Anupam Banerjee¹Priti Das²Orachorn Mekkerdchoo^{2,5}Chaleeda Borompichaichartkul³Subhadip Basu¹ Mita Nasipuri¹Georges Srzednicki⁴

Research Article

A soft computing tool for species classification and prediction of glucomannan content in *Amorphophallus* genus

The proposed work aims at designing a classification system for automatic identification of *A. muelleri* species, grown as a potential cash crop in many Asian countries, from the DNA fingerprints of *Amorphophallus* genus. Four sets of 48 DNA fingerprints belonging to 37 species of the *Amorphophallus* genus, developed with the help of four different primers are considered for the experiment, with an objective to identify only the fingerprints of the species of interest. A second experimental setup deals with the automatic classification of species containing high amounts of glucomannan from the same set of DNA fingerprints of the *Amorphophallus* genus. For each set of 48 DNA fingerprints generated with a specific primer, the DNA fingerprints are preprocessed to extract a 42 dimensional feature vector which is used to generate a k-Nearest Neighbor based classifier based on the Leave One Out Cross Validation protocol. Final classification based on outputs from individual classifiers constructed with respect to the four different primers is performed according to a n-star consensus strategy. The n-star consensus predicts species *A. muelleri* with cent per cent accuracy while it predicts species containing glucomannan with a more modest accuracy of 81.25%.

Keywords: *Amorphophallus* genus / Glucomannan / k-NN classification / Species classification

Additional supporting information may be found in the online version of this article at the publisher's web-site

Received: February 27, 2017; revised: May 12, 2017; accepted: August 11, 2017

DOI: 10.1002/elsc.201700040

¹Department of Computer Science & Engineering, Jadavpur University, India

²Program in Biotechnology, Faculty of Science, Chulalongkorn University, Bangkok, Thailand

³Department of Food Technology, Faculty of Science, Chulalongkorn University, Bangkok, Thailand

⁴School of Chemical Engineering, The University of New South Wales, Sydney, Australia

⁵Program in Industrial Fermentation Technology, Faculty of Agro-Industry, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand

1 Introduction

Species of the genus *Amorphophallus* are among the most striking and enigmatic of all plants. Famed for including some of the largest flower structures of the entire world's flora, many members of this beautiful and diverse genus remain poorly understood and seldom cultivated, despite their spectacular blooms. The genus was named after the Greek for 'misshapen (or deformed) penis', in reference to the shape of the prominent, saggy, spire-like spadix of the first described species [1].

The genus *Amorphophallus* comprises over 170 species occurring from West Africa, through Asia into the tropical portions of the Western Pacific and North Eastern Australia. Their morphological variation is unmatched by any of the other genera in same family and this extent of variation is found in the spathe, the appendix, odour, tuber, petiole, the individual female flower and

the shape of the inflorescences. Therefore, for the taxonomist, *Amorphophallus* is a true treasure of variation [2]. Several studies were carried out for identification of species belonging to this genus including morphology [2–5], palynology, [6], [7] and biochemistry related to its inflorescence odour [8], [9]. However, the morphology and palynology characters are highly variable and are therefore difficult to identify in the vegetative form without the help of an experienced professional taxonomist [7–10]. Nevertheless, if the diagnostic morphological character of a sample is lacking, it becomes difficult even for the specialists to recognize a species correctly [11]. Moreover, the morphological features are influenced by the environmental conditions, tissue specificity and development stage [12, 13]. Most of the conventional methods employed for species identification are laborious and time consuming, requiring large population sizes and space for testing [14, 15]. The DNA fingerprinting procedure overcomes such shortcomings and has been employed successfully for species identification. DNA fingerprints can detect DNA polymorphism between species [16, 17]. A plethora of DNA markers can be used to generate DNA fingerprints lending greater polymorphism than morphological or biochem-

Correspondence: A/Prof. Subhadip Basu (subhadip@cse.jdvu.ac.in), Department of Computer Science and Engineering, Jadavpur University, Kolkata-700032, India

Abbreviation: k-NN, k-Nearest Neighbor

ical markers. The DNA fingerprinting procedure is not affected by environmental conditions and can be carried out at any stage of the life cycle of plant and from almost any tissue [12, 13, 18].

Randomly Amplified Polymorphic DNA (RAPD) tests are used to generate DNA fingerprints of a particular plant. This technique can be used for a wide range of applications because of its sensitivity, simplicity, cost-effectiveness and because it does not require sequencing reactions [19–21]. Moreover, one primer allows the amplification of several bands which correspond to several DNA locations. The molecular technique is a simple method and is widely used to develop a specific DNA marker with universal RAPD primers. However, there are some drawbacks in the conventional techniques used for interpretation of results. The primary shortcoming is that it takes time for results to be confirmed and the identification outcome may still vary between laboratories. In order to address this shortcoming, automated prediction based on image analytics can be deployed to speed up the identification process and improve the accuracy of the results obtained. An image analytics based species classification system built with the help of a database of DNA fingerprints generated with RAPD primers can reduce manual intervention to a considerable extent. The identification process will necessarily require only the determination of the DNA pattern in unknown samples in order to compare it with existing patterns stored in the database of DNA fingerprints. The species classification system would allow a rapid identification of economically important species without the need of chemical analysis at initial stages. Such a system will have higher accuracy, consistency and will be faster in comparison to predictions performed solely on the basis of human observations.

In addition to species classification, the same prediction system can be extended to identify species (or samples corresponding to specific species) containing specific chemical components of interest. Corresponding to the *Amorphophallus* genus, glucomannan is one such chemical component of great economic importance. Glucomannan is a water-soluble polysaccharide that is considered a dietary fiber. It is a hemicellulose component present in the cell walls of some plant species. Glucomannan is a food additive that is used as an emulsifier and thickener. It is also sold as nutritional supplements for constipation, obesity, high cholesterol, acne vulgaris and type 2 diabetes [22, 23]. Glucomannan comprises 40% by dry weight of the rhizomes of some species of the *Amorphophallus* genus. The identification of the species rich in glucomannan can be of great commercial importance. Considering the fact that there is great variation in the time period required for inflorescence in certain species and that the intra species morphological differences can be often misleading, analysis of DNA fingerprints of the species of the *Amorphophallus* genus can facilitate their identification to a great extent. There lies the inspiration for the proposed work, the successful application of which can drastically eliminate the waiting period for the cultivators and can serve the farming and food processing industry to a considerable extent.

