Computer tools for spatial analysis of plant genetic resources data: 2. FloraMap

P.G. Jones[1]✉, L. Guarino[2†] and A. Jarvis[2]

[1] Centro Internacional de Agricultura Tropical (CIAT), Apartado Aereo 6713, Cali, Colombia
[2] International Plant Genetic Resources Institute (IPGRI), Americas Regional Office, c/o CIAT, Apartado Aereo 6713, Cali, Colombia
† Present address: Secretariat of the Pacific Community, Private Mail Bag, Suva, Fiji. Email: LuigiG@spc.int

Summary

Computer tools for spatial analysis of plant genetic resources data: 2. FloraMap

FloraMap is a specialized computer program (and associated data) that was developed to map the predicted distribution, or areas of possible climatic adaptation, of organisms in the wild. The climate at the locations where populations of a given taxon were recorded is assumed to be representative of the environmental range of the organism. The mean monthly values of 3 climatic variables at these locations are extracted from a set of interpolated climate grids and used to define a multivariate normal distribution, which is in turn used to calculate a probability surface for the occurrence of the species. The software can be used for a number of different plant genetic resource applications besides predicting species distributions. Some of these are described, along with future plans for development of the software.

Key words: climatic adaptation, geographic distribution, GIS, passport data, spatial analysis, wild species

Résumé

Outils informatiques pour l'analyse spatiale des données sur les ressources phytogénétiques : 2. FloraMap

FloraMap est un programme informatique spécialisé (avec des données associées) qui a été développé pour cartographier la distribution prévue ou les zones d'adaptation climatique possible d'organismes vivant dans la nature. Le climat associé aux endroits où des populations d'un taxon donné sont observées est censé être représentatif de la gamme d'habitats où se rencontre, l'organisme. Les valeurs mensuelles moyennes de trois variables climatiques en ces endroits sont extraites à partir d'un ensemble de grilles climatiques interpolées et utilisées pour définir une distribution multivariée Normal qui est, elle-même, utilisée pour calculer une surface de probabilité de distribution de l'espéce considérée. Le logiciel peut être utilisé dans un certain nombre d'applications relatives aux ressources phytogénétiques autres que la prédiction des distributions des espèces. Certaines de celles-ci sont décrites, de même que les plans futurs; de développement du logiciel.

Resumen

Sistemas de Información Geográficas para investigación en recursos fitogenéticos: 2. FloraMap

FloraMap es un programa especializado (con datos asociados) que ha sido desarrollado para predecir y mapear la distribución, o áreas de posible adaptación climática, de especies silvestres. Se asume que el clima de las localidades donde poblaciones de una determinada especie fueron observados es representativo del rango ambiental del organismo. El programe extrae los promedios mensuales de 3 variables climaticos por estas localidades de un conjunto de superficies climáticas interpoladas y usa estos datos para definir una distribucion Normal multivariada, la cual en su turno se usa para calcular una superficie de probabilidad de encontrar la especie. El programa puede ser usado para diferentes aplicaciones en recursos fitogenéticos además que para pronosticar la distribución de especies. Algunos de éstos son descritos, y tambien los planes futuros para el desarrollo del software.

Introduction

Management of plant genetic resources (PGR) both generates and uses data. The analysis of these data is crucial to the effectiveness of the PGR management process and can add significantly to the value of genetic resources. Many of these data are geo-referenced, i.e. they correspond to specific locations on the earth's surface, which means that they can be analysed—and linkages made between them and other geo-referenced data from sources external to the PGR process—using Geographic Information Systems (GIS). A GIS is a database management system that can simultaneously handle data representing spatial objects and their attribute data. GIS can play an important role in the management of large and complex PGR datasets (Guarino et al. 2001). However, examples of the use of GIS to address PGR problems are still relatively few. Reasons for this probably include the relatively high cost and complexity of much GIS software and the perception that geo-referenced data can be difficult and expensive to obtain.

While increasing familiarity with GIS within the PGR conservation community and advances in information technology will no doubt eventually overcome these problems, PGR workers need the power associated with these methods now. It is with this in mind that CIAT, CIP and IPGRI have been working to develop easy-to-use GIS applications (including software and data) aimed at solving specific PGR problems. This paper introduces one of these, FloraMap, and describes its possible usefulness to PGR conservation and use programmes. Another paper in this series describes the software tool DIVA developed by CIP and IPGRI.

