Aktuelle Themen der Sequenzanalyse

12.07.2021

Goethe Universität,
Master Bioinformatik
Sommersemester 2021

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

Prech Uapinyoying, Jeremy Goecks, Susan M. Knoblach, Karuna Panchapakesan, Carsten G. Bonnemann, Terence A. Partridge, Jyoti K. Jaiswal, and Eric P. Hoffman

Genome Research, 2020, 30(6), 885–897

doi: 10.1101/gr.259903.119

GRUPPE

Mustafa Massli, Henni Husso

Mustafa Massli
01.06.2021

Seminar „Aktuelle Themen der Sequenzanaylse, SoSe 2021“

Goethe-Universität Frankfurt am Main

Paper Abstarct:

Mammalian genome includes long coding exons and repetitive regions, which showed highly alternative spilcing. To alignment this kind of gens Short-read RNA seq has been used to determine the full-lengths and complete splicing pattern of the transcripts, but a precise prediction of a large protein with short-read sequencing is challenging. Therefor PacBio Isoform-seq were applied for a full-lengths Sequencing. Applying an exon phasing approach that leads to an accurate quantification of differential expression of ultralong transcripts. Extension of the identification approach of the novel exons and long transcript isoforms to compare their usage between samples. The approach that used in the papier reduced the difficulty to differential expression analysis of ultralong and difficult transcripts. Additionally, we found mismapped gens and unannotated exons, which make the Iso-Seq generally recommended for transcript isoforms. PSI and percentages of isoform values cannot be used for comparisons. The results of the developed approach showed that can be used with large and repetitive transcripts. The Data shows significance for clinical pathogenicity, which because of the identified of new exons and splicing pattern can improve the simplification of the pathogenicity.

Abstract of the Paper

Seminar "Aktuelle Themen der Sequenzanalyse SoSe 2021"

Henni Husso

May 30, 2021

Paper-ID: 01
Paper Title: A long-read RNA-seq approach to identify novel transcripts of very large genes.

Common RNA sequencing technologies (RNA-seq) are not fitted for ultralong or repetitive sequences. Traditional short-read methods are the most precise but turn ambigious for exceeding sequence lengths and number of repetitions. Especially the computation of possible alternative splicing patterns are difficult. While long-read platforms encompass longer read lengths, they are still capped and the issue remains. With both alternatives, does the quantification of transcriptomes persists to be challenging. Here, an exon-based analysis pipe line designed to use long-read isoform sequencing (Iso-Seq by PacBio) was successfully applied as comparison tool for the alternative expression of large transcripts accross multiple samples. The analysis of three transcripts coding for mammalian structural muscle proteins (Titin, Nebulin and *Nrap*) and sampled from three different muscle tissues of mouse confirmed annoted as also help discover unannoted splice variants. The identified exon usage and phasing yielded relative quantities of each transcript for each source tissue (cardiac, slow/fast skeletal muscle). All results were found to attest muscle-specific adaptions which were detected in former studies using short-read RNA-seq. Since rare and novel exons were identified, these findings can be used to extend or reassess previous knowledge of these muscle proteins as also their clinical relevance. Overall, does the presented approach prove to overcome shortcomings of long-read sequencing which could be utizilized for a comprehensive strategy for more profound differential gene expression studies.

PAPER SUMMARY

ADRESSED RESEARCH QUESTION

- difficulties in studying differential gen expression of long, repetitive sequences
- goal: accurate quantification of differential expression (ultralong transcripts)
  - overcome limitation from short reads sequencing (RNA-seq)
  - overcome shortcomings of long read sequencing
- main method: Iso-Seq
- comparison point: murine muscle types specific differential expression of transcripts (in mammalian mouse/human)
  - cardiac apex (heart)
  - extensor digitorum longus (EDL; fast skeletal muscle of the lower tight)
  - soleus (slow skeletal muscle of the lower tight))
- Key references publication: Savarese et al. (2018)

RELEVANT METHODS AND APPROACHES

- Iso-Seq (using PacBio protocol) + exCOVator & exPhaser analysis pipeline
  - Iso-Seq
    - corrects the random insertion and deletion of the PacBio's real-time sequencing technology errors
    - qualitative cluster reads → quantitative full-length reads
      - collapsing similar full-length reads into consensus (cluster) reads reads
  - exCOVator:
    - differential exon usage (PSI) analysis
    - identification of novel exons
  - exPhaser:
    - determine splicing patterns
    - identify transcript structure
- Short-read RNA-seq as comparison point
  - data processing and alignment
  - data obtained from NCBI sequence Read Archive
- RT-PCR and Sanger sequencing as confirmation method

RELEVANT RESULTS

Comparison: Long-Reads vs. Short-Reads Sequences

- long reads perform better than short reads at resolving intergenic and intragenic regions of the genome
  - short-read: substantial reads mismapped and spit accorss genes (myosin mapping, Supp. Fig. S2)
  - long-read: correct alignment (myosin mapping, Supp. Fig. S2)
- long-read sequencing strenth is phasing of mulitple neighboring exons
- Oligo(dt) internal priming shows beneficial use for covering very large gene

PacBio long-reads Iso-Seq detects novel muscle-type-specific splicing patterns

- can be used as **semi-quantitative approac**h for differential exon usage between these tissues
- phasing multiple neighbouring exons to determine transcript splice patterns **within the same read**

exCOVator and exPhaser analysis results

Nrap

- best read-average length match
- significant findings:
  - Nrap-c single expressed isoform in cardiac muscle
  - Nrap-c differentially expressed together with Nrap-s in skeletal muscle
    - Nrap-s and Nrap-c could be misnomers
  - rare transcript detected (lacking exon 2 and 12)
  - exCOVator results are close to previous data from other studies

Nebulin

- big transcript size (22 kb); bigger than max. read-length of Iso-Seq
- truncutated transcript: internal priming resolved issue
- significant findings:
  - novel transcripts found in Z-disk and super-repeat region at 3'
  - majority of differential splicing in Z-disk region

Titin

- biggest transcript size (103 kb)
- exons analyzed in groups based on proximity
- mouse findings partially conserverd in humans (CardioDB, TITINdb, etc.)
- significant findings:
  - exon 191: removed in subset of cardiac scripts (mouse)
  - exon 312/363: excluded more often in skeletal muscle transcripts than heart(human)
  - exon 45: expressed more in skeletal muscles
  - exon 11-13: fast EDL excludes more exons compared to soleus

CONCLUSION

- approach proved to work on very large and repetitive genes
- no additional short-read approach needed → cost reducing
- increase of sensitivity possible
  - ex. rare Nrap isoforms by phasing known casette exons
- supports relative quantification of same exon/isoform across multiple samples
  HOWEVER
  - "PSI and percentages of isoform values cannot be used for comparisons across genes"
  - absolute quantity of RNA can be identical even though percentage differ across samples
- not all gene support internal priming
- data shows significance for clinical pathongenicity

exCOVator (exon-based) approach

- enables differential exon usage (PSI) analysis & identification of novel exons
- (285 differentially used exon findings (14 novel across 51 genes; low number of artifacts))

exonPhaser approach

- able to determine splicing patterns
- able to identify transcript structure
- map to known annotations
- quantifies relative expression of genes accross muscle samples

Nrap

- exon 2 is speculated to modulate myofibrillogenesis during development

Nebulin

- suggests functional role of nebulin in regulating Z-disk width in different muscles
- novel alternative splicing of exons 127- 128
  - super repeat region bordering Z-disk anchorage point
    mutually exclusive expression
  - binding sites of KLH40 (loss causes nemaline-like myopathy)
  - KLH40 stabilizes nebulin (and LMOD3)

Titin

- exon 191
  - encodes for Ig-like domains in I-band region
  - Ig-like domain ˜elasticity and stiffness of sarcomeres in heart
  - suggests: casette exon spliced differently to adjust length of I-band and thus heart muscle stiffness
- exon 312/363 in humans:
  - may play a role in: mechanosensitivity, cleavage of C-terminal titin
  - local remodeling of sacromere/production of cleaved titin fragments for cellular signaling
- exon 45:
  - speculation: play role in myofibrillar signaling during muscle development and cardiac disease
- exon 11-13:
  - codes for Z-repeat domain 4-6 (N-terminal region of titin imbedded in Zdisk)
  - speculation: involved in assembling Z-disks of variable width correlation: Z-disk width

KEY ILLUSTRATION

We choose **figure 5** of the main publication, because it shows an analysis step of the proposed novel approach of the study.

Figure 5 shows the unannotated cassette exon 191 of the muscle gene titin. The graph A produced by exCOVater analysis, shows a line graph depicting the full length match divided by the total amount of reads. The red peak points to the exonic part 129 or 191. This indicates the expression of this exon in lesser amounts in cardiac muscles compared to the skeletal muscles. In the bottom shows the stacked graph with the total length of all exonic parts. In both skeletal muscles the exon 191 is present and only removed from the cardiac. (B) shows a Sashimi plot with all consensus reads, in which also showcases how exon 191 was less expressed in the cardiac muscle compared to the skeletal muscles. In (C) the Agarose gel of a RT-PCR run confirms the results of graph (A) and (B). In (D) a Sanger sequencing run also confirms the afformentioned results.

PAPER SUMMARY

ADRESSED RESEARCH QUESTION

- difficulties in studying differential gen expression of long, repetitive sequences
- goal: accurate quantification of differential expression (ultralong transcripts)
  - overcome limitation from short reads sequencing (RNA-seq)
  - overcome shortcomings of long read sequencing
- main method: Iso-Seq
- comparison point: murine muscle types specific differential expression of transcripts (in mammalian mouse/human)
  - cardiac apex (heart)
  - extensor digitorum longus (EDL; fast skeletal muscle of the lower tight)
  - soleus (slow skeletal muscle of the lower tight))
- Key references publication: Savarese et al. (2018)

RELEVANT METHODS AND APPROACHES

- Iso-Seq (using PacBio protocol) + exCOVator & exPhaser analysis pipeline
  - Iso-Seq
    - corrects the random insertion and deletion of the PacBio's real-time sequencing technology errors
    - qualitative cluster reads → quantitative full-length reads
      - collapsing similar full-length reads into consensus (cluster) reads reads
  - exCOVator:
    - differential exon usage (PSI) analysis
    - identification of novel exons
  - exPhaser:
    - determine splicing patterns
    - identify transcript structure
- Short-read RNA-seq as comparison point
  - data processing and alignment
  - data obtained from NCBI sequence Read Archive
- RT-PCR and Sanger sequencing as confirmation method

RELEVANT RESULTS

PAPER IMPACT

AUFGABE 1

REFERENCE PAPER 1

Authors: Marco Savarese, Per Harald Jonson, Peter Hackman
Title: The complexity of titin splicing pattern in human adult skeletal muscles
Journal: Skeletal Muscle volume 8, Article number: 11
Year of Publication: 2018
Number of citations (in the paper): 10

Reasons for choosing:
- comprehensive study of differential expression in skeletal muscle of adult human
→ describes complex and tissue specific splicing pattern of Titin
- uses short-read RNA-seq in combination with RT-PCR and Sanger sequencing

REFERENCE PAPER 2

Authors: Sean P. Gordon, Elizabeth Tseng, Feng Chen, Zhong Wang
Title: Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing
Journal: PLOS (online)
Year of Publication: 2015
Number of citations (in the paper): 3

Reasons for choosing:
- identify complex gene expression mechanism
- long-read sequencing strategy for polycistronic RNA
→ tool for precise characterization of complex transcriptomes

REFERENCE PAPER 3

Authors: Ravi K. Singh, Arseniy M. Kolonin, Thomas A. Cooper

Title: Rbfox Splicing Factors Maintain Skeletal Muscle Mass by Regulating Calpain3 and Proteostasis
Journal: Cell Reports volume 24 (1): 197 – 208
Year of Publication: 2018
Number of citations (in the paper): 3

Reasons for choosing:
- knockout study to determine alternative splicing of Capn3 protein in skeletal muscle
→ information EDL, soleus muscles
- RNA-seq reference data

REFERENCE PAPER 4

Authors: Kati Donner, Kristen J. Nowak, Carina Wallgren-Pettersson
Title: Developmental and muscle-type-specific expression of mouse nebulin exons 127 and 128
Journal: Eur J Hum Genet 12: 744-751
Year of Publication: 2006
Number of citations (in the paper): 4

Reasons for choosing:
- information for nebulin gene
- quantified the relative amounts of transcripts produced by alternative splicing

REFERENCE PAPER 5

Authors: Simon Anders, Alejandro Reyes, Wolfgang Huber
Title: Detecting differential usage of exons from RNA-seq data
Journal: Genome Res. 22(10): 2008–2017
Year of Publication: 2012
Number of citations (in the paper): 1

Reasons for choosing:
- information on bioinformatical analysis of exon usage (on a genome-wide scale)
→ assess biological variability

AUFGABE 2

According to Google Scholar, the paper was cited a total of 5 times. So, our pick of the 5 most influential papers that cited this paper are:

Robinson et al. (2021): Dissecting the transcriptome in cardiovascular disease. Cardiovasc. Res. cvab117. https://doi.org/10.1093/cvr/cvab117
Impact Factor 5.35 IF

Zhao X et al. (2021): Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. Am. J. Hum. Genet. 108(5): 919-928. https://doi.org/10.1016/j.ajhg.2021.03.014
Impact Factor 10.5 IF

Qin L et al. (2020): RJunBase: a database of RNA splice junctions in human normal and cancerous tissues. Nucleic Acids Res. 49(8): D201–D211. https://doi.org/10.1093/nar/gkaa1056
Impact Factor: 16.48 IF

Savarese M et al. (2020): Panorama of the distal myopathies. Acta Myol. 39(4): 245-265. https://dx.doi.org/10.36185%2F2532-1900-028
Impact Factor: 1.36 IF

Serrano MC et al. (2021): Biallelic loss-of-function OBSCN variants predispose individuals to severe, recurrent rhabdomyolysis. bioRxiv (pre-print). https://doi.org/10.1101/2021.06.04.447044
Impact Factor:-/- IF

AUFGABE 3

PAPER CHOICE 1: Most cited of all publicly available publications.

