What the protein!? Computational methods for predicting microbial protein functions

Susanna R. Grigson[1]*, Robert A. Edwards[1]

[1] Flinders Accelerator for Microbiome Exploration, College of Science and Engineering, Flinders University, Adelaide, South Australia, Australia

* Corresponding author: Susanna R. Grigson, Email: susie.grigson@flinders.edu.au

Abstract

The identification of protein functions is crucial for understanding microbial life at a molecular scale. While computational methods for annotating protein sequences have greatly advanced in recent years, 30% of all bacterial and 65% of all viral protein sequences cannot be attributed a known biological function. As a result, protein function inference remains a fundamental challenge in computational biology. This paper reviews various bioinformatics methods for annotating microbial and viral proteins, categorised into homology-based and homology-free approaches. Widely used homology-based methods encompass sequence similarity searches such as BLAST and profile hidden Markov models, both of which compare novel protein sequences to databases of protein sequences with known functions. These homology-based methods have limitations, particularly for viral sequences which are severely underrepresented in protein sequence databases. As a result, homology-free methods, including numerical feature extraction, language-based models, guilt-by-association, and protein structure prediction software, offer potential alternatives. In addition, it is also important to critically consider the functional labels used to describe protein functions, and the hierarchical organisation of functional labels, regardless of the annotation method implemented. This review highlights that a combination of multiple functional prediction strategies, including machine learning, may provide the best improvements for microbial protein annotation and alleviate the ever-expanding sequence-function gap affecting microbial proteins. Overall, we provide experimental biologists with a comprehensive overview of annotation methods and inform computational scientists of open challenges and future research avenues.

Keywords: microbiology, protein function prediction, sequence annotation, machine learning, proteomics

Issue Section: Review

Introduction

Bacteria live in almost every environment and perform vital processes, sustaining ecological cycles and influencing human health. While it has been estimated that there are 10^{30} bacteria on Earth [1], it has been estimated that there are ten times as many viruses (10^{31}), the majority of which are bacteriophages (phages), viruses that infect bacteria [2]. These viruses significantly impact bacterial populations by inhibiting or killing susceptible bacteria, acting as conduits of genetic exchange and sinks of essential nutrients. Together, bacteria and phages have profound influences across diverse environments, including environmental, clinical and industrial settings [3–5].

Like all organisms, bacteria and the viruses which infect them contain genetic material which encodes genes. These genes encode proteins that enable bacteria and viruses to undertake virtually all their biological processes by catalysing metabolic reactions, providing structure, replicating DNA, enabling responses to stimuli, and transporting molecules from one location to another [6]. Each protein can be described as a long chain composed of 22 possible amino acids, abbreviated by alphabetical characters, and organised with a specific arrangement. Biochemical and biophysical interactions between these amino acids cause the protein to fold into a three-dimensional structure which facilitates a specific biological function. With the advent of low-cost, high-throughput sequencing technology and the increased prevalence of metagenomic studies which sequence entire microbial communities, the number of available protein sequences is ever-increasing. However, determining the function of these proteins experimentally relies on time-intensive and resource-demanding techniques such as mass spectrometry [7], microscopy [8] and pull-down assays [9]. As a result, wide gaps persist between the number of known protein sequences and known protein functions. This sequence-function gap is profound for sequences of microbial origin due to the globally high abundance of microorganisms and their massive functional diversity [10]. Strategies have been devised to infer protein sequence functions computationally to increase the proportion of sequences annotated with a known biological function. While this has improved annotation

rates, 30% of bacterial sequences [11] and 65% of phage sequences remain unannotated (See Methods). Consequently, novel strategies for predicting microbial protein functions are required to attain higher levels of functional annotation.

In this review, we examine the various computational techniques for assigning functional labels to bacterial and viral proteins. To begin, we outline the process of identifying protein-coding regions within microbial genomes and the many annotation systems available for functional labelling. Then, we examine the use of homology-based methods, such as sequence similarity searches and profile hidden Markov models for assigning functional descriptions to novel protein sequences, as well as their limitations. We also compare these widely-employed strategies with homology-free annotation methods and highlight the benefits and disadvantages of each. In doing so, we outline potential computational approaches which may enable the annotation of currently unannotated sequences, broadening our understanding of bacteria and viruses across diverse environments.

Finding Proteins

To locate protein-coding regions within bacterial or viral genomes, it is necessary to identify the open reading frame (ORF) of each gene present. This can be achieved through the use of various gene calling algorithms that search for specific patterns, such as start and stop codons marking the start and end of a protein-coding sequence, the Shine-Delgarno sequence, a short sequence near the start codon for initiating translation, and ribosome binding sites which facilitate protein translation [12]. Additionally, GC frame-plot bias, which refers to the distribution of specific nucleotides across codons, can assist in gene calling by identifying the reading frame most likely to be transcribed. As some codons are more commonly used in ORFs than others, codon usage bias can also guide the identification of ORFs [12]. For bacterial genomes, several ORF calling tools have been developed using different combinations of these features, including Prodigal [12], GeneMark [13], and Glimmer [14]. Although these tools are commonly used for viral genomes [15], viruses have distinct genomic

properties, such as shorter and overlapping genes and minimal non-coding DNA [16,17]. Thus, phage-specific gene-calling software such as PHANOTATE [18] should be used to identify phage genes. After protein-coding regions have been identified, several methods can be utilised to describe and predict the functions of these regions.

Describing Protein Functions

Functional ontologies

Ontologies are hierarchical systems that aid in describing the function of a protein. These systems are structured as a directed acyclic graph, where proteins are organised into broader and broader categories at each level of the hierarchy. For example, a protein might first be classified as an enzyme, then as a specific type of enzyme, and finally as a specific enzyme involved in a particular metabolic pathway. One of the most widely used ontologies is Gene Ontology (GO) [19], which provides three different labels to describe gene products: cellular component, molecular function, and biological process. Other popular ontologies include the Kyoto Encyclopedia of Genes and Genomes (KEGG) [20] and Metacyc [21], which primarily focus on metabolic and regulatory pathways. However, these ontologies are not specifically designed for microbial sequences and include pathways exclusive to higher organisms such as humans and plants. To address this, the SEED (RAST) Subsystems ontology [22] uses four hierarchical levels of annotation and specifically targets bacterial functions (Figure 1), potentially enabling more complete annotation of bacterial functional pathways [23]. Despite the efforts of human annotation experts, these ontologies are not guaranteed to be complete or arranged in the most biologically meaningful configuration.