FloraMap (Jones and Gladkov 1999) is a GIS application developed by CIAT (International Centre for Tropical Agriculture, Cali, Colombia) primarily for the prediction of the distribution of organisms in the wild, when little or nothing is known of the physiology of the species involved, so that analytical or simulation models cannot be used. This is important to PGR workers because data on the geographic distribution of species are often scant and it would clearly be useful to be able to use what little is known to identify areas where a species has not been previously recorded, but where it might still be expected to occur. FloraMap assumes that the climate at the points of observation and / or collection of a species (herbarium specimens, germplasm accessions etc.) is representative

of the environmental range of the organism. The climate at these points is used as a calibration set to compute a climate probability model, which is then used to assess the likelihood of other sites being climatically suitable for the species.

Below we briefly describe the methods used in FloraMap, the software interface, its possible uses and future developments. The methods used in FloraMap are described in detail by Jones et al. (1997a), and also in the user manual (Jones and Gladkov 1999).

How FloraMap works

The base climate data

FloraMap includes—and uses in its analyses—climatic data from a 10-minute grid (corresponding to 18 km at the equator). The grids were derived by interpolation from thousands of meteorological stations. A simple algorithm based on the inverse square of the distance between the five nearest stations and the interpolated point is used. Thirty-six climatic variables are used: 12 monthly averages for temperature, rainfall, and diurnal temperature range. Temperature is standardised with elevation using the NOAA TGP-006 (NOAA 1984) digital elevation model (DEM) and a lapse rate model (Jones 1991). Rainfall and diurnal temperature range are independent of elevation.

In order to adjust for geographic differences in the timing of major seasons, a 12-point Fourier transformation is applied (for further information, see Jones et al. 1997a). This is done to allow for a better comparison among climates. For example, the climate of a location in the Northern Hemisphere may be very similar to that of a location in the Southern Hemisphere, except that the rainy season starts in a different month (e.g. July vs January). Six frequencies are fitted to the climate data. The phase angle of the first frequency for the rainfall and temperature records are combined as a function of latitude and the resulting angle is subtracted from each climate record. The Fourier transformation thus rotates the climate data to a common time frame to make them directly comparable.

FloraMap currently has climate grids for Latin America and the Caribbean, Africa, Asia, Europe and the coterminous USA. The European grid uses data courtesy of Dr Ian Makin of IWMI (International Water Management Institute) and the USA grid data from climate grids developed by Dr Chris Daly of Oregon State University.

Computing the climate probability model

The input database file provided by the user should include an accession identifier, latitude and longitude coordinates in decimal degrees and altitude if available. The 36 climate variables described above are first extracted for the pixel in which each accession in the input file is located, and principal components analysis (PCA) is carried out on the resulting data. PCA identifies new variables or dimensions (principal components) related to the original variables. These account for most of the variance in climates among the accession locations in the least number of dimensions. The PCA is performed on the variance-covariance matrix since the Fourier analysis has already transformed the variables to comparable scales. FloraMap then fits a multivariate normal distribution to the principal component scores of the accessions (the user specifies how many dimensions to include).

Mapping and exporting the probability surface

Finally, FloraMap uses the parameters of this multivariate normal distribution to calculate the probability of the climate at each pixel in selected continental coverage(s) belonging to the probability distribution defined by the calibration set, in the space of the dimension determined by the number of principal components chosen by the user. The final result is a probability surface for the continent(s) under consideration. These probability surfaces are ESRI® Shapefiles, with each pixel defined as a square. They can be readily exported to other GIS applications, a feature useful for carrying out further analyses and producing individually crafted output maps.

FloraMap analysis can thus be summarized as follows:

1. Input and checking of accession locality data
2. Extraction of data for 36 climate variables for each accession
3. PCA of the extracted climate data
4. Fitting of multivariate normal distribution to the resulting PCA scores
5. Calculation, for each pixel in the selected continent, of its probability of belonging to this multivariate normal distribution
6. Mapping of the probability surface for the selected continent

The process is managed by the user through a series of distinct but interrelated tools, as described in the next section.

The FloraMap user interface

The FloraMap user interface was developed using C++ and ESRI's MapObject software. It is described below in terms of each of its component tools.

