www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277-7695
ISSN (P): 2349-8242
NAAS Rating: 5.23
TPI 2023; SP-12(7): 01-06

www.thepharmajournal.com
Received: 02-03-2023
Accepted: 05-04-2023

Suman Dutta
ICAR-Indian Agricultural Research Institute, New Delhi, India

Rajkumar U Zunjare
ICAR-Indian Agricultural Research Institute, New Delhi, India

Vignesh Muthusamy
ICAR-Indian Agricultural Research Institute, New Delhi, India

Firoz Hossain
ICAR-Indian Agricultural Research Institute, New Delhi, India

Corresponding Author:
Firoz Hossain
ICAR-Indian Agricultural Research Institute, New Delhi, India

Prediction of CENH3 protein in maize using machine learning techniques

Suman Dutta, Rajkumar U Zunjare, Vignesh Muthusamy and Firoz Hossain

DOI: https://doi.org/10.22271/tpi.2023.v12.i7Sa.21185

Abstract
Centromere specific CENH3 gene encoding a variant for histone H3 protein causes *in-vivo* haploid induction in maize. Chromosome duplication caused by colchicine therapy in haploids causes inbreds to become fully fixed after just one generation, as opposed to 6-7 generations of selfing in traditional methods. For *in-vivo* haploid induction, understanding of CENH3 proteins in segregation of chromosomes during cell division is therefore of vital importance. There is currently no online resource that can categorise unknown proteins into CENH3 proteins. In this study, our goal was to build a machine learning-based system for predicting the CENH3 protein of unidentified origin. Amino acid composition (AAC) was employed to construct random forest, decision tree and logistic regression classifiers to predict the CENH3 proteins. A total of 618 protein sequences were examined, including 309 CENH3 sequences from different species and 309 Non-CENH3 sequences from *Zea mays*. The prediction of CENH3 proteins showed considerable promise using random forest and logistic regression classifiers. AAC achieved >98% prediction accuracies using random forest and logistic regression classifiers. Also, t-SNE technique could successfully separate two different classes of proteins in two-dimensional space. The average accuracy scores from the cross-validation of the logistic regression and random forest models were promising while 10 folds of cross-validation using the k-fold method was performed. Hence, the cross-validation score also showed that each model had a promising ability to predict CENH3 proteins. The findings of the study can be applied to different crops before any experiments are conducted.

Keywords: Haploid induction, CENH3, random forest, decision tree, logistic regression, and prediction

1. Introduction
Induction of haploid become a method of choice in many crop breeding programmes due to its logistic and economic viability (Gain *et al*. 2022). Haploid (n) stage in higher plants is generally considered a transition phase in the form of gametes produced from sporophytic diploid (2n) plants (Mahlandt *et al*. 2023) [17]. However, haploids may arise from several intergeneric, interspecific, and a few intraspecific crosses due to complete uniparental genome elimination (Watts *et al*. 2020) [27]. The first report of sporophytic haploid plants was published on the progeny of *Nicotiana tabacum* X *N. sylvestris* crosses by Clausen and Mann (1924) [3]. Haploids are of great practical significance in many plant breeding programmes as doubling of chromosome number yields a completely homozygous doubled haploid (DH) plant within two seasons as compared to 6 to 7 generations in conventional true breeding methods (Dutta *et al*. 2022) [6]. The haploids can be created through *in-vitro* culture of microspores, anthers or ovules in several plant species (Dunwell 2010) [5]; however, it could not bring economic and logistic solutions to produce complete homozygous plants due to technical reasons. Therefore, haploid inducer (HI) based *in-vivo* methods could be a viable alternative for large scale production of DH required for the plant breeding programme. In maize, maternal and paternal haploids can be generated using the naturally existing mutation from the induction crosses of Stock 6 and Wisconsin 23 derived mutants; respectively (Coe 1959; Kelliher *et al*. 2017) [4, 12]. High frequency barley haploids were generated from the interspecific cross *Hordeum vulgare* and *H. bulbosum* (Kasha and Kao 1970) [11]. Subsequently, several intergeneric crosses were documented in oat x maize (Marcinska *et al*., 2013) [18] and wheat x maize (Laurie and Bennett, 1988) crosses to yield oat and wheat haploids due to uniparental loss of maize chromosomes. Another breakthrough came when Ravi and Chan (2010) [22] were able to

generate haploids in crosses between wild-type *Arabidopsis* plants and transgenic expressing engineered centromeric histone H3 (CENH3) protein. Since then, new opportunities opened for the development of HI lines through genetic engineering and allow production of haploids progeny without the need of *in-vitro* technique.

