Seminar Algorithmen in der Sequenzanalyse (ASA-S)

Statement of Relevance (SoR) 1

The sequencing of genomes, transcriptomes and proteomes is now routine in applied biological and medical research.

SoR 2

There is an ever increasing focus on bioinformatics approaches to manage, analyse, integrate and interpret biological sequence data on a genome-wide scale.

SoR 3

As a consequence, bioinformatics sequence analysis has developed into a multi-faceted rapidly evolving reasearch field.

SoR 4

The current seminar is based on a selection of papers focussing on key aspects of sequence analysis, ranging from the intial data processing to sequence analysis and modeling in a biological context. It should deepen and connect the individual lecture topics and it should provide complementary information on theoretical and applied aspects of bioinformatic sequence analysis.

SoR 5

The seminar that accompanies the Master course Algorithms in Sequence Analysis will cover relevant papers from the area of biosequence informatics:

1. DNA sequencing and error correction
2. Sequence assembly and structural variant detection
3. Multiple Sequence Alignment
4. Sequence Annotation
5. Phylogeny reconstruction
6. Machine learning

Definition Seminar

Die Studienordnung des MSc Bioinformatik definiert ein Seminar wie folgt:

- Ein Seminar
  - ist eine Gruppenveranstaltung
  - dient der Erörterung wissenschaftlicher Probleme
  - führt in die selbständige Erarbeitung wissenschaftlicher Literatur ein.
- In der Regel muss von den Teilnehmerinnen oder Teilnehmern
  - ein gegebenes Thema bearbeitet
  - eine Ausarbeitung angefertigt
  - ein Vortrag gehalten werden.
- Von allen Teilnehmerinnen und Teilnehmern wird eine aktive Teilnahme an der Diskussion erwartet.

Definition Seminar 2

- Die Zahl der Teilnehmerinnen und Teilnehmer an einem Seminar ist auf 15 begrenzt. Über Ausnahmen entscheidet die Veranstaltungsleitung[1)].
- Für die Teilnehmerinnen oder Teilnehmer eines Seminars besteht Anwesenheitspflicht!

Modulbeschreibung

Die Beschreibung des Moduls ASA-S ist in Abbildung 1 wiedergegeben. Der Arbeitsaufwand im Kontaktstudium beträgt 30 h (1 CP) und im Selbststudium 120 h (4 CP). Die Umrechnung CP in Arbeitsaufwand erfolgt hier nach den Standards für Veranstaltungen der GU Frankfurt.

M-ASA-S: Aktuelle Themen der Sequenzanalyse: Algorithmen

Verwendbarkeit: Master Bioinformatik, Pflichtmodul im Vertiefungsgebiet Sequenzanalyse / Data Mining		
Credit Points: 5 (SL)	Rhythmus: jährlich (SS)	Dauer: einsemestrig
Veranstaltungen: Die Veranstaltung ASA-S ist Pflichtveranstaltung des Moduls.		
Zulassungsvoraussetzungen zur Modulprüfung: Keine.		
Modulabschlussprüfung: Schriftliche Ausarbeitung und Vortrag.		

Seminar Aktuelle Themen der Sequenzanalyse

Veranstaltungs-Nr.: ASA-S	SWS: 2 S	Rhythmus: jährlich (SS)	Kontaktstunden: 1 CP
Lehrform: Seminar	Unterrichtssprache (i.d.R.): Deutsch oder Englisch		Selbststudium: 4 CP

Inhalt: Aktuelle Themen im Bereich der Sequenzanalyse und Phylogenie, insbesondere bezüglich neuer Algorithmen, Methoden und Anwendungen, sind anhand von Originalarbeiten und ergänzender Literatur vorzustellen.

Lernziele: Das Kennenlernen neuester Forschungsergebnisse in der Genomanalyse und phylogenetischen Analyse, das Verstehen wissenschaftlicher Originaltexte, die Fähigkeit zur Einordnung der Inhalte und Aussagen sowie deren Wiedergabe in eigener Darstellung in einem begrenztem Zeitrahmen.

Teilnahmevoraussetzungen / erforderliche Kenntnisse: Keine.

Nützliche Vorkenntnisse: Keine.

Figure 1: Modulbeschreibung des Seminars Algorithmen der Sequenzanalyse M-ASA-S

Kurs OLAT

Die Seminaraufgaben finden Sie im Olat system

Slides stehen im Teaching WIKI
https://applbio.biologie.uni-frankfurt.de/teaching/wiki/doku.php?id=asa:seminar-main#seminartage

Semesterablauf 1

- 29. April 2025
  - Einleitung, Daten, was erwartet Sie:
  - Hauptthema: Was ist ein Seminar, wozu dient es eigentlich, was macht ein gutes Seminar aus?
- Mai 06
  - Was ist eine Wissenschaftliche Arbeit - wie ist sie aufgebaut → IMRAD Schema.
  - Kurzvorstellung der Artikel
  - Aufgabe: für die kommende Woche: Paper lesen und Score abgeben
  - Paper-Zuweisung via Auge-System. **Deadline: 11. Mai**
- Mai 13
  - Was ist ein Poster → was soll ein Poster leisten
  - Aufgabe: Paper lesen

Semesterablauf 2

- Mai 20
  - Wie fasst man ein Paper zusammen?
  - Aufgabe: Paper-Zusammenfassung inklusive Wahl und Beschreibung einer Schlüsselabbildung
  - Deadline: 25. Mai
- Mai 27
  - Schreiben einer Zusammenfassung (Abstract)
  - Aufgabe: Abstract schreiben für das zugewiesene Paper
  - Deadline: 01. Juni
- Juni 03 - Paper Impact
  - Impact-Faktoren und Journal Metrics
  - Aufgabe: Zusammenstellung der Zitationen, Metrics, etc für das zugewiesene Paper
  - Deadline: 8. Juni
- Juni 03 - Bewertung der Abstracts
  - Aufgabe: Lesen Sie die Abstracts Ihrer Kolleginnen und Kollegen und bewerten Sie die sieben nachfolgend angeführten Kriterien mit 1 (A), 2 (B) oder 3 (C). Sie müssen nur die Hälfte der Abstracts lesen: Bearbeiten Sie ein Paper mit einer Nummer < P08, bewerten Sie bitte die Abstracts der Paper 09 - P15. Ansonsten bewerten Sie bitte die Abstracts der Paper P02 - P07.
    1. Clarity[2)]
    2. Accessibility[3)]
    3. Entertainment[4)]
    4. Research Objective[5)]

5. Main Result[6)]
6. Relevance[7)]
7. Language[8)]

Geben Sie Ihre Voten gesammelt für alle Abstracts in **einem** Textfile ab, der dem folgendem Format entspricht:

Zeile 1: ###Papernummer[9)]
Zeilen 2-8: ##Kriteriennummer,Votum
Zeile 9: //
Zeilen 10-19: Bewertung für das zweite Abstract
Zeilen 20-29: Bewertung für das dritte Abstract, etc

```
###1.1
##1,1
##2,3
##3,2
##4,1
##5,1
##6,2
##7,1
//
###1.2
##1,2
##2,2
##3,1
##4,1
##5,2
##6,1
##7,2
//
```

- Deadline: 16. Juni

Semesterablauf 3

- Juni 10
  - Postererstellung, Flash-Talks und Poster-Abstract
- Juni 17 / 24; Juli 1 / 8 - Keine Präsenz (vorläufig)