The Map Wndow Tool (Figure 1)

The Map Window is the central tool in FloraMap. It displays the map and has controls for managing the map layers, including probability surfaces and accession points used in the model. The political boundaries of Latin America, Africa and Asia can be displayed as background to the probability layers. Roads, rivers and cities can be overlaid as aids in geographical representation of the study areas. Control of colours and characteristics of all

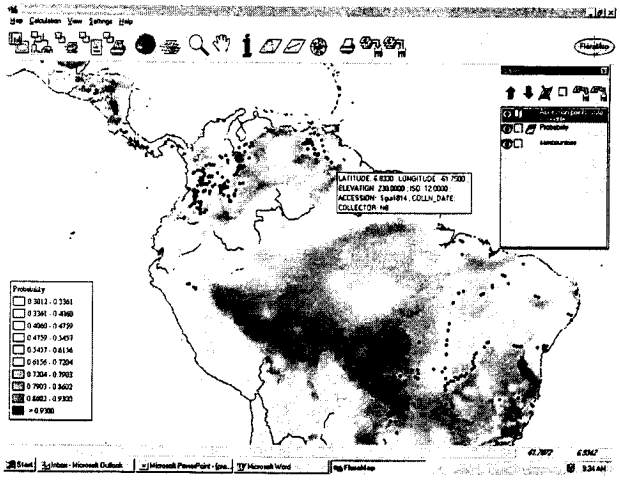


Figure 1. The map window.

layers is available through the Layer Control Window. There are facilities for saving and reloading complete maps, or individual map layers.

The Principal Components Window Tool (Figure 2)

This window controls the principal components analysis. The scaling, weighting and transformation of the variates, choice of components and map probability limits are all controlled by the user, using convenient sliders or by entering the values desired. A scatter diagram indicates the distribution of the accession points in any combination of the principal component dimensions. A powerful tool for selecting sub-sets of the data works by drawing around the points in question. This also works from the Map Window.

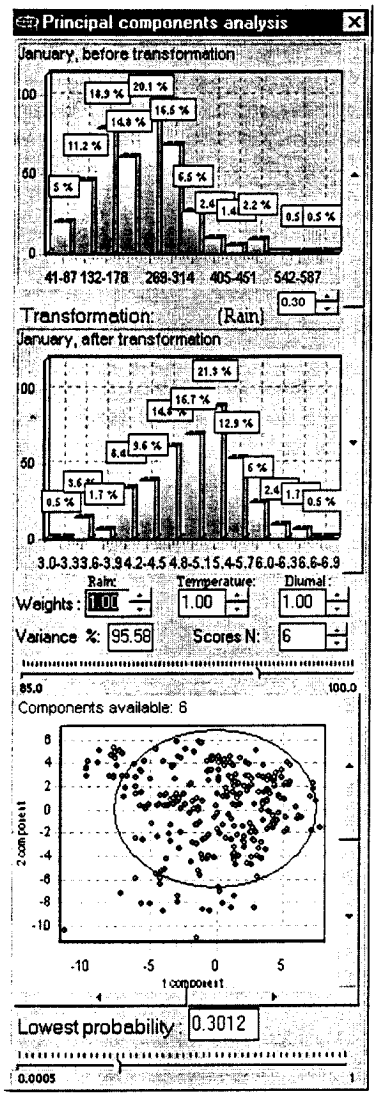


Figure 2. The principal components window.

The Cluster Analysis Tool (Figure 3)

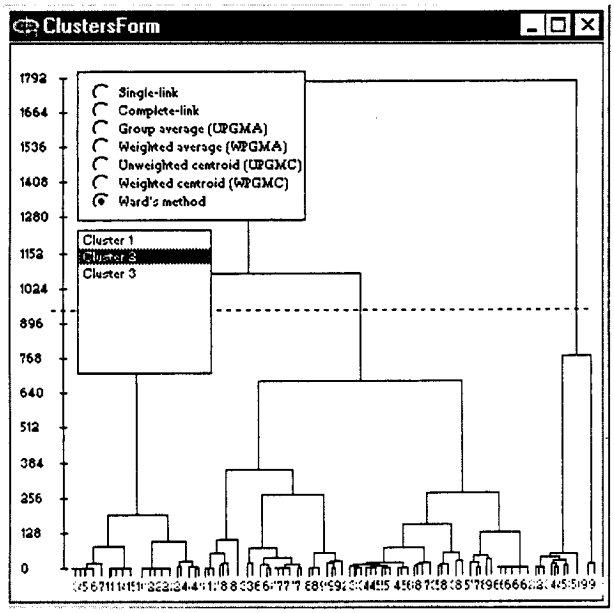
An accession set may be found to contain sub-groups with significantly distinct climatic ranges. In such cases, fitting a model to the data as a whole may give spurious results. The Cluster Analysis Tool gives the user access to seven powerful techniques to investigate the possibility of clustering. When clusters are found, the user can point and select a cluster and calculate the probability model for those accessions only, repeating the procedure for each cluster.