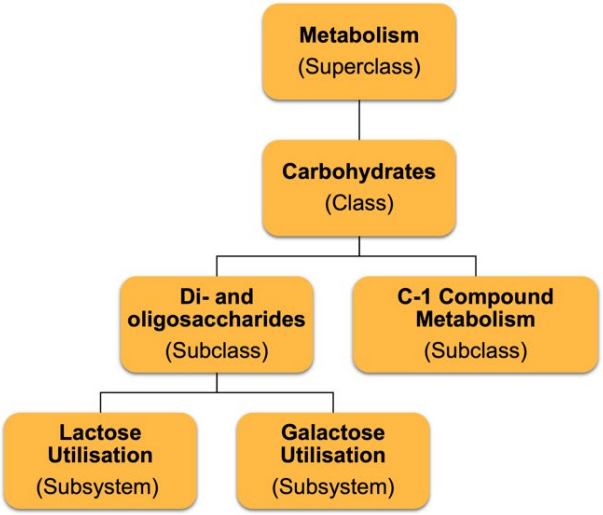

Figure 1. Example from the SEED subsystems hierarchy which demonstrates four levels of functional annotation: superclass, class, subclass and subsystem.

Orthologous groups

Orthologous groups offer an alternative approach for describing the function of protein sequences using evolutionary relationships. Orthologs are genes from different species that have evolved from a common ancestral gene and serve similar biological functions. These orthologs are grouped into orthologous groups using sequence alignments to evaluate sequence similarity and phylogenetic analysis to determine evolutionary relationships. This process is automated using pipelines and bioinformatic tools but may require some manual curation. Additionally, as new species and gene sequences become available, orthologous groups require revision and may change over time [24].

Detecting orthologs is complicated by convergent and divergent evolution. Two proteins with the same sequence might not have a common ancestor, and two proteins with a common ancestor may no longer have identifiable homology. Like functional ontologies, many different systems of orthologous groups have been devised. For instance, the bioinformatics tools eggNOG [25] and OrthoFinder [26] use different methods to identify orthologous groups in bacteria, which may lead to non-identical results. Additionally, the orthologous groups for viruses differ from those used for bacteria as they have different evolutionary histories and

belong to different domains of life. Thus, orthologous groups specific to viruses, such as PHROGs (Prokayotic Virus Remote Homologous Groups) [27], VOGs (Viral Orthologous Groups), and pVOGs (prokaryotic Virus Orthologous Groups) [28], have been developed. However, many orthologous groups are not associated with a known function; for example, 87% (33,747 out of 38,880) of PHROGs lack a known function.

Protein Domains

When examining protein functions, focusing on specific regions within sequences known as protein domains is possible. These are distinct, structurally and functionally independent protein regions that can have unique three-dimensional structures, perform specific biological functions, evolve, and can be inherited [29]. Identifying protein domains helps predict the potential function of a protein. This is executed by utilising databases that store protein sequences that confer activity, such as InterPro (formerly Pfam)[30], CATH [31] and SCOP (Structural Classification Of Proteins database) [32]. Once identified, the combination and arrangement of domains within a protein can be used to predict its function. These domains can also be used to categorise protein sequences, for example into superfamilies [33] or into FunFams [34], a system that groups domains with similar molecular functions, such as catalytic activity or ligand binding.

Annotating using homology

Sequence similarity searches

Sequence similarity searches are one of the most commonly used techniques for predicting protein functions. This process involves comparing a query sequence with large sequence databases to identify homologous sequences through sequence alignment. This functional annotation can be transferred to the query sequence if these homologous proteins have a known function.

The most cited similarity search tool is the Basic Local Alignment Search Tool (BLAST), developed in the 1990s, [35]. This algorithm uses heuristics to generate quick alignments and

comparisons of query sequences with huge databases, making it an indispensable tool in genomic research. BLAST uses a seed and extend algorithm, where *k*-mers (short subsequences of length *k*) are extracted from the query sequence and used to initiate (seed) searches in a predefined index of database sequences to identify candidate homologs. Alignments to the candidate sequences are then extended from the exactly matched subsequence. Next, each alignment is scored using a scoring matrix such as PAM (Point Accepted Mutation) [36] or BLOSUM (BLOcks SUbstitution Matrix) [37], which ranks the substitution of different amino acids, along with an affine scoring system that accounts for gaps, and the extension of gaps within the alignment [38]. The best alignment is determined by the bit-score, which measures the significance of the alignment, and the expectation value, or e-value, which represents the number of expected hits of similar quality by chance. Overall, this process is easily parallelised as each sequence scoring is independent.

The continued popularity of BLAST is partly due to its online availability via the National Center for Biotechnology Information (NCBI) website [39]. Despite this, alternatives have been developed, such as RAPSearch (Reduced Alphabet-based Protein similarity Search) [40,41], DIAMOND [42,43], and MMseqs2 [44]. These tools operate on similar principles to BLAST and offer increased speed with similar sensitivity.

Profile Hidden Markov Models

Functional labelling of protein sequence data can also be accomplished using probabilistic hidden Markov models that synthesise the progression of a series of observed states controlled by a set of unobservable hidden states. Profile hidden Markov models (pHMMs) are a particular application of hidden Markov models designed to capture the patterns and statistical properties of amino acid sequence alignments [45,46]. PHMMs allow the detection of homologs in protein sequence databases by converting multiple sequence alignments into position-specific scoring systems by representing each position in a multiple sequence alignment using three different types of hidden states; match states which represent a conserved amino acid, insertion states which model amino acids inserted between the *k*th and

(*k* + 1)th positions, and deletion states which represent the removal of the *k*th amino acid from the alignment (Figure 2). Each state transition is assigned a probability, and match and insertion states are assigned a separate probability for each possible amino acid. Using this process, a pHMM can be generated that statistically describes a family of protein sequences with a common function, such as an orthologous group. To annotate a protein sequence using pHMMs, the sequence is compared against a database of pHMMs representing known protein families and functional domains. The alignment score between the query sequence and the most similar pHMM is calculated, and if it exceeds a specified threshold, the functional annotation of the pHMM can be transferred to the query sequence [47,48].

PHMMs often produce more accurate results than traditional sequence similarity searches. They have been shown to detect three times more remote homologs than conventional pairwise methods such as BLAST [49], making them particularly useful for the annotation of viral sequences that are significantly underrepresented in public databases [50].