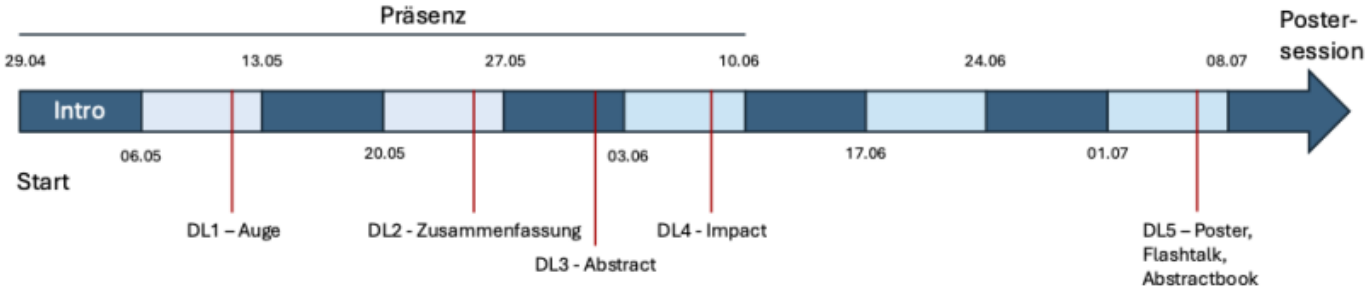
Deadlines

- Flash-Talk: 07. Juli 2025
- Poster: 07. Juli 2025
- Abstract Book: 07. Juli 2025

Seminartage

- Datum (QIS):
  - Montag - 14.07.2025 09:00 bis 17:00 Biologicum - Bio 3.206 AK Ebersberger (Postersession)
- Modus: 2er Teams bearbeiten eine Publikation
  - flash talk (Video): 7-8 min
  - Posterpräsentation: 7-8 min
  - Diskussion am Poster

Ablauf in der Übersicht

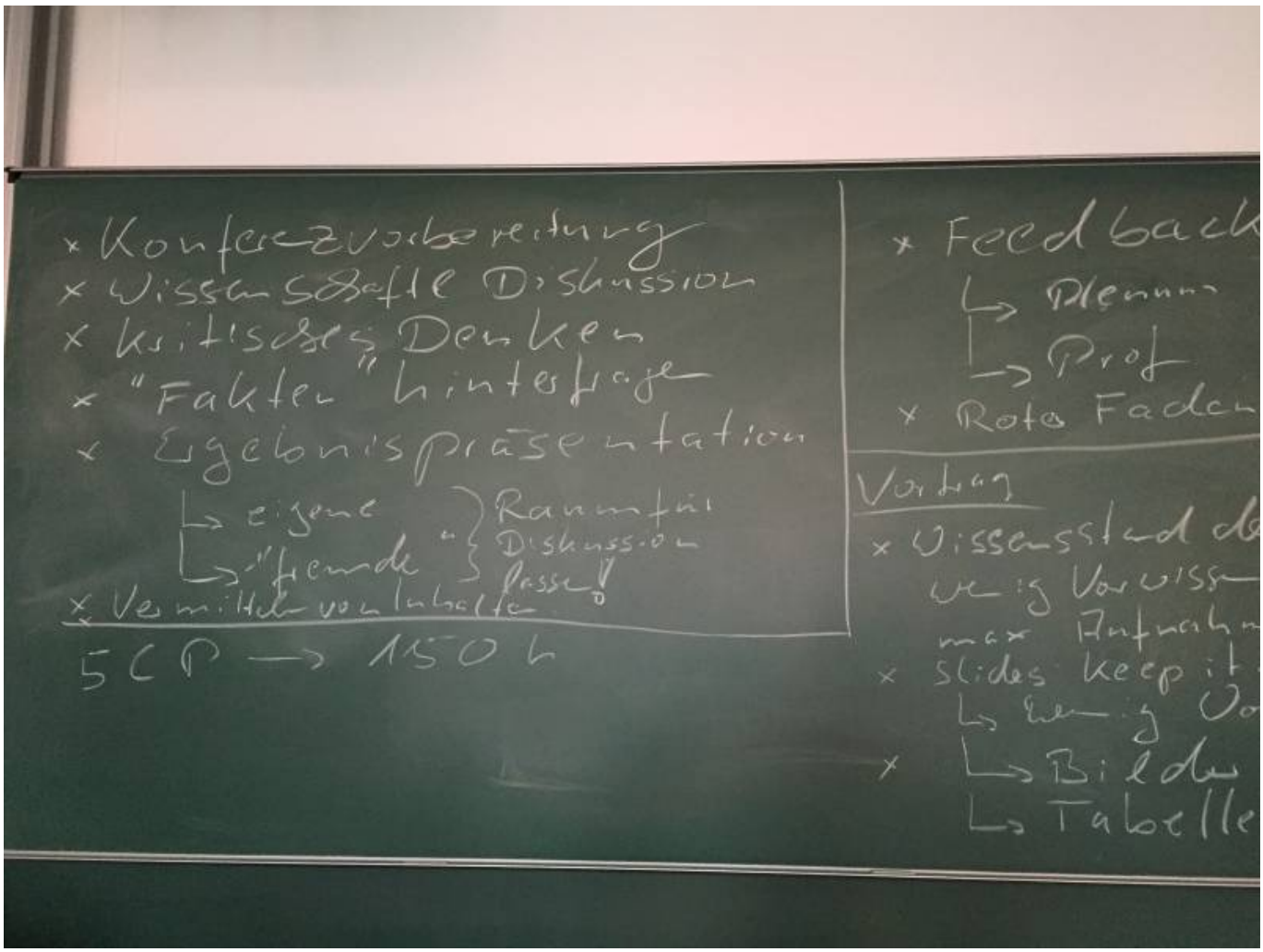


DL – Deadline für die Abgabe. Alle Abgabetermine liegen auf einem Sonntag um 23:59 Uhr

