ISSN:2790-1688 Volume-14-(2025)

Computer Extraction and Analysis of Genomic DNA of Mustard Plant

Zhenghao Huang

University of California Davis, Davis, 95616, United of America

zhhuang@ucdavis.edu

Abstract. Brassica juncea, a plant of the family Brassica, has high economic, medicinal and nutritional value and is an important subject of genomic research. With the development of bioinformatics and gene editing technologies, this study screened and annotated key gene sequences in the genome of Brassica juncea using platforms such as Gensas. Computational modeling enables researchers to predict gene function and protein structure without relying solely on experimental methods, greatly improving efficiency. The predictive power of these models is highly dependent on the availability of homologous data and must be rigorously validated with error thresholds considered. In addition, tools such as BLAST and UniProt were used for functional analysis and protein sequence alignment. Studies have shown that the addition of structural data can significantly improve the reliability of prediction models. Ultimately, this study helps to deepen the understanding of gene function and pave the way for advances in plant breeding, drug development, and biotechnology applications.

Keywords: Genome sequencing, Gene editing, Mustard plant, Homology modeling, Protein structure prediction.

1. Introduction

Mustard plants belong to the Brassicaceae family and are widely cultivated due to their economic, nutritional, and medicinal value. Their roots, seeds, and leaves are used not only in culinary applications but also in traditional medicine to treat various diseases such as indigestion, respiratory infections, and edema. In recent years, scientific interest in cruciferous plants has increased because of their high antioxidant content, particularly sulforaphane, which has been linked to cancer prevention and immune regulation. Despite their value, there is still a lack of comprehensive genomic studies that identify the key functional genes responsible for these beneficial traits.

This study focuses on the extraction and computer-based analysis of the genomic DNA of Bractococcus mustard plants using modern bioinformatics tools. The central research question is: Can computational gene editing and annotation methods reveal novel functional genes in mustard plants that are potentially beneficial for medical or agricultural use? To address this, we apply tools such as Gensas for genome annotation, BLAST for homology analysis, and UniProt for protein function identification. By constructing a computer model of the mustard plant genome, we aim to accurately predict gene sequences and protein structures without the need for exhaustive experimental procedures.

The purpose of this research is to establish a systematic workflow for genome extraction, gene screening, and functional modeling. The significance lies in improving our understanding of plant molecular biology, aiding in drug discovery, and informing plant breeding strategies. Furthermore, this research may serve as a model for genomic studies of other medicinally valuable plants, contributing to advances in agriculture, biotechnology, and human health.

2. Literature review

2.1 Research on the Plant Genes

The field of plant genomics has undergone a transformative development, thanks to advances in high-throughput sequencing technology and computational biology. Traditional genome sequencing methods, such as cloning-based methods, are labor-intensive and costly, but next-generation sequencing (NGS) technology has significantly accelerated the speed of plant genome analysis [1][2]. These developments have made it

possible to extract and analyze plant DNA on a large scale, allowing researchers to explore the function, structure, and variation of genes at unprecedented resolution.

The most critical tools in modern genome annotation include BLAST (Basic Local Alignment Search Tool) and UniProt. BLAST enables researchers to identify homologous regions between gene sequences in different species, which is crucial for understanding functional and evolutionary relationships [3]. At the same time, the UniProt database provides high-quality protein sequence and functional data and is a key resource for downstream protein analysis and gene function prediction [4].

Recent breakthroughs in the field of protein structure prediction have also revolutionized bioinformatics research. AlphaFold, developed by DeepMind, uses a neural network trained on a protein database to predict the three-dimensional structure of proteins with high accuracy based on their amino acid sequences [5]. This innovation is particularly important in plant genomics, where experimental structural data are often limited. Homology modeling is another important technique that uses known structures to infer unknown macromolecules, supporting drug discovery and crop trait optimization [6].

Despite these advances, research specifically focusing on mustard plants remains scarce. While genomic studies have been applied to major crops such as rice and wheat, genes that determine medicinal and agricultural traits in mustard have not been fully mapped or functionally characterized. Therefore, combining tools such as BLAST, UniProt, and AlphaFold with genome sequencing platforms such as Gensas may be able to fill this knowledge gap. This study aims to reveal the genetic basis of valuable traits in mustard and promote the wider application of plant biotechnology by integrating these technologies.

