2.1

Informed and automated K-mer size selection for genome assembly

Rayan Chikhi and Paul Medvedev

ABSTACT

Genome assembly and De Bruijn graph-based assemblers in particular, depend on selecting an optimal K-mer size, one of the most significant parameters determine both efficiency and accuracy of the assembly. Despite its importance, there are currently no automated tools to determine the best k or to assess the effectiveness of generating abundance histograms that would facilitate the assembly process. Consequently, the question this research becomes can an automated and accurate method be developed to estimate the optimal K-mer size, thus reducing manual effort and the time-consuming of parameter selection.

The authors present sampling-based method to count K-mer abundance much more quickly than traditional exhaustive counting methods. They then applied a fitting model to distinguish error-based from true genomic K-mer, approve an approximation the number of disting genomic K-mer in the histogram. The method selects the optimal value of k with maximum genomic K-mer. This technique was implemented in the tool KMERGENIE and tested on three genome datasets. The result indicate that KMERGINIE can predict an optimal K-MER size correctly and considerably improving assembly quality. The approach greatly reduces computational time, and assemblies generated using suggested K values demonstrate greater contiguity such as higher NG50 scores, compared to others methods.

Nevertheless, the paper demonstrates that KMERGINIE provides a pragmatic and effective means of selecting the optimal K value. The authors acknowledge limitations in coping with data of irregular coverage or complexes samples such as metagenomics. Future direction include improving the model to better handling heterozygosity and other complex genomic structures.

2.2

Abstract:

De Bruijn graph-based assemblers require k-mers to reconstruct genomes. First reads are split into k-mers and the graph is then constructed with (k-1)-mers as nodes and k-mers present in the reads as edges. The size of k-mers represent a trade-off between several effects. However, there is a lack of tools to estimate the optimum k automatically and to efficiently generate its abundances histograms while taking the repetitiveness of the genome, heterozygosity rate or read error rate into consideration. Here we present KMERGENIE, a method that we developed to find the best value of k. KMERGENIE uses a sampling-based approach instead of a time-consuming older approach. It then chooses the value of k that gives the maximum number of genomic k-mers by fitting a generative abundance model to each histogram. To validate its accuracy, we compared the assemblies of different datasets using a k value chosen by KMERGENIE to other assemblies. Our results demonstrate that KMERGENIE selects an informed k-mer size that leads to a good genome assembly. We anticipate that our method can be integrated into assembly pipelines so that the choice of k can be made automatically without user intervention. The method of automatically choosing k from non-uniform coverage could be tested and the accuracy of our statistical model can be improved ensuring efficient and precise sequencing data analysis.

3.1

Genome assembly is the process of reconstructing complete genomes from DNA sequencing reads. While long single-molecule sequencing (SMS) reads offer improved resolution over short-read data, assembling repetitive genomic regions, especially complex segmental duplications (SDs), is still challenging. Traditional assemblers (such as PacBio and ONT) often fail to fully resolve these repeats, limiting assembly quality. Previous methods struggle particularly with unbridged repeats and mosaic repeat structures, which are common in large genomes like humans. The key problem addressed by this study is how to improve the assembly of complex, highly repetitive genomic regions, using error-prone long reads. Here, the authors present Flye, a novel long-read assembly algorithm, which builds a repeat graph, and allowing it to resolve the repeats. Flye produces more contiguous and accurate assemblies compared to five state-of-the-art assemblers (Canu, Falcon, HINGE, Miniasm, MaSuRCA), nearly doubling the NGA50 for the human genome. Furthermore, Flye effectively reconstructs detailed mosaic SD structures of repeats. These results suggest that incorporating repeat graph approaches into long-read assembly has transformative potential, improving not just basic assembly metrics but also enabling deeper insights into genomic structures and variations. Flye's approach could significantly reduce the need for additional finishing experiments and enhance studies of genome evolution, disease, and structural variation. It provides a major step forward in leveraging long-read data for high-quality genome assemblies, paving the way for more comprehensive and accurate genomic research.

