchmarking tests (Fig. 1, A, B and C).

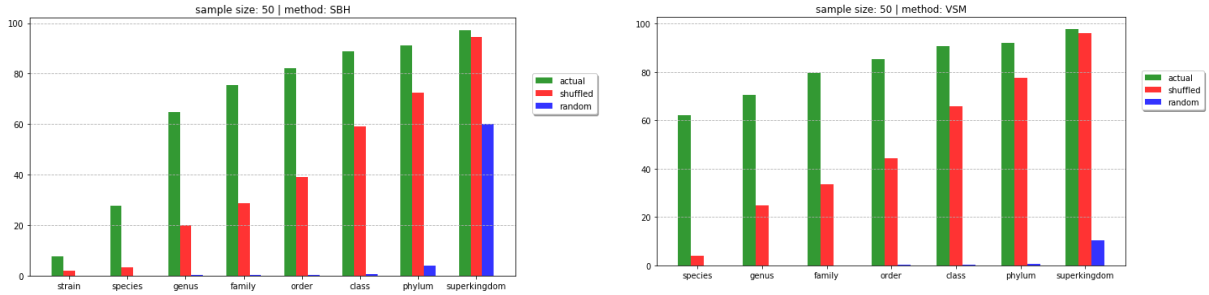
A



B



C



D

	strain	species	genus	family	order	class	phylum	superkingdom
SBH	100%	98.8%	90.4%	91.5%	93.1%	92.8%	98.6%	100%
VSM	0%	36.2%	84.1%	89.8%	92.6%	92.7%	98.4%	100%

Fig. 1. NCBI taxonomy benchmarking results for the following sample sizes: 500, 100 and 50 random gene translations, using SBH (left) and VSM (right). *In order to obtain controls for each actual protein sequence used, we created two additional sets of decoy sequences mirroring the actual ones. These were designated as "shuffled" and "random". Shuffled decoys were created by simple protein sequence shuffling creating a random order of initial amino acids i.e. the single amino acid frequency remains the same between actual and shuffled sequence, while the random decoy sequences use the same set of amino acids found in the actual counterpart and are of the same length, but the single amino acid frequency approaches the normal distribution due to random sampling. (**A**) Randomly sampled protein sequences ranging from (**A**) 500, (**B**) 100 and (**C**) 50 proteins per taxon were decomposed into constituent 3-peptides, folded-in and used to construct queries used to search against 54,526 taxa vectors in species LSA model, built from remaining sequences. (**D**) Table displaying percentage of taxa included in benchmarking*

tests for each taxonomic rank for both SBH and VSM method. Not all taxa have all ranks listed in their lineages and since VSM method includes majority vote among 5 best hits, it requires at least three taxa representatives for each query and for each rank. This excluded strain level from VSM benchmarking and significantly restricted number of query candidates on species level being benchmarked

To test this markedly different approach in comparison to sequence alignment, LSA species model was compared to widely used alignment-based method – BLAST (Altschul *et al*, 1990). Two non-overlapping, same-sized sets of sequences from 1,000 arbitrarily selected species were split 50/50 into training and test set (Supplementary Table 2). Taxonomy assignments were made same as before for LSA. However, for sake of comparison, because LSA utilizes multiple sequences forming a single query based on cumulative 3-peptide frequency opposed to BLAST, which performs solely sequence-to-sequence alignment, a concession had to be made where *a posteriori* only one good homology match between subject and query datasets was to be sufficient for taxonomic assignment made by BLAST (see Materials and Methods). Quite surprisingly, alignment-free natural language processing method clearly outperformed BLAST (Fig. 2, A and B), although the comparison allowed for BLAST to take full advantage of coincidental homology between query and subject protein sets.

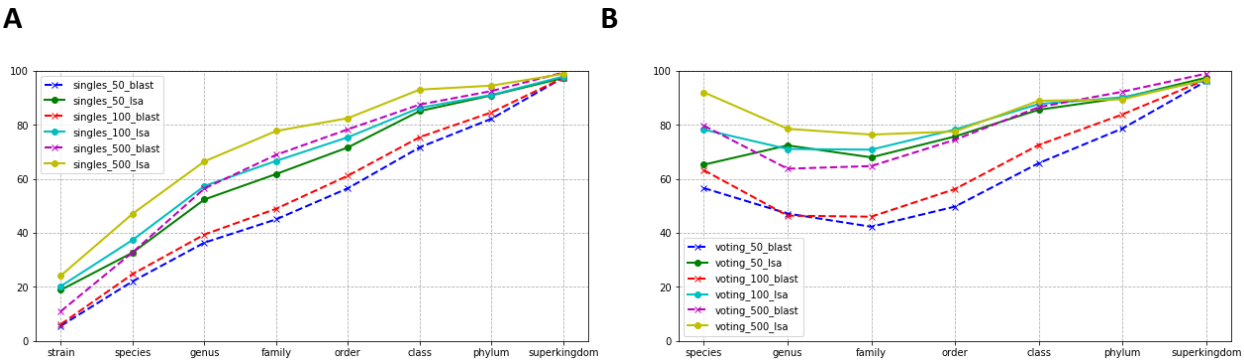


Fig.2. Taxonomy benchmarking results obtained by both LSA and BLAST for each of the three sample sizes (500, 100 and 50 randomly selected proteins) for both (A) SBH and (B) VSM method. *X-axis differs between the two, because the VSM method was applied only to those taxa that had more than 3 representative organisms, therefore the lowest level ("strain") is missing from the VSM results. Full lines represent LSA results and dotted lines represent BLAST results. Unlike in the results from the initial LSA benchmarking test, the VSM method displayed a slight drop in correlation between species and family taxonomic levels, most likely because there were fewer organisms with 3 or more species or genus representatives in the 1,000 organisms selected (Supplementary Table 3)*

This comparison of alignment-based method with natural language processing, suggested that LSA was able to utilize some additional information content stored within heterologous protein samples and not related to alignment-based homology.

Taxonomically restricted proteins exhibit shared semantic properties

In order to further explore this possibility, coincidental homology between protein sets being compared had to be removed. To achieve this, taxonomically restricted sequences were used. Amongst these, orphan sequences are the most limited, restricted to only one or just a few species. These sequences are also interesting because they are frequently associated with novel phenotypes (Khalturin *et al*, 2009; Tautz & Domazet-Lošo, 2011; Babonis *et al*, 2016). Therefore, the capabilities of LSA to extract knowledge from sequences beyond predictions obtainable using distant homology alignments (Rost, 1999) could be further explored. Declaring gene or protein sequence as an orphan is very challenging because genomic databases are incomplete (Basile & Elofsson, 2017). This is why LSA approach was tested on sequences obtained by both stringent and relaxed scenario to account for the level of homology remaining within the sequence sets. Relaxed scenario included the NCBI Protein Clusters (Tatusova *et al*, 2013) dataset. All sequences within our previously used taxa proteomes, which were not listed as parts of NCBI Clusters were regarded as orphan candidates and became parts of the relaxed orphan dataset (Supplementary Table 4). This reversal of the manner in which Clusters dataset was created is analogous to alignment-based methods and E-value threshold criteria used for orphan sequence identification (Domazet-Loso & Tautz, 2003; Sun *et al*, 2015) and significantly reduces overall homology between Clusters and orphan sequence sets used for LSA model and subsequent queries. Because orphan candidates were those sequences not listed in the NCBI Clusters, this approach was proclaimed as "relaxed". In total, 3,913 unique species were present both in our NCBI "nr" based taxa proteome collection and NCBI Clusters, with > 100 orphan candidates. Exactly 100 proteins randomly selected from both Clusters and orphan dataset represented each species twice, once as part of homology-rich Clusters and once as homology-poor orphans. Taxonomy-benchmarking tests comparing BLAST to LSA were performed in the same manner as previously described, with additional step in which datasets reversed their role order. Despite BLAST favoring conditions, the LSA predictions provided significantly better taxonomy benchmarking results (Fig. 3 A and B).