CENH3 based haploidization has other potential applications in many crop species including polyploid as it offers as a tool to accelerate the mutagenesis screen for minimising the ploidy label. CENH3 is a variant of histone H3 protein present in centromeric nucleosomes and consists of two major domains. N terminal tail domain of CENH3 shares little similarity with conventional histone H3 whereas the C terminal histone fold domain (HFD) shows significant similarity with conventional histones (Watts *et al*. 2020) [27]. Mutation in *CENH3* is lethal at the homozygous stage as chromosomes fail to segregate to poles at the time of cell division due to the loss of functional centromere. However, heterozygous *CENH3* mutants are viable in both animals and plants (Ravi *et al*. 2014) [23]. *Bulbosum* technique in barley was governed by incompatible CENH3 spindle-fiber interactions between two species *H. vulgare* × *H. bulbosum* which leads to uniparental genome elimination (Sanei *et al*., 2011) [24]. An attempt was also made in *B. juncea* to produce haploids using CENH3 mediated genome elimination (Watts *et al*. 2020) [27]. Therefore, CENH3-based haploid induction can be targeted in other plant species to harness the benefit of the DH technology for the production of large-scale homozygous lines (Dutta *et al*. 2022) [6]. Before starting any DH production facility, it needs a proper understanding of the responsive factors. The primary success of DH production mainly relies on a suitable haploid inducer system and can be exploited through CENH3-based genetic engineering in other crop species. However, there is limited information available in the computational facility that can predict the protein sequences with CENH3 properties. Therefore, the present study targeted for building a computational model that can allow us to predict the putative CENH3 protein in plant species.

Machine learning based approaches are now commonly used in many circumstances for the prediction of a particular protein of interest (Meher *et al*. 2019) [20]. Computational-based approaches are not restricted only to biological problems as it gets momentum in image processing, natural language processing and evolution (Meher *et al*. 2016) [19]. Machine learning and deep learning-based binary predictors were developed in the last two decades for the classification of a target protein of interest against the other proteins in a genome (Nielsen *et al*. 1999) [21]. There are several biological domains where machine learning techniques can be implemented in several biological domains including genomics, proteomics, microarrays, systems biology, evolution and text mining of biological sequences using natural language processing (NLP) (Larranaga *et al*. 2006) [15]. Therefore, the present context was relevant to develop a machine learning model for the prediction of CENH3 protein from the other unknown sequence. The developed model can be useful to gain prior knowledge before initiating any laboratory experiment.

2. Materials and Methods

2.1 Collection of datasets

For classification of protein, both datasets belonging to CENH3 and Non-CENH3 proteins were retrieved from Uniprot database (http://www.uniprot.org/). The dataset with CENH3 proteins were used as positive dataset, whereas, Non-CENH3 proteins were used as negative dataset for binary classification. Positive dataset was constructed with CENH3 proteins from all the plant species available at Uniprot database. For negative dataset, randomised protein sequences of Non-CENH3 type were considered for maize to avoid the biased model building. Therefore, a total number of 309 protein sequences of each positive and negative class of protein sequences (Total 618 protein sequences) were used for binary classification.

2.2 Feature generation

Feature generation from protein sequence is a critical step while building machine learning model. Numeric feature vectors were created from the strings of amino acids of each protein sequences. Here, sequence-based features were generated from each protein sequences to build the classification model. Feature include amino acid compositions (AAC) for classification of the proteins. AAC is the simplest and most widely used structural feature for representing a protein sequence (Bhasin and Raghava, 2004) [2]. It is the proportions of amino acid residues present in a protein sequence. For a protein sequence with N residues, AAC for the i^{th} amino acid can be computed as AAC (i) = fi/N, where i = 1 to 20. Therefore, every protein sequence can be transformed into a vector of 20 numeric observations. Distribution of the amino acids in the CENH3 and Non-CENH3 proteins were visualized using violin plot.

2.3 Model building for protein structure prediction

For the prediction of CENH3 proteins, a sample size of 247 protein sequences was employed, and the remaining 62 peptide sequences were included in the testing data set. Using various classifiers such as random forest, logistic regression and decision tree with default parameters, prediction accuracy for the protein was calculated. Balanced data set was used for classification of model to avoid the bias at the time training of the model.

