METK Barley SNP-Chip: Exploring the correlation between barley's genetic makeup and its protein content

Team members: Agnes Kivistik, Liselle Velner, Alexandra Voit
https://github.com/alexandravoit/BARLEY-SNP-CHIP.git

Business understanding

- ## Identifying your business goals

    - **Background:**

Barley is a nutritious grain that can grow even in harsher climates and contains considerable amounts of fiber, vitamins, and minerals. It is widely used as animal feed, however, it is also a staple in Estonia's traditional cuisine. We obtained very interesting private data from the METK biotechnology department about barley harvested in 2024 which in turn inspired our research: barley's agricultural significance led us to explore the potential connection between the plant's various attributes and its SNP composition (meaning its DNA makeup). We decided to narrow our scope to the grain's protein content, as plant-based diets are becoming increasingly popular, and the protein amount per 100g of barley is both nutritionally important for consumers as well as crucial for farmers when selecting crop varieties.

    - **Business goals:**

The business goal of this project is to lay the groundwork for improving barley's nutritional composition by hopefully identifying genetic factors linked to its protein content. By establishing a connection, we could facilitate larger studies that take into account barley's other attributes as well - these include grain size, proneness to diseases, growth time etc. This knowledge could prove useful in crop breeding.

    - **Business success criteria:**

The success of this project will be measured by some qualitative and quantitative criteria. First, we would like to see a clear correlation between barley's genetic attributes and its protein content. The ultimate goal would be to achieve a correlation threshold of at least 0.7. Second, if we decide to create a prediction model, it should have an accuracy rate of 70% or higher in forecasting protein content based on genetic data. Lastly, our models should be efficient enough that the data provided will not take longer than 30 minutes to process.

● Assessing your situation

- **Inventory of resources:**

The research project relies on two key datasets. The team consists of three people who possess some knowledge of data analysis. Additional guidance is available from METK mentors specializing in biotechnology. Computing resources include personal computers and data analysis tools.

- **Requirements, assumptions, and constraints:**

The project has a strict deadline of December 9th, and all work must be completed before this date. Constraints also include a 30-minute time limit for processing the data and the sensitive nature of the private datasets, necessitating careful handling and storage.

- **Risks and contingencies:**

Several risks could affect the progress of the project. First, there may be challenges in filtering data, particularly in ensuring compatibility between the two datasets. Second, if calculations exceed the time limit, we will need to optimize the code. Lastly, there could be data alignment issues, such as mismatched or incomplete barley variety names across datasets. This will be addressed by removing non-overlapping varieties during data cleaning.

- **Terminology:**

SNPs, or Single Nucleotide Polymorphisms, represent specific genetic variations in the barley genome, which may correlate with traits such as protein content. Protein content refers to the percentage of protein in 100g of barley and is a key focus of this study. Genome-wide association involves analyzing genetic markers across the genome to identify links with specific traits. DNA nucleotides — adenine (A), cytosine (C), guanine (G), and thymine (T) — are the building blocks of genetic data and are central to interpreting SNP information.

- **Costs and benefits:**

The main cost of the project is the time invested by the team, estimated at approximately 60 hours collectively. The study will provide valuable insights into the genetic factors influencing barley's protein content, laying a foundation for future research in crop improvement. Additionally, the project will give team members practical experience in data science and genetics, enhancing our skills and knowledge in these fields.

● Defining your data-mining goals

- **Data-mining goals:**

One of the goals is to create a phylogenetic tree that reflects the genetic diversity of grain samples using SNP data. Develop a machine learning model that can predict the protein content of grain based on SNP markers. Identify genetic markers that are key drivers of protein levels in the barley grain.The final deliverables will be a detailed poster and a presentation that convey the results of the phylogenetic analysis, the predictive model, and the SNP-protein associations.

- **Data-mining success criteria:**

The model should achieve a prediction error of no more than 10% when predicting the protein content in barley based on SNP data. The model should successfully identify the best SNPs that are most significantly associated with protein content in barley. The model should handle missing SNP data effectively, with imputation techniques ensuring at least 90% accuracy in filling in missing values. The training and operational phases of the model should not exceed 30 minutes in runtime for datasets with typical sizes.

Data understanding

Gathering data

- **Outline data requirements:**

The data requirements for this research include two datasets: one dataset containing information about different barley varieties' attributes (including the variety name, string, and protein content per 100g of barley, float) and the other dataset should have information about the genetic composition of these same varieties, specifically their SNP data as a char (a, t, c, g). The SNP dataset should include the genetic markers associated with each barley variety, allowing for the exploration of potential correlations between the varieties' genetic makeup and their protein content or other relevant traits. Naturally, both datasets should be about barley harvested in the year 2024. The datasets should be in a csv format as it includes less formatting data and should be faster to process.