A **B**

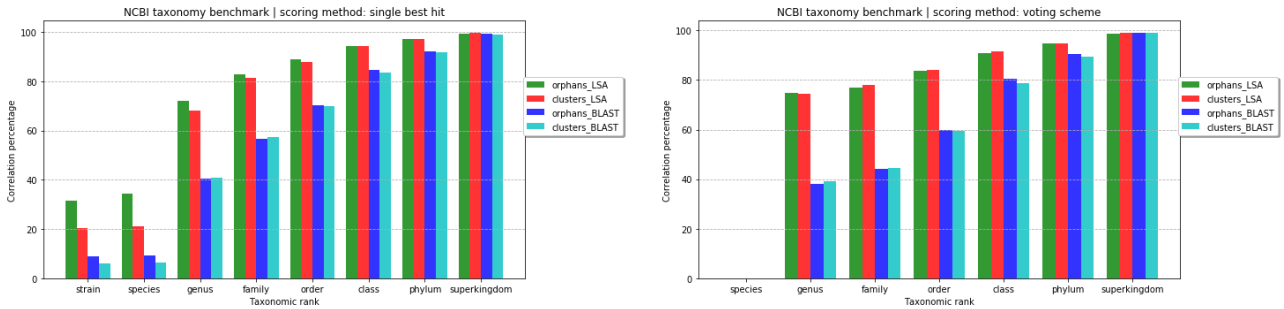


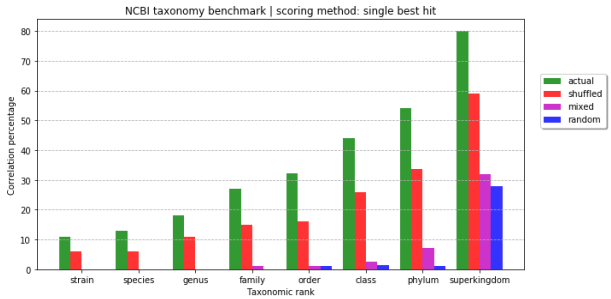
Fig.3. Relaxed orphan taxonomy benchmarking using (A) SBH and (B) VSM methods. *The results were obtained in a reciprocal manner, in which the subject and query sequence sets*

reversed their roles. The combination in which the relaxed orphan dataset was used to build both the LSA subject vectors and blastp database; and where the Clusters dataset was used for query is denoted with "orphans_" prefix. The reciprocal combination in which the Clusters dataset was used to the build both the LSA model and blastp database, which were queried using orphan dataset is denoted with "clusters_" prefix. Green and red bars display the correlation between the LSA benchmarking results and NCBI taxonomy database, whilst blue and cyan bars show the correlation between the results obtained using BLAST.

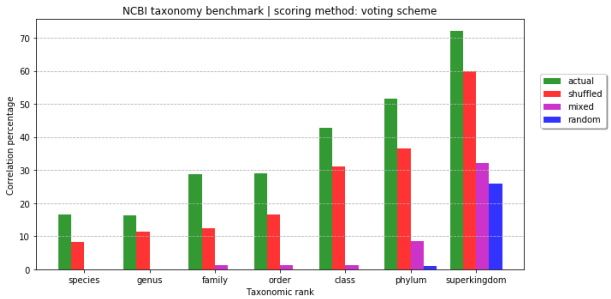
Semantic properties of species proteins are completely independent of alignment-based homology

In order to completely rule out the effect of coincidental homology, we utilized a method in which query sequence sets were stripped from all alignment-detectable homology. For this task, a sample of 100 species was randomly selected, with all four kingdoms of life having at least one representative. In this sample, a two-stage filtering process removed all alignment-detectable homology (see Materials and Methods). In first stage, protein family members could be detected and removed using Hidden Markov models (Eddy, 1998) and Pfam (Finn *et al*, 2008). In second, extensive BLAST of the remaining sequences against the 54,526 species proteomes was performed to remove the remaining homology. What proteins remained after this filtering became parts of the stringent orphan dataset (Supplementary Dataset 1). LSA model was built upon 54,526 species proteomes, with the orphan sequences removed. The benchmarking test was performed without BLAST comparison, because second step of stringent filtering made it redundant. The results showed that 84% of species represented solely through orphan sequences, when used as a queries against LSA model managed to display similarity to other related taxa (Fig. 4, A - E).

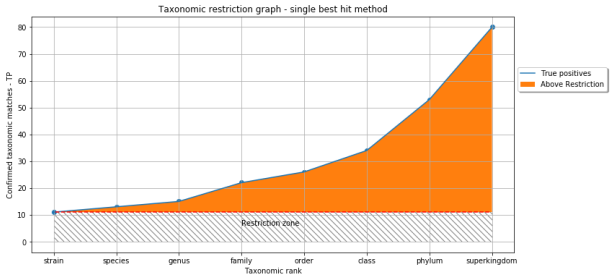
A



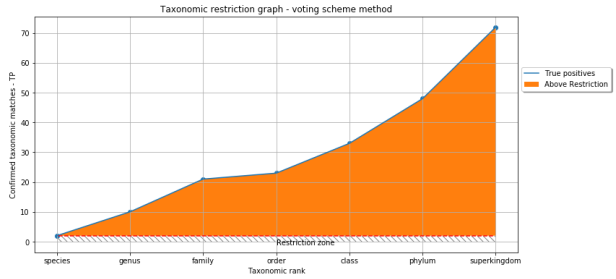
B



C



D



E

	strain	species	genus	family	order	class	phylum	superkingdom
TP - SBH	11	13	15	22	26	34	53	80
TP - VSM	-	2	10	21	23	33	48	72
Number of organisms used as query - SBH	100	100	83	81	81	77	98	100
Number of organisms used as query - VSM	-	12	61	73	79	77	93	100
Total number of unique taxId per taxonomy rank	100	98	79	68	58	32	29	4
Unique taxId coverage - SBH	11%	13.3%	16.5%	27.9%	34.5%	50%	65.5%	100%
Unique taxId coverage – VSM	-	2%	11.4%	22.1%	27.6%	46.9%	62.1%	75%

Fig.4. Stringent orphan taxonomy benchmarking results based on randomly sampled 100 organisms. *Results for (**A**) SBH and (**B**) VSM method. For each orphan sequence, three decoy sequences were included in the analysis. These decoys were used to create "shuffled", "random" and "mixed" datasets. Shuffling the entire orphan dataset and then re-populating each subject taxon container with the same number of proteins taken from this collective mix made "mixed" dataset. The total number of positive taxonomic matches (TP) for each rank is shown for (**C**) SBH and (**D**) VSM methods. Horizontal dashed line at the bottom of each plot denotes the expected taxonomic recognition threshold (as defined by query sequence being restricted on the lowest taxonomic level) (**E**) Tabular overview of results where the first two rows display TP count on each taxonomic level. The third and the fourth row show the number of organisms available as a query at each taxonomy level. The fifth row gives the number of unique taxonomic identifiers amongst the selected 100 organisms – displaying the taxonomic diversity of the sample. The sixth and seventh row display the percentage of taxonomic diversity within query organisms that was successfully covered by positive matches (total unique taxId per rank / unique taxId per positive matches) for both SBH and VSM methods.*

To rule out possibility of obtaining comparable results by chance, “mixed” and “random” decoy sequences were used as controls (Fig. 4, A and B). Majority of positive taxonomic assignments were achieved beyond the homology restriction level (Fig.4, C and D). This strongly implied that LSA established relationships between orphan proteins and related species proteomes, based on taxonomy and in the total absence of alignment-detectable homology. This common taxonomy-related denominator was further highlighted by the “mixed” query set, which was constructed from sequences of mixed origin. Unlike “shuffled” or “random” decoy sequences, by using actual sequences of mixed taxonomic origin, real information content of the protein sequences was used, but the taxonomic contribution was removed. Mixed sequence results were most comparable to those of random decoys and performed significantly worse than actual or shuffled sequences. This confirmed the importance of 3-peptide frequency patterns within species proteomes and indicated that taxonomically homogenous protein sets indeed include additional information content, even in the complete absence of alignment-based homology.

Species proteomes exhibit distinct 3-peptide frequency signatures

Different species have different codon usage preferences. A major difference between codon frequency and amino acids frequency is that different codons can code for the same amino acid, while different combinations of amino acids result in different molecules. Codon usage bias is a well-accepted phenomenon, but the consequence of biased DNA translated to biased proteins on a genome wide scale is somewhat less well known (Singer & Hickey, 2000). Different species may have different preferences for specific amino acids codons and consequently these codons will occur at uneven frequencies in the population causing preferences for specific amino acids to differ (Cannon *et al*, 2015; LaBella *et al*, 2019). Following this rationale, LSA was used to investigate existence of more systematic biases between entire collections of species proteins - proteomes. This was tested using both random sequence sets sharing solely coincidental homology and orthologous sets sharing maximal homology across species. Shuffling sequences is commonly used to create decoy datasets when estimating false discovery rates (Wu *et al*, 2018; Cottrell, 2011). Since LSA can utilize multiple sequences as query and notionally “amplify” signals coming from individual sequences, this method might not suffice. To test whether this might be the case, each 3-peptide in “actual”, “shuffled” and “random” datasets was counted for in > 27 million proteins, from each of the 54,526 taxa used throughout this study (Fig.5 A).