2.4 Model performance evaluation using confusion matrix

Model performance was evaluated through analysis of the confusion matrix where actual and predicted DMP and non-DMP proteins were presented as true positive (TP), false positive (FP), false negative (FN) and true negative (TN) categories. On the basis of actual and predicted observations, several scores were calculated (precision, recall, accuracy score, F1-score, matthew's correlation coefficient (MCC)) to evaluate the performance of the predicted model.

2.5 Cross validation of the model

We used k-fold cross validation to execute 10 folds of cross validation on each model to analyse its performance. Mean accuracy scores of random forest, decision tree, and logistic regression classifiers were calculated to assess the performance of the model. Accuracy scores obtained for 10-fold cross validation was visualized using the violin plot.

2.6 Visualisation of dataset in two-dimensional space using t-SNE

Data visualization has been considered as a powerful tool due to its efficiency in abstracting out the right information clearly and easily. For this purpose, t-distributed Stochastic Neighbour Embedding (t-SNE), an unsupervised, randomized algorithm, used for visualization. t-SNE introduced by van der

Maaten and Hinton (2008) [26] is a popular method for exploring high-dimensional data. The technique has become popular in the field of machine learning as it has the ability to embed hundreds or even thousands of dimensional data into two-dimensional space. t-SNE is the improvement of Stochastic Neighbour Embedding (SNE) introduced by van der Maaten and Hinton (2008) [26] that reduces Kullback-Leibler (KL) divergence of scaled similarities of the points i and j in high dimensional (p_{ij}) and low dimensional (q_{ij}) space in such a way that $KL(\mathrm{P}||Q)\ equals \sum p_{ij} log \frac{p_{ij}}{q_{ij}}$, where $i \neq j$.

The KL divergence of the joint probabilities between original and embedded space will be minimized using gradient descent. t-SNE converts affinities of data points to probabilities, where affinities in embedded space are denoted by student's t-distributions and affinities in original space are denoted by gaussian joint probabilities. t-SNE especially suitable to represent high dimensional and complex data in low dimensions due to heavy-tailed t-distribution as it preserves local neighbourhoods of the data efficiently and penalizes wrong embeddings of dissimilar points. t-SNE allows to group samples based on their local structure that might be useful to disentangle dataset visually. The low dimensional so generated is further used for the selected parameter for evaluation the performance. Two important hyperparameters namely perplexity and iteration were used for dimension reduction purpose. A perplexity value of 50 and an iteration value of 500 were used in combination for visualisation of the data.

2.7 Statistical software used for analysis

For data curation and labelling of the sample, Microsoft Excel Version 2019 was used. For feature generation using "protr" package of R programming language. All the statistical analysis was carried out in the Anaconda Jupiter Notebook integrated development environment (Python Version 3.7). Machine learning analysis was performed using numpy (1.13.1), sklearn (0.19.1), matplotlib (2.1.0), and pandas (0.20.1).

3. Results

3.1 Amino acid distribution of two different classes of proteins

ACC of the CENH3 and Non-CENH3 proteins were calculated using all the protein sequences under study. The symbolic code of each amino acid was presented using single letter code- G: Glycine, A: Alanine, L: Leucine, M: Methionine, F: Phenylalanine, W: Tryptophan, K: Lysine, Q: Glutamine, E: Glutamic acid, S: Serine, P: Proline, V: Valine, I: Isoleucine, C: Cysteine, Y: Tyrosine, H: Histidine, R: Arginine, N: Asparagine, D: Aspartic acid, T: Threonine (Figure 1). In case of CENH3 proteins, the highest proportion of amino acid was recorded for R (Arginine) and A (Alanine) with the mean value of 12.06 ± 2.59 % and 11.58 ± 3.98 %, respectively. In contrast, C (Cysteine), W (Tryptophan), and Y (Tyrosine) showed the lowest proportion of amino acid in CENH3 proteins with the mean value of 1.16 ± 1.13 %, 1.26 ± 0.89 %, and 1.3 ± 0.95 %, respectively (Figure 1A). When it comes to Non-CENH3 proteins, A (Alanine), G (Glycine), and L (Leucine) had the highest percentages of amino acids, with mean values of 9.77 ± 3.01% and 8.63 ± 2.97%, and 8.87 ± 2.17 %, respectively. The lowest percentage of amino acids in Non-CENH3 proteins was found in the proteins W (Tryptophan) and C (Cysteine) with mean values of 1.22 ± 0.85 % and 1.76 ± 1.37 %, respectively (Figure 1B). The distribution of other amino acids in the two different classes of proteins were visualized through box plot analysis in Figure 1.