The Climate Diagram Tool (Figure 4)

The user can view the climate of a point from the map, an accession point from the scatter diagram or map and also the mean climate for groups of points from the scatter diagram or the map. The Climate Diagram can be displayed in Cartesian or polar coordinates and can show the data either by real date, or in rotated form.

The varied uses of FloraMap

The primary use of FloraMap is, as the subtitle of the user manual states, predicting the distribution of plants and other organisms in the wild from very limited distribution data. Comparing such predicted distribution with the localities where accessions where collected could, for example, be useful to germplasm collectors trying to identify gaps in existing collections. There have been few attempts to test the accuracy of the predicted distributions, but Jones et al. (1997a) had some success with wild *Phaseolus* in South America.



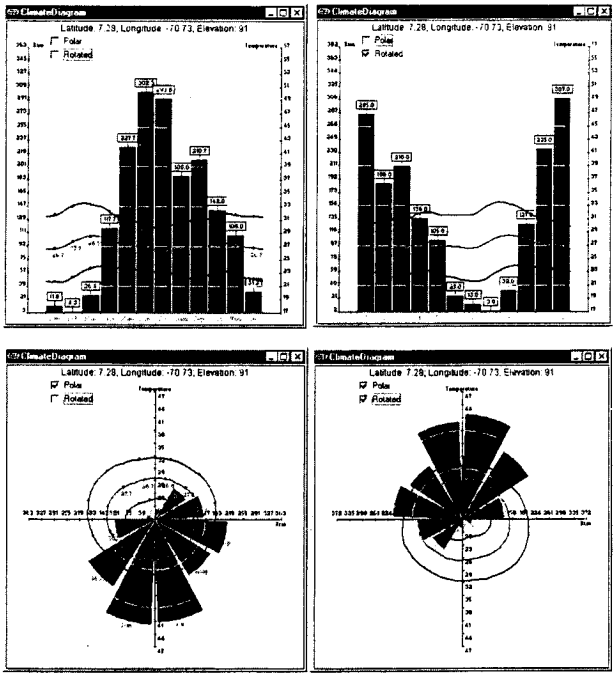


Figure 4. The Climate Diagram.

However, prediction of natural distributions and gaps is not the only use that PGR workers can make of FloraMap. The data that come with it, and the software's analytical facilities, also allow other PGR problems to be tackled, mainly through the 'retro-classification' of accessions. This is the term coined by Steiner and Greene (1996) to describe the process of obtaining data on the environment at a site of collection not in the field, but by overlaying the locations of the sites on different thematic base maps of the appropriate scale and reading off the map attributes, most conveniently by using GIS. Steiner and Greene (1996) give an example of the usefulness of retro-classification using the *Lotus* collection of the US National Plant Germplasm System (NPGS) (see also Greene et al. 1999). The following applications may be highlighted:

- checking data quality
- predicting climatic adaptation in other areas
- identifying groups of accessions with distinct climatic adaptations (ecotypes)
- comparing climatic adaptation among *a priori* groups of accessions.

Checking data quality

The PCA scatter diagram is useful for identifying errors in the data. Any accessions that are obvious outliers in the PCA diagram should be checked for any inaccuracies in the latitude and longitude information. FloraMap also flags accessions whose coordinates cause it to fall in a body of water. If required, it can move them to the nearest point on the coastline.

Predicting climatic adaptation in other areas

The model describing the climatic adaptation of the species need not be applied only to the putative region of its distribution in the wild. FloraMap allows the user to apply the model to other continents, for example. This means that the distribution of species that are promising as forages and are native to Africa, for example, can be used to predict where they might be expected to yield a good response on introduction into South America or Southeast Asia. Initial trial material could be targeted precisely to areas where it would be expected to be best adapted to the climate. Jones et al. (2000) presented a study of cross-continental climatic adaptation in *Desmodium* and Libreros et al. (2000) a study of adaptation of South American fruit species in Mexico.

