AMSI VACATION RESEARCH SCHOLARSHIPS 2020–21

Get a Thirst for Research this Summer

Hierarchical Clustering of Bacterial Protein Sequences

Susanna R. Grigson

Supervised by Jody C. McKerral, Robert A. Edwards, Jim G. Mitchell, Vlad Ejov

College of Science and Engineering, Flinders University

Vacation Research Scholarships are funded jointly by the Department of Education, Skills and Employment and the Australian Mathematical Sciences Institute.

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Australian Government
Department of Education, Skills and Employment

Abstract

Proteins are crucial for biological and molecular functions in living organisms. Due to the ever-expanding gap between the number of proteins being discovered and their functional characterisation, protein function inference remains a fundamental challenge in computational biology. Currently, protein functions are classified in hierarchical ontologies constructed from experimental observations. Recent advancements in natural language processing and machine learning have inspired the development of the Protvec method for embedding amino acid sequences as vectors in a protein space. While Protvec successfully groups proteins with shared functions, the underlying mechanisms facilitating these groupings remain unclear. By analysing the most informative vectors in Protvec models we determine that Protvec groups proteins with similar functions by learning biologically meaningful information. Through embedding protein sequences using Protvec, we identify discrepancies between existing classification procedures and systematic groupings relative to the biophysical and biochemical properties of proteins. By extending Protvec to group sequences with unidentified functions, we propose an alternative approach to select optimal candidate proteins to characterise experimentally. Protein sequence embeddings and hierarchical clustering may be beneficial for reorganising and completing classification frameworks used to label bacterial proteins.

Introduction

Proteins are biomacromolecules composed of strings of amino acids which enable biological functions essential for an organism to sustain life. The type and order of amino acids constituting a protein are encoded by the nucleotides in an organism's DNA. Interactions between amino acids determine the three-dimensional structure of the protein, dictating its specific biological function.

With the advent of low-cost, high throughput DNA sequencing technology, the amino acid sequences of millions of bacterial proteins have been obtained (Levy & Myers 2016). As the knowledge of protein sequence function is not increasing at the same rate, wide gaps persist between the number of known protein sequences and known protein functions (Koboldt et al. 2013). Traditional methods determine protein function using experimental procedures. Although reliable, this process is slow and expensive, therefore not feasible for characterising millions of unknown proteins. Alternatively, protein function can be inferred using sequence similarity to proteins with known functions. As singlecelled organisms are the most abundant of all organisms, bacteria have a rich and diverse range of proteins (Olanrewaju, Glick & Babalola 2017).

To organise the available protein functional information, computational biologists have devised hierarchies known as ontologies to label protein sequences which are likely to share similar functions. These ontologies group proteins using laboratory experiments and human expert annotation. The Subsystems ontology (Overbeek et al. 2014) is a popular ontology used to label protein sequences and contains four levels: superclass, class, subclass and subsystem.

Recently, methods from natural language processing have been utilised to represent amino acid sequences. Asgari and Mofrad (2015) devised Protvec models based on the *Word2V*ec algorithm which create word-embeddings by learning word associations from a large corpus of text (Mikolov et al. 2013). Protvec splits amino acid sequences into overlapping subsequences of length k to obtain 'words' known as k-mers. Using a skip-gram neural network which iterates through lists of words analysing the likelihood of neighbouring words, a distributed representation is created where interactions of k-mers with other k-mers is stored. This allows each possible k-mer to be represented as a 100-dimensional vector which is updated as the model cycles through training data. Using the resulting Protvec model, proteins can be embedded as 100-dimensional vectors by taking the sum of k-mer vectors for all k-mers present in the sequence. As protein function is encoded by the amino acid sequence of a protein, the region in the embedding space where the sequence is embedded is related to its biological role. Therefore, Protvec could be used to mathematically group proteins with similar functions. In many cases dissimilar protein sequences have similar biological roles as the location of a few key amino acids within sequences determine protein function. Protvec observes patterns within amino acid sequences rather than relying on sequence similarity.

It remains unclear whether hierarchies created using protein embeddings and hierarchical clustering approaches will partition sequences differently to existing classification methods. However, if successful, Protvec models grouping protein sequences with unidentified functions could be used to resolve the fundamental challenge of protein function inference.

This project aims to hierarchically cluster protein amino acid sequences embedded using Protvec models. Through evaluating models trained with protein sequence data from different bacterial groups, protein sequence properties that drive the embedding of amino acid sequences within the protein space are identified. These models were used to mathematically group proteins with shared biological roles, identifying inconsistencies between existing protein classification schemes. Finally, by grouping proteins with unknown functions, we indicate a promising approach to improve protein function prediction and to reduce the disparity between known protein sequences and known protein functions. The ability to accurately predict protein function has the potential to accelerate research in fields including human health and biotechnology (Burley et al. 2018).

Methods

Protvec model training

Protvec models were trained with 8743 protein sequences in the carbohydrate metabolism class for the bacterial groups, *Bacillus* and *Bacteroides.* These sequences were obtained from *the Genome Taxonomy Database* (GTDB), a database containing a diverse range of prokaryotic sequences (Parks et al. 2020). The sequences contained the standard 20 amino acids represented by the letters $\{\mathrm{R}, \mathrm{I}, \mathrm{E}, \mathrm{M}, \mathrm{W}, \mathrm{D}, \mathrm{P}, \ \mathrm{K}, \mathrm{C}, \mathrm{F}, \mathrm{G}, \mathrm{T}, \mathrm{L}, \mathrm{Y}, \mathrm{Q}, \mathrm{V}, \mathrm{A}, \mathrm{S}, \mathrm{N}, \mathrm{H}\}$. Using the Python *genism* package, sequences were converted to overlapping k-mers of length k=3, resulting in 20^3 unique 3-mers. These 3-mers were trained through a skip-gram neural network to create a model containing a 100-dimensional vector for each 3-mer. The resulting models were compared with the Protvec model trained with 324,018 protein sequences from the Swiss-Prot database in previous work (Asgari, McHardy & Mofrad 2019).

Protein Space Analysis

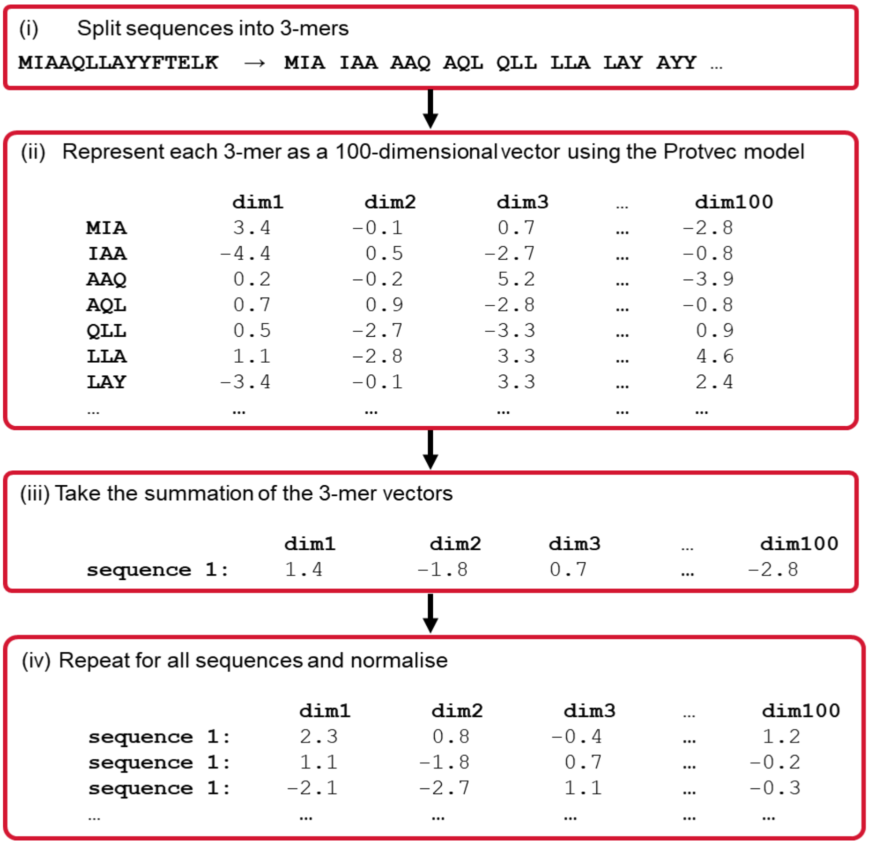
Comparisons were made between the trained Protvec models. Singular value decomposition was used to factorise each of the models, M into three matrices U, S and V satisfying, $M = USV^T$ where S is a diagonal matrix of the singular values.

The 100 3-mer vectors with the greatest Euclidean distance from the origin in the *Bacillus* and *Bacteroides* models were isolated. The occurrence of each amino acid in these 3-mers was plotted to identify whether the 3-mer vectors which contribute the most to sequence embeddings are associated with particular amino acids.

