ARTIFICIAL ANALYSIS OF MOLECULAR MARKER LOCI LINKED TO TREE RESISTANCE RESPONSE BY AN ARTIFICIAL NEURAL NETWORK

Jorge Fernández, Angel Castellanos, Juan Castellanos

Abstract: *Citrus tristeza virus (CTV) is one of the most important pathogen affecting citrus and no CTV resistant scion varieties are available. Since Chandler pummelo was found to be CTV resistant, this variety is being used as a donor of CTV resistance given that incorporation of resistance genes into commercial varieties will definitely offer an ultimate solution to CTV problem. To facilitate this breeding program through marker assisted selection (MAS), the analysis of the percentage of influence of molecular marker loci linked to CTV resistance response was done using an artificial neural network (ANN) model. Three main molecular marker loci associated with the Chandler pummelo resistant response to CTV inoculation were detected, allowing the MAS and decreasing the economic cost of breeding programs. Two of those molecular marker loci located at the same genomic region were the Poncirus trifoliata dominant gene responsible for its CTV resistance is located, supporting the theory of a common disease resistance gene cluster between those species that may supply a resource for P. trifoliata and citrus resistance to different pathogens including CTV.*

Keywords: *Citrus tristeza virus, tree resistance to virus, marker loci, neural network*

Introduction

Citrus trees are economically the most important fruit crop, with an annual production exceeding 149 million tons in 2012 [FAOSTAT, 2015]. The main cultivated species are sweet oranges (*Citrus sinensis* (L.) Osb.), mandarins (mainly *C. clementina* Hort. Ex Tan. and *C. unshiu* (Mak.) Marc.), grapefruits (*C. paradise* Macf.), pummelos (*C. grandis* (L.) Osb.) and lemons (*C. limon* L. Burm. f.). Cultivars of all these species are always vegetatively propagated by bud-grafting onto a seedling rootstock in order to obtain a more uniform and early yielding tree with tolerance to *Phytophthora*, parasitic nematodes and some viruses, and well adapted to the local edaphoclimatic conditions. Sour orange (*C. aurantium* L.) was the most common rootstock before 1930, while rough lemon (*C. jambhiri* Lush.) and trifoliate orange (*Poncirus trifoliate* (L.) Raf.) were used in areas where sour orange performed poorly, such as Australia and South Africa, because of tristeza disease (caused by Citrus tristeza virus, CTV). Therefore, two kinds of citrus breeding populations, groups of species, and target traits, are managed for rootstock and scion improvement.

CTV is a member of the genus Closterovirus, family Closteroviridae, and one of the most important pathogens affecting citrus. It has a genome of 19.2 kb, which is the largest among RNA plant viruses. CTV probably originated in Asia, which is also the centre of origin of Citrus, and has been disseminated to many countries by movement of infected plant material. Subsequent natural spread by aphid vectors has created major epidemics. CTV dispersal to other regions and its interaction with new scion varieties and rootstock combinations resulted in three distinct syndromes named tristeza, stem pitting and seedling yellows. The first, inciting decline of varieties propagated on sour orange, has forced the rebuilding of many citrus industries using tristeza-tolerant rootstocks. The second, inducing stunting, stem pitting and low bearing of some varieties, causes economic losses in an increasing number of countries. The third is usually observed by biological indexing, but rarely in the field. It was estimated that almost 100 million citrus trees had been killed by CTV [Moreno et al, 2008]. In the absence of exclusion of infection, there are no satisfactory management strategies against severe CTV-induced diseases [Bar-Joseph et al, 1989].