- **Verify data availability:**

The data is entirely available to us and has already been added to the project's github repository. Currently it is in a .xls format, however, it will be converted to csv. There might not be

an exact overlap between the barley varieties addressed in the two datasets so while cleaning the data we must address this concern and remove all varieties that are not present in both datasets.

- **Define selection criteria:**

The two datasets we will be using are “BARLEY_DATA.xls” and “SNIP_DATA.xlsx”. From “SNIP_DATA.xlsx” we will analyse all columns (variety name) that are present in the other dataset too. We will look at rows containing genetic info that do not have too many ‘failed’ values nor only contain one and the same allele across all barley varieties. From “BARLEY_DATA.xls” we will only analyse the last column that contains the protein percentage in the different varieties of barley.

● Describing data:

- **Dataset 1 “BARLEY_DATA.xls”:**
  Dataset 1 has 16 columns and 192 rows containing data about specific varieties + the last LSD row. First column basically serves as an id attribute and marks a variety’s number in the dataset (integer). Second column is for the variety’s name (string). Third column marks the total yield measured in kilograms per hectare (integer). Then there is a column for the stability rating of the barley seeds (1-9 scale). Seventh column addresses the mass of 1000 barley seeds, measured in grams (float). The next column records the height of the barley plants, measured in centimeters (integer).Then three columns for data concerning growth time periods measured in days (integer). Another four columns for data about the plant’s susceptibility to various diseases measured on a scale from 1-9 (integer). And finally a column regarding the variety’s protein content in percentages (float).

- **Dataset 2 “SNIP_DATA.xlsx”:**
  Dataset 2 has 3 800 rows and 189 columns of raw data of barley SNP-chip analysis. The columns are for names of different barley varieties and the rows indicate the presence of different alleles (T, C, G, A, or "failed"). Each allele corresponds to a specific genetic variation at that marker for the given barley variety.

● Exploring data:

For the protein content data from "BARLEY_DATA.xls," values range between 10.4% and 15.5%, with an average of approximately 12%, and median of 12%. The “SNIP_DATA.xlsx” dataset was explored to determine the frequency of each allele (T, C, G, A) and assess the prevalence of "failed" values, which represent data points that could not be read. Initial findings

indicated that certain SNPs had consistent alleles across all barley varieties, which might not contribute to genetic variation analysis.

- **Verifying data quality:**

For BARLEY_DATA.xls, the protein content data was validated to confirm they fell within the expected ranges, and there were no missing values. For SNIP_DATA.xlsx, the extent of "failed" SNP values was assessed to ensure that they did not compromise the integrity of the dataset. Rows with excessive failed values or limited genetic variability (e.g., SNPs with the same allele across all varieties) were flagged. Overall, no major issues were found that would prevent moving forward with the analysis.

Task 4. Project plan

TASK	METHODS / TOOLS	ESTIMATED TIME	WHO	COMMENTS
Convert datasets from .xls to .csv format for easier processing	Python / Pandas	1	Alexandra	These tasks are critical to ensure the datasets are properly aligned before analysis and model building.
Identify and address missing or "failed" SNP values in **SNIP_DATA.xlsx** and clean any inconsistencies	Python / Pandas	2	Agnes	
Align both datasets based on barley variety names to ensure consistency.	Python / Pandas	1	Liselle	
Remove non-overlapping barley varieties between the two datasets	Python / Pandas	2	Liselle	
For **SNIP_DATA.xlsx**, evaluate the frequency of alleles (T, C, G, A) for each SNP marker and look for consistent patterns	Python / Pandas	2	Alexandra	
Select SNP features that are most likely to have a significant impact on protein content.	Python / Pandas	2	Agnes	
Normalize SNP data for use in machine learning models.	Python / Pandas	3	Alexandra	The most time-consuming task, as it requires fine-tuning models and ensuring
Separate features (SNPs) and target variable (protein content) for training the	Python + libraries	2	Liselle	

prediction model.				optimal performance. We will prioritize model efficiency so that the runtime stays within the 30-minute limit
Develop machine learning models to predict protein content based on SNP data.	We will try different models such as Linear Regression, Random Forest etc	4	Alexandra, Liselle, Agnes	
Evaluate models using cross-validation and compare performance metrics	Accuracy, mean error, standard error (Scikit-learn)	3	Alexandra, Liselle, Agnes	
Localise important SNP markers on genome map	Manual list lookup	1	Agnes	
Create a phylogenetic tree of barley varieties based on their genetic data	Python (Matplotlib, Seaborn for visualizations)	2	Agnes	
Develop a poster for the final project presentation.	Adobe illustrator / Canva	2	Alexandra, Liselle	
Prepare for a final presentation		2	Agnes	