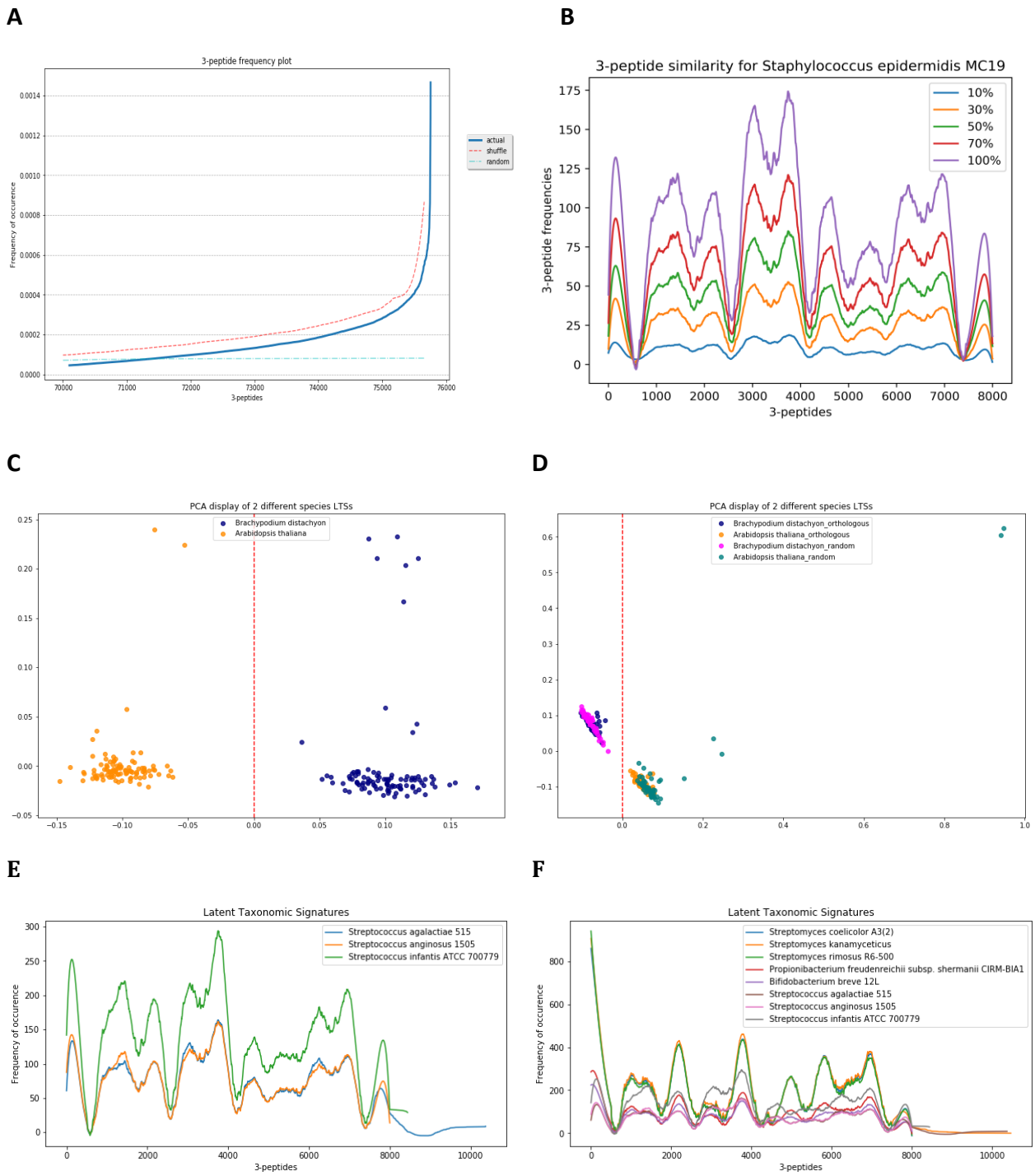


Fig.5. Species proteins exhibit similar 3-peptide frequencies. *3-peptide frequency distribution and selected taxa 3-peptide frequency plots. (**A**) 3-peptide frequencies were calculated from a sample of 27,263,000,00 proteins. The frequencies (Y-axis) were calculated for "actual" (blue), "shuffled" (red) and "random" (cyan) sequences. The X-axis represents 3-peptides sorted by*

*decreasing frequency using "actual" sequence dataset for reference. Shuffled and random 3-peptide frequencies were shifted equally on the x and y-axis to be discernible. Only the most frequent 5,658 3-peptides are being displayed. (**B**) Random protein samples containing different percentage of a complete proteome display a "signature" in form of a repeating frequency pattern, as seen here in a randomly selected bacteria (**C**) Two random plant species proteomes each represented by random selection of 100 proteins on 100 iterations. Each iteration was folded into LSA vector space and displayed using 2-dimensional PCA plot (**D**) The same two plant species, each represented 50 times using random 100 proteins (virtually no homology between two species proteins sets) and another 50 times by different sets of 100 shared orthologous proteins (highest possible degree of homology between two species protein sets) (**E**) 3-peptide frequency plots displaying similar signatures of 3 species of Streptococcus (**F**) 8 different taxa "signatures", 3 Streptococcus species, 3 Streptomyces species, 1 species of Bifidobacterium and 1 species of Propionibacterium. Savitzky-Golay filter was used for smoothing of all plots displayed. Closely related species have more similar signatures.*

When being displayed on the same plot, the "shuffled" sequences 3-peptide frequency exhibited similar pattern as the "actual" ones, strongly implying that shuffling was not sufficient for the collection of proteins to lose its taxonomic signature. This was confirmed using a two-sample Kolmogorov-Smirnov test (KS statistic=0.02474 and p_value=0.062596 for α=0.05 with critical value D_crit=0.0255695) calculated on 5,658 of the most frequent 3-peptides from the "actual" protein dataset. While sequence shuffling can make a single sequence sufficiently random for alignment-based methods, it was not sufficient for LSA, which uses more than one sequence to construct meaningful queries even in the absence of alignment between them. This explains the difference between "shuffled" decoy results compared to the "random" and "mixed" ones. The 3-peptide frequency pattern was shown to be an integral feature exhibited by all proteins making up a proteome (Fig. 5B). While previous benchmarking results indicated these signatures were species specific and robust, this had to be tested by repeated subsampling of different proteome samples. Although each sample was a completely different mixture of species proteins, the boundary between species remained stable (Fig. 5C). Perhaps the most striking confirmation of robustness was obtained when orthologous sequence samples were mixed with the random ones (Fig. 5D). Although same sets of orthologs were used, the species vector clustering remained stable despite ortholog sample protein family relatedness. Smoothed histograms of 3-peptide frequency data were used to demonstrate signature similarities between closely related taxa (Fig. 5 E and F). We named this proteome-distributed feature Latent Taxonomic Signatures (LTSs) because it is both hidden and discernible from alignment-based homology. The discovery of LTSs posed an intriguing question, how did genes belonging to different gene families manage to produce a feature distributed among different proteins constituting species proteomes?

Taxonomic signatures resulting from context dependent evolution

Besides originating from related organisms, it is difficult to explain any commonality between proteins exhibiting LTSs. Same LTSs have been exhibited by heterogeneous protein subsamples encoded by genes from different gene families sharing only genome and proteome context but no alignment-based homology. This suggested that all protein-coding genes within taxa evolve in a concerted manner, dependently of their taxonomic context. Previous results demonstrated that not even orphan sequences are being exempt of this (Fig. 4). However, these datasets represented heterogeneous collections of functionally and structurally unrelated species proteins. Although random sampling is usually the method of choice for testing hypotheses, in order to further validate universality of proposed constraint we also tested whether LTSs are exhibited by non-random, homogenous protein samples. To accomplish this, protein families were used to collect sequences, further sub-divided into taxonomic groups with each group folded-into a single LSA vector representing it. Each of these vectors was created using 3-peptides from highly homologous sets of proteins characterized by significant sequence similarity and a well-defined alignment. Because how they were made, these vectors represent both protein families and contributing taxa and were pairwise compared within the same protein family e.g. “intra-class”, and to entire collection of species vectors, based on whole proteome 3-peptide frequencies e.g. “inter-class”. Cosine based comparison was performed based on two possible outcomes, “selfish” or “altruistic” one. A null hypothesis assumed that there was no relationship between the taxonomic origin of the proteins and the outcomes. Hence, the 3-peptide frequency pattern for any given protein family taxonomic group is expected to be more similar when compared “intra-class” than “inter-class”. High degree of homology within protein family versus no-homology within a collection of proteins contributing to a proteome would certainly indicate this. To test this hypothesis, two control datasets were added, the "mixed” control dataset, with sequences randomly sampled from a protein family (mixed taxonomic origin, but possible bias), and an "HMM” control dataset where protein sequences were replaced by the consensus sequences generated from protein family's HMM model (Eddy, 1998), being completely taxonomically independent. In cases where observed intra-class similarity was greater than any of the inter-class similarities, the outcome was regarded as “selfish”. Opposite case, when one or both inter-class similarities prevailed, the outcome was regarded as “altruistic”. The results of these comparisons indicated both more similar and more abundant inter-class than intra-class similarity outcomes (Fig. 6A).

