World Academy of Science, Engineering and Technology
International Journal of Bioengineering and Life Sciences
Vol:1, No:12, 2007

Coefficient of Parentage for Crop Hybridization

Manpreet Singh, Parvinder Singh Sandhu, and Basant Raj Singh

Abstract**—**Hybridization refers to the crossing breeding of two plants. Coefficient of Parentage (COP) is used by the plant breeders to determine the genetic diversity across various varieties so as to incorporate the useful characters of the two varieties to develop a new crop variety with particular useful characters. Genetic Diversity is the prerequisite for any cultivar development program. Genetic Diversity depends upon the pedigree information of the varieties based on particular levels. Pedigree refers to the parents of a particular variety at various levels. This paper discusses the searching and analyses of different possible pairs of varieties selected on the basis of morphological characters, Climatic conditions and Nutrients so as to obtain the most optimal pair that can produce the required crossbreed variety. An algorithm was developed to determine the coefficient of parentage (COP) between the selected wheat varieties. Dummy values were used wherever actual data was not available.

Keywords**—**Coefficient of Parentage, Morphological characters, Pedigree, Genetic Diversity.

I. Introduction

GENETIC Diversity refers to the different genes and variation generally found within a species, for example, a variation among genes across different wheat varieties. Diverse genetic resources allow humans to select and breed plants with desired characteristics, thus increasing agricultural productivity.

Knowledge of genetic diversity is important for plant breeding. Monitoring of genetic diversity can form a basis for rational correction of breeding programs and the strategies in plant industry. With the computerization and digitalization techniques advancing, the cost of storage is decreasing and great amount of data are now available in the field of biotechnology. Such data may provide a rich resource for knowledge discovery and decision support.

In order to understand, analyze, and eventually make use of the huge amount of data, a multidisciplinary approach, data mining, is proposed to meet the challenge. Data mining is the process of identifying interesting patterns from large databases. The objective of our paper is to describe the data mining algorithm to compute the coefficient of parentage across wheat varieties. [4]

Manuscript received October 9, 2001.

Manpreet Singh is with the Department of CSE & IT, Guru Nanak Dev Engineering College, Ludhiana (e-mail: mpreet78@yahoo.com).

Parvinder Singh Sandhu is with Department of CSE, Rayat and Bahara Institute of Engineering and Technology, Sahauran, Distt. Mohali, Punjab, India (e-mail: parvinder.sandhu@gmail.com).

Basant Raj Singh is student of M.Tech CSE in Guru Nanak Dev Engineering College, Ludhiana (e-mail: sonigill_1@yahoo.co.in).

II. Problem Statement and Solution Approach

Bioinformatics is the science of managing, mining and interpreting information from biological sequences and structures. In this area of science, biology, computer science and information technology, all the three merges into single discipline. During the last few years, bioinformatics has been overwhelmed with increasing floods of data, both in terms of volume and in terms of new databases and new types of data.

The problem is to access such a large amount of data and get the useful information. Due to the growing size and complexity of the biological data, it is necessary to explore newer technologies to handle the large databases efficiently. There is a strong interest in employing methods of knowledge discovery and data mining to generate models of biological systems. Mining biological databases imposes challenges which knowledge discovery and data mining have to address. Analyzing data from biological databases often requires the consideration of data from multiple relations rather than from one single table.

Morphological characters are the various parameters related to the wheat varieties as shown in Table 1. If we take any desired morphological characteristics as input we get a list of varieties satisfying the conditions. This list can be processed, taking a pair of varieties at a time to find out the most optimal and probable pair of genetically diverse varieties. The results can be shown graphically, depicting the genetic diversity among the varieties based on the pedigree levels. We also get a percentage probability of getting the required hybrid breed as an output. [3]

III. Solution Methodology

Database was created for the different varieties of wheat. It contains the pedigree information and the morphological characters for the different varieties of wheat. First of all, the varieties are selected on the basis of morphological characters, climatic conditions, nutrients etc. This information is important to develop a variety with particular useful characters. Now we have to determine the genetic diversity between the varieties.