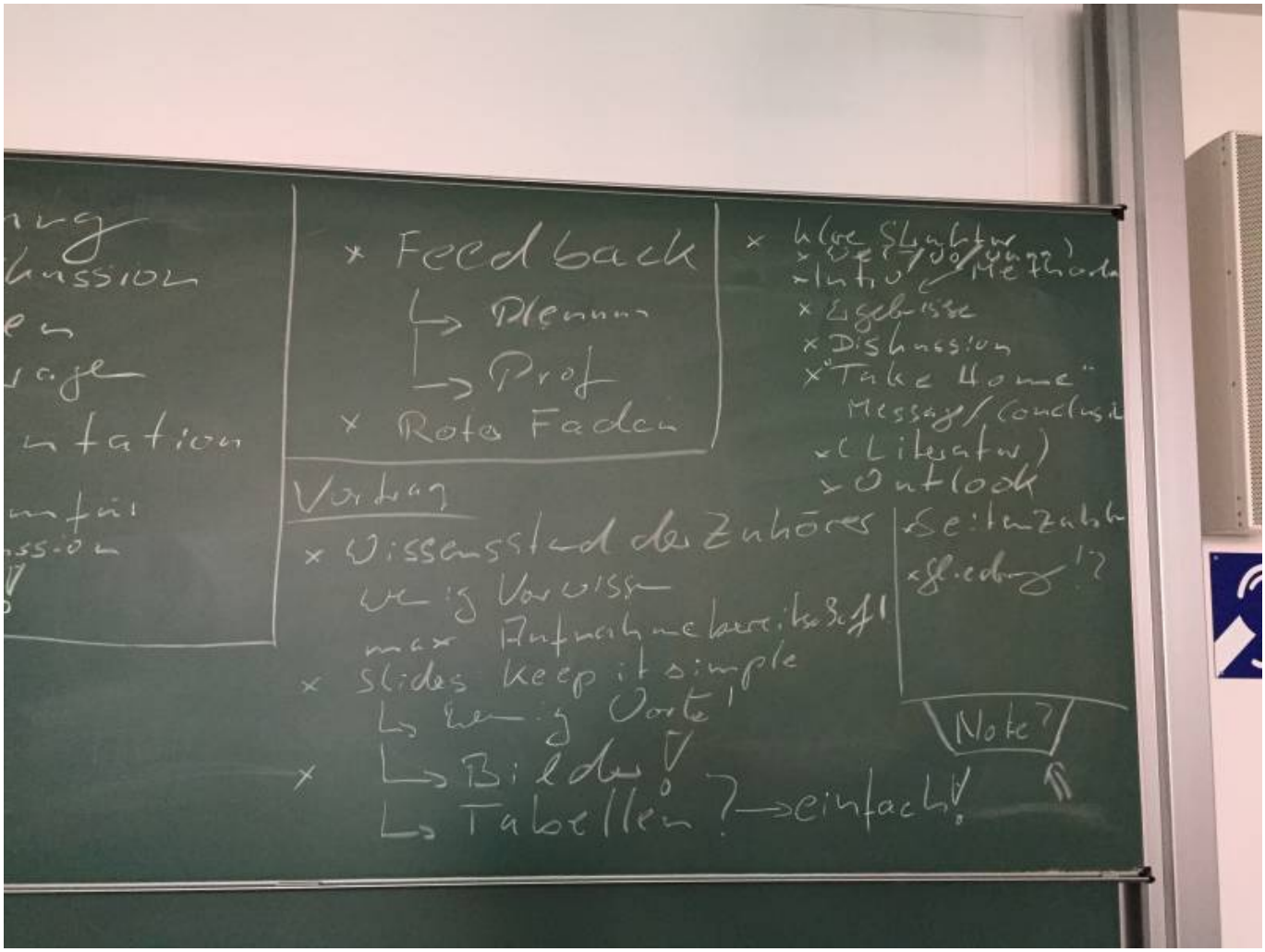
Was soll ein Seminar leisten?

Was soll ein 'gutes' Seminar - und die dazugehörigen Vorträge - leisten?

Was soll ein Seminar leisten - 2



Was soll ein Seminar leisten - 3



Was soll ein Seminar leisten - 4

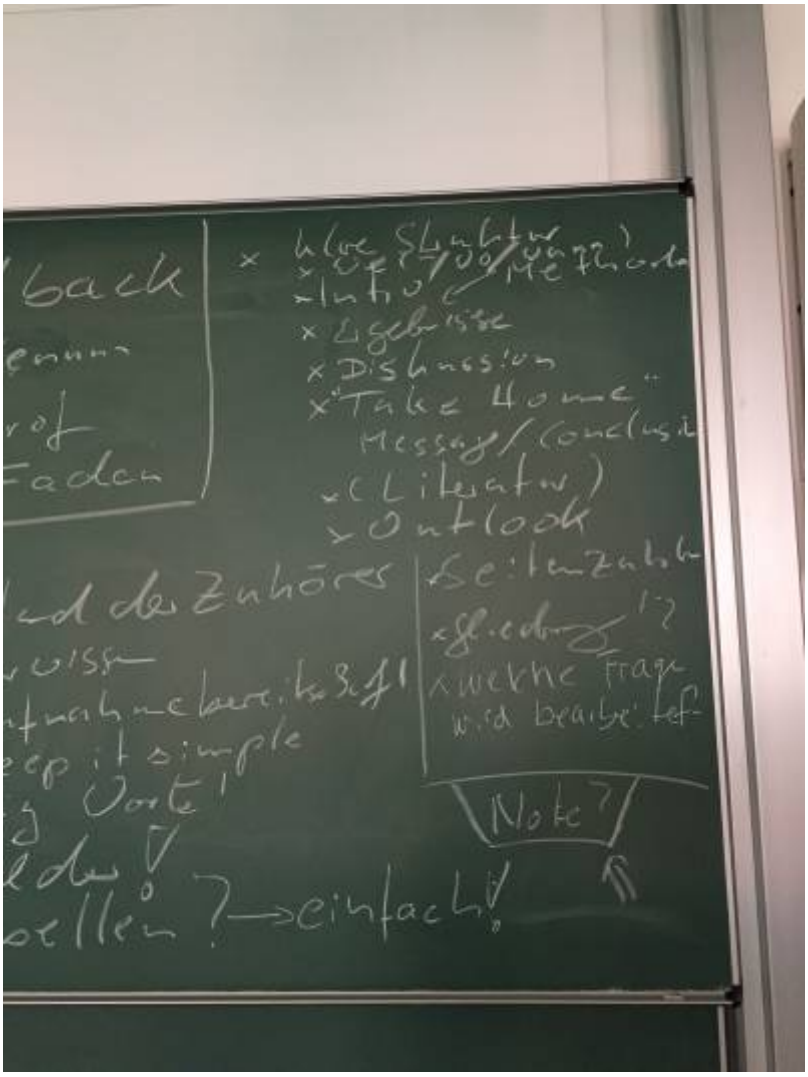


Figure 5: Was soll ein Seminar leisten, welche Punkte sind zu beachten.

Topics

1. Molecules as documents of evolutionary history. Zuckerkandel and Pauling 1965. Journal of Theoretical Biology 8(2):357-366

   Different types of molecules are discussed in relation to their fitness for providing the basis for a molecular phylogeny. Best fit are the “semantides”, i.e. the different types of macromolecules that carry the genetic information or a very extensive translation thereof. The fact that more than one coding triplet may code for a given amino acid residue in a polypeptide leads to the notion of “isosemantic substitutions” in genic and messenger polynucleotides. Such substitutions lead to differences in nucleotide sequence that are not expressed by differences in amino acid sequence. Some possible consequences of isosemanticism are discussed.

   Link to

   PDF

   Evolutionary Divergence and Convergence, in Proteins. Zuckerkandl and Paulin 1965. “Protides of Biological Fluids,” Proceedings of the 12th Colloquium (H. Peeters, ed.), p. 102, Bruges, 1964

KI-generated Summary: Zuckerkandl and Pauling's work, “Evolutionary Divergence and Convergence in Proteins,” explores the evolutionary processes of protein sequences and structures. They highlight how proteins evolve through divergence (where related proteins develop distinct structures and functions) and convergence (where unrelated proteins develop similar structures or functions due to similar selective pressures).

Link to

PDF

; Link to online version

2. Informed and automated k-mer size selection for genome assembly. Chikhi et al. Bioinformatics 2014, 30(1):31-7

Motivation: Genome assembly tools based on the de Bruijn graph framework rely on a parameter k, which represents a trade-off be- tween several competing effects that are difficult to quantify. There is currently a lack of tools that would automatically estimate the best k to use and/or quickly generate histograms of k-mer abundances that would allow the user to make an informed decision. Results: We develop a fast and accurate sampling method that con- structs approximate abundance histograms with several orders of magnitude performance improvement over traditional methods. We then present a fast heuristic that uses the generated abundance histo- grams for putative k values to estimate the best possible value of k. We test the effectiveness of our tool using diverse sequencing data- sets and find that its choice of k leads to some of the best assemblies. Availability: Our tool KMERGENIE is freely available at: http://kmergenie.bx.psu.edu.