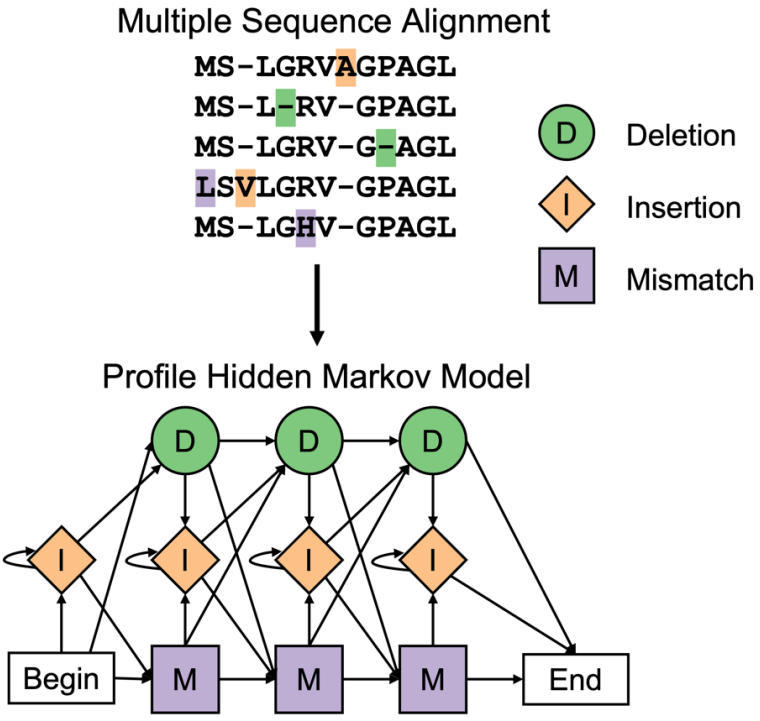

Figure 2. Profile hidden Markov model depicting a multiple sequence alignment of amino acid sequences. State transition probabilities are indicated as arrows.

Limitations of homology-based annotation

It is difficult to assign functions to many protein sequences using standard homology-based methods such as sequence similarity searches and profile hidden Markov models. Of the 230 million proteins deposited in UniProt [51], the leading protein sequence database, only 1.6 million protein sequences have been annotated with experimental evidence. Homology-based comparison has been used to assign functions to 73 million additional sequences, however, the majority of deposited sequences (103 million sequences) have no known function (Figure 3). Many of these sequences are of bacterial or viral origin, hindering our understanding of the most globally abundant organisms. Homology-free functional annotation methods, are in their infancy and not yet widely adopted, so homology-based strategies continue to dominate microbial functional annotation in bioinformatics.

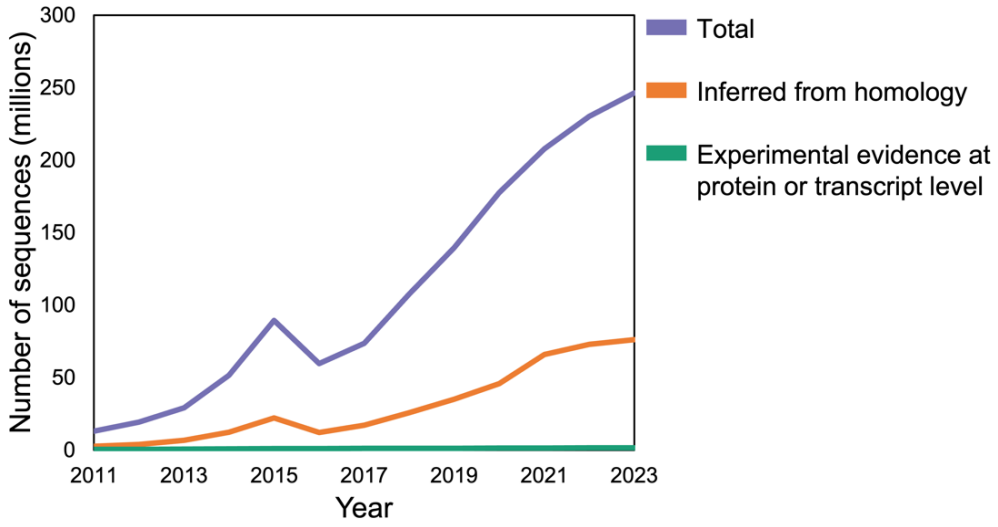

Figure 3. Total number of protein sequences, number of protein sequences inferred from homology and the number of protein sequences with experimental evidence at the protein or transcript level (millions) deposited in UniProtKB at the first major release of each year. Between 2015 and 2016, new procedures were implemented to identify and remove redundant sequences.

Homology-free annotation

Extracting Protein Sequence Features

One strategy for predicting the functions of viral and bacterial proteins without relying on sequence homology involves extracting numerical features from protein sequence data. These numerical features can be arranged to form a vector representing each protein sequence in a dataset, where each vector dimension reflects a different numerical property of the protein sequence.

Several different types of numerical features can be extracted from protein sequences. *K*-mer counts are among the most informative numerical features that can be extracted, as they provide information regarding a sequence's nucleotide or amino acid composition. *K*-mer counts can demonstrate the frequency with which specific amino acids are preceded or followed by other amino acids [52]. Different length *k*-mers can be easily extracted, and many tools use $2 \leq k \leq 12$ amino acids. However, storing *k*-mer counts generates sparse matrices, which are computationally intensive. A common solution is only to store *k*-mers that appear two or more times and to use a probabilistic data structure like a Bloom Filter or Counting Quotient Filter to remove singleton *k*-mers [53]. Other valuable numerical features include physicochemical properties, such as isoelectric point, aromaticity, molar extinction coefficient, instability index, molecular weight, polarity and hydrophobicity [54–56]. Each of these features can be calculated from the sequence alone using readily available bioinformatics tools [57]. Database-related features may also be extracted from protein sequence data, including the presence of motifs, protein subcellular location, binding preferences and the presence or absence of transmembrane regions, which are likely linked to protein function [55].

After a set of features has been extracted, feature selection procedures are implemented to eliminate redundant features and enhance accuracy. Once a set of features has been curated, machine learning techniques are applied to differentiate between protein functions. These techniques include Support Vector Machines [58–60] and Artificial Neural Networks [54], as

well as ensembles of multiple machine learning methods [55,56]. Several function predictors have been released as bioinformatic tools, including PhANNs (Phage Artificial Neural Networks) [54], which can classify phage proteins into one of ten structural classes with 86.2% accuracy. However, PhANNs can only annotate structural proteins, leaving many non-structural viral genes completely unannotated.