Suneil H et al. (2021): Toll/Interleukin-1 Receptor Domain-Containing Adapter Inducing Interferon-β Mediates Microglial Phagocytosis of Degenerating Axons. J. Neurosci. Res. 32(22): 7745-7757. https://doi.org/10.1523/JNEUROSCI.0203-12.2012

PAPER CHOICE 2: Second most cited of all publicly available publications.

Emily G et al. (2014): A selective thyroid hormone β receptor agonist enhances human and rodent oligodendrocyte differentiation. Glia. 62(9): 1513-1529. https://dx.doi.org/10.1002%2Fglia.22697

PAPER CHOICE 3: Third most cited of all publicly available publications.

Deanne M et al. (2019): The Pediatric Cell Atlas: Defining the Growth Phase of Human Development at Single-Cell Resolution. Dev. Cell 49(1): 10-29. https://doi.org/10.1016/j.devcel.2019.03.001.

PAPER CHOICE 4: P. Uapinyoying is the lead author and the paper has been cited by other publications.

Prech Uapinyoying et al. (2020): A long-read RNA-seq approach to identify novel transcripts of very large genes. Genome Res. 30: 885-897. https://doi.org/10.1101/gr.259903.119

PAPER CHOICE 5: The paper has been cited a number of times.

Jaya P et al. (2016): Targeted Re-Sequencing Emulsion PCR Panel for Myopathies: Results in 94 Cases. J. Neuromuscul. Dis. 3(2): 209-225. https://doi.org/10.3233/JND-160151

Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak

Author: Tao Zhang, Qunfu Wu, Zhigang Zhang

Current Biology, 2020, Volume 30, Issue 7, p1167 – 1356, R287 – R328

Presented by Yangyu Li, Jiexin Chen

Abstract 1

There are three deadly coronaviruses, SARS-CoV, MERS-CoV and SARS-CoV-2, have emerged in human population causing hundreds of thousands of deaths. The COVID-19 pandemics caused by SARS-CoV-2 broke out in Wuhan in China and has spread worldwide. The bat is the most likely species of origin for SARS-CoV-2 because BatCov RaGT13 96% identical to SARS-CoV-2 at whole genome level. But like SARS-CoV and MERS-CoV, SARS-CoV-2 may get first into intermediate hosts, then leap to human. It is essential to find the potential intermediate hosts of SARS-CoV-2 to control the outbreak of COVID-19 pandemic. They compared the sequences of Pangolin-CoV, RaGT13 and SARS-CoV-2 to determine whether pangolin is a potential intermediate host of SARS-CoV-2. Through their study of pangolins, they suggest that pangolins are a natural reservoir of SARS-CoV-2-like coronaviruses. At the whole genome level, Pangolin-CoV is 91.02% identical to SARS-CoV-2 and compared with RaTG13 (BatCov), it is the second closest relative of SARS-CoV-2. Five key amino acids in the receptor-binding domain are same between Pangolin-CoV and SARS-CoV-2, but only SARS-CoV-2 contains a potential cleavage site for furin proteases. Aside from RaTG13, the Pangolin-CoV is the CoV most closely related to SARS-CoV-2. However, whether pangolin species are well candidates for SARS-CoV-2 origin is still uncertain. Considering that coronaviruses are widespread in other species, such as bats, camels and so on, more coronaviruses data of potential species need to be analyzed to reveal the most likely intermediate hosts.

Abstract 2

A new coronavirus called SARS-CoV-2 was discovered in Wuhan, China, in late 2019 and soon ravaged the world. The virus is extremely contagious and threatens to mutate and cause great harm worldwide. But the source of the epidemic has still not been found. Members of this virus family have the longest genome of any RNA virus and express up to 29 proteins, for which genetic and protein sequencing is available for the possible origin of the virus. A sketch of the coronavirus genome of the target species was reconstituted by a reference-guided scaffolding approach. Protein-based BLAST analysis and Simplot analysis can be performed to obtain overlapping groups of target species and sample proteins as well as identity. Identifying the source of the virus, that is, the animal from which it was transmitted to humans, will help to identify the initial transmission route, mutation patterns and potential risks, so that we can take more targeted and effective preventive and control measures. Evidence for the pangolin as a possible intermediate host for the virus is shown here by sequence comparison. The authors recombined the viral and protein sequences carried by Pangolin through de novo assembly, and by comparison with the human SARS-CoV-2s genome, the nucleotides identity was 91.02% and the amino acid identity was 95.41%. The sequence was also compared with the coronavirus carried by bats, and the concordance was over 80%. Some pangolin-Cov genes are more consistent with SARS-CoV-2s than bats. Although the genetic sequence of the pangolin is highly

similar to that of SARS-CoV-2s, the limitations of the sample did not allow further confirmation of whether the pangolin was the species of origin of SARS-CoV-2s. Considering that coronaviruses are widespread in other species, such as bats, camels and so on, more species of coronaviruses data need to be analyzed to reveal the most likely intermediate hosts.

Summary

Researched questions

- Assessing the Probability of SARS-CoV-2-like CoV Presence in Pangolin Species
- Draft genome of Pangolin-CoV and its genomic characteristics.
- Phylogenetic Relationships among Pangolin-CoV, RaTG13, and SARS-CoV-2.
- Dualism of the S protein of Pangolin-CoV.
- Amino acid variations in the nucleocapsid (N) protein for potential diagnosis.

Relevant methodological approaches

- De novo assembly and Blast analysis against protein.
- Reference-guided scaffolding approach, Simplot analysis and sequence alignment.
- phylogenetic trees.
- amino acid phylogenetic tree.
- Phylogenetic analysis based on the N protein.

Relevant results

- 22 contigs were best matched to SARS-CoV-2s (70.6%–100% amino acid identity; average: 95.41%). And 12 contigs matched to bat SARS-CoV-like CoV (92.7%–100% amino acid identity; average: 97.48%).
- The assembled Pangolin-CoV draft is reliable. Pangolin-CoV showed high overall genome sequence identity to RaTG13 (90.55%) and SARS-CoV-2 (91.02%) throughout the genome. some Pangolin-CoV genes showed higher amino acid sequence identity to SARSCoV-2 genes than to RaTG13 genes, including the spike (S) protein (97.5%/95.4%).
- In all phylogenies, Pangolin-CoV, RaTG13, and SARS-CoV-2 were clustered into a well-supported group, here named the ‘‘SARS-CoV-2 group’’. Within this group, RaTG13 and

SARSCoV-2 were grouped together, and Pangolin-CoV was their closest common ancestor.

- S1 protein of Pangolin-CoV is more closely related to that of 2019-CoV than to that of RaTG13. Within the receptor-binding domain, Pangolin-CoV and SARS-CoV-2 were highly conserved, with only one amino acid change, which is not one of the five key residues involved in the interaction with human ACE2. But RaTG13 has changes in 17 amino acid residues, 4 of which are among the key amino acid residues. Only SARS-CoV-2 contains a potential cleavage site for furin proteases.

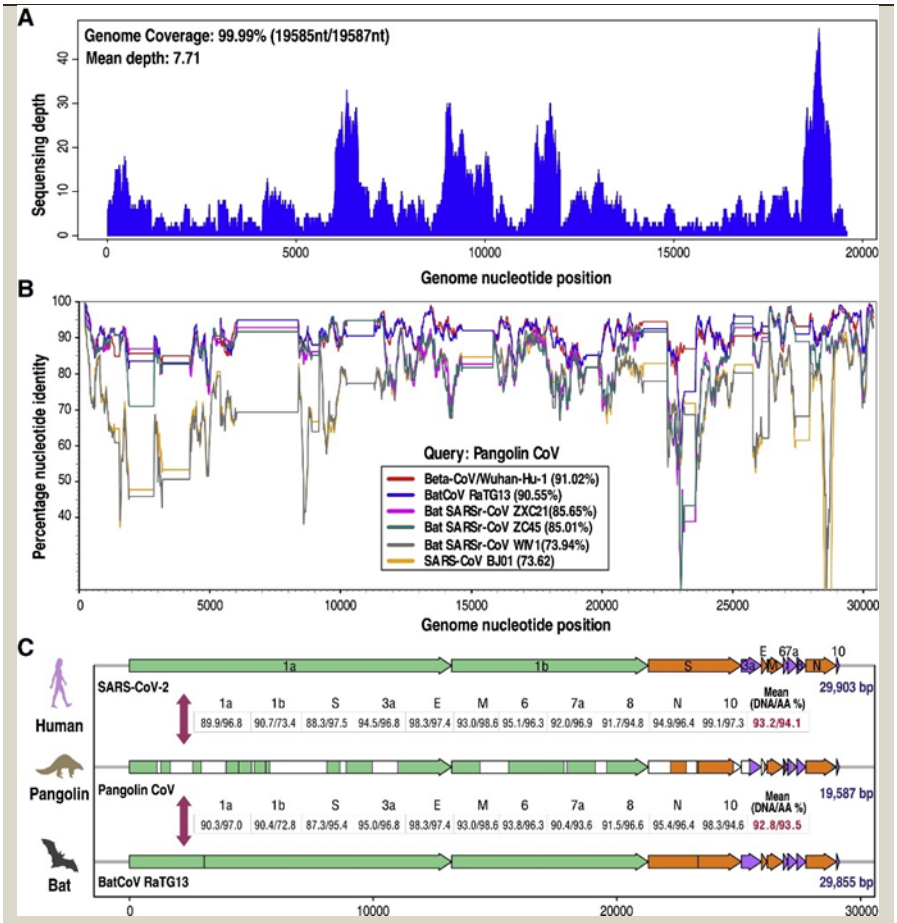
- The analysis supported the classification of Pangolin-CoV as a sister taxon of SARS-CoV-2 and RaTG13.

Conclusion

- Malasian pangolin may carry a novel coronavirus similar to SARS-CoV-2.
- Pangolin-CoV might be the common origin of SARS-CoV-2 and RaGT13. The high S protein amino acid identity implies functional similarity between Pangolin-CoV and SARS-CoV-2.
- This correspondence indicates that our Pangolin-CoV draft genome has enough genomic information to trace the true evolutionary position of Pangolin-CoV in CoVs.
- Pangolin-CoV could have pathogenic potential similar to that of SARS-CoV-2. But Whether the Pangolin-CoV or RaTG13 are potential infectious agents to humans remains to be determined.
- The observed amino acid changes in the N protein would be useful for developing antigens with improved sensitivity for SARSCoV-2 serological detection.

Figure 1: Genome-Related Analysis

Key illustration



We think the key figure is Figure 1. Because this figure shows the similarity between Pangolin-CoV and Bat-Cov, SARS-CoV-2. The Figure 1A shows the reliability of their Pangolin-CoV data. In particular, Figure 1B shows similarity based on the full-length genome sequence of Pangolin-CoV compared with for example BatCoV RaTG13. And Figure 1C is the Comparison of common genome organization similarity among SARS-CoV-2, Pangolin-CoV and BatCoV RaTG13.

Paper Impact

Reference Paper 1	
Title of the paper:	MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph
Authors:	Dinghua Li, Chi-Man Liu, Ruibang Luo, Tak-Wah Lam
Journal:	Bioinformatics, Volume 31, Issue 10, 15 May 2015, Pages 1674–1676
Publication year:	20 January 2015
Number of citations:	1721
Reason for choosing:	This reference provided the basis for the research. Through this literature, the authors have been able to rapidly reassemble the pangolin virus genome.

Table 1

Reference Paper 2	
Title of the paper:	Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan Pangolins
Authors:	Kangpeng Xiao, Junqiong Zhai, Yongyi Shen, Lihua Xiao, Wu Chen
Journal:	bioRxiv

Publication year:	
February 20, 2020	
Number of citations:	
174	
Reason for choosing:	
The literature provides results consistent with our paper as a comparative reference.	

Table 2

Reference Paper 3

Title of the paper:	
Viral Metagenomics Revealed Sendai Virus and Coronavirus Infection of Malayan Pangolins (Manis javanica)	
Authors:	
Ping Liu, Jin-Ping Chen	
Journal:	
Viruses, 2019, 11(11), 979	
Publication year:	
24 October 2019	
Number of citations:	
258	
Reason for choosing:	
This reference provides the original DNA sequence and offers research ideas.	

Table 3

Reference Paper 4	
Title of the paper:	
Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor	
Authors:	
Xing-Yi Ge, Jia-Lu Li and Xing-Lou Yang, Peter Daszak, Zheng-Li Shi	
Journal:	
Nature, volume 503, pages535–538 (2013)	
Publication year:	
30 October 2013	
Number of citations:	
1363	
Reason for choosing:	
The results of this reference are used as a comparative reference for our thesis.	

Table 4

Reference Paper 5	
Title of the paper:	
Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2	
Authors:	
Renhong Yan, Qiang Zhou	
Journal:	
Science, Vol. 367, Issue 6485, pp. 1444-1448	
Publication year:	
27 Mar 2020	
Number of citations:	

2699
Reason for choosing: This reference was used as a basis for comparison of the amino acid variation patterns of S1 proteins.

Table 5

The present and future of de novo whole-genome assembly

Jang-Il Sohn, Jin-Wu Nam

Briefings in bioinformatics, 2018, 19(1), 23–40.

ASA-Seminar

Presented by:

Yousef Alayoubi, Anastasiya Stepanenko

Abstract № 1

Sequencing has become an indispensable part of molecular biology, largely thanks to next generation sequencing technologies (NGS). The target genome is sliced into smaller fragments and amplified using PCR, then sequenced, generating billions of copies, or „reads".

Due to limitations of NGS technologies, the reads are mostly very short – in case of Illumina ~150 bp on average – and must be assembled to yield the target genome. Lacking a reference genome, the reads must be aligned by their overlapping regions to construct the genome, this method is called de novo assembly. To tackle this challenge, various algorithms and heuristics were developed, many of them are based on de Bruijn graphs.

In this paper, we review various assemblers and discuss their computational cost, assembly qualities and limitations. We show that to overcome the limitations of short-reads assembly, long reads only assembly or hybrid methods yield better results. While the short-read assemblers are very efficient, they struggle with repetitive structures and GC-biased regions, leading to misassembles. Moreover, constructing a de Bruijn graph is computationally expensive, thus demanding substantial memory access and long run time. Long read assemblers use overlap graphs (OLC) for assembly, which are generally less complex and more robust than de Bruijn graphs, however, the overlapping process for large genomes is highly demanding. Using long-read assembly methods proves to be more effective in solving the issues regarding repetitive structures and GC-biased regions, thus generating better, more contiguous results.

