ABSTRACT BOOK

Aktuelle Themen der Sequenzanalyse
[2022]

Betreuer des Seminars: Prof. Ingo Ebersberger

1 INHALTSVERZEICHNIS

Anmerkungen:

1. Die einzelnen Kapitel sind gegliedert in:
    a. Abstracts
    b. Zusammenfassung
    c. Impact
2. Die Abstracts wurden nach der Vorlage des Journals „Nature“ verfasst. Die Vorlage kann unter folgendem Link eingesehen werden:
    - http://www.cbs.umn.edu/sites/default/files/public/downloads/Annotated_Nature_abstract.pdf

Group: Dara Da Silva Weirich, Karolina Anna Kloss

Molecules as Documents of Evolutionary History

from

Emile Zuckerkandl, Linus Pauling

Abstract 1

When we speak of evolutionary history, it is very difficult to reproduce the step from our present point of view. It is still open to clarify how species have evolved over time and how relationships within and outside species have come about. In order to understand at what point of evolution we are today, we need a reliable bridge which, takes us from today to yesterday. One way to establish this is to examine the cells of today's living organisms. In this context, the focus is on macromolecules that we know carry genetic information. For this reason, we will ask the following question: Which type of molecules are the most appropriate for providing the basis for a molecular phylogeny? Here we explain why “semantides” (DNA, RNA, Polypeptides) are the best use for phylogenetic reconstruction and what effects semantic substitutions has on nucleotide sequences and amino acid sequences. Semantides seem to be the best choice for phylogenetic reconstruction, because unlike other molecule groups, they provide genetic information in firsthand. We also found that isosemantic substitutions have an impact on cryptic genetic polymorphism. We distinguish between changes in the nucleotide sequence that affect the amino acid sequence, which we call secondary crypticity, and changes in the nucleotide sequence that do not affect the amino acid sequence, which we have termed primary crypticity. The hypothesis is supported by signatures of different hemoglobin types. Possible consequences of these results are also investigated in terms of evolutionary stability and synthesis rate. In conclusion it seems to be likely that semantic substitutions have a far reaching impact on population genetics.

Group: Dara Da Silva Weirich, Karolina Anna Kloss

Molecules as Documents of Evolutionary History

from

Emile Zuckerkandl, Linus Pauling

Abstract 2

In phylogenetic research different molecules can provide an insight into the historical evolution of organisms. Living matter preserves the largest amount of its own past history and is therefore especially interesting to analyze in order to understand more about phylogeny. Furthermore, the degeneracy of the genetic code makes the task at hand more difficult. We aim to define molecules that are of most phylogenetic interest and attempt to find relationships between hemoglobin types, by taking the existence of isosemantic substitutions into account. Here we show, why semantides (DNA, RNA and the corresponding amino acid sequence) might be best fitted for our stated task and propose some consequences of isosemantic substitution. Examining semantopheric molecules further as well as the relationship between the different types of semantides is consequently of huge interest. Isosemantic substitution might influence the synthesis rate of polypeptides and even affect the evolutionary stability, however, the sole existence of them suggests that DNA can reveal more than amino-acid sequences could. Universal and informative molecular phylogeny will most likely be built on semantides alone since they preserve the highest amount of its own history. The influence of underlying processes when constructing a protein is still immense and very complex, but we believe that most answers to those processes as well as to evolutionary mysteries lie within the DNA, rather than molecules that contain only fragments of genetic information or none at all.

Group: Dara Da Silva Weirich, Karolina Anna Kloss

Molecules as Documents of Evolutionary History

from

Emile Zuckerkandl, Linus Pauling

1.) Research Question:

- How can we best trace evolutionary history?
- Why are semantides the best approach?
- What are the effects of isosemantic substitution on polypeptide chain construction?

2.) Methods

By comparing several previous papers and research results the topic is discussed and concluded.

3.) Results:

According to the paper, semantides are the best lead in order to understand the evolutionary development of an organism and to discover phylogenetic relationships. This is because semantides, rather than episemantic molecules, contain genetic information first hand, while the other molecules are only a product of it or do not contain genetic information at all.
The authors inspect semantides further and compare amino acid chains and the genetic code to emphazise the phylogenetic interest of those subcategories. The paper suggests that multiple codon triplets could encode the same amino acid. After the distinction between primary and secondary crypticity is established (primary = change of a nucleotide → change of amino acid, secondary = change of nucleotide → no change of amino acid), the theory gets backed up with previous findings on different hemoglobyn types and their conversions. Possible consequences and functions of said isosemantic substitutions are discussed, like an influence of the synhesis rate of polypeptide chains as well as the evolutionary stablity.

4.) Conclusion

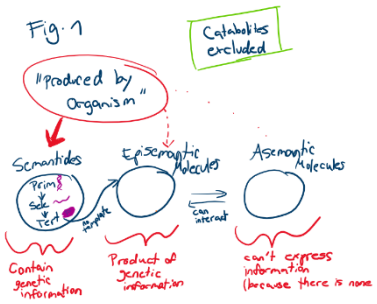
The paper indicates that there is more evolutionary history enscripted in nucleic acid sequences than in amino acid chains or even proteins. Instead of presenting the evolutionary history of certain molecules, the authors decided to present methods and principles that can help to extract those informations of various molecules. Although it has been claimed that nucleic acids should be preferred when doing so approaches for less complex molecules get discussed and evaluated as well.

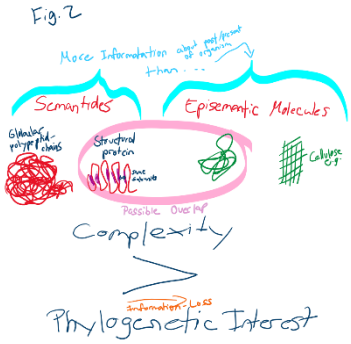
Considering that at the time of this publication, the nowadays well known code sun did not exist, the authors contributions are quite important and mostly acurate. However, some misleading connections and oversimplifications could not have been avoided.

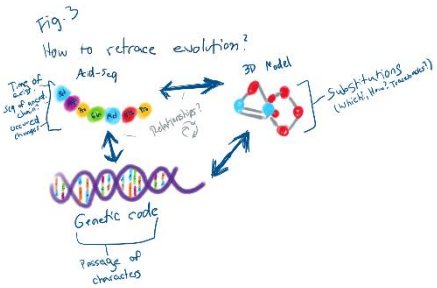
5.) Key-Figures:

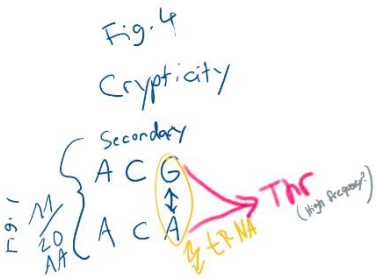
Since there are no actual graphics in the paper, we came up with some of our own. We decided that we will use Fig. 1 and Fig. 4 (maybe merge both) as key graphics. Fig. 1 provides us a good overview over the first part of the paper (The Chemical Basis for a Molecular Phylogeny) and shows the central three molecules discussed. Fig. 4 main focus are the two types of substitutions which are discussed in the second part of the paper (Cryptic Genetic Polymorphism through Isosemantic Substitution).

(Those are just doodles and concept ideas, the final graphics will be revised and created with the computer)

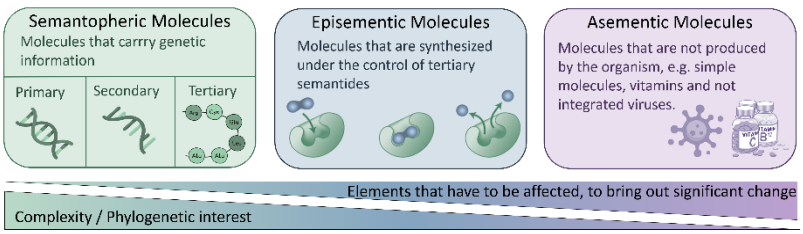








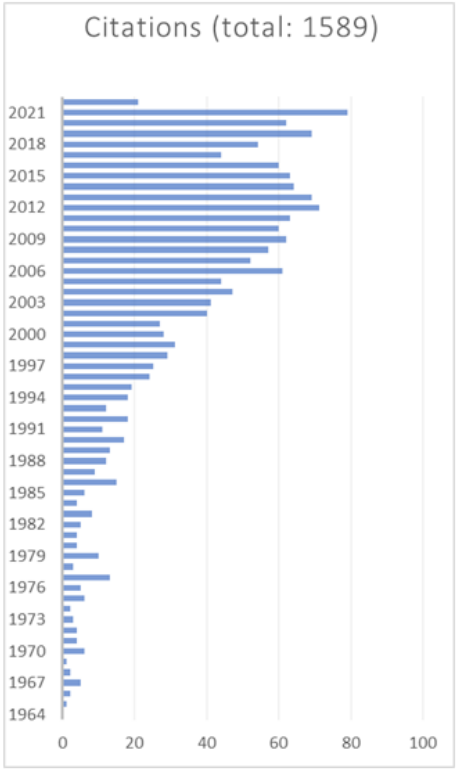
Finished Key-Figure:



Group: Dara Da Silva Weirich, Karolina Anna Kloss

Molecules as Documents of Evolutionary History from Emile Zuckerkandl, Linus Pauling

Authors	Cited Publications	Journal	Year	Citations	Relevance for the paper
Emile Zuckerkandl& Linus Pauling Michael Kasha & Bernard Pullman	Molecular Disease, Evolution, and Genic Heterozygosity	Horizons in Biochemistry	1962	37	These authors were chosen because they are not only the authors of the actual paper, but because the argumentation in the paper is often based on earlier works by E. Zuckerkandl & L. Pauling. For example, in earlier works the two authors have also discussed ways of gaining information about evolutionary history through the comparison of homologous polypeptide chains
Bernard Weisblum, Seymour Benzer, Robert W. Holley	A PHYSICAL BASIS FOR DEGENERACY IN THE AMINO ACID CODE	PNAS	1962	183	The paper was chosen because the fact that more than one base triplet can code for an amino acid was important knowledge at the time. And also some argumentations in the paper are based on this knowledge (e.g. the question is asked why this is so? This also explains why one can infer the amino acid sequence from the nucleotide sequence, but not necessarily vice versa). Furthermore, the study forms the basis for R. V. Eck's claims, which also play a role in the paper.
Richard V. Eck	Genetic Code: Emergence of a Symmetrical Pattern	Science	1963	90	The work of Eck mentioned here is based on B. Weisblum et al (A PHYSICAL BASIS FOR DEGENERACY IN THE AMINO ACID CODE, 1962). Eck proposes the thesis that the middle base in the base triplet is unimportant for transferRNAs. E. Zuckerkandl and L. Pauling take up this thesis and discuss the significance of this finding.
J. A. HUNT & Vernon Martin Ingram	Allelomorphism and the Chemical Differences of the Human Haemoglobins	Nature	1958	175	E. Zuckerkandl and L. Pauling discuss in their paper the different forms of human hemoglobin and how the different molecules are formed. J. A. HUNT & V. M. INGRAM form the basis for this discussion as they highlight the differences in their paper.
Harvey Akio Itano	The Human Hemoglobins: Their Properties and Genetic Control	Advances in Protein Chemistry	1957	206	H. A. Itano worked with L. Pauling on sickle cell anemia, the discussions between L. Pauling are fundamental for the part of the paper investigating human hemoglobin. Itano and L. Pauling used to think that the existence of isoalleles is probably linked to cryptic substitutions.
Seymour Benzer	ON THE TOPOGRAPHY OF THE GENETIC FINE STRUCTURE	PNAS	1961	782	Benzer's work has comparatively less impact for the paper, but the fact that GC are more common than AT pairs plays a major role in research today.



Following Publication	Journal	Year	5 year Impact Factor	Citations
Bacterial Evolution	Microbiological reviews	1987	0,34	9765
Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya	PNAS	1990	12,29	7768
A Molecular View of Microbial Diversity and the Biosphere	Science	1997	51,43	3551
From genomics to proteomics	Nature	2003	49,96	1503
Occurrence, Classification, and Biological Function of Hydrogenases: An Overview	Chemical reviews	2007	55,48	1605

Other publications from Zuckerkandl	Journal	Year	5 year Impact Factor	Citations	Position
Evolutionary divergence and convergence in proteins	Evolving genes and proteins	1965	(book)	3273	1
Molecules as documents of evolutionary history	Journal of theoretical biology	1965	2,69	1589	1
Chemical paleogenetics	Acta chemica Scandinavica	1963	3,4	281	2
A Comparison of Animal Hemoglobins by Tryptic Peptide Pattern Analysis	PNAS	1960	12,29	161	1
The appearance of new structures and functions in proteins during evolution	Journal of molecular evolution	1975	2,9	137	single

We decided that the number of citations is more important than the current impact factor due to journals merging and changes over the years. Zuckerkandl worked later mostly as a corresponding author and the chosen papers are therefore "older".

*all citations according to google scholar (29.5.2022)

Kristiyana Tsenova 6946704
Cem Bakisoglu 6888909

Informed and automated k-mer size selection for genome assembly

by Rayan Chikhi and Paul Medvedev

Abstract 1

Genome sequencing is crucial for studying the function and evolution of various organisms. Before analyzing genetic diversity or reconstructing evolutionary events, genome assembly is performed using overlap-based tools or de Bruijn graphs. The results of the de Bruijn graph-based assemblers depend on the size of the k-mers. In order to choose the best value for the parameter k, several conflicting aspects have to be considered, which complicates the decision. Currently, there are no tools that can automatically suggest the best possible value of k or quickly generate abundance histograms for many k-mer sizes. Here, we introduce the tool KmerGenie that estimates the number of distinct genomic k-mers and chooses the best k-mer size that yields the most distinct genomic k-mers. The decision for the best k is based on a quick sampling method, which generates approximate abundance histograms for many values of k. In comparison to other k-mer size optimization methods, KmerGenie is orders of magnitude faster and can be used for genomes of larger size. Its choice of k results in some of the best assemblies tested for different sequencing datasets. Therefore, KmerGenie could improve the quality of assemblies by automatically suggesting one of the best possible values of k. Moreover, it could help users to make an informed choice of the k-mer size by generating the abundance histograms. In addition, the abundance histograms could provide useful information for metagenome and transcriptome data, for which there is usually no single best k value.

Kristiyana Tsenova 6946704
Cem Bakisoglu 6888909

Informed and automated k-mer size selection for genome assembly

by Rayan Chikhi and Paul Medvedev

Abstract 2

Genome assembly refers to the reconstruction of the whole genome sequence from read fragments. The quality of the assembly with de Bruijn graph-based tools depends on the parameter k, of which the choice is a trade-off between different effects. Finding the optimal k for a data set is thus the key to the best assembly. However, there are currently hardly any tools to quickly determine the best k for a given data set. Here they show a fast sampling method for computing approximate abundance histograms and a fast heuristic that chooses the best k from the abundance histograms computed for many different k. They found that their method is not only fast but also fairly accurate in calculating abundance histograms for a k-value, making it superior to traditional exact counting methods in terms of speed. Moreover, they found that the heuristic optimally finds the parameter k, resulting in one of the best assemblies for each of the three test data sets they used. These results demonstrate that their tool KmerGenie is able to estimate an optimal k for different kinds of data sets, making it helpful to possibly increase the quality of an assembly. Hence, KmerGenie allows the user to learn in advance about the potentially optimal k-value for any given data set. Furthermore, approximate abundance histograms can be used in various ways, e.g., to obtain useful information for metagenome and transcriptome assembly

Kristiyana Tsenova 6946704
Cem Bakisoglu 6888909

Informed and automated k-mer size selection for genome assembly

by Rayan Chikhi and Paul Medvedev

1. Research Questions

- How to generate abundance histograms more quickly?
- How to estimate the best k-mer length for de Brujin graph-based genome assembly tools?
- How to estimate the number of distinct genomic k-mers for many putative k values quickly?