Language-Based Models

An alternative to manually extracting features from protein sequence data is to leverage language-based models used in natural processing to predict protein functions. These models, originally designed to evaluate text similarities, use a self-supervised approach to learn embedded representations of protein sequences which capture biophysical and biochemical properties of protein sequences [61]. As these properties are linked to protein function, proteins with similar functions have related embedded representations.

One of the most impactful models for embedding protein sequence data is Protvec [62,63], which utilises the *word2vec* language algorithm originally designed to learn the syntactic and semantic qualities of text [64,65]. This model treats entire protein sequences as sentences and overlapping *k*-mers as words to learn complex patterns relating to protein function using a skip-gram neural network. This allows protein sequences to be represented as *n*-dimensional vectors in a feature hyperspace where sequences with related functions are proximally located (Figure 4). With ongoing advancements in the field of natural language processing, there have been several advancements in protein embeddings since Protvec, resulting in advanced algorithms including SProBERTa [66], SeqVec [67] and ProteinBERT [68] which can distinguish between different families of protein sequences with high accuracy. Like numerical features extracted from protein sequences, machine learning classifiers can be trained using protein embedding vectors to assign functional labels to protein sequences [69]. Additionally, protein sequences may be clustered by calculating the distance between proteins within an embedding space. This clustering can help to generate more consistent protein families for functional annotation [70].

Although the features learnt by protein embeddings are biologically meaningful and outperform numerical features directly extracted from protein sequences [69,71,72], protein embeddings risk becoming 'black boxes' due to the difficulty associated with interpreting the learned features in terms of known protein properties. Nonetheless, these embeddings can still be used for clustering and classifying sequences as an alternative to sequence similarity searches [69,71,73]. Despite the potential of protein embeddings, studies that generate embeddings for viral and bacterial proteins are limited, and there are currently no bioinformatic tools that utilise embeddings for microbial gene annotation. Regardless, UniProt [51], the leading protein database, has begun including embeddings for all deposited protein sequences using the ProtNLM natural language processing model. This has led to the generation of functional predictions for approximately 49 million previously uncharacterised proteins [74].

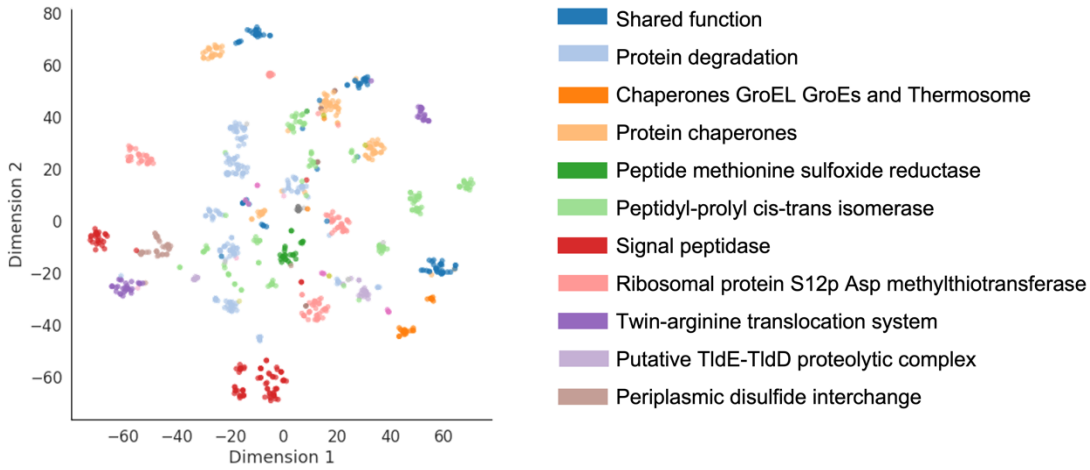

Figure 4. Example of a protein sequence embedding using the Protvec method to embed *Bacteriodes* proteins belonging to the 'Protein Fate (folding, modification, targeting, degradation)' SEED class. Each subclass is shown as a separate colour and visualised using *t*-distributed stochastic neighbourhood embedding [75]. It can be seen that the clusters formed are composed of sequences with related biological functions.

Guilt-by-association

Rather than using amino acid sequences to predict protein functions, interactions between proteins can be used to infer protein functions. Commonly referred to as ‘guilt-by-association’, this approach relies on the assumption that if an unknown protein physically interacts with a protein that has a known function, the unknown protein is likely to perform a related biological role [76,77].

Mathematical representations of the physical connections between proteins, known as protein-protein interaction (PPI) networks, are useful for implementing the guilt-by-association principle. Different strategies can be employed using these networks, such as neighbour-based methods, which annotate proteins based on the annotations of their neighbouring proteins, or module-based methods, which cluster PPI networks into groups of proteins with related functions [78,79]. The connectivity of proteins within a PPI network is another factor that can inform protein function as highly connected proteins are believed to perform essential biological functions and form a tightly connected core. In bacteria, these essential functions include processes related to translation, metabolism, transcription and replication [80]. Machine learning algorithms can also be trained on PPI networks to predict protein functions based using the interactions of known protein sequences [81]. These predictions can be strengthened by incorporating sequence homology information in conjunction with protein interaction data [81–84].

As there are limited experimentally determined protein interactions between phage and bacterial proteins, using guilt-by-association to predict microbial protein functions is challenging. To overcome this limitation, machine learning has been utilised to predict interactions between pairs of viral genes by extracting numerical features from pairs of protein amino acid sequences [85]. However, not all interacting proteins exhibit genomic similarities, making it difficult to predict all possible interactions using this strategy. Another method involves the use of the triadic closure, a social network theory that hypotheses that if two people have shared friends, then there is an increased probability they are also friends

themselves. This principle can be applied to discover additional pairs of interacting proteins based on existing connections within a PPI network [86]. Meanwhile, machine learning may be used to predict missing interactions within PPI networks using a generative adversarial network to predict probable interactions by extracting subnetworks of interacting proteins [87], or by embedding PPI networks using language models [88,89]. It is important to note that all of these interaction prediction methods require previously known protein interactions to build upon. An alternative method of predicting protein interactions is based on the co-evolution of amino acid residues in the sequences. This follows the premise that if one protein undergoes a mutation, its interacting partner must also undergo a compensatory mutation for a stable interaction between the proteins to persist. Hence, the co-evolution of two residues across different species or variants is a strong indication of a functional interaction between the corresponding proteins [90–93].