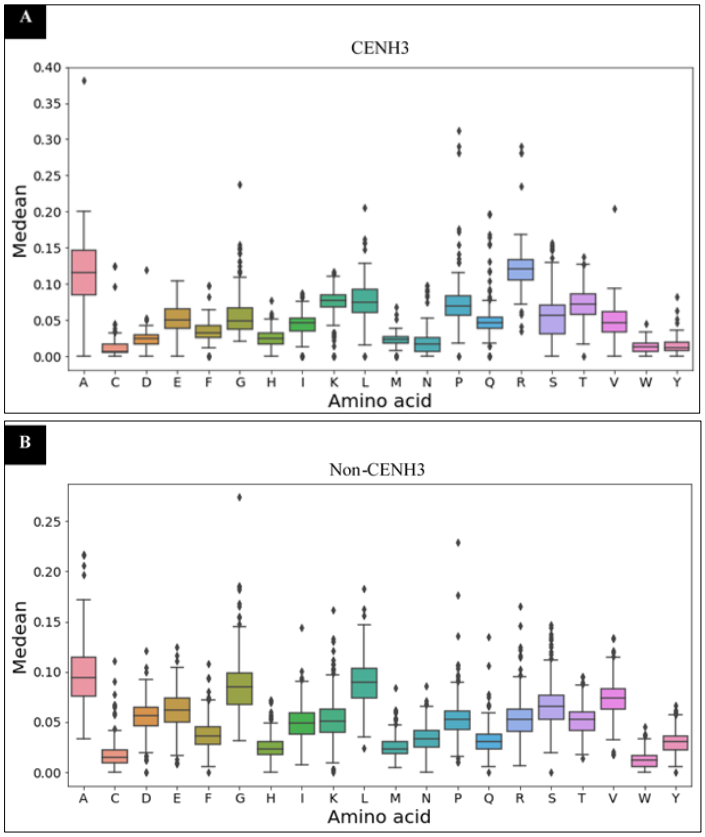

Fig 1: Distribution of twenty amino acids in the CENH3 (A) and Non-CENH3 (B) proteins.

3.2 Performance of different prediction models

A sample sizes of 247 protein sequences were used for prediction of CENH3 proteins and remaining 62 peptides sequences were included in the testing data set. Prediction accuracies for the protein in the selected crop plants were given using different models with default parameters (Table 1). The accuracy score using random forest, decision tree, and logistic regression were 98.40 %, 97.60%, and 98.40%, respectively. Precision scores were 96.90%, 96.80% and 96.90% respectively using random forest, decision tree, and logistic regression. The recorded recall values using random forest, decision tree, and logistic regression were 1.00, 0.98, and 1.00, respectively. F1-scores for newly developed models were 0.984, 0.976, and 0.984 for random forest, decision tree, and logistic regression, respectively. Brier score was low with the recorded value of 0.016, 0.024, and 0.016 for random forest, decision tree and logistic regression, respectively. Therefore, the performance of random forest and logistic regression was slightly better than the decision tree model in terms of all the score matrices (Table 1).

Table 1: Performance of the model using different classifiers

Score	Random forest	Decision tree	Logistic regression
Accuracy	0.984	0.976	0.984
Precision	0.969	0.968	0.969
Recall	1.000	0.984	1.000
F1-score	0.984	0.976	0.984
MCC	0.968	0.952	0.968
Brier score	0.016	0.024	0.016

3.3 Cross-validation of the model performance

To analyse the performance of each model, we performed 10 folds cross validation using k-fold cross validation. The model performance was promising using random forest classifier with mean accuracy score of 98.05 ± 2.40 %. Similarly, logistic regression shows a promising cross validation score with mean value of 97.57 ± 2.20 %. Cross validation score using the decision tree was 96.43 ± 3.55 % which is slightly lower than the other two models used under study. Therefore, cross-validation score also revealed the promising performance of each model to predict the CENH3 proteins. The results of the cross-validation score were visualized through a violin plot in Figure 2.