Open Science Index, Bioengineering and Life Sciences Vol:1, No:12, 2007 publications.waset.org/8942.pdf

TABLE I
MORPHOLOGICAL CHARACTERS

Varity	Days Flowering	Tiller Number	Plant Height	TG Weight	Grains per Ear	Bio Yield	Grain Yield	Harvest Index
KAVKAZ	95	7.0	87.55	25.58	41.12	17.6	5.11	28.89
TONORI	80.5	5.5	72.6	29.18	33.37	11.6	3.85	33.06
SONARA	81	5.0	76.0	38.1	31.37	11.4	3.83	33.27
BW 11	86	5.8	67.26	29.6	43.62	18.7	5.94	31.46
GENARO8	91.5	6.7	74.3	27.42	35.75	11.0	3.14	28.33

For this, we need to compare the pedigree information (parentage) of the varieties. The varieties which are selected they form a list. Different pairs of varieties are analyzed to calculate the probability percentage of obtaining the desired variety using the algorithm shown in the form of flow chart in Fig. 1. The formula used for calculating the results is given below:

$$P_{i+1} = P_i + \frac{100 * D_{i+1}}{(C_{i+1})^2} - \sum_{i \neq j} \eta_i \eta_j M_{ij} \text{ Where}$$

$\mathbf{P_i}$ is the percentage probability for the varieties to be genetically diverse up to level 'i'.

$\mathbf{C_i}$ and $\mathbf{D_i}$ correspond to the number of varieties in a level and the number of distinct varieties in a level respectively.
$\boldsymbol{\eta_i}$ and $\boldsymbol{\eta_j}$ are constant for pedigree level 'i' and 'j' respectively indicating the effect of that level on the genetic diversity.
$\mathbf{M}_{ij}$ indicates the similar number parents at ith level of one variety and jth level of another variety.
And

$$\eta_i = \frac{1}{2^{i+1}}$$

So higher the value of $\mathbf{P_i}$, greater will be the genetic diversity between the crop varieties.

Also the algorithm includes that if the value of i is maximum (top most level) and then it adds the value $100/C_i$ to the percentage probability (P_{max}) so that if the parents of two varieties are all distinct then it shows the value for P = 100% (that is the varieties are 100% genetically diverse from each other) because no more information about ancestors is available from which the %age of genetic diversity can be computed. This formula can be written down as:

$$P_{i_{max}} = P_{i_{max}} + \frac{100}{C_{i_{max}}}$$

where i_{max} is the top hierarchical level in the pedigree tree

and P_i is the percentage probability for the varieties to be genetically diverse up to level 'i'

MATLAB is utilized to plot a graph between pedigree levels and genetic diversity utilizing the formula for $\mathbf{P_i}$. 100% genetically diverse varieties are represented with the help of a straight horizontal line as shown in Fig. 2. The varieties that are not genetically diverse i.e. indicating the same crop variety is represented with a straight line across the pedigree levels as shown in Fig. 3. The most commonly obtained graph is shown in Fig. 4. The downward step indicates the similarity of some parents of the two varieties at that particular level.

IV. CONCLUSION

The present model computes the value of Parentage Coefficient and provides the result in the form of percentage that can be compared for different parent sets. The results are also shown graphically which further provides a detailed and complete description of the diversity between the parent varieties at their respective pedigree level.

Such a model will help the plant breeders in selecting the parent set which is most likely to produce the desired result. It will save both time and resources which are otherwise wasted in carrying out the actual crossbreeding process, repeatedly executed for different varieties until the potential crop variety with desired morphological characters is developed.