One of the most effective general means of managing plant diseases has been through the use of resistant varieties, but most citrus species are hosts of CTV. Citrus genetic resources are rich but underutilized in breeding because their complex reproductive biology and the scarceness of inheritance studies on agronomic traits. Up to now, the citrus and related genotypes where CTV resistance has been found are: trifoliate orange [Yoshida et al, 1983], the Meiwa kumquat (*Fortunella crassifolia*) [Mestre et al, 1997b] and the pummelo “Chandler” (*C. grandis*) [Garnsey et al, 1996]. All cultivars of *P. trifoliate* tested have been found resistant to most CTV isolates, making it the specie of choice as donor of CTV resistance in all breeding programs for citrus rootstocks. Nevertheless, breeding a marketable CTV-resistant scion cultivar by crossing trifoliate orange with citrus has not been possible because undesirable traits of trifoliate orange remain after several generations of backcrossing with citrus. Then, breeding strategies based on transgenic technology directed to introduce the dominant gene responsible for the CTV resistance of *P. trifoliata* in to scion varieties, and the obtaining of pathogen-derived resistance by plant transformation are presented as interesting open lines of research. Otherwise, the resistance reported in some pummel cultivar such as Chandler opens a possible way to breed CTV-resistant citrus scion cultivars by sexual hybridization, but the genetic control of such resistance has hardly been studied [Fang & Rose, 1999].

Citrus breeding takes a long time due to the long juvenility period of these species, and is very expensive because of the long time needed and the huge cultivation costs for maintaining and evaluating large segregating progenies. To overcome such limitations, marker assisted selection (MAS) within the progenies is a valuable tool. A first step towards obtaining those tools has been the genetic dissection and mapping of the resistance gene(s). Previous studies have reported quantitative trait loci (QTL) controlling CTV resistance in Chandler pummelo cultivars [Asins et al, 2012].

Computational tools such as artificial neural networks (ANNs) represent a new approach hardly employed in genetic studies. The attractiveness of ANNs comes from their remarkable information processing characteristics pertinent mainly to non linearity, high parallelism, fault and noise tolerance, and learning and generalization capabilities [Basheer & Hajmeer, 2000]. Neural networks [Anderson & Rosenfeld, 1988] are non-linear systems whose structure is based on principles observed in biological neuronal systems [Hanson & Burr, 1990]. Neural networks can predict any continuous relationship between inputs and the target. Then, a neural network could be seen as a system that can be able to answer a query or give an output as answer to a specific input. Similar to linear or non-linear regression, ANNs develop a gain term that allows prediction of target variables for a given set of input variables. Physical–chemical relationships between input variables and target variables mayor may not built in to the association of target and input variables. The in/out combination, i.e. the transfer function of the network is not programmed, but obtained through a training process on empiric datasets. In practice the network learns the function that links input together with output by processing correct input/output couples. Actually, for each given input, within the learning process, the network gives a certain output that is not exactly the desired output, so the training algorithm modifies some parameters of the network in the desired direction. Hence, every time an example is input, the algorithm adjusts its network parameters to the optimal values for the given solution: in this way the algorithm tries to reach the best solution for all the examples. These parameters we are speaking about are essentially the weights or linking factors between each neuron that forms our network.

Calibrating a neural network means to determinate the parameters of the connections (synapses) through the training process. Once calibrated there is needed to test the network efficiency with known datasets, which has not been used in the learning process. There is a great number of Neural Networks [Anderson, 1995] which are substantially distinguished by: type of use, learning model (supervised/non-supervised), earning algorithm, architecture, etc. Multilayer perceptrons (MLPs) are layered feed forward networks typically trained with static back propagation. These networks have found their way in to countless applications requiring static pattern classification. Their main advantage is that they are easy to use, and that they can approximate any input-output map. In principle, back propagation provides a way to train networks with any number of hidden units arranged in any number of layers. In fact, the network does not have to be organized in layers, any pattern of connectivity that permits a partial ordering of the nodes from input to output is allowed. In other words, there must be a way to order the units such that all connections go from earlier (closer to the input) to later ones (closer to the output). This is equivalent to stating that their connection pattern must not contain any cycles. Networks that respect this constraint are called feed forward networks; their connection pattern forms a directed acyclic graphordag.

The objective of our study was to appraise the percentage of influence of eleven molecular marker loci linked to Chandler pummelo cultivar CTV resistance response using an ANN.

Materials and methods

Plant materials

The segregating population of 201 *C. grandis* x *C. clementina* full-sib hybrids derived from a cross between Chandler (Ch) and Fortune (F) commercial varieties, was employed. Fortune is a hybrid mandarin derived from the cross between *C. clementina* Hort. ex Tan. and *C. tangerine* Hort. ex Tan. Chandler is a hybrid pummelo derived from a cross between two accessions of *C. grandis* (L) Osbeck. Genetic linkage maps of these species were previously built up using this segregating population [Bernet et al, 2010]. Quantitative trait locus (QTL) analysis of accumulation and distribution of CTV was also previously carried out with this segregating population [Asins et al, 2012].

