Genetic Diversity in Bioinformatics: A Novel Application of Biology

[1]Pankaj Bhambri, [2]Dr. O.P. Gupta

[1]Department of Information Technology
Guru Nanak Dev Engineering College
Ludhiana, Punjab

[2]Department of Information Technology
Punjab Agriculture University
Ludhiana, Punjab

Abstract

This paper deals with the various search and analysis techniques involving different possible pairs of varieties selected on the basis of morphological characters, Climatic conditions and Nutrients so as to obtain the optimal pair that can produce the required crossbreed variety. An algorithm has been developed to determine the genetic diversity between the selected wheat varieties. Data mining techniques have been used for retrieving the results. Dummy values have been assumed wherever actual data was not available. MATLAB has been deployed to demonstrate the results in a better visual perspective.

Keywords: Genetic Diversity, Morphological characters, Pedigree.

1. Introduction

The research in biotechnology has accumulated substantial data which can not be handled with conventional/classical techniques. With the advanced hardware technology, the cost of storing has decreased thereby leading to an urgent need for new techniques and tools that can intelligently, automatically assist in transferring the data into useful knowledge. Different techniques of data mining are developed which are helpful for handling these large size databases. Data mining is also finding its important role in the field of biotechnology. Pedigree refers to the associated ancestry of a crop variety. Genetic diversity is the variation in the genetic composition of individuals within or among species. Genetic diversity depends upon the pedigree information of the varieties. Parents at lower hierarchic levels have the more weightage for predicting genetic diversity as compared to the upper hierarchic levels. The weightage decreases as the level increases. For crossbreeding, the two varieties should be genetically diverse so as to incorporate the useful characters of the two varieties in the newly developed variety.

2. 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 merge into one 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 extract 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 and effectively. There is a strong interest in employing methods of knowledge based discovery and data mining to generate models of biological systems. Mining biological databases imposes challenges which knowledge based 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.

Data mining part of larger process called knowledge based discovery; specifically, the step in which advanced statistical analysis and modeling techniques are applied to data to find useful patterns and relationships. Recent progress in data mining research has led to the development of numerous efficient and scalable methods for mining interesting patterns in large databases.

This project has been undertaken to cover the data mining applications in existing knowledge buried in large biological texts and to infer results from them. Morphological characters are the various parameters related to the wheat varieties as summarized in Table 1. The user can assign any desired morphological characteristics as input obtain a list of varieties at a time to find out the most optimal and probable pair of genetically diverse varieties. The result can be shown graphically, depicting the genetic diversity among the varieties based on the pedigree levels. The program is user friendly and the percentage probability of getting the required hybrid breed is the output.

3. 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 and nutrients. This information is important to develop a variety with particular useful characters. Now it is required to determine the genetic diversity between the varieties. For this, the user needs to compare the pedigree information (parentage) of the varieties.

The selected varieties they formulated in 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 Figure 1. The formula used for calculating the results is given below:

$$P_{i+1} = P_i \quad + \quad \left(\frac{\eta_i + 50/L_{i+1}}{C_{i+1}}\right) D_{i+1}$$

Where,

P_i is the percentage probability for the varieties to be genetically diverse upto ith level.

η_i is constant for pedigree level 'i' indicating the effect of that level on the genetic diversity. L_i indicates the hierarchical level under observation. C_i and D_i correspond to the number of varieties in a level and the number of distinct varieties in a level. So higher the value of $\mathbf{P_i}$, greater will be the genetic diversity between the crop varieties.

MATLAB (MATrix LABoratory) has been utilized to plot a graph between pedigree levels and genetic diversity utilizing the formula for P_i. Genetically diverse varieties are represented with the help of a straight horizontal line as shown in Figure 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 Figure 3. The most commonly obtained graph is shown in Figure 4. The downward step indicates the similarity of some parents of the two varieties at that particular level.

4. Conclusion

The model was developed has incorporated knowledge discovery from large databases in the field of bioinformatics. The project has focused mainly on the design to determine the most optimal and probable parent varieties for a desired crossbreed wheat variety.

In addition to the above module, the database is provided for the morphological characters, climate conditions and nutrients for the given varieties. As India is agriculture oriented country and the Punjab State depends on the wheat crop, the project inherently reflects substantial potential for furthering the project.

5. References:

[1] Chen, Zhengxin and Zhu, Quiming (1998) Query construction for user-guided knowledge discovery in databases. Information Sciences 109 (1-4) pp 49-64.

