International Journal of Approximate Reasoning 55 (2014) 74–83

Contents lists available at SciVerse ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

A tool based on Bayesian networks for supporting geneticists in plant improvement by controlled pollination

Jens D. Nielsen [a], Antonio Salmerón [b,*], José A. Gámez [c]

[a] CLCbio, Finlandsgade 10-12, 8200 Aarhus N, Denmark
[b] Dept. of Mathematics, University of Almería, La Cañada de San Urbano s/n, 04120 Almería, Spain
[c] Dept. of Computing Systems, University of Castilla-La Mancha, Campus Universitario s/n, 02071 Albacete, Spain

ARTICLE INFO

Article history:
Available online 2 April 2013

Keywords:
Bayesian networks
Inference
Learning
Vegetal genetic improvement
Decision support systems

ABSTRACT

In this paper we describe a system designed for assisting geneticists in vegetal genetic improvement tasks. The system is based on the use of Bayesian networks. It has been developed under the industrial demands emerging from the area of *Campo de Dalías* in Almería (Spain), and is therefore oriented to producing new tomato varieties, which constitute the main product in the area. The paper concentrates on the main aspects of the design of the system.

1. Introduction

The area of *Campo de Dalías* (Almería, Spain) hosts the highest concentration of greenhouses in the world. Besides the extension of the cultivated area, its privileged weather conditions and the high technological level reached has made this region the main supplier of fresh vegetables during almost all the time in the year. Along with the agricultural development, a strong industry has emerged, involving all the auxiliary services connected to agriculture. A good example of that is the industry devoted to seeds production, when both international and local firms have established themselves in the area.

One of the central activities of seeds firms is the development of new varieties that show specially appealing features, like resistance to some plagues, special shape or taste, etc. This is achieved through *vegetal genetic improvement*.

Vegetal genetic improvement is based on various concepts and methodologies involving different areas like Genetics, Biology, Computer Science and Statistics [1]. Its main goal is to improve the adaptation of crops to the environmental conditions and attain the high standards in terms of quality imposed by consumers and producers. Phenotypic characters of commercial importance, like plant height or maturation time, are influenced by genetic inheritance but also by the environmental conditions [2].

Geneticists usually play the role of determining the impact of both factors (genetic and environmental) on the phenotypic variability of some character of interest. Out of this analysis, selected crossings are carried out in order to pursue improved hybrid varieties. Testing the success of such breedings is a costly operation, both in terms of money and time, what makes the use of a decision support system specially appropriate when selecting the hybridisations to be actually carried out, so that only those with high chances of success are chosen. In this sense, making use of any available information, including germplasm databases, is a crucial issue [3].

Bayesian networks [4,5] provide an appropriate framework for representing probabilistic knowledge about complex problems, as the interactions among the different variables are encoded by the network structure, and inference can be carried out in an efficient way taking advantage of the represented interactions. In this paper we describe a system for assisting geneticists in deciding trial crossings that maximise the probability of showing some desired phenotypic character.

* Corresponding author.
E-mail addresses: jnielsen@clcbio.com (J.D. Nielsen), antonio.salmeron@ual.es (A. Salmerón), jose.gamez@uclm.es (J.A. Gámez).

0888-613X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ijar.2013.03.010

It is a general tool aimed at working with any database provided by the final user. The typical final user is a geneticist that works for a seed company that owns a database gathering information about crossings that have already been tested by the company. Such kind of data is extremely valuable and seed companies coin them as one of their most important assets.

The software we describe helps to combine the knowledge of the geneticists (for instance, probability of inheriting genetic characters) with the knowledge contained in the database. The result of the combination is a Bayesian network that is used by the system to suggest crossings with high probability of producing the desired results, among all the possible crossings that can be carried out between pairs of seeds available to the company using the software. The system is implemented in Java, using some functionality of the ProGraMo API [6].