2.2 Plant Genomics and Research Tools

DNA is found in the nucleus of all cells in all organisms. It is a macromolecule made up of long chains of subunits called nucleotides. It stores a lot of information and controls all chemical changes that occur in cells, such as the formation of cells, muscles, blood, nerves, etc. Also controls the types of living things. Every organism starts with a single cell. A single cell contains either one or two copies of its DNA, depending on the cell, and then the information stored in that cell is enough for the organism to grow, making the cell differentiate. Different types of cells, tissues, and DNA give an organism its full complexity to form the final organism. The significance of our study of DNA is to study what all the components of the mustard plant are and how the mustard plant is different from other organisms, and then to study and edit genes based on DNA, because genes are triplets in DNA. Sequence that can be encoded into a complete protein. In other words, the DNA base sequence is the genetic code. A triplet is a set of three bases. It controls the production of specific amino acids in the cell cytoplasm. Different amino acids and their combinations The sequence determines the type of protein produced. People also read these three bases when people read DNA later. However, it should be noted that not all genes code for DNA, and some only code for RNA[7].

2.3 Genomic and Genetic Concepts

First, using gensas, a web-based genome structure and function annotation and curation platform, a pipeline can be provided for the entire genome structure and function annotation of Bractaceae plants. You can upload your genome sequence and choose from a variety of tools for repeat masking, predictive gene models and other structural features, and functional annotation tools to provide visualization and editing of genes.

With the accumulation of experimental data sets and the rapid emergence of novel omics data, gene sets have become more diverse, which is critical for the refinement of gene annotation at the multidimensional level. From the gensas gene, the person can see the final gene set, gene consensus, gene consensus and predictions, and protein alignments of mustard plants.A genome is the complete sequence of DNA, genes, or genetic material present in a cell or organism and contains all the information needed to build and maintain the organism. Another concept that needs to be distinguished is that genomics is a scientific discipline that studies the composition and function of the genome, involving genetics, molecular biology and bioinformatics methods. The goal of studying the genome is to determine the sequence of the chemical base pairs that make up human DNA, and to identify and map all genes in the human genome from a physical and functional perspective.

2.4 Gensas Platform for Genomic Analysis

The first genome sequencing method is clone-based sequencing, which requires a "tiled path" of clones across the genome, reducing the assembly problem to small fragments and "completed" genome sequences, which has the disadvantage of being expensive. Construction of cloning library, restriction endonuclease

digestion or random shearing can be ligated into cloning vectors (BAC, YAC, fosmid), the cloning vector is then propagated in a host (e.g., E. coli), each host colony containing a unique insert, and each library representing multiple genomic equivalents (redundant). However, some clones cannot reproduce in the host, and people need to pay attention to their toxicity to the host. Clones are anchored to a genetic map, and identification of genetic markers in the clones (e.g., by hybridization) fingerprints the clones and clusters them into groups based on overlapping fingerprints. This allows clone-by-clone sequencing of a minimal tiling path, reducing the number of redundant sequences.

3. Genome Sequencing Technologies

The second whole-genome shotgun sequencing does not require additional molecules or genetic resources, the complexity of sequence assembly is higher, and the draft genome sequence has the advantage that it is not that expensive. It can be applied to the entire genome and to selected clones, where several DNA fragments of similar size are sheared into small fragments (0.5 – 2 kb), and each fragment is sequenced starting from one end. Then there is “double-barreled” shotgun sequencing, where the DNA is sheared into larger fragments (2 kb – 20 kb), and each fragment is sequenced from both ends (“matepairs”).

3.1 Clone-Based Sequencing vs. Shotgun Sequencing

Sequencing is followed by sequence assembly, where the complete sequence ("read") of the DNA molecule is reconstructed from short sequence fragments, overlap layout is expanded, all-to-all sequence similarity is compared, and sequences with high sequence identity are aligned. It is not only based on the assembly of graphs, the construction of k-mer graphs, but also the extraction of linear paths from the graphs. Graph-based sequence assembly, with short read lengths and large data volumes, makes comprehensive comparisons impractical because of many mutually inconsistent overlaps. The biggest challenge in genome assembly is that duplications can be resolved with longer reads. If the reads are larger than the duplicates, it won't cause a problem. Many duplications are a few to tens of kb long...duplications can also be avoided by using pairing. If the pairing is larger than the duplication, people can connect the two sides of the duplication to each other. It allows us to locate repeating types/families within the scaffold if one end of the pair is unique.