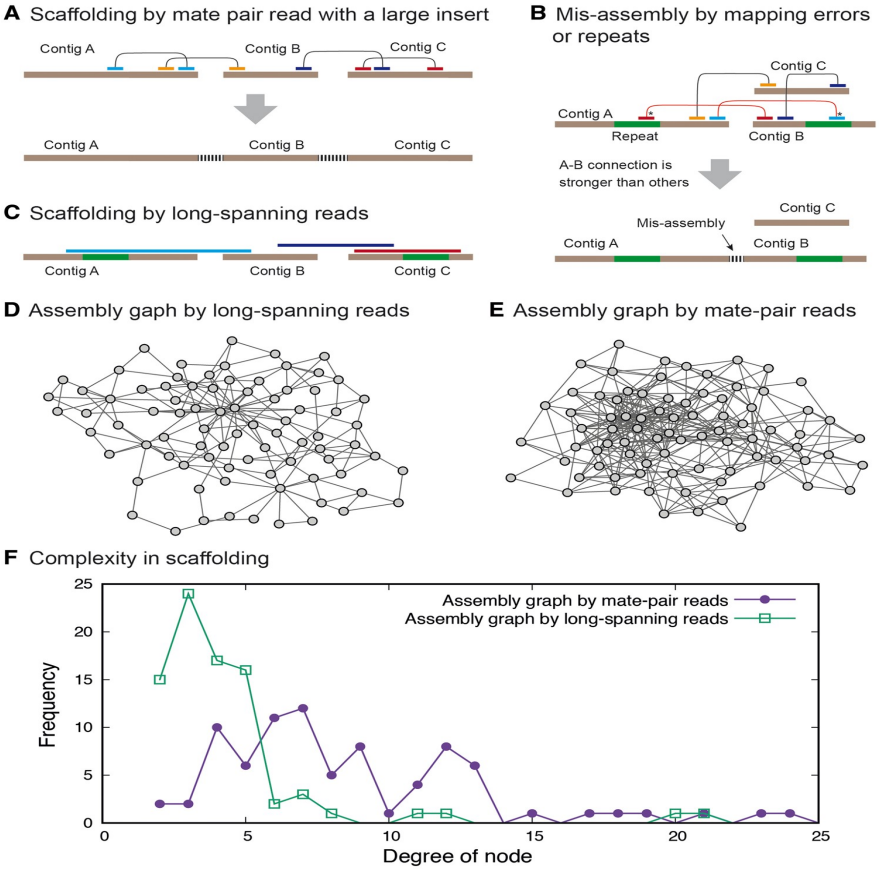
Our review provides an overview of variant assembly technologies and assemblers, which helps to choose an optimal approach for various input data and budgets. We show that long-read technologies provide better assemblies overall. However, the current long-read technologies have low-throughput and are much more expensive than short NGS technologies, therefore, sequencing and computational costs for large genomes are very high.

Abstract № 2

De novo assembly became a relevant theme lately because of the development of methods of next generation sequencing. Numerous algorithms of *de novo* assembly were developed based on two types of de Bruijn graph. Some challenges in *de novo* short read assembly still remain and numerous ideas and methods were developed to resolve them. Many challenges are occurring due to repetitive sequencies in big genomes, computation costs, sequencing errors, complexity of scaffolding and uneven read depth. These challenges can be resolved by changing the algorithm used to perform an assembly or changing the length of reads while sequencing. In this study, the main *de novo* assemblers are categorized due to type of de Bruijn graph and analyzed due to the computational costs, the performing *de novo* assembly using short and long reads and the resulting challenges are compared. The complexity of de Bruijn graph allows to find an optimal accuracy and computational costs of an assembly depending on a specific genome. Other challenges can be overcome by different methods of sequencing. The methods of *de novo* sequencing are becoming more and more different nowadays, that is why it is important to keep an overview of the main possible ways to perform *de novo* assembly. Long read assembly caused some changes in the *de novo* genome assembly methods, which should be considered while developing strategies of performing of *de novo* genome assemblies in future studies. In this review, the guidelines of an optimal *de novo* assembly are provided, regarding given data input and expected computational costs.

Summary

1. Research question: a review of current de novo assemblers and assembly methods, including short read, long read and hybrid methods, and the challenges with short read assemblers.
2. Relevant methodological approaches: reviewing and assessing papers.
3. Results:
   a) ALLPATHS-LG, the Eulerian de Bruijn graph assembler performed slightly better than other short-read de Bruijn assemblers, However, the Hamiltonian de Bruijn graph assemblers were relatively faster.
   b) The performance speed and the quality of assembly are also highly dependent on used hardware.
   c) The methods using short reads perform poorly for repetitive structures or GC-biased regions.
   d) Long-read-only methods performed better than hybrid methods.
   e) Long-read-only methods perform better for repetitive structures and GC-biased regions, thus generate better assemblies.
4. Conclusion:
   a) The computational cost and the accuracy of assemblers are dependent on the complexity of the de Bruijn graph.
   b) The assemblers based on the Hamiltonian de Bruijn graph have less execution time compared with Eulerian de Bruijn graph when the graph is simplified.
   c) Challenges resulting from complexity of the genome structure could be overcome using long SMS reads such as PacBio or Oxford Nanopore.
5. Figure: The graph shows that when using long reads for scaffolding, the de Bruijn graph is simplified compared to scaffolding with short reads, which reduces the computational complexity.



Paper Impact

Relevant references

1. Title: Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species
First author: Keith R Bradnam, **corresponding authors**: Keith R Bradnam, Ian F Korf, **last author**: Ian F Korf
Journal: Gigascience, **publishing year**: 2013, **number of citations**: 652
Relevance: Benchmarking of different assembly methods.
2. Title: Repetitive DNA and nextgeneration sequencing: computational challenges and solutions
First author: Todd J. Treangen, **corresponding author**: Steven L. Salzberg, **last author**: Steven L. Salzberg
Journal: Nature Reviews Genetics, **publishing year**: 2012, **number of citations**: 1447
Relevance: The challenges by repetitive regions can be solved with long reads sequencing.
3. Title: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler
First author: Ruibang Luo, **corresponding author**: Tak-Wah Lam, Jun Wang, **last author**: Jun Wang
Journal: Gigascience, **publishing year**: 2012, **number of citations**: 3509
Relevance: An example of Hamiltonian de Bruijn graph, which was often mentioned in the paper.
4. Title: High-quality draft assemblies of mammalian genomes from massively parallel sequence data
First author: Sante Gnerre, **corresponding author**: Eric S. Lander, David B. Jaffe, **last author**: David B. Jaffe
Journal: Proceedings of the National Academy of Sciences, **publishing year**: 2011, **number of citations**: 1676
Relevance: An example of Eulerian de Bruijn graph, which was often mentioned in the paper.
5. **Title:** Assembling large genomes with single-molecule sequencing and locality-sensitive hashing
First author: Konstantin Berlin, **corresponding author**: Sergey Koren, **last author**: Adam M Phillippy
Journal: Nature Biotechnology, **publishing year**: 2015, n**umber of citations**: 757
Relevance: An example of overcoming the challenges during the long reads only assembly.

Studies that cite the paper

Number of citations: 124
Citations per year 2018-2021: 124/3 ~ 41
Journal: Briefings in Bioinformatics
Journal Impact Score: 9.48
Ranked 2nd in Mathematical & Computational Biology

Olson, N. D., Treangen, T. J., Hill, C. M., Cepeda-Espinoza, V., Ghurye, J., Koren, S., & Pop, M. (2019). Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. Briefings in bioinformatics, 20(4), 1140-1150.
Jung, H., Winefield, C., Bombarely, A., Prentis, P., & Waterhouse, P. (2019). Tools and strategies for long-read sequencing and de novo assembly of plant genomes. Trends in plant science, 24(8), 700-724.
Voichek, Y., & Weigel, D. (2020). Identifying genetic variants underlying phenotypic variation in plants without complete genomes. Nature genetics, 52(5), 534-540.
Kono, N., & Arakawa, K. (2019). Nanopore sequencing: review of potential applications in functional genomics. Development, growth & differentiation, 61(5), 316-326.
Chebbi, M. A., Becking, T., Moumen, B., Giraud, I., Gilbert, C., Peccoud, J., & Cordaux, R. (2019). The genome of Armadillidium vulgare (Crustacea, Isopoda) provides insights into sex chromosome evolution in the context of cytoplasmic sex determination. Molecular Biology and Evolution, 36(4), 727-741.

Papers of corresponding author

1. Agarwal, V., Bell, G. W., Nam, J. W., & Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *elife*, *4*, e05005.
Number of citations: 3847, **Journal**: eLife, **Journal Impact Score**: 7.080
Position in the author list: 3
Reason: Most cited.
2. Kim, V. N., & Nam, J. W. (2006). Genomics of microRNA. *TRENDS in Genetics*, *22*(3), 165-173.
Number of citations: 1142, **Journal**: Trends in Genetics, **Journal Impact Score**: 11.333
Position in the author list: 2
Reason: The last author, many citations.
3. Nam, J. W., Shin, K. R., Han, J., Lee, Y., Kim, V. N., & Zhang, B. T. (2005). Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic acids research*, *33*(11), 3570-3581.
Number of citations: 291, **Journal**: Nucleic acids research, **Journal Impact Score**: 16.48
Position in the author list: 1
Reason: First author, corresponding author, microRNA is the main research field of Jin-Wu Nam, one of the first papers, published by this author.
4. Nam, J. W., Rissland, O. S., Koppstein, D., Abreu-Goodger, C., Jan, C. H., Agarwal, V., ... & Bartel, D. P. (2014). Global analyses of the effect of different cellular contexts on microRNA targeting. *Molecular cell*, *53*(6), 1031-1043.
Number of citations: 232, **Journal**: Molecular cell, **Journal Impact Score**: 15.584
Position in the author list: 1
Reason: First author, corresponding author, microRNA is the main research field of Jin-Wu Nam, microRNA is a relevant research field nowadays.
5. Nam, J. W., & Bartel, D. P. (2012). Long noncoding RNAs in C. elegans. *Genome research*, *22*(12), 2529-2540.
Number of citations: 220, **Journal**: Genome research, **Journal Impact Score**: 11.093
Position in the author list: 1
Reason: Long noncoding RNA is the second most researched field by Jin-Wu Nam, the topic long noncoding RNAs is becoming more and more relevant lately.

Standardized benchmarking in the quest for orthologs

Adrian M Altenhoff, Brigitte Boeckmann, Salvador Capella-Gutierrez, Daniel A Dalquen, Todd DeLuca, Kristoffer Forslund, Jaime Huerta-Cepas, Benjamin Linard, Cécile Pereira, Leszek P Pryszcz, Fabian Schreiber, Alan Sousa da Silva, Damian Szklarczyk, Clément-Marie Train, Peer Bork, Odile Lecompte, Christian von Mering, Ioannis Xenarios, Kimmen Sjölander, Lars Juhl Jensen, Maria J Martin, Matthieu Muffato, Quest for Orthologs consortium, Toni Gabaldón, Suzanna E Lewis, Paul D Thomas, Erik Sonnhammer & Christophe Dessimoz

Nature Methods volume 13, pages 425–430 (2016)

Presented by: Jonas Elpelt, Julian Rummel

Abstract

During the evolutionary process speciation events can occur, which lead to homologous DNA sequences across different species. As the true evolutionary history of these (typically functionally) related gene pairs (called orthologs) generally cannot be definitively determined, different approaches for orthology inference have been introduced. The importance of orthologs in modern computational genomic applications leads to the necessity of a systematic performance evaluation of the broad variety of orthology inference methods and resulted in the development of various benchmark tests. However, a standardized comparison of individual quality measures is made difficult due to application-dependent precision-recall trade-offs. Here we present a community-based web-service for automated orthology benchmarking. A comprehensive evaluation of about 70 million orthologous relationships and the inference of 233.000 phylogenetic trees was required to statistically analyze 15 common inference methods on 20 conventional benchmarks. To enable direct comparison a yearly-updated reference proteome dataset is provided by the 'Quest for Orthologs' consortium. The benchmark service allows the identification of the most effective inference algorithm for a given problem and gives users the opportunity to test new tools. Possible ways of improvement include for example the use of additional quality ratings and an extension of confidence scores or posterior probabilities. We suppose that it can serve as a new combined standard benchmark for prospective orthology inference methods and positively contribute to a comprehensive yet user-friendly validation process.

Abstract

Homologous genes are described as biologically similar in terms of common ancestry in the evolutionary history, originating from speciation events (orthologs), duplication events (paralogs) or horizontal gene transfer events (xenologs). Orthologs are critical for taxonomic and phylogenetic analysis since a particular pattern of genetic divergence is significant for the relatedness of organisms. Determining orthologous genes, and thus the proximity of two organisms, is usually accomplished by heuristic analysis or by phylogenetic methods. Today, dozens of tools/databases exist and therefore finding the right approach for a particular question might be a tough task. Here we demonstrate a standardized and facilitated orthology benchmarking as a web-based service, in order to provide a comprehensive evaluation of state-of-the-art orthology tools. Species discordance tests detected no obvious performance difference between tree-based and graph-based methods, refuting the assumption of an advantage due to species tree knowledge. In a benchmarking with reference gene trees, balanced precision-recall strategies performed best, while skewed precision-recall strategies might perform better on ambiguous phylogenies. Functional benchmarking shows that orthology inference methods have a clear trade-off between precision and recall. Our results demonstrate the functionality and robustness of our orthology benchmarking service and shows the benefits of systematic comparison of multiple benchmarks. This web-based service might be a first contact point comparing different orthology tools in a standardized way and offers many possibilities to do so. New benchmarks will be introduced in the future, by providing a way to automatically include new methods and publicly disseminate the result, in order to address small issues. Further improvements/extensions might be reconciled gene trees, hierarchical orthologous groups, confidence scores and posterior probabilities.