The proposed work deals with species of the *Amorphophallus* genus [1], and is intended as a case study to validate the hypothesis that species can be classified on the basis of their DNA fingerprints. The first experimental setup [24] deals with the identification of the *Amorphophallus muelleri* species on the basis of DNA fingerprints among thirty seven species of the *Amorphophallus*

genus. Four sets of 48 DNA fingerprints with dimension 42*48, developed with the help of four primers, namely- AB-04 [25], OPD-04 [26], OPC-02 [25], OPC-07 [26] are considered for the trial. In addition to the above discussed objective, a second experimental setup [27] deals with the identification of species containing a high amount of glucomannan from the same sets of DNA fingerprints of the *Amorphophallus* genus. The RAPD procedure can lead to unstable bands even after several repeats. However, for our analysis, the generated DNA fingerprints were replicated for and the bands were found to be consistent. Additionally, as we considered four sets of DNA fingerprints (each set containing multiple DNA fingerprints of our species of interest) corresponding to four different primers it was ensured that the band pattern was eventually well represented in the classifier. Moreover, RAPD was the most suitable method for our analysis as we had access to limited DNA data corresponding to the different species of the *Amorphophallus* genus.

The DNA fingerprints corresponding to individual species are initially extracted with the help of a semi-automatic procedure. A representative intensity plot corresponding to each of the input images is used to identify relevant bands in the DNA fingerprints. Depending on the local maxima identified from intensity plots corresponding to individual inputs, we identify positions that best represent the DNA fingerprint in consideration. The relative intensity and relative position of each of these pixel positions with respect to 21 fixed bands in the DNA ladder corresponding to specific markers help us to construct a 42 dimensional feature vector that represents a specific DNA fingerprint. On successful encoding of all DNA fingerprints, for the four datasets corresponding to the four different primers considered, a k-nearest neighbor (k-NN) classifier adhering to a Leave One Out Cross Validation (LOOCV) protocol is used to classify our species of interest. Finally an n-star consensus procedure sum up the result obtained from the four different classifiers.

The remaining portion of the paper is organized as follows: Section 2 briefly describes the database used in the proposed work along with the wet lab experimental procedure designed to generate the four different sets of input images. Section 3 describes the different image processing techniques adopted to encode the DNA fingerprints into representative feature vectors and discusses the design of the k-NN classifier adhering to the LOOCV protocol and the n-star consensus approach adopted to sum up the classification results. The experimental results are presented in the same section. Section 4 concludes by analyzing the proposed procedure and by providing scopes for future work.

2 Materials and methods

2.1 Database description

From the image analysis point of view, the proposed work deals with images of the DNA fingerprints that are processed in order to analyze the similarity in band patterns. The database of fingerprints was prepared by colleagues from Chulalongkorn University. The research group works on the determination of glucomannan content in different species of the *Amorphophallus* genus extracted from various geographic locations and at the same time develops molecular markers for species identification.

The DNA fingerprints are generated for species of the *Amorphophallus* genus, which are obtained from various accessions. Different sets of DNA fingerprints are obtained by treating the DNA fragments with a range of different primers. A primer is a strand of nucleic acid that serves as a starting point for DNA synthesis. These primers are usually short, chemically synthesized oligonucleotides, with a length of about ten to twenty bases. They are hybridized to a target DNA, which is then copied by the polymerase enzyme. The introduction of different primers adds dimensionality to the problem of species identification, enabling the formulation of a consensus strategy by virtue of which a robust classifier is constructed. Plant samples of *Amorphophallus* genus were obtained from different Thai provinces. The samples consisted of wild (14 species) and cultivars (23 species) specimens. All leaf materials were taken from living plants and were subsequently subjected to freeze drying. Genomic DNA was extracted by DNeasy Plant Mini Kit (Qiagen, USA). RAPD-PCR was used to generate banding patterns for the different *Amorphophallus* accessions using 4 different RAPD primers:

OPD-04 (5' -TCTGGTGAGG -3') [26] OPC-07 (5' -CACACTCCAG -3') [26] AB-04 (5' -GGCACGCGTT-3') [28] and OPC-02 (5' -GTGAGGCGTC -3') [28].

Amplification of genomic DNA was performed in a 15 μ L reaction mixture containing 1 $\times$ Thermopol buffer (2mM MgCl₂) (Bio Lab Inc., USA), 2% bovine serum albumin (BSA, 10 mg/mL) (Bio Lab Inc., USA), 1 mM dNTPs (Bio Lab Inc., USA), 1.25 μ L of primer (Operon, USA) and 0.4 U of Taq DNA polymerase (Bio Lab Inc., USA) and 30 ng of template DNA. PCR application conditions for OPD-04 and OPC-07 were an initial denaturation at 94°C for 5 min, followed by 45 cycles of 94°C for 60 s, 36°C for 60 s, and 72°C for 2 min, followed by 72°C for 4 min. PCR application conditions for AB-04 and OPC-02 were an initial denaturation at 94°C for 5 min, followed by 44 cycles of 94°C for 60 s, 35°C for 60 s, and 72°C for 2 min, followed by 72°C for 5 min. PCR products were separated by electrophoresis in a 1% (w/v) agarose gel in TAE buffer (Thermo Scientetific, USA) with DNA ladder (1 kbp and 50 bp) (Fermentus, USA) used for size estimation.