3.2 Sequence Assembly and Repetitive Sequence Handling

Secondly, the blast is used to discover similar regions between biological sequences. This program compares nucleotide or protein sequences from Bractaceae plants to sequence databases and calculates statistical significance. Used to infer functional and evolutionary relationships between Bractaceae plant sequences and to aid in the identification of members of gene families. Repeat sequence identification and masking of repeat sequences containing coding elements (such as reverse transcriptase and polymerase in retrotransposons) leads to many "wrong" gene predictions, such as 16,220 of the 56,797 genes predicted in rice Related to repetition! Therefore, repeated sequences should be masked before gene prediction. This also requires a database of known repeat sequences, and sequence similarity.

4. Gene Function Annotation

After that, use UniProt to search for protein sequences and functional information.Finally, the first way to check the extracted DNA information is to use Solexa sequencing, which not only controls the size but also destroys and separates the DNA. First, prepare the DNA sample of our bractose mustard plant, fragment a DNA, connect the adapter to both ends of the fragment, denature the DNA from double-stranded to single-stranded, and then randomly bind the single-stranded DNA fragment to the flow cell On the inner surface of the channel, double bonds can be formed by circulating molecules to form a polymerase reaction. After that, both ends are in contact with the terminus. At this time, the molecules have two copies. After the continuous amplification is completed, each channel of the flow cell will Generating millions of dense double-stranded DNA clusters; the first sequencing cycle can be initiated by adding all labeled reversible terminators and DNA polymerase to the flow cell.

4.1 Gene Editing Validation and Structural Prediction

People can also use homology-based protein modeling to visualize the results, because this protein structure not only adds context to its function, but also quantitatively predicts atomic coordinates, which can provide

new insights for people when viewing the model. Insights, most importantly, such structures could serve as novel building blocks or enhance protein function and targeted therapies. So, first of all, people can select genome-based targets, then isolate, express, purify, crystallize, and perform structure determination after data collection. Finally,people can obtain the protein structure through PDB Deposition & Release. People will find many available structures. However, this method is not widely used by the public because Obtaining structures experimentally is challenging, time-consuming and expensive. Experimental methods in structural biology currently lag far behind sequencing,PDB has 185,541 entries (as of January 1, 2022).In comparison, the previously mentioned UniProt contains over 190,000,000 unique entries.

4.2 Homology-Based Protein Modeling

The principle of this method is that sequence drives protein folding. Every protein is made up of a series of amino acids bound together that interact locally to form helical and sheet-like structures (alpha helices and beta-pleated sheets) that fold up on larger scales to form A complete three-dimensional protein structure (pleated sheet and Alpha helix), the protein can interact with other proteins, perform functions such as signal transmission and transcribe DNA to form a quaternary structure. Of course, this is why this method is difficult to be widely used. Protein folding landsacpes are complex. People also need to differentiate between viable folds. Many times, there are more sequences than unique structures. People can broadly group proteins based on the spatial arrangement of protein secondary structure elements. Currently people know of approximately 1,400 unique folds, while there are only an estimated 2,700 unique folds in nature, so the distinction Wrinkles can also be a huge amount of work.

4.3 Challenges in Protein Structure Prediction

But people can use the structures people already know to constrain the possible solutions. Proteins are polymers of amino acids, made up of atoms, each atom is a certain distance (bond length) from its neighboring atoms, and each triplet of atoms is passed through a set of angles (bond angles and dihedral (torsion) angles). The bond lengths between pairs of atoms are consistent, and the bond lengths can be thought of as more or less fixed, with some slight vibrations, which leaves fewer parameters that people need to adjust to effect changes in the protein structure. Dihedral angles determine the geometry, and the dihedral angles of the peptide backbone are the source of nearly all interesting variations in protein conformation, with phi and psi (either side of the carbon) being the most important. For amino acids other than glycine and proline, the number of possible phi/psi angles is limited, and people cannot just freely rotate phi and psi as they would introduce conflicts - thus increasing the free energy of folding. The Ramachandran diagram defines the range of angles that phi and psi can occupy for most side chains, with phi and psi approaching -60 for α-helical structures. For beta-sheet structures, phi is negative, and psi is positive (-140 and 130).