Open Science Index, Bioengineering and Life Sciences Vol:1, No:12, 2007 publications.waset.org/8942.pdf

Open Science Index, Bioengineering and Life Sciences Vol:1, No:12, 2007 publications.waset.org/8942.pdf

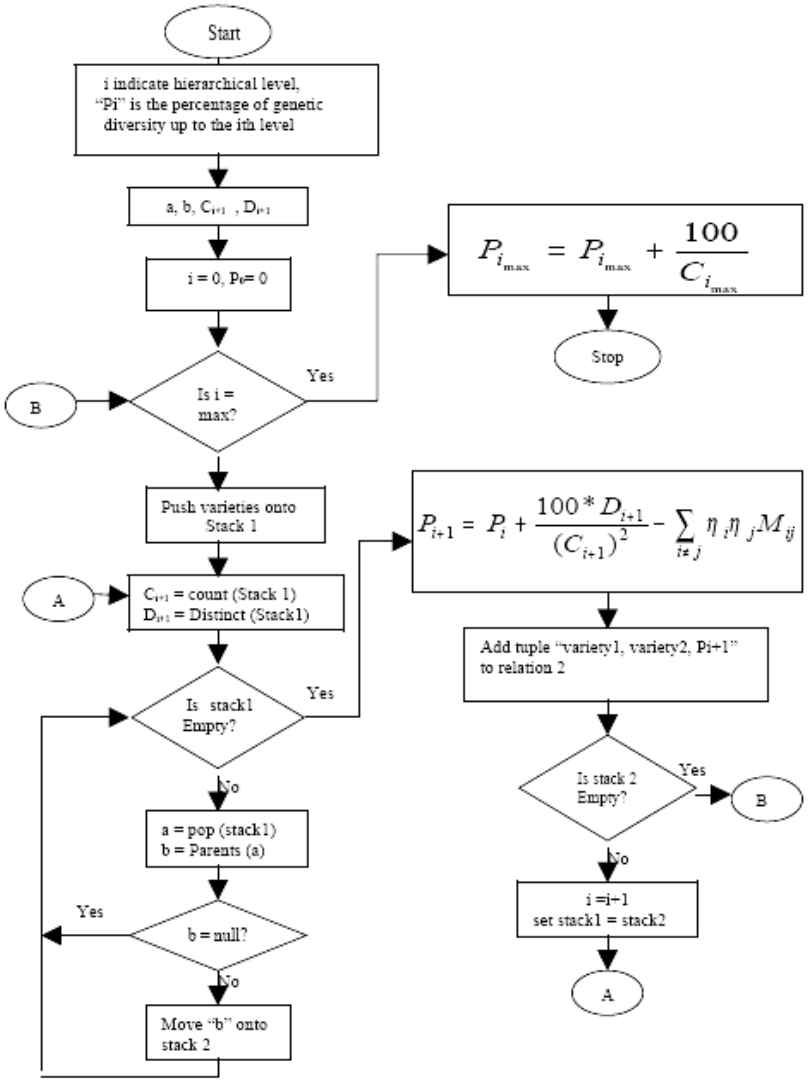

Fig. 1 Predicting Coefficient of Parentage based on Pedigree Information

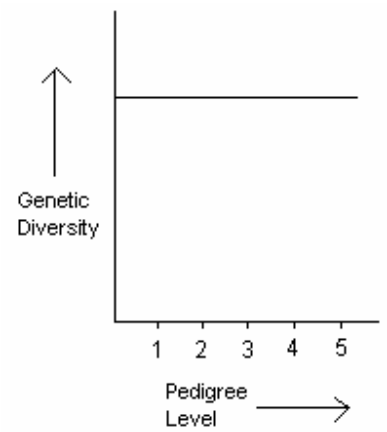

Fig. 2 Graph Showing 100% Genetically Diverse varieties

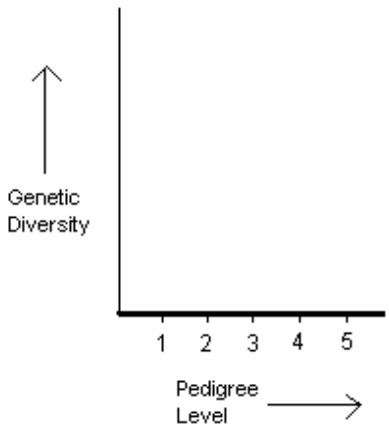

Fig. 3 Graph showing 0 % Genetically Diverse varieties

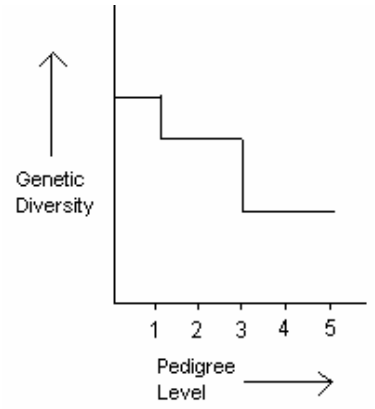

Fig. 4 Graph showing varieties having a random amount of genetic diversity

REFERENCES

[1] Chen, Zhengxin and Zhu, Quiming, "Query construction for user-guided knowledge discovery in databases", Information Sciences 109 (1-4), 1998, pp. 49-64.
[2] Fan, Jianhua and Li, Deyi (1998) "Overview of data mining and knowledge discovery", Journal of Computer Science and Technology. 13 (4), 1998, pp 348-368.
[3] H.K Grewal, Parvinder Singh and Manpreet Singh "A Bioinformatics Approach to Genetic Diversity" IEEE International Conference held at Balauchistan, Pakistan, 2006, pp.
[4] Jagdeep Singh (2002) "Development of Biotechnology Information System using a Web Server", M.Tech Thesis PAU, Ludhiana, 2002, pp. 1-73.
[5] Lee, L.E.J.; Chin, P.; Mosser, D.D. (1998). Biotechnology and the Internet. Biotechnology Advances 16 (5-6). pp 949-960.
[6] Manpreet Singh (2003) Development of Data Mining model for bioinformatics system, M.Tech Thesis PAU, Ludhiana
[7] M.S. chen, J. Han, and P.S. yu "Data mining: an overview from a database perspective", IEEE transaction on knowledge and data engineering, 1996, pp. 866-883.