For each primer, 48 DNA fingerprints corresponding to the *Amorphophallus* genus are prepared in three gel plates. In the first gel plate, DNA fingerprints corresponding to accession number 1–14 (For species corresponding to index positions refer to Supporting Information Table 1) is prepared while in the second and third gel plates, DNA fingerprints corresponding to accession number 15–33 and 34–48 are prepared respectively. As per experimental modalities, for each gel plate a marker is developed. To sum up, a total of 192 DNA fingerprints and 12 markers corresponding to the four primers were generated.

Supporting Information Table 1 provided contains general information about the different species of the *Amorphophallus* genus considered for the experiment and collected location. The last column of Supporting Information Table 1 contains information regarding the presence or absence of glucomannan in the sample of that particular species. While Fig. 1 portrays the 48 DNA fingerprints corresponding to the 37 different species of the *Amorphophallus* genus generated with OPC-02 primer. Supporting Information Figs. 1–4 section illustrates the four different sets of DNA fingerprints used in the experiment.

Figure 1. DNA fingerprints of various species of the *Amorphophallus* genus (Primer OPC-02). M represents the marker in individual agar plates, while the numerical indices annotate the fingerprints of different species.

2.2 DNA fingerprint preprocessing

The 12 images consisting of 192 DNA fingerprints and 12 markers is considered for segmentation. The image consisting of all the DNA fingerprints corresponding to each gel plate undergo manual segmentation, by virtue of which they are segregated. The inter fingerprint spacing is identified from the complete image, and a manual segmentation procedure is adopted to slit the image according to the spacing specifications. Owing to the non-uniformly illuminated and extremely noisy image quality, we refrain from adopting an automated segmentation procedure. Moreover, the starting point of each DNA fingerprint is non-uniform and quite difficult to identify, resulting in height varying segmented DNA fingerprints. Corresponding to each primer, DNA fingerprints are developed in three different gel plates. Due to minor variations in the experimental condition and owing to the fact that different gel plates contain different number of DNA fingerprints in them, the dimension of the images of individual fingerprints corresponding to a particular primer varies to a small extent. In order to accommodate the variability, we have not performed any size normalization on the input data. The dimension of the extracted images of the DNA marker and that of the DNA fingerprints corresponding to a particular gel plate is constant but varies across different gel plates. Figure 2 presents some extracted DNA fingerprints of the *Amorphophallus* genus prepared with primer OPC-07.

Figure 2. Example of some extracted DNA fingerprints of the *Amorphophallus* genus. Primer OPC-07. M (1-14) and M (14-33) denote markers in two consecutive agar plates containing DNA fingerprints. FP# stands for the fingerprint indices of DNA fingerprints belonging to different species of the *Amorphophallus* genus.

The extracted fingerprint images have different range of intensities due to inherent variations in experimental conditions while performing DNA fingerprinting in different gel plates. An intensity normalization [29] procedure using Eqn. (1) is used to ensure common range intensities between 0 and 255 in all DNA fingerprints.

$$I = (i - i_{\min}) / (i_{\max} - i_{\min}) * 255 \quad (1)$$

where, i = grayscale value in original image, I = grayscale value in normalized image, $i_{\max}$ = maximum grayscale value in original image, $i_{\min}$ = minimum grayscale value in original image.

Contrast limited adaptive histogram equalization (CLAHE) [30] is performed on the intensity normalized DNA fingerprint images. The CLAHE operator distributes the histogram across all intensity values thereby generating high contrast output images. The operator being adaptive works locally on a predefined neighborhood. At each location, the histogram of the points in the neighborhood is computed and histogram equalization is performed but only intensity of the central pixel in the neighborhood is mapped to the transformed value. The center of the neighborhood region is then moved to an adjacent pixel location and the procedure is repeated. The intensity transformation takes place according to Eqn. (2) in which each pixel in the input image with intensity r_k is mapped into a corresponding pixel with intensity s_k in the output image using a histogram equalization transformation function $T(r_k)$.

$$s_k = T(r_k) = (L - 1) \sum_{j=0}^k p_r(r_j) = \frac{(L - 1)}{MN} \sum_{j=0}^k n_j \quad (2)$$

where, $k = 0, 1, 2, \dots, L - 1$, such that, L is the number of possible intensity level in the image, $M \times N$ is the total number of pixels in the image, $p_r(r_k)$ is the probability of occurrence of level r_k in the image, and n_j is the number of pixels with intensity level j . The size of the neighborhood is empirically determined to be 5, where a 5×5 sized window is considered and the contrast limit is set as 255.

2.2.1 Annotation of relevant bands in DNA fingerprints

A series of procedures convert individual fingerprints to a 42 dimensional feature vector that represent each DNA fingerprint based on the relative intensity and relative position of prominent bands with respect to prominent bands in corresponding molecular-weight size markers.

2.2.1.1 Identification of bands based on local maxima in vertical intensity plots.

In order to construct an initial feature vector corresponding to a DNA fingerprint, a 5 pixel wide mid-section is selected from it. The mean intensity value M_i of the 5 columns for the i^{th} row forms a feature. The number of features in the feature vector is determined by the number of rows present in the size normalized DNA fingerprint (i.e. the length of the DNA fingerprint in pixels) generated by a specific primer. Thus the initial feature vector can be expressed as $\mu = \{M_1, M_2, M_3, \dots, M_n\}$ where M_i denotes the i^{th} feature or the mean intensity value of the i^{th} row. The last feature being M_n suggests the DNA fingerprint under consideration has to be of length n . Corresponding to each mean intensity feature vector, we evaluate local maxima or peaks that represent prominent bands in a given fingerprint. The local maxima considered should have corresponding intensity greater than an empirically determined threshold τ_b ($\tau_b = 60$). The row positions and the intensity values of the relevant local maxima are represented by vectors $\mathbf{p} = \{R_1, R_2, R_3, \dots, R_l\}$ and $\mathbf{n} = \{I_1, I_2, I_3, \dots, I_l\}$ where, R_k and I_k correspond to the row position and intensity value of the k^{th} maximum and l denotes the number of maxima considered for a particular DNA fingerprint. Additionally, we evaluate the local maxima or peaks corresponding to an ideal marker. While we can have variable number of local maximas corresponding to arbitrary fingerprints, we get 21 distinct local maxima corresponding to the 21 prominent bands in an ideal marker. Figure 3 represents a vertical intensity plot of the mean intensity vector corresponding to an ideal marker.