2. Relevant Methos

- Build approximate histograms for several putative k values by sampling from the k-mers (state-less 64 bits hash function).
- Estimation of the number of distinct genomic k-mers by modeling the genomic k-mers as a mixture of Gaussians and the erroneous k-mers as a Pareto distribution
- Fitting a generative model to the histograms (maximum likelihood estimation, optim function in R (BFGS algorithm))

3. Relevant Results

The statistical model generates approximate abundance histograms which are close to exact abundance histograms and distinguishable from nearby k values. Testing KmerGenie for the three datasets S.aureus, human chr.14 and B.impatiens, its choice of k results in the best assemblies of S.aureus and B.impatiens based on NG50 and assembly size, while for human chr.14 a tradeoff between NG50/assembly size and errors is suggested. Regarding computation time, the generation of approximate abundance histograms is 6–10 times faster than traditional exact k-mer counting methods. In comparison to other k optimizing methods, KmerGenie is by orders of magnitude faster and applicable for genomes of larger sizes.

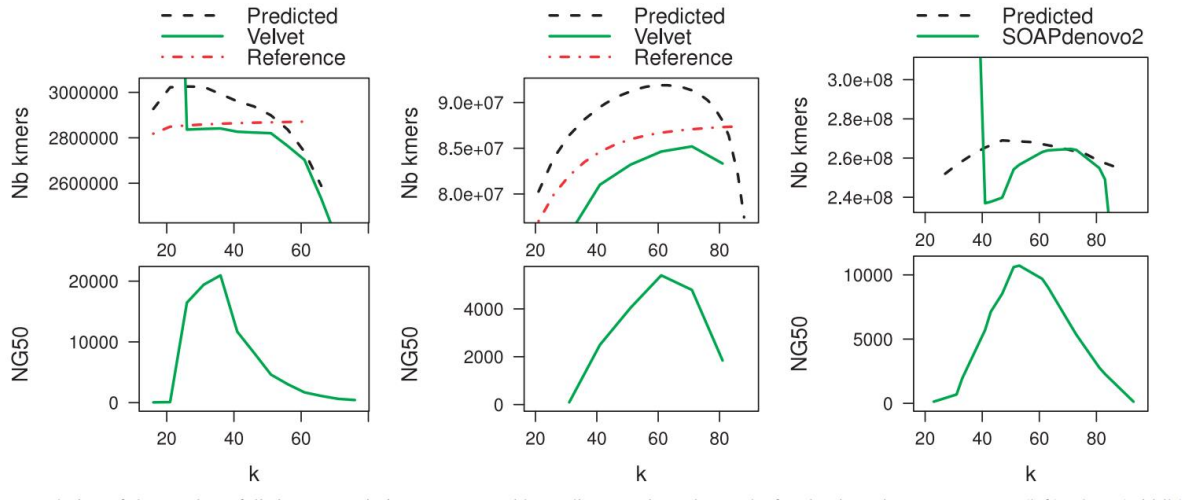
4. Conclusion

Overall, KmerGenie is a fast tool that is useful for estimating the best k-mer length for de Brujin graph-based assemblers. However, there are few limitations to KmerGenie. For instance, KmerGenie requires data with an even coverage. Therefore, it cannot find the best value of k for data from, for example, single cell experiments. Moreover, although there is not a single best k value for metagenome and transcriptome assembly, the approximate abundance histograms created for these kinds of data could still contain useful information

5. Key Figure

In the following figure, the number of predicted distinct genomic k-mers is put in relation to the quality of the assemblies based on the NG50 value for the datasets S.aureus (left), chr14 (middle) and B.impatiens (right). Furthermore, the predicted number of distinct genomic k-mers is compared to those of Velvet and the reference genome. For

all three organisms, the NG50 rises and falls in accordance with the number of predicted distinct genomic k-mers.



We have chosen this figure as the key figure because it shows the ability of KmerGenie to predict the number of distinct genomic k-mers and its relation to the quality of the assembly based on the NG50 value. The latter is an important aspect, because the authors assume in the model, that the best value of k provides the most distinct genomic k-mers.

Kristiyana Tsenova 6946704
Cem Bakisoglu 6888909

Top 5 relevant references:

Authors*	Title	Journal	Year	Citations	Reasoning
Pavel A. Pevzner[F] Michael S. Waterman[C,L]	An Eulerian path approach to DNA fragment assembly	PNAS 98.17 S. 9748-9753	2001	G. Scholar: 1,676	The paper forms the foundation of de Bruijn graph-based assembler.
Can Alkan[F], Evan E Eichler[C,L]	Limitations of next-generation genome sequence assembly	Nature Methods 8.1 S. 61-65	2011	G. Scholar: 765	It shows aspects of genome assembly that need to be improved
Guillaume Marçais[F,C] Carl Kingsford[L]	A fast, lock-free approach for efficient parallel counting of occurrences of k-mers	Bioinformatics 27.6 S. 764-770	2011	G. Scholar: 2,250	The running time of the approximate abundance histogram is compared with this method
David R Kelley[F,C], Steven L Salzberg[L]	Quake: quality-aware detection and correction of sequencing errors	Genome Biology 11.11 R116	2010	G. Scholar: 682	The generative model of the paper was adopted for the abundance histograms.
Daniel R. Zerbino[F], Ewan Birney[C,L]	Velvet: Algorithms for de novo short read assembly using de Bruijn graphs	Genome Research 18.5 S. 821-829	2008	G. Scholar: 10,233	Velvet is a de Bruijn based assembler. Velvet was used for the analysis KmerGenie

Total citations of our Paper (Google Scholar): 610, 66/year

Top 5 citations from the most influential studies:

1. Küpper, C., Stocks, M., Risse, J. et al. A supergene determines highly divergent male reproductive morphs in the ruff. Nat Genet 48, 79–83 (2016).
2. Shen, X. X. et al. Tempo and mode of genome evolution in the budding yeast subphylum. Cell 175, 1533–1545 (2018)
3. Ranallo-Benavidez, T.R., Jaron, K.S. & Schatz, M.C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun 11, 1432 (2020)
4. Hall, M., Kocot, K., Baughman, K. et al. The crown-of-thorns starfish genome as a guide for biocontrol of this coral reef pest. Nature 544, 231–234 (2017).
5. Meyer, A., Schloissnig, S., Franchini, P. et al. Giant lungfish genome elucidates the conquest of land by vertebrates. Nature 590, 284–289 (2021).

*First Author = F, Corresponding Author C, Last Author L

Kristiyana Tsenova 6946704
Cem Bakisoglu 6888909

Top 5 relevant publications of Dr. Paul Medvedev[C]

1. Medvedev, P., Stanciu, M. & Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 6, S13–S20 (2009). (2-year Impact Factor: 28.467, Google Scholar: 647, first author)
2. Sahlin, K., Medvedev, P. Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. Nat Commun 12, 2 (2021) (2-year impact factor: 14.919, Google Scholar: 52, corresponding author)
3. Sahlin, K., Tomaszkiewicz, M., Makova, K.D. Medvedev, P. Deciphering highly similar multigene family transcripts from Iso-Seq data with IsoCon. Nat Commun 9, 4601 (2018) (2-year Impact Factor: 28.467, Google Scholar: 40, corresponding author).
4. Minkin, I., Medvedev, P. Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. Nat Commun 11, 6327 (2020) (2-year Impact Factor: 28.467, Google Scholar: 20, Methodology, Validation, Writing, Funding acquisition).
5. Rayan Chikhi, Paul Medvedev, Informed and automated k-mer size selection for genome assembly, Bioinformatics, Volume 30, Issue 1, 1 January 2014, Pages 31–37(Impact Factor: 5.610, Google Scholar: 610, corresponding autor))

*First Author = F, Corresponding Author C, Last Author L

Group: Martin Brand, Dominik Sens

Assembly of long, error-prone reads using repeat graphs

from

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

Abstract

Reassembling DNA sequence reads into their correct order is called genome assembly. This task comes with numerous challenges like how to deal with repetitive regions and sequencing errors. It is especially difficult to assemble short reads e.g. Illumina reads accurately around those repetitive regions. Although single molecule sequencing (SMS) long-read technologies resolves this challenge to some extent, long repetitive regions stay problematic. Using de Bruijn assembly-graphs improves the accuracy for short read assemblies. However long-reads do not fully satisfy the requirements for this approach and therefore current long-read assemblers do not provide the necessary repeat characterization. Here we show our long-read assembly algorithm Flye which is able to construct an accurate repeat graph from disjointigs (arbitrary paths of a unknow repeat graph). In comparison to other current long-read assemblers Flye performed better or with at least the same accuracy while being significantly faster and contiguous. In the case of the human genome assembly the contiguity (NGA50 value) was almost doubled. Our results demonstrate that long-read assemblies are improved significantly with repeat characterization. The resulting assembly graph can be used to resolve bridged and unbridged repeats what leads to a more accurate assembly. Our results suggests that there is still room to improve the contiguity of SMS assemblies. Flye provides an assembly graph that is suitable as foundation for future genome finishing algorithms.

Group: Martin Brand, Dominik Sens

Assembly of long, error-prone reads using repeat graphs

from

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

Abstract

DNA sequencing is a complex and multifaceted field of research that uses methods from life science and computer science. One approach in DNA sequencing is to cut the DNA into many fragments called reads, sequence them individually and put them back together in the original order. The last step remains a challenge because of sequencing errors and repetitive regions, which make a clear allocation difficult. We want to develop an algorithm that can handle both sequencing errors and repetitive regions using long-read technologies while delivering accurate and trustful assemblies. Flye is a long-read assembly algorithm that addresses both problems. The algorithm generates initial assemblies from the reads and combines them into one single string. The resulting string is used to build an accurate repeat graph, with which we can reconstruct the sequence. We benchmarked Flye against 5 state-of-the-art assembly algorithms and presented the results. Flye performs better or at least as well as other state-of-the-art assembly algorithms while reducing runtime. DNA sequencing is necessary to understand evolution and has important applications in medicine, among other fields. Reliable assembly can further improve DNA sequencing.

Group: Martin Brand, Dominik Sens

Summary of the research paper „Assembly of long, error-prone reads using repeat graphs“

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

1.) Research Question:

How can we build an assembly algorithm, which handles elevated error-rates from long single molecule sequencing reads and resolves repeat regions?

2.) Methods:

- repeat graphs: compactly represent all repeats in a genome and reveals their mosaic structure

-> through arbitrary paths of the unknown repeat graph (called disjointigs)

- repeat characterization

3.) Results:

- compared to 5 other assemblers
- repeat characterization improves genome assembly
- is better or as good as other algorithms
- higher contiguity of assemblies
- order of magnitude faster

4.) Conclusion:

Flye was able to achieve an improved assembly (NGA50-value) of the human genome and therefore suggests that there is still room to improve the contiguity of SMS assemblies.

5.) Key-Figure:

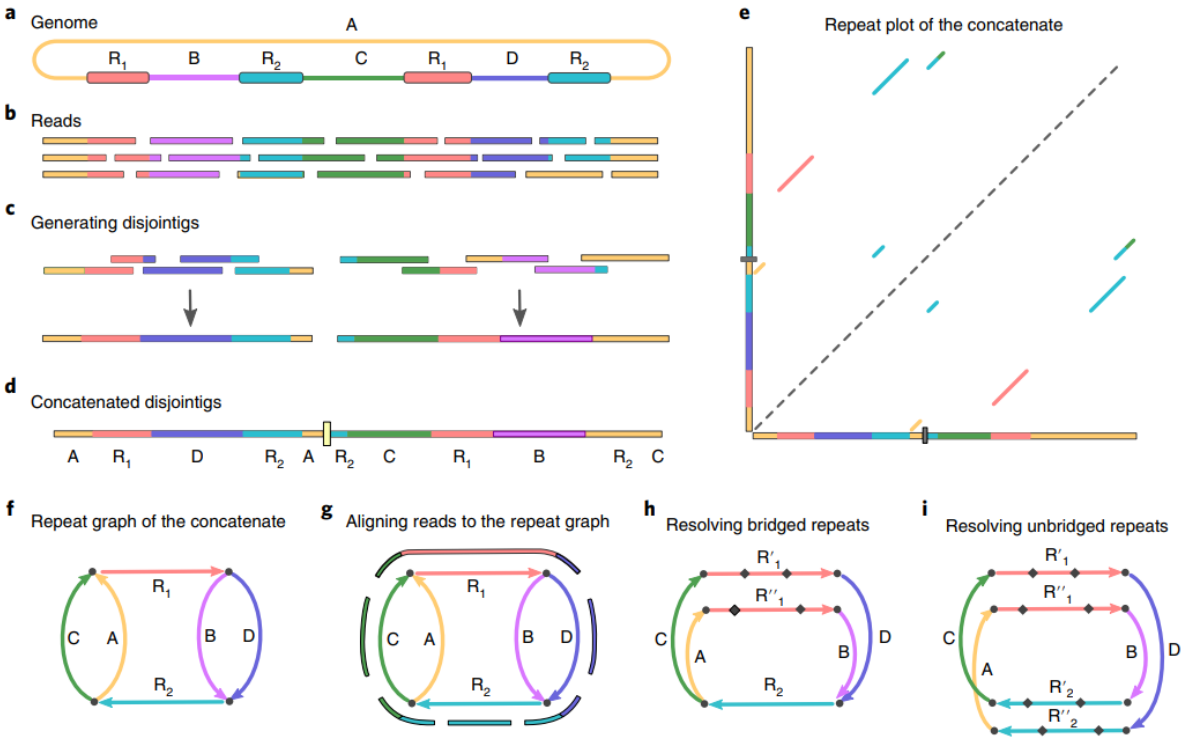


Figure 1:

The figure shows the most important steps of the algorithm. We have chosen the figure because it serves as a good overview from building those arbitrary paths of the unknown repeat graph (disjointigs) to resolving unbridged/bridged repeats.