Identifying groups within collections

We have seen how the clustering tool allows the user to investigate the existence of groups of accessions with distinct climatic adaptations within a collection. When such groups are strongly defined, it is necessary to fit individual probability distributions to the accessions in each group, rather than to the collection as a whole, in order for the model to work correctly. However, the PGR worker may well be interested in the existence of such groupings quite apart from their importance in fitting the probability model, because such groups may be associated with genetic differences of significance for genetic resources conservation or use. In studies of *Stylosanthes* (Jones et al. 1997b) and *Passiflora* (Segura et al. 1998, 1999), for example, climatic clusters were found to differ with regard to specific isozymes. Climatic groupings could be used in developing core collections. Also, material with distinct climatic adaptation within the collection may need to be multiplied, characterized and evaluated at different sites.

Comparing climatic adaptations

Groupings may already have been defined within a collection on the basis of characterization or evaluation data. In such cases, it might be interesting to the PGR worker to compare the climatic adaptation of these *a priori* groups of accessions. For example, FloraMap was used in conjunction with data on a collection of wild *Arachis* species to define and compare the climatic adaptations of different species (Ferguson et al. in preparation). Users interested in adaptation to particular climatic conditions would then know which of the species are most suitable for their purposes.

What FloraMap does NOT do

FloraMap is proving useful for various applications, but users should be fully aware of its limitations. The software uses a particular model to map the potential climatic 'envelopes' where an organism could exist. First, it is worth reiterating that the 'envelope' that is developed is purely a climatic one. FloraMap takes no account of factors such as soil and dispersal mechanism. Though climate is widely considered to be one of the primary determinants of plant distribution, it is clearly not the only one. However, the probability surfaces can be imported into other GIS, where they can by overlaid with other geo-referenced information, for example on soils, natural vegetation, human intervention, physical barriers etc., to refine the analysis.

Second, it should also be remembered that a different model could result in different predictions. Apart from FloraMap, other software also address the problem of predicting species distributions, though following different models, and therefore potentially with different results. These include BIOCLIM[1] (Busby 1991; see also GARP[2], Genetic Algorithm for Rule-set Production, an extension of the BIOCLIM approach) and DOMAIN[3] (Carpenter et al. 1993). CIFOR's DOMAIN software uses map layers of environmental factors such as climate, soil, land use etc., to construct an environmental habitat envelope or 'domain' on the basis of points for the known distribution points of a species. A map is then produced showing the similarity between different areas within the target region with the species' domain. The purpose of BIOCLIM is to find a single rule that identifies all areas with a similar climate to that in which the species is located. To do this, the basic BIOCLIM algorithm finds the climatic range of the points for each climatic variable individually, which contrasts with the FloraMap multivariate (principal components analysis) approach.

An instructive, large-scale example of the application of GIS to the problem of "objective prediction of the full distribution of a species from incomplete point distribution maps, based on its ecological preferences" is the work being carried out by the Royal

1 http://dino.wiz.uni-kassel.de/model_db/mdb/bioclim.html
2 http://kaos.erin.gov.au/general/biodiv_model/ERIN/GARP/home.html
3 http://www.cifor.cgiar.org/domain/index.htm

Botanic Gardens, Kew[4] in the context of Madagascar's Environmental Action Plan.

Finally, FloraMap does not provide a once-and-for-all solution. As more distribution data become available, the results of the analysis can change significantly, in particular if the original accessions were in some way a climatically biased sample of the real distribution of the species. For more detailed examination of geographic biases in accessions datasets see Hijmans et al. (2000).

The FloraMap user community

The Users Group

FloraMap currently has a users group of over 100 individuals throughout the world. To join the group listserver, it is simply necessary to send a message to: listserv@cgiar.org with the following text in the body of the message: subscribe FloraMap <your email address>. The listserver is used to answer user questions, disseminate new information about FloraMap and as a platform for users world-wide to exchange experiences and hints on its use.

The Web site

FloraMap has a Web site at http://www.floramap-ciat.org with information on the software, examples of its use, a growing list of Frequently Asked Questions (FAQ) with their answers, and extra climate grids not supplied with the original software package. The solution to your difficulties with FloraMap could be merely a few mouse clicks away. The software can be ordered from CIAT through the Web site and costs US$100 ($25 for educational purposes).

The future of FloraMap

CIAT is committed to supporting FloraMap as a member of its stable of Climate Applications, and for the foreseeable future will be developing the product in response to user requests and as the algorithms are further refined. One of the present drawbacks to the system is the precision of the climate grids. At 10 minutes (18 km) for Latin America, Africa and Europe, these are far from satisfactory in mountainous areas and territory with broken relief. The aim is to provide grids based on the USGS 1 km Digital Elevation Model ETOPO30 as soon as possible.