2.2.1.2 Mapping DNA fingerprints according to an ideal marker.

Along with the ideal marker, we consider 12 other markers generated with the help of the 4 designated primers in 12 agar plates. For all the markers we identify the top most and the bottom most row positions that have relevant marker information. In order to identify the specified positions in a marker, we consider the mean intensity feature vector μ_{Mi} where i stands for the i^{th} marker. By convention, we consider μ_{M0} as the mean intensity feature vector of the ideal marker. From each marker we identify row index *top* and *bottom* that satisfies the following criterion-

$$\begin{aligned} \text{top} &= (j | \text{the smallest row position } j, \text{ s.t } \tau_e \\ &< (M_j + M_{j+1} + M_{j+2} + M_{j+3} + M_{j+4}) / 5 < \tau_b) \end{aligned} \quad (3)$$

$$\begin{aligned} \text{bottom} &= (j | \text{the largest row position } j, \text{ s.t } \tau_e \\ &< (M_j + M_{j-1} + M_{j-2} + M_{j-3} + M_{j-4}) / 5 < \tau_b) \end{aligned} \quad (4)$$

where M_j is the mean intensity value of the j^{th} row of the marker under consideration. τ_b and τ_e are empirically

Figure 3. Vertical intensity plot of the mean intensity vector corresponding to an ideal marker.

determined thresholds and at assigned values 60 and 15 respectively.

Prominent row positions in an ideal marker are represented by vector $\mathbf{p}_{im} = \{R_{0im}, R_{1im}, R_{2im}, \dots, R_{21im}, R_{22im}\}$ where R_{0im} and R_{22im} represent the top and bottom row indices in an ideal marker and $R_{1im}, R_{2im}, \dots, R_{21im}$ represent the 21 maxima corresponding to the ideal marker. Figure 4B illustrates the top, bottom and the 21 band positions in between with the help of white lines.

Hence, we have determined *top* and *bottom* indices corresponding to the ideal marker and the 12 markers developed in the 12 different agar plates. We map the relevant row positions $\mathbf{p}_{fv} = \{R_{1fv}, R_{2fv}, R_{3fv}, \dots, R_{1fv}\}$ denoting the maxima in an arbitrary fingerprint $\mathbf{fv}$ to mapped row positions $\mathbf{p}'_{fv} = \{R'_{1fv}, R'_{2fv}, R'_{3fv}, \dots, R'_{1fv}\}$ as follows:

$$R'_{kfv} = R_{kfv} * \left(\frac{bottom_{M0} - top_{M0}}{bottom_{Mi} - top_{Mi}} \right) + (top_{M0} - top_{Mi}) \quad (5)$$

where top_{M0} and $bottom_{M0}$ denote the *top* and *bottom* row indices of the ideal marker while top_{Mi} and $bottom_{Mi}$ denote the *top* and *bottom* row indices of the marker developed in the agar plate to which the relevant DNA fingerprint belongs. R_{kfv} correspond to the k th maximum in fingerprint $\mathbf{fv}$ while R'_{kfv} corresponds to the mapped k th maximum in the same DNA fingerprint. With the help of Eqn. (5) we successfully map prominent bands in any fingerprint corresponding to row indices in the ideal marker.

2.2.1.3 Feature vector generation. Once we have represented all prominent band positions in any given DNA fingerprint according to the row indices of the ideal marker, we construct a 42 dimensional feature vector λ corresponding to each DNA fingerprint. For each band (local maxima) in the ideal marker we try to find the best band (if any) in a given DNA fingerprint that supports it. By support we mean that we have a relatively prominent band corresponding to a prominent band position in the DNA fingerprint. Corresponding to each prominent band in the ideal marker we assign a relative intensity and relative position value that supports that particular maxima with respect to a

Figure 4. a- Bands corresponding to an ideal marker, b- row positions in an ideal marker (top bottom and the 21 in between) denoted by single pixel wide white lines, c- bands of a marker developed in an agar plate, d- intensity normalized and CLAHE operated marker; top and bottom positions denoted by single pixel wide white lines.

given fingerprint. For a given fingerprint $\mathbf{fv}$, the 42 dimensional feature vector $\lambda_{fv} = \{rp_{1fv}, ri_{1fv}, rp_{2fv}, ri_{2fv}, \dots, rp_{21fv}, ri_{21fv}\}$ can be computed as follows

$$rp_{kfv} = rp_{indfv} \quad (6)$$

$$ri_{jfv} = ri_{indfv} \quad (7)$$

$$ind = \operatorname{argmax}_j \forall j \in R_{(l-1)im} \leq R'_{jfv} \leq R_{(l+1)im} (rp_{jfv} * ri_{jfv}) \quad (8)$$

$$rp_{jfv} = \begin{cases} 1 - \frac{R'_{jfv} - R_{lim}}{R_{(l+1)im} - R_{lim}} & \text{if } (R'_{jfv} \geq R_{lim}) \\ 1 - \frac{R_{lim} - R'_{jfv}}{R_{(l)im} - R_{(l-1)im}} & \text{if } (R'_{jfv} < R_{lim}) \end{cases} \quad (9)$$

$$ri_{jfv} = \frac{M_{jfv}}{M_{maxfv}} \quad (10)$$

$$M_{maxfv} = \max_{vj} (M_{jfv}) \quad (11)$$

Once we have calculated the 42 dimensional feature vector λ_{fv} for each DNA fingerprints corresponding to the four markers, we move towards the classification procedure. The feature vector for all 48 DNA fingerprints for primer AB-04, OPD-04, OPC-02 and OPC-07 is illustrated in Supporting Information Figs. 5, 6, 7 and 8.