- Basic introduction
- Detailed background
- General problem
- "Here we show"
- Main result
- General context
- Broader perspective

P03: Assembly of long, error-prone reads using repeat graphs

Have we de-mystified repeat untangling in genome assembly? Yes and no – short-read data relies on tried-and-true approaches to reconstruct repetitive regions as well as conceivably possible given the inherent limitations in working with 300bp segments; long-read data, while indispensable for better repeat resolution, are a different – and yet undefeated – beast. Firstly, single-molecule sequencing reads fall short of being tractable to the classic de Bruijn graph method by one unfulfilled requirement – most k-mers in the genome must be preserved in multiple reads. Secondly, error-prone reads complicate the distinguishing of repeat copies with divergence below 10%. Both challenges torpedo attempts at resolving bridged and unbridged repeats. Here we show that long-read genome assembly can be significantly improved in quality and runtime by exploiting one of its main flaws – repeat copy variants. Counter-intuitively, we assemble reads sloppily and concatenate the resulting "disjointigs" arbitrarily, but arrive at a repeat graph guaranteed to be same as if derived from the complete genome. Finally, we use errors in repeats to find a Eulerian tour trough the graph – an assembly that is on par with or better than those produced by five state-of-the-art assemblers. Our algorithm doubled the contiguity of the human genome assembly. The potential for improvement in genome reconstruction seems far from exhausted, and our results show a direction worth further exploring. Algorithms tackling mosaic segmental duplications are needed; so are approaches resolving repeats with divergence below 3%. Once achieved, these milestones could elevate assembly contiguity, as measured by NGA50, by an order of magnitude.

BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA

Lars Gabriel, Tomáš Bruna, Katharina J. Hoff,Matthis Ebel, Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke

1 Abstract

Gene prediction in eukaryotic genomes is still a difficult task. Many genomes differ in the quality and amount of available data, making it hard to create accurate annotations. Previous tools like BRAKER1 and BRAKER2 could use either RNA-seq or protein data, but not both simultaneously. Some newer technologies try to enhance gene prediction by combining various kinds of data. However, there is still no easy and fully automatic tool that can use both RNA-seq and protein data to give precise gene predictions. Here we show that BRAKER3, a pipeline that runs GeneMark-ETP and AUGUSTUS for gene prediction and then uses TSEBRA to select the best transcripts with joint extrinsic evidence, enables accurate genome annotation. BRAKER3 performs more effectively than prior tools such as MAKER2, Funannotate and FINDER. It finds genes and transcripts more correctly, with F1-scores up to 20% higher. Although BRAKER3 is slower than some other tools, it works well even with large genomes and big protein databases. These results indicate that BRAKER3 is an effective tool for large genome studies such as the Earth BioGenome Project. The pipeline is provided as a Docker container for easy deployment, making it accessible for a wide range of users. This makes BRAKER3 a useful option not only for scientists working on genome research, but also for substantial genome projects that need fast and accurate results.

BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA

Lars Gabriel, Tomáš Bruna, Katharina J. Hoff,Matthis Ebel, Alexandre Lomsadze, Mark Borodovsky, and Mario Stanke

Abstract

Accurate gene annotation is essential for understanding the biological functions encoded in genomes, as well as for supporting research in many areas of the life sciences. As genome sequencing becomes faster and more affordable, the demand for accurate and highly automated annotation tools continues to increase. Gene annotation can be supported by both intrinsic and extrinsic evidence. Intrinsic evidence is based on computational models and statistical features of the genome sequence itself, while extrinsic evidence includes external data sources such as transcriptome data or known protein sequences. The previous annotation tools BRAKER1 and BRAKER2 use either RNA-seq data or protein data as evidence, but not a combination of both. Here we present BREAKER3, a new gene annotation tool for eukaryotic genomes, which integrates GeneMark-ETP, continues the annotation with AUGUSTUS, and combines the results with TSEBRA. With GeneMark-ETP, it is now possible to use both RNA-seq and protein data to improve annotation accuracy compared to previous models. In addition, the integration of the ab initio gene prediction tool AUGUSTUS enables the training of statistical models on high-confidence genes to further optimize the annotation. This approach of BRAKER3 outperforms both its predecessors and other gene annotation tools in benchmark tests. Evaluation on 11 well-annotated genomes. BRAKER3 achieved an average F1-score that was 20% higher than that of the compared tools, with the largest advantage observed on large and complex genomes. With its improved accuracy and ability to handle complex genomes, BRAKER3 enables more comprehensive and meaningful gene annotations.