A

Protein family, taxonomic grouped protein sets	Total number of Pfam families tested	% of SELFISH outcomes	% of ALTRUISTIC outcomes	Total number of outcomes
SPECIES	468	590,563 (29,3%)	1.424,886 (70,7%)	2.015,449
GENUS	201	26,300 (25,1%)	78,556 (74,9%)	104,856

MIXED SPECIES	468	1.600,332 (79,4%)	415,117 (30,6%)	2.015,449
MIXED GENUS	201	75,798 (72,3%)	29,058 (37,7%)	104,856
HMM SPECIES	468	2.015,449 (100,0%)	0 (0,0%)	2.015,449
HMM GENUS	201	104,856 (100,0%)	0 (0,0%)	104,856

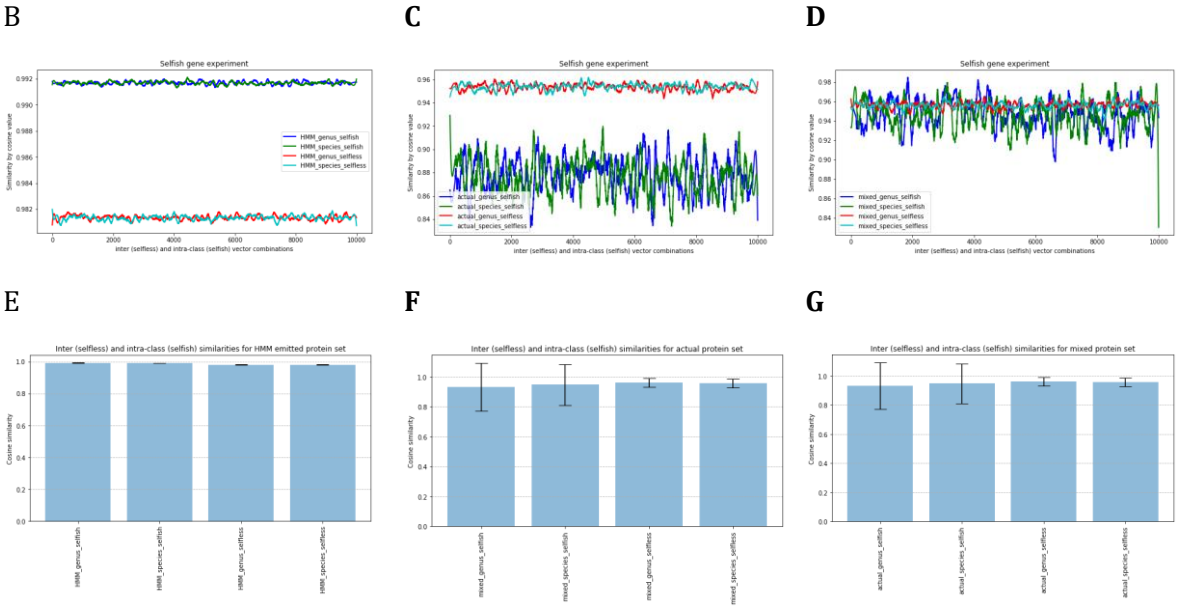


Fig.6. Taxonomy context dependence tested on 468 Pfam protein families in which proteins were grouped according to species, and 201 families where proteins were grouped according to genus *(Supplementary Table 5).* ***(A)*** *Table summarizing the experimental outcomes for "actual" (no prefix), "mixed" (prefix MIXED) and HMM (prefix HMM) datasets. HMM consensus sequences produced exclusively "selfish" outcomes, whilst the majority of actual sequence outcomes were "altruistic".* ***(B)*** *HMM taxonomy independent inter- and intra-class cosine similarity plots constructed from a random selection of 10,000 family-taxa combinations showing best recorded intra-class (selfish) and inter-class (altruistic) cosine values for each of the 10,000 pairwise comparisons. Selfish outcomes are characterized by significantly higher recorded cosine values as expected* ***(C)*** *Cosine plot from a random selection of the 10,000 pairwise family-taxa actual protein set combinations representing "selfish" and "altruistic" outcomes. Selfish outcomes are characterized by significantly lower and more variable recorded cosine similarities compared to "altruistic" ones which indicates evolutionary constraint imposed by species* ***(D)*** *"Mixed" species inter- and intra-class cosine plot for a random selection of 10,000 pairwise comparisons* ***(E)*** *HMM consensus sequences error bar plot* ***(F)*** *Mixed proteins error bar plot* ***(G)*** *Actual proteins error bar plot. In all error bar plots, the height of the bars represents the mean intra-class and inter-class similarity represented as cosine value, for both species and genus grouped sequences. The error bars represented +1/-1 standard deviation about the mean. For the inter-class similarities, an average of two best cosine values obtained by comparison*

with species LTSs are given. For all three datasets (HMM, actual and mixed) all possible pairwise combinations were used in the table and only 10,000 random ones for the plots.

Intra- and inter-class cosine similarity measures for all three datasets gave the same frequency distributions based on Mann-Whitney U Test (p = 0). A Chi-Squared Test on the outcomes from all three datasets (Fig.6, A) rejected the null hypothesis (p=0 with a significance level of 0.05). Experimental outcomes obtained from HMM dataset clearly represented gene-centered view expectation. However, in the case of actual sequences, this was not the case. Inter-class similarity clearly outcompeted intra-class homology (Fig.6 B and C). This was further supported by mean values from the cosine similarity measures. The mean cosine value for the HMM control dataset was significantly higher (0.992) for both species and genus groups in selfish outcomes, compared to 0.981 in altruistic outcomes. Conversely, with actual sequences afforded mean cosine values were significantly lower, 0.879 (species) and 0.886 (genus) in selfish, compared to altruistic 0.954 (species) and 0.960 (genus) outcomes. The mixed dataset was more aligned with the HMM dataset outcomes. However, because some families had taxonomic biases, the outcomes were also largely mixed (Fig.6, A and D). Data used for this experiment were evaluated by constructing error bar charts to display inter- and intra-class cosine similarities for all three datasets (Fig. 6, E, F and G). These charts suggested that the amount of uncertainty was highest in the actual dataset, when "selfish" outcomes were allocated. When "altruistic" outcomes were observed, the degree of uncertainty was much lower and comparable with the control HMM dataset.

Discussion

Statistical analysis of homology between biological sequences is a standard approach to identify conserved sites and motifs that correlate with biological knowledge, such as structure and how it relates to function (Altschuh *et al*, 1987). Function related knowledge is dominantly being uncovered relying solely on sequence alignments. However, these provide limited insight, for example, complementation studies between mutations in a protein can infer interactions between amino acid residues not readily observed within an alignment (Göbel *et al*, 1994). Thus, a longstanding problem in bioinformatics is inferring the information encoded in a biological sequence, when there is little or no sequence homology. A promising new field of research in biology is natural language processing, with the original task of determining the meaning of a word (i.e. semantics) from the contexts in which the words appear (Harris, 1954) being translated to biological sequences. Natural language processing and distributional semantics have increasingly been adopted in biology, as an alternative to alignment-based homology (Bengio *et al*, 2003; Riesselman *et al*, 2018; Rives *et al*, 2021). In this study, Latent Semantic Analysis (LSA) was applied on large number of different species proteomes, spanning wide range of taxonomic diversity in current databases. Major accomplishment of this LSA based species model was the extraction of conceptual content from heterogeneous collections of proteins contributing to proteomes, by establishing associations between 3-peptides that