To evaluate these 201 hybrids for CTV resistance, they were propagated on sweet orange rootstocks, given that CTV replicates and accumulates abundantly in sweet orange. Every plant was grown in a separate container, in the same greenhouse (25 ± 10ºC). Each propagation was inoculated at the rootstock by grafting two patches of infected sweet orange with CTV isolate T-346, a common Spanish isolate, kept at the bank of CTV isolates at Instituto Valenciano de Investigaciones Agrarias (IVIA) [Ballester-Olmos et al, 1993].

Evaluation of CTV accumulation and distribution was done as described in [Asins et al, 2012]. CTV was monitored and its titer evaluated in each tree by two recommended serological ELISA methods [EPPO, 2004] using specific monoclonal antibodies 3CA5 and 3DFI together, as described in [Cambra et al, 1993]. A plant was declared resistant when CTV was detected at the original inoculum but not at any branch, and its DAS-ELISA (Double Antibody Sandwich ELISA) values through years were similar to those of un-inoculated plants (negative control). Those hybrids where the virus was detected by both serological methods were considered susceptible.

Molecular markers

Eleven of the molecular markers loci analyzed in the segregating population were selected, because of their linkage with the resistance response to CTV inoculation. Simple Sequence Repeats (SSRs), Sequence Characterized Amplified Regions (SCARs), Inter-Retrotransposon Amplified Polimorphism (IRAPs) and one resistance gene analogue were included. Four of those markers were common between both parents and behaved as codominant. Given that they allowed the unambiguous classification of hybrids into four possible genotypes, were considered independently for the neural networks analysis obtaining a total of 15 molecular marker loci.

Artificial neural networks analysis

We use neural networks models with analysis of sensibility because the process of finding relevant data components is based on the concept of sensitivity analysis applied to trained neural networks. This model predicts more accurately the relationship existing between variables. The suitable way to find the individual effects of forecasting variables (15 molecular marker loci) over the variable to forecast (plant response to CTV inoculation), and the way to find a set of forecasting variables (additional marker loci) to include in the new model generated. We have studied different analysis for detecting relationships between those 15 molecular marker loci analyzed in the segregating population and the response to CTV inoculation. In order to study the relationships between those different variables, neural networks models MLP (multilayer perceptron) with a two hidden layer with 4 axons and a Tanh transfer function were used.

Results

Chandler and 13 of its hybrids (6.47%) were found to be resistant to CTV isolate T-346, since it was not detected after the inoculation during the whole experiment.

After training the network, a study of sensitivity was made obtaining the percentages of influence of those molecular markers loci considered to CTV inoculation response (Table 1).

Table 1. Percentages of influence of the molecular markers loci considered to CTV inoculation response

Chandler			Fortune		
PI	Marker	LG	PI	Marker	LG
16,24	1R	Gr4b	8,98	CAG01	Cl4b
8,6	CAG01	Gr4b	8,93	Py28	Cl4b
8	CMS20	Gr12	6,36	Py65	Cl4b
5,77	CMS48,700	Gr4b	6,17	CR19	Cl2
4,59	C11intCrt100	Gr7	5,94	CK16	Cl4b
3,75	CMS47,160	Gr4a	5,41	1R	Cl4b
			5,03	CMS20	Cl12
			4,36	CMS48	Cl4b
			1,87	AintCrt235	Cl12

PI: percentage of influency, ***LG:*** linkage group

The General performance probe displays the Mean Squared Error (MSE), the Normalized Mean Squared Error (NMSE), the Correlation Coefficient (r), and the Percent Error. See Figures 1 and 2 below:

Figure 1. Table of mistake obtained.