Link to PDF

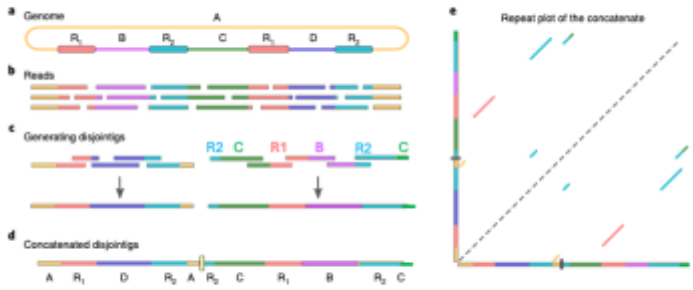
3. Assembly of long, error-prone reads using repeat graphs. Kolmogorov et al. 2019 Nature Biotech 73:540-546

Accurate genome assembly is hampered by repetitive regions. Although long single molecule sequencing reads are better able to resolve genomic repeats than short-read data, most long-read assembly algorithms do not provide the repeat character- ization necessary for producing optimal assemblies. Here, we present Flye, a long-read assembly algorithm that generates arbitrary paths in an unknown repeat graph, called disjointigs, and constructs an accurate repeat graph from these error-rid- dled disjointigs. We benchmark Flye against five state-of-the-art assemblers and show that it generates better or comparable assemblies, while being an order of magnitude faster. Flye nearly doubled the contiguity of the human genome assembly (as measured by the NGA50 assembly quality metric) compared with existing assemblers.

Link to

PDF

Figure 1 has mistakes in the colouring of the fragments. The corrected figure[10)] is shown below.



4. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Gabriel et al. 2024. Genome Res 34(5):769-777.

Gene prediction has remained an active area of bioinformatics research for a long time. Still, gene prediction in large eukaryotic genomes presents a challenge that must be addressed by new algorithms. The amount and significance of the evidence available from transcriptomes and proteomes vary across genomes, between genes, and even along a single gene. User-friendly and accurate annotation pipelines that can cope with such data heterogeneity are needed. The previously developed annotation pipelines BRAKER1 and BRAKER2 use RNA-seq or protein data, respectively, but not both. A further significant performance improvement integrating all three data types was made by the recently released GeneMark-ETP. We here present the BRAKER3 pipeline that builds on GeneMark-ETP and AUGUSTUS, and further improves accuracy using the TSEBRA combiner. BRAKER3 annotates protein-coding genes in eukaryotic genomes using both short-read RNA-seq and a large protein database, along with statistical models learned iteratively and specifically for the target genome. We benchmarked the new pipeline on genomes of 11 species under an assumed level of relatedness of the target species proteome to available proteomes. BRAKER3 outperforms BRAKER1 and BRAKER2. The average transcript-level F1-score is increased by about 20 percentage points on average, whereas the difference is most pronounced for species with large and complex genomes. BRAKER3 also outperforms other existing tools, MAKER2, Funannotate, and FINDER. The code of BRAKER3 is available on GitHub and as a ready-to-run Docker container for execution with Docker or Singularity. Overall, BRAKER3 is an accurate, easy-to-use tool for eukaryotic genome annotation.

Link to BRAKER3:PDF; Link to BRAKER2:

PDF

5. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics Levy et al. *Microbiome* volume 8, Article number: 48 (2020)

Background Metagenomics is revolutionizing the study of microorganisms and their involvement in biological, biomedical, and geochemical processes, allowing us to investigate by direct sequencing a tremendous diversity of organisms without the need for prior cultivation. Unicellular eukaryotes play essential roles in most microbial communities as chief predators, decomposers, phototrophs, bacterial hosts, symbionts, and parasites to plants and animals. Investigating their roles is therefore of great interest to ecology, biotechnology, human health, and evolution. However, the generally lower sequencing coverage, their more complex gene and genome architectures, and a lack of eukaryote-specific experimental and computational procedures have kept them on the sidelines of metagenomics.

Results MetaEuk is a toolkit for high-throughput, reference-based discovery, and annotation of protein-coding genes in eukaryotic metagenomic contigs. It performs fast searches with 6-frame-translated fragments covering all possible exons and optimally combines matches into multi-exon proteins. We used a benchmark of seven diverse, annotated genomes to show that MetaEuk is highly sensitive even under conditions of low sequence similarity to the reference database. To demonstrate MetaEuk's power to discover novel eukaryotic proteins in large-scale metagenomic data, we assembled contigs from 912 samples of the Tara Oceans project. MetaEuk predicted >12,000,000 protein-coding genes in 8 days on ten 16-core servers. Most of the discovered proteins are highly diverged from known proteins and originate from very sparsely sampled eukaryotic supergroups.

Conclusion The open-source (GPLv3) MetaEuk software (https://github.com/soedinglab/metaeuk) enables large-scale eukaryotic metagenomics through reference-based, sensitive taxonomic and functional annotation.

Link to PDF

6. A long-read RNA-seq approach to identify novel transcripts of very large genes. Uapinyoying et al. 2020 Genome Res 30: 885-897

RNA-seq is widely used for studying gene expression, but commonly used sequencing platforms produce short reads that only span up to two exon junctions per read. This makes it difficult to accurately determine the composition and phasing of exons within transcripts. Although long-read sequencing improves this issue, it is not amenable to precise quantitation, which limits its utility for differential expression studies. We used long-read isoform sequencing combined with a novel analysis approach to compare alternative splicing of large, repetitive structural genes in muscles. Analysis of muscle structural genes that produce medium (Nrap: 5 kb), large (Neb: 22 kb), and very large (Ttn: 106 kb) transcripts in cardiac muscle, and fast and slow skeletal muscles identified unannotated exons for each of these ubiquitous muscle genes. This also identified differential exon usage and phasing for these genes between the different muscle types. By mapping the in-phase transcript structures to known annotations, we also identified and quantified previously unannotated transcripts. Results were confirmed by endpoint PCR and Sanger sequencing, which revealed muscle-type-specific differential expression of these novel transcripts. The improved transcript identification and quantification shown by our approach removes previous impediments to studies aimed at quantitative differential expression of ultralong transcripts.

Link to

PDF

7. PureCLIP: capturing target-specific protein–RNA interaction footprints from single-nucleotide CLIP-seq data. Krakau et al. Genome Biology 2017, 18:240

The iCLIP and eCLIP techniques facilitate the detection of protein–RNA interaction sites at high resolution, based on diagnostic events at crosslink sites. However, previous methods do not explicitly model the specifics of iCLIP and eCLIP truncation patterns and possible biases. We developed PureCLIP (https://github.com/skrakau/PureCLIP), a hidden Markov model based approach, which simultaneously performs peak-calling and individual crosslink site detection. It explicitly incorporates a non-specific background signal and, for the first time, non-specific sequence biases. On both simulated and real data, PureCLIP is more accurate in calling

crosslink sites than other state-of-the-art methods and has a higher agreement across replicates.

Link to PDF

8. A Pan-plant Protein Complex Map Reveals Deep Conservation and Novel Assemblies McWhite et al. *Cell* Volume 181, Issue 2, 16 April 2020, Pages 460-474.e14

Summary
Plants are foundational for global ecological and economic systems, but most plant proteins remain uncharacterized. Protein interaction networks often suggest protein functions and open new avenues to characterize genes and proteins. We therefore systematically determined protein complexes from 13 plant species of scientific and agricultural importance, greatly expanding the known repertoire of stable protein complexes in plants. By using co-fractionation mass spectrometry, we recovered known complexes, confirmed complexes predicted to occur in plants, and identified previously unknown interactions conserved over 1.1 billion years of green plant evolution. Several novel complexes are involved in vernalization and pathogen defense, traits critical for agriculture. We also observed plant analogs of animal complexes with distinct molecular assemblies, including a megadalton-scale tRNA multi-synthetase complex. The resulting map offers a cross-species view of conserved, stable protein assemblies shared across plant cells and provides a mechanistic, biochemical framework for interpreting plant genetics and mutant phenotypes.