occur in similar contexts. Resulting 3-peptide signatures emerged as a discernible feature characterizing each species proteome and more than that, they appear to be imprinted on all protein subsets within a proteome. Major difference between this approach and sequence alignment is that the alignment is primarily analytical method, restricted to comparison of contiguous sequences and subsets of conserved residues. Conversely, natural language processing and LSA represent a more systemic approach, which places the dispersed motifs into the context of entire proteomes and allows for species to be modeled as a collection of sequences as opposed to reducing it to a single gene or protein family representative. The resultant vectors that represented the species proteomes emerging from the model, provided valid taxonomic knowledge even when in the total absence of meaningful sequence alignment (Fig 3. A and B, Fig 4. A and B). Evolution is biased towards selecting conserved amino acids that are consistent with fitness (Yanofsky *et al*, 1964). Taking a gene-centered view, evolutionary success is directly linked with the ability to generate as many near-identical copies in a population as possible. Hamilton mathematically set foundations for this reproductive fitness as a measure of evolutionary success in his "Inclusive Fitness" theory of 1964 (Hamilton, 1963). Our results indicate that notion of a successful gene carrying solely self-contained information is incomplete. No gene or protein exists as a sole unit but resides within a taxonomic context. In this context it would not be able to traverse generations successfully without carrying additional information that describes the strength of reciprocal interactions between single gene and genome and conversely, a protein and proteome. Thus, total information content in the genome and proteome is inextricably linked with the integration between these salient features. Latent Taxonomic Signatures (LTSs) reflect this contextual information, which is equally important as the self-contained. LTSs indicate bi-directional relationship between the species and its protein content. In this relationship it is clear that genes affect species but the genomic and proteomic context of species puts a constraint on individual gene and protein evolution requiring a concerted change in order to preserve both the primary function but also the majority of contextual information. Restrictions on the independence of chromosome evolution and examples of concerted evolution phenomena are well known (Brown & Dover, 1980a, 1980b), but have been less well demonstrated at the level of entire genome or proteome (Koonin & Wolf, 2010). Another concerted but only recently explained process is the conservation of replication timing order, which has been shown to orchestrate the global epigenetic state of individual cells. The principle behind this process is the continuation of previously established functional interpretation of the information stored in the genes (Klein *et al*, 2021), which follows our rationale. Beside random protein samples, we showed that even highly homologous groups of proteins categorized under different protein families' express more affinity towards proteome context than their family counterparts when being modeled semantically. Finally, most surprising confirmation of this constraint came from orphan protein sequences. Although these sequences appear to be both taxonomically restricted and *de novo,* with no apparent relatedness to established protein families, we believe this notion is incomplete because it was based solely on sequence alignment. Our results indicate these

sequences should not have been proclaimed taxonomically restricted *a priori*, since even the most stringently defined orphan proteins in our study exhibited LTS patterns placing them in proximity of taxonomically related organisms outside the alignment-imposed "restriction zone" (Fig 4. C and D). Since protein evolution is a direct consequence of changes and selection of DNA polymorphisms and mutations, the real question is what could have caused this both constrained and concerted evolutionary pattern observed through LTSs? The only plausible evolutionary process that acts to maximize genomic and phenotypic cohesiveness, which is connected to speciation and can act on a sufficiently high level to be able to cause LTSs to become embedded in proteins contributing to a species' proteomes is Molecular drive. Major evolutionary forces - natural selection and genetic drift provide explanation for homology-based instances that give rise to unidirectional relationships between gene sequences and traits. Understanding how information in the genome is inextricably linked to a context and gene interactions are conserved between subsequent generations warrants further experimentation to more fully connect the three molecular modes of evolutionary theory – natural selection, genetic drift and molecular drive. Einstein has been quoted saying "It is theory which decides what we can observe", in case of molecular biology, alignment-based methods have dictated what we can observe for a long time, but they also have their limitations and perhaps it is time to start looking for a wider view.

MATERIALS AND METHODS

Obtaining data

All protein sequences were collected from NCBI "nr" database for each of the species included in LSA model (https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/). Proteins were linked to taxa using python programs we have written, which relied on accession to taxId mapping files provided by NCBI Taxonomy database (https://ftp.ncbi.nlm.nih.gov/pub/taxonomy). NCBI Taxonomy database was used as a reference taxonomy database to benchmark LSA against. All used protein data was stored in a local copy of a simple relational database we made using SQLite and python scripts. A Python framework for the analysis and visualization of trees (http://etetoolkit.org/) was used for dealing with the NCBI Taxonomy database (Huerta-Cepas *et al*, 2016). It was used to make a local copy of the NCBI Taxonomy database, and it was later used in our programs to convert from taxId to scientific names (and vice versa) and to establish taxonomic ranks, names and lineages. For relaxed orphan dataset, NCBI Protein Clusters repository was used (https://ftp.ncbi.nih.gov/genomes/CLUSTERS/) in combination with the local copy of "nr" based SQLite relational database. For stringent orphan dataset, we have downloaded current release of Pfam-A HMM models (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.hmm.gz) and used HMMER v.3 (http://hmmer.org/) to scan protein family HMM models against previously selected proteome sequence sets which were extracted from the local database and written in form of simple multi-FASTA protein files. Pfam-A dataset was also used for the "selfish" vs "altruistic" genes experiment as a source of protein family HMM models.

Transforming proteomes into "words"

Let AA be set of 20 natural amino acids. Let $p = a_1 a_2 a_3 \ldots a_n$ represent protein, where $a_i \in$ AA and n is length of protein.

We define triplet t_i as

$$t_i = a_i a_{i+1} a_{i+2}\text{, for each } i < n - 2$$

Truncated singular value decomposition

After $m \times n$ matrix A has been created and properly weighted, a rank-k approximation ($k \ll \min(m, n)$) of matrix A, A_k, is computed using truncated singular value decomposition (T-SVD). Singular value decomposition is a technique closely related to eigenvector decomposition and factor analysis, used to model the associative relationships. This allows for both integrative proteome comparison and topic modeling of 3-peptide motifs within protein sequences. With truncated SVD, matrix A is factored into the product of 3 matrices: $m \times k$ term-concept vector matrix U_k, a $k \times k$ singular values matrix Σ_k and $n \times k$ document-concept vector matrix V_k. Singular value decomposition allows the arrangement of the space to reflect the major associative patterns in the data, and ignore the smaller, less important influences. As a result, terms (in our case 3-peptides) that did not actually appear in a document may still end up close to the document, if that is consistent with the major patterns of association in the data. The truncated SVD captures most of the important underlying structure in the interrelation of terms (3-peptides) and documents (proteomes) and at the same time much of the noise that causes poor retrieval performance is eliminated. In the reduced space, semantically related terms and documents probably lay near each other since the SVD attempts to obtain the fundamental, semantic structure of the term-document (3-peptide – proteome) space. Matrices and process are illustrated below (Fig. 7).

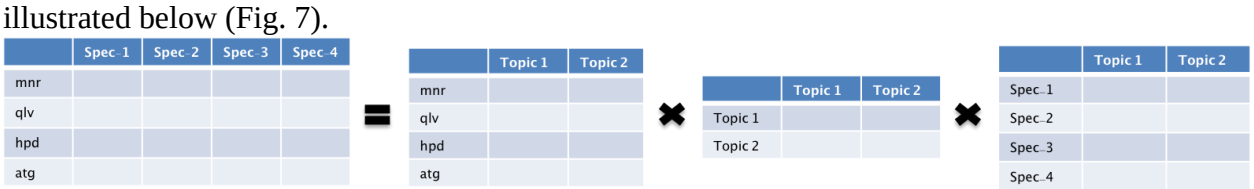


Fig.7. An overview of T-SVD of 3-peptide (terms) – species proteomes (documents) matrix. *With truncated SVD, 3-peptide - species matrix is factored into the product of 3 matrices in the following order (left to right): $m \times k$ 3-peptide-topic (e.g. concept) vector matrix U_k followed by $k \times k$ singular values matrix Σ_k and final $n \times k$ species-concept vector matrix V_k which is the source of our species proteome vectors.*

Latent semantic analysis

Typically, database is organized so that information is retrieved by literally matching exact terms in documents with those of a query. Since there is usually great amount of common 3-peptides in different species proteome, the literal terms in a query may not match those of a relevant document. A better approach would allow users to retrieve information based on the meaning of

a document. Latent Semantic Analysis (LSA) (Deerwester *et al*, 1990) is an extension of the vector space retrieval method in which the dependencies between terms are explicitly considered in the representation and exploited in retrieval. This is done by simultaneously modeling all the interrelationships among terms and documents. The LSA is a retrieval method that builds upon the prior research in information retrieval - Vector Space Model, using the Singular Value Decomposition (SVD) (Golub & Reinsch, 1971) to reduce the dimensions of the term-document space. In vector-space retrieval, a document is represented as a vector in m-dimensional space, where m is the number of terms in the lexicon being used. Each component of the vector reflects a concept associated with the given document. More formally, let m be the number of words (3-peptides) in the dictionary, n the number of documents (species) and $d_i \in \mathbb{R}^m$, $i = 1, \dots, n$, tf-idf representation of *i-th* document. Then the term-document matrix A has the following form:

$$A = (d_1, \dots, d_n)$$

A is usually very large and very sparse matrix since the number of terms in each document is significantly less than the number of terms in the entire document collection. Once a term-document matrix is constructed, local and global weightings are applied to increase or decrease the relative importance of terms within documents. We must also represent the query by the tf-idf scheme, the vector $q \in \mathbb{R}^m$. We would now like to calculate the singular decomposition of the matrix $A = U\Sigma V^T$ which has the following shape:

$$(d_1, \dots, d_n) = (u_1, \dots, u_m)\Sigma \underbrace{(\tilde{d}_1, \dots, \tilde{d}_n)}_{V^T}$$