1. Research question(s) addressed

- standardized benchmarking in orthology inference

- identifying most effect methods for a given problem

- creation of a comprehensive assessment of state-of-the-art orthology tools

2. Relevant methodological approaches

Benchmark service:

- using the Quest for Orthologs (QfO) reference proteome dataset for creating a benchmark
- accepting OrthoXML or tab-delimited format
- Performance measured by precision (proportion of correct prediction) and recall (sensitivity)

Investigated tools:

- Tree-based methods:
    - Ensembl Compara, PANTHER 8.0, PhylomeDB
- Graph-based methods:
    - Best Reciprocal Hits, Reciprocal Smallest Distance (RSD), EggNOG, Hieranoid, InParanoid, OMA, OrthoInspector
- Combined:
    - MetaPhOrs

Generalized species tree discordance test:

- Evaluates the accuracy of orthologs in terms of the accuracy of the species tree that can be reconstructed from them
- Overcoming the limitation of species tree comb topology and a small number of taxa, through any tree topology and larger reference trees from the SwissTree initiative
- Avoidance of sampling orthologs among species separated by branches shorter than 10 million years

Reference gene trees:

- Uses evolutionary relationships of gene pairs derived from annotated high-quality gene trees
- Through combination of computational inference and expert curation (results of each step are individually inspected)
- Poor-quality sequences are excluded from the analysis
- Elucidates gene phylogenies with high statistical support and topological consistency

Functional benchmarks:

- Orthology in terms of functional similarity
- Using the approach of Schlicker et al. → similarity measured using Lin´s Metric

Online Methods:

- More detailed description of the different used tools and benchmarks

3. Relevant Results

- Benchmarking results for species discordance test, reference gene trees, functional trees:

Generalized species tree discordance test:

- Trade-off between precision (average discordance measured with Robinson-Foulds distance) and recall (number of trees that can be sampled; qualitative same results for number of inferred orthologs or other clades)
- Eukaryotes: highest precision, lowest recall = OMA; lowest precision, highest recall = PANTHER 8.0. Overall good performance of Orthoinspector, In Paranoid, PANTHER (LDO only). No obvious performance difference between tree-based and graph-based methods.
- Consistent results for vertebrates, a spanning tree across archaea, bacteria and eukaryotes and including short-branches (minor ranking differences, but overall trends are similar) → shows robustness of benchmark

Reference gene trees:

- Predictions made with SwissTree and TreeFam-A give quite similar results
- Balanced precision-recall strategies performed best (especially MetaPhOrs)
- Skewed precision-recall strategy (OMA or PANTHER) might perform better on ambiguous phylogenies (as they provide better error-correction)

Functional benchmarks:

- Trade-off between precision (average Schlicker semantic similarity of functional annotations associated with orthologs) and recall (number of ortholog relationships predicted)
- Consistent results for UniProt-GOA and ENZYME database (except for MetaPhOrs, whose missing taxa have a negative effect on recall)

4. Conclusion

- **Orthology Benchmark service enables systematic comparison of multiple benchmarks** (can be used for example for quality control, method development testing, etc.)

- There is an unavoidable precision-recall trade-off (final decision depends on application)

- Issue: Circularity (Overfitting) → new benchmarks will be introduced in the future

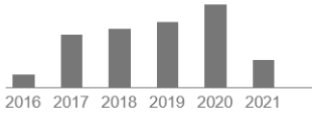
- Possible Improvements (extensions): reconciled gene trees, hierarchical orthologous groups, confidence scores, posterior probabilities

5. Main Figure

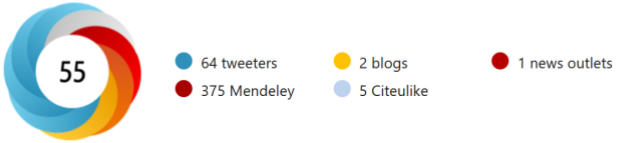
Figure 1: Schematic overview of the Orthology Benchmark service. Gives a visual insight about the workflow of the developed software, which is the main result of this paper. Thus, this figure describes the main work of Altenhoff et al. in a vivid way and briefly explains how their software works.

2016: 7; 2017: 28; 2018: 31; 2019: 35; 2020: 44; 2021: 15

2)

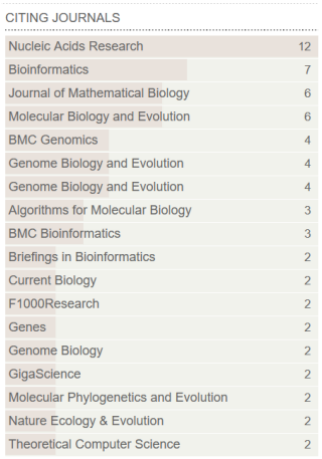


Online attention



This article is in the 95th percentile (ranked 11,665th) of the 271,711 tracked articles of a similar age in all journals and the 86th percentile (ranked 14th) of the 95 tracked articles of a similar age in *Nature Methods*

Citations:



CITING JOURNALS

Nucleic Acids Research	12
Bioinformatics	7
Journal of Mathematical Biology	6
Molecular Biology and Evolution	6
BMC Genomics	4
Genome Biology and Evolution	4
Genome Biology and Evolution	4
Algorithms for Molecular Biology	3
BMC Bioinformatics	3
Briefings in Bioinformatics	2
Current Biology	2
F1000Research	2
Genes	2
Genome Biology	2
GigaScience	2
Molecular Phylogenetics and Evolution	2
Nature Ecology & Evolution	2
Theoretical Computer Science	2

The paper is most frequently cited in studies in Nucleic Acids Research, whose 2019 impact factor was 11.5.

Top 5:

- UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*, *47*(D1), D506-D515. *(2805 Citations)*
- Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Von Mering, C., & Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular biology and evolution*, *34*(8), 2115-2122 *(873 Citations)*
- Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology*, *20*(1), 1-14. *(490 Citations)*
- Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., ... & Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research*, *47*(D1), D309-D314. *(469 Citations)*
- Fernández, R., Kallal, R. J., Dimitrov, D., Ballesteros, J. A., Arnedo, M. A., Giribet, G., & Hormiga, G. (2018). Phylogenomics, diversification dynamics, and comparative transcriptomics across the spider tree of life. *Current Biology*, *28*(9), 1489-1497. *(126 Citations)*

1) **Top 5:**

[6] Gabaldón, T. *et al.* Joining forces in the quest for orthologs. *Genome Biol.* **10**, 403 (2009).

[7] Dessimoz, C. *et al.* Toward community standards in the quest for orthologs. *Bioinformatics* **28**, 900–904 (2012).

[8] Sonnhammer, E.L.L. *et al.* Big data and other challenges in the quest for orthologs. *Bioinformatics* **30**, 2993–2998 (2014).

These above sources, from members of the QfO consortium, show the relevance, years of study of the problem, and previous efforts of some of the co-authors. The benchmarks are then built on this data.

[15] Altenhoff, A.M. & Dessimoz, C. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput. Biol.* **5**, e1000262 (2009).
This source also demonstrates the relevance of the topic and shows appropriate methods for evaluating the accuracy of orthologs, i.e., a benchmarking approach. The main authors are the same as the authors from our paper.

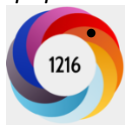
[43] Boeckmann, B. *et al.* Quest for Orthologs (QfO) entails Quest for Tree of Life (QfToL): in search of the gene stream. *Genome Biol. Evol.* **7**, 1988–1999 (2015).
This source is also from members of the Qfo consortium and some co-authors. This is the reference to the reference species used.

3) **Corresponding author:** Prof. Christophe Dessimoz

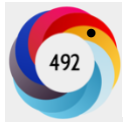
Swiss Institute of Bioinformatics, Biophore Building, Lausanne, Switzerland

Top 5:

The decision was based on journal impact factor, the number of citations and the author's position in the author list. The first two criteria are well represented in the Attention Score (Altmetric.com). The author's position was evaluated for each paper individually.



Nick Goldman, Paul Bertone, Siyuan Chen, Christophe Dessimoz, Emily M. LeProust, Botond Sipos, Ewan Birney: *Towards practical, high-capacity, low-maintenance information storage in synthesize DNA*: Nature, 2013, 494:7435, 77-80
Nature (42.8), 438 Citations, 4th, Score: 1216



Emma B Hodcroft et al.: *Want to track pandemic variants faster? the bioinformatics bottleneck*: Nature, 2021, 591:7848, 30-33.
Nature (42.8), 8 Citations, last author, Score: 492

Christophe Dessimoz and Nives Škunca, Editors: *The Gene Ontolo Handbook*: Methods in Molecular Biology, 2017, Springer (New York), Vol. 1446
Methods in molecular biology (1.17), 45 Citations, first of two authors, Score: 216

Liam P Shaw, Alethea D Wang, David Dylus, Magda Meier, Grega Pogacnik, Christophe Dessimoz, Francois Balloux: *The phylogenet range of bacterial and viral pathogens of vertebrates*: Molecular Ecology, 29(17):3361-3379
Molecular Ecology (5.2), 17 Citations, pre-last author, Score: 142

Philippe et al.: *Mitigating Anticipated Effects of Systematic Errors Supports Sister-Group Relationship between Xenacoelomorpha ar Ambulacraria*: Current Biology, 2019, 29:1–9
Current Biology (9.6), 50 Citations, 4th, Score: 113

Measuring the distance between multiple sequence alignments

Benjamin P. Blackburne and Simon Whelan

Bioinformatics, Volume 28, Issue 4, 15 February 2012, Pages 495–502
https://doi.org/10.1093/bioinformatics/btr701

Presented by Melanie Mößer and Dominik Stroh

Paper abstracts

Abstract Melanie Mößer

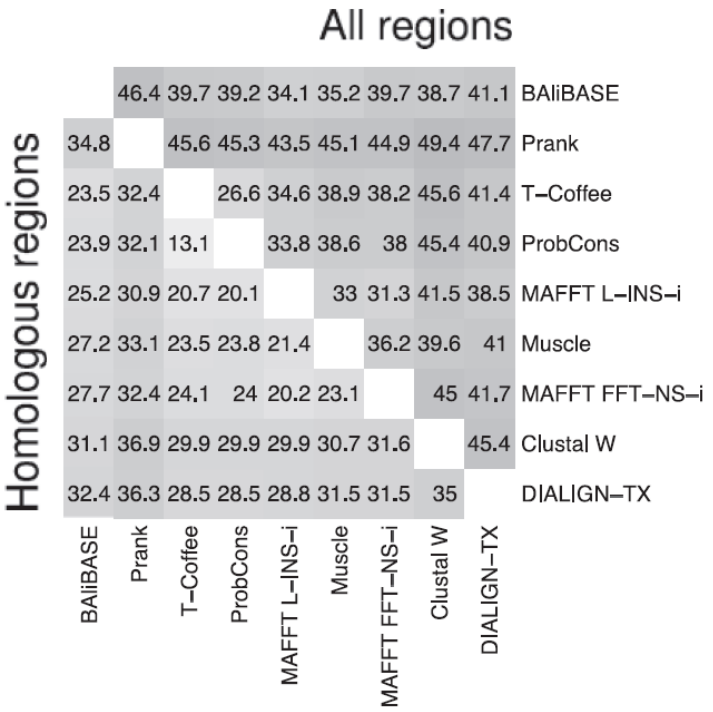
Multiple sequence alignment (MSA) is a common method to investigate phylogenetic proximity of different protein, DNA or RNA sequences. For a given set of sequences, various alignment methods use different heuristics and produce different alignments. An assessment of differences of MSAs could help to compare alignment methods and quantify the evolutionary distance of multiple sequences. The task of quantitative comparison of distances, however, has not been fully solved yet as the most common measurements, such as the Sum-of-Pairs or Total-Column scores, are no true metrics. Here, we present MetAl, a free implementation of four true metrics in Haskell, to accurately compute differences of MSAs. The metrics include positional and evolutionary information of the occurrence of gaps (insertion/deletion events) and are tested on real-world and synthetic data. In contrast to previous assumptions, the metrics predict large differences between commonly used alignment software and between them and a reference alignment. The biological significance of many downstream analyses, such as protein structure prediction or studies on phylogenetic inference, highly depend on MSA accuracy. Our metrics provide a valuable tool to evaluate the accuracy of MSAs and prevent misalignments. The metrics can help to investigate the origin of the distances between the heuristics of different aligners. Our study sets the foundation for further work on discovering biases that common alignment methods and heuristics may share and might facilitate the development of more accurate aligners.

Abstract Dominik Stroh

Multiple sequence alignments (MSAs) are commonly used for phylogenetic tree reconstruction, structure prediction and functional annotation. Multiple sequence aligners need to make assumptions on when insertions and deletions in the phylogenetic history of the genomes occurred. Differences in these assumptions lead to different gap placements for each alignment algorithm and unique solutions for complex MSAs. Correctly quantifying the difference between these alignments remains challenging, as existing scores are not suitable for distance measurements in MSAs, because they do not fulfil at least one of the criteria of true metrics being the identity of indiscernibles, symmetry or the triangle inequality. Here we show four true metrics which can accurately measure the dissimilarity between MSAs, based on incorporation of gap information. It was previously thought, the most common sequence aligners produce relatively equal MSAs. Through testing these aligners on real-world as well as in silico datasets and quantifying the pairwise dissimilarity between algorithms with our metrics, we reveal that even the most established aligners produce largely different alignments. This benchmarking provides a reliable quantification of dissimilarity between algorithmic calculated MSAs and a true or manually curated alignment facilitating the research of more accurate algorithms and helping to limit effects in the downstream analysis. Through our Haskell implementation of the metrics, we anticipate the testing of established sequence aligners on further real-world data, revealing biases of the sequence aligners in special occasions.

