Computational and Structural Biotechnology Journal 20 (2022) 434–442

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini review

Using metagenomic data to boost protein structure prediction and discovery

Qingzhen Hou [a,b,1], Fabrizio Pucci [c,d,1], Fengming Pan [a,b], Fuzhong Xue [a,b], Marianne Rooman [c,d], Qiang Feng [e,f,*]

[a] Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Shandong 250012, China
[b] National Institute of Health Data Science of China, Shandong University, Shandong 250002, China
[c] Computational Biology and Bioinformatics, Université Libre de Bruxelles, 1050 Brussels, Belgium
[d] Interuniversity Institute of Bioinformatics in Brussels, 1050 Brussels, Belgium
[e] Shandong Provincial Key Laboratory of Oral Tissue Regeneration & Shandong Engineering Laboratory for Dental Materials and Oral Tissue Regeneration, Department of Human Microbiome, School of Stomatology, Shandong University, Jinan, Shandong Province 250012, China
[f] State Key Laboratory of Microbial Technology, Shandong University, Qingdao, Shandong Province 266237, China

ARTICLE INFO

Article history:
Received 12 July 2021
Received in revised form 17 December 2021
Accepted 21 December 2021
Available online 3 January 2022

Keywords:
Metagenomics
Multiple sequence alignment
Enzyme design
CRISPR-Cas system
Antibiotic resistance
Microbiome

ABSTRACT

Over the past decade, metagenomic sequencing approaches have been providing an ever-increasing amount of protein sequence data at an astonishing rate. These constitute an invaluable source of information which has been exploited in various research fields such as the study of the role of the gut microbiota in human diseases and aging. However, only a small fraction of all metagenomic sequences collected have been functionally or structurally characterized, leaving much of them completely unexplored. Here, we review how this information has been used in protein structure prediction and protein discovery. We begin by presenting some widely used metagenomic databases and analyze in detail how metagenomic data has contributed to the impressive improvement in the accuracy of structure prediction methods in recent years. We then examine how metagenomic information can be exploited to annotate protein sequences. More specifically, we focus on the role of metagenomes in the discovery of enzymes and new CRISPR-Cas systems, and in the identification of antibiotic resistance genes. With this review, we provide an overview of how metagenomic data is currently revolutionizing our understanding of protein science.

Contents

* Corresponding author at: Shandong Provincial Key Laboratory of Oral Tissue Regeneration & Shandong Engineering Laboratory for Dental Materials and Oral Tissue Regeneration, Department of Human Microbiome, School of Stomatology, Shandong University, Jinan, Shandong Province 250012, China.
E-mail address: fengqiang@sdu.edu.cn (Q. Feng).
[1] Contributed equally to this work.

https://doi.org/10.1016/j.csbj.2021.12.030

1. Introduction

The use of metagenomic sequencing is dramatically improving our understanding of the evolution and ecology of microbial systems in various environments, from water and soil to the human body [1–3]. For example, metagenomics has been essential in revealing the mechanism of certain human diseases by detecting changes in the gut microbiome [4–6] and in identifying and controlling pathogens [7]. These advances have been made possible by metagenomics high-throughput sequencing techniques, through which billions of protein sequences have been characterized. Meanwhile, their number continues to grow at an impressive rate. These huge amounts of data are an invaluable source of information that has a big impact in different areas of protein science.

Protein three-dimensional (3D) structure prediction is one of these areas. Since a seminal article [8], metagenomic sequence data has been widely used to construct multiple sequence alignments (MSA) of target proteins, which are used as inputs to deep learning models for structure prediction. Metagenomic information has significantly contributed to improving the accuracy of the predictors [8], which have achieved astonishing scores in recent years [9,10]. Many studies have also been devoted to understanding the functions of proteins from metagenomic assembly, even though only a small part is functionally annotated. Metagenomic data constitute a huge reservoir of information that can be exploited to discover new proteins with specific functions. Indeed, they have proven to be a fundamental resource for discovering new enzymes with given stability properties [11–18], exploring antibiotic resistance genes in different microbial communities [19–23] and identifying new CRISPR-Cas systems [24–28].

In the next sections, we review widely known metagenomic databases and their characteristics, and show how this huge amount of information is used to improve the abovementioned research fields, as schematically depicted in Fig. 1.

2. Metagenomic resources and databases

We start by reviewing the widely known and curated metagenome resources and databases: IMG/M [29], MGnify [30], MetaClust [31] and BFD [32]. These databases are extensively used by the research community in a wide range of studies, *e.g.*, protein structure prediction [8,9], metabolic gene cluster discovery [33], enzyme discovery [11], and gene function prediction [34]. Their characteristics and content, dated December 2021, are described below; further details can be found in Table S1 of Supplementary Material:

- **IMG/M.** The Integrated Microbial Genomes and Microbiomes [29] is a comprehensive data management resource for the analysis of annotated genomic and metagenomic sequence data. It is increasing rapidly, reaching about 360 million genes from isolated genomes and 66 billion genes from metagenomes. The latter mainly come from human gut microbiome and from marine and freshwater microbial systems (see Fig. 2.a). The genomes and metagenomes with their metadata attributes were collected from the manually curated GOLD database [35] and then annotated with the IMG annotation pipeline [36]. Protein-coding genes were identified from (meta) genomic data by the prediction program Prodigal [37], and functionally annotated by using hidden Markov model (HMM)-based homologous sequence searches [38].
  IMG/M includes a set of genomic tools for data analysis, such as IMG/ABC for the study of biosynthetic gene clusters and secondary metabolites, and IMG/VR for the analysis of viral genome fragments derived from metagenomic samples. It also provides multi-search capabilities to search the database for, *e.g.*, homologous proteins of a target sequence via BLAST [39], KEGG enzyme classes and pathways [40,41], CATH families [42] and Pfam domains [43].
- **MGnify.** It is a comprehensive hub for the analysis, exploration and archiving of microbiome information [30]. It is one of the world's largest resources of microbiome data, and a user-friendly platform integrating multiple genomic tools, which makes MGnify widely used. A total of about 4,000 publicly available studies corresponding to about 325,000 samples and 437,000 analyses were deposited in the database. These numbers are constantly growing and have doubled in the last two years.
  MGnify provides a non-redundant protein set generated from the analysis of all the assembled datasets, which contains more than 1 billion sequences [30]. It also uses Linclust [31] to cluster the protein sequences with a sequence identity and coverage of 90%; the cluster representative is chosen to be the longest sequence. Moreover, it provides very useful tools to, for example, query the non-redundant protein dataset for sequence homologs using the HMM profile-based tool HMMER [38]. Note that the user can choose to query only a subset of the full set of proteins, corresponding to a type of microbial niche (also called biome). As shown in Fig. 2.b, most of the entries come from human microbiomes, but marine, animal, plant, and soil biomes also contribute significantly to the dataset.
- **MetaClust.** The MetaClust database contains about 1.6 billion protein sequence fragments, predicted by the gene prediction program Prodigal [37] from about 1,800 metagenomic and 400 metatranscriptomic datasets obtained from multiple resources [29,44,45]. These sequences were clustered into 424 million classes using Linclust [31], a fast protein sequence clustering algorithm able to cluster huge sets of sequences. The thresholds used for the clustering is 50% sequence identity and 90% sequence coverage. MetaClust is a ready-to-use tool, providing 424 million representative sequences.
- **BFD.** Unlike other databases, the Big Fantastic Database (BFD) is a sequence profile database. It contains about 65 million families represented as MSAs and hidden Markov models (HMMs). It is one of the largest and most used metagenomic databases, as MSA and HMM representations are sometimes more convenient to work with than non-redundant representative protein sequence sets. It has been constructed by collecting about 2.5 billion protein sequences from UniProt/Trembl [46], SwissProt [47], MetaClust[31], as well as the Soil Reference Catalog and the Marine Eukaryotic Reference Catalog, assembled using the *de novo* protein-level assembler PLASS [32], which is able to recover more protein sequences from metagenomes than classi-