Protvec Sequence Embedding

26,547 *Bacillus* and 16,764 *Bacteroides* carbohydrate metabolism sequences were obtained from the Pathosystems Resource Integration Center to test the Protvec models (PATRIC) (Wattam et al. 2014). Sequences which also occurred in GTDB and used to train the Protvec models were excluded. Additionally, sequences with a length below 30 amino acids and greater than 1024 amino acids were removed. These parameters have been used in previous studies as proteins shorter than 30 amino acids are unlikely to form a structure allowing a biological function and proteins longer than 1024 amino acids are uncommon (Abu-Doleh, Al-Jarrah & Alkhateeb 2012; Rives et al. 2019; Villegas-Morcillo et al. 2020; Yang et al. 2018).

Sequences that also occurred in GTDB and were used to train the Protvec models were excluded. Additionally, sequences with a length below 30 amino acids and greater than 1024 amino acids were removed. These parameters have been used in previous studies as proteins shorter than 30 amino acids are unlikely to form a structure allowing a biological function and proteins longer than 1024 amino acids are uncommon (Abu-Doleh, Al-Jarrah & Alkhateeb 2012; Rives et al. 2019; Villegas-Morcillo et al. 2020; Yang et al. 2018).



(i) Split sequences into 3-mers

MIAAQLLAYYFTELK → MIA IAA AAQ AQL QLL LLA LAY AYY …

(ii) Represent each 3-mer as a 100-dimensional vector using the Protvec model

	dim1	dim2	dim3	…	dim100
MIA	3.4	-0.1	0.7	…	-2.8
IAA	-4.4	0.5	-2.7	…	-0.8
AAQ	0.2	-0.2	5.2	…	-3.9
AQL	0.7	0.9	-2.8	…	-0.8
QLL	0.5	-2.7	-3.3	…	0.9
LLA	1.1	-2.8	3.3	…	4.6
LAY	-3.4	-0.1	3.3	…	2.4
…	…	…	…	…	…

(iii) Take the summation of the 3-mer vectors

	dim1	dim2	dim3	…	dim100
sequence 1:	1.4	-1.8	0.7	…	-2.8

(iv) Repeat for all sequences and normalise

	dim1	dim2	dim3	…	dim100
sequence 1:	2.3	0.8	-0.4	…	1.2
sequence 1:	1.1	-1.8	0.7	…	-0.2
sequence 1:	-2.1	-2.7	1.1	…	-0.3
…	…	…	…	…	…

Fig. 1 Procedure to embed sequences using Protvec models.

Sequence embeddings were visualised using Principal Component Analysis (PCA) and t-distributed stochastic neighbourhood embedding (t-SNE) (Van der Maaten & Hinton 2008). Prior to visualisation, all sequence vectors were standardised using Z-score normalisation using the python *scikit-learn* package (Pedregosa et al. 2011). Sequences were coloured by their subclass in the Subsystems ontology.

K-mer Frequency

As an alternative to embedding sequences using Protvec models, the *Bacillus* and *Bacteroides* carbohydrate metabolism test sets were embedded using the frequency of each k-mer within the amino acid sequences. Sequences were converted to the murphy10 reduced amino acid alphabet containing the letters $\{\mathrm{F, E, C, G, L, S, A, K, H, P}\}$ (Murphy, Wallqvist & Levy 2000) and represented as overlapping k-mers of length k=3. Using the reduced alphabet, the number of possible 3-mers was reduced from 20^3 to 10^3, lowering computational requirements.

To embed sequences using k-mer frequency, an identity matrix of size $n = 10^3$ was created with the rows and columns corresponding to each possible 3-mer. This allowed each 3-mer to be represented as zero vector with a 1 at a unique position (Fig. 2). Amino acid sequences were embedded as vectors of length 10^3, where each position denotes the presence or absence of a 3-mer. Embedding vectors were obtained by converting amino acid sequences to overlapping 3-mers, matching each 3-mer to its corresponding vector in the identity matrix and taking the sum of these vectors. To adjust for sequences containing different numbers of 3-mers, the sequence vectors were normalised by dividing each sequence vector by the length of the sequence. The resulting embeddings were visualised using PCA and t-SNE.

$$\begin{array}{c|ccccc|} & FFF & FFE & FFC & \cdots & PPP \\ FFF & 1 & 0 & 0 & \dots & 0 \\ FFE & 0 & 1 & 0 & \dots & 0 \\ FFC & 0 & 0 & 1 & \dots & 0 \\ \cdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ PPP & 0 & 0 & 0 & \dots & 1 \end{array}$$

Fig. 2: Matrix used to embed amino acid sequences using k-mer frequency.

Number of Clusters

The number of clusters in the sequence embeddings was estimated by creating clusters for different values of K and calculating the within-cluster sum of squares (WSS) and the between cluster sum of squares (BSS).

$$WSS(K) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} z_{ik}(x_i - \overline{x_k})^2$$

$$BSS(K) = \sum_{k=1}^{K} \frac{n_k}{n}(\overline{x_k} - \overline{x})^2$$

Where n is the total number of elements, K is the number of clusters, $n_k$is the number of elements in the kth cluster, $\overline{x_k}$ is the mean of the kth cluster , $\overline{x}$ is the sample mean and z_{ik} is an indicator function,

$$z_{ik} = \begin{cases} 1 & x_i \in clusterk \\ 0 & x_i \notin clusterk \end{cases}$$

Using WSS and BSS the Calinski-Harabasz index (CH) was calculated (Caliński & Harabasz 1974).

$$CH(K) = \frac{WSS(K)}{BSS(K)} \frac{(n-k)}{(k-1)}$$

The Calinski-Harabasz index was plotted for *Bacillus* and *Bacteroides* carbohydrate metabolism sequences embedded with the Protvec models for $k = 2{:}\,100$.

Dendrograms

The Euclidean and Manhattan distances were calculated between each of the sequences embedded with the Protvec models. Hierarchies were constructed using 1,000 randomly selected sequences from each embedding. These hierarchies were constructed using agglomerative (bottom-up) and divisive (top-down) clustering. Comparisons were made with Subsystems classifications of these sequences by constructing tanglegrams. Tanglegrams were untangled using the *step2side* method to minimise entanglement between the dendrograms. The resulting hierarchies were visualised using the R *dendextend* package (Galili 2015).

Clustering Unknowns

Bacillus protein sequences without functional annotations in Subsystems were obtained from the PATRIC database. Sequences were dereplicated at 70% sequence identity using CD-HIT (Fu et al. 2012). Sequences with a length less than 120 amino acids were removed. This is the average number of amino acids required to form a protein domain, thus, ensuring each sequence encodes at least one function (Xu & Nussinov 1998). Additionally, sequences longer than 1024 amino acids and sequences containing 'X' were removed. A random subset of 425,000 unknown sequences were used to train a Protvec model. Using this model, a separate test set of 25,000 unknown sequences were embedded. The embedded sequences were visualised using t-SNE and clustered using k-means clustering. The optimal number of clusters was determined using the within sum of square errors.

Results

To understand the properties of Protvec models driving the embedding of protein sequences, the singular values of the 3-mer vectors in each Protvec model were calculated and plotted (Fig. 3). The singular values of the *Bacillus* and *Bacteroides* models were similar in value, indicating little clustering behaviour between the 3-mers. In comparison, the Swiss-Prot model had 12 singular values ranging between 100 and 250. This indicates that a series of multiple coarse-grained clusters are present between the 3-mer vectors. These results are visible in the PCA of the 3-mers of each of the models (Fig. 3). Using k-means clustering the Swiss-Prot vectors form 3 distinct clusters (App. 1). These clusters are caused by the presence of 'X' which denotes an unknown amino acid and 'C' which encodes the amino acid cysteine.