[8] Raghavan, V. Vijay, Deogun and S. Jitender "Introduction to Data Mining" Journal of the American Society for Information Science 49 (5), 1998, pp. 397-402.
[9] Sanjay Soni, Zhaohui Tang and Jim Yang "Performance Study of Microsoft Data Mining Algorithms" Microsoft White Paper pages 10, 2000.
[10] Stahl, Earl "Employing intelligent agents for knowledge discovery" Proceedings – International Conference on Data Engineering 1998. IEEE Comp Soc, Los Alamitos, CA, USA, 1998, pp. 100-104.

Manpreet Singh received the B.Tech. Electronics & Electrical Communication from Guru Nanak Dev Engineering College, Ludhiana and M.Tech. in Computer Science & Engineering from P. A. U., Ludhiana. He is presently working with Department of CSE & IT, Guru Nanak Dev Engineering College, Ludhiana. His current research interests are Bioinformatics, Distributed Computing and Data Mining. He has published around 20 research papers in various National and International conferences.

Parvinder Singh Sandhu is working as Professor in the Department of Computer Science and Engineering with Rayat & Bahra Institute of Engineeirng & Bio-technology, Sahauran, Mohali and previously he was with Guru Nanak Dev Engineering College, Ludhiana (Punjab). He is Master of Engineering in Software Engineering (Thapar University, Patiala), M.B.A. and Bachelor in Computer Engineering from National Institute of Technology (NIT), Kurukshetra. He has published 18 research papers in referred International journals and 17 papers in renowned international conferences. His current research interests are Software Reusability, Bio-informatics, Software Maintenance and Machine Learning.

Open Science Index, Bioengineering and Life Sciences Vol:1, No:12, 2007 publications.waset.org/8942.pdf