Suppose we construct the matrices U_k, Σ_k *i* V_k. The representation of the document (species) d_i in the vector space $\{u_1, \dots, u_k\}$ is

$$d_i = \sum_{j=1}^{k} \underbrace{\sigma_j \, (\tilde{d}_i)_j}_{coefficient} * \underbrace{u_j}_{basis\ vector}$$

The singular value of σ_i represents the importance of *i-th* basis vector, u_i. The columns of the matrix $\Sigma_k {V_k}^T$ T represent the coefficients of documents (species) in the newly created lower dimension space, i.e. latent space. In order to be able to compare the query (a set of proteins sharing a proteome) with documents (species), we must represent the query within the latent space. Namely, we look for the coordinates of this vector in the space spanned by the columns of the matrix U_k . If we denote by $\hat{q}$ projection of the vector q on that space, then we know that:

$$\hat{q} = U_k {U_k}^T q$$

Thus, the vector ${U_k}^T q$ represents the coefficients of the vector q in the latent space, i.e., the space spanned by the columns of the matrix U_k. The coefficients of other documents (species) in

the latent space are given by the columns of the matrix $\Sigma_k {V_k}^T$. In order to avoid explicit computation of the matrix $\Sigma_k {V_k}^T$ we define

$$\tilde{q} = \Sigma_k^{-1} {U_k}^T q$$

This is also called *folding-in* method (Brand, 2006). Once the query is projected into the term-document space, cosine similarity measures is applied to compare the position of the new-created query to the positions of the terms or documents in the reduced term-document space. We compare the query q with the documents (species) by comparing $\tilde{q}$ with the columns of the matrix ${V_k}^T$. In the end, we count

$$\cos \Theta_j = \frac{\tilde{d}_j \, \tilde{q}}{\|\tilde{d}_j\|_2 \, \|\tilde{q}\|_2}$$

for each $j = 1, \ldots, n$.

Practical application

All LSA related modeling was performed using Gensim. Gensim is an open-source library for unsupervised topic modeling and natural language processing, using modern statistical machine learning. Gensim is implemented in Python and Cython (https://radimrehurek.com/gensim/) and was installed as part of a python virtual environment with all of the dependencies. Beside Gensim, numpy (https://numpy.org/) and scipy (https://www.scipy.org/) were used for vector operations and statistical analyses. Figures were created using matplotlib (https://matplotlib.org/) and Jupyter Notebook, which were in the same virtual environment Gensim was installed in. All of the libraries and code written was deployed on x86_64 Debian based Linux Ubuntu 18.04.5 LTS local server machine with 2 Intel Xeon E5-2650 2.00GHz CPUs, 8 cores per socket and 2 Threads per core and 264 GB RAM. This server also hosts the web accessible version of LSA based species model. This model was deployed as a web application making it easy for anyone to explore taxonomic context of selected proteins, which is simple to use, requiring no programming or NLP knowledge and no alignment-based homology in the proteins being compared: http://matrix.pbf.hr/.

LSA model of species

LSA uses a term-document matrix, which describes the occurrences of terms in documents. In our model, we replace documents with taxa, and words (also called terms) with all occurring 3-peptides extracted from proteins within species proteomes. In order to transform proteomes into LSA vector space, we decomposed the entire constituent proteins into 3 consecutive amino acid residues e.g. 3-peptides. This is the same length BLAST (Altschul *et al*, 1990) uses as default "word" size for a protein sequence. In our case, this was accomplished by first lowercasing all the protein sequences followed by decomposition into consecutive 3-peptides using a sliding window method, with frame equaling 3 amino acids and "sliding" step equaling to one (Supplementary Figure 1A). Applying dimensionality reduction (SVD) we have embedded

distributional information on all occurring 3-peptides into 400-dimensional LSA vector space. This allowed us to compare entire species proteomes and explore their distributional semantic similarity by correlating it with established taxonomic classification using cosine similarity as metric (Supplementary Figure 1B). Final result aggregates proteome information on tens of thousands of taxa collected from both NCBI Taxonomy database and non-redundant protein database. There are more taxa in these databases; however, we have included only those represented with more than 100 proteins. In order to handle big protein data, we restricted the maximal number of protein sequences to be included in the LSA model to no more than 5,000 proteins per taxon. This means that for all taxa which have proteomes represented with more than 5,000 proteins, 5,000 proteins were randomly selected, while for the rest of the taxa having protein datasets between 100 and 5,000 proteins, entire protein sets were tokenized (Supplementary Table 1). This model is an example of the "bag of words" model (Harris, 1954), which disregards 3-peptide ordering but keeps their multiplicity. On this model, first a term frequency - inverse document frequency (TF-IDF) weighing scheme was implemented (Cullum & Willoughby, 2002). During TF-IDF transformation, 3-peptides which were rare in the training corpus had their weights increased without changing the vector dimensionalities. Finally, we have used LSA to transform from TF-IDF weighted space into a latent space of a lower dimensionality. In our case, we have chosen 400 dimensions empirically (Bradford, 2008). The entire process is summarized below (Fig. 8).

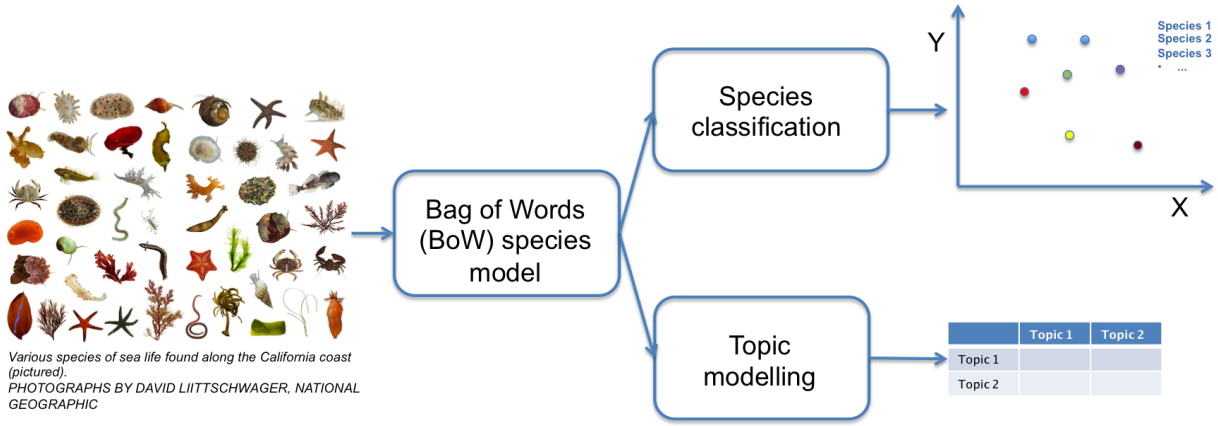


***Fig.8. An overview of LSA based species model**. Species proteomes are represented as "bags of 3-peptides". First step of transforming species proteomes into 400-dimensional vector space (here depicted in simple 2D) includes building a dictionary, which simply counts each 3-peptide occurrence in each species proteome. In the second step, 3-peptide counts are weighted by TF-IDF that gives more weight to those 3-peptides occurring frequently within species proteome and seldom within the rest of the species proteomes in the initial matrix. Finally, SVD is used to transform weighted 3-peptide counts into 400-dimensional species vector space.*