2.3 Experimental design

2.3.1 k-NN classification and the LOOCV protocol

In pattern recognition, the k-Nearest Neighbors algorithm is a non-parametric method used for classification and regression. [31] In both cases, the input consists of the k closest training examples in the feature space. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

In the present context, the DNA fingerprints are represented in a 42 dimensional feature space. For the *A. muelleri* classification problem, the class membership of each DNA fingerprint is determined by ascertaining whether the fingerprint in question belongs to *A. muelleri* or not. In order to classify species containing glucomannan, the class membership of each DNA fingerprint is determined by ascertaining whether the fingerprint in question contains glucomannan or not. The training and testing of the k-NN classifiers (for different values of k) is carried out using the LOOCV protocol. In the LOOCV protocol, considering n instances of input data, (n-1) instances of input data is used for training purpose, while 1 instance of the input data is used for testing purpose. Adhering to the LOOCV protocol we test all instances of input data.

In spite of the presence of sophisticated machine learning algorithms, our reliance on the k-NN algorithm with LOOCV protocol is due to the fact that limited numbers of input samples are present for training and testing purposes. The samples considered for the experiment belong to diverse geographical locations and performing multiple wet lab experiments using multiple primers is a costly affair.

2.3.2 The n-star consensus strategy

In order to develop the consensus strategy for the present work, we assign a n-star quality positive prediction score [32–34] to a test sequence, where n is the number of classifiers agreeing for the test sample (DNA fingerprint) under consideration to be positive. Here positive means correct classification, i.e. an *A. muelleri* positive sample is classified as positive and an *A. muelleri* negative sample is classified as negative. For example, when a test sequence is classified as positive by all classifiers under consideration, the strength of positive prediction is said to be of n-star quality. On the other hand, if only one network predicts the test sequence to be positive, the prediction quality is 1-star. In order to test an unknown sequence, a user may tune the quality of prediction by choosing a specific value of n. It is to be noted that, for any value of n = m; m ≥ 1, quality consensus of the order (m – 1)–star is considered as a negative prediction.

In general, we define a n-star quality consensus scheme as C_n^N , where N is the number of classifiers participating in the specific consensus strategy, and n (1 ≤ n ≤ N) determines the quality of prediction. More categorically, a 1-star prediction says that any one of possible N classifiers predicts the test sample (fingerprint) to be positive, and a N-star represents that all classifiers agreed to the decision. Along this principle, we define the 4-star (for the 4 classifiers generated with the help of 4 different primers) quality consensus prediction C_n^4 as the consensus over 4 classifiers.

Based on the values of the statistical parameters precision and recall, the performance of the consensus strategy is evaluated. Although the n-star consensus strategy is user defined, i.e. it solely depends on the qualitative needs of the user, the value of n for which C_n^4 performs best, can be thought of as the best classifier combination strategy.

2.3.3 Overview of statistical parameters

In the context of statistical analysis, Recall is referred to as the True Positive Rate or Sensitivity. It calculates the proportion of actual positives which are correctly classified.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (12)$$

Precision also referred to as the positive predictive value is the proportion of positive test results that are true positives. A small precision value, say of the order of 15% or below, indicates that many of the positive results from this testing procedure are false positives. This indicates the fact that it will be necessary to follow up any positive result with a more reliable test to obtain a more accurate assessment as to whether the null hypothesis holds or not. Precision is the ratio of the true positives to the summation of true positives and false positives.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (13)$$

Accuracy can be defined as a systems degree of veracity and is defined as

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}} \quad (14)$$

A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score. F-score is a factor in which the Recall and Precision scores are evenly weighted. F-score is an important measure while considering two class problems, and often outscores the measure of Accuracy. F-score is calculated as follows-

$$F = 2 \cdot \left(\frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \right) \quad (15)$$

The Matthews correlation coefficient (MCC) is another important statistical measure and is used in machine learning to evaluate the quality of binary (two-class) classifications. It takes into account true positive, false positive, true negative, and false negative and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The Matthews correlation coefficient is meticulously used to analyze the statistical parameters and to eliminate the skewness of the accuracy measure in uneven sized two class problems.

$$\text{MCC} = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (16)$$

where, *TP*, *TN*, *FP* and *FN* denote the true positives, true negatives, false positives and false negatives classified by a particular classifier.

3 Results and discussion

3.1 Classification of *A. muelleri* species

Table 1 corresponds to the accuracy, precision, recall and F-Score measures while classifying species *A. muelleri* from DNA fingerprints of the *Amorphophallus* genus. Here PR 1, PR 2, PR 3 and PR 4 denote primers AB-04, OPB-04, OPC-02 and OPC-07 respectively. Each of the 48 fingerprints belonging to species of the *Amorphophallus* genus has been classified based on 1-NN, 3-NN, 5-NN and 7-NN classifiers adhering to the LOOCV protocol. While considering the N star consensus strategy, the test sample is classified as positive only if all the N classifiers (in our case the classifiers corresponding to the four primers) classify it as positive. As we have four classifiers corresponding to the four different primers, the best we can ask for is a four star consensus wherein all the four classifiers agree. The Best N star implies N star consensus considering the best 3 classifiers. While considering the performance of the primers alone, Primer 4 (KNN1) performs best with an accuracy of 95.83%. Considering N star consensus the accuracy of the reasonably confident 3 STAR with KNN1 classification is 91.67%. Taking into account performance based on the precision parameter, considering the performance of the primers alone, Primer 2 (KNN5 & KNN7) performs best

with a Precision of 100%. Considering N star consensus the Precision of the reasonably confident 1 STAR, 2 STAR, 3 STAR & 4 STAR with KNN3, KNN5 & KNN7 classification is 100%. In view of performance based on the recall parameter, considering the performance of the primers alone, Primer 2 performs best with a recall of 100%. Considering N star consensus the recall of the reasonably confident 3 STAR with KNN7 classification is 100%. The F-score measure provides a balanced insight to the performance evaluation of the classifiers. While considering the performance of the primers alone, Primer 2 (KNN7) performs best with a F-Score of 100%. Considering N star consensus the F-Score of the reasonably confident 3 STAR with KNN7 classification is 100%. The graphs pertaining to the accuracy, precision, recall, F-Score and MCC measures are provided in Supporting Information Figs. 9 and 10 (best of 1 STAR, 2 STAR and 3 STAR).