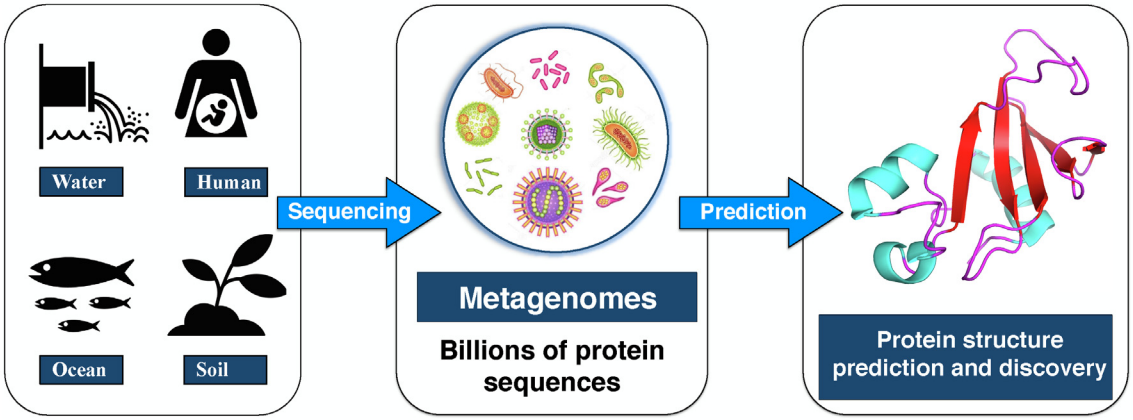

Fig. 1. Schematic representation of the pipeline from biomes, metagenome samples to protein structure prediction and discovery.

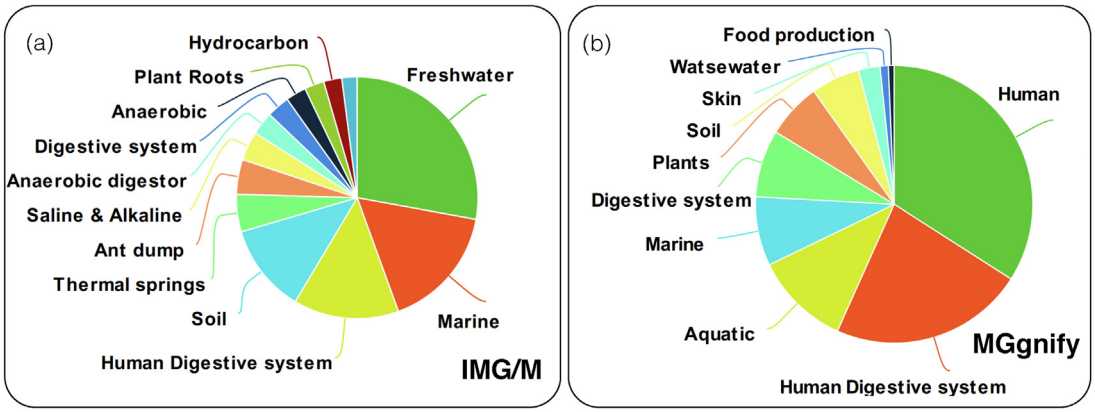

Fig. 2. Sources of metagenomic data in (a) IMG/M and (b) MGnify databases. For more details, see Table S1 of Supplementary Material..

cal assembly methods. The sequences were clustered using a strict sequence identity cut-off of 30% and a coverage threshold of 90% using MMseqs2/Linclust [31]. Clusters with less than three entries were removed.

Here we only described metagenomics databases that were commonly used in the last rounds of the Critical Assessment of Structure Prediction (CASP) [48], a community-wide experiment in which the competitors blindly predict the 3D structure of target proteins and the accuracy of the predictions is evaluated by a group of assessors. There are also other metagenomic databases available in the literature, which are listed in Table S2 of Supplementary Material. Among them, one of the most complete repositories is the MetaGenomics Rapid Annotation using Subsystems Technology (MG-RAST) [49], which allows storage, annotation, phylogenetic study and functional analysis of metagenomes. Other resources are mainly databases that collect eukaryotic metagenomic data such as TOPAZ [50], SMAGs [51] and MetaEuk [52] or viral metagenomes such as MetaVir [53], VIROME [54], Metagenomic Gut Virus (MGV) [55] and Gut Phage Database (GPD) [56].

Despite the huge amounts of sequences in the different metagenomic databases described above, their overlap with standard protein sequence databases such as UniProtKB [57] is very limited. Therefore, the combination of metagenome and genome sequence databases has the enormous potential to provide improved biological information.

3. Integrating metagenomics data to structure prediction pipelines

3.1. Boosting protein structure prediction accuracy

In the last two decades, huge improvements have been made in the field of protein structure prediction. Many predictors have been developed with a steady increase in performance, achieving amazing prediction accuracy in CASP14, the last round of the CASP competition [48]. AlphaFold2 [58] was the best performing method, reaching an accuracy close to that of the experimental methods, even for difficult targets for which no structural templates were available [48].

The improved performance of structure predictors is due to several technological advances, among which novel machine learning algorithms such as deep learning and end-to-end prediction models; for more detailed analyses of these approaches, we refer to excellent reviews on the subject [9,10]. Another breakthrough is certainly the incorporation of data from large metagenomic databases, which allows building deeper and better quality MSAs than when limiting the search to genomic sequences; this is especially true when the number of genomic sequences is low. In turn, these enhanced MSAs are used to gain more precise protein structure information through the application of coevolutionary approaches.