Benchmarking tests

Random subsampling was used to create proteome subsets used for benchmarking. In all four cases, the actual proteins being tokenized and folded into query vectors were searched against the LSA species model and two different methods have been applied in order to assign

taxonomic information to query vectors. First method named "single best hit" (SBH), directly maps taxonomic information belonging to most similar subject vector (by cosine value) to the query. If this assignation correlates with the NCBI Taxonomy taxId of query proteome it is regarded as positive match, otherwise it is negative. Second method is the "voting scheme" (VSM), which is based on taxonomic information from 5 highest-ranking subject vectors (based on cosine) and used in a voting scenario. This scenario requires unambiguous taxonomic majority among the highest-ranking subjects. If this majority correlates with NCBI Taxonomy taxId for the query proteome it is regarded as positive match otherwise it is negative (Supplementary Figure 1C). In case there is no majority vote, the benchmarking result for this query is regarded as negative. Other difference between SBH and VSM methods is the total number of valid query proteomes. In case of SBH the query space is the same as the number of taxa in initial LSA matrix (subject space). In case of VSM, query size is somewhat smaller. The reason for this discrepancy is the fact that the voting scheme relies on majority vote. This majority in worst-case scenario means 3:2 ratio between subject taxa and therefore, query space is restricted only to those taxa, which have at least three representatives in all taxonomic ranks benchmarked. This also implies that VSM will not be including the lowest "strain" taxonomic level, into benchmarking results, unlike SBH. Since cosine value is used as taxa proximity metrics, we have created two new proximity values called "adjacency score" and "similarity interval". These values serve the role of making taxonomic assignations based on cosine similarity between query and subject taxa vectors simpler and more intuitive. Adjacency score is calculated using formula below:

$$adjacency_score = \frac{100}{\max(cos_value) - \min(cos_value)} \, (cos_value - \min(\text{cos_value}))$$

Because cosine range of [0,1] is very narrow, when querying LSA species model we get very dense distribution of subject cosines. This is why in our web application "Matrix of Life" (http://matrix.pbf.hr) only the most similar 100 subject vectors taxonomy information is being displayed. Therefore we decided to "stretch" obtained cosine values and present them on a scale ranging from 0 to 100. The formula above transforms cosines into adjacency scores in a way that the minimal reported cosine maps to 0 and the maximal one to 100. Similarity interval is even simpler to interpret. Essentially, we have divided the [min(cos_value), max(cos_value)] interval on 5 equal parts and counted subjects contained within each part. All subjects within this new interval get a score ranging from 1 to 5 (1 for being the first interval, 5 for being in the last).

Comparison between LSA and BLAST

For BLAST comparison, we have used locally compiled and installed NCBI BLAST 2.9.0+ with "makeblastdb" application to produce BLAST databases from the same proteome sets, which were used to construct LSA vector spaces. Same query proteome sets were used for both LSA queries and "blastp" queries. In order to compare LSA benchmarking results using previously described SBH and VSM methods for taxonomic assignations we have decided to use BLAST favoring conditions. For comparison with SBH, in case of BLAST obtained results, best E-value (lowest) overall HSP (High-scoring Segment Pair) obtained by searching all query proteins

against all subject taxa proteins was used. For comparison with VSM method same "blastp" search was performed, but in this case best five HSPs overall (based on E-values) were used to infer votes and assign taxId to query if there is a majority vote.

This concession had to be made because unlike LSA, which utilizes information stored within multiple query sequences, performing a single search across all subject vectors, BLAST is limited to information stored within a single, contiguous sequence and performs $M \times N_{sample}$ pairwise comparisons for each query, N_{sample} being the sample size (500, 100 or 50 sequences) and $M = N_{taxa} \times N_{sample}$, where in this case N_{taxa}= 1,000. This concession meant that for BLAST only, the best overall pairwise-comparison was used to make taxonomic assignations, and all other matches were disregarding.

Stringent orphan dataset

Using Python scripts we wrote, two-stage orphan candidate filtering was performed on randomly selected 100 taxa proteomes. Since the taxonomic composition of the initial dataset is dominantly Bacterial (92% of taxa), while the other three superkingdoms represent 5,5% (Eukaryota), 2,5% (Archaea) and 0,5% (Viruses). We have sampled 100 taxa more equally from all four superkingdoms to ensure they all have their representatives. First stage orphan candidate filtering included screening for protein family members using Pfam database. Pfam is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (https://pfam.xfam.org/). For this screening, Pfam-A collection of HMM profiles was used since it is based on the same non-redundant ("nr") protein sequences from the NCBI ftp site (https://ftp.ncbi.nlm.nih.gov/refseq/), which we used to create LSA species model. Screening was performed using HMMER "hmmscan" search program and Pfam-A collection of HMM models to search against each of the 100 taxa proteomes in 100 selected taxa. Default "hmmscan" E-value threshold of 10.0 was used. Results of this search were saved in tabular formatted files, which were parsed, and only proteins with no detectable protein family signal were retained (proteins which were not included in the "hmmscan" report). These proteins were selected as first stage or "candidate" orphans. In the second stage "candidate" orphan proteins from the first stage were used as queries in a "blastp" search. This second step served to remove all those proteins, which passed first stage, but still might have homologs that are not assigned to particular protein family. In order to construct blast database for this second step, from each of the 54,526 taxa having more than 1,000 proteins, 500 proteins were randomly sampled and written as a single FASTA file that was turned into a "blastp" searchable database using "makeblastdb" command. In this manner we could perform BLAST search with all the first stage candidate orphan proteins in reasonable time on our server machine, without compromising opportunity to catch stray homologs within entire taxa range we plan to query against. The most commonly used "blastp" E-value threshold for orphan identification is greater or equal than 1e-3 (Tatusova *et al*, 2013). We made this threshold 1,000x more stringent and set it to >= 1.0 for all BLAST hits outside of the organism orphan belongs to. Each of the 100 taxa used for construction of query vectors had to be represented with at least 50 orphan

proteins left after both stages of orphan filtering process. These proteins made the "stringent" orphan dataset. Probability of obtaining the same or greater number of taxonomic matches by chance, for each taxonomic rank can be calculated based on multinomial distribution. More formally, if we have r + 1 subsamples, $A_1, A_2, \dots, A_r, A_{r+1}$, then the probability of drawing at least $i_1, i_2, \dots, i_r$ of elements $A_1, A_2, \dots, A_r$ with replacement, in n repeated draws equals:

$$p = \sum_{\substack{k_1,k_2,\dots,k_r,k_{r+1} \\ k_1 \geq i_1, k_2 \geq i_2, \dots, k_r \geq i_r \\ k_1+k_2+\dots+k_r+k_{r+1}=n}} \binom{n}{k_1, k_2, \dots, k_r,\ k_{r+1}} p_1^{k_1}\, p_2^{k_2} \dots\, p_r^{k_r} p_{r+1}^{k_{r+1}}$$

This formula considers the fact that dataset is imbalanced, meaning taxa distribution is not uniform. Due to our limited computational resources we could not calculate probabilities for taxonomic ranks that had more diverse composition, therefore we have calculated probabilities of a more probable event. This event can be described as drawing all positively identified taxa at least once. The same multinomial distribution applies, and if we have r + 1 subsamples, $A_1, A_2, \dots, A_r, A_{r+1}$, then the probability of drawing at least one of elements $A_1, A_2, \dots, A_r$ with replacement, in n repeated draws can be described as:

$$p = \sum_{\substack{(k_1,k_2,\dots,k_r) \in \mathbb{N},\ \ k_{r+1} \in \mathbb{N}_0 \\ k_1+k_2+\dots+k_r+k_{r+1}=n}} \binom{n}{k_1, k_2, \dots, k_r,\ k_{r+1}} p_1^{k_1}\, p_2^{k_2} \dots\, p_r^{k_r} p_{r+1}^{k_{r+1}}$$

Latent Taxonomic Signatures

To produce illustrations of newly discovered shared sequence feature – Latent Taxonomic Signatures (LTSs), Savitzky-Golay filter was utilized (Savitzky & Golay, 1964). This filter first convolves data by fitting successive sub-sets of adjacent data points with a low-degree polynomial by the method of linear least squares. Python includes this filter in Scipy (savgol_filter) so we implemented it within our code to generate plots showing LTSs of several selected organisms. For these plots, 3-peptide frequency data used for LSA species model training was used, smoothed by Savitzky-Golay windows length 999, and 4th order polynomial.

Context dependent evolution

In order to make this experiment, we collected protein sequences from those Pfam families, which could be decomposed taxonomically into groups having at least 100 protein sequences per taxa. There were 468 such families fulfilling this condition on the species level. We collected an additional dataset consisting of protein family sequences grouped under genus level. Genus level dataset contained 201 Pfam families, and it was constructed to be independent from the species level dataset, meaning that it did not include proteins from taxa already included in the species dataset. Together, these two datasets shared 188 Pfam families; therefore difference in the genus dataset was 13 additional families. Species dataset aggregated protein sequences

from 2,350 taxa, while the genus dataset aggregated protein sequences from additional 570 taxa, not included in species dataset. In total, 292,000 protein sequences (randomly sampled 100 proteins per taxa) and 2,920 taxa were part of this experiment.

