[17] The Role of Geographic Analysis in Locating, Understanding, and Using Plant Genetic Diversity

By ANDY JARVIS, SAM YEAMAN, LUIGI GUARINO, and JOE TOHME

Abstract

The genetic structure of an organism is shaped by various factors, many of which vary significantly over space. In this chapter, we provide insight on how studying geographic patterns may contribute to an improved understanding of variability in genetic structure. We first review the theoretical background on how differences in genetic structure may be generated through processes that are inherently variable over space. We then present novices with some basics on how geographic information systems (GIS) may be adopted to study this variation, including advice on software, data, and the type of research questions that might be addressed. The chapter finishes with a brief review of how spatial analysis has contributed to the conservation and use of plant genetic resources, through an understanding of spatial patterns in species distribution and genetic structure. We conclude that spatial variation is a factor often overlooked in genetic studies and one that merits greater consideration. With the advent of functional genomics and improved quantification of adaptive traits, spatial analysis may be key in understanding variation in genetic structure through careful analysis of genotype–environment interactions.

Introduction

A genome arises through the processes of mutation, selection, gene flow, and genetic drift. Differences in how these processes unfold from one place to another result in the development and maintenance of distinct genotypes, population differentiation, and ultimately, in some cases, the emergence of new species with distinct adaptations and distributions. Spatial considerations are, therefore, key to understanding evolution. This chapter outlines a potential role for spatial analysis using geographic information systems (GIS) in understanding spatial variation in biological data. By visualizing and analyzing spatial patterns in genetic and ecological data, GIS can provide a tool for investigating the processes that shape genomes and for conserving and using genetic diversity as effectively and efficiently as possible. Although applying spatial analysis to the conservation and use of agricultural plant genetic diversity is our focus, the principles can be applied to other organisms and different objectives.

METHODS IN ENZYMOLOGY, VOL. 395 0076-6879/05 $35.00

The chapter is aimed at the GIS novice, outlining the tools and potential data sources available, and providing some examples of their use. We start by introducing some theory of how genetic structure may vary over space and how some researchers have analyzed it. We provide some practical information, describing the types of georeferenced biological and nonbiological data available to plant genetic resources (PGR) workers. Next, we outline some of the PGR questions that can be investigated through spatial analysis, providing the reader with an overview of the kinds of insights GIS can provide. We outline the GIS tools available and conclude with a set of examples of GIS-based geographic analysis applied to PGR research.

Understanding Variation in Genetic Structure

Genetic structure is shaped by the interaction between the evolutionary forces of selection, gene flow, mutation, and drift. Variation in the effects of these forces in different parts of a species' range results in characteristic spatial patterns in the genetic structure of populations. Understanding these can be of practical utility in prioritizing where and how to conserve and use genetic resources.

The simplest of genetic structure tends to occur when there is no spatial variation in any of the evolutionary processes. In this case, genetic drift causes the gradual accumulation of genetic differences between different areas of a species range, and all areas appear equally related. Departures from this simple pattern are caused by spatial variations in the evolutionary processes of drift, selection, and gene flow, and can provide information relevant to conservation planning.

While drift can be influenced by spatial variation in population density, selection and gene flow often are explicitly shaped by environmental and geographic variables. Selection, in particular, may be related directly to landscape features acting on traits that confer adaptive responses to environmental stresses. These can include climatic factors such as cold or drought stress (Bekessy *et al.*, 2003), edaphic factors such as soil texture, mineral availability, or toxicity (Wu and Antonovics, 1976) and biological factors such as vegetation cover (Brouat *et al.*, 2003). Hedrick *et al.* (1976) and Linhart and Grant (1996) have reviewed links between genetics and environment extensively. Selection also can result in indirect correlation between genotype and environment in cases in which neutral genes or gene complexes are linked to selected alleles. The results of numerous studies by Nevo (2001) and colleagues suggest extensive linkage and indirect selection on many nonadaptive regions of the genome in a wide array of species and environments.

Gene flow results in the redistribution of alleles between populations or areas of a species' range, modifying patterns caused by drift and selection.

This can take place either through the physical movement and reproduction of individuals or through the dispersal of seeds or pollen. Although the degree of displacement will be affected primarily by geographical distance, other factors can modify the manner in which gene flow redistributes alleles, including barriers or restrictions to migration (Arnaud, 2003; Pfenninger, 2002), regional differences in phenology and pollination time (Galen *et al*., 1997), and seed/pollen dispersal mechanisms (Ennos, 1994). Since genetic patterns resulting from drift and/or selection can be greatly modified by gene flow, it is important to account for both.