The idea that coevolutionary models extracted from MSAs can be used to gain information on protein structure dates back almost

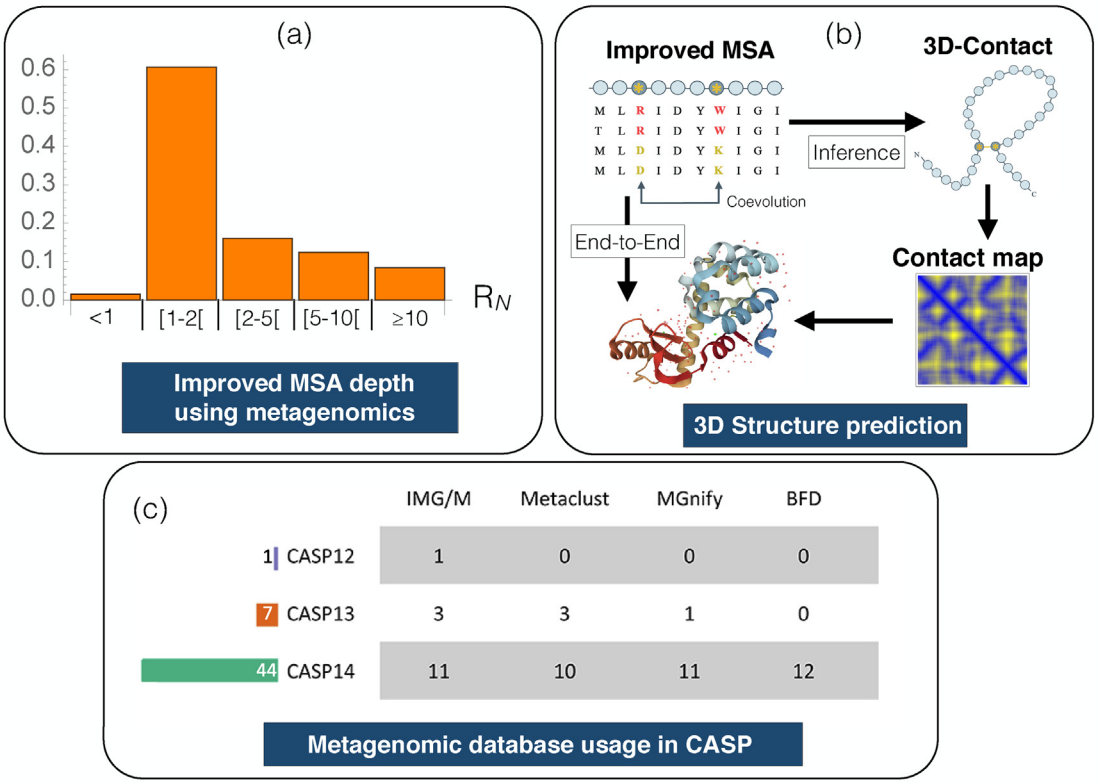

Fig. 3. Metagenomics in protein structure prediction. (a) Quantitative MSA enrichment when adding metagenomic sequences: probability distribution of R_N, which is defined as the ratio between the number of effective sequences N_{eff} in MSAs constructed from both metagenomic and genomic sequence databases, and from genomic sequences only; the values come from the study of 5,721 Pfam families [8]; (b) Schematic representation of the two types of protein structure prediction pipelines based on MSAs: the optimization of multiple intermediate steps such as the identification of coevolutionary signals and the prediction of contact maps, and an end-to-end differentiable model which enables a single optimization from the input MSA to the output 3D structure; (c) Number of times metagenomic databases have been used in structure prediction methods in the last three CASP experiments [48,77,78].

thirty years [59], even though it only started to be a focus of the research community when a series of seminal papers introduced the coevolution formalism of direct coupling analysis (DCA) [60–62] (see also [63] for a recent review). Basically, residues that are close in the 3D protein structure tend to coevolve along the evolutionary history. Indeed, if an interaction between two residues is essential for the stability of the protein structure, the mutation of one of the residues causes an evolutionary pressure on the second residue, which favors compensatory mutations to restore the original interaction, thus maintaining the molecular function of the protein [64].

Since these early approaches, many studies have been devoted to extracting coevolutionary signals from MSAs and using them to predict protein contact maps [65–70]. These contacts are then used as constraints in modeling tools to guide the protein structure prediction pipeline. A key step of these pipelines is to build and curate MSAs. Widely known algorithms for MSA construction are PSI-BLAST, which uses position-specific sequence profiles [71], and HHblits [72] and HMMsearch [73], which use HMM profiles. Note, however, that low-quality or shallow MSAs can lead to inaccurate predictions, when the substitution statistics are not well estimated.

The first time metagenomic data was used to improve MSA quality was in 2017 [8]. It was shown that substantially deeper MSAs can be obtained by combining the Integrated Microbial Genomes and Microbiomes (IMG/M) database [29] to the genomic sequence cluster database UniRef30 [74]. Indeed, the addition of metagenomic data leads to the increase of the effective number of sequences N_{eff} (as defined in [61]) by a factor of about 3.5 on average for the approximately 5,000 protein families from the Pfam database [43]. In particular, about 500 families show an increase in N_{eff} by a factor of 10, and a few families, by a factor of 100 [8] (Fig. 3.a). This improvement led to more accurate predictions of the protein contact maps for about 20% of the Pfam families considered using the contact map predictor GREMLIN [75], which in turn led to more accurate 3D structures generated via the *de novo* structural modeling tool Rosetta [76].

After this first study [8], multiple structure prediction tools integrating metagenomic data have been developed. Already in the CASP13 experiment [77], several methods used this source of information to predict residue-residue contacts [79–81], 3D structure [82–85] and structure refinement [86]. DeepMSA [87], an open-source automated pipeline for the construction of deep alignments using metagenomic information, has also been introduced in CASP13. It is based on different types of sequence sources, two genomic sources (UniClust30 [88] and UniRef90 [74]) and a metagenomic source (MetaClust [31]). These databases are queried via a hybrid homology-detection approach including HHblits [72] and HMMER [89]. The high quality MSAs generated from DeepMSA has been shown to significantly improve long-range contact prediction accuracy [87].

In a later investigation [90], metagenomic sequence data collected by the Tara Oceans expeditions [45] and from MetaClust [31], another metagenomic database, were used in addition to UniRef [91] to increase the number of effective sequences N_{eff} of about 400 Pfam families. For 27 of them, an enriched MSA was obtained with an N_{eff} increase by a factor of two, which, again, led to an improvement of their predicted 3D protein structures.

The metagenomics contribution to the field has become even more important in the last CASP rounds (CASP14), where the majority of the methods used metagenomics sequences either for predicting inter-residue contacts or distances as a preliminary step in protein structure modeling [92,81,93], or directly using them in the end-to-end structure prediction model without intermediate steps [58,94,95,83,96] (see Fig. 3).

Some of the prediction methods in CASP14 used the DeepMSA pipeline to query the target sequence against metagenomics databases. However, the best performing methods such as AlphaFold2 [58], D-I-Tasser [97] and RoseTTaFold [95] developed new, improved pipelines for homologous sequence search which combine multiple methods to mine metagenomics databases. For example, AlphaFold2 [58], which dominated CASP14 and achieved astonishing prediction accuracy, employs homologous searches in UniRef90 and MGnify using JackHMMER, and in BFD and Uniclust30 using HHBlits. The output MSAs of these searches are then deduplicated and stacked together to further improve the amount of homologous sequences collected. This pipeline led to an average improvement of the structure prediction performance of approximately 6% in terms of the global distance test score [58].