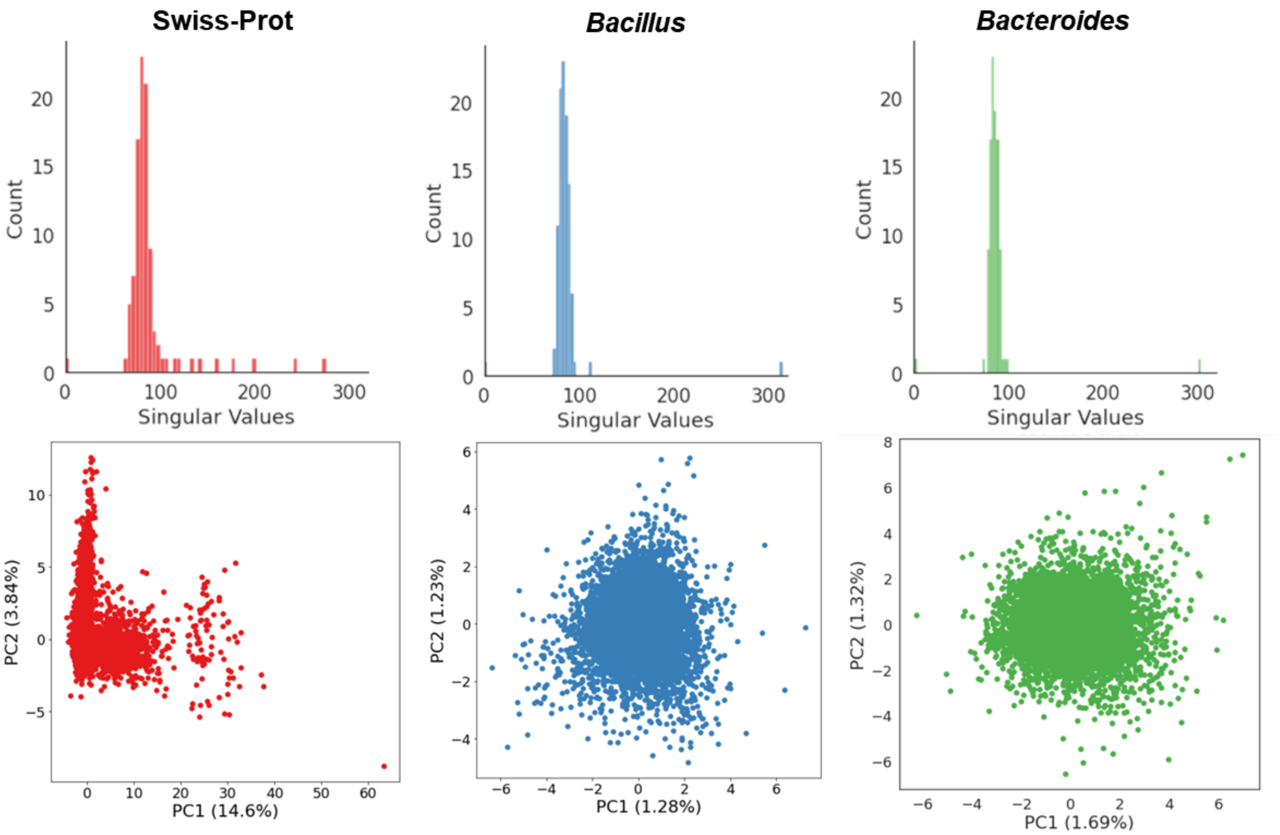
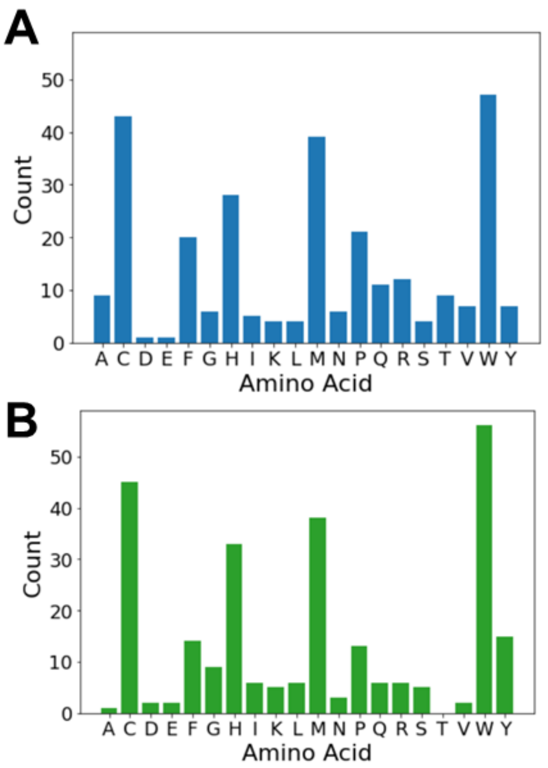


Fig 3. Singular values and PCA of Protvec models trained with Swiss-Prot sequences, *Bacillus* carbohydrate metabolism sequences and *Bacteroides* carbohydrate metabolism sequences.

To further investigate the distribution of 3-mer vectors, the Euclidean distance of each 3-mer vector to the origin was determined for the *Bacillus* and the *Bacteroides* Protvec models. The 100 vectors with the greatest distance from the origin for both models had a high occurrence of the amino acids tryptophan (W), cysteine (C), methionine (M) and histidine (H) (Fig 3A, 3B). These amino acids have high values in the Blocks Substitution Matrix (BLOSUM) used to score protein sequence alignments (Fig 3C). The values in this matrix provide data on the conservation amino acids between proteins (McGinnis & Madden 2004). Additionally, the distributions of these amino acids are different between the Protvec models trained with sequences from different bacterial groups.



A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	-1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**

Fig 3. Number of occurrences of each amino acid the 100 k-mers with the greatest Euclidean distance from the origin in **A:** The *Bacillus* Protvec model and **B:** The *Bacteroides* Protvec model. **C:** The BLOSUM62 matrix used to score protein alignments with the 4 most abundant amino acids in **A** and **B** highlighted.

Bacillus carbohydrate metabolism sequences embedded using Protvec models and k-mer frequency demonstrated grouping of sequences belonging to the same subclass (Fig. 4). Greater noise was present for the embedding using the Swiss-Prot Protvec model compared to the *Bacillus* trained model. Despite this, some poorly differentiated sequences remain clustered at the centre of the t-SNE visualisation for sequences embedded using the *Bacilllus* Protvec model. The k-mer frequency embedding was similar to the *Bacillus* Protvec embedding, though some sequences did not form clusters in the *t*-SNE of the k-mer frequency embedding. Similar patterns were observed for *Bacteroides* carbohydrate metabolism sequences embedded using k-mer frequency and the Protvec models trained with *Bacteroides* and Swiss-Prot sequences (App. 2).

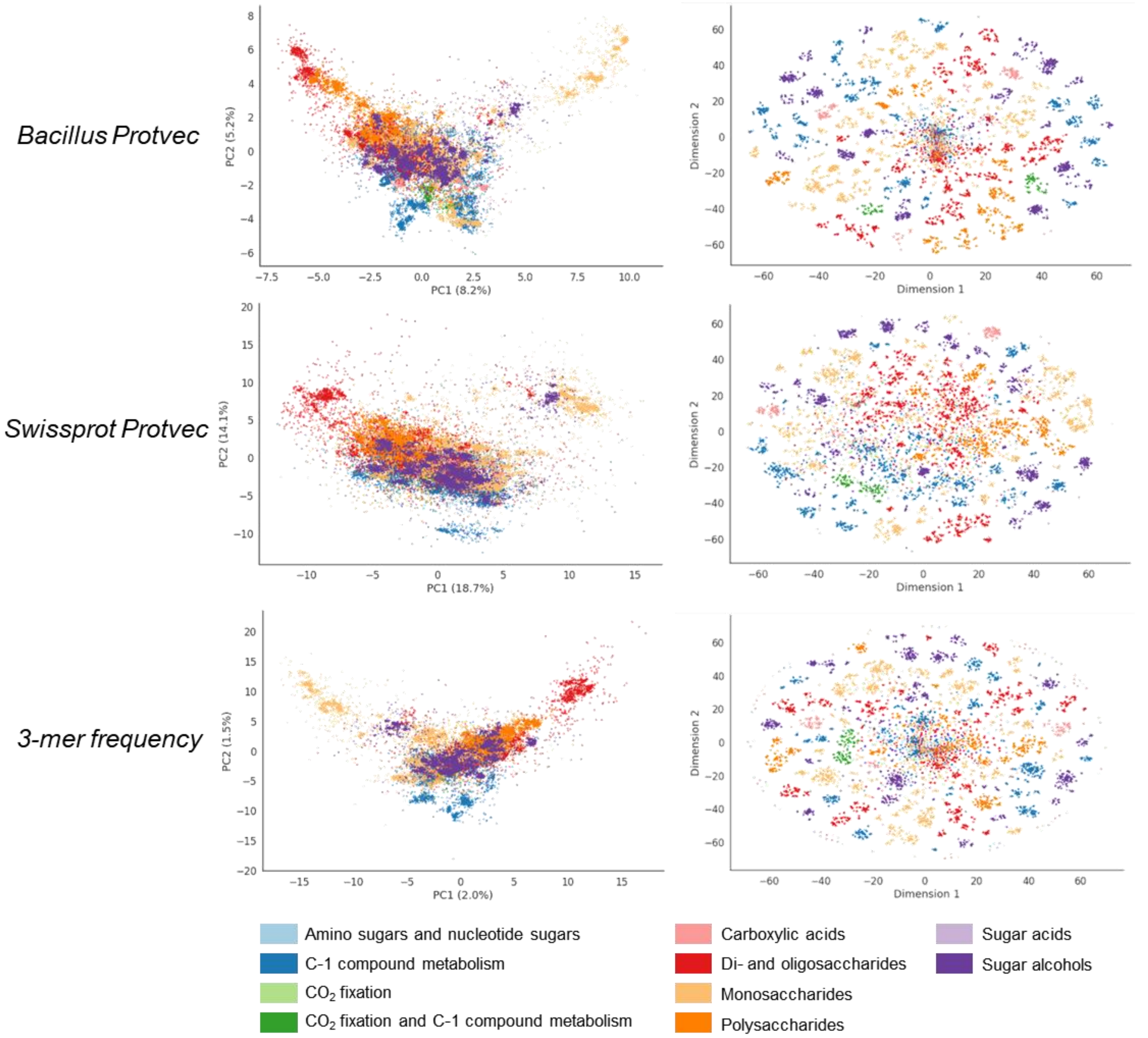


Fig 4: Bacillus carbohydrate metabolism sequences embedded using a Protvec model trained with Bacillus Carbohydrate metabolism sequences, Protvec model trained with Swiss-Prot and k-mer frequency of the embedded sequences. Sequences are coloured by subclass and visualised with PCA and tSNE (perplexity = 30, learning rate = 100).