Because evolutionary and conservation biologists are interested in understanding these processes, many methods have been developed to model their effect on spatial genetic patterns. For the most part, these rely on a hypothesis about how a given spatial process or feature should affect genetic structure and then compare landscape features to observed patterns in genetic structure.

An often examined hypothesis in the study of geographic effects on genetic structure is that of isolation by distance, suggesting that more geographically distant populations will also tend to be more genetically distinct (Wright, 1943). Areas of a species' range that are fragmented and not connected by gene flow tend to evolve distinct genetic patterns by the gradual accumulation of genetic differences by drift and mutation. Mutual exchange of alleles by gene flow tends to counteract such differentiation. A range of analytical techniques have been developed to analyze this effect of gene flow by comparing genetic structure and relatedness among several populations with measures of geographic or biological distance between them.

One of the simplest ways to analyze patterns of isolation by distance has been through comparing measures of relatedness or genetic differentiation between populations with the geographic distance separating them. Both Mantel tests and spatial autocorrelation methods use this approach to analyze whether populations have genetic structure expected from the isolation by distance model (Escudero *et al*., 2003; Heywood, 1991). These methods have been applied to many natural populations and have been used to identify minimum sampling distances for collection of neutral diversity in wild soybeans (*Glycine soja* Siebold and Zucc.) (Jin *et al*., 2003) and to study gene linkage in Norway spruce (*Picea abies* [L.] H. Karst.) (Bucci and Menozzi, 2002). Often, however, migration is not linear or equal in all directions, and euclidean distances are not appropriate. To account for the effects of barriers to migration, studies have used connectivity networks (Arnaud, 2003; Pfenninger, 2002). Although this approach has not been applied widely to plant populations, geographical barriers such as mountains and phonological barriers such as differences in climate or altitude could be incorporated. Other methods such as wombling and

the Monmonier algorithm also can be used to analyze observed genetic structure for evidence of barriers to gene flow (Manel *et al*., 2003). It is important to note, however, that all of these methods rely on the use of neutral molecular markers. Because selection can alter spatial patterns in adaptive traits, isolation by distance typically is studied only for neutral traits or for adaptive traits in which selection pressure is homogenous across a species' range (Lande, 1991; Nagylaki, 1994).

Another very extensively investigated hypothesis is that of the link between adaptive diversity and environmental heterogeneity. Although there is no standard methodology for such investigations, identifying such interactions typically has required the identification of significant differences between population trait means and the detection of correlations between trait means and selection pressures (but see Volis *et al*., 2004). Reciprocal transplant experiments can also be used to assess local adaptation. Varying approaches to this have been reviewed extensively (Hedrick, 1986; Hedrick *et al*., 1976; Linhart and Grant, 1996; Nevo, 2001).

The methods described here are only a small sample of a wide variety of approaches used for interpreting ecological effects on genetic structure. They have proven useful for understanding why genetic structure tends to look the way it does and have provided the basis for a secondary type of analysis, namely reversing the approach and using ecological and geographical information to predict patterns in spatial genetic structure. This approach has obvious uses in conservation, where the collection of genetic data is often prohibitively time-consuming and expensive. Although rules of thumb are derived easily from the results of the aforementioned investigations (i.e., genetic differences will be high between populations that are geographically isolated), an increasing number of methods are being developed to quantitatively predict and assess these patterns. Whereas spatial autocorrelation and similar methods rely on explicit sampling and analytical design to account for geographical patterns, mathematical approaches based on ecological mapping are a means of implicitly incorporating geography into analysis. Some of these advances are outlined in this chapter after some practical information on how to analyze biological data in a GIS.

Georeferenced Biological Data

Before discussing tools and analyses, it is important to describe the types of data generated and used in PGR conservation and use. PGR collections are typically sets of samples (or accessions) of seeds, live plants, or tissue cultures of cultivated plants or of wild crop relatives, sometimes with associated herbarium specimens. The information normally associated

with these collections includes so-called ''passport data'' (i.e., species name and perhaps local land race name, plus data about the plant collector, the date of collection, and descriptive information about the collecting site, including its geographic coordinates). Many PGR collection databases also include characterization data, which may refer to the phenotype (morphology, phenology) and/or genotype (molecular markers, isoenzymes, quantitative trait loci [QTLs], etc.) of the accessions. If geographic coordinate data are available or can be obtained, they may be said to be georeferenced in that the passport, characterization, and other associated data may all be linked to a particular location on the earth's surface.

Nonbiological Georeferenced Data

Both environmental factors such as climate, topography, and soils and anthropogenic factors such as habitat destruction and artificial selection help shape genetic patterns and structure in crops and related wild species.