5.1

Unicellular eukaryotes inhabit every ecological niche, playing vital roles in global ecosystems. 18S ribosomal DNA metabarcoding has revealed huge diversity among them, uncovering new taxons with significant potential for biotechnology and biomedicine. Prediction of eukaryotic protein-coding genes in metagenomic assemblies is often complicated by their exon-intron structure, typically large genome sizes and low content in samples. Existing methods usually require taxonomic binning or species-specific training data, limiting their use to poorly characterized or *de novo* discovered species with very few references. Therefore, there is a need for tools that can predict and reconstruct eukaryotic genes without prior knowledge of the organisms present in metagenomes.

Here we introduce MetaEuk, a scalable, reference-based tool for sensitive identification and annotation of eukaryotic protein-coding genes in metagenomes. MetaEuk accurately recovers complex gene structures within different eukaryotic clades, predicting over 12 million genes from 1,35 million contigs of Tara Oceans metatranscriptomic datasets. It uses 6-frame translation, rapid homology searches and exon-chaining to reconstruct gene models, achieving over 90% sensitivity in benchmarked genomes, even with low similarity to references. This approach outperforms traditional tools, dependent on genome binning or closely related training data.

These results show that MetaEuk enables efficient, large-scale discovery of eukaryotic genes from metagenomic data, making it particularly valuable for studying uncultured and diverse eukaryotes in natural environments. As environmental sequencing is becoming more common, it has a potential for uncovering new protein families, improving reference databases and advance functional and evolutionary analyses of eukaryotic microbial life.

244 words

6.1

Alternative Splicing in eukaryotes is the ability to express different versions of the same gene to adapt to different circumstances. Albeit short reads have a high quality, they lack the ability to cover multiple splice junctions. This limitation makes it hard to discover rare isoforms. On the other hand, long reads exceed in this filed but have lower quality and coverage leading to limited usability for differential expressions analysis. Here we show a novel analysis pipeline using long reads generated by PacBios Hifi Iso-Seq technology to not only discover new isoforms but to open the door for differential expression analysis using long reads. As an example, we are able to find a Nrap isoform, previously thought to only be expressed in cardiac tissue, to be present in all muscle tissue types. Additionally, we have a high enough coverage to compute meaningful PSI values for all tissue types. This information enables us to identify the isoforms using Sanger sequencing. Our results demonstrate how long read sequencing can provide not only information of novel splice events but also correct older believes based on insufficient data. The increasing accessibility of high quality long read data gives us more and more opportunity not only make new discoveries of splicing events but also to challenge our knowledge thus far. Thereby, we are not limited to the identification of splice events, but can perform quantitative analysis for a wast number of splicing events.