To mathematically determine the number of groups present in the carbohydrate metabolism subclass, carbohydrate metabolism sequences were embedded and the number of clusters formed calculated. Using *Bacillus* sequences embedded with the Protvec *Bacillus* model, the Calinski-Harabasz index peaks when the embedding is partitioned into 48 clusters, indicating that the within cluster variance is minimised and the between cluster variance is maximised (Fig 5A). The human built Subsystems ontology groups the same sequences into 29 subsystems. This implies that the mathematical grouping of protein sequences organises sequences differently than currently used approaches to label proteins. The same sequences embedded with the Swiss-Prot Protvec model did not form discrete clusters. The Calinski-Harabasz index does not reach a peak and there is no clear elbow visible for the within cluster sum of squares (Fig 5C, 5D). The Calinski-Harabasz index and within cluster sum of squares was also calculated for *Bacteroides* sequences embedded with Swiss-Prot and the *Bacteroides* trained Protvec model (App. 3).

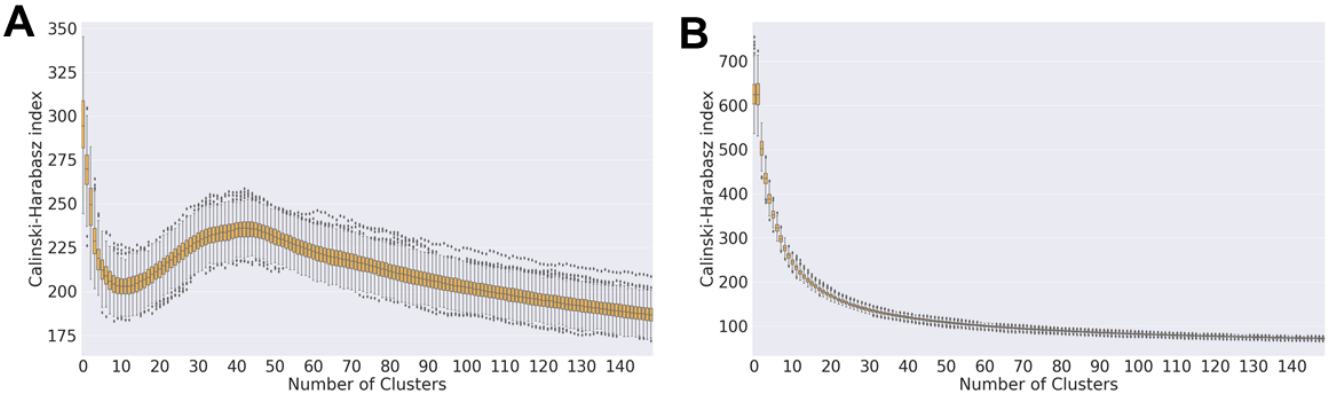


Fig 5. Calinski-Harabasz index of 5,000 Bacillus carbohydrate metabolism sequences embedded with **A:** Protvec model trained with Bacillus carbohydrate metabolism sequences and **B:** Protvec model trained with Swiss-Prot. 500 bootstrap iterations were used for each value of k.

The hierarchical organisation of *Bacillus* sequences embedded with the *Bacillus* Protvec model using Euclidean distance and agglomerative clustering was resolved differently than the same sequences under the Subsystems ontology (Fig. 6). In many cases, sequences which clustered together in the embedding hierarchy also clustered within the same subsystem in the Subsystems ontology. However, the higher-level structure was not preserved between hierarchies as distantly related groups in the hierarchy built using the embedded sequences belonged to different subclasses in the Subsystems hierarchy. In some cases, subsystems contain two distinct groups of sequences which belong to different regions of the protein sequence embedding. Hierarchies were also constructed using divisive clustering (App. 4) and using Manhattan distance (App. 5) though these demonstrated less similarity with the Subsystems classifications.

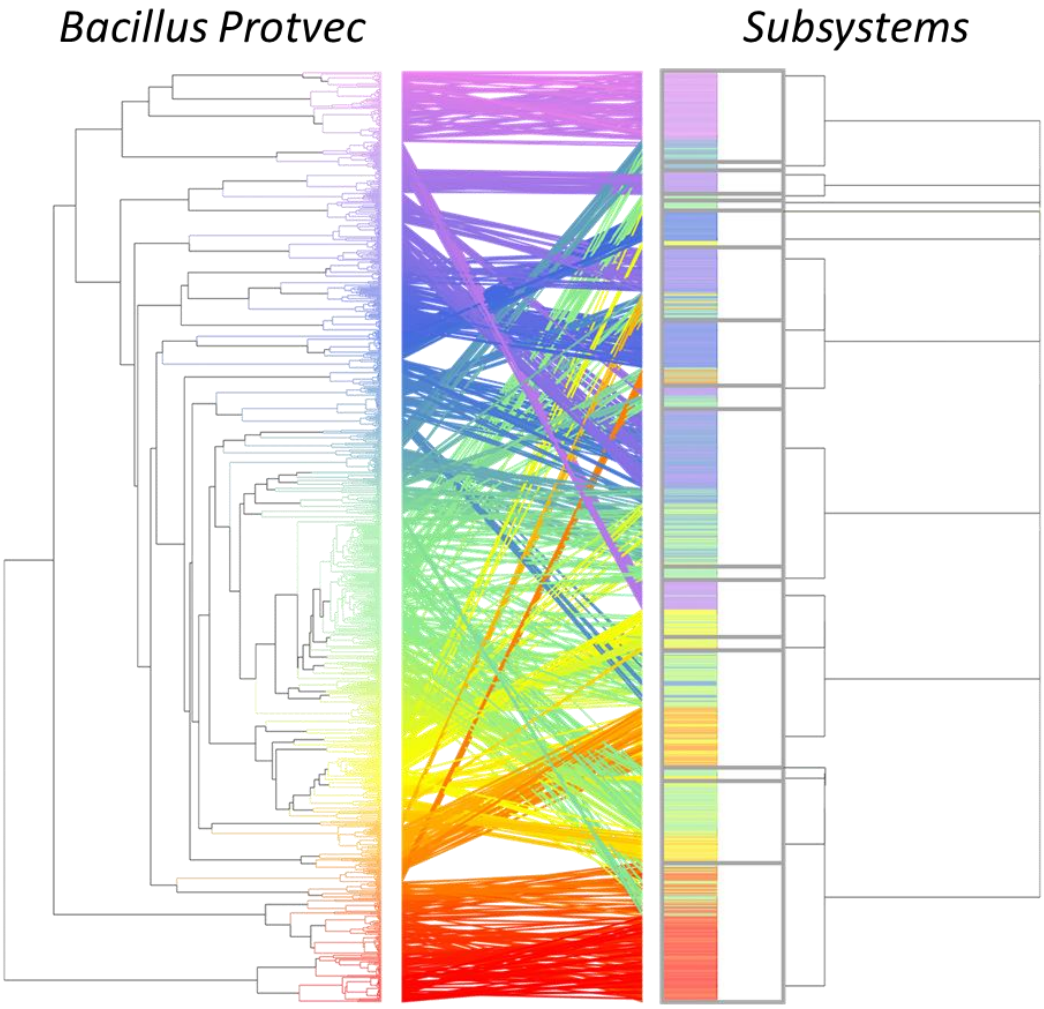


Fig 6 A: Tanglegram of 1,000 Bacillus carbohydrate metabolism protein sequences. Hierarchy on the left is built using agglomerative clustering on the Euclidean distances between sequences embedded using a Protvec model trained with Bacillus sequences. Hierarchy on the right is the classification of sequences in the existing subsystems ontology.

To determine whether sequence embeddings can be used to group sequences with unknown functions, *Bacillus* protein sequences with unknown functions were embedded with the Protvec model also trained with unknown *Bacillus* sequences. The within cluster sum of squares plot has an elbow an 12 clusters, indicating that there were 12 groups of unknown sequences in the embedding set (Fig 7B). Using k-means with 12 clusters, the embedded sequences formed discrete groups (Fig 7A). Sequences belonging to the same cluster experience low sequence similarity as indicated by the percentage identity between the clustered sequences (App. 6). Whilst the function of the sequences within these groups unknown, sequences within the same cluster likely share similar biological roles.