Paper summary

1. Research Questions
    a. How similar/different are different multiple sequence alignment (MSA) approaches and heuristics?
    b. How accurate is an MSA in comparison to a reference alignment?
    c. Which metrics are suitable for the computation of the distance of different alignments?
    d. What are the effects of large distances between MSAs on downstream analyses?
2. Methodical Approaches
    a. Definition of four new metrics for MSAs with different approaches of gap incorporation
        i. Symmetrized Sum-of-Pairs Score (SSP): based on the Jaccard distance, ignores gaps
        ii. Seq: based on the Hamming distance, records in which sequence a gap occurs
        iii. Pos: based on the Hamming distance, records in which sequence and at which position a gap occurs
        iv. Evol: based on the Hamming distance, records in which sequence, at which position and where in the phylogenetic tree a gap occurs
    b. Definition of a unique representation of MSAs
        i. The order of non-overlapping characters with gaps is arbitrary and can be adapted to a specific format for comparability
    c. Application of different MSA aligners and common software
        i. Progressive, consistency and phylogenetic tree-based aligners
    d. Testing of performance of aligners on real-world data
    e. Construction of a phylogenetic tree and tree-related synthetic test data
    f. Implementation of the metrics in Haskell
3. Results
    a. Proof that existing SP, total column and overlap scores are not true metrics
    b. Many alignment methods are similar (but not equal) to a reference alignment, but still substantially differ from each other
    c. Differences between aligners are larger for synthetic than for real-world data
    d. The similarity between d_{pos} and d_{evol} is much higher than between d_{pos} and d_{seq}
    e. The differences of MSAs can have effects on downstream analysis, such as protein structure prediction or studies of non-coding DNA
4. Conclusion
    a. Existing scores are not sufficient to compare different MSA aligners
    b. The new four metrics enable the comparison between different MSAs and different aligners
    c. The d_{evol} metric includes evolutionary information
    d. Existing MSA aligners differ in composition and accuracy
5. Key Figure
    a. Comparison of different MSAs via the distance metric d_{evol}. Shown are dissimilarity values for each MSA method in percentage.
    b. All alignment methods produce similarly accurate MSAs compared to the manually curated reference alignment of BAliBASE of real-world data.
    c. The different methods highly vary. In average, Prank has the largest differences to all other methods and the reference alignment.



Paper impact

1.	Author	Last author	Title	Journal	Year	Citations	Relevance
	Cédric Notredame*		Recent Evolutions of Multiple Sequence Alignment Algorithms	PLoS Comput Biol 3(8): e123	2007	158	The paper reviews different alignment software and discusses the differences of the aligners in terms of methodology and heuristics. This aspect of MSAs is important for the impact of our paper, as it highlights the broad variety of software, which produce different results and implies the importance of a real measurement of dissimilarity.
	Tanya Golubchik	Lars S. Jermiin*	Mind the Gaps: Evidence of Bias in Estimates of Multiple Sequence Alignments	Molecular Biology and Evolution, Volume 24, Issue 11, Pages 2433–2442	2007	75	The paper describes the basic methods of MSA and its impact on bioinformatic analyses. With this fundamental topic, the paper provides a methodological basis for our paper.
	Robert Edgar*	Serafim Batzoglou	Multiple sequence alignment.	Current opinion in structural biology vol. 16,3: 368-73.	2006	232	The paper demonstrates the importance of the incorporation of gaps in an MSA to avoid biases. These findings illustrate the need of a measurement between MSAs that includes gaps just as the authors in our paper have done.
	Julie D. Thompson*	Olivier Poch	BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark	Proteins, 61: 127-136	2005	252	Benchmarking is an important step to test the performance of software. In this case, it helped to test the distances computed by the metrics and provided testable real-world data.
	William Fletcher	Ziheng Yang*	INDELible: A Flexible Simulator of Biological Sequence Evolution	Molecular Biology and Evolution, Volume 26, Issue 8, Pages 1879–1888	2009	295	Benchmarking is an important step to test the performance of software. In this case, it helped to test the distances computed by the metrics and provided testable synthetic data.

2.	Citations	Citation/Year	Journal	Impact factor	Title
	16,311		Molecular Biology and Evolution	11.062	MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability
	0		Methods in Molecular Biology	1.17	Phylogeny-Aware Alignment with PRANK and PAGAN.
	141		Protein Science	3.876	The interface of protein structure, protein biophysics, and molecular evolution
	75		Methods in Ecology and Evolution	6.36	phyloGenerator: an automated phylogeny generation tool for ecologists
	24		Methods in Molecular Biology	1.17	Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment.
	42	4.2 (19% in last two years)	Bioinformatics	5.61	Measuring the distance between multiple sequence alignments

3.	Title	Journal	Impact factor	Citations	Author position	Relevance
	Initial sequencing and comparative analysis of the mouse genome	Nature	42.778	4,734	Consortium	Publication with the most citations, work in an international collaboration.
	Physicochemical amino acid properties better describe substitution rates in large populations	Molecular Biology and Evolution	11.062	13	2/2	Recent publication (2019). Proposes a new model for amino acid substitution rates.
	A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach	Molecular Biology and Evolution	11.062	1,988	1/2	Combines two existing models of amino acid substitution and extracts the advantages of both approaches.
	PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences	Bioinformatics	5.61	40	1/3	New method that uses explicit probabilistic model which investigates non-homologous characters.
	Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution	Genome Research	11.093	246	13/18	Benchmarking of six measures of evolutionary change.

BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences

Minoru Kanehisa, Yoko Sato, Kanae Morishima

February 2016
Journal of Molecular Biology - Elsevier
Volume 428, Issue 4, Pages 726-731

presented by
Bo Zheng, Franziska Hicking

1 Abstracts

An important step in the sequencing of unknown genomes is the high-throughput annotation of the biological functions of all genes in the genome. The tool presented in this paper is the KEGG-based annotation web service. In the KEGG database, proteins with similar functions are grouped together in the same group and then labelled with a KO number. Through similarity matching, KO numbers can be annotated for protein sequences of unknown function. At the same time KEGG already contains nearly 4000 complete genomes, constituting a non-redundant KEGG GENE database to be used as a comparative database. The rapid and accurate assignment of KO numbers to genes was then the main purpose of developing this tool.Here we present two KEGG-based web services, BlastKOALA and GhostKOALA.These two algorithms perform KO assignments to characterise the function of individual genes. The GHOSTX alignment is similar to BLAST alignment in that it detects remote homologues and is approximately 100 times faster than BLAST. Both algorithms can download and annotate the result files for further KEGG Mapper analysis, including comparative pathway analysis using multiple BlastKOALA results.With the KO number, it is possible to reconstruct the KEGG pathways in the KEGG database as well as the BRITE hierarchy and KEGG modules. The data in the results file can also be used as a chemical reaction network analysis and phylogenetic analysis of small molecules. In addition, the comparison with human orthologues allows the analysis of diseases and drugs.

The high number of sequenced genomic material puts forth the issue of how to interpret these findings and assign biological functions to sequenced genomes and metagenomes. The KEGG Orthology (KO) database stores biological functions which are identifiable through assigned Knumbers. The associated ortholog groups, i.e. genes associated with the respective function and the molecular networks in which they are represented, are included. An open approach for genome annotation is to assign functionality based on sequence homology in a fast and easily accessible manner.Here we present two automatic genome annotation servers, BlastKOALA and GhostKOALA, which perform K-number assignments to sequenced genomes and metagenomes. BlastKOALA utilizes the BLASTP algorithm to search a non-redundant database of pangenomes which is created from the KEGG GENES database. GhostKOALA uses the faster GHOSTX algorithm to search the database which is supplemented with Cd-hit clusters. Both servers subsequently assign K-numbers using the internal KOALA algorithm. BlastKOALA can perform searches at a species, genus or family level which makes it suited for genome annotation, while GhostKOALA is suited for annotation of metagenomes and assigns taxonomic categories to the query genes. The presented servers are freely available on the KEGG website to characterize gene functions. Additionally, they provide information about associated molecular networks such as KEGG pathways, BRITE hierarchies and KEGG modules which can give important insights on possible high-level functionality of organisms or ecosystems. All result files can be downloaded and allow the possibility to perform additional KEGG Mapper analysis.

2 Summary

2.1 Bearbeitete Forschungsfrage

- How to assign functions to sequenced genomes and metagenomes?

2.2 Relevante Methodische Ansätze

- Development of automatic annotation servers in KEGG.
- Assign K numbers from KEGG Orthology to sequencing results to characterize the function of individual genes.
- Create non-redundant dataset of pangenome sequences to improve
- Make use of the internal KOALA algorithms BLASTP and GHOSTX to search the non-redundant GENES database.

in BlastKOALA after a BLAST search of a fully redundant dataset of whole genome sequences at the species, genus or family level containing the KO content of each taxonomic category from the KEGG GENES database.

2.3 Relevante Ergebnisse

- The tools BlastKOALA and GhostKOALA are available web services on KEGG to perform genome annotation.
- Both tools provide the option to reconstruct the KEGG pathway, BRITE hierarchy and KEGG modules through the KEGG Mapper links.
- GhostKOALA is faster and more suitable for annotating metagenome sequences, while BlastKOALA is more suitable for annotating fully sequenced genomes and has a higher accuracy.
- For both algorithms, the result files can be downloaded and used for various types of analyses.

2.4 Schlussfolgerung

- BlastKOALA and GhostKOALA are available web services that make it possible for everyone to obtain annotation results.
- These tools can help make sense of large amounts of sequencing data in very short time.

2.5 Schlüsselabbildung

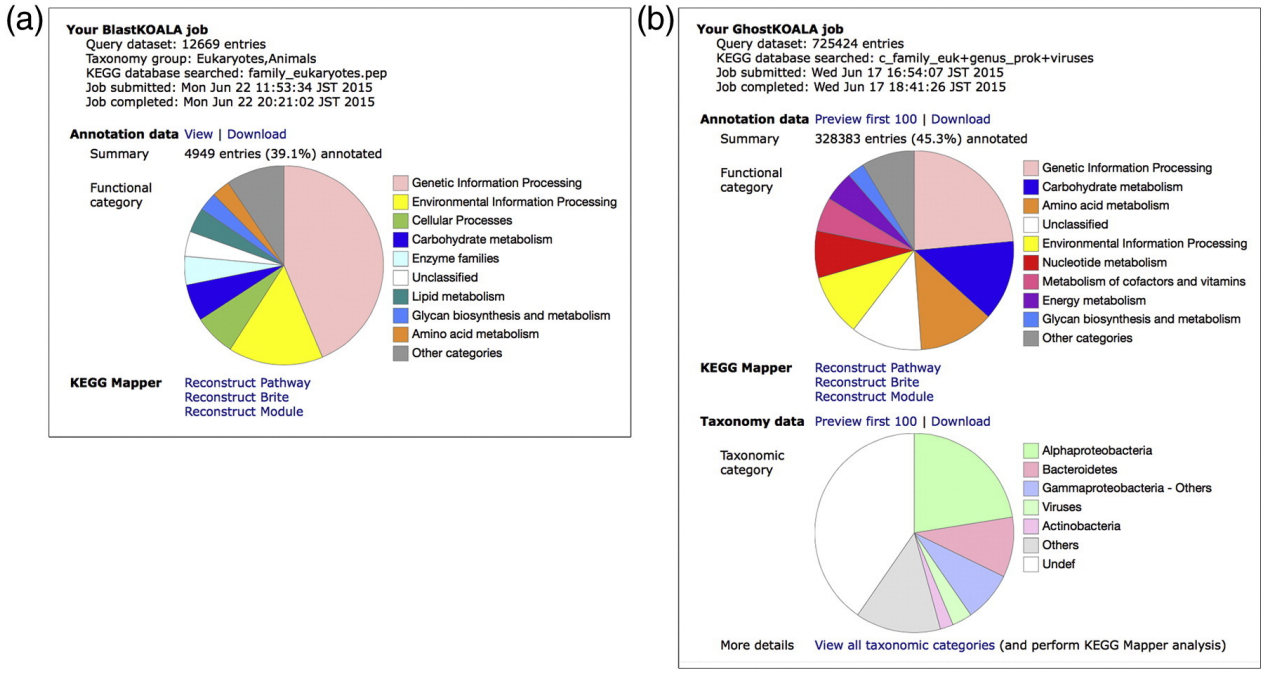


Figure 1: Results of BlastKOALA and GhostKOALA

There is only 1 figure in this paper showing the results of BlastKOALA and GhostKOALA. Taking GhostKOALA in figure (b) as an example, there are three parts in the figure.
- Annotation data: KEGG annotation classification results.
- KEGG Mapper: KEGG comparison pathway, hierarchy, module results.
- Taxonomy data: species classification results.

3 Impact

3.1 Important references

- 1. M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe
  2. Data, information, knowledge and principle: Back to metabolism in KEGG
  3. Nucleic Acids Research, 42, pp. D199-D205, year 2014, 2755 citations
  4. This publication is in our opinion the most influential one out of all citations. Here, the current state of the database is described, which above all serves as the essential foundation for this work.
- 1. Weizhong Li, Adam Godzik
  2. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences
  3. Bioinformatics, Volume 22, Issue 13, pp. 1658-1659, year 2006, 6532 citations
  4. We included this citation because the creation of Cd-hit clusters is an essential step in the annotation of metagenomes presented in this work. The algorithm presented in the cited paper is applied to genes which have no assigned K numbers.
- 1. Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, David J. Lipman
  2. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs
  3. Nucleic Acids Research, Volume 25, Issue 17, pp. 3389-3402, year 1997, 78823 citations
  4. This citation is more relevant than other citations because the improved BLAST program presented here serves as an essential basis for the BlastKOALA server. The algorithm is used during the annotation of genomes by BlastKOALA it relies on it heavily as no results could be achieved without it. Additionally, the number of citations for this paper is extremely high which shows its general relevance very well.
- 1. S. Suzuki, M. Kakuta, T. Ishida, Y. Akiyama
  2. GHOSTX: An improved sequence homology search algorithm using a query suffix array and a database suffix array
  3. PLoS One, Volume 9, Issue 8, p. e103833, year 2014, 61 citations
  4. We included this citation, despite its comparatively low number of citations, for similar reasons as the previous mentioned citations. The GHOSTx algorithm is used for the annotation of metagenomes by the GhostKOALA web server.
- 1. W.R. Pearson
  2. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms
  3. Genomics, Volume 11, pp. 635-650, year 1991, 708 citations
  4. The KEGG Ortholog Clusters which are presented in this publication serves as a base of information for the KO assignment. The SSDB database which is linked to the GENES databank is basically based on the information presented in this paper. And the SSDB database is very important for the K number assignment of the annotation servers BlastKOALA and GhostKOALA.

3.2 Important citing papers

- The total number of citations is 1313. The number of citing articles is 633 (ncbi) and yearly citations range from 21 in 2016 to 205 in 2020 and increases each year. In 2021 there have been 88 citing articles so far according to the ncbi.