A long-read RNA-seq approach to identify novel transcripts of very large genes

Prech Uapinyoying, Jeremy Goecks, Susan M. Knoblach, et al. 2020

Alternative splicing allows individual genes to generate multiple mRNAs. Many of these mRNAs encode functionally distinct protein isoforms, thereby bridging the gap between genome and proteome[1,2]. Short-read sequencing struggles with alternative splicing analysis because individual reads typically cover no more than two exon junctions. This limits the ability to resolve full exon compositions and their phasing within transcripts.[3,4] Long-read sequencing addresses this by capturing entire transcripts, but it lacks the precision needed for accurate quantification, making it less suitable for differential expression analysis[5,6]. Here we show that combining PacBio long-read isoform sequencing with a novel analysis approach enables comparison of alternative splicing in large, repetitive structural genes in muscle. We found that Nrap isoforms excluding exon 12, previously considered cardiac-specific, are also expressed in skeletal muscle, and a rare isoform lacking both exons 2 and 12 is present in cardiac and soleus muscle. In Nebulin, we identified a novel exon (u-002), mutually exclusive splicing of exons 127 and 128, and several unannotated phased isoforms showing fiber-type–specific patterns. In Titin, we discovered exon 191 as a likely unannotated cassette exon present in all skeletal but only a subset of cardiac transcripts. Our quantitative analysis with full-length reads enabled isoform-level comparisons across tissues, demonstrating how long-read sequencing reveals complex, tissue-specific splicing and uncovers unannotated isoforms. Improved transcript identification and quantification from our approach removes prior barriers to quantitative differential expression of ultralong transcripts. Unannotated exons and splicing patterns may directly impact clinical sequencing and interpretation of muscle disease variants.

References

[1] Arianne J Matlin, Francis Clark, and Christopher WJ Smith. Understanding alternative splicing: towards a cellular code. *Nature reviews Molecular cell biology*, 6(5):386–398, 2005.

[2] Eddo Kim, Amir Goren, and Gil Ast. Alternative splicing: current perspectives. *Bioessays*, 30(1):38–47, 2008.

[3] Douwe Schulte, Weiwei Peng, and Joost Snijder. Template-based assembly of proteomic short reads for de novo antibody sequencing and repertoire profiling. *Analytical chemistry*, 94(29):10391–10399, 2022.

[4] Stacey A Simon, Jixian Zhai, Raja Sekhar Nandety, Kevin P McCormick, Jia Zeng, Diego Mejia, and Blake C Meyers. Short-read sequencing technologies for transcriptional analyses. *Annual Review of Plant Biology*, 60(1):305–333, 2009.

[5] Medhat Mahmoud, Daniel P Agustinho, and Fritz J Sedlazeck. A hitchhiker's guide to long-read genomic analysis. *Genome Research*, 35(4):545–558, 2025.

[6] Renato Santos, Hyunah Lee, Alexander Williams, Anastasia Baffour-Kyei, Sang-Hyuck Lee, Claire Troakes, Ammar Al-Chalabi, Gerome Breen, and Alfredo Iacoangeli. Investigating the performance of oxford nanopore long-read sequencing with respect to illumina microarrays and short-read sequencing. *International Journal of Molecular Sciences*, 26(10):4492, 2025.

7.1

During the last few years, techniques like iCLIP and eCLIP made it possible to study protein–RNA interactions with high resolution. We use specific patterns in the sequencing data, like truncation events, to find where proteins bind to RNA. However, many existing tools don't fully take into account effects like background noise or how often a transcript is present. This can make the results less reliable or lead to too many false positives. A main challenge is that, even with these high-resolution methods, it is still difficult to clearly identify the real binding sites of RNA-binding proteins, especially in noisy or biased data. Here we show PureCLIP, a program based on Hidden Markov Models that tries to improve the detection of crosslink sites by combining different types of signals in the data. Compared to other tools like CITS, CLIPper, or Piranha, PureCLIP seems to do better in many cases. In the paper, it's shown that PureCLIP finds binding sites more precisely, is more reproducible between experiments, and works even when the protein doesn't bind very strongly. Because of that, PureCLIP could be a useful method for future studies that want to look at RNA-binding proteins across the whole transcriptome. It can also be adapted to work with other similar kinds of data. In general, it might help to study things like long non-coding RNAs or weak interactions that are harder to detect. This could give new insights into how RNA and proteins work together inside cells.