The DeepMSA approach has been generalized to DeepMSA2 [97] in which, in addition to the Uniclust30 and UniRef90 genomic sequence databases, the four widely known metagenomic sequence databases described in the previous sections are mined (MetaClust, BFD, MGnify and IMG/M). The full pipeline consists of a complex series of steps including multiple rounds of database mining with JackHammer, HHBlits and HMMsearch (see [97] for technical details). It provides MSAs that are 40% to 150% deeper than the original DeepMSA pipeline, which in turn leads to statistical significant improvements in both residue-residue distance and protein structure predictions [97].

Recently, a more computation-efficient pipeline for MSA generation has been introduced [98]. It employs MMseqs2 [31] to mine UniRef30 and, using the sequence profile generated, performs an iterative search against two new databases: BFD/MGnify and ColabFoldDB. The former was created by merging BFD and MGnify databases through a MMseqs2 search of MGnify sequences among the BFD clusters of representative sequences. The aligned matches were assigned to the corresponding BFD clusters; the non-matching MGnify sequences were used to generate new clusters. The latter database (ColabFoldDB) was essentially constructed in the same way but includes, in addition to BFD/MGnify, sequences retrieved from other metagenomic databases such as MetaClust, GPD, MGV, TOPAZ, MetaEuk and SMAGs.

The improved accuracy of these different methods compared to standard approaches that do not rely on metagenomic information demonstrates the central role played by metagenomics in the field of protein structure prediction. This is due to the fact that the current sequence databases are far from complete, despite their rapid growth, and that they contain too few homologous sequences for too many target proteins. Metagenome sequence databases have the advantage of filling this gap. Note that the combined use of multiple metagenomic databases with different mining algorithms and parameters further improves homologous sequence search and thus helps construct deeper MSAs and identify more accurate evolutionary information needed for protein structure prediction.

3.2. Is more metagenomics always better?

We underlined in the previous subsection the rapid accumulation of metagenomic sequences and the impressive size of metagenomics databases with *e.g.*, the IMG database containing more than 60 billion microbial genes [29]. Although these databases represent an invaluable source of information, deep MSA construction by querying them is becoming computationally expensive and memory-demanding. The precise identification of MSA characteristics that may improve the accuracy of contact and structure prediction is still an open question in the community. Indeed, having more sequence homologs in the alignment is not always better [87] considering that there is a trade-off between the effective number of sequences, the sequence coverage, and the alignment accuracy.

In an interesting recent study [99], the link between microbial niches and homologous protein families was investigated for a set of about 2,000 Pfam families with no structural templates. Four different microbial biomes, from gut, lake, soil and fermentor, were used in turn for MSA enrichment to test their ability to improve 3D structure prediction. It turned out that the structural modeling of the Pfam families is more precise when only one or a few specific biomes linked to the target protein family are used.

This has led to propose a prediction model called MetaSource which is able to identify one biome or a set of biomes which allows better MSA construction and modeling of a given Pfam family [99]. Note that this approach yields not only an improved accuracy but also a significant increase in computational efficiency: it is around 3.3-fold faster than considering all sets of metagenomic information [99].

4. Integrating metagenomics data for functional annotation and validation

4.1. Boosting enzyme discovery using metagenomics

Metagenomic sequencing data started to be used to identify new proteins with specific enzymatic activity and stability properties. The use of huge amounts of sequence data extracted from a wide variety of different environments, from animal rumen to marine, water and soil, has revolutionized the discovery process of novel enzymes in the last decade [11]. We can estimate from previous reviews on the topic that at least 500 new enzymes have been identified using metagenomics-based approaches; this underlines the deep impact of metagenomics in this important biotechnological research field [11]. Here we provide a non-exhaustive list of enzyme types whose development has been boosted by using of such approcahes.

The HotZyme project [100], for example, has been devoted to the extensive screening and analysis of metagenomes from thermal springs around the world, with the aim to first discover and then characterize novel thermostable hydrolases of industrial interest. Metagenomics screening resulted in 100 potentially new hydrolases, of which 12 have been biochemically and structurally characterized, including carboxylesterases, lactonases and cellulases. Metagenomics data has also been widely used for the identification of lignin-degrading enzymes such as laccase, xylanase, β-glucosidase, acetyl xylan esterases, arabinofuranosidases, and lyases [12–15]. These enzymes, which catalyze the depolymerization of lignin, have been discovered by different consortia such as RAS [101] and LigMet [12] through the mining of metagenomes from different environments such as rice straw compost, sugarcane soil samples, bovine rumen and insect intestinal tracts.

Marine-related metagenomics data also provides huge amounts of information. Large expeditions collecting marine samples such as Tara Ocean [45,102] and GEOTRACES [103] have led to the metagenome assembly of more than 25,000 genomes. These efforts have contributed to the discovery and functional characterization of a wide series of novel enzymes [16–18], such as cold-adapted lipases and esterases which are of key importance in food and biotechnology industry. Other discovered enzymes of interest are novel thermostable biocatalysts including lipolytic enzymes, hydrolases, fumarase and β-glucosidase, as well as a series of enzymes that are tolerant to salt, acid or basic pH, or heavy metals.

There are basically two main approaches for the discovery of new enzymes from metagenomic data, which are function-based and sequence-based screenings [104]. In the former, DNA fragments from environmental metagenomes are first cloned and expressed using expression vectors to produce proteins which are then screened *in vivo* or *in vitro* for enzymatic activities [105]. This is the most frequently used approach and allows the identification of enzymes that do not share sequence similarity with known counterparts. However, it is laborious and requires reliable and high-throughput screening methods that are difficult to generalize [104].

The second approach is sequence-based and applicable when the enzymes sought are closely related in terms of sequence similarity to enzymes collected in known databases. This *in silico* method queries large metagenomic datasets with HMM profiles [38] constructed from sequences of known enzymes and their close homologs. While this method is more efficient than function-based methods, it is more limited in terms of sequence space explored. It can also lead to false hits, due to misannotations in poorly curated datasets. Examples of automatic computational pipelines for metagenomic enzyme discovery through sequence-based screening are MetaHMM [106] and ANASTASIA [107].

4.2. CRISPR-Cas system identification in microbiomes

CRISPR-Cas, where Cas is an enzyme and CRISP the acronym of Clustered Regularly Interspaced Short Palindromic Repeats, is a system employed by most archaea and bacteria as immunological defense against invading DNA [108]. In brief, short fragments of foreign DNA are integrated into CRISPR loci, which causes the memorization of the infection. These are transcribed into CRISPR RNA, which are then used as guides for Cas proteins to specifically interfere with invading nucleic acids upon reoccurring infection.

Due to its huge potential for genome editing, the CRISPR-Cas system is used as a precise technology in biological and clinical research and applications [109], and metagenomic datasets are therefore mined to discover new such systems [24]. For example, three sources of metagenomic data were used for this purpose [24], from two soil environments and one water environment. As many as 155 million protein-coding genes were extracted from these, using Prodigal, a protein-coding gene predictor for prokaryotic genomes [37]. This set of sequences were searched for Cas protein homologs using HMMER [38], while CRISPR arrays were identified using the CrisprFinder detection tool [110]. This analysis led to the identification of novel CRISPR-Cas systems: CRISPR-Cas9 in archaea, and CRISPR-CasX and CasY in bacteria, which are among the most compact CRISPR-Cas systems known to date [24].