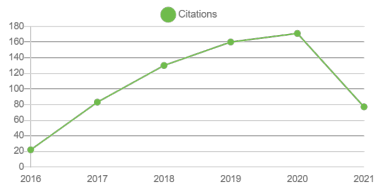


Figure 2: Amount of citations per year (Europe PMC)

- **most important citing articles:**
  1. Kanehisa, Minoru, et al. "KEGG: new perspectives on genomes, pathways, diseases and drugs." Nucleic acids research 45.D1 (2017): D353-D361.
  Impact factor: 11.501

2. Kanehisa, Minoru, et al. "New approach for understanding genome variations in KEGG." Nucleic acids research 47.D1 (2019): D590-D595.
   Impact factor: 11.501
3. Almeida, Alexandre, et al. "A new genomic blueprint of the human gut microbiota." Nature 568.7753 (2019): 499-504.
   Impact factor: 42.778
4. Shen, Xing-Xing, et al. "Tempo and mode of genome evolution in the budding yeast subphylum." Cell 175.6 (2018): 1533-1545.
   Impact factor: 38.637
5. Gu, Yanyun, et al. "Analyses of gut microbiota and plasma bile acids enable stratification of patients for antidiabetic treatment." Nature communications 8.1 (2017): 1-12.
   Impact factor: 12.121

3.3 Important work by Minoru Kanehisa

1. Kanehisa, Minoru, and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes." Nucleic acids research 28.1 (2000): 27-30.
   Impact factor: 11.501
   Number of citations: 19107
   Position in author list: 1
   This work by Minoru Kanehisa has by far the highest number of citations and Kanehisa is number one in the list of authors. The impact factor is 11.501, as are most of his works with a higher number of citations.
2. Kanehisa, Minoru, et al. "KEGG for linking genomes to life and the environment." Nucleic acids research 36.suppl_1 (2007): D480-D484. Impact factor: 11.501
   Number of citations: 4686
   Position in author list: 1
   Just like the previous entry, Kanehisa is first in the list of authors and the number of citations is the second highest out of all his publications.
3. Kanehisa, Minoru, et al. "KEGG for integration and interpretation of large-scale molecular data sets." Nucleic acids research 40.D1 (2012): D109-D114.
   Impact factor: 11.501
   Number of citations: 4133
   Position in author list: 1
   This publication is very similar to previous one concerning the ranking criteria of this list. Solely the number of citations is a little bit lower.
4. Kanehisa, Minoru, et al. "The KEGG resource for deciphering the genome." Nucleic acids research 32.suppl_1 (2004): D277-D280.
   Impact factor: 11.501
   Number of citations: 4064
   Position in author list: 1
   This work by Kanehisa is again very similar criteria wise to the previous one. The number of citiations is just slightly lower.
5. Kuroda, Makoto, et al. "Whole genome sequencing of meticillin-resistant Staphylococcus aureus." The Lancet 357.9264 (2001): 1225-1240.
   Impact factor: 60.392
   Number of citations: 2274
   Position in author list: 28
   This publication is the only one which is fairly different to the previous entries. The number of citations for this entry is not as high as multiple of Kanehisas other publications and he is only at the 28th position of the author list. The reason that this publication is included, is due to the very high impact factor of the journal.

   We found a few other publications in journals with higher impact factors than 11.501 which the first four items on this list share. However, the number of citations was a lot lower, and for most of them Kanehisa was not the first name on the list of authors.

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

Nature Biotechnology, VOL 37, May 2019, Pages: 540-546

Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin and Pavel A. Pevzner

Johannes Hausmann, Luis Kress
17.06.2021

Abstract 1: Luis Kress

With the improvement of sequencing longer DNA fragments, the demand for better algorithms reconstructing the genome out of these fragments is growing. Current long read sequencing techniques still have high error rates which makes it more difficult to align these reads. Up to date long read assemblers are still not able to resolve all repeating regions correctly. Especially segmental duplications, long and highly homologous sequences resulted from duplications, are still problematic to resolve correctly. The aim is to generate an algorithm that is able to resolve repeating regions in the genome and beeing able to assemble long reads correctly. Here we present Flye a *de novo* assembler for long-error prone reads, by creating a precise repeat graph, built in a new manner using so called disjointigs. Flye could achive two times better contiguity for the assembly of a human Oxford Nanopore test dataset in combination with short read Illumina data in contrast to the assembler Canu. In the created repeat graph many segmental duplications are represented from which the simple ones are already resolved by the algorithm. Our results represent a new way of constructing repeat graphs which leads to high quality assemblies and precise repeat graphs out of long error-prone reads and even increase the assembly speed. We created a tool which is able to represent complex repeating regions in a repeat graph. Further research is needed to find algorithms resolving these regions correctly. Better assembly results than state of the art assemblers can still only be generated by adding high quality short read data, which shows the demand for long read sequencing method optimization.

Abstract 2: Johannes Hausmann

With the emergence and the spreading adoption of long-read technologies, the assembly of genomes has improved, which may also become the "gold-standard" for *de novo* assemblies. A major difficulty in assembly is the resolution of repeat-rich genomic regions and the reconstruction of complex segmental duplications. For short-reads, assemblers use de Bruijn graphs to build a consensus sequence from reads and resolve repeats using bridging read pairs. Assemblers for long-reads largely rely on an overlap-layout-consensus approach and use different heuristics to resolve repeats. While genomic repeats can be better resolved using long reads, assembly with them is still challenging and not straightforward due to their error-prone nature. Here we present Flye, a new *de novo* assembly pipeline for long error-prone reads, that provides a solution to correctly resolve repetitve regions and to reconstruct segmental duplications in the assembly using a repeat graph built from disjointigs. We demonstrate that Flye generates high quality assemblies from nanopore and SMRT reads. Compared to state-of-the-art assemblers, e.g. Canu, Flye was able to generate partially more accurate or better contiguos assemblies, as shown by the metrics NGA50 and reference percentage identity. When combined with short-reads, Flye generated a more contiguous and accurate assembly in a human test dataset. Our assembler shows that a genome can be accurately assembled by repeat characterization using repeat graphs. This information can also help in improving existing assemblies. With the presented algorithm, a possibility is provided to improve the *de novo* assembly of a genome. Short segmental duplications are already resolved by Flye, while long and complex ones need further adaptations of the algorithm.

Summary

1. Adressed Research Questions

- create a de novo assembler for long error prone reads
- find a solution to resolve repetetive regions correctly which is still not straight forward with long reads

2. Methods

- Flye (the assembly algorithm described in the article) is using repeat graphs, built with approximate sequence matches -> tolerating higher noise of single molecule sequencing reads
- using disjointigs (concatenate overlapping reads without attempting to resolve repeats and be aware that contigs might contain misassemblies) to create a repeat graph

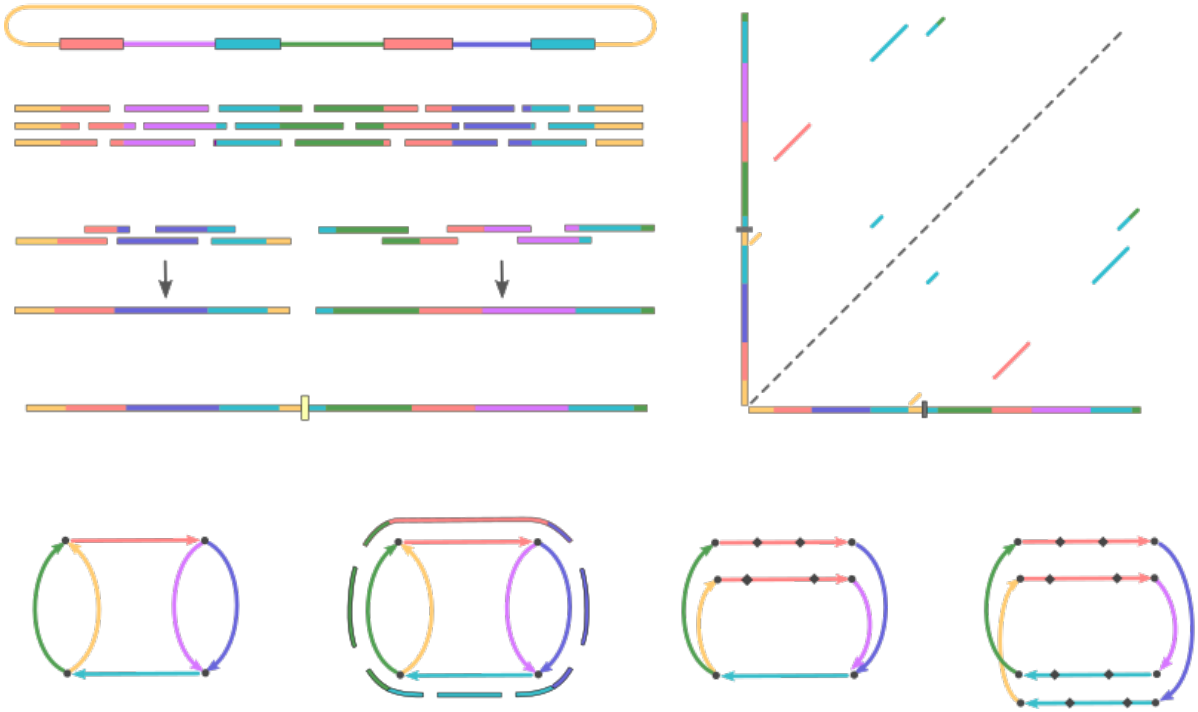
3. Relevant Results

- Flye generates in combination with short read NGS data higher quality assemblies for higher eucaryotes than canu and other assemblers (comparing NGA50 and reference percentage identity)
- in the generated repeat graph, many long segmental duplications (SDs) are represented -> the simple ones are already resolved by the algorithm

4. Conclusion

- with the repeat graph the possibility to represent SDs is given and simple SDs are already resolved by the algorithm, but further algorithms are needed to resolve complex mosaic structures like chromosomal regions arround the centromer
- while Flye generates comparatively good results for higher eucaryotes, short read NGS data is still needed for indel correction

5. Key Figure



In the figure (Figure 1 in the article, corrected version by the author) the main aspects of the flye pipeline are displayed. We chose this figure since it shows all essential steps of the algorithm from disjointig creation to repeat resolving. The other figures in the article only display single aspects of the algorithm or results. Due to the fact, that a new algorithm is presented in the article we wanted to emphasize on the pipeline rather than the results.

Assembly of long error-prone reads using repeat graphs (paper impact)

Johannes Hausmann[1], Luis Kress[1]

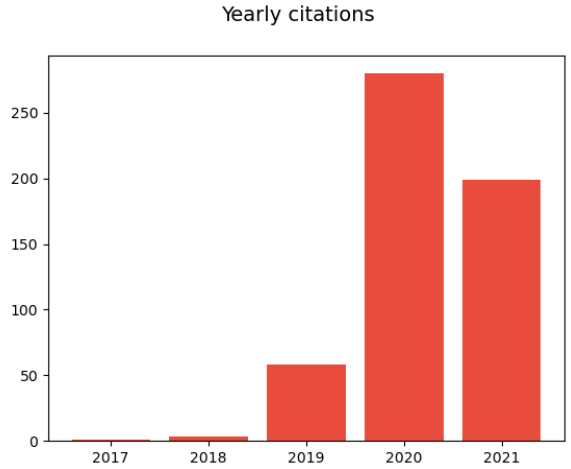
[1] Goethe Universität Frankfurt

Key References

- Assembly of long error-prone reads using de Bruijn graphs.
  - Authors: Yu Lin (First), Pavel A. Pevzner (Last and corresponding)
  - Journal: PNAS (113 (52), 8396-8405)
  - Year: 2016
  - Citations: 126 (Google scholar)
  - Reason: The presented algorithm Flye is an improvement of the tool presented in this article.
- De novo repeat classification and fragment assembly.
  - Authors: Paul A. Pevzner (First), Glenn Tesler (Last and corresponding)
  - Journal: CSH Press, Genome Research (volume 14, 1786-1796)
  - Year: 2004
  - Citations: 294 (Google scholar)
  - Reason: Basics on the repeat graph and the repeat classification problem.
- Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution.
  - Authors: Zhaoshi Jiang (First), Evan E Eichler (Last and corresponding)
  - Journal: Nature Genetics (volume 39, 1361-1368)
  - Year: 2007
  - Citations: 192 (Google scholar)
  - Reason: The information, that repeat graphs can be used to represent mosaic structures.
- What is the difference between the breakpoint graph and the de Bruijn graph?
  - Authors: Yu Lin (First), Pavel A Pevzner (Last and corresponding)
  - Journal: BMC Genomics volume 15, Suppl 6 (S6)
  - Year: 2014
  - Citations: 23 (Google scholar)
  - Reason: Groundwork publication from Pevzner. Information on the graph structure used to implement Flye. Assembly graphs created from the repeat graph are special cases of breakpoint graphs.
- Haplotype and Repeat Separation in Long Reads.
  - Authors: Tischler-Höhle G.
  - Journal: Computational Intelligence Methods for Bioinformatics and Biostatistics, CIBB 2017, 103-114
  - Year: 2017
  - Citations: 3 (Google scholar)
  - Reason: The publication describes methods for repeat & haplotype seperation with long reads. Resolving unbriged and highly similar repeats is realted to the challenge of overlap-filtering repeat resolution described in this publication.

Citations of our article

- Citations: 564 (Google scholar); 264 (PubMed); 273 (Web of Science); 356 (Cross Ref)
- Yearly
  - PubMed: 2019: 24; 2020: 141; 2021: 108
  - Altmetric: 2017: 1; 2018: 3; 2019: 58; 2020: 280; 2021: 199



Citations per year (Altmetric)

- Articles which cite the given paper:
  - Fast and accurate long-read assembly with wtdbg2 (Nature Methods) (cited by: 313, scholar), (impact score: 10.93, 2020)
  - De novo assembly of haplotype-resolved genomes with trio binning (Nature Biotechnology) (cited by: 151, scholar), (impact score: 10.71, 2018)
  - Telomere-to-telomere assembly of a complete human X chromosome (Nature) (cited by: 180, scholar) (impact score: 42.778, 2019)
  - Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome (Science) (cited by: 160, scholar), (impact score: 12.84, 2019)
  - Opportunities and challenges in long-read sequencing data analysis (Genome Biology) (cited by: 180, scholar) (impact score: 11.71, 2020)

Corresponding author Pevzner

- SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing
  - Reason: last author, 12583 citations, widely used assembly tool, (impact score 2.09; 2013)
- Initial sequencing and comparative analysis of the mouse genome
  - Reason: first sequencing and assembly of the mouse genome, 7438 citations, (impact score 36; 2010)
- De novo identification of repeat families in large genomes
  - Reason: corresponding author, groundwork for the given paper (Assembly of long-error prone reads) /groundwork for assembly of repeat rich genomes, 1211 citations, (impact score 5.15; 2013)
- An Eulerian path approach to DNA fragment assembly
  - Reason: first author, groundwork for further short read assembly tools, 1571 citations, (impact score 9.81; 2013)
- De Novo Peptide Sequencing via Tandem Mass Spectrometry
  - Reason: last author, development of de novo peptide sequencing which is still used today, 747 citations, (impact score 2.09; 2013)

Rayan Chikhi and Paul Medvedev:

Informed and automated k-mer size selection for genome assembly

1. *Scientific questions:*

- How can a good parameter k for de Bruijn be found in feasible time?