The classification accuracy of the proposed classifier while classifying species *A. muelleri* outperforms the classifier proposed in [27]. The classifier proposed in [27] uses the same dataset and with the help of a simple correlation coefficient based consensus classifier predicted species *A. muelleri* with an accuracy of 90%. Whereas the classifier designed with the feature vector definition and classification protocol proposed in the present work has higher classification accuracy considering both single primers and N star consensus.

3.2 Classification of species containing glucomannan

Table 2 corresponds to the accuracy, precision, recall and F-Score measures while classifying species containing Glucomannan from DNA fingerprints of the *Amorphophallus* genus. Here PR 1, PR 2, PR 3 and PR 4 denote primers AB-04, OPB-04, OPC-02 and OPC-07 respectively. Each of the 48 fingerprints belonging to species of the *Amorphophallus* genus has been classified based on 1-NN, 3-NN, 5-NN and 7-NN classifiers adhering to the LOOCV protocol. While considering the performance of the primers alone, both Primer 3 and Primer 1 perform best with an accuracy of 85.41%. Considering N star consensus, the accuracy of the reasonably confident 3 STAR with KNN1 classification is 91.66%. Taking into account performance based on the precision parameter, considering the performance of the primers alone, Primer 1 (KNN1 & KNN3) perform best with a precision of 80% and 81.8%. Considering N star consensus, the precision of the reasonably confident 3 STAR & 4 STAR with KNN1, and KNN1, KNN3 classification is 100%. In view of performance based on the recall parameter, considering the performance of the primers alone, both Primer 3 and Primer 1 perform best with a recall of 75%. Considering N star consensus, the recall of the reasonably confident 3 STAR with KNN1 classification is 76.47%. While considering the performance of the primers alone, both Primer 3 and Primer 1 performs best with a F-Score of 77.41%. Considering N star consensus the F-Score of the

Table 1. Accuracy, precision, recall and F-score measures when classifying *A. muelleri*

Specs.	Accuracy				Precision				Recall				F-Score			
	NN1	NN3	NN5	NN7	NN1	NN3	NN5	NN7	NN1	NN3	NN5	NN7	NN1	NN3	NN5	NN7
PR1	83.3	85.4	75	83.3	97.2	97.4	100	100	83.3	86.0	85.7	95.2	89.7	91.3	92.3	97.5
PR2	64.5	79.1	85.4	87.5	83.8	100	100	100	73.8	90.5	97.6	100	78.5	95.0	98.8	100
PR3	85.4	89.6	85.4	85.4	92.7	93.0	100	100	90.5	95.2	97.6	97.6	91.6	94.1	98.8	98.8
PR4	95.8	87.5	85.4	79.1	97.6	95.2	97.6	100	97.6	95.2	95.2	90.5	97.6	95.2	96.4	95
1STR	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
2STR	97.91	100	100	100	100	100	100	100	97.67	100	100	100	98.94	100	100	100
3STR	85.41	91.67	97.91	100	90.69	95.23	100	100	92.85	95.23	97.91	100	91.75	95.23	98.79	100
4STR	47.91	66.67	79.16	85.41	79.31	88.23	97.05	100	54.76	71.42	78.57	83.34	64.78	78.94	86.83	90.90
B1SR	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
B2SR	93.75	93.75	100	100	97.56	95.34	100	100	95.23	97.61	100	100	96.36	96.45	100	100
B3SR	70.83	75	89.58	93.75	88.89	92.30	97.43	100	76.19	80	90.47	92.85	82.05	85.71	93.82	96.29

PR*, *STR and B*SR stands for the primer number, n-star and best of n-star measures. NN* stands for k-NN*

Table 2. Accuracy, precision, recall and F-score measures while classifying species containing glucomannan

Specs.	Accuracy			Precision			Recall			F-Score		
	NN1	NN3	NN5	NN1	NN3	NN5	NN1	NN3	NN5	NN1	NN3	NN5
PR1	85.4	68.7	68.7	80	81.8	56.2	75	56.2	56.2	77.4	66.7	56.2
PR2	60.4	68.7	64.6	42.1	52.6	50	50	62.5	43.7	45.7	57.1	46.7
PR3	85.4	72.9	60.4	80	75	33.3	75	37.5	6.6	77.4	50	11.0
PR4	75	75	79.2	64.7	64.7	73.3	68.7	68.7	68.7	66.7	66.7	71.0
1STR	95.83	97.91	100	93.75	100	100	93.75	93.75	100	90.62	96.34	100
2STR	89.58	87.5	87.5	86.67	85.71	91.66	81.25	75	68.75	83.87	79.98	78.57
3STR	79.16	75	58.83	66.67	66.67	16.67	75.0	50	6.25	70.58	57.14	9.09
4STR	41.67	41.67	43.75	15.78	7.14	0	20	6.25	0	17.64	6.66	0
B1SR	93.75	95.83	100	93.34	93.75	100	87.5	93.75	100	90.32	93.75	100
B2SR	85.41	81.25	79.16	80	76.92	68.75	75	62.5	68.75	77.41	68.96	68.75
B3SR	68.75	43.75	39.58	52.94	13.34	0	56.25	12.5	0	54.54	12.90	0

PR*, *STR and B*SR stands for the primer number, n-star and best of n-star measures. NN* stands for k-NN*

reasonably confident 3 STAR with KNN1 classification is 86.63%. The graphs pertaining to the accuracy, precision, recall, F-Score and MCC measures are provided in Supporting Information Figs 11 and 12 (best of 1 STAR, 2 STAR and 3 STAR).