Link to PDF

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

New genome assemblies have been arriving at a rapidly increasing pace, thanks to decreases in sequencing costs and improvements in third-generation sequencing technologies1,2,3. For example, the number of vertebrate genome assemblies currently in the NCBI (National Center for Biotechnology Information) database4 increased by more than 50% to 1,485 assemblies in the year from July 2018 to July 2019. In addition to this influx of assemblies from different species, new human de novo assemblies5 are being produced, which enable the analysis of not only small polymorphisms, but also complex, large-scale structural differences between human individuals and haplotypes. This coming era and its unprecedented amount of data offer the opportunity to uncover many insights into genome evolution but also present challenges in how to adapt current analysis methods to meet the increased scale. Cactus6, a reference-free multiple genome alignment program, has been shown to be highly accurate, but the existing implementation scales poorly with increasing numbers of genomes, and struggles in regions of highly duplicated sequences. Here we describe progressive extensions to Cactus to create Progressive Cactus, which enables the reference-free alignment of tens to thousands of large vertebrate genomes while maintaining high alignment quality. We describe results from an alignment of more than 600 amniote genomes, which is to our knowledge the largest multiple vertebrate genome alignment created so far.

Link to PDF

10. Highly accurate protein structure prediction with AlphaFold Jumper et al. *Nature* volume 596, pages 583–589 (2021)

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort1,2,3,4, the structures of around 100,000 unique proteins have been determined5, but this represents a small fraction of the billions of known protein sequences6,7. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence—the structure prediction component of the 'protein folding problem'8—has been an important open research problem for more than 50 years9. Despite recent progress10,11,12,13,14, existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14)15, demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.

Link to PDF

11. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Steinegger and Salzberg *Genome Biology* Vol. 21 115 (2020)

Genomic analyses are sensitive to contamination in public databases caused by incorrectly labeled reference sequences. Here, we describe Conterminator, an efficient method to detect and remove incorrectly labeled sequences by an exhaustive all-against-all sequence comparison. Our analysis reports contamination of 2,161,746, 114,035, and 14,148 sequences in the RefSeq, GenBank, and NR databases, respectively, spanning the whole range from draft to "complete" model organism genomes. Our method scales linearly with input size and can process 3.3 TB in 12 days on a 32-core computer. Conterminator can help ensure the quality of reference databases. Source code (GPLv3): https://github.com/martin-steinegger/conterminator

Link to PDF

12. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. Baid et al. *Nature Biotechnology* Vol. 41 (2023)

Circular consensus sequencing with Pacific Biosciences (PacBio) technology generates long (10–25 kilobases), accurate 'HiFi' reads by combining serial observations of a DNA molecule into a consensus sequence. The standard approach to consensus generation, pbccs, uses a hidden Markov model. We introduce DeepConsensus, which uses an alignment-based loss to train a gap-aware transformer–encoder for sequence correction. Compared to pbccs, DeepConsensus reduces read errors by 42%. This increases the yield of PacBio HiFi reads at Q20 by 9%, at Q30 by 27% and at Q40 by 90%. With two SMRT Cells of HG003, reads from DeepConsensus improve hifiasm assembly contiguity (NG50 4.9 megabases (Mb) to 17.2 Mb), increase gene com- pleteness (94% to 97%), reduce the false gene duplication rate (1.1% to 0.5%), improve assembly base accuracy (Q43 to Q45) and reduce variant-calling errors by

24%. DeepConsensus models could be trained to the general problem of analyzing the alignment of other types of sequences, such as unique molecular identifiers or genome assemblies.

Link to

PDF

13. Clustering predicted structures at the scale of the known protein universe. Barrio-Hernandez et al. Nature 622: 637–645 (2023)

Proteins are key to all cellular processes and their structure is important in understanding their function and evolution. Sequence-based predictions of protein structures have increased in accuracy1, and over 214 million predicted structures are available in the AlphaFold database2. However, studying protein structures at this scale requires highly efficient methods. Here, we developed a structural-alignment-based clustering algorithm—Foldseek cluster—that can cluster hundreds of millions of structures. Using this method, we have clustered all of the structures in the AlphaFold database, identifying 2.30 million non-singleton structural clusters, of which 31% lack annotations representing probable previously undescribed structures. Clusters without annotation tend to have few representatives covering only 4% of all proteins in the AlphaFold database. Evolutionary analysis suggests that most clusters are ancient in origin but 4% seem to be species specific, representing lower-quality predictions or examples of de novo gene birth. We also show how structural comparisons can be used to predict domain families and their relationships, identifying examples of remote structural similarity. On the basis of these analyses, we identify several examples of human immune-related proteins with putative remote homology in prokaryotic species, illustrating the value of this resource for studying protein function and evolution across the tree of life.

Link to online version and

PDF

14. Fast and sensitive taxonomic assignment to metagenomic contigs. Mirdita et al. Bioinformatics 37(18):3029–3031 (2021)

Summary
MMseqs2 taxonomy is a new tool to assign taxonomic labels to metagenomic contigs. It extracts all possible protein fragments from each contig, quickly retains those that can contribute to taxonomic annotation, assigns them with robust labels and determines the contig’s taxonomic identity by weighted voting. Its fragment extraction step is suitable for the analysis of all domains of life. MMseqs2 taxonomy is 2–18× faster than state-of-the-art tools and also contains new modules for creating and manipulating taxonomic reference databases as well as reporting and visualizing taxonomic assignments.
Availability and implementation
MMseqs2 taxonomy is part of the MMseqs2 free open-source software package available for Linux, macOS and Windows at https://mmseqs.com.

link to Paper

15. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Cheng et al. Science 381:eadg7492 (2023)

The vast majority of missense variants observed in the human genome are of unknown clinical significance. We present AlphaMissense, an adaptation of AlphaFold fine-tuned on human and primate variant population frequency databases to predict missense variant pathogenicity. By combining structural context and evolutionary conservation, our model achieves state-of-the-art results across a wide range of genetic and experimental benchmarks, all without explicitly training on such data. The average pathogenicity score of genes is also predictive for their cell essentiality, capable of identifying short essential genes that existing statistical approaches are underpowered to detect. As a resource to the community, we provide a database of predictions for all possible human single amino acid substitutions and classify 89% of missense variants as either likely benign or likely pathogenic.