- How can the choice of the parameter k be automated and put in a automated assembly pipeline?

- How to construct an approximate abundance histogram in feasible time

- How to deal with heterozygosity and repetitivness of genomes

- What are the limitations of this technique of chosing k and what limitations do other techniques have?

2. *Relevant methods:*

- Generating an approximate histogram using an algorithm with hashing

- Estimate the number of genomic k-mers using the abundance histogram

- Define a haploid model with the free parameters (μ_1, σ_1^2, s, α, p_e).

- Extend the haploid model to a diploid model

- Estimate the parameters with the maximum likelihood estimation (BFGS algortithm).

- Extrapolating the optimal value k.

3. *Relevant results*

- Approximation of the histograms is faster than with common methods.

- Good results for the choice of k in benchmarks and comparison with common methods

4. *Conclusion*

- A way was found to chose for the most cases, a well working value k in feasible time

- There are limits of this approach e.g. data of single cell sequencing is not suitable for this method.

5. *Key figure:*

- We chose this table because it shows the immense reduction of time needed to estimate value k.

- The time and energy needed to compute those approximations is a bottleneck in building and optimizing assemblies and therefore an important factor to improve.

Table 2. Resource utilization of KmerGenie compared with a *k*-mer counting-based approach (DSK)

Organism	CPU time		Memory usage of KmerGenie (GB)
	DSK	KmerGenie	
S.aureus	2 min	11 s	0.1
chr14	48 min	7 min	0.1
B.impatiens	7.5 h	1.2 h	0.4

Note: We executed KmerGenie and DSK for a single value of k (81) using one thread. KmerGenie was executed with a sampling frequency of $\epsilon = 1000$. DSK used 5 GB of memory.

Informed and automated k-mer size selection for genome assembly

Chikhi, R., Medvedev, P. (2013)

Abstract

In genome assembly, many of the most common tools rely on constructing so called De Bruijn graphs. The construction of these graphs is heavily influenced by the value of parameter *k*. De Bruijn graphs are directed graphs, constructed by chopping the reads of a genome study into substrings of length *k*, and using these *k-mers* as edges between its prefix and suffix. Then, all nodes with identical lables are combined into one. Finding a Eulerian cycle in this graph, meaning every edge has to be visited exactly once, will give a possible sequence of the assembled genome. However, there is currently no way of automatically determining the best value for *k* for any given sequencing dataset. In this study we present a new software tool, which allows the user to quickly estimate the best *k* for their data. We achieve this by generating approximated abundance histograms for all possible k-mers, and then using a maximum likelihood estimation to fit a generative model to those histograms. This allows us to estimate how many of the sequenced *k-mers* are actually genomic (error-free). Finally, the tool chooses the value for *k* which maximizes the number of correct *k-mers*. We also show that our tool estimates some of the best values for *k*. The tool we present provides a way to drastically simplify the process of genome assemblies.

Abstract:
Informed and automated k-mer size selection for genome assembly

To assemble genetic sequences, different approaches can be used. For instance, the reads of a sequencing method can be assembled by de Bruijn based assemblers. To use de Bruijn based assemblers, there are parameters, that must be set to complete the assembly. The choice of the word length k, which produces the k-mers in the de Bruijn graph is a crucial factor for the quality of the assembly. To determine an adequate k by computing, a huge amount of computational effort is necessary. In this paper, a method is proposed, to determine a good choice for the parameter k in assembly problems with moderate computational requirements. Here we show a technique, that uses statistical modeling to propose the value k, which is capable to outperform other methods in benchmarks. In previous algorithms, the prediction of k was done by calculating the assembly for many different k values and chose the one that optimizes e.g., the scaffold N50 value. By using statistical modeling and circumventing the calculation of every possible assembly, the runtime of this approach is a lot lower with still good results. Now that a good k can be found in a moderate time with moderate computational effort, more data can be analyzed in data centers and can lead to a significant speedup in assembly research. As a perspective, this approach of statistical modeling for the optimizing problem of the k-mer choice can be applied to not uniform coverage problems such as assemblies of single cell sequencing experiments. Overall, there is still a lot of potential improvement and possibilities for the future of this project.

Pavel A. Pevzner, Michael S. Waterman	David R Kelley, ,Steven L Salzberg
An Eulerian path approach to DNA fragment assembly	Quake: quality-aware detection and correction of sequencing errors
PNAS vol. 98 9748–9753	Genome Biology, 11 :R116
August 14 2001	2010
865	370
Wegen der Ausführungen zum k-mer based Assembies.	Wegen der Fehlererkennungssoftware

QUAST: quality assessment tool for genome assemblies	Erstautor: Guillaume Marçais, korrespondierender Autor: Guillaume Marçais, Letztautor: Carl Kingsford
Alexey Gurevich, ,Glenn Tesler	A fast, lock-free approach for efficient parallel counting of occurrences of k-mers
BIOINFORMATICS Vol. 29 no. 8	Bioinformatics, Volume 27, Pages 764-770
February 19, 2013	15 March 2011
2,382	1140 Citations
Wegen der Referenz für das Benachmarking	Dieses Paper bietet einen Absprungs punkt zum Entwickeln einer Methode zum Zählen von k-mers.

Erstautor: Guillaume Rizk, korrespondierender Autor: Rayan Chikhi, Letztautor: Rayan Chikhi
DSK: k-mer counting with very low memory usage
Bioinformatics, Volume 29, Pages 652-653
1 March 2013
131 Citations
Dieses Paper bietet einen Absprungs punkt zum Entwickeln einer Methode zum Zählen von k-mers.

GenomeScope: fast reference-free genome profiling from short reads [HTML] oup.com
GW Vurture, FJ Sedlazeck, M Nattestad... - ..., 2017 - academic.oup.com
GenomeScope is an open-source web tool to rapidly estimate the overall characteristics of a genome, including genome size, heterozygosity rate and repeat content from unprocessed short reads. These features are essential for studying genome evolution, and help to choose ...
☆ 99 Cited by 479 Related articles All 16 versions

[HTML] A supergene determines highly divergent male reproductive morphs in the ruff [HTML] nature.com
C Küpper, M Stocks, JE Risse, N Dos Remedios... - Nature ..., 2016 - nature.com
Three strikingly different alternative male mating morphs (aggressive'independents', semicooperative'satellites' and female-mimic'faeders') coexist as a balanced polymorphism in the ruff, Philomachus pugnax, a lek-breeding wading bird 1, 2, 3. Major differences in ...
☆ 99 Cited by 304 Related articles All 32 versions

Constitutive activation of the Ras/mitogen-activated protein kinase signaling pathway promotes androgen hypersensitivity in LNCaP prostate cancer cells [PDF] aacrjournals.org
Free from Publisher
RE Bakin, D Gioeli, RA Sikes, EA Bissonette, MJ Weber - Cancer research, 2003 - AACR
Progression of prostate cancer ultimately results in a disease that is refractory to hormone ablation therapy but nevertheless continues to require the androgen receptor. Progression to hormone refractory disease is often correlated with overexpression of growth factors and ...
☆ 99 Cited by 257 Related articles All 7 versions

BUSCO: assessing genome assembly and annotation completeness
M Seppey, M Manni, EM Zdobnov - Gene prediction, 2019 - Springer
Genomics drives the current progress in molecular biology, generating unprecedented volumes of data. The scientific value of these sequences depends on the ability to evaluate their completeness using a biologically meaningful approach. Here, we describe the use of ...
☆ 99 Cited by 418 Related articles All 5 versions

Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction [HTML] oup.com
D Laehnemann, A Borkhardt... - Briefings in ..., 2016 - academic.oup.com
Characterizing the errors generated by common high-throughput sequencing platforms and telling true genetic variation from technical artefacts are two interdependent steps, essential to many analyses such as single nucleotide variant calling, haplotype inference, sequence ...
☆ 99 Cited by 238 Related articles All 13 versions

Computational methods for discovering structural variation with next-generation sequencing 614 2009
P Medvedev, M Stanciu, M Brudno
Nature methods 6, S13-S20

Informed and automated k-mer size selection for genome assembly 535 2014
R Chikhi, P Medvedev
Bioinformatics 30 (1), 31-37

Computability of models for sequence assembly 194 2007
P Medvedev, K Georgiou, G Myers, M Brudno
International Workshop on Algorithms in Bioinformatics, 289-301

Detecting copy number variation with mated short reads 190 2010
P Medvedev, M Fiume, M Dzamba, T Smith, M Brudno
Genome research 20 (11), 1613-1622

Computational pan-genomics: status, promises and challenges 178 2016
Computational Pan-Genomics Consortium
Briefings in Bioinformatics, bbw089

Wir haben uns für diese 5 Paper entschieden, da diese bei Google Schoolar am häufigsten zitiert wurden.

Sequence alignment using machine learning for accurate template-based protein structure prediction

Shuichiro Makigaki, Takashi Ishida

Bioinformatics, 2020, 36(1), 104–111

Abstract book created by
Aayush Marishi and Magdalena Weber

Abstract

Aayush Marishi

Proteins are one of the key molecules in biological sciences, there have been countless methods developed to predict a protein's functions. One of the most popular methods is Template-based modelling which predicts structures based on templates and the sequence alignment to a target protein. Long-term homology detection studies have detected homology with very high accuracy. The inputs are the target protein's amino acid sequence and another amino acid sequence that was detected as a template by an homology detection method. The output is alignment which is more suitable for homology modelling. However, The sequence alignments generated by the homology detection methods are dissimilar to those generated by structural alignment, especially for remote homologs. Here we propose a Machine learning based model that generates more accurate alignments than other models. This machine learning based model learns the structural alignment of known homologs. It uses dynamic programming during sequence alignment to dynamically predict a substitution score from the learned model. The improved model accuracy over the other models model clearly show the prowess of this model. However, it is difficult to judge whether this improvement is useful for advanced applications such as protein function estimation. Our model has provided, what we would anticipate as the ground work for more machine learning based models for protein structure and function prediction. This also lays some ground work for more complex models such as Deep artificial neural networks for improved computational efficiency.

Abstract

Magdalena Weber

For research concerning protein function and simulation of ligand docking it is esstential to know the structure of a protein. There are different computational approaches to generate protein structures from sequence data, the de novo simulations and template based modeling (TBM). The common approach for TBM is homology detection, sequence alignment and finally the modeling of the target protein based on the template protein. The homology detection methods have been primarily improved, since it is important to find the best possible template, but for higher accuracy of the target protein structure the alignment quality is crucial. Here we focus on providing an automated alignment method for accurate TBM using machine learning. It learns structural alignment using known homologs and predicts a substitution score during alignment using dynamic programming. As structural alignment minimizes the structural difference between target and template protein it is ideal for TBM. The structures predicted from our method are very close to those generated from structural alignment while other methods like Smith-Waterman, HHsearch and DELTA-BLAST are less precise. This method can be incorporated into template based structure prediction, it enhances the accuracy of 3D models and consequently may serve for protein function estimation and homology detection. The efficiency of our method can still be increased by using faster kNN algorithms including approximate schemes. It is a good base for the developement of further methods employing machine learning or even convolutional neural networks.

Summary

1 Research Question

Can template based modeling be optimized?

One of the most accurate method of predicting protein structure is Template based modelling (TBM). This paper aims to optimize the process and introduce a novel Machine Learning based method that performs better in terms of prediction accuracy and computational speed as compared to already existing state of art methods for TBM.

2 Relevant Methodological Approaches

There are a number of different methods used in this paper to solve the aforementioned research question. The relevant methods are listed below:

- Alignment of sequences using the Smith-Waterman Algorithm
- Using BLOSUM62 or PAM250 to evaluate the matches between residue pairs
- Profile comparison methods (to improve alignment accuracy) like, FORTE and FFAS
- Training of a prediction model with pairwise structural alignment data. (kNN)
- Prediction of substitution scores for each residue pair.

The training of the model itself requires a series of different approaches in order to find the best machine learning model the methods involved are listed below:

- Making the dataset to train, test and validate the ML model.
  - Generating structural alignments of every domain pair using TM-align.
  - PSSM generation using three-iteration PSI-BLAST with UniRef90 database.
  - Reducing the size of training dataset by random selection
- Encoding information about residue pairs in a numerical vector representation for easy input.
  - inputs are made to be a pair of query and template PSSMs and a residue position.
  - outputs are predicted label and normalized confidence score
- Parameter optimization
  - Grid search was used to find the best combination of hyper-parameters to tune the model. (Hyper-parameters: Number of nearest neighbors, gap open penalty, gap extend penalty)

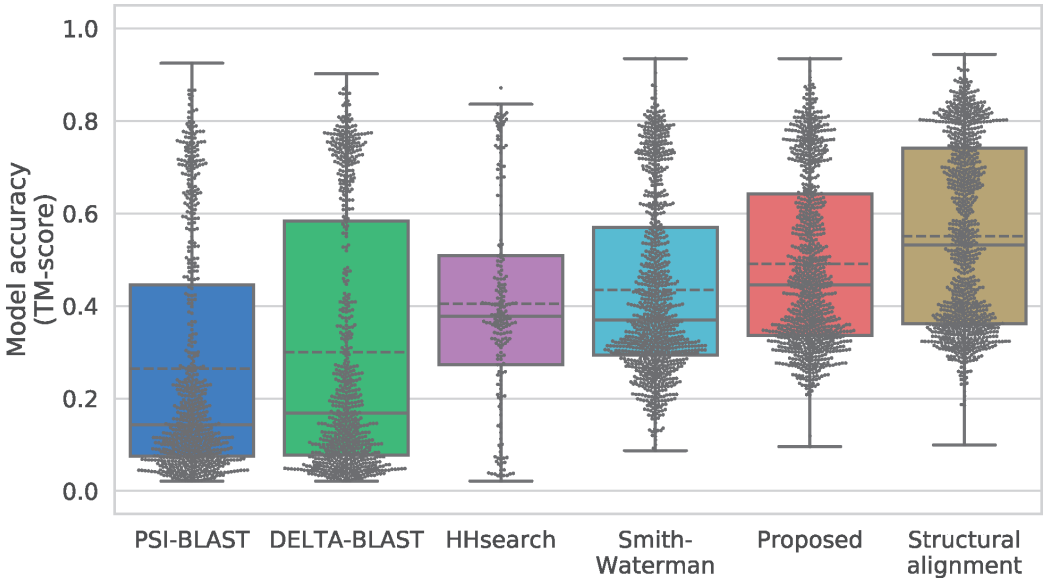
3 Relevant Results

- The proposed method predicted labels accurately except for d2axto1 (showed random prediction).
- The proposed method attained an accuracy(TM-score) of 0.499 which is very close to structural alignments (0.551) which are the most accurate models, It also has a higher score than most of the state of the art models: PSI-BLAST, DELTA-BLAST, HHsearch, Smith-Waterman
- The proposed method succeeded in aligning almost a whole protein as compared to methods like HHsearch who failed to do so.