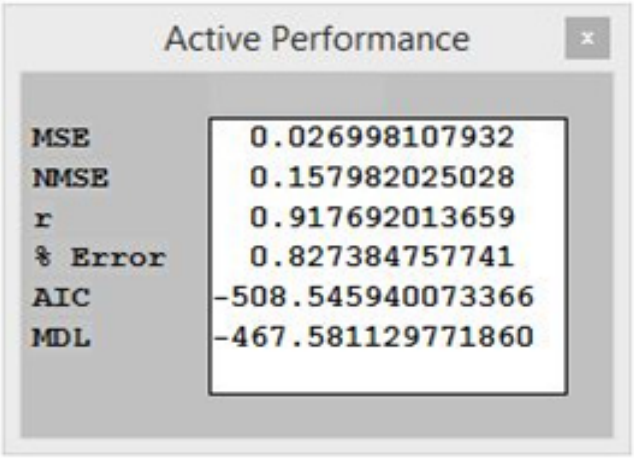
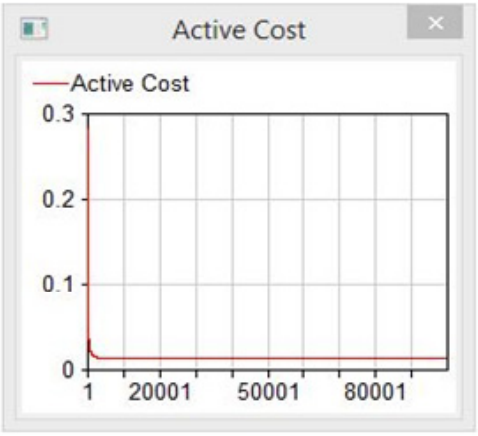


Figure 2. Graph of mistake.



ANN analysis indicates that the main molecular markers loci linked to CTV inoculation response are 1R and CAG01. Both markers loci are located in linkage group 4 of *C. grandis* and *C. clementina* maps. CMS20 marker locus, located on linkage group 12, is also important. This methodology has allowed the detection of two main genomic regions, one of them associated with candidate gene 1R [Bernet et al, 2004], involved in the plant response to CTV inoculation.

Discussion

Pummelo is commercially cultivated in many countries, such as China, Thailand, Japan, Mexico and Israel, because pummelo fruits and juice have abundant antioxidant compounds, including vitamin C, carotenoids, flavonoids and limonoids [Mokbel & Hashinaga, 2006]. Pummelo is also considered a valuable germoplasm for citrus breeding. Since Chandler pummelo was found to be CTV resistant, this variety is being used as a donor of CTV resistance for scion improvement given that incorporation of resistance gene(s) into commercial varieties will definitely offer an ultimate solution to CTV problem. To facilitate this breeding program through MAS, the genetic analysis of CTV resistance was pursued. The ANNs analysis reveal the detection of two main genomic regions, one of them associated with candidate gene 1R [Bernet et al, 2004], involved in the plant response to CTV inoculation. The region with the higher percent of influence located at linkage group 4, in the same genomic region where the major CTV resistance QTL from *P. trifoliata* and *C. aurantium* were previously mapped [Asins et al, 2004]. Our results disagree with those reported by [Fang & Rose, 1999], as they indicate that Chandler CTV resistance was controlled by a single dominant gene not allelic with *Ctv* (P. trifoliata gene). Sequence analyses of the genomic region were the *P. trifoliate* dominant gene responsible for its CTV resistance is located reveal the presence of a disease resistance-gene cluster including at least five functional R genes [Yang et al, 2003]. On the other hand, comparative analysis of that genomic region sequence of Poncirus, C. grandis, C. Clementine and C. sinensis detected the presence of a diverse group of retrotransposable elements (REs) suggesting that their activity has led to the considerable variation in

localization and resistance-gene copy number between those species [Rawat et al, 2015]. Additionally, hypothetical chromosomal rearrangements affecting this genomic region, as those reported in the evolution of this group of species [Raghuvanshi, 1962] can also explain those variations observed. The clustering of disease resistance genes is a common occurrence in plant genomes [Michelmore & Meyers, 1998]. This disease resistance gene cluster may supply a resource for *P. trifoliata* and citrus resistance to different pathogens including CTV. Therefore, further genetic and physical mapping of that genomic region would provide important information on the evolution and function of disease-resistance genes in *Poncirus* and *Citrus*.