7.2

PureCLIP: capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data

RNA-binding proteins (RBPs) are central to gene regulation, influencing various cellular processes. Understanding where RBPs bind to RNA is crucial for deciphering the complexities of post-transcriptional control. CLIP-seq methods are used to map these protein-RNA interactions at high resolution. However, existing methods don't fully account for biases and truncation patterns specific to iCLIP and eCLIP data, potentially leading to inaccurate identification of binding sites. This poses a challenge for researchers aiming to precisely define RBP binding landscapes. This study address this challenge by developing a computational method that improved the accuracy of CLIP-seq analysis. Here we show that PureCLIP, a hidden Markov model-based approach, simultaneously addresses peak-calling and individual crosslink site detection with enhanced accuracy. Unlike previous methods, PureCLIP explicitly models non-specific background signals and sequence biases, significantly reducing false positives and improving the precision of crosslink site identification. This advance offers a more reliable means of exploring protein-RNA interaction networks. Accurate identification of RBP binding sites is fundamental to understanding gene regulatory networks and will likely advance our knowledge of many biological processes. The algorithm and model presented can be used in other genomic applications.

9.1

Progressive Cactus is a multiple-genome aligner for the thousand-genome era

Joel Armstrong, Glenn Hickey, Mark Diekhans, Ian T. Fiddes, Adam M. Novak, Alden Deran, Qi Fang, Duo Xie, Shaohong Feng, Josefin Stiller, Diane Genereux, Jeremy Johnson, Voichita Dana Marinescu, Jessica Alföldi, Robert S. Harris, Kerstin Lindblad-Toh, David Haussler, Elinor Karlsson, Erich D. Jarvis, Guojie Zhang & Benedict Paten
2020
Nature, Vol. 587, 12 November 2020, Seiten 246–251
https://doi.org/10.1038/s41586-020-2871-y

More cost-effective genome sequencing technologies are leading to a rapid increase in available genomic data, which poses new challenges for alignment tools in terms of their performance and accuracy. What is needed now are tools for precise alignment in large and complex datasets.
However, current alignment tools often perform poorly with increasing numbers of genomes and often experience reference bias. In addition, they lack the ability to represent multiple orthologous regions correctly, especially in regions with duplications or rearrangements.
This raises the question of how multiple-genome alignments can be efficiently and accurately performed for tens to thousands of large vertebrate genones.
Considering this, Progressive Cactus was developed.
It is a scalable, reference-free whole-genome aligner designed to meet the demands of large-scale comparative genomics. It uses a progressive alignment strategy, breaking down the problem into smaller sub-alignments using phylogenetic guide trees. These sub-alignments are then combined step by step, while reconstructed ancestral genomes serve as bridges between related species, allowing for a better representation of evolutionary history.
Advantages over previous aligner tools include higher accuracy in benchmark tests (e.g. Alignathon) and good tolerance to poor assemblies. It is suitable for large-scale comparative studies and supports downstream applications such as gene annotation or variant discovery. It is also suitable for pangenomics and population comparisons.
In the era of high-throughput genomics, the ability to align many genomes precisely and without reference bias is key to understanding genome evolution, function, and diversity.
Furthermore, opens progressive Cactus new perspectives for future studies such as pangenome construction or population-scale genome comparison.

9.2

Abstract: **Progressive Cactus – improved and better - a multiple-genome aligner for the thousand-genome era**

Multiple-genome Alignment is a necessary method to find to find differences and similarities within DNA sequences. It is an essential tool to identify structural variations. The number of genome assemblies is growing more and more. Therefore, there is still the need to improve multiple-genome alignments to make it more efficient and accurately, because of the large amount of available data. Here we present progressive Catus, an improved version of the Cactus aligner, which enables a reference free multiple genome alignment with high accuracy, in context of vertebrate genomes. Existing aligners created reference bias or can not scale beyond a limited number of genomes because of computational constrains. Progressive Cactus overcomes these problems, using a progressive alignment strategy with ancestorial reconstruction with a linear runtime scale. This strategy improves the alignment accuracy and flexibility significant. This way, Progressive Cactus enables a new era of comparative genomics. It lets researchers analyze genome evolution and variation across whole clades. This study provides a foundation for better comparisons of vertebrate genomes and other species, which highlights its flexibility.