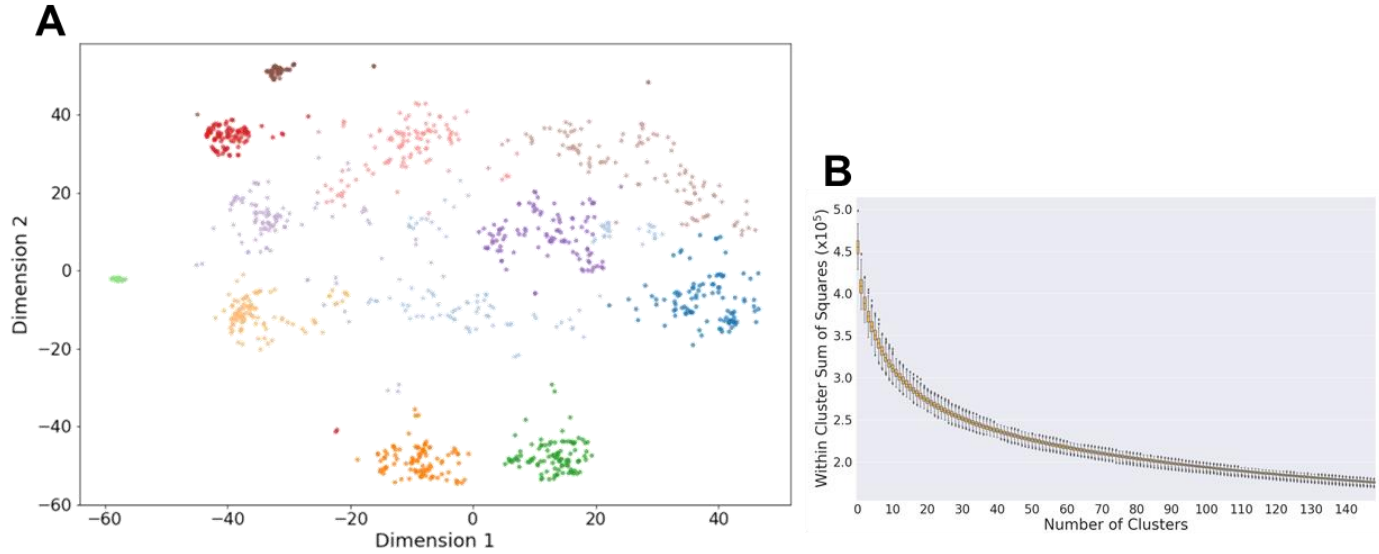


Fig 7. A: k **-**means clustering of Bacillus sequences from PATRIC without functional annotations in Subsystems. Sequences were embedded using a Protvec model trained with 500,000 unknown Bacillus sequences. The 100 sequences with the smallest Euclidean distance to the centroid of each cluster are shown. **B:** Within sum of square errors of 5,000 Bacillus sequences from PATRIC without functional annotations in the Subsystems ontology for 1-100 clusters.

Discussion

This work investigated how biological sequence embeddings operate to group proteins with similar functions. Previous studies focus on developing embedding methods for proteins sequences but do not interpret the underlying mathematics which enable proteins to be grouped respective of their biological roles (Nambiar et al. 2020; Villegas-Morcillo et al. 2020). Instead, these studies use embeddings to develop tools for protein prediction tasks. As existing classification frameworks are based on experimental observations, they may label sequences differently to mathematical approaches utilising statistics and linear algebra. Therefore, protein embeddings are helpful for evaluating hierarchies used to classify protein function.

The BLOSUM matrix has been developed to determine the similarity between proteins by aligning amino acid sequences and calculating a quality score for the alignment. Substitution between dissimilar amino acids results in a penalty and conservation of similar amino acids increases the score. The amino acids tryptophan, cysteine, histidine and methionine played the greatest role in driving proteins to different regions of the protein space. This is consistent with the BLOSUM matrix which allocates high scores when cysteine, tryptophan, and histidine residues are conserved between proteins (Eddy 2004). Furthermore, methionine and cysteine are the only sulphur containing amino acids allowing the formation of disulphide bonds which play a critical role in protein folding (Lim, Kim & Levine 2019). The observation that these amino acids drive sequence embedding means that Protvec operates by considering conserved amino acids and amino acids important for protein folding. Protvec learns biologically meaningful information rather than relying on the similarity of protein sequences.

Training Protvec with sequences from different bacterial groups altered the amino acid composition of the k-mers driving the sequence embedding. Additionally, embedding sequences with the Swiss-Prot model trained with a diverse set of sequences did not form distinct groups of clusters. Applications of Word2Vec in natural language processing have determined that generic word vectors constructed from generalised corpuses are less effective than domain specific word vector models (Chiu et al. 2016; Ghosh et al. 2016). Therefore, Protvec models trained using a diverse set of sequences may not distinguish between closely related proteins. Ideally, Protvec models should be trained with sequences similar to the intended embedding set to ensure that model vectors embed sequences based on the specific properties of the embedding sequences. Alternatively, k-mer frequency embeddings do not require this consideration as training is not required to embed sequences.

Bacillus carbohydrate metabolism sequences embedded using the *Bacillus* Protvec model produced 48 discrete clusters. Under the Subsystems classification system, these sequences belonged to 29 different groups. This implies that Subsystems does not contain all the labels required to classify these sequences completely. As Subsystems was constructed from experimental observations, this introduces error to the labelling accuracy (Laukens, Naulaerts & Berghe 2015; Overbeek et al. 2014). Some subsystems within the Subsystems ontology likely contain proteins with more than one function and therefore could be partitioned into separate groups.

The dendrograms built using sequence embeddings can be used to evaluate how closely hierarchies built using embeddings resemble the Subsystems Ontology. Several groups of sequences which were clustered together in the *Bacillus* carbohydrate metabolism dendrogram belonged to the same subsystem in the Subsystems ontology. Despite this, sequences belonging to different subsystems within the same subclass were mapped to different sections of the embedding hierarchy.

This indicates that low-level groupings in subsystems are somewhat consistent with those derived from sequence embeddings but are incorrectly divided into the overarching hierarchies. Therefore, existing classification schemes including Subsystems may not sufficiently group protein sequences by their biological roles. Improved ontology design should follow a systematic approach. This could be achieved using the mathematical properties of protein sequences rather than experimental observations.

Tools have been developed which predict the function of proteins by utilising Protvec models (Cai, Wang & Deng 2020). As one-third of bacterial proteins are too dissimilar to proteins which have been previously characterised (Price et al. 2018), their function cannot be accurately predicted using these methods. Training and embedding unknown *Bacillus* sequences using Protvec revealed 12 clusters of proteins with unknown functions. These clusters may contain further 'subclusters' which could be identified by using Protvec to train and embed sequences from each cluster. As Protvec groups proteins with similar functions, the sequences in each of these clusters likely share similar biological functions. Consequently, prime experimental candidates could be selected from each cluster and characterised to inform the function of the sequences in the unknown clusters. As characterising proteins is expensive and labour intensive, grouping proteins with unknown functions is an efficient strategy for determining the function of unknown proteins (Seo et al. 2018). By characterising unknown proteins grouped using sequence embeddings, additional labels could be included in protein classification schemes to decrease the number of proteins with unknown functions.

Recent advances in natural language processing have developed more advanced algorithms for word embeddings. This includes Embeddings from Language Models (ELMo) and transformer models such as Bidirectional Encoder Representations from Transformers (BERT). These methods have been successfully used to embed protein sequences (Heinzinger et al. 2019; Nambiar et al. 2020). Future work could utilise these embedding methods to construct hierarchies and group unknown sequences.

Conclusion

This study embedded bacterial protein sequences using word embeddings to evaluate protein sequence classification. Dissection of Protvec models revealed that Protvec models embed amino acid sequences using biologically meaningful information. By grouping proteins with shared functions in a vector space, inconsistencies were identified between the existing Subsystems ontology used to label proteins and hierarchical classifications utilising sequence embeddings. Further, amino acid sequences with unknown biological functions were organised into groups based on the underlying mathematical properties of protein sequences, providing an alternative approach for resolving the fundamental challenge of protein function inference. Redesigning protein classification schemes using a systematic approach would help overcome the limitations of currently used experimental approaches.

Code accessible at https://github.com/susiegriggo/ProtvecHierachy

Authorship Statement

The workload was divided as follows:

JCM, RAE & JGM conceived and supervised the work.

RAE & JCM gathered the data.

SRG trained the models, composed sequence embeddings, did the analysis and wrote the report.

JCM, VE & JGM proofread the report.

Funding was provided by AMSI and the Australian Department of Education.

References

Abu-Doleh, AA, Al-Jarrah, OM & Alkhateeb, A 2012, 'Protein contact map prediction using multi-stage hybrid intelligence inference systems', *Journal of Biomedical Informatics*, vol. 45, no. 1, pp. 173-183.

Asgari, E, McHardy, AC & Mofrad, MR 2019, 'Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (DiMotif) and sequence embedding (ProtVecX)', *Scientific reports*, vol. 9, no. 1, pp. 1-16.

Asgari, E & Mofrad, MR 2015, 'Continuous distributed representation of biological sequences for deep proteomics and genomics', *PloS one*, vol. 10, no. 11, p. e0141287.

Burley, SK, Berman, HM, Bhikadiya, C, Bi, C, Chen, L, Di Costanzo, L, Christie, C, Dalenberg, K, Duarte, JM, Dutta, S, Feng, Z, Ghosh, S, Goodsell, DS, Green, RK, Guranović, V, Guzenko, D, Hudson, BP, Kalro, T, Liang, Y, Lowe, R, Namkoong, H, Peisach, E, Periskova, I, Prlić, A, Randle, C, Rose, A, Rose, P, Sala, R, Sekharan, M, Shao, C, Tan, L, Tao, Y-P, Valasatava, Y, Voigt, M, Westbrook, J, Woo, J, Yang, H, Young, J, Zhuravleva, M & Zardecki, C 2018, 'RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy', *Nucleic Acids Research*, vol. 47, no. D1, pp. D464-D474.