The rest of the paper is organised as follows. The problem addressed is described in Section 2. In Section 3 we explain how the problem is modelled using Bayesian networks. Section 4 is devoted to the process followed to validate pieces of information provided by the user. The way of obtaining candidate crossings is described in Section 5. An example of the system operation using simulated data is given in Section 6. The paper ends with conclusions in Section 7.

2. The addressed problem

The system described in this paper is oriented to assists geneticists in obtaining new tomato varieties that meet some fixed restriction, connected to some phenotypic character. For instance, a possible restriction could be to obtain a variety of tomatoes such that the plants reach a given height, so that the volume in use inside a greenhouse is maximised. The system is able to process that kind of restrictions as an input, and then provides suggestions on tentative breedings highly likely to produce a variety meeting such restrictions.

The importance of obtaining accurate suggestions is evidenced by the process in which breedings are obtained:

1. Male and female parent plants are selected.
2. When selected plants flower, pollen is extracted from the male plant and the female plant is pollinated.
3. After the pollinated plant produces tomatoes, the success of the hybridisation is evaluated.

This is a costly and time consuming process, as a single plant is not enough to validate the success of the operation, and therefore a full line of plants is used as a test, and of course growing a plant takes time.

The system we describe here is able to work with germplasm databases containing information about the seeds of plants available for hybridisation. The main goal can be formally stated as follows. Let $\mathbf{C}$ be the variables representing the properties of a child plant, $\mathbf{P}$ be the variables representing the properties of the parent plants, and $\mathbf{e}$ be a joint configuration of variables $\mathbf{E} \subseteq \mathbf{C}$. Then, the system returns a configuration $\mathbf{p}^*$ for variables $\mathbf{P}$ such that

$$\mathbf{p}^* = \arg\max_{\mathbf{p}\in\mathbf{U}} P(\mathbf{P} = \mathbf{p}|\mathbf{E} = \mathbf{e}) . \tag{1}$$

In Eq. (1), $\mathbf{U}$ can be either the set of all possible joint configurations of the variables in $\mathbf{P}$, or all configurations existing in the seed bank.

3. Modelling the problem with BNs: learning from data

As mentioned in Section 1, to approach the problem addressed in this paper, we chose the Bayesian networks formalism [4,5] because of its capabilities for knowledge representation, uncertainty modelling and decision making.

A Bayesian network (BN) is a mathematical object composed by two different parts which respectively accounts for the qualitative and quantitative parts of the model. The qualitative part of the model is represented by a directed acyclic graph (G) whose nodes represent the random variables in the problem domain and whose edges codify relevance relations (frequently cause-effect) between the variables they connect. The whole graph encodes the (in)dependence relations among the variables, and can be used for qualitative or relevance analysis (e.g. reading independence sentences, identifying information flows, etc.). The quantitative part of the model is a set of conditional probability distributions (CPDs). There is a CPD (usually a conditional probability table or CPT), $P(X_i|pa_G(X_i))$, associated with each node X_i, $pa_G(X_i)$ being the set of parents of X_i in the graph G. Because of the independencies codified in the graph, the joint probability distribution (JPD) defined over the variables in the problem domain, $P(X_1, \ldots, X_n)$, factorises as $\prod_{i=1}^{n} P(X_i|pa(X_i))$.

Once a BN has been constructed for a given problem domain, it becomes a powerful tool for probabilistic and/or evidential reasoning. However, constructing a BN is not an easy task. We can try to build it by hand with the help of experts by following a knowledge engineering process [7, Part III][8, Part II], or we can try to automatically learn it from data [9][10, Part III], or to combine these two options [11–13].

In our case, we adopt the hybrid approach. More precisely, the tool we present here facilitates to incorporate expert knowledge into the machine learning process. Below, we describe the process followed to construct a prototype of the model: first, we describe the type of data available to afford the task, then we propose the type of structure to be learnt, and finally, we describe how to integrate some expert knowledge into the network construction process.