[2] Jagdeep Singh (2002) Development of Biotechnology Information System using a Web Server. M.Tech Thesis PAU, Ludhiana.

[3] Lee, L.E.J.; Chin, P.; Mosser, D.D. (1998). Biotechnology and the Internet. Biotechnology Advances 16 (5-6). pp 949-960.

[4] Manpreet Singh (2003) Development of Data Mining model for bioinformatics system, M.Tech Thesis PAU, Ludhiana

[5] Pongor-S; Landsman-D (1999) Bioinformatics and the developing world. Biotechnology-and-Development-Monitor. No. 40, pp10-13.

[6] Sanjay Soni; Zhaohui Tang; Jim Yang (2000) Performance Study of Microsoft Data Mining Algorithms. Microsoft White Paper pages 10.

[7] Stahl, Earl (1998) Employing intelligent agents for knowledge discovery. Proceedings – International Conference on Data Engineering 1998. IEEE Comp Soc, Los Alamitos, CA, USA. pp 104.

[8]

Table 1: Morphological Characters

Variety	DOF	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

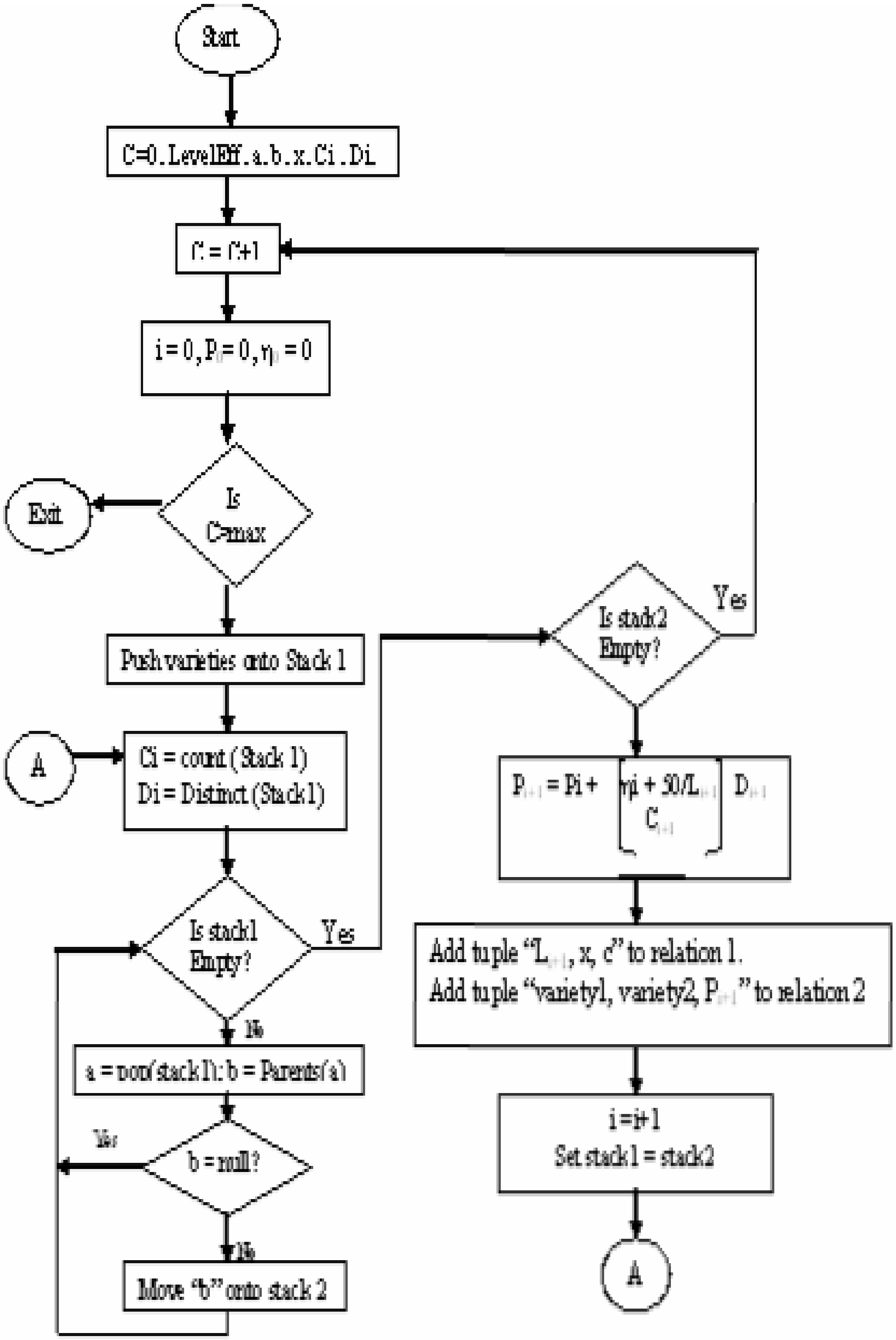
Start
C=0, LevelEff, a, b, x, Ci, Di
C = C+1
i = 0, P0 = 0, η0 = 0
Is C>max
Exit
Push varieties onto Stack 1
A
Ci = count (Stack 1)
Di = Distinct (Stack1)
Is stack1 Empty?
Yes
No
a = pop(stack1); b = Parents(a)
Yes
b = null?
No
Move "b" onto stack 2
Is stack2 Empty?
Yes
Pi+1 = Pi + [ηi + 50/Li+1 Di+1 / Ci+1]
Add tuple "Li+1, x, c" to relation 1.
Add tuple "variety1, variety2, Pi+1" to relation 2
i = i+1
Set stack1 = stack2
A