While guilt-by-association methods have been developed, their use for inferring microbial functions appears underutilised. Most guilt-by-association models rely on protein-protein interaction databases. These databases depend on experimentally determined interactions and are therefore limited.

Using computed protein structures

This review would not be complete without discussing software that infers the three-dimensional structures of proteins using just the amino acid sequence of a protein. This software includes AlphaFold [94], Colabfold [95], and RoseTTAFold [96] which utilise deep neural networks trained on multiple sequence alignments to extract interconnections between amino acid residues based on evolutionary and geometric constraints of protein structures. While AlphaFold produces more accurate structures, RoseTTAFold structures are generated faster. This has led to both tools being used in tandem to balance the benefits of both [97]. Regardless, many protein structures can be downloaded from online databases without needing to fold them, including over 200 million folded protein structures available through the AlphaFold Protein Structural database [98]. As protein structure is directly related to protein

function, these computed structures are beneficial for inferring the function of unknown proteins.

One approach to inferring protein functions involves aligning novel protein structures to known protein structures in large ‘foldome’ databases. This process relies on protein structure alignment tools [99,100] and has been successfully applied to predict the function of various pore-forming proteins [101]. If these folded structures do not align with any known protein structures, the structures may be used to determine interacting protein pairs via computational docking. Such docking predicts the most probable orientation in which two protein structures bind and the stability of this interaction [102,103] and can be used to make functional predictions using guilt-by-association. Alternatively, the coevolution of protein structures may be used to determine the likelihood that a pair of proteins interact by evaluating the extent to which the changes to one protein covary with changes to another protein structure [91,93]. This operates by aligning ortholog sequences of two proteins across multiple genomes to create a paired multiple-sequence alignment and computationally folding the alignment to determine the probability that an interacting interface is formed between the two proteins. This process has been successfully used to predict protein interactions in yeast [97], suggesting a similar approach may be possible for microbial proteins.

Despite the advancement of protein structure prediction software, AlphaFold and RoseTTAFold cannot learn the physical dynamics of protein folding or protein folds poorly represented in protein structural databases [104]. Therefore, structural predictions may be limited for proteins with novel functions. Regardless, predicted protein structures have significantly broadened the structural coverage of the human proteome [105] and offer the potential to minimise the proportion of uncharacterised microbial sequences.

Conclusions and perspectives

Despite multiple annotation strategies, accurately inferring microbial protein functions remains a significant challenge impacting microbial research. Here, we have reviewed existing

computational methods for predicting the function of bacterial and viral proteins. These strategies include sequence similarity searches such as BLAST which are commonly used for identifying homologous sequences, though frequently fail to identify homologs for divergent sequences. Meanwhile, pHMMs extract relationships between amino acid residues and protein structures, enabling the detection of distant sequences with related functions. Despite this, pHMMs require large volumes of high-quality and diverse training data and can be computationally intensive [106]. While these homology-based methods are widely utilised for sequence annotation they are successful for only a subset of all known microbial sequences.

Homology-free strategies, while underutilised, offer potential solutions to close the ever-growing protein annotation gap. Despite this, each homology-free annotation method has associated benefits and limitations. For instance, extracting numerical features from protein sequences is fast but requires human input to identify suitable sequence features. Alternatively, language-based models can self-learn protein features, however, these features are not explainable in a biological context. Guilt-by association does not require protein sequences and rather utilises protein interaction data, but is limited by data scarcity as there are limited known bacterial and viral protein interactions. Alternatively, protein structural prediction software has the potential to predict microbial functions using computed protein structures, but performance may be reduced for novel or poorly represented folds (Table 1). As a result, a combination of multiple protein annotation methods may be necessary to successfully discover unknown microbial functions.

With the emergence of advanced machine learning algorithms including neural networks, machine learning has been increasingly utilised to predict protein functions. These algorithms can consider a large range of sequence features and self-learn complex features. Where possible, these models should be analysed carefully to interpret which sequence properties machine learning algorithms extract to gain more significant biological insights into protein function. Furthermore, multiple machine learning architectures should be compared to

establish the most optimal predictor for each prediction task. Overall, machine learning offers monumental potential for protein annotation tasks.

Method	Homology-based	Advantages	Disadvantages	Literature
Sequence Similarity	Yes	Easily available, fast.	Does not always provide the best hit.	[35,40,42–44]
Profile Hidden Markov Models	Yes	Robust to divergence. Capture relationships between amino acid residues.	Require large volumes of training data. May overfit.	[47,48]
Numerical protein features	No	Fast. Reduced reliance on databases.	Features require extraction. Can only extract known features.	[54–56]
Language-based models	No	Fast. Extract complex features relevant to protein function.	Learnt features cannot be easily interpreted.	[62,66–68,73]
Guilt-by-association	No	Does not rely on amino acid sequences.	Limited known protein-protein interactions.	[77–79,81]
Protein Structures	No	Incorporate deep protein structural information.	Folding protein structures is computationally expensive. Many proteins have novel structures.	[94–97,100]

Table 1. Summary of strategies for predicting unknown microbial functions

In summary, the ability to annotate protein functions provides useful information on fundamental microbial functions. Improving the annotation of microbial proteins will enable a greater understanding of bacteria, and the viruses which infect them, enhancing our understanding of complex biological phenomena, and providing insights into human and environmental health.

Key Points

- We review computational function prediction methods for assigning functional labels to bacterial and viral protein sequences.
- Homology-based protein function prediction strategies are routinely used, however, are unsuccessful for millions of bacterial and viral proteins.
- Homology-free annotation methods, many of which utilise machine learning, have been devised and can potentially be utilised to better understand fundamental microbial processes.
- Novel computational methods should be further explored to reduce the ever-expanding sequence function gap affecting bacteria and viruses.

Methods

Estimating unknown viral proteins

All proteins present in the 24,106 phage genomes available in NCBI Genbank were annotated as either hypothetical or phage-like. Hypothetical proteins were defined using the script https://github.com/linsalrob/EdwardsLab/blob/master/roblib/functions.py and phage-like proteins were defined using the script https://github.com/linsalrob/PhiSpy/blob/master/PhiSpyModules/protein_functions.py. A total of 1,373,232 proteins were annotated as hypothetical (65%), 277,157 proteins were annotated as phage-like (13%) and the remaining 463,612 proteins were not phage-like or hypothetical (22%).