Paper Impact of
Assembly of long, error-prone reads using repeat graphs

Authors: Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin & Pavel A. Pevzner

by Martin Brand, Dominik Sens

Most fundamental research papers of this research paper

1st reference:
Authors: Paul A. Pevzner, Haixu Tang, Glenn Tesler
Title: De Novo Repeat Classification and Fragment Assembly
Journal: Genome Research, Pages: 1786-1796
Publication Year: 2004
Number of Citations: 306
Rationale: about repeat graphs

2nd reference:
Authors: Zhaoshi Jiang, Haixu Tang, Evan E Eichler
Title: Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution
Journal: Nature, Pages: 1361-1368
Publication Year: 2007
Number of Citations: 206
Rationale: about repeat graphs

3rd reference:
Authors: Alla Mikheenko, Andrey Prjibelski, Alexey Gurevich
Title: Versatile genome assembly evaluation with QUAST-LG
Journal: Bioinformatics, Pages: i142–i150
Publication Year: 2018
Number of Citations: 339
Rationale: about evaluation of assembly algorithms

4th reference:
Authors: Yu Lin, Sergey Nurk, Pavel A Pevzner
Title: What is the difference between the breakpoint graph and the de Bruijn graph?
Journal: BMC Genomics
Publication Year: 2014
Number of Citations: 25
Rationale: Connection between breakpoint graphs and de Bruijn graph graphs

5th reference:
Authors: Yu Lin, Jeffrey Yuan, Pavel A. Pevzne
Title: Assembly of long error-prone reads using de Bruijn graphs
Journal: PNAS
Publication Year: 2016
Number of Citations: 176
Rationale: Fundamentals of error-prone long-read SMS assemblers

Metrics of our research paper

Citations total: 782 Citations p.a: ~261

Most impactful research papers that cite our research paper:

1st reference:
Authors: Jue Ruan & Heng Li
Title: Fast and accurate long-read assembly with wtdbg2

2nd reference:
Authors: Shanika L. Amarasinghe, Shian Su, Xueyi Dong et. al
Title: Opportunities and challenges in long-read sequencing data analysis

3rd reference:
Authors: Karen H. Miga, Sergey Koren, Arang Rhie et. al
Title: Telomere-to-telomere assembly of a complete human X chromosome

4th reference:
Authors: Haoyu Cheng, Gregory T. Concepcion, Xiaowen Feng et. al
Title: Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm

5th reference:
Authors: Víctor J. Carrión, Juan Perez-Jaramillo, Viviane Cordovez et. al
Title: Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome

Most impactful research papers of Mikhail Kolmogorov

Title: Assembly of long, error-prone reads using repeat graphs
Rationale: IF: 54.908, Position: 1, Citations: 782

Title: Assembly of long error-prone reads using de Bruijn graphs
Rationale: IF: 11.205, Position: 3, Citations: 176

Title: The complete sequence of a human genome
Rationale: IF: 47.728**,** Citations: 166

Title: Ragout—a reference-assisted assembly tool for bacterial genomes
Rationale: IF: 6.937, Position: 1, Citations: 144

Title: metaFlye: scalable long-read metagenome assembly using repeat graphs
Rationale: IF: 28.467**,** Position: 1, Citations: 130

Hania Ahmed

BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database

by

Tomáš Brůna, Katharina J Hoff, Alexandre Lomsadze et al.

Abstract 1

New methods of next generation sequencing have made it easy and accessible to sequence the whole genome of eukaryotes. However, genome annotation is crucial to find out more about the features and the functional role of the sequenced genome. Therefore, it is important to find tools that deliver accurate annotation of the eukaryotic genomes including possible isoforms of the genes. The authors of this paper had developed a pipeline called BRAKER1, which combines GeneMark-ET and AUGUSTUS and uses RNA-Seq data to annotate a novel genome. Here the authors present an extension of BRAKER1, the fully automatic pipeline BRAKER2. It combines the powerful properties of GeneMark-EP+ and AUGUSTUS for training and gene prediction. The pipeline is supported by massive protein databases. The goal of BRAKER2 is to extract information about the exon-intron structures from a set of different proteins. For a novel genome no close relative may be available in the protein set therefore this pipeline also utilizes remote homologs. However, the accuracy of the prediction is dependent on the protein set. Here the number of proteins and the taxonomical distance of the proteins plays a significant role. The authors were able to show that BRAKER2 delivers very accurate results that are better than some existing gene prediction tools (e.g. MAKER2). The BRAKER2 pipeline lends itself well to the annotation of protein-coding genes. For the future, a combination of BRAKER1 and BRAKER2 offers a good prospect to determine an even more accurate and precise gene prediction.

Hania Ahmed

BRAKER2 Zusammenfassung

1. Bearbeitete Forschungsfrage(n)

Wie kann man protein-codierende Gene eines eukaryotischen Genoms besser und genauer annotieren?

Wie gut und genau ist die Genvorhersage der Pipeline BRAKER2?

Wie genau ist BRAKER2 im Vergleich zu MAKER2, BRAKER1 und GeneMark-ES?

2. Relevante Methodische Ansätze

Ablauf der Genvorhersage:

Zunächst misst die Pipeline Prothint die Protein hints:

1.) GeneMark-ES: Vorhersage von Proteinen (seed proteins) und Vorhersage von Proteinregionen (seed regions) aus dem Genom

2.) DIAMOND: sucht in einer Proteindatenbank nach möglichen Homologen zu den seed proteins

3.) Spaln: Spliced Alignment (Exons, Introns, Start- und Stoppcodons)

Danach trifft BRAKER2 die Genvorhersagen. BRAKER2 ist eine Pipeline, die GeneMark-EP+ und AUGUSTUS miteinander verbindet. BRAKER2 arbeitet in zwei Iterationen:

a.) GeneMark-EP+: Nutzt Proteindatenbanken und nutzt Proteine mit einer beliebig großen taxonomischen Entfernung, um Genvorhersagen mit den Hints aus Prothint zu treffen

- Es werden anchored genes selektiert:
  - Multi-Exon-Gene, die alle Introns aufweisen
  - Single-Exon-Gene, deren Start- und Stoppcodon mit den Proteinhints übereinstimmt

b.) AUGUSTUS: trifft die Genvorhersage mit möglichen Isoformen aus dem Set der anchored genes

3. Relevante Ergebnisse

In einigen Fällen war BRAKER2, welches Proteindatenbanken zur Annotation benutzt, genauer bei der Genvorhersage als BRAKER1, welches Spezies-spezifische RNA-Seq-Daten benutzt.

Die BRAKER2 Pipeline liefert ein genaueres Ergebnis für die Genvorhersage für die Genome *A. thaliana, C. elegans und D. melanogaster* als MAKER2.

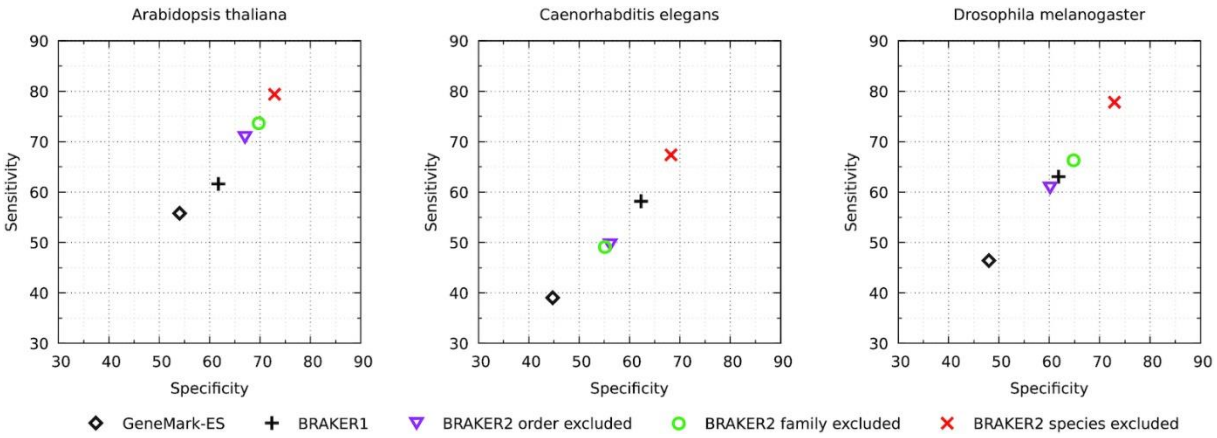
Die Genlänge hat keine Auswirkung auf die Sensitivität und Spezifität der Exon level, die mit BRAKER2 bestimmt werden. Die Sensitivität der Genlevel hingegen nimmt mit der Länge des Genoms ab.

Die Genauigkeit von BRAKER2 nimmt zu, wenn die Anzahl der Spezies, dessen Proteine genutzt wurden, um Proteinhints zu erstellen, höher ist. Außerdem nimmt die Genauigkeit zu, wenn ein naher Verwandter der untersuchten Spezies im Proteinset enthalten ist.

4. Schlussfolgerung

BRAKER2 ist eine Pipeline, die Genvorhersagen in Eukaryoten mithilfe von Proteindatenbanken durchführt, um über Genstrukturen aufzuklären. Die Pipeline funktioniert sehr genau und besser als einige bereits vorhandene Pipelines (MAKER2).

5. Schlüsselabbildung



Als Schlüsselabbildung habe ich die Abbildung 4 ausgesucht. Hier wurde die Genauigkeit der Genvorhersage von BRAKER2, welches Proteindaten verwendet, im Vergleich zu BRAKER1, welches RNA-Seq-Daten verwendet, gezeigt. Es zeigt die Sensitivität und Spezifität der Genlevel für drei verschiedene Arten. Es ist zu sehen, dass BRAKER2 bei *Arabidopsis thaliana* ein deutlich genaueres Ergebnis liefert als BRAKER1. Grund dafür kann sein, dass man mit RNA-Seq nur die Lokalisierung der Introns bekommt, während die Proteindatenbanken die genauen C- und N-Terminus der Proteine liefern. So kann besser gemappt werden, wo die proteincodierenden Regionen beginnen und enden. Bei *C. elegans* und *D. melogaster* ist BRAKER1 genauer als BRAKER2, wenn das Proteinset die Ordnung und Familie der Art nicht enthält. Sobald Proteine aus der gleichen Familie oder Ordnung nicht im Genset vorhanden sind, kann BRAKER1 bei der Genvorhersage besser abschneiden. Diese Abbildung habe ich deswegen ausgewählt, weil sie von allen Abbildungen die meisten Informationen liefert.

Hania Ahmed

BRAKER2 Paper Impact

Aufgabe 1: Die fünf relevantesten Referenzen des Papers BRAKER2 (Bruna et al. 2021)

Titel	Autoren	Journal	Ausgabe	Seitenzahl	Erscheinungsjahr	Zitationen[3]	Warum ist die Referenz relevant?
BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS	Katharina J. Hoff[1], Mark Borodovsky[*], Mario Stanke[2]	*Bioinformatics*,	Vol. 32, Issue 5	S. 767-769	2015	671	Das Paper baut auf diese Referenz auf und ist ein Ausbau der BRAKER1 Pipeline aus dieser Referenz. Beide Pipelines werden auch im Paper miteinander verglichen.
GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins	Tomas Bruna[1], Mark Borodovsk[2*],	*NAR Genomics and Bioinformatics*	Vol. 2, Issue 2	-	2020	67	Die Referenz wurde ausgewählt, weil GeneMark-EP+ ein Teil dieser Pipeline ist und in diesem Paper vorgestellt wird.
Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources	Mario Stanke[1*], Stephan Waack[2]	*BMC bioinformatics*	Vol. 7, Issue 1	S. 1-11	2006	827	Diese Referenz benutzt AUGUSTUS, einen Teil der BRAKER2 Pipeline, und zeigt, wie die Gene für AUGUSTUS ausgewählt werden. Diesen Ansatz benutzt auch unser Paper.
MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects	Carson Holt[1], Mark Yandell[2*],	*BMC bioinformatics*	Vol. 12, Issue 1	S. 1-14	2011	911	BRAKER2 wird im Paper mit MAKER2 verglichen, um aufzuzeigen, welche Pipeline bessere Genvorhersagen trifft.
BUSCO: Assessing Genome Assembly and Annotation Completeness,	Mathieu Seppey[1], Evgeny M. Zdobnov[2*],	*Gene prediction*	-	*S.* 227-245	2019	582	Diese Referenz wurde benutzt, um eine Genauigkeitsanalyse für die vorhergesagten Gene durchzuführen.

[1]Erstautor, [2]Letzter Autor, *korrespondierender Autor, [3]Quelle: Google Scholar

Aufgabe 2: Zitationsanalyse BRAKER2 (Bruna et al. 2021)

Anzahl Zitationen: 144 (2021: 79, 2022: 65)[3]
Einflussreiche Studien, die das Paper zitieren:
1. Haplotype-resolved genome assembly enables gene discovery in the red palm weevil *Rhynchophorus ferrugineus* (Dias et al. 2021, Sci Rep Nature)
2. Underwater CAM photosynthesis elucidated by *Isoetes* genome (Wickell et al. 2021, Nature Communications)
3. The dark proteome: translation from noncanonical open reading frames (Wright et al. 2022, Trends Cell Biol.)
4. FINDER: an automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences (Banerjee et al. 2021, BMC Bioinformatics)
5. Molecular Characterization of *Candida auris* Isolates at a Major Tertiary Care Center in Lebanon. (Reslan et al. 2022, Front Microbiol.)

Aufgabe 3: Korrespondierender Autor des Papers BRAKER2: Mark Borodovsky

Relevanteste Publikationen von Mark Borodovsky	**Begründung**
The complete genome sequence of the gastric pathogen Helicobacter pylori	Dieses Paper wurde in Nature veröffentlicht und 4222-mal zitiert, auch wenn der Autor relativ am Ende der Autorenliste erwähnt wird, ist das eine der relevantesten Publikationen des Autors.
Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii	Dieses Paper wurde in Science veröffentlicht und 3260-mal zitiert, daher kann man es eine relevante Publikation nennen.
GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions	Dieses Paper wurde 1827-mal zitiert und Borodovsky steht hier an letzter Stelle der Autorenliste.
GeneMark. hmm: new solutions for gene finding	Dieses Paper wurde auch sehr oft zitiert (1851) und der Autor steht an zweiter und letzter Stelle.
GENMARK: parallel gene recognition for both DNA strands	Das Paper wurde oft zitiert (897). Außerdem ist er der Hauptautor. Das Paper ist auch inhaltlich relevant und die zukünftigen Publikationen des Autors bauen darauf auf.

Group: Rutian Zhou, Christian Ickes

MetaEuk - sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics

from

Eli Levy Karin, Milot Mirdita, Johannes Söding

Abstract 1

Metagenomics is the study of genetic materials that are extracted directly from samples taken from the environment without the need for prior cultivation. Unicellular eukaryotes are present in almost all environments and due to their vast diversity, still hold invaluable secrets for biotechnology and biomedicine. Gene prediction in eukaryotes with high accuracy acquires large enough genomic coverage and enough reference genomes. To date, two types of information will be considered when training models: intrinsic sequence signals (e.g., CpG Islands) and extrinsic data, such as transcriptomics or an annotated genome from a closely related organism. However, the metagenomic data is severely limited and the phylogenetically related organisms are usually not available, which aggravate the trained annotation models. Here they showed that the open-source software MetaEuk enables large-scale eukaryotic metagenomics through reference-based, sensitive taxonomic and functional annotation. They evaluated against the UniRef90 database seven annotated unicellular eukaryotic organisms from various categories with an average run time of 42min per genome. The gene predictions on the benchmarks covered the majority (77-91%) of exons annotated proteins and has very high precision (>99.9%). Furthermore, they generated two resources for the analysis of eukaryotes: a protein profile database and the MetaEuk marine protein collection which is an expanded dataset after approaching MetaEuk on the Tara Oceans project. Their results demonstrate MetaEuk is a sensitive reference-based algorithm for large-scale discovery of protein-coding genes in eukaryotic metagenomic data. Applying MetaEuk to large metagenomic data sets is expected to significantly enrich databases with highly diverged eukaryotic protein-coding genes. They anticipated their algorithm could improve sequence homology searches, protein profile computation, functional annotation and even protein structure prediction. These, in turn will allow for further exploration of eukaryotic activity in various environments.