3.3 Feature ranking for classification of *A. muelleri*

This section aims at identifying the 3 best bands (6 features) that help best in distinguishing of *A. muelleri* from other species of the *Amorphophallus* genus. The objective is to help in the wet lab experimentation process by enumerating the 3 best bands that one should look for in a DNA fingerprint to identify the *A. muelleri* species. In order to estimate the 3 best bands, we have performed the classification experiment with the help of k-NN classifiers adhering to the LOOCV protocol by considering just 2 features (relative position and relative intensity) corresponding to each band. The 3 best feature pairs that provide us with the highest classification accuracy are considered. The ranking is done on the 3 best primers as earmarked by our experimental collaborators. Figure 5 presents the best, second best and the third best bands (for the classification of *A. muelleri*) for primers OPB-04, OPC-02 and OPC-07.

3.4 Feature ranking for classification of species containing glucomannan

Similar to the previous section, this section provides information regarding the 3 best bands that one should look for while classifying species containing glucomannan. In order to estimate the three best bands, we have performed the classification experiment with the help of k-NN classifiers adhering to the LOOCV protocol by considering just two features (relative position and relative intensity) corresponding to each band. The three best feature pairs that provide us with the highest classification accuracy are considered. The ranking is done on the three best primers as earmarked by our experimental collaborators. Figure 6 presents the best, second best and the third best bands (for the classification of species containing glucomannan) for primers OPB-04, OPC-02 and OPC-07.

4 Concluding remarks

The research work intends to device a robust, state of the art automatic species identification system relying solely on

Figure 5. Ranking of features for *A. muelleri* classification, where the red band(s) refer to the most significant band(s), the orange band(s) refer to the 2nd most and the yellow band(s) refer to the third most significant band(s).

Figure 6. Ranking of features for glucomannan classification, where the red band(s) refer to the most significant band(s), the orange band(s) refer to the 2nd most and the yellow band(s) refer to the 3rd most significant band(s).

computational analysis of DNA fingerprints. The work derives its motivation from the non-existence of an automatic technique to classify species based on non-morphological traits. The existing techniques depending on morphological or bio-chemical factors of a species are seldom automated, and are plagued by inconsistencies. The classifier designed to classify the species *A. muelleri* belonging to the *Amorphophallus* genus displays a reasonable classification accuracy and can be easily extended to predict any other species in a similar taxonomic system. Additionally and predominantly, a similar kind of classification system designed to classify species based on their glucomannan content has been tested with reasonable accuracy. High recall and F-Score values corresponding to performance evaluation of both classifiers further established the robustness of the proposed system. Additionally, the feature ranking system that has been devised with respect to the three different primers for the classification of both *A. muelleri* and glucomannan provides a ranking of the bands according to their relative importance in the classification procedure. Similar classification systems adhering to identical protocols can be used in future to classify other species of interest from species of any particular genus. Moreover, this system can be useful as universal database program to identify high KGM species for tests involving several laboratories. It could also be used as prototype development program to identify other living organisms in future. In a nutshell, image-analytic approach to automatic classification of DNA fingerprints may be considered as a relatively unexplored application area, and the current work is an attempt to address the real-life problem with reasonable accuracy and scopes for further research.

Practical application

The present body of work aims at automatic classification of a species of interest from DNA fingerprints of species of a given genus. A classification system has been constructed considering DNA fingerprints of the *Amorphophallus* genus. The same system can be extended to classify other species provided there are sufficient fingerprints available to train the classifier. The proposed classifier is robust to poor quality DNA fingerprint images and can efficiently classify species from fingerprints prepared in different lab setups. Apart from identifying species the same classifier has been used to classify species based on their glucomannan content from DNA fingerprint images of species belonging to the *Amorphophallus* genus.

This project is partially supported by the CMATER research laboratory of the Computer Science and Engineering Department, Jadavpur University, India, PURSE-II and UPE-II project and Research Award (F.30-31/2016 (SA-II)) by UGC, Government of India.

The authors have declared no conflict of interest.