Cai, Y, Wang, J & Deng, L 2020, 'SDN2GO: An Integrated Deep Learning Model for Protein Function Prediction', *Frontiers in Bioengineering and Biotechnology*, vol. 8, no. 391.

Caliński, T & Harabasz, J 1974, 'A dendrite method for cluster analysis', *Communications in Statistics*, vol. 3, no. 1, pp. 1-27.

Chiu, B, Crichton, G, Korhonen, A & Pyysalo, S 'How to train good word embeddings for biomedical NLP', pp. 166-174.

Eddy, SR 2004, 'Where did the BLOSUM62 alignment score matrix come from?', *Nature Biotechnology*, vol. 22, no. 8, pp. 1035-1036.

Fu, L, Niu, B, Zhu, Z, Wu, S & Li, W 2012, 'CD-HIT: accelerated for clustering the next-generation sequencing data', *Bioinformatics*, vol. 28, no. 23, pp. 3150-3152.

Galili, T 2015, 'dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering', *Bioinformatics*, vol. 31, no. 22, pp. 3718-3720.

Ghosh, S, Chakraborty, P, Cohn, E, Brownstein, JS & Ramakrishnan, N 'Characterizing diseases from unstructured text: A vocabulary driven word2vec approach', pp. 1129-1138.

Heinzinger, M, Elnaggar, A, Wang, Y, Dallago, C, Nechaev, D, Matthes, F & Rost, B 2019, 'Modeling aspects of the language of life through transfer-learning protein sequences', *BMC Bioinformatics*, vol. 20, no. 1, pp. 1-17.

Koboldt, DC, Steinberg, KM, Larson, DE, Wilson, RK & Mardis, ER 2013, 'The next-generation sequencing revolution and its impact on genomics', *Cell*, vol. 155, no. 1, pp. 27-38.

Laukens, K, Naulaerts, S & Berghe, WV 2015, 'Bioinformatics approaches for the functional interpretation of protein lists: from ontology term enrichment to network analysis', *Proteomics*, vol. 15, no. 5-6, pp. 981-996.

Levy, SE & Myers, RM 2016, 'Advancements in next-generation sequencing', *Annual review of genomics and human genetics*, vol. 17, pp. 95-115.

Lim, JM, Kim, G & Levine, RL 2019, 'Methionine in proteins: it's not just for protein initiation anymore', *Neurochemical Research*, vol. 44, no. 1, pp. 247-257.

McGinnis, S & Madden, TL 2004, 'BLAST: at the core of a powerful and diverse set of sequence analysis tools', *Nucleic acids research*, vol. 32, no. suppl_2, pp. W20-W25.

Mikolov, T, Chen, K, Corrado, G & Dean, J 2013, 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781*, vol.

Murphy, LR, Wallqvist, A & Levy, RM 2000, 'Simplified amino acid alphabets for protein fold recognition and implications for folding', *Protein engineering*, vol. 13, no. 3, pp. 149-152.

Nambiar, A, Heflin, M, Liu, S, Maslov, S, Hopkins, M & Ritz, A 'Transforming the language of life: Transformer neural networks for protein prediction tasks', pp. 1-8.

Olanrewaju, OS, Glick, BR & Babalola, OO 2017, 'Mechanisms of action of plant growth promoting bacteria', *World Journal of Microbiology and Biotechnology*, vol. 33, no. 11, pp. 1-16.

Overbeek, R, Olson, R, Pusch, GD, Olsen, GJ, Davis, JJ, Disz, T, Edwards, RA, Gerdes, S, Parrello, B & Shukla, M 2014, 'The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)', *Nucleic acids research*, vol. 42, no. D1, pp. D206-D214.

Parks, DH, Chuvochina, M, Chaumeil, P-A, Rinke, C, Mussig, AJ & Hugenholtz, P 2020, 'A complete domain-to-species taxonomy for Bacteria and Archaea', *Nature Biotechnology*, vol. 38, no. 9, pp. 1079-1086.

Pedregosa, F, Varoquaux, G, Gramfort, A, Michel, V, Thirion, B, Grisel, O, Blondel, M, Prettenhofer, P, Weiss, R & Dubourg, V 2011, 'Scikit-learn: Machine learning in Python', *the Journal of machine Learning research*, vol. 12, pp. 2825-2830.

Price, MN, Wetmore, KM, Waters, RJ, Callaghan, M, Ray, J, Liu, H, Kuehl, JV, Melnyk, RA, Lamson, JS, Suh, Y, Carlson, HK, Esquivel, Z, Sadeeshkumar, H, Chakraborty, R, Zane, GM, Rubin, BE, Wall, JD, Visel, A, Bristow, J, Blow, MJ, Arkin, AP & Deutschbauer, AM 2018, 'Mutant phenotypes for thousands of bacterial genes of unknown function', *Nature*, vol. 557, no. 7706, pp. 503-509.

Rives, A, Goyal, S, Meier, J, Guo, D, Ott, M, Zitnick, CL, Ma, J & Fergus, R 2019, 'Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences', *bioRxiv*, vol., p. 622803.

Seo, S, Oh, M, Park, Y & Kim, S 2018, 'DeepFam: deep learning based alignment-free method for protein family modeling and prediction', *Bioinformatics*, vol. 34, no. 13, pp. i254-i262.

Sievers, F & Higgins, DG 2014, 'Clustal Omega, accurate alignment of very large numbers of sequences', in *Multiple sequence alignment methods*, Springer, pp. 105-116.

Van der Maaten, L & Hinton, G 2008, 'Visualizing data using t-SNE', *Journal of machine learning research*, vol. 9, no. 11.

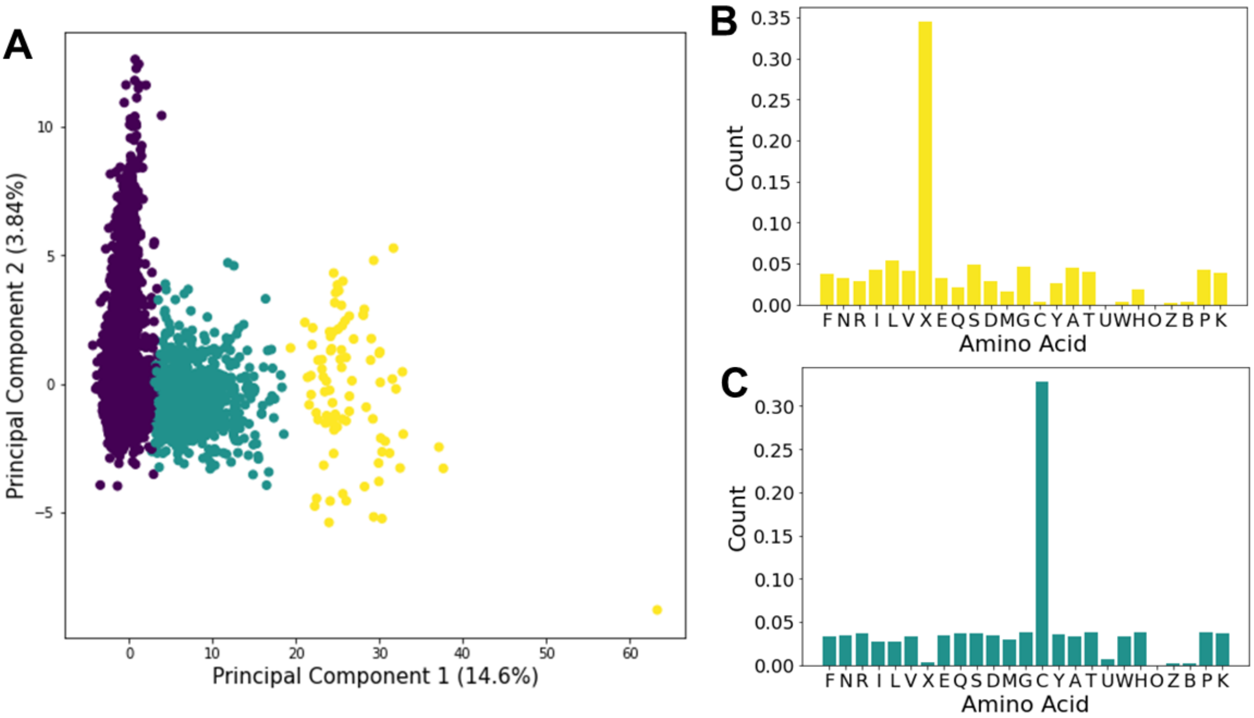
Villegas-Morcillo, A, Makrodimitris, S, van Ham, RCHJ, Gomez, AM, Sanchez, V & Reinders, MJT 2020, 'Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function', *Bioinformatics*, vol.

Wattam, AR, Abraham, D, Dalay, O, Disz, TL, Driscoll, T, Gabbard, JL, Gillespie, JJ, Gough, R, Hix, D & Kenyon, R 2014, 'PATRIC, the bacterial bioinformatics database and analysis resource', *Nucleic acids research*, vol. 42, no. D1, pp. D581-D591.