10.1

Proteins are fundamental to life because they carry out essential tasks inside cells. Understanding how they fold into their 3D shape is important, since their structure is linked to their function.

So far, scientists have solved the structures of about 100,000 proteins, but there are billions of known protein sequences. Determining a single structure experimentally is slow and costly, which limits our overall understanding. Computational approaches have tried to predict 3D structures from sequences, using physics-based models or evolutionary information. But until recently, these methods were not accurate enough.

This study addresses the long-standing problem of how to predict protein structures with high accuracy when there is no template to rely on.

Here we show that AlphaFold, can predict protein structures at atomic-level accuracy in most cases, even without known homologous structures.

AlphaFold was tested in the CASP14 challenge and clearly outperformed all other methods. It achieved a median backbone accuracy below 1 Å, which is close to experimental precision. This represents a major leap forward compared to previous computational approaches.

These findings suggest that deep learning can capture the rules of protein folding better than traditional approaches. With AlphaFold, it is now possible to predict the structure of many proteins that were previously inaccessible.

This breakthrough could transform structural biology, as it makes it possible to analyze proteins on a much larger scale. It will likely support future research in medicine, biotechnology, and basic biology by helping to understand diseases, design drugs, or discover new functions of proteins.

10.2

Abstract of "Highly accurate protein structure prediction with AlphaFold by Jumper et al. in Nature (2021)

Introducing into this field of understanding three-dimensional structures of proteins is essential for insights into their biological functions and mechanisms. Normally, determining these structures required a lot of experimental effort through techniques such as X-ray crystallography, NMR etc. that limited the determination of the entire proteins out there. Therefore, computational methods have been developed to fill this gap, using either physics-based simulations or evolutionary analyses, but with limited success, especially when no homologous structures are available. The development of deep learning has recently enabled substantial improvements in structure prediction.[1] However, existing methods fall short of atomic accuracy, particularly when faced with novel folds or blurred evolutionary and experimental data. To overcome this limitation, this study introduces a new deep learning model called 'AlphaFold' that leverages evolutionary, geometric, and physical constraints to predict protein structures directly from amino acid sequences.[2] **The research question of this study** is based on whether 'AlphaFold' can computationally predict protein structures at atomic accuracy even in the absence of homologous structures.[3] Here we show that 'AlphaFold' was validated through the CASP14 blind assessment, where it achieved a median backbone accuracy of 0.96 Å r.m.s.d.95, outperforming all other methods by a substantial margin. The model also demonstrated a strong generalization to novel proteins, high side-chain fidelity, and scalability to long sequences, while providing confidence estimates that closely correlate with the actual accuracy.[4] These results establish 'AlphaFold' as the first computational method to routinely achieve near-experimental accuracy in protein structure prediction. The study concludes that 'AlphaFold' represents a groundbreaking

[1] Overview to the field of research
[2] Addressing the knowledge gap and introducing the filler for this gap
[3] Research question of this study
[4] Key results of this study

advancement in structural biology, enabling applications in drug discovery and large-scale proteome analysis by applying this method on the entire human proteome.[5]

[5] Conclusion of this study

11.1

Public sequence repositories have grown rapidly, but the extent of mislabeled or contaminant sequences remains largely unknown. Existing quality controls like VecScreen and BLAST against known vectors can detect common synthetic contaminants, but they do not manage to find cross-kingdom mislabeling. Without modern approaches, contaminated sequence entries can remain undetected, potentially misleading future analyses in genomic research. In this paper, we aim to find out how much are the nucleotide and protein sequence databases contaminated by incorrect sequence labeling. Here we show that, in the sequence databases GenBank and RefSeq we detected over 2.16 million and 114000 contaminated nucleotide entries. The strongest source for contamination we found out to be human nucleotide sequences. Also we found 14148 protein contaminants in NR. This result shows that new methods for detecting cross kingdom contamination in nucleotide and protein databases are needed. Additionally the findings show a fundamental challenge in modern biology, as vast amounts of data are generated, ensuring its accuracy becomes a critical challenge. Contamination in sequences can not only affect genomic studies but can also mislead research in other fields like medicine, ecology and biotechnology. Therefore cooperation between different fields is needed to develop and implement robust strategies and standards for contamination detection, to safeguard the validity of scientific research.