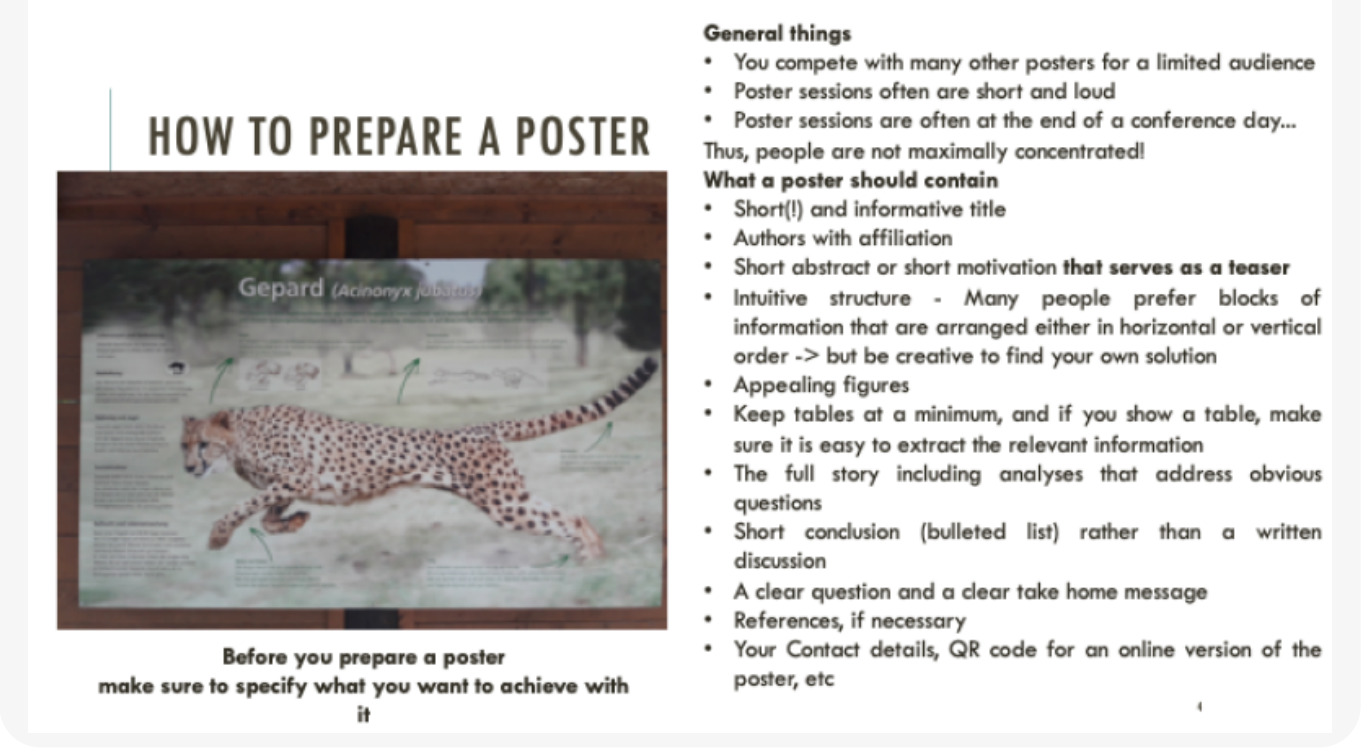
link to Paper

Paper Auswahl

1. Schauen Sie sich bitte die vorgestellten Publikationen genau an
2. Wählen Sie dann bitte drei Publikationen aus, mit denen Sie sich auseinandersetzen wollen
3. Melden Sie sich bitte im Auge System der Goethe Universität an. Sie finden hier die Publikationen aufgeführt
4. Treffen Sie Ihre Erst-, Zweit-, und Drittwahl

- Wenn Sie schon ein Team gebildet haben, können Sie uns das gerne mitteilen (bis zum 11. Mai). Dennoch müssen beide Teammitglieder bitte abstimmen.
- Es hilft, wenn Sie sich in der Gruppe kurz vorab besprechen, damit sich nicht alle auf zwei oder drei Paper stürzen

Erstellung einer Präsentation



Sobald es an die Erstellung einer Präsentation geht stellt sich automatisch die Frage:“*Was muss ich dabei beachten?*”. Auf diese Frage gibt es, meiner Meinung nach, keine allgemeingültige Antwort. Vielmehr sollte man sich zunächst fragen:“Was möchte ich mit meiner Präsentation erreichen?”. Ob die präsentierende Person sich darüber im Klaren ist, merkt man daran, wie präzise Fragen zur Erstellung einer Präsentation gestellt werden. Fragen wie “*Wie schaffe ich dies und das durch meine Präsentation zu erreichen?*” zeugen davon, dass sich jemand mit der Präsentationsproblematik schon klar auseinandergesetzt hat.

Im angehängten

PDF

sind nun darüber hinaus ein paar generelle Punkte angemerkt.

Paper Zuweisung

Paper-Id	User
P01	
P02	Boudouassel, Ioab
P03	Deng, Sarach
P04	Biesecker, Inciler
P05	Kolos, Tahir
P06	Fischer, Le
P07	Barié, Dieter
P08	
P09	Bernshausen, Chanthirakanthan
P10	Alkanat, Paraparan
P11	Berger, Voss

Paper-Id	User
P12	
P13	Batman, Zeng
P14	Ashirov, Qin
P15	Li, Sahin

Paper summaries

Erstellen Sie nun gemeinsam mit Ihrer Partnerin oder Ihrem Partner eine Zusammenfassung des Papers. Unterteilen Sie Ihre Zusammenfasssung in

1. Bearbeitete Forschungsfrage(n)
2. Relevante Methodische Ansätze
3. Relevante Ergebnisse
4. Schlussfolgerung
5. Schlüsselabbildung → Wählen Sie hier eine Abbildung aus der Publikation die Sie als die wichtigste Abbildung ansehen, und beschreiben Sie kurz mit eigenen Worten was die Abbildung aussagt, und warum Sie diese als Schlüsselabbildung ansehen. Für den Fall, dass Ihr Paper keine relevanten Abbildungen hat, erstellen Sie bitte eine neue Abbildung, die den Zweck einer Schlüsselabbildung übernehmen kann.

Format der Abgabe: PDF

Deadline für die Abgabe: siehe oben

Ort der Abgabe: OLAT

Die eingegangen Zusammenfassungen finden Sie hier:

PDF

Paper abstracts

Die vorige Aufgabe hatte zum Ziel die wesentlichen Kernaspekte 'Ihres' Papers herauszuarbeiten, und die Ergebnisse finden im Kurswiki. Wenn Sie sich die Zusammenfassungen einmal genau anschauen, werden Sie feststellen, dass eine kurze Einführung in das Thema fehlt.

Die Zeitschrift *Nature* stellt ein sehr hübsches Beispiel zur Verfügung wie man eine Zusammenfassung (Abstract) aufbaut

Figure 8: Nature journal's guideline of how to write an abstract

Das Schreiben einer Zusammenfassung (Abstract)

Das Schreiben einer Zusammenfassung ist vermutlich der schwierigste Teil bei der Erstellung eines Manuskripts, und man sollte das nur angehen, wenn man einen vollen Überblick über das Projekt hat. Ihr Abstract dient quasi als Angelhaken, mit dem Sie potentielle Leser und Leserinnen 'fangen'. Entsprechend ansprechend und informativ sollte Ihr Abstract sein!

Ein gutes Abstract sollte folgende Inhalte haben:

- Einen kurzen Überblick über das Forschungsfeld in das sich Ihre Arbeit einbettet (3-4 Sätze)
- Das Präzisieren einer Wissenslücke, die Sie mit Ihrer Studie füllen wollen (1 Satz)
- Ihre Forschungsfrage (1 Satz)
- Die wesentlichen Ergebnisse (2-3 Sätze)
- Die wesentlichen Schlussfolgerungen (2-3 Sätze)

Insgesamt haben Sie bei Manuskripten in der Regel nur zwischen 150 und 300 Wörter für Ihr Abstract.