4 Conclusion

- The method delivers more accurate 3D models than other comparable methods.
- current execution time of the proposed model is high so it does not provide so much of a computational speed edge over the other models.

5 Key Figure



The key figure of this paper was figure 6 which compared the proposed model against the state of art models. The figure clearly shows the prowess of the novel Machine learning model over the other models.

Paper Impact

Aayush Marishi, Magdalena Weber

Import references:

1. Hijikata,A. et al. (2011) Revisiting gap locations in amino acid sequence alignments and a proposal for a method to improve them by introducing solvent accessibility. Proteins Struct. Funct. Bioinform., 79, 1868–1877. - **cited: 12**
2. Manavalan,B. and Lee,J. (2017) SVMQA: support–vector-machine-based protein single-model quality assessment. Bioinformatics, 33, 2496–2503. - **cited: 41**
3. Wang,S. et al. (2016) Protein secondary structure prediction using deep convolutional neural fields. Sci. Rep., 6, srep18962. - **cited: 101**
4. Wei,L. and Zou,Q. (2016) Recent progress in machine learning-based methods for protein fold recognition. Int. J. Mol. Sci., 17, 2118. - cited: 15
5. Cao,R. et al. (2016) Deepqa: improving the estimation of single protein model-quality with deep belief networks. BMC Bioinformatics, 17, 495 - **cited: 42**

Cited by – A. Jain and S. Tiwari, "Prediction and Visualization of Viral Genome Antigen Using Deep Learning & Artificial Intelligence," *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021, pp. 1430-1437, doi: 10.1109/ICCMC51019.2021.9418356.

Other relevant works by the author:

Works in English:

1. Sequence alignment generation using intermediate sequence search for homology modeling. Shuichiro Makigaki, Takashi Ishida; DOI: 10.1016/j.csbj.2020.07.012, Published: 2020
2. Improvement of template-based protein structure prediction by using chimera alignment. Shuichiro Makigaki, Takashi Ishida; DOI: 10.1145/3180382.3180405, Published: 2018

Works in Japanese:

1. 最近傍法を用いた構造予測向け配列アラインメント生成手法の高速化
   English: Acceleration of sequence alignment generation method for structure prediction using k nearest neighbors' method. Published: 2009
2. 機械学習を用いたホモロジーモデリングのための配列アラインメント生成手法の高速化
   English: Acceleration of machine learning-based sequence alignment generation for homology modeling.
3. 配列類似性ネットワークに基づく高感度な遠縁タンパク質検索
   English: Highly sensitive distant protein search based on sequence similarity network.
4. 機械学習を用いたホモロジーモデリングのための配列アライメント手法
   English: Sequence alignment method based on k-Nearest Neighbor for improving homology modeling.

Reasons why the references are selected as important: (numbering with respect to Important references)

1. This paper by Hijikata et al. supports S.Makigaki's paper as it proposed an automated method (similar to theirs) where they improve upon the prediction of alignment by optimizing gap penalties. This served as a comparative model for S.Makigakis's paper.
2. This paper by Manavalan et al. lays the groundwork for the S.Makigaki's approach in using machine learning based methods for protein model quality assessment. In a way saying that they employed some methodology that was indirectly inspired by Manavalan's work.
3. This paper by Wang et al. is quite similar to the Manavalan et al. providing more support to the idea that newer machine learning based methods are required to solve the protein structural analysis problems.
4. This paper by Wei et al. summarizes all the progress made in the field of machine learning based methods for protein fold recognition. This provides more leway to the fact that the paper written by S.Makigaki is quite relevant to the topic that has intrigued other scientists, and the area of research (machine learning based models for protein structure prediction) requires more work.
5. This paper by Cao et al. This paper on the other hand provides the expert knowledge in tuning the hyperparameters of a machine learning based model which is used to estimate protein model quality. This paper might have inspired the hyper parameters of the machine learning model

DeepNOG: fast and accurate protein orthologous group assignment

Roman Feldbauer, Lukas Gosch, Lukas Lüftinger, Patrick Hyden, Arthur Flexer, Thomas Rattei

Bioinformatics, Volume 36, Issue 22-23, 1 December 2020, Pages 5304–5312

Presented by:
Jeff Gower & Nico Bohlinger

Abstract 1

The primary structure of proteins, i.e. the sequence of amino acids, largely determines the conformation of these, which in turn influences functional and phylogenetic information. Orthologous genes, genes that can be found in different species and can be traced back to common ancestors, also provide much information about protein function. There are are several public resources that provide information about millions of orthologous groups. In order to to classify protein sequences mappings against these resources are performed. Alignment-based methods are currently the classical approach to do these mappings. These methods are relatively slow and with the ever-increasing homology databases the alignment-based methods become a computational bottleneck. DeepNOG provides a solution that can achieve almost similar results in accuracy, precision and recall with exceptionally better computation time. Previously the inference time has been around multiple minutes per 1000 sequences worsening with larger datasets. DeepNOG cannot achieve similar results to the classical alignment-based methods but they are almost as good, especially for larger datasets. The computational time stays linear with larger datasets, around 20 to 30 seconds per 1000 sequences using the CPU and 0.6 seconds using the GPU. DeepNOG outperforms the other non-classical alternative DeepFam, which is also faster than the alignment-based methods but slower than DeepNOG and performs way worse than DeepNOG in terms of accuracy, precision and recall.

Abstract 2

Assigning amino acid sequences to orthologous groups forms important genetic relationships to extract functional information about proteins and it enables further phylogenetic analysis. State of the art alignment-based methods are commonly used to create high sensitivity assignments. With the exponentially growing surge of sequenced genetic data, those methods won't be able to handle the sheer volume of this data and output assignments in a reasonable time. Efficient alignment-free deep-learning-based algorithms could breach the gap between high enough sensitivity and fast enough assignments.

We have developed DeepNOG, an alignment-free convolutional neural network model, that enables highly accurate and fast assignments within large orthology databases. We found that our model clearly outperforms the accuracy scores of similar previous attempts like DeepFam on really big datasets like eggNOG 5. Whereas alignment-based methods like DIAMOND and HMMER still reign supreme in the category for highest precision and recall, deepNOG creates assignments with comparable accuracy scores while being an order of magnitude faster. Our work shows that on GPUs DeepNOG can easily assign 1000 sequences in less than one second. Assigning sequences with DeepNOG can scale to even bigger future databases because the interference time won't change with the number of datapoints and the architecture can utilize pre-trained models to enable a flexible and fast training process. We believe that in developing and publishing DeepNOG open source, assigning sequences to orthologous groups will be easier and especially faster for everyone working on such data. With advances in deep learning methods and the field of machine learning in general the usefulness of these kinds of algorithms will only increase for genetic data.

Summary

1. Research questions

- - How can current deep learning architectures using convolutional neural networks for the assignment of homologous groups (e.g. DeepFam) be enhanced to better certain performance indicators such as accuracy and speed?
- - How do deep learning approaches scale regarding ever-increasing homology databases? How do these compare to classical alignment-based approaches where, looking at the amount of low-cost high-throughput sequencing technologies, they represent a bottleneck

2. Relevant methodological approaches

- - Performance indicators used: accuracy, precision, recall, interference time
- - Sequences of any length can be input, the amino acids are represented in a variable-

  dimensional vector and the dimension is also learned

- - The resulting word embedding is passed through a 1D convolutional layer and then

  through a 1-max-pooling layer

- - A softmax layer assigns the sequence to exactly one homology group
- - An assignment confidence threshold avoids false positives

3. Relevant results

- - Accuracy in smaller databases (e.g. COG) similar to DeepFam, still worse than alignment-based algorithms
- - Accuracy in larger databases (e.g. eggnog) 10-20% better than DeepFam, still worse than alignment-based algorithms
- - Consistently low inference time, even with larger databases, which is way better then classical alignment-based algorithms, that scale with the amount of data
- - 0.6 seconds needed for classifying 1000 sequences (using GPU)

4. Conclusion

- - DeepNOG outperforms DeepFam in accuracy on large orthology databases like eggNOG and that only at a fraction of the computational cost of classical alignment-based methods like pHMMs
- - DeepNOG allows for an automatic assignment of sequences to homology classes
- - Users can use the open-source architecture to train their own models

- - DeepNOG enables the combination with other homology tools in order to classify

  sequences that fall under the assignment confidence threshold differently and to enable full coverage

5. Key figure

Jeff Gower, Nico Bohlinger

Table 2. Inference time (seconds/1000 sequences) for COG and eggNOG 5 (bacteria level)

	COG-500	COG-100	NOG_2^5-500	NOG_2^5-100
DIAMOND	161.7	214.5	781.6	810.0
pHMMs	96.3	207.0	218.9	253.7
DeepFam	49.0	50.2	n/a	n/a
DeepFam light	32.7	35.0	34.9	38.7
DeepNOG (CPU)	**24.3**	**26.0**	**26.4**	**28.9**
pHMMs (parallel)	4.8	5.1	9.5	14.4
DeepNOG (GPU)	0.6	0.6	0.6	0.6

Note:Fastest method **bold** (single core). Averages over three replicates. Parallel pHMMs used 29x16 CPU cores.

- - The table shows DeepNOGs gain in speed compared to DeepFam and especially in comparison to the classical alignment-based algorithms (DIAMOND, pHMMs)
- - The classical algorithms are not only significantly slower from the outset, but their speed also scales disadvantageously with the amount of data, therefore the larger the amount of data, the greater the speed gain through DeepNOG

Information & Data collected: 08.06.2021

Impact

1. Five most relevant references

1.
- Author: First author: Seokjun Seo; Corresponding author: Sun Kim; Last author: Sun Kim
- Title: DeepFam: deep learning based alignment-free method for protein family modeling and prediction
- Journal: *Bioinformatics*, Volume 34, Issue 13, 01 July 2018, Pages i254–i262
- Publishing date: 2018
- Number of citations: 32
- Explanation: Uses deep learning as an alignment-free method for orthologous group assignment too. Its architecture serves as reference for DeepNOG.

2.
- Author: First author: Sean R Eddy; Corresponding author: Sean R Eddy; Last author: Sean R Eddy
- Title: Accelerated Profile HMM Searches
- Journal: PLoS Comput Biol. 2011 Oct 7(10)
- Publishing date: 2011
- Number of citations: 2795
- Explanation: Explains profile hidden markov models (Profile HMMs) and introduces HMMR, which the DeepNOG paper uses for benchmarking against an alignment-based method.

3.
- Author: First author: Edoardo Pasolli; Corresponding author: Nicola Segata; Last author: Nicola Segata
- Title: Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle
- Journal: Cell Volume 176, Issue 3, 24 January 2019, Pages 649-662.e20
- Publishing date: 2019
- Number of citations: 444
- Explanation: Serves as an example for the main hypothesis and the reasons methods like DeepNOG are needed that there will be soon billions of protein sequences waiting for analysis.

4.
- Author: First author: Jaime Huerta-Cepas; Corresponding author: Jaime Huerta-Cepas, Peer Bork; Last author: Peer Bork
- Title: eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses
- Journal: *Nucleic Acids Research*, Volume 47, Issue D1, 08 January 2019, Pages D309–D314
- Publishing date: 2019
- Number of citations: 462
- Explanation: Introduces eggNOG 5 a database for orthology relationships, functional annotation and gene evolutionary histories. Used for the benchmarking of DeepNOG.

5.
- First author: Michael Y. Galperin; Corresponding author: Eugene V. Koonin; Last author: Eugene V. Koonin
- Title: Expanded microbial genome coverage and improved protein family annotation in the COG database
- Journal: *Nucleic Acids Research*, Volume 43, Issue D1, 28 January 2015, Pages D261–D269
- Publishing date: 2015
- Number of citations: 737
- Explanation: Explains how to use the COG database for protein family annotation. Used for the benchmarking of DeepNOG.

2. Referencing DeepNOG

Number of citations: 1

Citations per year: 1

Five most influential citations:
- Deep hierarchical embedding for simultaneous modeling of GPCR proteins in a unified metric space

3. Author

1. Genome sequencing and analysis of the model grass Brachypodium distachyon
2. Deciphering the evolution and metabolism of an anammox bacterium from a community genome
3. The dynamic genome of Hydra
4. Complete nitrification by Nitrospira bacteria
5. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea

Explanation:
First we took the Journal Impact Factor (JIF), the number of citations and the position of the author from the twenty most cited papers that were published. We tried to consider the order of the parameters of the assignment. The JIF was therefore a major factor in calculating the authors papers relevance. All of the papers published in "Nature" are therefore, and, because they are all amongst the most cited ones, very relevant. The last paper we chose has been published in "Nature Biotechnology", which has still a high JIF and a high amount of citations. Twice he has been last author, once even the corresponding author, but these papers have been published in "Bionformatics" and "PLOS Pathogens" which have way lower JIFs and the numbers of citations.