This is quite a challenge in two respects. First the sheer effort of compiling the climate data and fitting and checking the grids is daunting. However, there is a much more important hurdle to clear. At present, FloraMap will take a considerable time, and may fail, on all but the largest, fastest, modern PCs when presented with the 2.5 arc minute grid for South East Asia. A 30 arc second (1 km) grid contains 25 times the data and FloraMap will certainly not work for such a large grid at this precision. The solution to this is to restructure the data storage and retrieval for the climate grids. Work has started on this and the ambitious aim has been to increase the speed of FloraMap 10 000 times while reducing storage space needed for the data. This is not going to be achieved quickly, but will hopefully be accomplished for the next release in 2002.

Other enhancements may include better geographic background files, automatic variable transforms, alternative climate models (the user may like to see what the analysis would look like in DOMAIN) and Boolean operations to combine multiple probability layers.

References

Busby JR. 1991. BIOCLIM—a bioclimate prediction system. In: Margules CR, Austin MP, eds. Nature conservation: cost effective biological surveys and data analysis. CSIRO, Melbourne, Australia, pp. 4–68.

Carpenter G, Gillison AN, Winter J. 1993. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. Biodiversity and Conservation 2:667–680.

Greene SL, Hart T, Afonin A. 1999. Using geographic information to acquire wild crop germplasm: II. Post collection analysis. Crop Science 39:843–849.

Guarino L, Jarvis A, Hijmans RJ, Maxted N. 2001. Geographic Information Systems (GIS) and the conservation and use of plant genetic resources. International Conference on Science and Technology for Managing Plant Genetic Diversity in the 21st Century (SAT21), Kuala Lumpur, Malaysia, 12–16 June, 2000.

Hijmans RJ, Garrett KA, Huaman Z, Zhang DP, Schreuder M, Bonierbale M. 2000. Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. Conservation Biology 14(6):1755–1765.

Jones PG. 1991. The CIAT climate database version 3.41. Machine readable dataset. Centro Internacional de Agricultura Tropical, Cali, Colombia.

Jones PG, Gladkov A. 1999. FloraMap Version 1. A computer tool for predicting the distribution of plants and other organisms in the wild. CD-ROM and Manual. Centro Internacional de Agricultura Tropical, Cali, Colombia.

Jones PG, Galwey N, Beebe SE, Tohme J. 1997a. The use of geographical information systems in biodiversity exploration and conservation. Biodiversity and Conservation 6:947–958.

Jones PG, Sawkins MC, Maass BL, Kerridge PC. 1997b. GIS and genetic diversity case studies in *Stylosanthes*. Poster presented at XVIII International Grassland Congress, June 8–19, Winnipeg, Canada.

Jones PG, Segura S, Guarino L, Peters M. 2000. FloraMap: a computer tool for predicting the distribution of plants and other organisms in the wild. Poster presented at the International Conference on Science and Technology for Managing Plant Genetic Diversity in the 21st Century (SAT21), Kuala Lumpur, Malaysia, 12–16 June, 2000.

Libreros DF, Segura S, Guarino L. 2000. El potencial de algunos frutales Suramericanos en México: Una nueva herramienta SIG para una antigua intención. XVIII Congreso de la Sociedad Mexicana de Fitogenética, Irapuato, México, 15–20 Octubre 2000.

NOAA (National Oceanographic and Atmospheric Administration). 1984. TGPO006 D. Computer tape. Boulder, Colorado, USA.

Segura SD, Coppens d'Eeckenbrugge G, Ollitrault P. 1998. Isozyme variation in five species of *Passiflora* subgenus *Tacsonia* and *Passiflora manicata*. In: XLIV Annual Meeting of the InterAmer. Soc. Trop., September 28–October 2 1998, Barquisimeto, Venezuela.

Segura SD, Guarino L, Coppens d'Eeckenbrugge G, Grum M, Ollitrault P. 1999. Mapping the distribution and regions climatically suitable for four species in *Passiflora* subgenus *Tacsonia* (Passifloraceae) and *P. manicata*. Poster presented at II Simposio de Recursos Genéticos para América Latina e Caribe—SIRGEALC, 21–26 November 1999, Brasília, Brazil.

Steiner JJ, Greene SL. 1996. Proposed ecological descriptors and their utility for plant germplasm collections. Crop Science 36:439–451.

[4] http://www.rbgkew.org.uk/herbarium/madagascar/plant_dis.html