This is the first time CTV resistance has been genetically analyzed by an ANN system, a form of machine learning from the field of artificial intelligence utilized in many areas of bioinformatics, biotechnology and medicine [Basheer & Hajmeer, 2000]. Previous studies have reported QTL controlling CTV resistance. However, before embarking upon a QTL analysis, two conditions need to be fulfilled to ensure the data are suitable [Gupta, 2002]:

1. The molecular marker whose association with the trait of interest is being examined can't exhibit any segregation distortion since it may lead to biased estimate of marker-trait association.
2. The phenotypic data on the quantitative trait should show a normal distribution among the segregating population, and in case normality is not present, the data need to be transformed on a scale that will achieve normality of distribution.

On the other hand, ANNs have many advantages in their ability to derive meaning from large complex datasets. First, they do not rely on data to be normally distributed, an assumption of classical parametric analysis methods. They are able to process highly dimensional datasets, data containing complex (non-linear) relationships and interactions that are often too difficult or complex to interpret by conventional linear methods. Another advantage is that they are fault tolerant they have the ability of handling noisy or fuzzy information, whilst also being able to endure data which is incomplete or contains missing values. Nevertheless, before embarking upon ANN analysis care needs to be exercised as the addition of a given variable into a forecasting model does not implies that this variable will have an important effect over the response of the model. That is, if a researcher identifies a set of forecasting variables, he must check if they really affect the response. A frequent problem is that some of the forecasting variables are correlated. If the correlation is small, then consequences will be less important. However, if there is a high correlation between two or more forecasting variables, then the model results will be ambiguous but not for obtain a bad prediction, the problem is the high correlation between variables (high lineal association) decrease in a drastic way the individual effect over the response for each correlation variable and sometimes is difficult to detect and is not possible measure the real effect for each variable over the output.

Conclusion

Here we report the two main genomic regions involved in the Chandler Pummelo CTV resistant response detected by an ANN analysis. Those results allow the molecular marker assisted selection in citrus breeding programs based on the sexual hybridization of Chandler Pummelo with commercial varieties, in order to obtain CTV resistant scion cultivars.

Bibliography

[Anderson & Rosenfeld, 1988] Anderson JA, Rosenfeld E. Neurocomputing: Fundations of research. The MIT Press. 1988.

[Anderson, 1995] Anderson JA. An introduction to neural networks. The MIT Press. 1995.

[Asins et al, 2004] Asins MJ, Bernet GP, Ruiz C, Cambra M, Guerri J, Carbonell EA. QTL analysis of Citrus Tristeza Virus-citradia interaction. Theor Appl Genet. 2004, 108:603-611.

[Asins et al, 2012] Asins MJ, Fernández-Ribacoba J, Bernet GP, Gadea J, Cambra M, Gorris MT, Carbonell EA. The position of the major QTL for Citrus tristeza virus resistance is conserved among Citrus grandis, C. aurantium and Poncirus trifoliata. Mol Breeding 2012, 29:575–587.

[Ballester-Olmos et al, 1993] Ballester-Olmos JF, Pina JA, Carbonell EA, Moreno P, Hermoso de Mendoza A, Cambra M, Navarro L. Biological diversity of citrus tristeza virus (CTV) isolates in Spain. Plant Pathology 1993, 42:219-229.

[Bar-Joseph et al,1989] Bar-Joseph M, Marcus R, Lee RF. The continuous challenge of Citrus Tristeza Virus control. Annu Rev Phytopatol 1989, 27:291-316.

[Basheer & Hajmeer, 2000] Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. J Microbiol Methods 2000, 43:3–31.

[Bernet et al, 2004] Bernet GP, Bretó MP, Asins MJ. Expressed sequence enrichment for candidate gene analysis of Citrus Tristeza Virus resistance. Theor Appl Genet 2004, 108:592-602.

[Bernet et al, 2010] Bernet GP, Fernández-Ribacoba J, Carbonell EA, Asins MJ. Comparative genome-wide segregation analysis and map construction using a reciprocal cross design to facilitate citrus germplasm utilization. Molecular Breeding 2010, 25:659-673.