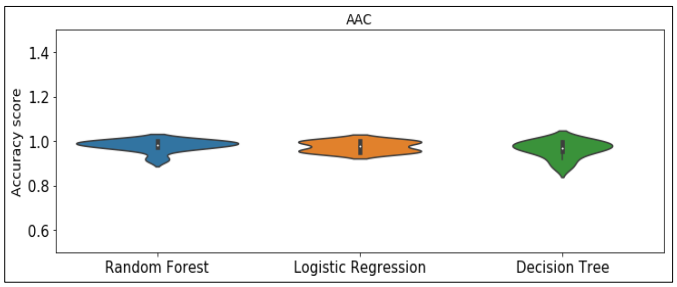

Fig 2: Visualisation of cross validation of different models using violin plot.

3.4 Two-dimensional visualization of two different classes of proteins using t-SNE

Due to its effectiveness in clearly and quickly abstracting out the appropriate information, data visualisation has been regarded as a valuable tool. One of the key elements supporting the field of data analysis is the descriptive visualization of the full dataset. Nevertheless, because our data visualization is normally limited to two dimensions, dealing with datasets with multi dimension start generating complications. An unsupervised, randomized approach called t-distributed Stochastic Neighbour Embedding (t-SNE) was employed for visualization in this process. Perplexity and iteration, two crucial hyperparameters, were applied for dimension reduction. For the data visualization, a perplexity value of 50 and an iteration value of 500 were combined to visualize the CENH3 and Non-CENH3 properly (Figure 3). With few exceptions, t-SNE method could effectively separate two different kinds of data sets.

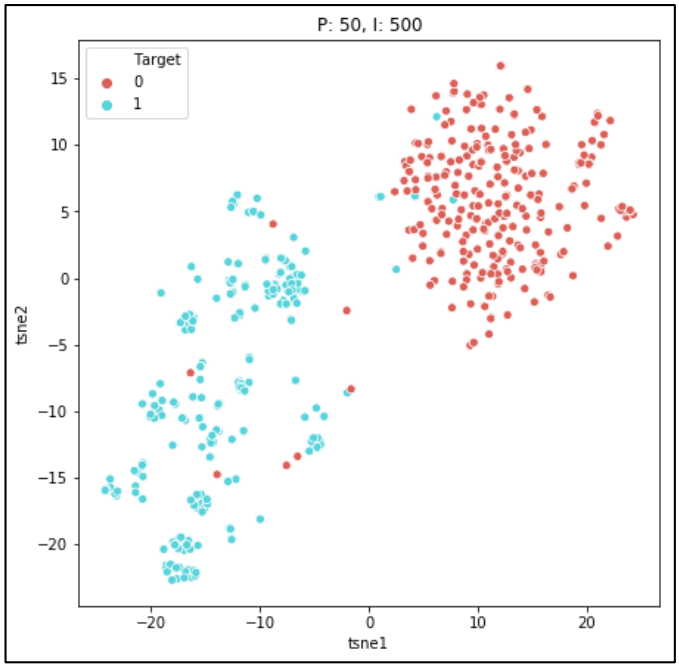

Fig 3: Visualisation of twenty amino acids in two-dimensional space, 1: CENH3; 0: Non-CENH3.

4. Discussion

The possible uses for CENH3-based haploidization include a method to speed up mutagenesis screening for reducing the ploidy label. The centromeric nucleosome contains CENH3, a form of the histone H3 protein that has two primary domains. Whereas the C terminal histone fold domain (HFD) of CENH3 exhibits significant similarities with ordinary histones, the N terminal tail region of CENH3 displays little similarity with regular histone H3 (Watts *et al*. 2020) [27]. When CENH3 is mutated, it is fatal because the absence of a functioning centromere prevents chromosomes from segregating to the poles during cell division. Nonetheless, heterozygous CENH3 mutants live in both plants and animals (Ravi *et al*. 2014) [23]. Incompatible CENH3 spindle-fibre interactions between the two species *H. vulgare* and *H. bulbosum*, which result in the removal of the uniparental genome, rule the *bulbosum* technique in barley (Kasha and Kao 1970; Sanei *et al*., 2011) [11, 24]. With this understanding in mind, Kelliher *et al*. (2016) [13] attempted to use an RNAi construct to downregulate the maize CENH3 gene in order to induce haploidy in the plant. HIR, however, fell short of the desired range needed for DH generation. Furthermore, haploid production using CENH3-mediated genome deletion was attempted in *B. juncea* (Watts *et al*. 2020) [27]. The differentiation of haploid from diploid cells can be done using engineered green fluorescent protein (Ravi and Chan 2010; Kelliher *et al*. 2016) [22, 13].