Protvec embedding

Bacteriodes amino acid sequences were obtained from the Pathosystems Resources Integration Center (PATRIC) [107] and annotated with their SEED annotations [22]. Sequences that were annotated with the 'Protein Fate (folding, modification, targeting, degradation)' SEED class (3,396 sequences) were embedded using a Protvec model trained

on all sequences in the SwissProt database [62] using embedding scripts from previous work [73]. Embedded proteins were visualised using *t*-distributed stochastic neighbourhood embedding [75].

Data Availability

All data used to generate figures is publicly available. All of the phage genomes available on NCBI Genbank were obtained from the INPHARED data set (1st of March 2023, https://millardlab-inphared.s3.climb.ac.uk/1Mar2023_data.tsv.gz) [108].

The proportion of sequences with functional annotations in the UniProt database (Figure 2) was obtained using the release notes for the UniProtKB/TrEMBL protein database via the UniProt FTP site (https://ftp.uniprot.org/pub/databases/uniprot/).

Acknowledgements

We acknowledge the Flinders University Deep Thought High-Performance Computing platform [109] for computing resources used to generate figures.

Funding

This work was supported by an award from the National Institutes of Health (NIH), National Institute of Diabetes and Digestive and Kidney Diseases [RC2DK116713 to R.A.E] and an award from the Australian Research Council [DP220102915 to R.A.E].

Authors' contributions

S.R.G wrote the manuscript, performed analysis and made figures. R.A.E performed the analysis and reviewed the manuscript.

Author Biographies

Susanna R. Grigson is a bioinformatics PhD student at the Flinders Accelerator for Microbiome Exploration, Flinders University. Her current research focuses on machine learning methods for the functional prediction of microbial proteins.

Robert A. Edwards is the Matthew Flinders Professor of Bioinformatics at the Flinders Accelerator for Microbiome Exploration, Flinders University. His current research focuses on using metagenomics and viromics to understand how viruses control bacteria in the environment.

References

1. Magnabosco C, Lin L-H, Dong H, et al. The biomass and biodiversity of the continental subsurface. *Nature Geoscience* 2018; 11:707–717.

2. Breitbart M, Rohwer F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* 2005; 13:278–284.

3. Carlson-Jones JAP, Kontos A, Kennedy D, et al. The microbial abundance dynamics of the paediatric oral cavity before and after sleep. *J. Oral Microbiol.* 2020; 12:1741254.

4. Xu Z, Walker ME, Zhang J, et al. Exploring the diversity of bacteriophage specific to Oenococcus oeni and Lactobacillus spp and their role in wine production. *Appl. Microbiol. Biotechnol.* 2021; 105:8575–8592.

5. Silveira CB, Coutinho FH, Cavalcanti GS, et al. Genomic and ecological attributes of marine bacteriophages encoding bacterial virulence genes. *BMC Genomics 2020*; 21:126.

6. Typas A, Sourjik V. Bacterial protein networks: properties and functions. *Nat. Rev*. *Microbiol.* 2015; 13:559–572.

7. Liu R, Xia S, Li H. Native top-down mass spectrometry for higher-order structural characterization of proteins and complexes. *Mass Spectrom.* Rev. 2022; e21793.

8. Hu J, Worrall LJ, Hong C, et al. Cryo-EM analysis of the T3S injectisome reveals the structure of the needle and open secretin. *Nat. Commun.* 2018; 9:3840.

9. Louche A, Salcedo SP, Bigot S. Protein–Protein Interactions: Pull-Down Assays. Bacterial *Protein Secretion Systems: Methods and Protocols* 2017; 247–255.

10. Escalas A, Hale L, Voordeckers JW, et al. Microbial functional diversity: From concepts to applications. *Ecology and Evolution* 2019; 9:12000–12016.

11. Vanni C, Schechter MS, Acinas SG, et al. Unifying the known and unknown microbial coding sequence space. *Elife* 2022; 11: e67667.

12. Hyatt D, Chen G-L, Locascio PF, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010; 11:119.

13. Borodovsky M, Mills R, Besemer J, et al. Prokaryotic gene prediction using GeneMark and GeneMark.hmm. *Curr. Protoc. Bioinformatics* 2003; 1:4-5.

14. Delcher AL, Harmon D, Kasif S, et al. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* 1999; 27:4636–4641.

15. McNair K, Aziz RK, Pusch GD, et al. Phage Genome Annotation Using the RAST Pipeline. *Methods Mol. Biol.* 2018; 1681:231–238.

16. Kang HS, McNair K, Cuevas DA, et al. Prophage genomics reveals patterns in phage genome organization and replication. *bioRxiv* 2017; 114819.

17. Cahill J, Rajaure M, O'Leary C, et al. Genetic Analysis of the Lambda Spanins Rz and Rz1: Identification of Functional Domains. *G3* 2017; 7:741–753.

18. McNair K, Zhou C, Dinsdale EA, et al. PHANOTATE: a novel approach to gene identification in phage genomes. *Bioinformatics* 2019; 35:4537–4542.

19. The Gene Ontology Consortium The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* 2021; 49:D325–D334.

491 20. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids*
492 *Res.* 2000; 28:27–30.

493 21. Caspi R, Billington R, Keseler IM, et al. The MetaCyc database of metabolic pathways
494 and enzymes - a 2019 update. *Nucleic Acids Res.* 2020; 48:D445–D453.

495 22. Overbeek R, Olson R, Pusch GD, et al. The SEED and the Rapid Annotation of microbial
496 genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* 2014; 42:D206–14.

497 23. Shrestha P, Kim M-S, Elbasani E, et al. Prediction of trehalose-metabolic pathway and
498 comparative analysis of KEGG, MetaCyc, and RAST databases based on complete genome
499 of Variovorax sp. PAMC28711. *BMC Genom Data* 2022; 23:4.

500 24. Glover N, Dessimoz C, Ebersberger I, et al. Advances and Applications in the Quest for
501 Orthologs. *Mol. Biol. Evol.* 2019; 36:2157–2164.

502 25. Hernández-Plaza A, Szklarczyk D, Botas J, et al. eggNOG 6.0: enabling comparative
503 genomics across 12 535 organisms. *Nucleic Acids Res.* 2023; 51:D389–D394.

504 26. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative
505 genomics. *Genome Biol.* 2019; 20:238.

506 27. Terzian P, Olo Ndela E, Galiez C, et al. PHROG: families of prokaryotic virus proteins
507 clustered using remote homology. *NAR Genom Bioinform* 2021; 3:lqab067.