Abstract

Proteins have a wide range of crucial functions in the body, including structural support, acting as enzymes and hormones, and transporting molecules. They also play a vital role in the immune system, fluid balance, and energy production. In recent years, the explosive growth in protein structure predictions, especially via AlphaFold, has led to the availability of over 214 million predicted protein structures. To make sense of this vast structural space, this study developed a structure-based clustering approach called Foldseek cluster to identify millions of proteins by structural similarity. Using this method, they clustered the entire AlphaFold database into 2.30 million non-singleton structural clusters, 31% of which lack known annotations.

This study highlights that while most clusters are conserved and likely ancient, 4% of them appears species-specific, indicating possible de novo gene birth or prediction artifacts. This study also inferred that putative domain families play a important role in evolutionary links across species. Furthermore, remote structural homologies were identified between human immunity proteins and bacterial counterparts, suggesting ancient shared mechanisms. Additionally, thousands of unannotated clusters with confident structural predictions were discovered, some of which may function as enzymes or membrane-bound transporters.

This work provides a crucial framework for exploring protein structure space at scale and deepens our understanding of evolutionary relationships, immune system evolution, and functional diversity across life.

13.2

Fast clustering of large number of predicted protein structures for analysis

Proteins are essential for life and in order to better understand cell processes, it is important to have knowledge of proteins structures as this can provide important insights into their function and evolution.

In the past decades thousands of protein strcutures have been experimentally determined by biologists. Since the introduction of AlphaFold in 2021 it is possible to computationally predict highly accurate three-dimensional protein structures from protein sequences. The AlphaFold database contains nowadays over 214 million predicted protein structures.

In turn the research and analysis of such a big number of protein structures requires a comprehensive and efficient method.

For the purpose of this we developed a structural clustering algorithm named FoldSeek which can compare hundred millions of protein structures at once and group them into clusters based on their similarity at an improved speed.

We identified ~2,3 million clusters consisting of non-singleton structures of which 31% have no similarity to previously known structures. Distribution after mapping of cluster-members in the tree of life: Cellular organism (23%), bacterial (16.1%), Eukaryota (13.5%) and Archaea (0.5%).

“Transporter activity“ detected as the most widespread protein function. Several immunity-related proteins with structural similarty in bacterial proteins have been found.

Assuming that the number of records in publicly accessible protein strcuture databases is constantly increasing a highly scalable structure-based clustering algorithm is needed to be able to explore billions of structures in future.

Clustering of known and unknown protein structures may help scientists studying proteins function and evolution more precisely and can pave the way for the development of new drugs.

Abstract for Fast and sensitive taxonomic assignment to metagenomic contigs

Metagenomic studies have revolutionized our understanding of microbial diversity, yet accurately assigning taxonomic labels to assembled contigs remains a computational challenge.Current tools like CAT use prokaryote designed gene prediction (Prodigal) and require manual parameter tuning, limiting their performance on eukaryotic contigs and scalability. While short read classification is well established, contig level assignment should improve accuracy but demands faster, domain covering methods.To address these limitations, the authors developed MMseqs2 taxonomy, a novel protein-search-based tool that overcomes prokaryotic bias and automation challenges. The key question was whether a more comprehensive approach could achieve faster, more accurate taxonomic assignment across all domains of life.MMseqs2 taxonomy extracts all possible protein fragments from contigs (suitable for all domains of life), retains fragments with database similarity, and implements an approximate 2bLCA strategy for taxonomic assignment. Benchmarking showed it classifies bacterial contigs 18× faster than CAT with equal accuracy (CAMI-I dataset). For eukaryotic SAR contigs, it achieved 62% assignment (vs CAT's 47%) with 46% species-level precision (vs 28%). The tool automatically determines search parameters, eliminating manual tuning.This work presents MMseqs2 taxonomy as a fast, sensitive solution for metagenomic contig annotation that outperforms existing methods in both bacterial and eukaryotic datasets. Its open source implementation, compatibility with standard databases (UniProt, GTDB…), and included utility modules make it a versatile tool for advancing metagenomic research across the tree of life.