The composition of amino acids or the frequency of amino

acids in proteins is well conserved between species (Gilis *et al*. 2001; Itzkovit and Alon 2007) [9, 10]. Compositional changes have been connected to integral membrane proteins, cellular architecture, gene expression, and the rise in protein stability in response to environmental challenges including sulphur deprivation and high ambient temperatures (Zeldovich *et al*. 2007; Sterner and Liebel 2001; Friedman *et al*. 2004) [28, 25, 7]. The distribution of amino acids within a protein is not solely governed by its functional needs. Hence, it is unlikely that a historical incident led to a specialized makeup of amino acids found in natural proteins. The natural composition is expected to lower the metabolic cost of producing amino acids in some animals (Akashi *et al*. 2002) [1]. An amino acid frequency and the number of codons that correspond to it are tightly correlated, which raises the possibility that the composition is a product of the genetic code (King and Jukes 1969) [14] Yet, even in the presence of a stable genetic code, modifications to the underlying genome sequence can have an impact on the ratio of amino acids. In the current study, peptide sequences were mapped onto numeric feature vectors using the AAC dataset. These feature vectors were then used as input in the different models to predict CENH3 proteins. Also, it would be useful to understand how the AAC of CENH3 proteins function in relation to haploid induction and chromosome segregation is related to their compositional property. AAC was found to be predictive in this analysis when applied to three different models. In addition, t-SNE approach was capable of effectively separating two distinct data sets. We used k-fold cross-validation to execute 10 folds of cross-validation on each model to analyse its performance. Mean accuracy scores obtained from cross-validation of random forest and logistic regression were encouraging. Hence, the cross-validation score also demonstrated promising ability of each model to predict CENH3 proteins.

The inability to conduct in-depth investigations of several such proteins in crops, particularly those implicated in *in-vivo* haploid induction, is significantly hampered by the lack of an online tool currently available to detect proteins with CENH3 activity. Here, we also present the first computational machine learning model for differentiating the two protein groups (CENH3 and Non-CENH3) in maize. The established methodology is anticipated to be a supplement to transcriptional profiling and comparative genomics studies for the identification and functional annotation of genes important for *in-vivo* maternal haploid induction. The model will aid in the identification of CENH3 and Non-CENH3 proteins as well as the functional annotation of CENH3 genes found in the genomes of maize. For the vast majority of experimental scientists working on *in-vivo* haploid induction research, the established model is important since it not only shows the direction in which future computational approaches will be created. This is the first instance of machine learning being used to find proteins in plants that act like CENH3. A web-based server portal for the discovery of unidentified proteins with CENH3-like activity can be made using the developed model. Using the random forest and logistic regression models with predefined parameters, researchers can easily discover CENH3 proteins throughout the proteome without diving into the nuances of the statistical methods used to build the strategy.

5. Acknowledgements

The first author is thankful to Human Resource Development Group (HRDG) division of Council of Scientific & Industrial Research (CSIR), New Delhi, India for Junior Research Fellowship to pursue the Ph.D. programme.

6. Contribution of the authors

Conduct of the experiment: SD, Collection, curation and arrangement of data: SD and VM, Programming: SD, Statistical analyses: RUZ and VM, Manuscript writing: SD and FH, Design of experiment: FH

7. Funding

We are thankful to the ICAR-IARI, New Delhi for financial support.