508 28. Grazziotin AL, Koonin EV, Kristensen DM. Prokaryotic Virus Orthologous Groups
509 (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids*
510 *Res.* 2017; 45:D491–D498.

511 29. Das S, Orengo CA. Protein function annotation using protein domain family resources.
512 *Methods* 2016; 93:24–34.

513 30. Paysan-Lafosse T, Blum M, Chuguransky S, et al. InterPro in 2022. *Nucleic Acids Res*.
514 2023; 51:D418–D427.

515 31. Sillitoe I, Bordin N, Dawson N, et al. CATH: increased structural coverage of functional
516 space. *Nucleic Acids Res.* 2021; 49:D266–D273.

517 32. Andreeva A, Kulesha E, Gough J, et al. The SCOP database in 2020: expanded
518 classification of representative family and superfamily domains of known protein structures.
519 *Nucleic Acids Res.* 2020; 48:D376–D382.

520 33. Wu CH, Nikolskaya A, Huang H, et al. PIRSF: family classification system at the Protein
521 Information Resource. *Nucleic Acids Res.* 2004; 32:D112–4.

522 34. Scheibenreif L, Littmann M, Orengo C, et al. FunFam protein families improve residue
523 level molecular function prediction. *BMC Bioinformatics* 2019; 20:400.

524 35. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J. Mol. Biol.* 1990;
525 215:403–410.

526 36. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. In
527 Dayhoff MO, Ech RV (eds), *Atlas of Protein Sequence and Structure.* Maryland: National
528 Biomedical Research Foundation; 1978, 345–52.

529 37. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc. Natl.*
530 *Acad. Sci. U. S. A.* 1992; 89:10915–10919.

531 38. Altschul SF. Generalized affine gap costs for protein sequence alignment. *Proteins* 1998;
532 32:88–96.

533 39. Johnson M, Zaretskaya I, Raytselis Y, et al. NCBI BLAST: a better web interface. *Nucleic*
534 *Acids Res.* 2008; 36:W5–9.

535 40. Ye Y, Choi J-H, Tang H. RAPSearch: a fast protein similarity search tool for short reads.
536 *BMC Bioinformatics* 2011; 12:159.

537 41. Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity
538 search tool for next-generation sequencing data. *Bioinformatics* 2012; 28:125–126.

539 42. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using
540 DIAMOND. *Nat. Methods* 2021; 18:366–368.

541 43. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND.
542 *Nat. Methods* 2015; 12:59–60.

543 44. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the
544 analysis of massive data sets. *Nat. Biotechnol*. 2017; 35:1026–1028.

545 45. Yoon B-J. Hidden Markov Models and their Applications in Biological Sequence
546 Analysis. *Curr. Genomics* 2009; 10:402–415.

547 46. Vijayabaskar MS. Introduction to Hidden Markov Models and Its Applications in Biology.
548 *Methods Mol. Biol.* 2017; 1552:1–12.

549 47. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998; 14:755–763.

550 48. Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* 2011; 7:e1002195.

551 49. Park J, Karplus K, Barrett C, et al. Sequence comparisons using multiple sequences
552 detect three times as many remote homologues as pairwise methods. *J. Mol. Biol*. 1998;
553 284:1201–1210.

554 50. Mahmoudabadi G, Phillips R. A comprehensive and quantitative exploration of
555 thousands of viral genomes. Elife 2018; 7: e31955.

556 51. The UniProt Consortium UniProt: the Universal Protein knowledgebase in 2023. *Nucleic*
557 *Acids Res.* 2023; 51:D523–D531.

558 52. Déraspe M, Boisvert S, Laviolette F, et al. Flexible protein database based on amino
559 acid k-mers. *Sci. Rep.* 2022; 12:9101.

560 53. Pandey P, Bender MA, Johnson R, et al. Squeakr: an exact and approximate k-mer
561 counting system. *Bioinformatics* 2018; 34:568–575.

562 54. Cantu VA, Salamon P, Seguritan V, et al. PhANNs, a fast and accurate tool and web
563 server to classify phage structural proteins. *PLoS Comput. Biol.* 2020; 16:e1007845.

564 55. Mishra S, Rastogi YP, Jabin S, et al. A deep learning ensemble for function prediction of
565 hypothetical proteins from pathogenic bacterial species. *Comput. Biol. Chem.* 2019;
566 83:107147.

567 56. Ahmad S, Charoenkwan P, Quinn JMW, et al. SCORPION is a stacking-based ensemble
568 learning framework for accurate prediction of phage virion proteins*. Sci. Rep.* 2022; 12:4106.

569 57. Cock PJA, Antao T, Chang JT, et al. Biopython: freely available Python tools for
570 computational molecular biology and bioinformatics. *Bioinformatics* 2009; 25:1422–1423.

571 58. Saha S, Raghava GPS. VICMpred: an SVM-based method for the prediction of
572 functional proteins of Gram-negative bacteria using amino acid patterns and composition.
573 *Genomics Proteomics Bioinformatics* 2006; 4:42–47.

574 59. Han H, Zhu W, Ding C, et al. iPVP-MCV: A Multi-Classifier Voting Model for the Accurate
575 Identification of Phage Virion Proteins. *Symmetry* 2021; 13:1506.

576 60. Manavalan B, Shin TH, Lee G. PVP-SVM: Sequence-Based Prediction of Phage Virion
577 Proteins Using a Support Vector Machine. *Front. Microbiol.* 2018; 9:476.

578 61. Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling
579 unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* 2021;
580 118: e2016239118.

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

583 63. Asgari E, McHardy AC, Mofrad MRK. Probabilistic variable-length segmentation of
584 protein sequences for discriminative motif discovery (DiMotif) and sequence embedding
585 (ProtVecX). *Sci. Rep*. 2019; 9:3577.

586 64. Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in
587 Vector Space. *arXiv* 2013; *1301.3781*.

588 65. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases
589 and their compositionality. *arXiv* 2013*:*1310.4546.

590 66. Wu L, Yin C, Zhu J, et al. SPRoBERTa: protein embedding learning with local fragment
591 modeling. *Brief. Bioinform.* 2022; 23.

592 67. Heinzinger M, Elnaggar A, Wang Y, et al. Modeling aspects of the language of life
593 through transfer-learning protein sequences. *BMC Bioinformatics* 2019; 20:723.

594 68. Brandes N, Ofer D, Peleg Y, et al. ProteinBERT: A universal deep-learning model of
595 protein sequence and function. *Bioinformatics* 2022; 38:2102–2110.