Abstract 2

Metagenomic is the study of genetic material derived from environmental samples. Studying metagenomics is improving ecology, biotechnology, pharmacy, and human health since many organisms and organism behavior cannot be reproduced in cell culture. Identifying genes in metagenomic data, raises the knowledge about the number of species, their roles, and the relationships between them in Low sequence coverage, a high variety of genomes and gene structures, as well as missing experimental procedures increase the complexity. However, computational procedures for this problem are needed since current algorithms rely on well-designed training data for identifying genes in genomes, that are not capable of low-quality data. Here they developed a high-throughput algorithm based on references that is suitable for identifying protein-coding genes in eukaryotic, metagenomic data. The shown algorithm can find multi-exon proteins in low-quality data by preforming a sensitive database search. A benchmark data set, including seven diverse eukaryotic organisms, revealed an astonishing performance even under difficult conditions. Furthermore, the Tara Oceans data set got studied by using MetaEuk which predicted more than 12,000,000 different protein-coding genes in 912 samples. Their results demonstrated that MetaEuk is highly recommended for exploring eukaryotic metagenomic data at the genome level. Metagenomic projects will benefit from this algorithm due to its high accuracy and sensitivity. Taxonomic assignments, gene calling, and classification concerning the roles of organisms will be more precise if this approach is used. So could MetaEuk reveal new important facts when applied to metagenomic projects.

MetaEuk Summary

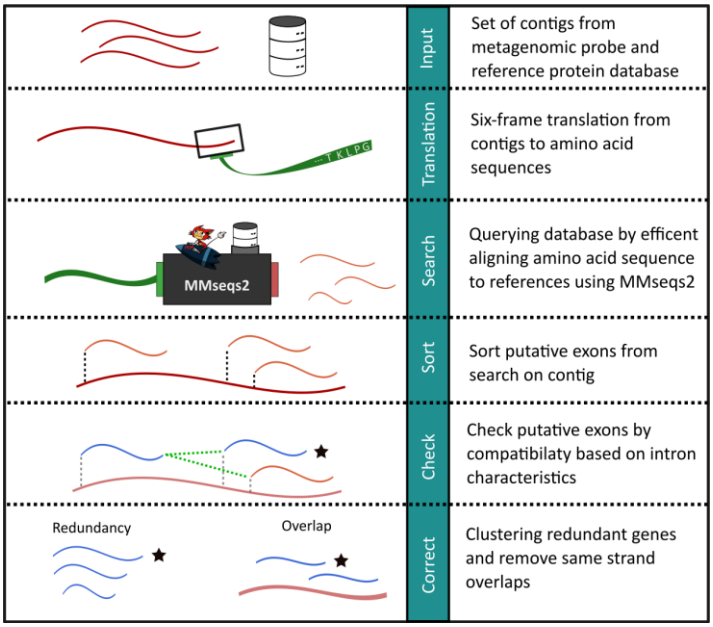
Research question
How to sensitively and possibly efficiently identify single- and multi exon protein-coding genes in eukaryotic metagenomic data?

Relevant Methods
- MMseq2 / Dynamic Programming for speedup
- Redundancy Reduction by clustering gene calls
- Scoring System using Bit-Score and E-value
- Protein-profile reference database

Relevant Results
- Software / Algorithm – high performance with respect to speed and sensitivity
- Marine Protein Collection databases based on the Tara Ocean project

Conclusion
MetaEuk is a fast, sensitive tool which predicts protein-coding genes in contigs based on reference database. Its key ability aims on the prediction of genes in metagenomic data which has to deal with low coverage, chimeric contigs and high numbers of sequencing errors. It could play a positive role not only in enriching the databases, but also in improving homology-based function annotation and even in predicting de novo protein structure.

Keyfigure
(Figure 1 – own creation)



Description:
Displays the procedure of MetaEuk. Using contigs and a reference database, MetaEuk predicts putative exons by querying the reference database. An optimal set of exons per contig is generated and evaluated.

Reason:
For the reason that the paper aims to introduce the MetaEuk toolkit, figure 1 plays the role of a facilitated access to not only the algorithm but also the understanding of the results, which is a crucial factor for evaluating the toolkit.

1. 5 Most important references

KEYWORD	AUTHORS	TITLE	JOURNAL	PUBLISCHED IN	CITATIONS (PUBMED)	REASON
MMSEQ2	M. Steinegger, J. Söding	"MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets"	Nature Biotechnology	2017	308	Leads to fast alignments against reference database.
REFERENCE DATABASE	The UniProt Consortium	„UniProt: the universal protein knowledgebase"	Nucleic Acids Research	2017	1812	Database for benchmarking.
BENCHMARK DATABASE	D.A. Benson, E.W. Sayers, et al	„GenBank"	Nucleic Acids Research	2018	192	Database for assembly data
BIT-SCORE	S. Karlin, S. F. Altschul	"Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes"	PNAS*	1990	353	Basis of call evaluation
OVERVIEW EUKARYOTIC GENOM	A. Kumar	"An overview of nested genes in eukaryotic genomes"	Eukaryot Cell		32	Resolves the contradicting predictions.

2. 5 Most influential studies based on the paper

Number of citationen according to PubMed (2020-2022): 15
Citations per year: 15/3 = 5

Studies**: Title (Journal: Impact Factor JIF)

-- "MetaPlatanus: a metagenome assembler that combines long-range sequence links and species-specific features" (Nucleic Acids Research: 16971)
-- "Genomic and metabolic adaptations of biofilms to ecological windows of opportunity in glacier-fed streams" (Nature Communications: 14447)
-- "Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC" (Genome Biology: 13583)
-- "BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes" (Molecular Biology and Evolution: 11062)
-- "Using metagenomic data to boost protein structure prediction and discovery" (Computational and structural Biotechnology: 7271)

3. The corresponding author – Johannes Söding (Max-Planck Institute for Biophysical Chemistry, Göttingen, Germany)

Selected Publications**: Title (Journal: Impact Factor JIF)

-- "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets" (Nature Biotechnology: 54904)
-- "Big-data approaches to protein structure prediction" (Science: 47728)
-- "Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold" (Nature Methods: 28467)
-- "A high-throughput screen for transcription activation domains reveals their sequence characteristics and permits reliable prediction by deep learning" (Mol Cell: 17970)
-- "Clustering huge protein sequence sets in linear time" (Nature Communications: 14447)

The Evaluated Paper is published in journal "Microbiome" with JIF 11607.

*Proceedings of the National Academy of Sciences of the United States of America
** Ranked with Journal Impact Factor (JIF).

Group: Kassoum Wonogo, Carina Zschammer

OrthoInspector: comprehensive orthology analysis and visual exploration

from

Benjamin Linard*, Julie D Thompson, Oliver Poch, Odile Lecompte

Abstract 1

Homologous genes are described as biologically similar in the sense of sharing a common ancestry in the evolutionary history, originating from speciation events (orthologs) and gene duplication events (paralogs). Due to the rapidly growing amount of sequencing data, sophisticated methods and software applications for the detection of orthology and inparalogy are necessary. Existing software tools are available to predict orthology/inparalogy relationships between different proteomes, with results stored in simple databases or flat files. However, many of these applications require computer expertise to perform large-scale queries and are limited in intuitiveness and flexibility, which makes them inaccessible to non-specialists. Here, the authors developed a stand-alone software application using a novel algorithm called OrthoInspector for the rapid detection of orthology and inparalogy. A comprehensive large-scale proteome analysis with 59 different eukaryotic species was performed. In an example test case OrthoInspector was used to predict orthology in proteins of the myotubularin family with no false-negative results compared to other software tools. Additionally, a benchmarking with different existing methods showed that Orthoinspector improves detection sensitivity with a minimal loss of specificity for the inference of orthology. The results demonstrate the functionality and robustness of OrthoInspector in comparison to other existing methods. The development of OrthoInspector showed that a broad and generalized software that reliably predicts orthology and inparalogy relationships could be achieved. The quality of the results is mostly as good or better than the best currently available methods. The inclusion of a graphical interface and representation of the results makes OrthoInspector extremely intuitive and user friendly.

Abstract 2

Due to the continuous improvement of today's sequencing methods, the amount of available protein sequence data continues to increase. In order to efficiently handle this amount of data and to accurately predict orthology and inparalogy relationships, methods such as their reconstruction via phylogenetic tree- based inference are no longer sufficient. For evolutionary studies, comparative sequence analysis and functional gene annotation, the recognition of orthologous and inparalogous relationships of sequence data is essential. Therefore, methods for processing sequence data currently need to be permanently developed in order to be able to perform reliable orthology analyses despite the ever-increasing amount of these sequence data.

The graph-based software OrthoInspector developed here is one of these new methods and provides efficient data management using a pairwise distance-based algorithm for orthology and inparalogy prediction between different species.

OrthoInspector was used to validate its functionality and correctness for predicting orthology and inparalogy relationships within 59 eukaryotic organism proteomes. It was also compared to five other existing orthology prediction methods and the algorithmic differences with OrthoInspector, especially the handling of reciprocal best hits, were evaluated. OrthoInspector impressed with improved sensitivity and reliable specificity. With the addition that a combination of the OrthoInspector approach with advantages of other methods could also be beneficial.

Most of the methods already available for detecting orthology relationships can only be used to their full extent by users with extensive programming knowledge. OrthoInspector instead provides two user interfaces: a command-line client and a graphical interface. Thanks to the latter in particular, OrthoInspector reveals as a user-friendly orthology prediction method that can be used effectively not only by specialists but now also by non-specialists.

Paper Summary

Research Question

Can a complete software (OrthoInspector) be created for orthology and inparalogy prediction between different species and their analysis?

Relevant Methodic Approaches

- Implementation of a software system “OrthoInspector“ (for Java-supporting platforms)
- Basis of data processing: Blast-all-versus-all search
- Algorithm based on the principle of calculating pairwise distances
- Three main components of OrthoInspector:
  1. Installation:
     → Creation of the database (which will later be filled with orthologs)
     → Calculation of orthologous/inparalogous groups
     → Optional creation of precalculated data
  2. Queries: Search for orthologous relations in the data (text, batch queries, ...) and export of the results in fasta, csv, xml formats
     → Possible via 2 different user interfaces:
     Command line version: Fast information retrieval for high throughput studies; using the OI software in other packages
     Graphical interface (see component 3): Can be used for all steps of the analysis process; visualization is not possible with command line version
  3. Data visualization: For more detailed analysis
     → Performing more complex queries is possible by manually selecting data to be included or not included in the analysis
     → Presentation of results cross-referenced to other databases
     → Visualized summary of the data (e.g. as a graph)
- Analysis by OrthoInspector in three main steps:
  1. **Results of a blast-all-versus-all search as input**
     → Based on this forming of inparalog groups + **validation** (validation of such a group if the blast output for all proteins of a group to be validated each also leads to this grouping)

2. **pairwise comparison of the inparalogous groups**: Use of blast best hits to define the potential relationships between inparalogous groups
   → *1-to-1-relationship*: Reciprocal best hit (RBH) (between a protein from organism 1 and a protein from organism 2 and back)
   → *1-to-many-relationship*: Best hit between a protein from organism 1 to any protein of an inparalog group from organism 2; and best hit back from any protein in the inparalog group from organism 2 to the particular protein from organism 1
   → *Many-to-many-relationship*: Best hit between any proteins of two inparalogous groups from organism 1 and organism 2, respectively
3. **Recognition of contradictory information**

Relevant Results

- Compared to other existing methods OrthoInspector has improved detection sensitivity with only a minimal loss of specificity
- Evidence for recognition of orthologous/inparalogous relationships by OrthoInspector: description of a large-scale proteomic analysis with 59 organisms
  → Result: number of inparalogous groups generally correlates with proteome size
- Comparison of OrthoInspector with other existing methods of ortholog prediction - **algorithmic differences**
  - Comparison with “Inparanoid“ und “OrthoMCL“ algorithm: both use RBHs as a basic condition for the detection of potential inparalogous groups → partially false-negative results
    → OrthoInspector does not use purely RBHs as a basis, but derives inparalogy groups in each organism directly; followed by the second step of pairwise comparisons of inparalog groups
    → Using both RBH and BH does not provide false-negative results
- Comparison of OrthoInspector with other existing methods of ortholog prediction - **via benchmark datasets**
  - Comparisons between OMA, Inparanoid, OrthoMCL, Ensembl Compara, eggNGO and OrthoInspector
    - 4 benchmarks: TKL & CMGC kinases benchmarks + 2 literature benchmarks
    - subsequent division of the six methods into two groups:
      - OrthoMCL, Ensembl compara, eggNOG: higher sensitivity than specificity
      - Inparanoid, **OrthoInspector** and OMA: **higher specificity than sensitivity**

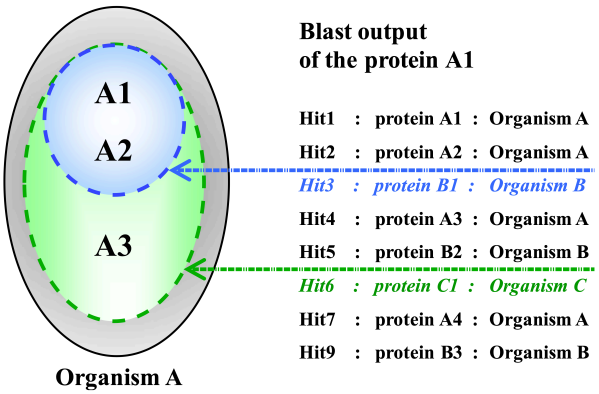
- Interpretation of the results: the six methods tested here provide complementary approaches to orthology inference
  → Advantages of different alternatives should be combined
  → e.g., OrthoInspector as a starting tool for deriving orthology relationships, as its sensitivity and specificity are well balanced compared to most other methods used here; in a subsequent step, the user could deal with true/false positives from the different methods by integrating and refining them all into each other

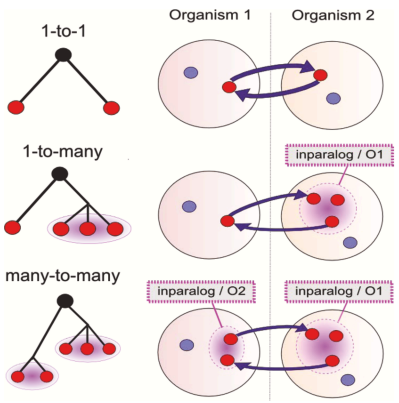
Conclusion

With the development of OrthoInspector, the research question of creating a software that reliably predicts orthology and inparalogy relationships between different proteomes could be answered positively. Problems concerning the runtime - especially when subsequent updates are made - are known and concrete approaches for their improvement are already being planned.
Overall, the user-friendly OrthoInspector software provides reliable, safe predictions, while it could probably achieve even better results in combination with other, already existing methods for ortholog prediction.

Key Figure

In our opinion there is no clear key illustration. Since the main topic of the paper is the software "OrthoInspector" itself, the key figure should best represent an overview of the structure or the procedure of OrthoInspector.
Figure 1 shows the basic structure of OrthoInspector including the different analysis possibilities between command line and graphical version, but this figure is not very intuitive if the principle of OrthoInspector itself is not exactly known. In principle a graphical representation of the procedure for the formation of inparalogous groups and their pairwise comparisons would be more interesting. This is the main principle of OrthoInspector, which distinguishes it from other tools. A clear, easy-to-understand combination of Figures 2 and 3 would probably make the most sense as a key figure.