In order to produce taxonomically unbiased control samples, to serve as positive control for actual proteins encoded by genes of different organisms, we have used the Pfam protein family HMM models in order to generate 100 protein sequences per taxon, from a fully configured HMMER protein family search profile. This was accomplished using "hmmemit -p" option which is available in new HMMER3 program suite installed locally on our machine. We did this in order to obtain sequences fully consistent with a sequence family consensus, thus representing taxonomically unbiased, e.g. "selfish" behavior. This option allowed us to sample sequences from a fully configured HMMER search profile, so we could obtain homologous sequences by HMMER implemented HMM definition, including non-homologous flanking sequences, allowing us to come as close as possible to the theoretical template of the protein family. After collecting and sorting all protein sequences, taxonomically grouped Pfam family proteins were folded into LSA 400-dimensional vector space. Intra-class comparisons refer to comparisons between taxa vectors belonging to the same protein family. On the other hand, inter-class comparisons correspond to comparisons between vectors made from protein family proteins and Latent Taxonomic Signatures extrapolated from entire proteomes.

Acknowledgments

We thank all those who keep contributing by making biological sequence data publicly available. We also thank prof. John Cullum, prof. emeritus Daslav Hranueli and prof. Walt Dunlap for discussions and constant encouragement. We thank late prof. Gabriel Dover for his work on molecular drive and concerted evolution. Without his scientific courage and vision, we would have great difficulties interpreting results herein presented.

This work was supported by Croatian Science Foundation (grant number IP-2016-06-3509) and Scientific Centre of Excellence for Marine Bioprospecting – BioProCro (grant number KK.01.1.1.01).

Author contributions:

E.M. and T.C. equally contributed to formal mathematics sections. J.D. contributed by developing web based search retrieval system based on LTSs, which can be readily accessed at http://matrix.pbf.hr/, J.Z. and P.F.L contributed to the writing of the manuscript and performed major revisions, A.S. conceived the manuscript, wrote the code and performed experiments. All authors analyzed the results and reviewed the manuscript.

Competing interests:

The authors declare no competing interests.

Data and materials availability:

All of the datasets generated and analyzed during this study are available from the corresponding author (A.S.) on reasonable request with no restrictions since their source is already referenced and publicly available. All computer code generated during this study is also available from the corresponding author (A.S.) on reasonable request and from a GitHub repository: https://github.com/astarsky2016/Latent-Semantic-Signatures. Web application based on LSA taxa model is freely accessible at: http://matrix.pbf.hr/

Supplementary Materials:

Supplementary Table 1 – Information on taxa included in LSA species model

Supplementary Figure 1 – Protein tokenization scheme, cosine similarity and taxonomy assignation in voting scenario

Supplementary Table 2 – Table containing download links for FASTA files with "train" and "test" protein sequence sets used.

Supplementary Table 3 – Percentage of initial taxa query space as defined by available lineage data – used in SBH and VSM method-benchmarking tests

Supplementary Table 4 – Relaxed orphan sequence dataset

Supplementary Dataset 1 – Download link for multi-FASTA formatted sequences comprising "stringent" orphan dataset

Supplementary Table 5 – Protein family taxonomy based groups used in "selfish" vs "altruistic" mode of evolution experiment

References and Notes:

Altschuh D, Lesk AM, Bloomer AC, Klug A (1987) Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J Mol Biol* 193: 693–707

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–10

Asgari E, Mofrad MRK (2015) Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLoS One* 10: e0141287

Babonis LS, Martindale MQ, Ryan JF (2016) Do novel genes drive morphological novelty? An investigation of the nematosomes in the sea anemone Nematostella vectensis. *BMC Evol Biol* 16: 114

Baranašić D, Zucko J, Diminic J, Gacesa R, Long PF, Cullum J, Hranueli D, Starcevic A (2014) Predicting substrate specificity of adenylation domains of nonribosomal peptide synthetases and other protein properties by latent semantic indexing. *J Ind Microbiol Biotechnol* 41:

461–467

Basile W, Elofsson A (2017) The number of orphans in yeast and fly is drastically reduced by using combining searches in both proteomes and genomes. *bioRxiv*: 185983

Bengio Y, Ducharme R, Vincent P, Jauvin C (2003) A neural probabilistic language model. *J Mach Learn Res* 3: 1137–1155

Bradford RB (2008) An empirical study of required dimensionality for large-scale latent semantic indexing applications. In *CIKM '08: Proceedings of the 17th ACM conference on Information and knowledge management* pp 153–162.

Brand M (2006) Fast low-rank modifications of the thin singular value decomposition. *Linear Algebra Appl* 415: 20–30

Brown SD, Dover GA (1980a) The specific organisation of satellite DNA sequences on the X-chromosome of Mus musculus: partial independence of chromosome evolution. *Nucleic Acids Res* 8: 781–792

Brown SD, Dover GA (1980b) Conservation of segmental variants of satellite DNA of Mus musculus in a related species: Mus spretus. *Nature* 285: 47–49

Cannon JGD, Sherman RM, Wang VMY, Newman GA (2015) Cross-species conservation of complementary amino acid-ribonucleobase interactions and their potential for ribosome-free encoding. *Sci Rep* 5: 18054

Chen H, Martin B, Daimon CM, Maudsley S (2013) Effective use of latent semantic indexing and computational linguistics in biological and biomedical applications. *Front Physiol* 4: 8

Cottrell JS (2011) Protein identification using MS/MS data. *J Proteomics* 74: 1842–1851

Cullum JK, Willoughby RA (2002) Lanczos Algorithms for Large Symmetric Eigenvalue Computations Society for Industrial and Applied Mathematics

Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. *J Am Soc Inf Sci* 41: 391–407

Domazet-Loso T, Tautz D (2003) An evolutionary analysis of orphan genes in Drosophila. *Genome Res* 13: 2213–2219

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* 14: 755–63

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz H, Ceric G, Forslund K, Eddy SR, Sonnhammer ELL, *et al* (2008) The Pfam protein families database. *Database* 36: 281–288

Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins* 18: 309–317

Golub GH, Reinsch C (1971) Singular Value Decomposition and Least Squares Solutions. In *Handbook for Automatic Computation. Die Grundlehren der mathematischen Wissenschaften*, Bauer FL Householder AS Olver FWJ Rutishauser H Samelson K & Stiefel E (eds) pp 134–151. Springer, Berlin, Heidelberg

Hamilton WD (1963) The Evolution of Altruistic Behavior. *Am Nat* 97: 354–356

Harris ZS (1954) Distributional Structure. WORD 10: 146–162

Huerta-Cepas J, Serra F, Bork P (2016) ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol* 33: 1635–1638

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG (2009) More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet* 25: 404–413

Klein KN, Zhao PA, Lyu X, Sasaki T, Bartlett DA, Singh AM, Tasan I, Zhang M, Watts LP, Hiraga S, *et al* (2021) Replication timing maintains the global epigenetic state in human cells. *Science (80-)* 372: 371 LP – 378

Koonin E V, Wolf YI (2010) Constraints and plasticity in genome and molecular-phenome evolution. *Nat Rev Genet* 11: 487–498

LaBella AL, Opulente DA, Steenwyk JL, Hittinger CT, Rokas A (2019) Variation and selection on codon usage bias across an entire subphylum. *PLoS Genet* 15: e1008304

Riesselman AJ, Ingraham JB, Marks DS (2018) Deep generative models of genetic variation capture the effects of mutations. *Nat Methods* 15: 816–822

Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, *et al* (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci* 118: e2016239118

Rost B (1999) Twilight zone of protein sequence alignments. *Protein Eng* 12: 85–94

Rusakovica J, Hallinan J, Wipat A, Zuliani P (2014) Probabilistic Latent Semantic Analysis Applied to Whole Bacterial Genomes Identifies Common Genomic Features. *J Integr Bioinform* 11: 93–105

Savitzky A, Golay MJE (1964) Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal Chem* 36: 1627–1639

Singer GAC, Hickey DA (2000) Nucleotide Bias Causes a Genomewide Bias in the Amino Acid Composition of Proteins. *Mol Biol Evol* 17: 1581–1588

Sun W, Zhao X-W, Zhang Z (2015) Identification and evolution of the orphan genes in the domestic silkworm, Bombyx mori. *FEBS Lett* 589: 2731–2738

Tatusova T, Zaslavsky L, Fedorov B, Haddad D, Vatsan A, Akoadjei D, Blinkova O, Ghazal H (2013) Protein Clusters. In *The NCBI Handbook [Internet]* pp 331–342. Bethesda (MD): U.S. National Center for Biotechnology Information

Tautz D, Domazet-Lošo T (2011) The evolutionary origin of orphan genes. *Nat Rev Genet* 12: 692–702

Wu G, Wan X, Xu B (2018) A new estimation of protein-level false discovery rate. *BMC Genomics* 19: 567

Yanofsky C, Horn V, Thorpe D (1964) Protein Structure Relationships Revealed by Mutational Analysis. *Science* 146: 1593–1594