The International Metagenomics and Metadesign of Subways and Urban Biomes (MetaSUB) consortium provided 4,728 metagenomic samples from mass-transit systems of 60 cities around the world [111]. These data led to the discovery of 838,532 CRISPR arrays predicted by an improved version of CrisprFinder [25], of which 3,245 had unambiguous annotations. More recently, 2.9 million CRISPR loci have been functionally and taxonomically profiled from 2,355 body-wide human microbiomes from 17 different body sites [26], thus increasing of one order of magnitude the number of known CRISPR in the human microbiome. The Crass tool has been used for that purpose, which identifies and reconstructs CRISPR from unassembled metagenomic data [112]. Also, the abundance of different Cas proteins was profiled and associated with CRISPR subtypes to obtain information about the functional and evolutionary role of CRISPR-Cas systems in human microbiomes.

The studies that identified CRISPR-Cas systems by mining metagenome resources are not limited to these few examples but include other studies that use completely different metagenomic environments such as the irrigation of water sources [27], various human microbiomes from skin to oral microbiomes [113,114] and extreme environments ranging from antarctic snow to hot springs [28,115]. To discover CRISPR repeats in these metagenomes, several bioinformatics tools have been developed, among which MinCED (github.com/ctSkennerton/minced), MetaCRAST [116], Crass [112], and metaCRT [113].

Finally note that metagenomic databases can also by mined to explore anti-CRISPRs, *i.e.* natural inhibitors of the CRISPR-Cas system [117]. For example, a high-throughput approach has been developed to discover anti-CRISPR genes from metagenomics data based on their functional activity [118]. The action of eleven DNA fragments from soil, animal, and human metagenomes were identified and tested *in vitro* to decrease Cas9 activity in *Streptococcus pyogenes*.

4.3. Functional annotation and analysis of the resistome using metagenomics data

Antimicrobial resistance is another central problem in microbiology where metagenomic data plays a fundamental role. The identification of antibiotic resistance genes (ARG) in soil-dwelling bacteria, human gut microbiota and other microbial communities, which can potentially act as ARG reservoirs [119–121], is important to fully understand the origin, evolution and maintenance of antibiotic resistance. Indeed, these genes can be exchanged through lateral gene transfer and confer antibiotic resistance to pathogens. For example, infant gut microbiome was investigated and revealed a cohort of resistance genes in fecal microbiota of pediatric patients, even without their prior exposure to the selective pressure of antibiotics [122]. These findings explain how a healthy human gut can act as a reservoir for ARGs.

An interesting method, taking advantage of 3D protein structures, was developed to predict ARGs in gut microbiota [20]. This method, based on a combination of homology modeling and machine learning techniques, is able to correctly identify ARGs: from the 71 predicted ARGs, antibiotic resistance activity was detected *in vitro* in 51 of them. The method was also tested on an experimentally validated functional metagenomic dataset from soils, highlighting very good performance, especially in terms of sensitivity. Furthermore, metagenomic sequencing from respiratory specimens of patients with and without chronic respiratory diseases such as severe asthma, chronic obstructive pulmonary disease and bronchiectasis showed that respiratory tract microbiota also harbors a core of ARGs dominated by genes resistant to macrolide antibiotics [21]. This finding was independent of the health status of the patients and of their previous exposure to antibiotics.

Soil is certainly another reservoir of ARGs, since it is in direct contact with antibiotics used in livestock farming and agriculture. Evidence of ARG exchange between soil-dewling bacteria and clinical pathogens was shown from functional metagenomics analyses of soil-derived bacteria cultures [22]. Multidrug-resistant soil bacteria were shown to harbor ARGs against five important classes of antibiotics: β-lactams, aminoglycosides, amphenicols, sulfonamides, and tetracyclines. Furthermore, the analysis of metagenomic data from gut microbiota of migratory birds revealed about 1,000 ARGs that can be classified into about 200 different types associated to specific antibiotic resistance [23]. Compared to environmental metagenomes, microbiota of migratory birds have a lower phylogenetic diversity but more antibiotic resistance proteins, thus suggesting the possible role of birds as ARG reservoir.

Finally, possible differences in the ARG distribution according to the ecological niches were investigated [19]. The analysis of human, animal, water, soil, plant and insect metagenomes from the MG-RAST database [123] led to conclude that the human

microbiome is characterized by the highest relative ARG abundance.

5. Conclusion

The use of metagenomics data has become essential in different domains of research during the last decade. Indeed, considerable efforts to improve the standardization of data analyses and metagenomic databases have resulted in impressive developments in enzyme discovery, 3D protein structure prediction and function annotation. The study of the role of human microbiota in disease, aging and antibiotic resistance has also greatly benefited from these developments.

The explosion of the amount of metagenomic data is currently creating a challenge for bioinformatics tools, especially in the data storage and analysis and in the integration of different metagenomic techniques, including metatranscriptomics, metaproteomics and metabolomics. The improvements of these tools will lead in the near future to further advances in these fields, but will also boost or continue to fuel a series of other applications that have not been analyzed in this mini-review such as protein function prediction [124], predictions of protein–protein interactions and protein complex structures [95,125,126] and the detection and tracing of novel viral pathogens [127,128].

CRediT authorship contribution statement

Qingzhen Hou: Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Fabrizio Pucci:** Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Fengming Pan:** Data curation, Formal analysis. **Fuzhong Xue:** Funding acquisition, Project administration, Supervision. **Marianne Rooman:** Formal analysis, Funding acquisition, Supervision, Validation, Writing - original draft, Writing - review & editing. **Qiang Feng:** Funding acquisition, Project administration, Supervision, Investigation, Software, Writing - original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

Q.H. was supported by the Young Scholars Program of Shandong University (21320082064101). F.X. was supported by the National Natural Science Foundation of China (81773547) and the National Key Research and Development Program of China (2020YFC2003500). F.P. and M.R. are Postdoctoral Researcher and Research Director, respectively, at the F.R.S.-FNRS Fund for Scientific Research and acknowledge financial support from the same Fund through a PDR and a PER project. Q.F. was supported by the National Natural Science Foundation of China (No.82071122), The Construction Engineering Special Fund of 'Taishan Scholars' of Shandong Province (tsqn201909180) and the Program of Excellent young scholars of Shandong University. The authors declare that there is no conflict of interest.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.csbj.2021.12.030.