Paper impact - OrthoInspector (OI)

Kassoum Wonogo, Carina Zschammer

Most relevant references

Authors (first, last)	Title	Journal	Year	Citations	Relevance
Eugene Koonin (only him)	Orthologs, paralogs, and evolutionary genomics	Annual Review of Genetics (39, 309-338)	2005	486	Good biological basis: technical terms used in the OI paper are defined more precisely and the (biological) motivation for why the identification of orthologs and paralogs is so important, especially for evolutionary studies, is provided.
Mark E. Peterson, Andrej Sali*	Evolutionary constraints on structural similarity in orthologs and paralogs	Protein Science (18, 1306-1315)	2009	32	Classification of the effect of orthology on the relationship between protein sequence and structure in the background of evolution. Emphasizes the relevance of detecting orthologies.
Arnold Kuzniar, Jack A.M. Leunissen*	The quest for orthologs: finding the corresponding gene across genomes	Trends in Genetics (24, 539-551)	2008	124	Reasons why nowadays many (bioinformatic) approaches for finding orthology relationships are no longer sufficient. Description of concrete alternative algorithmic approaches, some of which have also been used in the development of OI.
Gabriel Östlund*, Erik L. L. Sonnhammer	InParanoid 7: new algorithms and tools for eukaryotic orthology analysis	Nucleic Acids Research (38, D196-D203)	2009	340	OI is compared with the orthology prediction tool InParanoid (and 4 others). Mainly based on these comparisons, the conclusions about dis- and advantages of OI are given. InParanoid is most similar to OI, as both use pairwise distance calculations as the basis of their algorithms; InParanoid is referred to most strongly of all the comparison tools.
Andrey Alexeyenko, Erik L. L. Sonnhammer*	Overview and comparison of ortholog databases	Drug Discovery Today: Technologies (3, 137-143)	2006	15	Comparison of different Ortholog databases already existing at the time of publication. Three of the five tools with which OI is also compared are also discussed in this paper.

Most influential studies citing the paper - Total number of citations: 37 (PubMed), Citations per year: 3

Title	Citations	Journal	Impact factor
OrthoFinder: Phylogenetic orthology inference for comparative genomics (2019)	596	Genome Biology	13.583
eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges (2012)	267	Nucleic Acids Research	16.971
An integrative approach to ortholog prediction for disease-focused and other functional studies (2011)	294	BMC Bioinformatics	3.169
A New Comparative Genomic Analysis of Human and *Caenorhabditis elegans* Genes (2018)	81	Genetics	6.150
OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs (2012)	166	Nucleic Acids Research	16.971

Corresponding author: Benjamin Linard - 25 publications (PubMed, 2010-2022)

Title	Citations	Journal	Impact factor	Author position	Relevance
Standardized benchmarking in the quest for orthologs (2016)	78	Nature Methods	28.467	8/28	Publication with the most citations.
Controversies in modern evolutionary biology: the imperative for error detection and quality control (2012)	16	BMC Genomics	3.969	2/5	Highlighting of the relevance of improving the quality of sequencing process results (impact on subsequent evolutionary studies).
Big data and other challenges in the quest for orthologs (2014)	62	Bioinformatics	6.937	Part of Consortium	Work as part of the "Quest for orthologs" Consortium. Results of their third meeting.
A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives	71	PLOS One (2011)	3.24	2/4	Second most citations. Improved alignment design by connecting existing methods and suggesting future improvements.
Rapid alignment-free phylogenetic identification of metagenomic sequences (2019)	7	Bioinformatics	6.937	1/3	First author*. Development of a k-mere based database: precise assignment of sequences to their most probable origin in phylogenetic trees is possible.

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

by

Sabrina Krakau, Hugues Richard and Annalisa Marsico

Abstract 1

Interactions between proteins and RNA molecules are vital for various regulatory processes, especially for those controlling the transcription and post-transcriptional regulation of genes. The landscape analysis of these interactions is most commonly achieved by a UV induced crosslinking of RNA binding proteins (RBPs) to the RNA, followed by an immunoprecipitation (IP) to pull down specific RBP-RNA complexes. The resulting data can be used to study these interactions based on the enrichment of aligned reads in a certain area (peak-calling) and the accumulation of read start sites. While there are many tools capable of performing such analyses, there exists no method yet which is able to simultaneously call peaks and detect crosslink sites, while also taking into account biases like transcript abundance and non-specific crosslinking motifs. This paper introduces PureCLIP, a tool capturing protein-RNA interactions from single nucleotide resolution crosslinking immunoprecipitation (CLIP) data using a non-homogenous hidden Markov model (HMM). PureCLIP is able to find specific crosslink sites based on read enrichment and read start accumulation with a higher precision than previous methods. Given an input signal, it successfully distinguishes real peaks from regions with a strong background signal. Incorporating biases such as this, or known non-specific crosslink motifs, PureCLIP reduces the amount of false positives found compared to other methods. The ever-increasing amount of CLIP data calls for methods to analyze protein-RNA interaction landscapes with high precision. PureCLIP accomplishes this using multiple factors to stabilize its results. With the help of this tool, gathered knowledge regarding crosslink motifs can finally be incorporated into the analyses of protein-RNA interactions. This could lend a hand in the discovery of various regulatory mechanisms for transcription as well as the post-transcriptional processing of RNAs.

Abstract 2

iCLIP and eCLIP are methods for identifying protein-RNA interactions by detecting cross-linking sites caused by UV light covalently binding proteins and RNA molecules. Both methods provide single nucleotide resolution by detecting truncated cDNA, but eClip additionally improves the specificity of the proclaimed binding regions. It is still critical to consider various sources of bias, such as transcript abundance, preferences for cross-linking sequences, and mappability. Bias in the form of sequence preferences at crosslink sites can be systematically captured and used for correction in the form of crosslink-associated (CL) motifs, certain k-mers enriched in both input and target eCLIP data at read start sites. Known background binding regions are used as to validate called binding sites and lower false positives. However previous computational analysis methods for CLIP-seq data do not fully account for said sources of biases. Here we show that our hidden Markov model-based approach called PureCLIP, which simultaneously performs peak-calling and individual crosslink site detection and corrects biases by incorporating a control experiment and non-specific sequence biases, is more accurate than previous methods in determining crosslinking sites and features higher inter-repeat agreement. We have designed a realistic iCLIP/eCLIP simulation framework and demonstrated on a wide range of simulation parameters that PureCLIP is up to 7-15% more precise than other methods in detecting detect target-specific crosslinking sites and determining bona fide binding site locations. This was consistently validated using four datasets of published iCLIP/eCLIP data where the RBP motif or predominant binding region of RBP is known. Notably, the inclusion of covariates, such as the input signal and CL motifs, increases the precision of PureCLIP by up to 8-10% compared to prior methods. Our results show that PureCLIP is able to improve upon the limitations posed by high false positive rates of existing methods by incorporating control data and known CL motifs. While high-resolution CLIP-seq datasets are becoming more and more commonplace, the accurate determination of protein- RNA interaction sites from iCLIP/eCLIP data has still been a challenge. Our comprehensive evaluations have demonstrated that PureCLIP outperforms various prior approaches in the detection of target- specific cross-linking sites, across both simulated data and real datasets. It is capable of accurately detecting protein-RNA interaction footprints avoiding exclusive reliance on the highest peaks, and is able to correct for biases such as transcript abundances, background binding, and preferences for cross-linking sequences. As a result, it represents a valuable and promising method for the analysis of these data sets, but also for proteins with lower binding affinity or those that bind to low-abundant RNAs, such as lncRNAs.

Group: Jonas Busam, Jens Mayer

Paper Summary

1. Research Questions:

How can protein-RNA interactions be identified from single nucleotide resolution CLIP experiments with high precision, while correcting for biases?

2. Relevant Methods:

- non-homogenous hidden Markov Model to incorporate
    - known cross-link motifs
    - non-specific background signal
- comparison to other CLIP-seq analysis tools
- simulation of realistic iCLIP/e-CLIP –seq datasets including background noise
- evaluation on real datasets with known binding regions

3. Relevant Results:

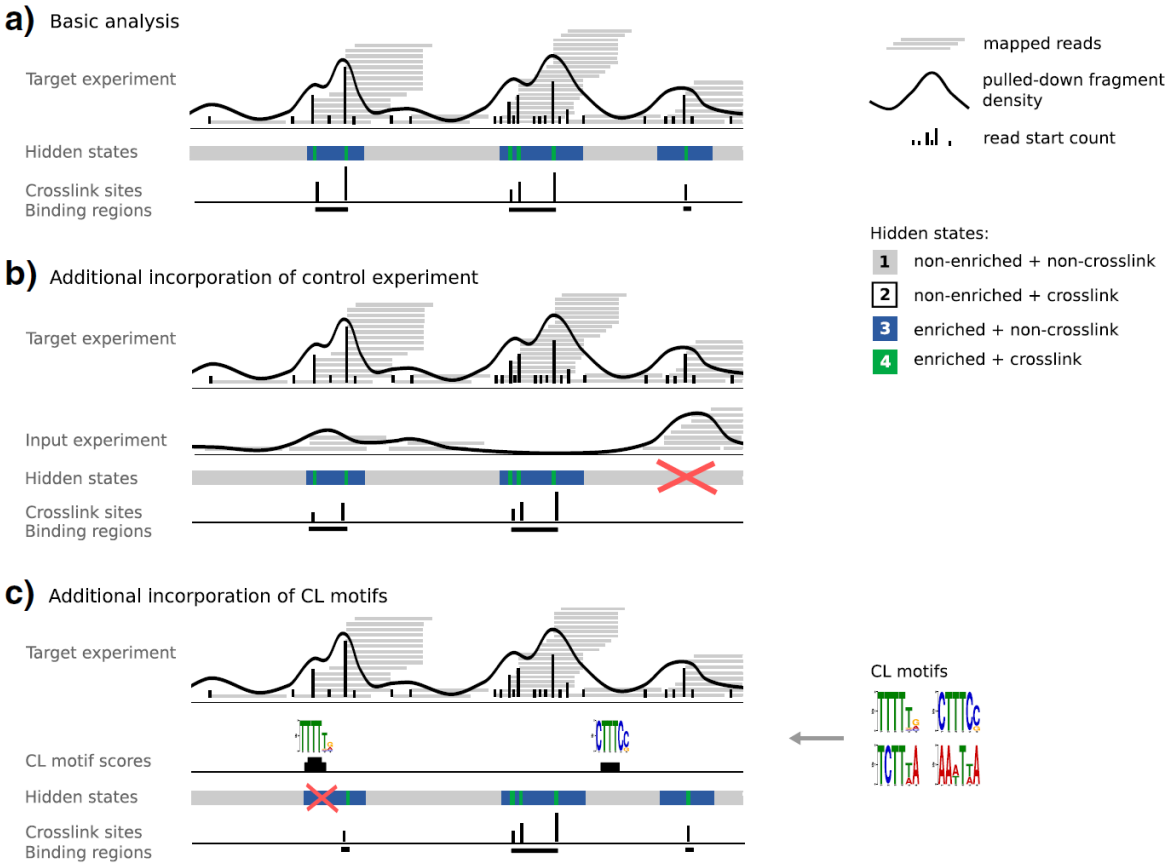
PureCLIP is able to capture protein-RNA interaction footprints from single nucleotide resolution CLIP data. It outperforms several other similar tools in terms of precision when detecting target-specific cross-link sites as well as replicate agreement. PureCLIP succeeds in correcting for false positives while maintaining a high accuracy.

4. Conclusion:

PureCLIP represents a promising new approach to optimize the performance of the computational analysis of CLIP-seq data, especially iCLIP/eCLIP. It is also suited for detecting proteins with low binding affinities or RNAs with lower read counts. In the future, more validation could be done to assure the wide applicability of the tool and the reproducibility of its results.

5. Key Figure:

Figure 1 is the key figure, because it gives a brief overview of the workflow, showing the import steps the tool undergoes in determining peaks and crosslink sites. It also shows the actions PureCLIP takes to account for sequence biases, such as the incorporation of control experiments and cross-link motifs.



Paper Impact

Group: Jonas Busam, Jens Mayer

Most important references of the publication

1. Yoichiro Sugimoto, Jernej Ule,
   Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions,
   Genome Biology 13 (2012, Aug 3), R67, 2012, 155 citations
   We chose this paper, because it contains important basic information about CLIP data and what to consider in their analysis.

2. Eric L Van Nostrand, Gene W Yeo,
   Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP),
   Nature Methods 13, 508–514, 2016, 550 citations
   We chose this paper, because it represents a state-of-the-art approach to generate single nucleotide resolution CLIP data.

3. T.L. Bailey,
   DREME: motif discovery in transcription factor ChIP-seq data,
   Bioinformatics Vol. 27, Issue 12, Pages 1653–1659, 2011, 801 citations
   We chose this paper, because it is an essential part of the methods and key to what sets PureCLIP apart from similar approaches, the incorporation of known CL motifs.

4. P. H. Reyes-Herrera, S. Herrera,
   BackCLIP: a tool to identify common background presence in PAR-CLIP datasets
   Bioinformatics, Vol. 31, Issue 22, Pages 3703–3705, 2015, 9 citations
   We chose this paper, because it provides a set of background signals, which PureCLIP uses as another means to correct biases.

5. L.R. Rabiner,
   A tutorial on hidden Markov models and selected applications in speech recognition
   Proceedings of the IEEE, vol. 77.2, pp. 257-286, 1989, 13153 citations
   We chose this paper, because it talks about essential methods of generating and using hidden-Markov models, which PureCLIP uses as a basis.

Most important citations (in total 34 citations (CrossRef) (Ø 6.8/year))

1. Xiaoyong Pan et al., Recent methodology progress of deep learning for RNA–protein interaction prediction,
   WIREs RNA (IF: 9.96), 2019, doi:10.1002/wrna.1544
2. Ping Xuan et al., Graph Convolutional Network and Convolutional Neural Network Based Method for Predicting lncRNA-Disease Associations,
   Cells (IF: 6.6), 2019, doi:10.3390/cells8091012
3. Nadine Körtel et al., Deep and accurate detection of m6A RNA modifications using miCLIP2 and m6Aboost machine learning,
   Nucleic Acids Research (IF: 16.97), 2021, doi:10.1093/nar/gkab485
4. Mahsa Ghanbari et al., Deep neural networks for interpreting RNA-binding protein target preferences
   Genome Research (IF: 9.04), 2020, doi:10.1101/gr.247494.118
5. Evgenia Ntini et al., Functional impacts of non-coding RNA processing on enhancer activity and target gene expression
   Journal of Mol. Cell Biology (IF: 6.22), 2019, doi:10.1093/jmcb/mjz047

Most relevant publications of corresponding author

1. Annalisa Marsico et al, PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs, Genome Biology, 2013, doi:10.1186/gb-2013-14-8-r84 **(First author, Citations: 81, IF: 13.58)**
2. Sabrina Krakau wt al., PureCLIP: capturing target-specific protein–RNA interaction footprints from single-nucleotide CLIP-seq data, Genome Biology, 2017, doi:10.1186/s13059-017-1364-2 **(Last author, Citations: 34, IF: 13.58)**
3. Stefan Budach et al., pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks, Bioinformatics, 2018, doi:10.1093/bioinformatics/bty222 **(Last author, Citations: 56, IF: 6.95)**
4. Marsico et al., A novel pattern recognition algorithm to classify membrane protein unfolding pathways with high-throughput single-molecule force spectroscopy, Annalisa Bioinformatics, 2007, doi:10.1093/bioinformatics/btl293 **(First author, Citations: 25, IF: 6.95)**
5. Thomas Conrad et al., Microprocessor Activity Controls Differential miRNA Biogenesis In Vivo, Cell Reports, 2014, doi:10.1016/j.celrep.2014.09.007 **(Second author, Citations: 54, IF: 9.42)**

Group: Annika Bamberger, Ann-Kathrin Otto

A Pan-plant Protein Complex Map Reveals Deep Conservation and Novel Assemblies

by

Claire D. McWhite, Ophelia Papoulas, Kevin Drew, et al.