4.4 Structural Parameters and Ramachandran Plot

Homology modeling uses existing structural information to infer unknown macromolecules. Align unknown structures (targets) to similar known structures (templates). Typically, this is an evolutionarily relevant protein molecule, and the end goal is to predict the structure with an accuracy comparable to experimental methods. Its basic steps are Align the amino acid sequence, generate the backbone, and then loop and side-chain modeling, so that model optimization can be generated to get the predicted structure people want. For the loop modeling - homology The bane of homology modeling is a challenge - there are very few rules that are important for understanding biological functions, interfaces and interactions. If loops longer than 12 – 14 residues are difficult to accurately predict using current methods, it is possible to computationally sample all loop conformations, but differentiation is difficult. Obtaining the desired assembled 3D structure in the final step requires an in-depth understanding of the residue contact points. For example, identifying neighboring residues is crucial to establishing a 3D fold. Knowledge of the interactions between residues is required, but these can be extracted from evolutionary sequence data. Residues in contact tend to be mutated together. For example, mutation of a large, bulky amino acid may be accompanied by a proximal mutation of a small amino acid to create sterically preserved protein structure and function.

This method can be applied to neural networks. These rules are "learned" using neural networks, an information-processing paradigm inspired by the way biological nervous systems process data. Groups of interconnected nodes (or computers) that process information and can be used to infer functionality from data - i.e. in real life. An example is what makes a protein a protein. These neural networks for homology modeling are trained on PDB structures and sequences from the Protein Data Bank, extracting coevolutionary

information to predict interresidue distances. It uses knowledge of protein parameters to extrapolate residue distances into a 3D structural model.

4.5 Deep Learning in Structural Prediction and AlphaFold Pipeline

The AlphaFold structure prediction pipeline generates prediction models using protein sequences of unknown structures. People first input the amino acid sequence to be folded. If people only input a single sequence, use the monomer model. If people input multiple sequences, use the multimer model. Searching against a genetic database, once this cell is executed, you will see statistics about the multiple sequence alignment (MSA) that AlphaFold will use. In particular, you will see how well each residue is covered by similar sequences in MSA. Once executed, a protein model is generated, and the structure can then be evaluated. Assessing the structure requires the local distance difference test (l-DDT) and approximate local errors. In general, l -DDT is influenced by lower l -DDT scores in disordered regions (loops) and regions lacking coevolutionary data, although the same challenges plague all homology-based modeling. AlphaFold has a wide range of applications and 992316 available structures to make up for the lack of structural data. Macromolecular crystal structures can be solved either by direct solution methods (pending high-resolution data) or by molecular substitution, where solving the structure involves using known related structures to provide an initial atomic model for structural characterization. AlphaFold predictions can be used as structural models to characterize and predict proteins in the absence of high-resolution data because they increase the efficiency and scope of experimental structure determination.

Alphafold is a tool that can be used to predict protein-protein interactions. It uses a variety of structures in a general pipeline, which is evaluated based on several criteria. These criteria include the amount of sequence information obtained, unique aligned sequences, sufficient alignments of all protein regions, analysis of prediction errors, and whether the errors are related to coevolutionary data or disordered regions. It is also essential to consider whether this model makes sense based on our knowledge and to analyze the biophysical properties that control protein folding.

5. Conclusion

This study aimed to investigate whether computational gene annotation and editing tools can effectively identify functional genes in Brassica juncea plants, especially those with medicinal or agronomic significance. By using platforms such as Gensas, BLAST, UniProt, and AlphaFold, the author successfully constructed computational pipelines for DNA sequence extraction, gene annotation, and protein structure prediction. The results support the hypothesis that computer-based modeling can improve the efficiency and accuracy of gene discovery in plant genomes. This approach reduces the reliance on time-consuming experimental methods and provides valuable insights into the molecular basis of traits in Brassica juncea plants.

However, some limitations were also noted. The accuracy of homology-based models is highly dependent on the availability of high-quality reference genomes, which are still limited for some plant species. In addition, some predicted gene functions are still speculative and have not been experimentally verified. To address these limitations, future studies should combine wet-lab experiments with computational predictions, especially in functional gene expression analysis and protein validation. In addition, future studies can focus on extending this approach to other understudied medicinal plants, improving the integration of machine learning models, and improving genome assembly quality through long-read sequencing technology. Ultimately, this research lays the foundation for genome-assisted plant breeding, bioengineering, and the development of plant-derived drugs.

References

[1] Mardis, E. R. (2008). Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics, 9, 387–402. https://doi.org/10.1146/annurev.genom.9.081307.164359

[2] Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics, 17(6), 333–351. https://doi.org/10.1038/nrg.2016.49

[3] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

[4] UniProt Consortium. (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research, 49(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

[5] Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2

[6] Varshney, R. K., Terauchi, R., & McCouch, S. R. (2014). Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. PLoS Biology, 12(6), e1001883. https://doi.org/10.1371/journal.pbio.1001883

[7] Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324