References

[1] Hiraoka S, Yang C-C, Iwasaki W. Metagenomics and bioinformatics in microbial ecology: current status and beyond. Microbes Environ 2016: ME16024.
[2] Taş N, de Jong AE, Li Y, Trubl G, Xue Y, Dove NC. Metagenomic tools in microbial ecology research. Curr Opin Biotechnol 2021;67:184–91.
[3] Hudson ME. Sequencing breakthroughs for genomic ecology and evolutionary biology. Mol Ecol Resour 2008;8(1):3–17.
[4] Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, Zhang D, Xia H, Xu X, Jie Z, et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. Nature Commun 2015;6(1):1–13.
[5] Shreiner AB, Kao JY, Young VB. The gut microbiome in health and in disease. Current Opinion Gastroenterol 2015;31(1):69.
[6] Durack J, Lynch SV. The gut microbiome: relationships with disease and opportunities for therapy. J. Exp. Med. 2019;216(1):20–40.
[7] Chiu CY, Miller SA. Clinical metagenomics. Nat Rev Genet 2019;20(6):341–55.
[8] Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D. Protein structure determination using metagenome sequence data. Science 2017;355(6322):294–8.
[9] Laine E, Eismann S, Elofsson A, Grudinin S. Protein sequence-to-structure learning: Is this the end(-to-end revolution)? Proteins: Structure, Function, Bioinform 2021;89(12):1770–86.
[10] AlQuraishi M. Machine learning in protein structure prediction. Curr Opin Chem Biol 2021;65:1–8.
[11] Robinson SL, Piel J, Sunagawa S. A roadmap for metagenomic enzyme discovery. Natural Product Rep 2021;38(11):1994–2023.
[12] Moraes EC, Alvarez TM, Persinoti GF, Tomazetto G, Brenelli LB, Paixão DA, Ematsu GC, Aricetti JA, Caldana C, Dixon N, et al. Lignolytic-consortium omics analyses reveal novel genomes and pathways involved in lignin modification and valorization. Biotechnol Biofuels 2018;11(1):1–16.
[13] Wang C, Dong D, Wang H, Müller K, Qin Y, Wang H, Wu W. Metagenomic analysis of microbial consortia enriched from compost: new insights into the role of Actinobacteria in lignocellulose decomposition. Biotechnol Biofuels 2016;9(1):1–17.
[14] Ferrer M, Beloqui A, Golyshin PN. Screening metagenomic libraries for laccase activities. In: Metagenomics. Springer; 2010. p. 189–202.
[15] Brennan Y, Callen WN, Christoffersen L, Dupree P, Goubet F, Healey S, Hernández M, Keller M, Li K, Palackal N, et al. Unusual microbial xylanases from insect guts. Appl Environ Microbiol 2004;70(6):3609–17.
[16] Barone R, De Santi C, Palma Esposito F, Tedesco P, Galati F, Visone M, Di Scala A, De Pascale D. Marine metagenomics, a valuable tool for enzymes and bioactive compounds discovery. Front Marine Sci 2014;1:38.
[17] Alma'abadi AD, Gojobori T, Mineta K. Marine metagenome as a resource for novel enzymes. Genomics, Proteomics Bioinform 2015;13(5):290–5.
[18] Popovic A, Tchigvintsev A, Tran H, Chernikova TN, Golyshina OV, Yakimov MM, Golyshin PN, Yakunin AF. Metagenomics as a tool for enzyme discovery: hydrolytic enzymes from marine-related metagenomes. Prokaryotic Syst Biol 2015:1–20.
[19] Fitzpatrick D, Walsh F. Antibiotic resistance genes across a wide variety of metagenomes. FEMS Microbiol Ecol 2016;92(2):fiv168.
[20] Ruppé E, Ghozlane A, Tap J, Pons N, Alvarez A-S, Maziers N, Cuesta T, Hernando-Amado S, Clares I, Martínez JL, et al. Prediction of the intestinal resistome by a three-dimensional structure-based method. Nature Microbiol 2019;4(1):112–23.
[21] Mac Aogáin M, Lau KJ, Cai Z, Kumar Narayana J, Purbojati RW, Drautz-Moses DI, Gaultier NE, Jaggi TK, Tiew PY, Ong TH, et al. Metagenomics reveals a core macrolide resistome related to microbiota in chronic respiratory disease. Am J Respiratory Critical Care Med 2020;202(3):433–47.
[22] Forsberg KJ, Reyes A, Wang B, Selleck EM, Sommer MO, Dantas G. The shared antibiotic resistome of soil bacteria and human pathogens. Science 2012;337 (6098):1107–11.
[23] Cao J, Hu Y, Liu F, Wang Y, Bi Y, Lv N, Li J, Zhu B, Gao GF. Metagenomic analysis reveals the microbiome and resistome in migratory birds. Microbiome 2020;8(1):1–18.
[24] Burstein D, Harrington LB, Strutt SC, Probst AJ, Anantharaman K, Thomas BC, Doudna JA, Banfield JF. New CRISPR–Cas systems from uncultivated microbes. Nature 2017;542(7640):237–41.
[25] Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, Rocha EP, Vergnaud G, Gautheret D, Pourcel C. CRISPRCasFinder, an update of CRISRFinder includes a portable version, enhanced performance and integrates search for Cas proteins. Nucl Acids Res 2018;46(W1):W246–51.
[26] Münch PC, Franzosa EA, Stecher B, McHardy AC, Huttenhower C. Identification of natural CRISPR systems and targets in the human microbiome. Cell Host Microbe 2021;29(1):94–106.
[27] Chopyk J, Nasko DJ, Allard S, Bui A, Treangen T, Pop M, Mongodin EF, Sapkota AR. Comparative metagenomic analysis of microbial taxonomic and functional variations in untreated surface and reclaimed waters used in irrigation applications. Water Res 2020;169:115250.
[28] Snyder JC, Bateson MM, Lavin M, Young MJ. Use of cellular CRISPR (clusters of regularly interspaced short palindromic repeats) spacer-based microarrays for detection of viruses in environmental samples. Appl Environ Microbiol 2010;76(21):7251–8.
[29] Chen I-MA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, Hajek P, Ritter S, Varghese N, Seshadri R, et al. The IMG/M data management and