Abstract 1

Plants are crucial for sustaining global economic and environmental systems. Therefore, it is important to understand the functions of plant genes and proteins. This understanding can be gained by determining protein-protein interactions, which reveal information about the phenotypes of genes and the functions of proteins and protein complexes. This information can then provide opportunities to study and manipulate critical cellular processes. However, because plant genomes are usually complex and polyploid, most gene and protein functions of plants are unknown. In this work, we constructed a protein complex map of 13 diverse green plant species, providing a global overview of the organization, functions, and interactions of plant proteins. Most of our observations coincide with known proteome characteristics of plants or other organisms, including a correlation between protein abundance and RNA transcript levels, and highly conserved protein complexes. However, we also found out that in some cases, plants show alternative multiprotein assemblies for homologous gene products. That indicates that sequence homology alone does not always suffice to predict the structure and function of a protein complex. In general, we provided a meaningful snapshot of the conserved and expressed proteome of plants, using co-fractionation mass spectrometry and orthogroup-based proteomics. The dataset and the results of this work can be used for further research across the vast landscape of plant biology. Possible applications could be fundamental plant research, research on pathogen resistance and plant health, and biofuel production.

Group: Annika Bamberger, Ann-Kathrin Otto

A Pan-plant Protein Complex Map Reveals Deep Conservation and Novel Assemblies

by

Claire D. McWhite, Ophelia Papoulas, Kevin Drew, et al.

Abstract 2

Identifying protein complexes is important for understanding the organization and interaction of cellular systems. Protein interaction networks as well as the systematic mapping of protein complexes are essential for discovering protein functions, characterizing proteins and understanding disease-related pathways. Plant proteins are like proteins in general the basis of processes allowing organism to function. However, a majority of plant proteins are still uncharacterized and a systematic way of defining multiprotein assemblies in plant cells has not been done yet. Here we selected a set of 13 diverse plant species from which we recovered known protein complexes and identified novel complexes conserved for over a billion years. The number of already known stable protein complexes in plants was significantly increased by our study. Many protein complexes that have only been derived from gene contents in the past could be observed here. Furthermore, we showed that protein sequence homology alone does not indicate an identical protein assembly. Our resulting protein complex map provides an overview of conserved, stable protein complexes and a general overview of protein organization in plants. It brings the opportunity to interpret plant genetics and mutant phenotypes. Our large and diverse proteomics dataset was used to show how to connect gene products with phenotypes, as well as testing specific functional hypotheses. This dataset can be the foundation of various other research questions in the field of plant biology.

Group: Annika Bamberger, Ann-Kathrin Otto

Paper Summary

Research questions

- How are proteins organized in plants?
- Which protein complexes are conserved in different plant species?
- How can orthology mapping be conducted for complex and polyploid plant genomes?

Relevant methodic approaches

- Co-fractionation mass spectrometry (CF-MS): high throughput method to detect interacting proteins
    - proteomics interpretation in terms of orthogroups (OGs) for orthology mapping
- Machine learning: to control false discovery rates, to quantitatively score co-elution behaviour

Relevant results

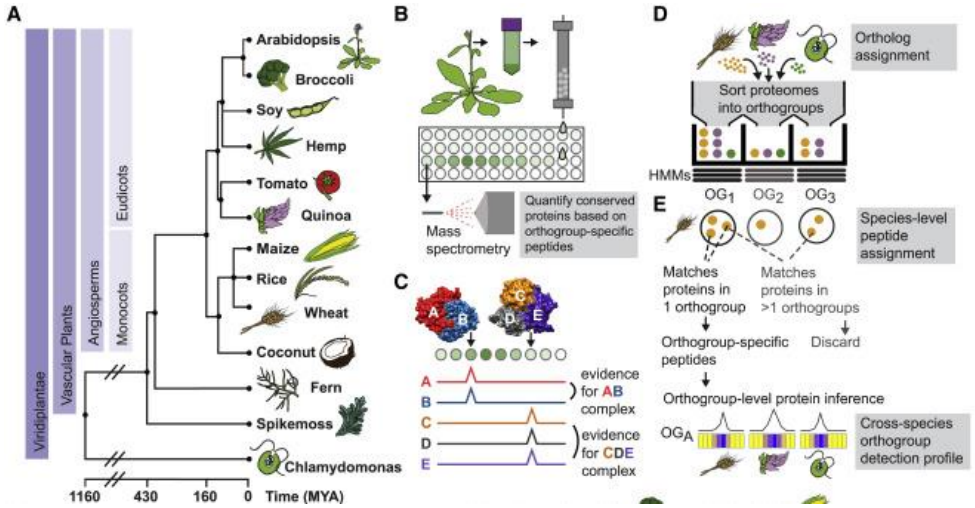
- Correlation between protein abundance and RNA transcript levels
- For some animal complexes, there are analogs in plants: e.g.: tRNA multi-synthetase complex
- Genetic absence of homologs does not necessarily predict the absence of functionally similar complexes: e.g.: proteasome assembly chaperones (plants lack a gene compared to humans but still express protein)
- Adaption of conserved molecular modules with plant-specific proteins: e.g.: NADH dehydrogenase-like complex
- Link between protein interactions and phenotype: multiple complexes are involved in vernalisation and pathogen defense

Conclusion

- The protein complex map constructed by the usage of MS proteomics gives a global snapshot of protein organization in plants.

Group: Annika Bamberger, Ann-Kathrin Otto

Key figure



- The figure shows the selected plant species and how the protein extracts of those species got separated and analyzed via mass spectrometry. It is seen how the mass spectral observations were used to sort the proteomes into orthogroups and create a detection profile.
- This figure is a key figure because it illustrates the main methods that were used and gives an overview of the steps leading to the resulting protein complex map.

Group: Annika Bamberger, Ann-Kathrin Otto

Paper Impact

First author **Corresponding author** **Last author**	Seung Yon Rhee Marek Mutwil Marek Mutwil	Jaime Huerta-Cepas Peer Bork Peer Bork	Cuihong Wan Andrew Emili Andrew Emili	Hunter B Fraser Joshua B Plotkin Hunter B Fraser	Christine Vogel Edward M. Marcotte C Vogel and EM Marcotte
Title	Towards revealing the functions of all genes in plants	Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper	Panorama of ancient metazoan macromolecular complexes	Using protein complexes to predict phenotypic effects of gene mutation	Insights into the regulation of protein abundance from proteomic and transcriptomic analyses
Journal	Trends in Plant Science, 19, Issue 4, 212-221	Molecular Biology and Evolution, 34, Issue 8, 2115-2122	Nature, 525, 339-334	Genome Biology, 8, article no R252	Nature Reviews Genetics, 13, 227-232
Year	2014	2017	2015	2007	2012
Citations	187	1334	411	102	3466
Explanation	This paper declares that network-based function inference is underutilized in plant biology and many plant proteins are uncharacterized. Therefore, this paper shows why our paper, in which protein complexes are determined to help solve this issue, is relevant.	The eggNOG-Mapper presented in this paper was used for the orthogroup assignment in our paper and is therefore an important foundation.	Wan et al. also created a conservation map using co-elution, but for animals. The work of McWhite et al. is based on several methods explained and used in this paper.	The authors state that interacting proteins are more likely to share phenotypes. Some important results of McWhite et al. are based on this statement.	McWhite et al. observed a correlation between protein abundance and RNA transcript levels. Similar correlations have already been observed by Vogel and Marcotte in other organisms, showing the conservation and importance of this correlation.

Group: Annika Bamberger, Ann-Kathrin Otto

The paper has been cited 73 times in total and approximately 33.7 times per year. It has been published on the Journal “Cell”, which has an impact factor of 41.587. We consider the following papers citing McWhite et al. as the most impactful:

- A phase-separated nuclear GBPL circuit controls immunity in plants, Shuai Huang et al., 2021: 16 citations, nature (impact factor ca. 50)
- Metabolons, enzyme-enzyme assemblies that mediate substrate channeling, and their roles in plant metabolism, Youjun Zhang and Alisdair R. Fernie, 2021: 39 citations, Plant Communications (partner journal of CellPress, impact factor unknown)
- Crop biotechnology and the future of food, Michael A. Steinwand and Pamela C. Ronald, 2021: 39 citations, nature
- Meta-analysis defines principles for the design and analysis of co-fractionation mass spectrometry experiments, Michael A. Skinnider and Leonard J. Foster, 2021: 7 citations, nature methods (impact factor ca. 28.5)
- Integrating multi-omics data for crop improvement, Federico Scossa et al., 2021: 25 citations, Journal of Plant Physiology (impact factor ca. 3.5)

Corresponding author: Edward M. Marcotte

1. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses, Vogel and Marcotte, 2012: This paper has the most citations (3466) and he is the last author there.
2. Detecting protein function and protein-protein interactions from genome sequences, Marcotte et al., 1999: This paper has 2000 citations and the Science journal has a high impact factor (41). He is the first author here.
3. A combined algorithm for genome-wide prediction of protein function, Marcotte et al., 1999: This is a nature paper (impact factor 50), has 1000 citations and he is the first author
4. A probabilistic functional network of yeast genes, Lee et al., 2004: This paper also with 1000 citatios a lot of citations and he is the last author.
5. Prioritizing candidate disease genes by network-based boosting of genome-wide association data, Lee et al., 2011: This paper has 700 citations and he is again the last author.

Group: Sam Gimbel, Tabea Merlevede

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

Shuichiro Makigaki and Takashi Ishida
Bioinformatics, 36(1), 2020, 104-111

Abstract 1

Although the experimental methods for determining protein structures evolved, the speed at which amino acid sequences can be resolved heavily outperforms the ability to find the corresponding protein structures. Computational methods, based on machine learning, are now considered to replace conventional structure determination techniques to save time and resources. A common computational method for tertiary structure prediction represents template-based modeling, which uses known homologous protein structure templates and their sequence alignments to the query protein. The underlying theory is that a notable similarity between proteins at their sequence level can lead to the assumption, that the corresponding protein structures are also similar. Sometimes, however, sufficiently accurate structure models cannot be obtained because of poor sequence alignment quality. Researchers tried to improve alignments manually based on their knowledge of biology, but fully automated methods are still missing. Here the autors show a relevant improvement in sequence alignment generation and accuracy of the resulting tertiary structure models based on their supervised machine learning method combined with the conventional dynamic programming approach for sequence alignment. The model accuracy of the alignment generation method has been found to outperform state-of-the-art methods like PSI-BLAST, DELTA-BLAST and HHsearch. They also evaluated their method for homology detection and found that their approach had the lowest detection sensitivity. The authors considered that their method is currently useful for the alignment generation phase of template-based modeling, after template detection. The likely occurence of many false-positive values when trying to detect homologs with their method made the authors discuss, that it may not be suitable for homology detection. For the proposed method, the amount of training data had to be reduced to reach a reasonable runtime. The authors propose more complex machine learning algorithms to increase the accuracy and efficiency of the approach.

Group: Sam Gimbel, Tabea Merlevede

Abstract 2

In Biology, it is crucial to determine the protein structure in order to understand it's function. Evolutionary closely related proteins often share a similar function, making template-based approaches based on sequence alignments to target proteins a valid method to reveal the protein's structure. However, remote homologous proteins that have been used in the tamplate-based approach cause much less accurate structures, even though sensitive detection methods for remote homologs exist. In order to achieve highly accurate structure predictions, high quality sequence alignments and templates, as well as detection methods for remote homologs, are needed. With machine learning methods on the rise, they might be an effective approach for sequence alignment generation based on structural alignments of homologs. Here the authors present a new approach for pairwise sequence alignment generation for template-based modeling based on the k-nearest neighbors algorithm. The proposed alignment model was more accurate than the compared state-of-the-art methods and build relatively accurate 3D models. However, this method failed to detect remote homologs in an accurate manner and performed worse than the benchmark methods. Also, reasonable computation time was only achieved by reduced the training data, suggesting that other machine learning algorithms may be more suitable here. All in all, this new approach was able to replace a substitution matrix with fixed values and caused accurate results in case of alignment prediction. Nevertheless, this proposed method showed that the problems of 3D protein structure prediction, as well as sequence alignment generation, can be tackled with machine learning based approaches that yield accurate results and models, if enough computational power and good training data is provided.

Zusammenfassung

1. Forschungsfragen:

Die Autoren versuchen, paarweise Sequenzalignments mithilfe einer Supervised Machine Learning Methode, welche strukturelle Alignments von Homologs als Trainingsdaten verwendet, durchzuführen um eine bessere Alignment-Qualität zu erzielen. Diese neue Methode soll dazu verwendet werden, um eine möglichst genaue Strukturvorhersage für eine Funktionsvorhersage zu erhalten und zudem noch Homologien erkennen.

2. Relevante methodische Ansätze:

- Strukturelle Alignments für Template-basierte Vorhersage
- Feature vector encoding, um einen Input für die Machine Learning Methode zu generieren
    - Binäre Klassifikation
- k-NN zur Vorhersage der Matching Scores
- Smith-Waterman, um Alignments zu erzeugen
- Pipeline: Vorhersage von Scores => Alignment => Template-basiertes Modellieren

3. Relevante Ergebnisse:

- Lange Ausführungszeit
- Feste Substitutionsmatrix konnte ersetzt werden, um Substitutionswerte vorherzusagen
- Das Alignment-Modell der vorgestellten Methode hatte eine höhere Genauigkeit als die ausgewählten Benchmarks (PSI-BLAST, HH-Search, etc.)
- Erkennung von entfernten Homologien war schlechter als die ausgewählten Benchmarks

4. Schlussfolgerung:

Die vorgestellte Methode zur Generierung von Alignments konnte mithilfe des Machine Learning Ansatzes eine feste Substitutionsmatrix umgehen, indem die Substitutionswerte für das Alignment vorhergesagt werden. Die hohe Genauigkeit des Models war höher als die gezeigten State-of-the-Art Methoden, allerdings war die Methode für die Erkennung entfernter Homologien ungeeignet. Um die Genauigkeit und die Effizienz zu erhöhen, schlagen die Autoren schnellere bzw. komplexere Machine Learning Methoden vor.