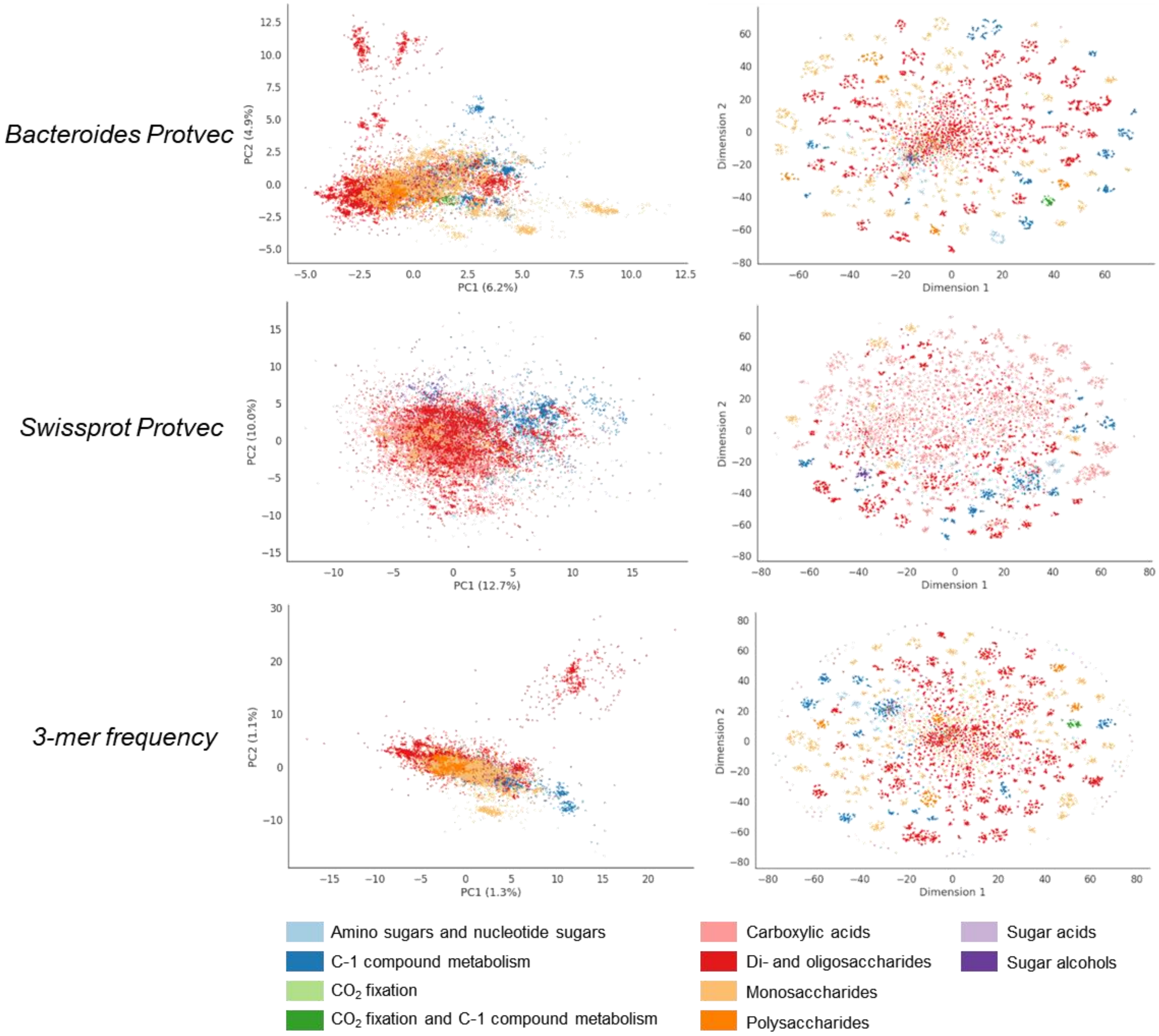
Xu, D & Nussinov, R 1998, 'Favorable domain size in proteins', *Folding and Design*, vol. 3, no. 1, pp. 11-17.

Yang, KK, Wu, Z, Bedbrook, CN & Arnold, FH 2018, 'Learned protein embeddings for machine learning', *Bioinformatics (Oxford, England)*, vol. 34, no. 15, pp. 2642-2648.

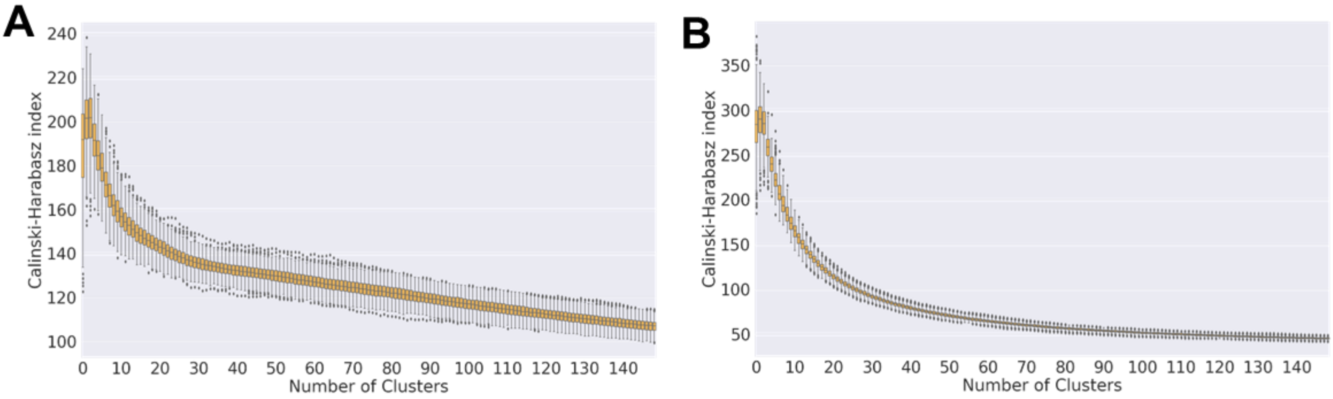
Appendices



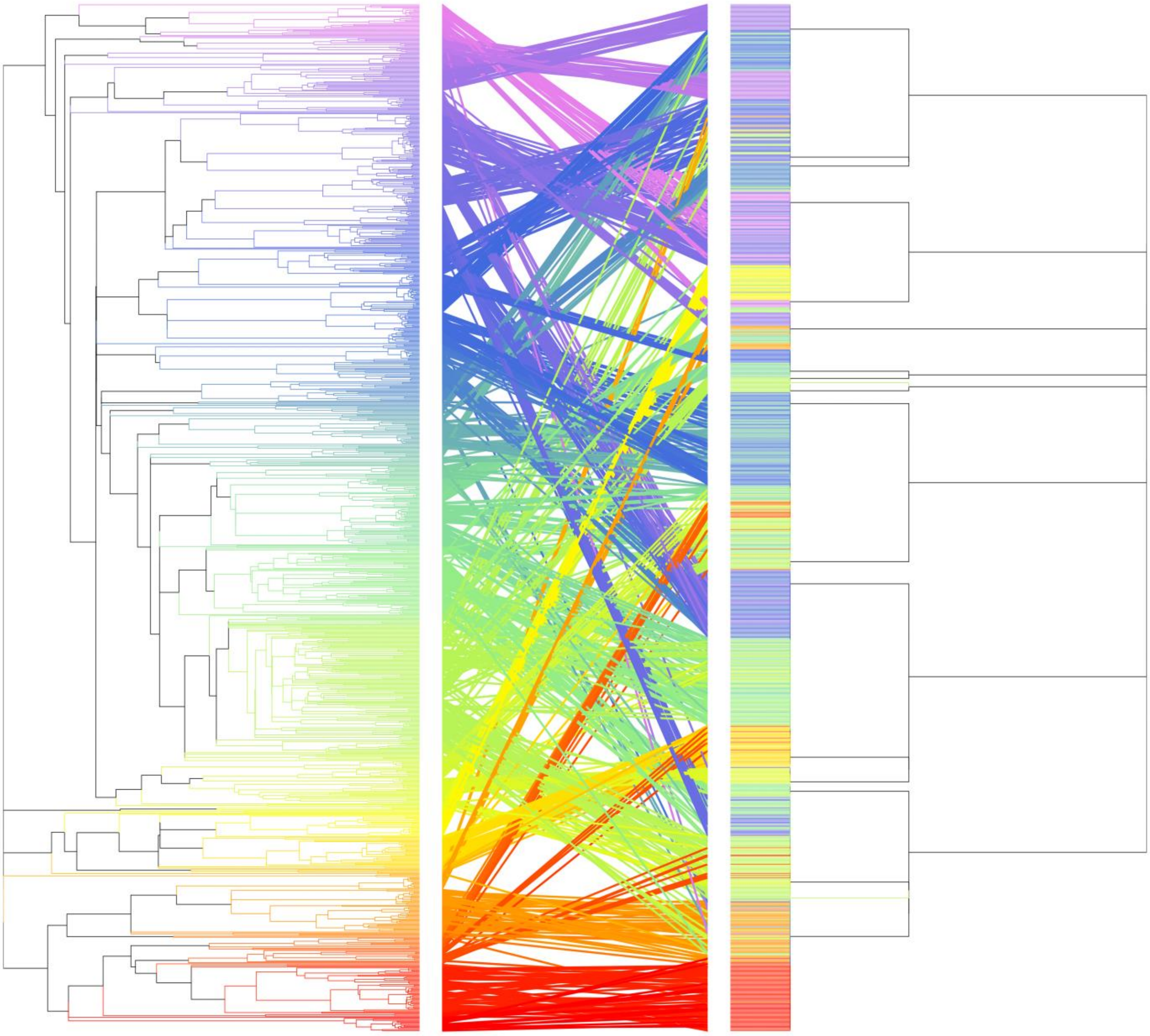
Appendix 1: A: k-means clustering of Swiss-Prot 3-mer vectors, **B:** Amino acid frequency of the k-mers in the yellow cluster in A, **C:** Amino acid frequency of the k-mers in the teal cluster in A.