8. Conflicts of interest Authors declare that no conflict of interest exists.

9. References

1. Akashi H, Gojobori T. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proceedings of the National Academy of Sciences. 2002;99:3695-3700.
2. Bhasin M, Raghava GP. Classification of nuclear receptors based on amino acid composition and dipeptide composition. Journal of Biological Chemistry. 2004;279:23262-23266.
3. Clausen RE, Mann MC. Inheritance in *Nicotiana tabacum*: V. The occurrence of haploid plants in interspecific progenies. Proceedings of the National Academy of Sciences. 1924;10:121-124.
4. Coe Jr EH. A line of maize with high haploid frequency. The American Naturalist. 1959;93:381-382.
5. Dunwell JM. Haploids in flowering plants: origins and exploitation. Plant Biotechnol Journal. 2010;8:377-424.
6. Dutta S, Muthusamy V, Zunjare RU, Hossain F. Accelerated generation of elite inbreds in maize using doubled haploid technology. Plant Breeding-New Perspectives. Intechopen, 2022.
7. Friedman R, Drake JW, Hughes AL. Genome-wide patterns of nucleotide substitution reveal stringent functional constraints on the protein sequences of thermophiles. Genetics. 2004;167:1507–1512.
8. Gain N, Chhabra R, Chandra S, Zunjare RU, Dutta S, Chand G, *et al*. Variation in anthocyanin pigmentation by *R1-navajo* gene, development and validation of breeder-friendly markers specific to *C1-inhibitor* locus for *in-vivo* haploid production in maize. Molecular Biology Reports. 2023;50:2221-2229.
9. Gilis D, Massar S, Cerf NJ, Rooman M. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. Genome Biology. 2001;2:1–12.
10. Itzkovit S, Alon U. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Research. 2007;17:405–412.
11. Kasha KJ, Kao KN. High frequency haploid production in barley (*Hordeum vulgare* L.). Nature. 1970;225:874-876.
12. Kelliher T, Starr D, Richbourg L, Chintamanani S, Delzer B, Nuccio ML, *et al*. Matrilineal, a sperm-specific phospholipase, triggers maize haploid induction. Nature. 2017;542:105-109.
13. Kelliher T, Starr D, Wang W, McCuiston J, Zhong H, Nuccio ML, *et al*. Maternal haploids are preferentially induced by *CENH3*-tailswap transgenic complementation

in maize. Frontiers in plant science. 2016;7:414.
14. King JL, Jukes TH. Non-darwinian evolution. Science. 1969;164:788–798.
15. Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, *et al*. Machine learning in bioinformatics. Briefings in bioinformatics. 2006;7:86-112.
16. Laurie DA, Bennett MD. The production of haploid wheat plants from wheat x maize crosses. Theoretical and Applied Genetics. 1988;76:393–397.
17. Mahlandt A, Singh DK, Mercier R. Engineering apomixis in crops. Theoretical and Applied Genetics. 2023;136:131.
18. Marcinska I, Nowakowska A, Skrzypek E, Czyczyło-Mysza I. Production of double haploids in oat (*Avena sativa* L.) by pollination with maize (*Zea mays* L.). Central European Journal of Biology. 2013;8:306-313.
19. Meher PK, Sahu TK, Rao AR. Prediction of donor splice sites using random forest with a new sequence encoding approach. BioData mining. 2016;9:1-25.
20. Meher PK, Sahu TK, Mohanty J, Gahoi S, Purru S, Grover M, *et al*. nifPred: Proteome-wide identification and categorization of nitrogen-fixation proteins of diaztrophs based on composition-transition-distribution features using support vector machine. Frontier in Microbiology. 2019;9:1100.
21. Nielsen H, Brunak S, von Heijne G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein engineering. 1999;12:3-9.
22. Ravi M, Chan SW. Haploid plants produced by centromere-mediated genome elimination. Nature. 2010;464:615-618.
23. Ravi M, Marimuthu MPA, Tan EH, Maheshwari S, Henry IM, Marin-Rodriguez B, *et al*. A haploid genetics toolbox for Arabidopsis thaliana. Nature communications. 2014;5:1-8.
24. Sanei M, Pickering R, Kumke K, Nasuda S, Houben A. Loss of centromeric *histone H3* (*CENH3*) from centromeres precedes uniparental chromosome elimination in interspecific barley hybrids. Proceedings of the National Academy of Sciences. 2011;108:E498-E505.
25. Sterner RH, Liebl W. Thermophilic adaptation of proteins. Critical Reviews in Biochemistry and Molecular Biology. 2001;36:39-106.
26. Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9:2579-2605.
27. Watts A, Sankaranarayanan S, Raipuria RK, Watts A. Production and application of doubled haploid in Brassica improvement. Brassica Improvement: Molecular, Genetics and Genomic Perspectives, 2020, pp.67-84.
28. Zeldovich KB, Berezovsky IN, Shakhnovich EI. Protein and DNA Sequence Determinants of Thermophilic Adaptation. PLoS Computational Biology. 2007;3:e5.