5. Schlüsselabbildung:

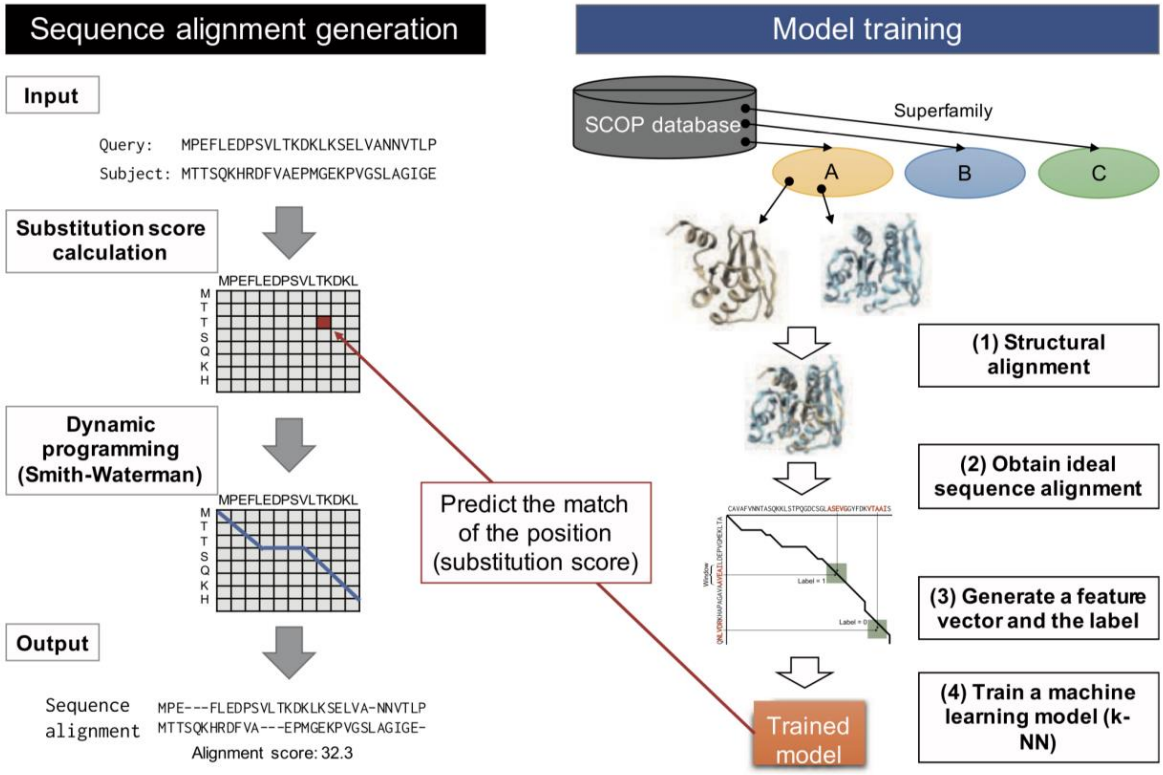


Fig. 2. Overview of the proposed method. Two sequences are aligned using the Smith–Waterman algorithm and substitution scores used in the process are estimated by a prediction model. The prediction model is trained to output an alignment similar to the structural alignment

Wir haben Abbildung 2 als Schlüsselabbildung gewählt, da der Ansatz der Autoren hier schematisch verdeutlicht und mit dem regulären Workflow der Sequenz-Alignment Generierung verglichen wird. Durch diese Abbildung wird gezeigt, dass beide Ansätze das gleiche Ziel haben und sich in nur einem Zwischenschritt unterscheiden, nämlich dem Berechnungsschritt der Substitutions Matrix Scores. Die anderen Abbildungen des Papers zeigen Proteinstrukturen oder die Ergebnisse des neuen Ansatzes im Vergleich zu State-of-the-Art Methoden. Die Abbildungen mit den Ergebnissen sind ebenfalls sehr relevant, allerdings können diese, im Gegensatz zu unserer Schlüsselabbildung, nicht ohne Kontext verstanden werden.

Impact

Die wichtigsten Referenzen unseres Papers:

- Recent Progress in Machine Learning-Based Methods for Protein Fold Recognition
  - Leyi Wei (Erstautor) und Quan Zou (korrespondierender Autor)
  - International journal of Molecular Science 2016, 17 (12), 2118; https://doi.org/10.3390/ijms17122118
  - Publiziert am 16. Dezember 2016, 85 Citations
  - Dieses Paper setzt sich mit dem gleichen Grundproblem (3D Struktur Vorhersage) auseinander und bietet eine Einführung in die Machine Learning Pipeline. Außerdem werden verschiedene Benchmarks vorgestellt, mit denen man das eigene Modell vergleichen kann

- SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures
  - Naomi K. Fox (Erstautor), Steven E. Brenner (Korrespondierender Autor), John-Marc Chandonia (Letztautor)
  - Nucleic Acids Research, Volume 42, Issue D1, 1 January 2014, Pages D304 D309, https://doi.org/10.1093/nar/gkt1240
  - Publiziert 3. Dezember 2013, 644 Citations
  - Dieses Paper stellt den verwendeten Protein-Datensatz bzw. die verwendete SCOP-Databse vor. Diese Daten werden als Trainings und Testdaten für den vorgestellten Algorithmus eingesetzt.

- Gapped BLAST and PSI-BLAST: a new generation of protein database search programs
  - Stephen F. Altschul (Erstautor und Korrespondent), David J. Lipman (Letztautor)
  - Nucleic Acids Research, Volume 25, Issue 17, 1 September 1997, Pages 3389 3402, https://doi.org/10.1093/nar/25.17.3389
  - Publiziert 1. September 1997, 80920 Citations
  - Diese Methode gilt als Benchmark für Homology-Erkennung und wird daher mit der vorgestellten Methode verglichen

- Identification of common molecular subsequences
  - T.F. Smith (Erstautor), M.S. Waterman (Letztautor, Korrespondierender Autor)
  - Journal of Molecular Biology, Volume 147, Issue 1, Pages 195-197, https://doi.org/10.1016/0022-2836(81)90087-5
  - Publiziert 1981, 13403 Citations
  - Die in unserem Paper vorgestellte Methode baut auf dem Smith-WatermanAlgorithmus auf. Das, was hier verändert wurde, war lediglich die

Substitutionsmatrix, deren Werte vom Machine Learning Ansatz vorhergesagt wurden.

- TM-align: a protein structure alignment algorithm based on the TM-score
  - Yang Zhang (Erstautor und Korrespondent), Jeffrey Skolnick (Letzautor)
  - Nucleic Acids Research, Volume 33, Issue 7, 1 April 2005, Pages 2302–2309, https://doi.org/10.1093/nar/gki524
  - Publiziert 1. Januar 2005, 2476 Citations
  - Der TM-Score wird hier als Evaluierungsmethode für strukturelle Alignments eingeführt. Dieser wird verwendet, um die vorgestellten Ergebnisse unseres Papers zu evaluieren.

Citations unseres Papers:

- Sequence alignment using machine learning for accurate template-based protein structure prediction
  - Shuichiro Makigaki (Erstautor und Korrespondent), Takashi Ishida (Letztautor und Korrespondent)
  - Bioinformatics, Volume 36, Issue 1, 1 January 2020, Pages 104–111, https://doi.org/10.1093/bioinformatics/btz483
  - Publiziert 14. Juni 2019, 9 Citations (laut Google Scholar):
    - 2019: 2 Citations (davon eine von den gleichen Autoren!)
    - 2020: 1 Citation
    - 2021: 5 Citations
    - 2022: 1 Citation

Die 5 relevantesten Studien, die unser Paper zitieren:

- A survey on the algorithm and development of multiple sequence alignment:
  - Yongqing Zhang, Qiang Zhang, Jiliu Zhou, Quan Zou
  - Publiziert am 10. März 2022
  - Briefings in Bioinformatics, Volume 23, Issue 3, May 2022, bbac069, https://doi.org/10.1093/bib/bbac069, höchster Impact-Factor mit 8.99

- Challenges in the Computational Modeling of the Protein Structure— Activity Relationship:
  - Gabriel Del Rio
  - Publiziert 4. Februar 2021
  - MDPI Computation 2021, 9 (4), 39; https://doi.org/10.3390/computation9040039 , Impact-Factor 2.64

- Prediction and Visualisation of Viral Genome Antigen Using Deep Learning & Artificial Intelligence:
  - Akshat Jain, Shamik Tiwari

    - Publiziert 6. May 2021
    - IEEE Xplore, DOI: 10.1109/ICCMC51019.2021.9418356 , Impact-Factor 2.4828

- Sequence Alignment with Q-Learning Based on the Actor-Critic Model:
    - Yarong Li
    - Publiziert 30.Juni 2021
    - ACM Transactions on Asian and Low-Resource Language Information ProcessingVolume 20Issue 5 , Impact Factor 1.86

- Metaheuristics for multiple sequence alignment: A systematic review:
    - Anderson Rici Amorim, Liria Matsumoto Sato et al.
    - Publiziert 16. August 2021
    - Science Direct: Computational Biology and Chemistry, Impact-Factor 1.479

Anmerkung: Unser Paper hat zum jetzigen Zeitpunkt lediglich 9 Citations. Zwei davon waren in fremder Sprache und eine weitere wurde auf bioRxiv gefunden, was kein Journal ist. Alle davon befinden sich im gleichen Themenbereich, aber mit unterschiedlicher Gewichtung der Methode (Deep Learning/Machine Learning) oder Proteinstrukturanalyse.

Autoren und ihre Publikationen:

Beide Autoren sind als korrespondierende Autoren gelistet und haben einige Publikationen gemeinsam veröffentlicht. Allerdings hat der Zweitautor, Takashi Ishida, deutlich mehr Publikationen veröffentlicht. Daher wird er im Folgenden als Hauptkorrespondent angesehen und seine Publikationsliste wird weiter evaluiert.

- Defucosylated Anti-CCR4 Monoclonal Antibody (KW-0761) for Relapsed Adult TCell Leukemia-Lymphoma: A Multicenter Phase II Study:
    - Takashi Ishida ist hier Erstautor
    - Publiziert 6. Februar 2012, 611 Citations
    - journal of clinical oncology, Impact Factor 44.544 (Top 1%)
    - Krebstherapiestudie ⇒ Sehr hohe Relevanz

- Multicenter phase II study of mogamulizumab (KW-0761), a defucosylated anticc chemokine receptor 4 antibody, in patients with relapsed peripheral T-cell lymphoma and cutaneous T-cell lymphoma:
    - Takashi Ishida ist hier Zweitautor
    - Publiziert 2004, 330 Citations
    - journal of clinical oncology, Impact Factor 44.544 (Top 1%)
    - Krebstherapiestudie ⇒ Sehr hohe Relevanz

- PrDOS: prediction of disordered protein regions from amino acid sequence:
    - Takashi Ishida ist hier Erstautor

    - Publiziert 1. July 2007, 711 Citations
    - Nucleic Acids Research, Impact Factor 16.971
    - Diese Publikation setzt sich ebenfalls mit Machine Learning Verhersagen auf Protein-Daten auseinander, allerdings werden hier “disordered regions” im Protein vorhergesagt

- D2P2: database of disordered protein predictions:
    - Takashi Ishida ist hier Drittautor
    - Publiziert 29. November 2012, 516 Citations
    - Nucleic Acids Research, Impact Factor 16.971
    - Relevanter Datensatz mit “disordered Proteins” wird durch die Publikation veröffentlicht, bezieht ich zT auch auf die oben erwähnte SCOP Database

- Prediction of disordered regions in proteins based on the meta approach:
    - Takashi Ishida ist hier Erstautor
    - Publiziert am 20. April 2008, 281 Citations
    - Bioinformatics, Impact Factor 4.531
    - Ishida hat relativ viel zu “disordered regions in proteins” (und deren Vorhersage) publiziert. Vor allem in diesem Paper wird die Motivation dahinter deutlich: Diese Regionen erschweren die Struktur-Vorhersage und haben eine wichtige Rolle in Molekül-Interaktionen.

Zusammenfassend kann man sagen, dass Takashi Ishida an sehr relevanten Krebsstudien maßgeblich mitgewirkt hat und in hochkarätigen Journals publiziert hat. Ishida unterstützt maßgeblich den Erstautor Shuichiro Makigaki. Makigaki’s Fokus liegt eher auf Sequenz Alignments, Homologien und Machine Learning. Allerdings wirkt Ishida in diesem Bereich mit, alle Publikationen von Makigaki (Stand 2022: 5st) enthalten Ishida als Letztautor.

Highly accurate protein structure prediction with AlphaFold

Proteins are the building blocks of all life. To gain a deeper understanding of the purpose and functionality of a protein, it is essential to understand the underlying structure. The structure of a protein is determined by the three-dimensional arrangement of the amino acid sequence.[**?**] The question dealing with this three-dimensional arrangement of molecules is also called the "protein folding problem" and has been one of the most important questions in structural biology for more than 60 years.[**?**] However, the resolution of protein structures is connected to a high experimental effort. Here the authors present AlphaFold, the first computational method that regularly predicts protein structures with atomic accuracy. AlphaFold is a machine learning approach that uses biological and physical knowledge to predict protein structure. Multiple sequence alignment as well as paired residue information are used to make the first prediction. The three-dimensional structure is subsequently realized and adapted in a following process. The efficiency and correctness of AlphaFold was further confirmed by the results of the 14th Critical Assessment of protein Structure Prediction (CASP14), making it a state of the art tool to use for protein structure research. AlphaFold enables the structure prediction of proteins and protein complexes for which no structural information is available to date. This should lead to breakthroughs in structural bioinformatics, as AlphaFold provides a way to keep up with the rapid development of genomic sequencing techniques and the acquisition of sequencing data. AlphaFold and future approaches using the technology of AlphaFold will be an essential part of modern biology.

Highly accurate protein structure prediction with AlphaFold

Proteins are the gearwheels of every organism, in form of small functional structures, facilitating the essential task in each living system. The precise decryption of their structures is essential for a detailed understanding how living systems are working. The precise and reliable determination of a individual protein structure is currently only possible by the usage of extensive experimental capacities and costs. This is a bottleneck regarding to the billions of known proteins sequences discovered by the usage of DNA sequencing. To circumvent this bottleneck, the protein structure prediction based on determined amino acid sequences is a known to be unresolved problem of the last 50 years of research. Regardless of the big advantages that were made in the past two decades, pushed by the inclusion of powerful computational algorithms and resources, there is still a low accuracy in cases where no homolog structure is known. Here we show AlphaFold, our pioneering deep learning based structure prediction approach, that provides a significantly higher accuracy compared to existing methods, even in cases where no homologous sequence is known. The performance of AlphaFold was validated during the challenging 14th Critical Assessment of protein Structure Prediction (CASP14), revealing an unseen accuracy improvement, greatly outperforming other methods. With our algorithm a big step, on the long and difficult way towards precise high throughput protein structure prediction, is done. Our latest version AlphaFold2 is a novel deep learning based structure prediction approach that comprises physical and biological knowledge about protein structures. Enabling accurate sequence-based structure prediction with comparably low costs and resources needed.

Highly accurate protein structure prediction with AlphaFold

(https://doi.org/10.1038/s41586-021-03819-2)

Paper summary

What are the treated research questions?

Development of a novel machine learning based protein structure prediction, that embeds physical and biological knowledge in a new deep learning algorithm.

What are the relevant methodical approaches?

→ New model structure that combines Multiple Sequence Alignment and pairwise feature information.
↪ Evoformer:
- exploits spatial information and evolutionary relationships.
- contains number of attention-based and non-attention-based components.
↪ Structure Module:
- breaking the chain structure for simultaneous local refinements.
- Equivariant transformer to reason about unrepresented side-chain atoms.
- Loss term that places substantial weight on orientational correctness.
→ Iterative refinement of predictions
→ Training based on evolutionary, physical and geometric constraints of protein structures.

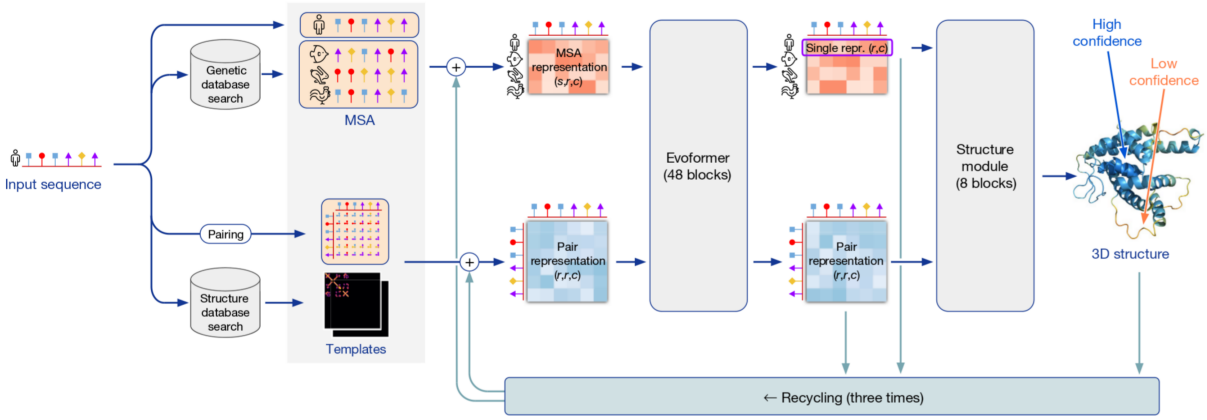
What are the relevant results?