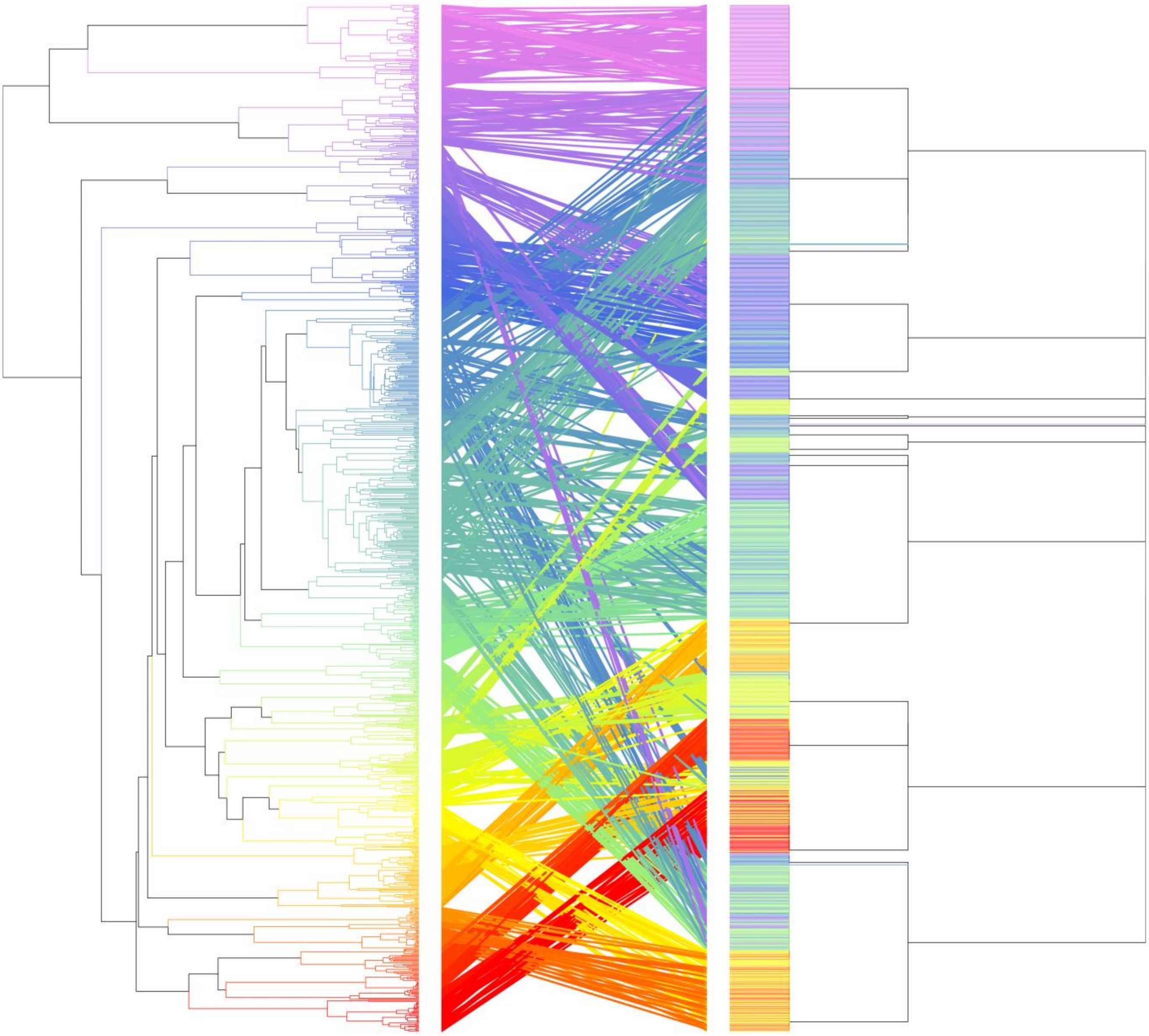
Appendix 2: Bacteroides carbohydrate metabolism sequences embedded using a Protvec model trained with Bacteroides Carbohydrate metabolism sequences, Protvec model trained with Swiss-Prot and k-mer frequency of the embedded sequences. Sequences are coloured by subclass and visualised with PCA and tSNE (perplexity = 30, learning rate = 100).



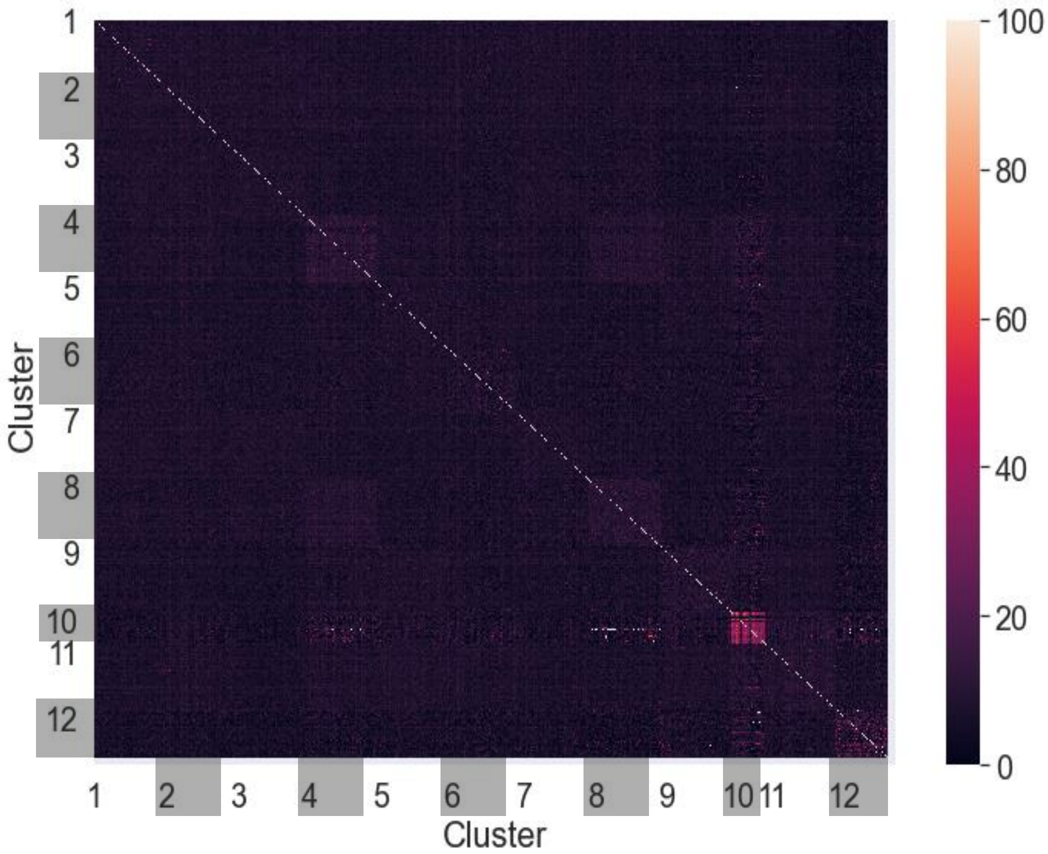
Appendix 3: Calinski-Harabasz index of 5,000 *Bacteroides* carbohydrate metabolism sequences embedded with **A:** Protvec model trained with *Bacteroides* carbohydrate metabolism sequences and **B:** Protvec model trained with Swiss-Prot. Within Cluster Sum of Squares of 5,000 Bacillus carbohydrate metabolism sequences embedded with **B:** Protvec model trained with *Bacteroides* carbohydrate metabolism sequences and **D:** Protvec model trained with Swiss-Prot. 500 bootstrap iterations were used for each value of k.



Appendix 4: Tanglegram of 1,000 Bacillus carbohydrate metabolism protein sequences. Hierarchy on the left is built using divisive clustering on the Euclidean distances between sequences embedded using a Protvec model trained with Bacillus sequences. Hierarchy on the right is the classification of sequences in the existing subsystems ontology



Appendix 5: Tanglegram of 1,000 Bacillus carbohydrate metabolism protein sequences. Hierarchy on the left is built using agglomerative clustering on the Manhattan distances between sequences embedded using a Protvec model trained with Bacillus sequences. Hierarchy on the right is the classification of sequences in the existing subsystems ontology.



Appendix 6: Heatmap of percent identity matrix of the unknown *Bacillus* sequences clustered in Fig. 7 using Clustal Omega (Sievers & Higgins 2014). Axes labels denote sequences from each cluster in Fig.7.