analysis system v. 6.0: new tools and advanced capabilities. Nucleic Acids Res 2021;49(D1):D751–63.
[30] Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, Crusoe MR, Kale V, Potter SC, Richardson LJ, et al. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res 2020;48(D1):D570–8.
[31] Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nature Commun 2018;9(1):1–8.
[32] Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nature Methods 2019;16(7):603–6.
[33] Sugimoto Y, Camacho FR, Wang S, Chankhamjon P, Odabas A, Biswas A, Jeffrey PD, Donia MS. A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. Science 2019;366(6471):eaax9176.
[34] Vidulin V, Šmuc T, Džeroski S, Supek F. The evolutionary signal in metagenome phyletic profiles predicts many gene functions. Microbiome 2018;6(1):1–21.
[35] Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Sundaramurthi JC, Lee J, Kandimalla M, Chen I-MA, Kyrpides NC, Reddy T. Genomes OnLine Database (GOLD) v. 8: overview and updates. Nucleic Acids Res 2021;49(D1):D723–33.
[36] Clum A, Huntemann M, Bushnell B, Foster B, Foster B, Roux S, Hajek PP, Varghese N, Mukherjee S, Reddy T, et al. DOE JGI Metagenome Workflow. Msystems 2021;6(3):e00804–20.
[37] Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform 2010;11(1):1–11.
[38] Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol 2011;7(10):e1002195.
[39] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215(3):403–10.
[40] Kanehisa M. Enzyme annotation and metabolic reconstruction using kegg. In: Protein Function Prediction. Springer; 2017. p. 135–45.
[41] Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res 2021;49(D1):D545–51.
[42] Sillitoe I, Dawson N, Lewis TE, Das S, Lees JG, Ashford P, Tolulope A, Scholes HM, Senatorov I, Bujan A, et al. CATH: expanding the horizons of structure-based functional annotations for genome sequences. Nucleic Acids Res 2019;47(D1):D280–4.
[43] Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, Tosatto SC, Paladin L, Raj S, Richardson LJ, et al. Pfam: The protein families database in 2021. Nucleic Acids Res 2021;49(D1):D412–9.
[44] Kodama Y, Shumway M, Leinonen R. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 2012;40(D1):D54–6.
[45] Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, et al. Structure and function of the global ocean microbiome. Science 2015;348(6237):1261359.
[46] . Nucleic Acids Res 2017;45(D1):D158–69.
[47] Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement trembl in 2000. Nucleic Acids Res 2000;28(1):45–8.
[48] Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM, Lupas AN. High-accuracy protein structure prediction in CASP14. Proteins: Structure, Function, Bioinform 2021;89(12):1687–99.
[49] Meyer F, Bagchi S, Chaterji S, Gerlach W, Grama A, Harrison T, Paczian T, Trimble WL, Wilke A. MG-RAST version 4-lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. Briefings Bioinform 2019;20(4):1151–9.
[50] H. Alexander, S.K. Hu, A.I. Krinos, M. Pachiadaki, B.J. Tully, C.J. Neely, T. Reiter, Eukaryotic genomes from a global metagenomic dataset illuminate trophic modes and biogeography of ocean plankton, biorxiv, doi:10.1101/2021.07.25.453713..
[51] T.O. Delmont, M. Gaia, D.D. Hinsinger, P. Fremont, C. Vanni, A.F. Guerra, A.M. Eren, A. Kourlaiev, L. d'Agata, Q. Clayssen, E. Villar, K. Labadie, C. Cruaud, J. Poulain, C. Da Silva, M. Wessner, B. Noel, J.-M. Aury, T.O. Coordinators, C. de Vargas, C. Bowler, E. Karsenti, E. Pelletier, P. Wincker, O. Jaillon, Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics, bioRxiv, doi:10.1101/2020.10.15.341214..
[52] Karin EL, Mirdita M, Söding J. MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome 2020;8(1):1–15.
[53] Roux S, Tournayre J, Mahul A, Debroas D, Enault F. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC Bioinform 2014;15(1):1–12.
[54] Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, Furman M, Jamindar S, Nasko DJ. VIROME: a standard operating procedure for analysis of viral metagenome sequences. Standards Genomic Sci 2012;6(3):421–33.
[55] Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, Proal AD, Fischbach MA, Bhatt AS, Hugenholtz P, et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nature Microbiol 2021;6(7):960–70.
[56] Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. Massive expansion of human gut bacteriophage diversity. Cell 2021;184(4):1098–109.
[57] UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Research 49 (D1) (2021) D480–D489..
[58] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873):583–9.
[59] Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins: Structure, Function, Bioinform 1994;18(4):309–17.
[60] Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein–protein interaction by message passing. Proc Nat Acad Sci 2009;106(1):67–72.
[61] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Nat Acad Sci 2011;108(49):E1293–301.
[62] Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. Protein 3D structure computed from evolutionary sequence variation. PLoS One 2011;6(12):e28766.
[63] Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M. Inverse statistical physics of protein sequences: a key issues review. Rep Prog Phys 2018;81(3):032601.
[64] Ivankov DN, Finkelstein AV, Kondrashov FA. A structural perspective of compensatory evolution. Current Opinion Struct Biol 2014;26:104–12.
[65] Buchan DW, Jones DT. Improved protein contact predictions with the MetaPSICOV2 server in CASP12. Proteins: Structure, Function, Bioinform 2018;86:78–83.
[66] Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de dovo prediction of protein contact mmap by ultra-deep learning model. PLOS Comput Biol 2017;13(1):1–34.
[67] Li Y, Zhang C, Bell EW, Yu D-J, Zhang Y. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. Proteins: Structure, Function, Bioinform 2019;87(12):1082–91.
[68] Michel M, Menéndez Hurtado D, Elofsson A. PconsC4: fast, accurate and hassle-free contact predictions. Bioinformatics 2018;35(15):2677–9.
[69] Liu Y, Palmedo P, Ye Q, Berger B, Peng J. Enhancing evolutionary couplings with deep convolutional neural networks. Cell Systems 2018;6(1):65–74.
[70] Zerihun MB, Pucci F, Peter EK, Schug A. pydca v1. 0: a comprehensive software for direct coupling analysis of RNA and protein sequences. Bioinformatics 2020;36(7):2264–5.
[71] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389–402.
[72] Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods 2012;9(2):173–5.
[73] Eddy SR. Profile hidden markov models, 14. England): Bioinformatics (Oxford; 1998. p. 755–63.
[74] Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, Consortium U. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 2015;31(6):926–32.
[75] Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. Proc Nat Acad Sci 2013;110(39):15674–9.
[76] Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66–93.
[77] Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-Round XIII. Proteins: Structure, Function, Bioinform 2019;87(12):1011–20.
[78] Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)-Round XII. Proteins: Structure, Function, Bioinform 2018;86:7–15.
[79] Kandathil SM, Greener JG, Jones DT. Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. Proteins: Structure, Function, Bioinform 2019;87(12):1092–9.
[80] Li Y, Zhang C, Bell EW, Yu D-J, Zhang Y. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. Proteins: Structure, Function, Bioinform 2019;87(12):1082–91.
[81] Wu Q, Peng Z, Anishchenko I, Cong Q, Baker D, Yang J. Protein contact prediction using metagenome sequence data and residual neural networks. Bioinformatics 2020;36(1):41–8.
[82] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AW, Bridgland A, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020;577(7792):706–10.
[83] Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proc Nat Acad Sci 2020;117(3):1496–503.
[84] Zheng W, Li Y, Zhang C, Pearce R, Mortuza S, Zhang Y. Deep-learning contact-map guided protein structure prediction in CASP13. Proteins: Structure, Function, Bioinform 2019;87(12):1149–64.
[85] Robertson JC, Nassar R, Liu C, Brini E, Dill KA, Perez A. Nmr-assisted protein structure prediction with meldxmd. Proteins: Structure, Function, Bioinform 2019;87(12):1333–40.
[86] Park H, Lee GR, Kim DE, Anishchenko I, Cong Q, Baker D. High-accuracy refinement using rosetta in CASP13. Proteins: Structure, Function, Bioinform 2019;87(12):1276–82.