14.2

Metagenomics enables the identification and analysis of uncultivated microorganisms in natural environments, leading to the discovery of novel microbial species. This is of great significance for understanding the functional potential of microbial ecosystems. Metagenomic sequencing of environmental samples generates large numbers of contigs derived from complex microbial communities. However, simply assembling these contigs does not reveal which organisms they originate from. Accurate taxonomic annotation of contigs is therefore essential to assign fragments to specific taxa. Yet, the state-of-the-art tools for taxonomic annotation of metagenomic contigs have limitations. Here the authors show that the new tool, MMseqs2, is faster, more automatic, and similarly accurate to the previous tool on bacterial and eukaryotic datasets. They found that this new tool is suitable for eukaryotic reads, multithreaded performance in metagenomic applications, and does not require manual selection of a key parameter. Moreover, it is ~ 18x faster on a bacterial dataset and ~ 4x faster than existing tools on a eukaryotic dataset. MMseqs2 supports microbial ecology, metagenomics, and epidemiology by making it easier to analyze microbial diversity and taxonomy. For example, MMseqs2 promotes the discovery of new microbial species and their functions. It can detect pathogenic microorganisms in the human intestinal tract, advancing the understanding of microbial diversity and enabling the development of new therapies. Moreover, it identifies antibiotic- and antifungal-producing microbes in soil, which could lead to the development of novel treatments.

15.1

Abstract of

“Accurate proteome-wide missense variant effect prediction

with AlphaMissense”

Missense variants change single amino acids in proteins. Many cause genetic diseases. Over 4 million human missense variants exist, but 98% lack clinical classification (VUS). Current tools like PolyPhen-2 rely heavily on limited data. This gap delays rare disease diagnosis and therapy development. Here we introduce AlphaMissense, a new prediction model. We show this model classifies 89% of 71 million variants across 19,233 human proteins as pathogenic or benign with 90% precision. It outperforms tools like EVE by 3.2% (AUC-ROC 0.940 vs. 0.911). Our method combines protein evolution patterns and human population data, avoiding biased clinical annotations. It works well even for challenging proteins like transmembrane domains. These results resolve key limits in VUS interpretation. All predictions are freely available online. They provide immediate help for diagnosing rare diseases. This framework also offers new tools for developing precision treatments.

15.2

Abstract

A key challenge is identifying the functional impact of the large number of missense variants found in human genome sequencing. Most of these variants are still clinically unclassified. This complicates rare disease diagnostics and limits progress in personalized medicine. To show this, the study introduces AlphaMissense, a machine learning tool based on AlphaFold, fine-tuned using human and primate variant population frequencies. The model predicts pathogenicity for all possible single amino acid substitutions across the human proteome. AlphaMissense integrates structural context and evolutionary conservation to assign a pathogenicity score for each variant. Here we show that it performs well on diverse genetic and experimental benchmarks like ClinVar and MAVE datasets, This indicates strong generalization and helps reduce bias from known variant databases. A notable finding is that the average pathogenicity score per gene correlates with gene essentiality, particularly benefiting short essential genes. The study fills an important knowledge gap in variant interpretation, especially for regions which are hard to classify. Most of the possible human missense variants are categorized as benign (57%) or likely pathogenic (32%), based on thresholds set for 90% accuracy using the ClinVar dataset. This accessible resource is anticipated to accelerate research in molecular biology and human genetics, support clinicians in disease diagnosis, and offer valuable insights into complex human traits.