Fig 1: Predicting Genetic Diversity based on pedigree information

DOF: Days of Flowering, TG: Tiller Grain.

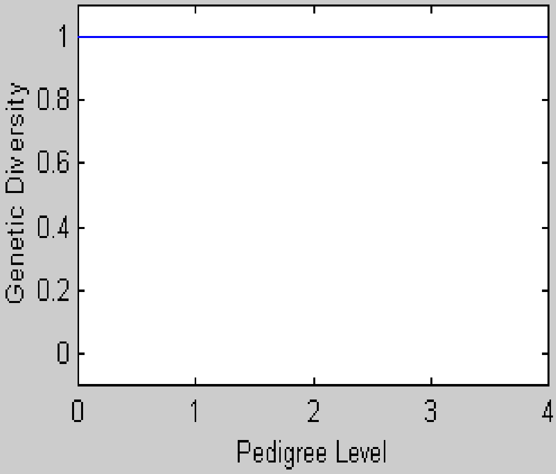

Figure 2:Graph showing Genetically diverse varities

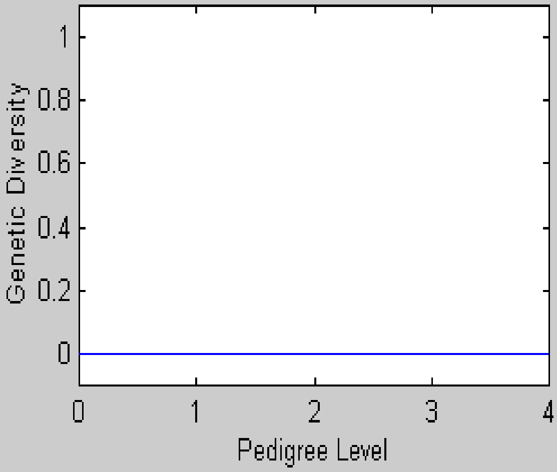

Figure 3:Graph showing Genetically similiar varities

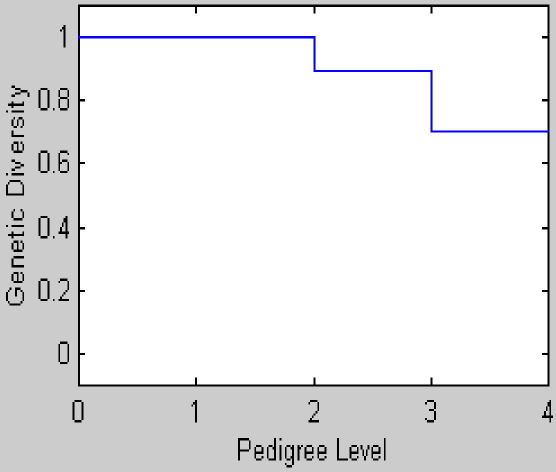

Figure 4: Graph showing varities that are not fully Genetically Diverse

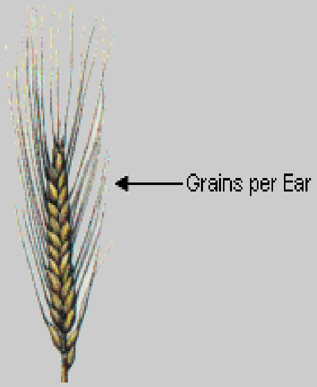

Figure 5: Picture of Wheat Ear