→ High accuracy that greatly outperforms existing prediction methods.
→ Precise novel structure prediction
→ high side chain accuracy when back bone prediction is accurate.

Conclusion

→ Revolutionary accuracy improvement in high throughput structure prediction.
→ Way for improvements at the prediction of larger less homotypic complexes and sequences with limited MSA data available.

What is the key illustration?



This figure shows a schematic structure overview for the neural network approach, specifying the used input data and embedding the revolutionary structure prediction process. We chose this as key figure because it illustrates the approach of combining multiple sequence alignment and residue pairing for protein structure prediction and all the components that affect this prediction.

Highly accurate protein structure prediction with AlphaFold

5 Most Fundamental References

- **Protein Data Bank: the single global archive for 3D macromolecular structure data**
  - *Authors*: wwPDB consortium (Stephen K Burley; Yannis E Ioannidis)
  - *Journal*: Nucleic Acids Res.; 47(Database issue): D520–D528
  - *Publication date*: 08.01.2019
  - *Number of citations* of citations: 330
  - *Reason*: The Protein Data Bank is the foundation of all AlphaFold training data and is therefore instrumental in the success of AlphaFold.
- **The Protein Folding Problem**
  - *Authors*: Ken A. Dill; Ken A. Dill, S. Banu Ozkan, M. Scott Shell, Thomas R. Weikl; Thomas R. Weikl
  - *Journal*: Annu Rev Biophys.; 37: 289–316.
  - *Publication date*: 09.06.2008
  - *Number of citations*: 1140
  - *Reason*: The basic problem solved by AlphaFold is the protein folding problem.
- **Principles that govern the folding of protein chains**
  - *Authors*: C B Anfinsen
  - *Journal*: Science.; 181(4096): 223-30
  - *Publication date*: 20.07.1973
  - *Number of citations*: 9433
  - *Reason*: Describes the basic mechanisms underlying protein folding and demonstrates the relevance of this problem over time.
- **rawMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments**
  - *Authors*: Claudio Mirabello; Björn Wallner; Yang Zhang
  - *Journal*: PLoS One.; 14(8): e0220182
  - *Publication date*: 15.08.2019
  - *Number of citations*: 33
  - *Reason*: Handles one of the two fundamental ideas (MSA) on which AlphaFold's predictions are based
- **Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks**

  - *Authors*: Yang Li; Dong-Jun Yu, Yang Zhang; Yang Zhang
  - *Journal*: PLoS Comput Biol.; 17(3): e1008865
  - *Publication date*: 26.03.2021
  - *Number of citations*: 36
  - *Reason*: Handles the other of the two basic ideas (Paired Residue) on which AlphaFold's predictions are based.

Paper Impact

Citations: 2976 → 2976 per year

- Highly accurate protein structure prediction for the human proteome (Nature)
- Altered TMPRSS2 usage by SARS-CoV-2 Omicron impacts infectivity and fusogenicity (Nature)
- ColabFold-Making protein folding accessible to all (Researchsquare)
- De novo protein design by deep network hallucination (Nature)
- Improved prediction of protein-protein interactions using AlphaFold2 (Nature)

Corresponding author

Corresponding author: John Jumper

- Highly accurate protein structure prediction with AlphaFold → Revolutionary paper and high impact (Nature 2021, 2976 citations)
- Atomic-level characterization of the structural dynamics of proteins → High impact (Science 2010, 1784 citations)
- Improved protein structure prediction using potentials from deep learning → Deep Learning approach and paper impact (Nature 2021, 1611 citations)
- Highly accurate protein structure prediction for the human proteome → Use of AlphaFold and paper impact (Nature 2021, 526 citations)
- Oncogenic Mutations Counteract Intrinsic Disorder in the EGFR Kinase and Promote Receptor Dimerization → paper impact (Cell 2012, 338 citations)

Group: Jessica Bender, Jonas Schuck

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

Tao Zhang, Qunfu Wu, and Zhigang Zhang [2020]

Abstract 1

The outbreak of the 2019 novel coronavirus (SARS-CoV-2) in the city of Wuhan in China and its extensive impact on human health requires further research on affected animal species with CoV-like viruses. Phylogenetic analysis is essential to examine key information about CoV-like viruses that could potentially help combat the spread of COVID-19. To discover evolutionary relatives, it is necessary to determine similarities in the genomic organization, for example the presence and similarity of the spike (S) protein as well as specific cleavage motifs used for host cell entry. However, apart from BatCoV RaTG13, there are more known species affected by potentially highly similar CoV-like viruses, with their affinity to the SARS-CoV-2 virus being still unexplored. Here they show that Pangolin-CoV is the second-closest known relative to SARS-CoV-2, right behind BatCoV RatG13, with an identity of 91.02 % on the whole genome level. Additionally, they analyzed the identity of the (S) protein, which is around 97.5 %, indicating functional similarities between the Pangolin-CoV and the SARS-CoV-2 virus. Moreover, they found that the highly conserved nucleocapsid (N) protein is also present in the Pangolin-CoV. The (N) protein is often used as a marker in diagnostic assays, while the (S) protein is essential for the cell entry process of the virus, giving information about the pathogenic potential of Pangolin-CoV. With these findings, the pangolin species can be classified as a natural reservoir of SARS-CoV-2-like CoVs. Further experiments should be performed to deepen the knowledge about the interconnection of potentially related viruses to be able to block interspecies transmission.

Group: Jessica Bender, Jonas Schuck

Abstract 2

The new Coronavirus 2019 (SARS-CoV-2) has been spreading rapidly worldwide since the 2019 outbreak in China. To control and contain the spread of the disease, research into potential intermediate hosts is essential. By finding the intermediate host, with the bat still a plausible intermediate host, it would allow interspecies transmission to be blocked. Thus, previously published lung samples from pangolins with a similar COV variant in the genome and as such potential intermediate hosts were analyzed again. Here they show that Pangolin-CoV is the most closely related COV to human SARS-CoV-2 besides the rat COV. At the genomic level, Pangolin-CoV is 91.02% identical to SARS-CoV-2. Furthermore, five key amino acids in the RBD match between Pangolin-CoV and SARS-CoV-2. However, the putative furin-recognition sequence motif only occurs in the human SARS-CoV-2. The results demonstrate that Pangolin, in addition to the bats, could also be considered a repository of CoVs similar to SARS-CoV-2. It is expected that the assay will increase the awareness of potential intermediate hosts and that the knowledge gained from the assay will help in the search for intermediate hosts of the virus. Conversely, this could prevent a global pandemic.

Paper summary of

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

Tao Zhang, Qunfu Wu, Zhigang Zhang

1: Research question

The paper is about the exploration of potential intermediate hosts of SARS-CoV-2 to control COVID19 spread and identify species that act as natural reservoirs of SARS-CoV-2 like CoVs performed on the pangolin.

2: Relevant methods

- **Data collection** (using raw RNA-seq. data from lung samples of dead pangolin individuals generated in a different study) + **pre-processing** (adapter/quality trimming) and **mapping against a reference genome** of the pangolin species to identify virus reads.
- **De novo genome assembly** + **contig annotation** with BLAST against different CoV reference genomes.
- Use **sequence alignment to generate different phylogenetic trees** to **determine and analyze evolutionary relationships** (different datasets from various CoV genes were used, e.g. whole genome, non-structural protein genes, and structural protein genes)

3: Relevant results

- Pangolin-CoV is 91.02 % identical to SARS-CoV-2 on the whole genome level and 90.55 % identical to BatCoV RaTG13 determining that Pangolin-CoV is the second-closest known relative to SARS-CoV-2.
- Structural similarities to SARS-CoV-2 in the spike protein and nucleocapsid protein determined by the comparison of the genome organization are indicators for the probable pathogenic potential of Pangolin-CoV.
- SARS-CoV-2 inherits a cleavage motif in the spike protein sequence which is used in the cleavage/priming phase by the host cell proteases to enable cell entry of the virus, which is lost in the Pangolin-CoV and RaTG13 variants.
- Further experiments could not be performed due to the unavailability of the original data.

4: Conclusion

Even though the authors could confirm 91.02 % of similarity between the Pangolin-CoV and the SARS-CoV-2, at the time of the release of this study it was still under debate if the pangolin species is a validate candidate to trace back the SARS-CoV-2 outbreak from a natural reservoir of CoV-like viruses. Similarities in the spike and n protein could be observed but for example, the cleavage motif in the spike protein sequence is not present in the pangolin variant, which is an essential part of the cell entry mechanism of SARS-CoV-2 making it pathogenic and transmissible to humans. Still, pangolins can be added to the species that act as a natural reservoir of SARS-CoV-2-like CoVs.

5: Key figure of the paper

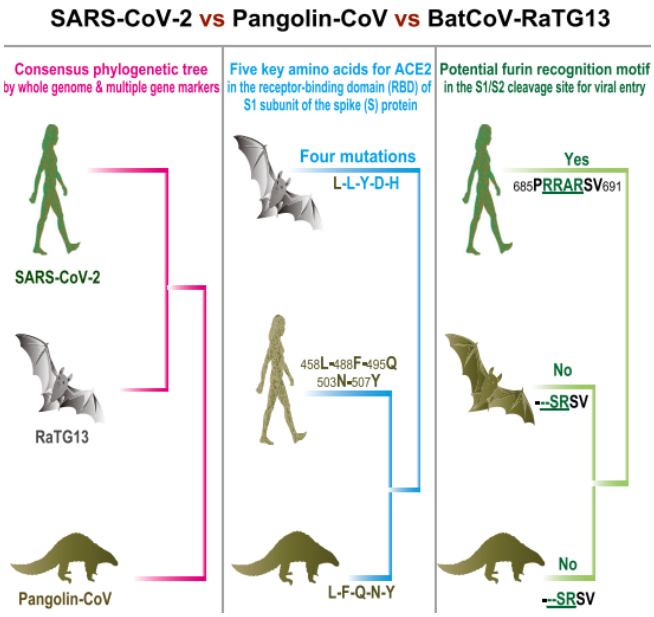


Figure 1: Graphical abstract taken from the paper.

The figure shows the comparison of the two potential intermediate hosts (Pangolin & Bat) of SARS-CoV-2 in relation to humans. Three smaller figures illustrate relevant results based on phylogenetic trees. The pink phylogenetic tree describes the analysis and evolutionary relationship of humans, bats, and pangolins of the whole genome and several gene markers. The result was that the whole genome was almost identical between SARS-CoV-2 and BatCoV RaTG13. The blue phylogenetic tree shows that five key amino acids in the receptor-binding domain of the spike protein correlate between SARS-CoV-2 and Pangolin-CoV. In contrast, only four key amino acids match between the BatCoV and SARS-CoV-2 with several changes in essential residues. The green tree shows the occurrence of the furin recognition motif found exclusively in SARS-CoV-2.

From our point of view, a key illustration should present the results in a visualized, clear, and comprehensible way. The illustration should be understandable with as limited text as possible. Therefore, this figure can be interpreted as a key figure, as it achieves these qualities and presents relevant results in a simply visualized way.

Paper impact of Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak by Tao Zhang, Qunfu Wu, Zhigang Zhang [2020]

1: Five relevant references

1: Authors (scheme: first, corresponding, (last)): Ping Liu, Jin-Ping Chen.

2: Title: Viral Metagenomics Revealed Sendai virus and Coronavirus Infection of Malayan Pangolins (*manis javanica*)

3: Journal: Viruses, Vol. 11, p. 979.

4: Release year: 2019

5: Citations: 185 [Scopus]

6: Cause of relevance: Pangolin lung sample data from this study got used in the proposed paper; Reinvestigation and continuation of the findings.

1: Authors: Peng Zhou, Zheng-Li Shi

2: Title: A pneumonia outbreak associated with a new coronavirus of probable bat origin

3: Journal: Nature, Vol. 579, p. 270-273.

4: Release year: 2020

5: Citations: 9886 [Nature]

6: Cause of relevance: Findings about the evolutionary relationship of BatCoV RaTG13 to SARS-CoV-2. Findings that the proposed paper uses to compare Pangolin-CoV with BatCoV.

1: Authors: Wendong Li, (Zhengli Shi, Shuyi Zhang, Lin-Fa Wang), Lin-Fa Wang

2: Title: Bats Are Natural Reservoirs of SARS-Like Coronaviruses

3: Journal: Science, Vol. 310, No. 5748, p. 676-679

4: Release year: 2005

5: Citations: 1185 [Pubmed]

6: Cause of relevance: Identifying natural reservoirs for CoV-Like viruses.

1: Authors: Wendong Li, (Zhengli Shi, Shuyi Zhang, Lin-Fa Wang), Lin-Fa Wang

2: Title: Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins

3: Journal: Nature, Vol. 583, p. 282-285

4: Release year: 2020

5: Citations: 802 [Nature]

6: Cause of relevance: Similar study with different data used; independent confirmation of very similar results.

1: Authors: Jean Kaoru Millet, Gary R. Whittaker

2: Title: Host cell entry of Middle East respiratory syndrome coronavirus after two-step, furin-mediated activation of the spike protein.

3: Journal: PNAS, Vol. 111, p. 15214-15219

4: Release year: 2014

5: Citations: 589 [Google Scholar]

6: Cause of relevance: Findings about the Spike protein of the MERS-CoV. Used to examine the pathogenic potential.

2: Most influential studies citing the paper

Citations: 498 results at **Pubmed (2020-2022).**

Citations per year: 498/3 = 166 citations/year.

Citing papers with highest impact factors [Pubmed]:

- SARS-CoV-2s' origin should be investigated worldwide for pandemic prevention. [Lancet Journal; IF: 79.32]
- The proximal origin of SARS-CoV-2 [Nature Medicine; IF: 53. 44]
- Characteristics of SARS-CoV-2 and COVID-19 [Nature Review Microbiology; IF: 60.63]
- Prospects of SARS-CoV-2 diagnostics, therapeutics and vaccines in Africa. [Nature Review Microbiology; IF: 60.63]
- Six months of coronavirus: the mysteries scientists are still racing to solve. [Nature; IF: 49.96]

3: Corresponding author: Zhigang Zhang

(Zhigang Zhang: State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, Yunnan University)

Release of 12 publications in the years 2014-2020. The following publications are the five studies released in the most cited journals.

- γδT17 cells promote the accumulation and expansion of myeloid-derived suppressor cells in human colorectal cancer [Immunity; IF: 31.745]
- NLRC3, a Member of the NLR Family of Proteins, is a Negative Regulator of Innate Immune Signaling Induced by the DNA Sensor STING [Immunity; IF: 31.745]
- Convergent Evolution of Rumen Microbiomes in High-Altitude Mammals [Current Biology; IF: 10.83]
- NLRX1 Sequesters STING to Negatively Regulate the Interferon Response, Thereby Facilitating the Replication of HIV-1 and DNA Viruses. [Cell Host & Microbe: IF: 21.02]
- A Pharmacogenomic Landscape in Human Liver Cancers [Cancer Cell; IF: 26.02]