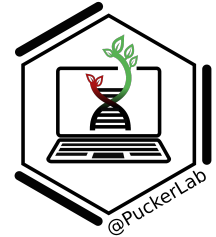
UNIVERSITÄT BONN



Prof. Dr. Boas Pucker

Python - Individual projects

UNIVERSITÄT BONN

P1: FASTA cleaner (easy)

- Read FASTA-like files and convert them into properly formatted FASTA files
- Remove non-standard characters (e.g. '\r' introduced on Macs)
- Include support for docx files
- Present basic statistics about the file (number of sequences, lengths, …)
- Opportunities: inclusion in webserver (https://www.pbb-tools.de/)

UNIVERSITÄT BONN

P2: CDS extractor (moderate)

- Extract all coding sequences (CDS) from a given GFF3+FASTA files
- GFF3 (generic feature format) are highly diverse and require considerations of numerous special cases and exceptions
- Run a test based on various NCBI and other data sets
- Opportunities: enables numerous data-driven projects

UNIVERSITÄT BONN

P3: Update seqex3.py (easy)

- Seqex3.py can extract a specified target region from a given FASTA file
- Update is needed to supply numerous target regions through a text file table
- Read table and extract all specified regions

UNIVERSITÄT BONN

P4: Impact Factor (IF) inflation? (moderate)

- Are the IFs of all (established) journals increasing?
- Collect data from the internet and store these in a text file
- Visualize the data with Python
- Perform statistical analyses to support your hypotheses
- Opportunity: commentary article about findings

UNIVERSITÄT BONN

P5: Ortholog expression plotter (moderate)

- Orthologs are the same genes in different species
- Studying gene expression across multiple species is very enlarges data sets
- Visualize and analyze gene expression across multiple species based on identified orthologs in all/some of these species
- Opportunity: basis for advanced gene expression studies in the future

UNIVERSITÄT BONN

P6: Screen KIPEs results for multifunctions (challenging)

- Enzyme families like FLS/ANS/F3H can have overlapping enzymatic functions
- Enzymes of one lineage can have side activities of the others
- Some enzymes might have multiple functions (promiscuity)
- Process KIPEs results to identify and highlight multifunctional enzyme candidates
- Opportunity: additional module for next KIPEs version

UNIVERSITÄT BONN

P7: MYB quizz (easy)

- Update the MYB quiz (https://colab.research.google.com/drive/1gBvRBkO-aGf3VJWkKc3aG-oVuBo80D2z?usp=drive_link)

- Nice graphical user interface, Highscore, Multiplayer, …

UNIVERSITÄT BONN

P8: BUSCO result discrepancies (moderate)

- BUSCO (https://busco.ezlab.org/) assesses the completeness of sequence data sets
- Better scores for genome sequences than for corresponding annotation
- What causes this difference?

UNIVERSITÄT BONN

P9: splice site conservation check (challenging)

- Splice sites (GT…AG) determine splitting points in mRNAs
- Non-canonical splice sites are almost unexplored
- Splice sites can be checked across species borders
- Conserved non-canonical splice sites would indicate functional relevance

UNIVERSITÄT BONN

P10: snp2marker (easy)

- Sequence variants are stored in VCF files
- Genetic markers can be developed based on sequence variants
- Develop a script for the automatic design of KASP markers (suitable for web server)
- Input: VCF + genomic FASTA file

UNIVERSITÄT BONN

P11: intron extraction for phylo tree (challenging)

- Phylogenetic trees are often based on coding sequences
- Intron sequences are more variable i.e. have more phylogenetic signal
- Extract corresponding intron sequences from a range of species and construct a phylogenetic tree based on these sequences

UNIVERSITÄT BONN

P12: tissue-specific expression analysis (challenging)

- Tissue-specific gene expression analysis
- Example: glucosinolate biosynthesis genes in *Arabidopsis thaliana*
- Process metadata from the Sequence Read Archive
- Generate heatmap
- Identify subfunctionalization of gene copies based on expression patterns

UNIVERSITÄT BONN

P12: map data to tree (moderate)

- Mapping of data onto phylogenetic trees
  - Boolean values (0/1)
  - Values (integers, float)
- Input: table with data + phylogenetic tree (names need to match)
- Tree visualization with iTOL possible (annotation file)
- Data inclusion in phylogenetic tree file (e.g. NEXUS)

UNIVERSITÄT BONN

P13: Assembly stats collector (challenging)

- Identify publications describing genome assemblies (database searches)
- Collect important statistics (N50, contig number, genome size, biological motivation)
- Generate a table and summarize data in figure

UNIVERSITÄT BONN

P14: Variant position visualization (moderate)

- Show sequence variants relative to genes
- Load positions of genes from GFF3 file
- Load positions of variants from VCF file
- Comparison of multiple samples (genotypes)
- Visualization with matplotlib/plotly

UNIVERSITÄT BONN

P16: Primer designer (easy)

- Automatically design primers for a given input sequence or region
- Check against a reference genome sequence for unique binding
- Predict various primer properties
- Enable batch mode for large number of primer designs

UNIVERSITÄT BONN

P17: pep2cds (easy)

- Phylogenetic tree construction requires replacement of amino acids by codons
- Input: CDS FASTA + PEP FASTA + PEP alignment
- Include checks for invalid inputs and offer solutions for errors

UNIVERSITÄT BONN

P18: TE illustrator (easy)

- Display all repeats and transposons around a gene of interest
- Display gene structure and details about the TEs
- Automatic labeling of features
- Support for different figure types

UNIVERSITÄT BONN

Project presentation expectations / grading

- Completeness and complexity of the project
- Quality of the documentation
- Quality of the presentation
  - Motivation / problem
  - Strategy
  - Innovative features
  - Clarity of presentation (overview + interesting details)
  - Length of presentation (10 minutes)
- Upload PDF and script in cloud folder afterwards
- Competency answering questions

UNIVERSITÄT BONN

Time for questions!