596 69. Bileschi ML, Belanger D, Bryant DH, et al. Using deep learning to annotate the protein
597 universe. *Nat. Biotechnol.* 2022; 932-937.

598 70. Littmann M, Bordin N, Heinzinger M, et al. Clustering FunFams using sequence
599 embeddings improves EC purity. *Bioinformatics* 2021; 3449-3455.

71. ElAbd H, Bromberg Y, Hoarfrost A, et al. Amino acid encoding for deep learning applications. *BMC Bioinformatics* 2020; 21:235.

72. Villegas-Morcillo A, Makrodimitris S, van Ham RCHJ, et al. Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function. *Bioinformatics* 2021; 37:162–170.

73. Grigson SR, McKerral JC, Mitchell JG, et al. Organizing the bacterial annotation space with amino acid sequence embeddings. *BMC Bioinformatics* 2022; 23:385.

74. Gane A, Bileschi ML, Dohan D, et al. ProtNLM: Model-based Natural Language Protein Annotation. 2022. https://www.uniprot.org/help/ProtNLM (1 April 2023, date last accessed).

75. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 2008; 9.

76. Oliver S. Guilt-by-association goes global. *Nature* 2000; 403:601–603.

77. Piovesan D, Giollo M, Ferrari C, et al. Protein function prediction using guilty by association from interaction networks. *Amino Acids* 2015; 47:2583–2592.

78. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. *Mol. Syst. Biol.* 2007; 3:88.

79. Vazquez A, Flammini A, Maritan A, et al. Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.* 2003; 21:697–700.

80. Dilucca M, Cimini G, Giansanti A. Bacterial Protein Interaction Networks: Connectivity is Ruled by Gene Conservation, Essentiality and Function. *Curr. Genomics* 2021; 22:111–121.

81. Zhang F, Song H, Zeng M, et al. DeepFunc: A Deep Learning Framework for Accurate Prediction of Protein Functions from Protein Sequences and Interactions. *Proteomics* 2019; 19:e1900019.

82. Kulmanov M, Khan MA, Hoehndorf R, et al. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* 2018; 34:660–668.

83. Piovesan D, Tosatto SCE. INGA 2.0: improving protein function prediction for the dark proteome. *Nucleic Acids Res*. 2019; 47:W373–W378.

84. Li M, Shi W, Zhang F, et al. A deep learning framework for predicting protein functions with co-occurrence of GO terms. IEEE/ACM Trans. *Comput. Biol. Bioinform.* 2022.

85. Pappas N, Dutilh BE. Finding functional associations between prokaryotic virus orthologous groups: a proof of concept. *BMC Bioinformatics* 2021; 22:438.

86. Kovács IA, Luck K, Spirohn K, et al. Network-based prediction of protein interactions. *Nat. Commun.* 2019; 10:1240.

87. Balogh OM, Benczik B, Horváth A, et al. Efficient link prediction in the protein–protein interaction network using topological information in a generative adversarial network machine learning model. *BMC Bioinformatics* 2022; 23:1–19.

88. Zhong X, Rajapakse JC. Graph embeddings on gene ontology annotations for protein–protein interaction prediction. *BMC Bioinformatics* 2020; 21:560.

89. Tsukiyama S, Hasan MM, Fujii S, et al. LSTM-PHV: prediction of human-virus protein–protein interactions by LSTM with word2vec. *Brief. Bioinform.* 2021; 22:bbab228.

90. Caporaso JG, Smit S, Easton BC, et al. Detecting coevolution without phylogenetic trees? Tree-ignorant metrics of coevolution perform as well as tree-aware metrics. *BMC Evol. Biol.* 2008; 8:327.

91. Cong Q, Anishchenko I, Ovchinnikov S, et al. Protein interaction networks revealed by proteome coevolution. *Science* 2019; 365:185–189.

92. Croce G, Gueudré T, Ruiz Cuevas MV, et al. A multi-scale coevolutionary approach to predict interactions between protein domains. *PLoS Comput. Biol.* 2019; 15:e1006891.

93. Green AG, Elhabashy H, Brock KP, et al. Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nat. Commun.* 2021; 12:1396.

94. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021; 596:583–589.

95. Mirdita M, Schütze K, Moriwaki Y, et al. ColabFold: making protein folding accessible to all. *Nat. Methods* 2022; 19:679–682.

96. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021; 373:871–876.

97. Humphreys IR, Pei J, Baek M, et al. Computed structures of core eukaryotic protein complexes. *Science* 2021; 374:eabm4805.

98. Varadi M, Anyango S, Deshpande M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 2022; 50:D439–D444.

662 99. Holm L. Dali server: structural unification of protein families. *Nucleic Acids Res.* 2022;
663 50:W210–5.

664 100. van Kempen M, Kim SS, Tumescheit C, et al. Foldseek: fast and accurate protein
665 structure search. *bioRxiv* 2022; 2022.02.07.479398.

666 101. Bayly-Jones C, Whisstock JC. Mining folded proteomes in the era of accurate structure
667 prediction. *PLoS Comput. Biol.* 2022; 18:e1009930.

668 102. Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions
669 using AlphaFold2. *Nat. Commun.* 2022; 13:1265.

670 103. Papudeshi B, Vega AA, Souza C, et al. Novel crAssphage isolates exhibit conserved
671 gene order and purifying selection of the host specificity protein. *bioRxiv* 2023;
672 2023.03.05.531146.

673 104. Outeiral C, Nissley DA, Deane CM. Current structure predictors are not learning the
674 physics of protein folding. *Bioinformatics* 2022; 881-1887.

675 105. Porta-Pardo E, Ruiz-Serra V, Valentini S, et al. The structural coverage of the human
676 proteome before and after AlphaFold. *PLoS Comput. Biol.* 2022; 18:e1009818.

677 106. Reyes A, Alves JMP, Durham AM, et al. Use of profile hidden Markov models in viral
678 discovery: Current insights. *Advances in Genomics and Genetics* 2017; 7:29.

679 107. Wattam AR, Abraham D, Dalay O, et al. PATRIC, the bacterial bioinformatics database
680 and analysis resource. *Nucleic Acids Res.* 2014; 42:D581–91.

681 108. Cook R, Brown N, Redgwell T, et al. INfrastructure for a PHAge REference Database:
682 Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes.
683 *PHAGE* 2021; 2:214–223.

684 109. Flinders University. Deep Thought (HPC). 2021.