[Cambra et al, 1993] Cambra M, Camarasa E, Gorris MT, Garnsey SM, Gumpf DJ, Tsai MC. Epitope diversity of citrus tristeza virus isolates in Spain. Proc 12th Conf IOCV 1993, Riverside, California, USA. pp 33-38.

[EPPO, 2004] Standards PM 7/31 diagnostic protocol for citrus tristeza closterovirus. OEPP/EPPO Bull 2004, 34:155–157

[Fang & Rose, 1999] Fang DQ, Roose ML. A novel gene conferring Citrus Tristeza Virus Resistance in Citrus maxima (Burm.) Merril. HortScience 1999, 34:334-335.

[FAOSTAT, 2015] FAOSTAT http://faostat.fao.org/site/567/DesktopDefault.aspx Cited 4 May 2015.

[Garnsey et al, 1996] Garnsey SM, Su HJ, Tsai MC. Differential susceptibility of Pummelo and Swingle citrumelo to isolates of citrus tristeza virus. Thirteenth IOCV Conference 1996, 138–146.

[Gupta, 2002] Gupta PK. Molecular marker and QTL analysis in crop plants. Current Science 2002, 83:113-114.

[Hanson & Burr, 1990]. Hanson SJ, Burr DJ. What connectionist models learn: Learning and representation in connectionist networks. Behavioral and Brain sciences 1990, 13:471-518.

[Mestre et al, 1997b] Mestre PF, Asins MJ, Pina JA, Navarro L. Efficient search for new resistant genotypes to the citrus tristeza closterovirus in the orange subfamily Aurantioideae. Theor Appl Genet 1997b, 95:1282-1288.

[Michelmore & Meyers, 1998] Michelmore RW, Meyers BC. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res 1998, 8:1113–1130.

[Mokbel & Hashinaga, 2006] Mokbel MS, Hashinaga F. Evaluation of the antioxidant activity of extracts from buntan (Citrus grandis Osbeck) fruit tissues. Food Chemistry 2006, 94:529-534.

[Moreno et al, 2008] Moreno P, Ambros S, Biach-Marti MR, Guerri J, Peña L. Plant diseases that changed the world – Citrus tristeza virus: a pathogen that changed the course of the citrus industry. Molecular Plant Pathology 2008, 9(2):251-268.

[Raghuvanshi, 1962] Raghuvanshi SS. Cytogenetical studies in the genus Citrus IV. Evolution in genus Citrus. Cytologia 1962, 27:172–188

[Rawat et al, 2015] Rawat N, Deng Z , Gmitter FG. Genomic structure and evolution of the Citrus Tristeza Virus (CTV) resistance locus in Poncirus and Citrus. XXIII International plant and animal genome conference 2015, P0927.

[Yang et al, 2003] Yang ZN, Ye XR, Molina J, Roose ML, Mirkov TE. Sequence Analysis of a 282-Kilobase Region Surrounding the Citrus Tristeza Virus Resistance Gene (Ctv) Locus in Poncirus trifoliata L. Raf. Plant Physiol 2003, 131:482-492.

[Yoshida et al, 1983] Yoshida T, ShichijoT, Ueno I, Kihara T, Yamada Y, Hirai M, Yamada S, Leki H, Kuramoto T. Survey for resistance of citrus cultivars and hybrid seedlings to citrus tristeza virus (CTV). Bull Fruit Tree Res Stn B 1983, 10:51-68.

Authors' Information

Jorge Fernández *– Ph.D. student at faculty of Biological Sciences, Universidad Complutense de Madrid, Ciudad Universitaria, 28040, Madrid, Spain; e-mail: jribacob@gmail.com*

Angel Castellanos *– Applied Mathematics Department. Universidad Politécnica de Madrid, Madrid; Spain; e-mail: angel.castellanos@upm.es*

Major Fields of Scientific Research: Artificial Intelligence, applied mathematics

Juan Castellanos *– Head of Natural Computing Group, Universidad Politécnica de Madrid, Campus de Montegancedo s.n., 28660 Boadilla del Monte, Madrid, Spain; e-mail: jcastellanos@fi.upm.es*

Major Fields of Scientific Research: Natural computing, formal language and automata theory.