[87] Zhang C, Zheng W, Mortuza S, Li Y, Zhang Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics 2020;36(7):2105–12.
[88] Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res 2017;45(D1):D170–6.
[89] Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic Acids Res 2018;46(W1):W200–4.
[90] Wang Y, Shi Q, Yang P, Zhang C, Mortuza S, Xue Z, Ning K, Zhang Y. Fueling ab initio folding with marine metagenomics enables structure and function predictions of new protein families. Genome Biology 2019;20(1):1–14.
[91] Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 2007;23(10):1282–8.
[92] Xu J, Mcpartlon M, Li J. Improved protein structure prediction by deep learning irrespective of co-evolution information. Nature Machine Intell 2021:1–9.
[93] Liu J, Wu T, Guo Z, Hou J, Cheng J. Improving protein tertiary structure prediction by deep learning and distance prediction in casp14. Proteins: Structure, Function, Bioinform 2022;90(1):58–72.
[94] S.M. Kandathil, J.G. Greener, A.M. Lau, D.T. Jones, Deep learning-based prediction of protein structure using learned representations of multiple sequence alignments, bioRxiv, doi:10.1101/2020.11.27.401232..
[95] Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 2021;373(6557):871–6.
[96] AlQuraishi M. ProteinNet: a standardized data set for machine learning of protein structure. BMC Bioinform 2019;20(1):1–10.
[97] Zheng W, Li Y, Zhang C, Zhou X, Pearce R, Bell EW, Huang X, Zhang Y. Protein structure prediction using deep learning distance and hydrogen-bonding restraints in casp14. Proteins: Structure, Function, Bioinform 2021;89(12):1734–51.
[98] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold-Making protein folding accessible to all, biorxiv, doi:10.1101/2021.08.15.456425..
[99] Yang P, Zheng W, Ning K, Zhang Y. Decoding the link of microbiome niches with homologous sequences enables accurately targeted protein structure prediction. Proc Nat Acad Sci 2021;118(49):e2110828118.
[100] Wohlgemuth R, Littlechild J, Monti D, Schnorr K, van Rossum T, Siebers B, Menzel P, Kublanov IV, Rike AG, Skretas G, et al. Discovering novel hydrolases from hot environments. Biotechnol Adv 2018;36(8):2077–100.
[101] Hossain MS, Dai J, Qiu D. European eel (Anguilla anguilla) GI tract conserves a unique metagenomics profile in the recirculation aquaculture system (RAS). Aquacult Int 2021;29:1529–44.
[102] Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh H-J, Cuenca M, Field CM, Coelho LP, Cruaud C, Engelen S, et al. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. Cell 2019;179(5):1068–83.
[103] Biller SJ, Berube PM, Dooley K, Williams M, Satinsky BM, Hackl T, Hogle SL, Coe A, Bergauer K, Bouman HA, et al. Marine microbial metagenomes sampled across space and time. Sci Data 2018;5(1):1–7.
[104] Simon C, Daniel R. Metagenomic analyses: past and future trends. Appl Environ Microbiol 2011;77(4):1153.
[105] Ngara TR, Zhang H. Recent advances in function-based metagenomic screening. Genomics, Proteomics Bioinform 2018;16(6):405–15.
[106] Szalkai B, Grolmusz V. MetaHMM: A webserver for identifying novel genes with specified functions in metagenomic samples. Genomics 2019;111(4):883–5.
[107] Koutsandreas T, Ladoukakis E, Pilalis E, Zarafeta D, Kolisis FN, Skretas G, Chatziioannou AA. ANASTASIA: an automated metagenomic analysis pipeline for novel enzyme discovery exploiting next generation sequencing data. Front Genetics 2019;10:469.
[108] Van Der Oost J, Westra ER, Jackson RN, Wiedenheft B. Unravelling the structural and mechanistic basis of CRISPR–Cas systems. Nat Rev Microbiol 2014;12(7):479–92.
[109] Anzalone AV, Koblan LW, Liu DR. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. Nature Biotechnol 2020;38(7):824–44.
[110] Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 2007;35:W52–7.
[111] Danko D, Bezdan D, Afshin EE, Ahsanuddin S, Bhattacharya C, Butler DJ, Chng KR, Donnellan D, Hecht J, Jackson K, et al. A global metagenomic map of urban microbiomes and antimicrobial resistance. Cell 2021;184(13):3376–3393.e17.
[112] Skennerton CT, Imelfort M, Tyson GW. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. Nucleic Acids Res 2013;41(10). e105–e105.
[113] Rho M, Wu Y-W, Tang H, Doak TG, Ye Y. Diverse CRISPRs evolving in human microbiomes. PLoS Genetics 2012;8(6):e1002441.
[114] Toyomane K, Yokota R, Watanabe K, Akutsu T, Asahi A, Kubota S. Evaluation of CRISPR diversity in the human skin microbiome for personal identification. Msystems 2021;6(1):e01255–20.
[115] Lopatina A, Medvedeva S, Shmakov S, Logacheva MD, Krylenkov V, Severinov K. Metagenomic analysis of bacterial communities of antarctic surface snow. Front Microbiol 2016;7:398.
[116] Moller AG, Liang C. MetaCRAST: reference-guided extraction of CRISPR spacers from unassembled metagenomes. PeerJ 2017;5:e3788.
[117] Zhang F, Song G, Tian Y. Anti-CRISPRs: The natural inhibitors for CRISPR-Cas systems. Animal Models Exp Med 2019;2(2):69–75.
[118] Uribe RV, van der Helm E, Misiakou M-A, Lee S-W, Kol S, Sommer MO. Discovery and characterization of Cas9 inhibitors disseminated across seven bacterial phyla. Cell Host Microbe 2019;25(2):233–41.
[119] D'Costa VM, McGrann KM, Hughes DW, Wright GD. Sampling the antibiotic resistome. Science 2006;311(5759):374–7.
[120] Hu Y, Yang X, Qin J, Lu N, Cheng G, Wu N, Pan Y, Li J, Zhu L, Wang X, et al. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. Nature Commun 2013;4(1):1–7.
[121] Sommer MO, Dantas G, Church GM. Functional characterization of the antibiotic resistance reservoir in the human microflora. Science 2009;325(5944):1128–31.
[122] Moore AM, Patel S, Forsberg KJ, Wang B, Bentley G, Razia Y, Qin X, Tarr PI, Dantas G. Pediatric fecal microbiota harbor diverse and novel antibiotic resistance genes. PLoS One 2013;8(11):e78822.
[123] Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, et al. The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinform 2008;9(1):1–8.
[124] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rihawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, et al., ProtTrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing, arXiv preprint arXiv:2007.06225..
[125] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A.W. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, et al., Protein complex prediction with AlphaFold-Multimer, bioRxiv, doi:10.1101/2021.10.04.463034..
[126] D. Burke, P. Bryant, I. Barrio-Hernandez, D. Memon, G. Pozzati, A. Shenoy, W. Zhu, A. Dunham, P. Albanese, A. Keller, et al., Towards a structurally resolved human protein interaction network, bioRxiv, doi:10.1101/2021.11.08.467664..
[127] Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, et al. Virus taxonomy in the age of metagenomics. Nat Rev Microbiol 2017;15(3):161–8.
[128] Sommers P, Chatterjee A, Varsani A, Trubl G. Integrating viral metagenomics into an ecological framework. Ann Rev Virol 2021;8(1):133–58.