Aufgabenstellung

Ihre nächste Aufgabe besteht nun darin, ein Abstract für Ihr Paper zu schreiben. Um die Struktur in Ihrem Abstract hervorzuheben,  verwenden Sie bitte eine Farbkodierung, die der der Vorlage entspricht.

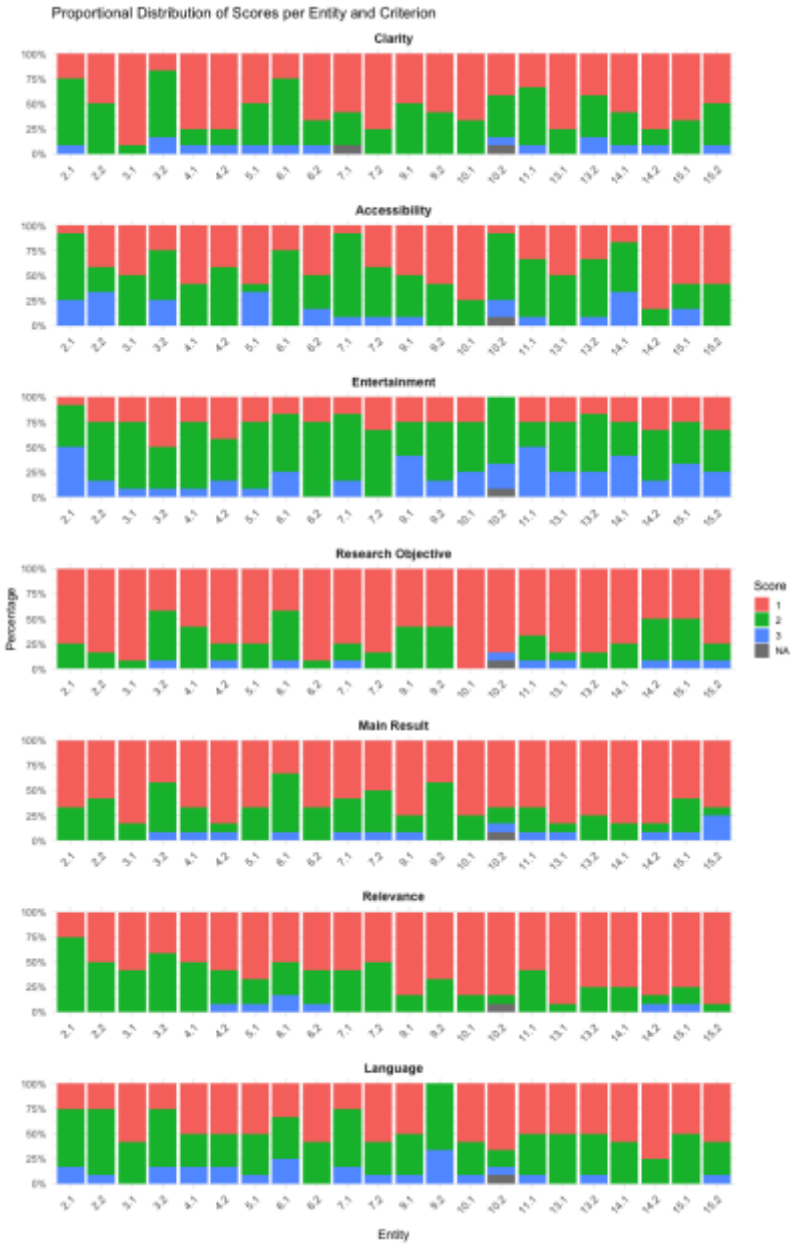
Schreiben müssen wir alle lernen, entsprechend erstellen Sie das Abstract bitte NICHT in Gruppenarbeit! Ich erwarte in diesem besonderen Fall Einzelabgaben.

Alle Abstracts sind hier zum Download verfügbar:

LINK

Peer review

Ihre Aufgabe war nun, die Abstracts Ihrer Gruppe zu lesen und zu bewerten. Das Ergebnis entnehmen Sie bitte der folgenden Abbildung.



Paper impact

Problem: Es gibt Unmengen an Publikationen zu einem Thema, welche sollen wir lesen?

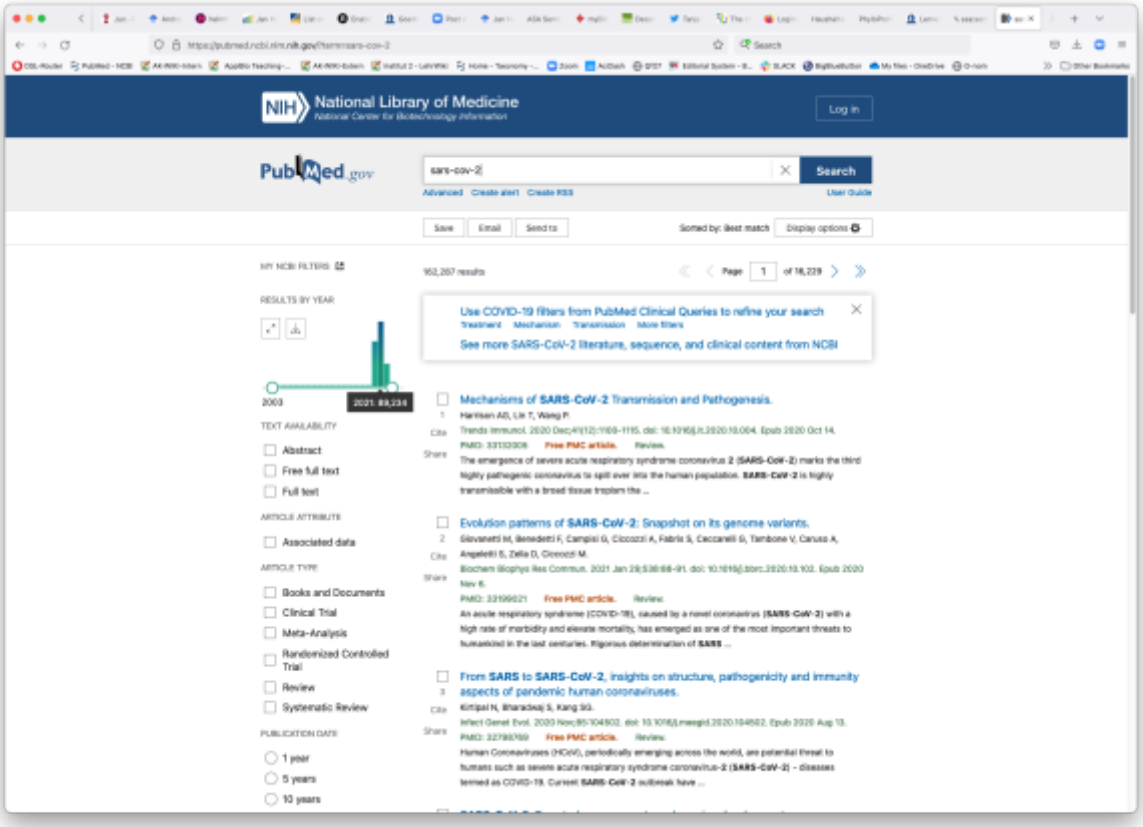


Figure 9: Number of publications connected to SARS-Cov-2 listed in NCBI Pubmed(accessed May 24 2022)