5 References

- [1] McPherSon, S., Hetterscheid, W. L. A., *Amorphophallus* in the wild and in cultivation. *The Plantsman Magazine, Royal Horticult. Soc.* 2011, 10, 90–97.
- [2] Hetterscheid, W., Ittenbach, S., Everything you always wanted to know about *Amorphophallus*, but were afraid to stick your nose into. *Aroidiana* 1996, 19, 7–131.
- [3] Hetterscheid, W. L. A., New species from SE Asia. *Aroidiana/Journal Int. Aroid Soc.*, 2006, 29, 53–79.
- [4] Sedayu, A., Eurlings, M. C., Gravendeel, B., Hetterscheid, W. L., Morphological character evolution of *Amorphophallus* (Araceae) based on a combined phylogenetic analysis of trnL, rbcL and LEAFY second intron sequences. *Bot. Stud.* 2010, 51, 473–490.
- [5] Hetterscheid, W., Claudel, C., Flottbek, B. K., A new miniature *Amorphophallus* (Araceae) from Eastern Thailand. *Aroidiana* 2012, 35, 47–50.
- [6] Giordano, C., Karyological and Palynological Observations on *Amorphophallus titanum* (Becc.) Becc. ex Arcangeli (Araceae). *Caryologia* 1999, 52, 65–73.
- [7] Punekar, S. A., Kumaran, K. P. N., Pollen morphology and pollination ecology of *Amorphophallus* species from North Western Ghats and Konkan region of India. *Flora-Morphology, Distrib. Funct. Ecol. Plants* 2010, 205, 326–336.
- [8] Kite, G. C., Hetterscheid, W. L. A., Inflorescence odours of *Amorphophallus* and *Pseudodracontium* (Araceae). *Phytochemistry* 1997, 46, 71–75.
- [9] Kite, G. C., Hetterscheid, W. L. A., Lewis, M. J., Boyce, P. C. et al., Inflorescence odours and pollinators of *Arum* and *Amorphophallus* (Araceae). *Reprod. Biol.* 1998, 295–315.
- [10] Grob, G. B. J., Gravendeel, B., Eurlings, M. C. M., Hetterscheid, W. L. A., Phylogeny of the tribe Thomsonieae (Araceae) based on chloroplast matK and trnL intron sequences. *Syst. Bot.* 2002, 27, 453–467.
- [11] Song, J., Yao, H., Li, Y., Li, X. et al., Authentication of the family Polygonaceae in Chinese pharmacopoeia by DNA barcoding technique. *J. Ethnopharmacol.* 2009, 124, 434–439.
- [12] Kumar, L. S., DNA markers in plant improvement: An overview. *Biotechnol. Adv.* 1999, 17, 143–182.
- [13] Dixon, R. A., Bouton, J. H., Narasimhamoorthy, B., Saha, M., Beyond structural genomics for plant science. *Adv. Agron.* 2007, 95, 77–161.
- [14] Sharma, A., Namdeo, A. G., Mahadik, K., Molecular Markers: New Prospects in Plant Genome Analysis. *Pharmacogn. Rev.* 2008, 2, 23–34.
- [15] Khan, S., Mirza, K. J., Al-Qurainy, F., Abdin, M. Z., Authentication of the medicinal plant *Senna angustifolia* by RAPD profiling. *Saudi J. Biol. Sci.* 2011, 18, 287–292.
- [16] Wenbing, C., Chunlin, L., Jianfu, Z., Study on genetic diversity of RAPD markers in *Amorphophallus*. *J. Agric. Biotechnol.* 2001, 4, 24.
- [17] Venkatachalam, L., Sreedhar, R. V., Bhagyalakshmi, N., The use of genetic markers for detecting DNA polymorphism, genotype identification and phylogenetic relationships among banana cultivars. *Mol. Phylogenet. Evol.* 2008, 47, 974–985.
- [18] Daud, S. K., Zaidi, N. S. M., Fadzillah, N., Mahmood, M., The use of LP-RAPD for assessing genetic relatedness among selected banana cultivars. *Pertanika J. Trop. Agric. Sci.* 2006, 29, 25–34.
- [19] Williams, J. G. K., Kubelik, A. R., Livak, K. J., Rafalski, J. A. et al., DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.* 1990, 18, 6531–6535.
- [20] Bardakci, F., Random amplified polymorphic DNA (RAPD) markers. *Turkish J. Biol.* 2001, 25, 185–196.
- [21] Li, J. J., Pei, G. L., Pang, H. X., Bilderbeck, A. et al., A new method for RAPD primers selection based on primer bias in nucleotide sequence data. *J. Biotechnol.* 2006, 126, 415–423.
- [22] Chen, H.-L., Cheng, H.-C., Wu, W.-T., Liu, Y. J. et al., Supplementation of konjac glucomannan into a low-fiber Chinese diet promoted bowel movement and improved colonic ecology in constipated adults: a placebo-controlled, diet-controlled trial. *J. Am. Coll. Nutr.* 2008, 27, 102–108.
- [23] Walsh, D. E., Yaghoobian, V., Behforooz, A., Effect of glucomannan on obese patients: a clinical study. *Int. J. Obes* 1984, 8, 289–293.
- [24] Banerjee, A., Basu, S., Mekkerdchoo, O., Nasipuri, M. et al., Automatic classification of *A. paeoniifolius* species from DNA fingerprints of *Amorphophallus* genus, in Communications, Devices and Intelligent Systems (CODIS), 2012 International Conference on, 2012, 580–583.
- [25] G. Study, Chokenamchai, S. 2003 Genetic study and variety improvement of elephant yams (*Amorphophallus* sp.) by colchicine. K. U. Department of Genetics.
- [26] Poerba, Y. S., Martanti, D., Genetic variability of *Amorphophallus muelleri* Blume in Java based on Random Amplified Polymorphic DNA. *Biodiversitas J. Biol. Div.* 2008, 9, 245–249.
- [27] Banerjee, A., Basu, S., Mekkerdchoo, O., Srzednicki, G. et al., Automatic Classification of *A. muelleri* Species from DNA Fingerprints of *Amorphophallus* Genus, in Proceedings of International Symposium on Agri-Foods for Health and Wealth (AFHW2013), 2014, 101–107.
- [28] Chokenamchai, S., Genetic study and variety improvement of elephant yams (*Amorphophallus* sp.) by colchicine (Masters Thesis). Kasetsart University, Bangkok, Thailand, 2003.
- [29] Mulyono, D., Jinn, H. S. J. H.S., A study of finger vein biometric for personal identification. *Int. Symp. Biometrics Secur. Technol.* 2008, 1–8.
- [30] Zuiderveld, K., Contrast limited adaptive histogram equalization, in *Graphics gems IV*, 1994, 474–485.
- [31] Altman, N. S., An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* 1992, 46, 75–185.
- [32] Plewczynski, D., Basu, S., Saha, I., AMS 4.0: consensus prediction of post-translational modifications in protein sequences. *Amino Acids* 2012, 43, 573–582.
- [33] Chatterjee, P., Basu, S., Kundu, M., Nasipuri, M. et al., PSP_MCSVM: brainstorming consensus prediction of protein secondary structures using two-stage multiclass support vector machines. *J. Mol. Model* 2011, 17, 2191–2201.
- [34] Chatterjee, P., Basu, S., Kundu, M., Nasipuri, M. et al., Improving Prediction of Interdomain Linkers in Protein Sequences Using a Consensus Approach, in Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012, 2012, 111–118.