Lesen Sie Ihr Paper erneut und berücksichtigen Sie nun die Referenzen, die herangezogen werden, um Aussagen, die im Paper gemacht werden zu belegen. Erstellen Sie eine Liste mit den 5 Referenzen, die sie als die relevantesten betrachten, die also das wesentliche wissenschaftliche Fundament des Papers darstellen! Geben Sie folgende Informationen:

```
      die Autoren (Erstautor, korrespondierender Autor und Letztautor)
      Titel der Arbeit
      Journal (Ausgabe und Seitenzahl)
      Erscheinungsjahr
      Anzahl der Zitationen
      Begründen Sie kurz warum Sie dieses Paper als relevant erachten
  Die Bedeutung eines Papers wird häufig daran gemessen in wievielen
nachfolgenden Studien auf dieses Paper verwiesen wird. Ermitteln Sie die
Gesamtanzahl der Zitationen Ihres Papers, und bestimmen Sie die Zitationen
pro Jahr. Nennen Sie die ihrer Meinung nach 5 einflussreichsten Studien, die
Ihr Paper zitieren. Berücksichtigen Sie hierbei den Impact Factor des
Journals. NCBI Pubmed und Google Scholar werden Ihnen hier gute Dienste
leisten.
  Lernen Sie 'Ihren' korrespondierenden Autor besser kennen. Schauen Sie
sich ihre/seine Publikationsliste an und ermitteln Sie die 5 relevantesten
Publikationen dieser Person. Berücksichtigen Sie hierbei den Journal Impact
Factor, die Anzahl der Zitationen und die Position in der Autorenliste.
Begründen Sie Ihre Entscheidung kurz.
```

Bitte stellen Sie Ihre Angaben so zusammen, dass Sie möglichst auf eine DinA4 Seite im Querformat

passen. Das Ziel wird sein unsere Paper-Summary weiter auszubauen.

Dateiformat: PDF

Link zur Abgabe:

PDF

Was haben wir geschafft?

- Übersicht des Artikels
- Zusammenfassung des Artikels (Abstract)
- Impact des Artikels und des Autors

Was bleibt zu tun?

- Zusammenfassung aller Dokumente in einem 'Abstract book'. **Deadline: siehe oben**
- Erstellung eines Flash talk Videos inklusive eines Quiz. **Deadline: siehe oben**
- Erstellung des Posters. **Deadline: siehe oben**

Flash talks

- 2022
- 2024
- 2025

Poster presentation

- Poster 2024

For the compilation of your poster make sure to follow the guidelines provided in the OLAT. Find below the information (in German only)

```
Erstellen Sie bitte für 'ihre' Publikation ein wissenschaftliches Poster auf dem Sie die wissenschaftlichen Ziele/Fragen und die relevanten Methoden und Ergebnisse zusammenfassen. Dieses Poster wird dann von Ihnen im Rahmen einer Postersession vorgestellt und dient dann als Grundlage für die Diskussion Ihres Themas. Bei der Erstellung achten Sie bitte noch einmal explizit auf folgende Punkte
```

```
Obligatorisch

    Das Format muss Din A0 im Hochformat sein (841 x 1189 mm)
    Das Dateienformat muss PDF sein.
    Schnittmarkierungen (Schnittrahmen oder Schnittmarker) müssen eingefügt
werden, da sonst ein automatisches Schneiden durch die Druckerei nicht
möglich ist
    Das Poster trägt den Titel der Publikation und die Orignial-Autoren
müssen unter dem Titel aufgeführt sein. Sollten mehr als vier Autoren auf
dem Paper vertreten sein, dann nennen Sie die erste drei und fassen die
restlichen Autoren unter 'et al.' zusammen
    Die erstellenden Personen müssen auf dem Poster genannt werden
    Jedes Poster beginnt mit einer Zusammenfassung bzw einem Abstract, der
nicht mehr als 300 Wörter betragen darf. Die Struktur des Abstracts folgt
der Vorgabe von Nature:
https://cbs.umn.edu/sites/cbs.umn.edu/files/public/downloads/Annotated_Natur
e_abstract.pdf
    Die Forschungsfrage muss klar erkenntlich sein
    Abbildungen müssen von ausreichend hoher Qualität sein, dass sie auf
einem A0 Ausdruck nicht unscharf werden
    Das Poster soll den roten Faden durch die Studie darstellen
    Text soll sich auf das maximal notwendige beschränken (Anmerkung: Ein
Poster lebt davon präsentiert zu werden). Langer Fließtext sollte vermieden
werden (Ausnahme: Abstract)
    Die Schriftgröße muss so groß sein, dass sie auch aus 1 1/2 - 2 m
Entfernung gelesen werden kann
    Tabellen sind möglich, sie sollten aber einfach verständlich gehalten
werden
    Der 'Weg' durch das Poster muss sich dem Betrachter auch ohne Erklärung
erschließen
    Abbildungen und Tabellen müssen in der Regel mit Nummer und Titel und
ggf. einer kurzen Beschreibung versehen sein. In begründeten(!) Einzelfällen
(je nach Layout-Wahl) kann dies aber entfallen
    Eine exzessive Verwendung von Farben sollte vermieden werden

Optional

    Relevante Referenzen können in einer Referenzliste zusammengefasst
werden
    Mittels eines QR-Codes kann auf eine elektronische Version des Posters
verwiesen werden
```

See

this presentation

for some nice further hints about what to consider when making a poster.

Erstellung und Einreichung eines Posters für eine

Postersession

Beachten Sie bitte die folgenden Vorgaben im Zusammenhang mit der Erstellung und Einreichung eines Posters für eine Postersession:

- Deadline: Die Deadline für das einreichen der Poster ist zwingend einzuhalten. Verspätet eintreffende Poster werden nicht auf AK Kosten gedruckt
- Größenvorgabe: **DIN A0 im Hochformat (841 x 1189 mm)**.

Poster, die nicht diesem Format entsprechen verursachen zusätzliche Arbeit und Kosten, die der AK nicht übernehmen wird!

- Schnittrahmen / Schnittmarker: Fügen Sie auf ihrem Poster einen Schnittrahmen oder Schnittmarker hinzu, damit die Druckerei den korrekten Zuschnitt durchführen kann (eine Erläuterung zu diesem Thema, finden Sie auch in dieser Anleitung)
- Abbildungen: Achten Sie auf eine **ausreichend hohe Auflösung der Abbildungen**, die auch bei 100% Skalierung ihres Posters nicht 'verpixelt' sind
- Dateiformat: PDF
- Der Druck Ihres Posters erfolgt durch die Universität/Lehrstuhl
  - und **darf nur nach separater Aufforderung** der Universität durch Ihre eigene Beauftragung einer Druckerei erfolgen
  - Liegt keine separate Aufforderung vor, werden die anfallenden Kosten für den Posterdruck nicht erstattet.

Abstract Book

Für die Erstellung des Abstract Books, folgen Sie bitte den Anweisungen im OLAT.

1)

Worin liegt das Problem, wenn mehr Personen an der Veranstaltung teilnehmen?

2)

Struktur und Text des Abstracts ermöglichen es die relevanten Aspekte des Papers klar zu erfassen

3)

Der Text des Abstracts ermöglicht es auch weniger erfahrenen Personen die relevanten Aspekte des Papers zu erfassen

4)

Es macht Spaß das Abstract zu lesen

5)

Das Ziel der Arbeit ist klar aus dem Abstract zu entnehmen

6)

Das wesentliche Ergebnis der Arbeit ist einfach zu erfassen

7)

Die Bedeutung der Studie für das Forschungsfeld (oder darüber hinaus) ist klar erkennbar

8)

Wie schätzen Sie die Qualität des Englischen ein?

9)

Diese finden Sie auf jedem Abstract in der linken oberen Ecke
[10)]

by Ingo Ebersberger