Over the past decade, many global georeferenced datasets of such environmental and socioeconomic variables have been produced. For example, WORLDCLIM (http://biogeo.berkeley.edu/worldclim/worldclim.htm) is a global database of monthly climate variables (maximum temperature, minimum temperature, and rainfall) in the form of grid surfaces with a spatial resolution (cell size) of 1 km. These surfaces have been produced through the interpolation of up to 46,000 meteorological stations distributed across the globe. Topography is a fundamental environmental factor that affects soil characteristics, hydrology, and climate, among others. Global datasets of topography are available at very high resolution from the Shuttle Radar Topography Mission (SRTM) (http://srtm.usgs.gov/). This is a global dataset with a cell size of 3-arc s ($\sim$100 m in the tropics) and is available from a number of sources (U.S. Geological Survey (USGS) FTP server [ftp://edcsgs9.cr.usgs.gov/pub/data/srtm/], USGS National Map Seamless Distribution System [http://seamless.usgs.gov/], or the Land Use Project of the International Center for Tropical Agriculture [CIAT] [http://srtm.csi.cgiar.org]). Global datasets of soil exist but are considered fairly crude and inexact. The most popularly used is the Food and Agriculture Organization (FAO) Digital Soil Map of the World (http://www.fao.org/ag/agl/agll/dsmw.htm), which includes variables such as soil classification unit, pH, organic carbon content, C/N ratio, clay mineralogy, and soil depth. However, when working at the regional scale, this dataset does not contain sufficient detail, and national or regional scale soil surveys should be consulted.

Land cover also is important when attempting to locate habitats or understand the degree of fragmentation to which a habitat might have been subjected. Land-cover datasets are available at a variety of scales.

Global datasets exist with 1-km cell sizes that classify the land surface into land-cover classes, including forest (evergreen/deciduous, needleleaf/broadleaf), savannas, grasslands, water bodies, croplands, and others. Some examples of these are the USGS 1-km Land Cover Characterization dataset (http://edcdaac.usgs.gov/glcc/glcc.asp) and the European Union Joint Research Council Global Land Cover 2000 Project (http://www.gvm.sai.jrc.it/glc2000/defaultGLC2000.htm). Other products are available on a more regional scale, often derived from Landsat or SPOT imagery. The Global Land Cover Facility (http://glcf.umiacs.umd.edu/index.shtml) provides an array of land-cover measurements and free satellite images to download. The National Aeronautics and Space Administration (NASA) Mr. SID Image Server offers Landsat data for the entire globe (https://zulu.ssc.nasa.gov/mrsid/).

Humans also play an important role in shaping genetic diversity and species distributions of both wild plants and cultivated species. Basic socioeconomic datasets on the distribution of roads, administrative boundaries and towns and cities are available form the Digital Chart of the World (http://www.maproom.psu.edu/dcw/). Human influence can play a diversifying role for cultivated species but can be a cause of genetic erosion. The human footprint dataset (http://wcs.org/humanfootprint) is an integrated map with global coverage, at 1-km cell resolution, rating the degree to which human activities have influenced the land surface. Population surfaces exist for most parts of the world (the Consortium for International Earth Science Information Network's [CIESIN's] Global Gridded Population, http://sedac.ciesin.columbia.edu/plue/gpw/index.html; 1-km gridded population for Latin America, http://gisweb.ciat.cgiar.org/population/). Livestock grazing is a documented cause of genetic erosion (Williams, 2001), and 1-km datasets of cattle density exist for Asia and Africa (http://ergodd.zoo.ox.ac.uk/livatl2/).

This is only a brief review of the most fundamental spatial datasets of environmental and socioeconomic variables on a global scale. The Internet is a huge resource for locating spatial datasets and should be consulted to find the most up-to-date and detailed datasets for the study. Web portals, such as http://www-sul.stanford.edu/depts/gis/bookmark.htm and http://unr.edu.homepage/daved/gislinks.html are useful starting points for novice users.

Some Spatial Questions

- Where might I find a given species?
- Where might I find the greatest intraspecific and/or interspecific diversity?
- Where might I find germplasm with a specific genetic adaptation?

• Where should I collect samples of a species to accurately reflect its intraspecific diversity?

Plant genetic resource workers always are asking themselves these kinds of ''where?'' questions. Given the resource constraints under which the PGR community operates, it clearly is important to be able to target interventions as accurately as possible in space, to prioritize areas for germplasm collecting or target a particular region for the introduction and testing of a new improved cultivar. These spatial questions are important because answers to them—and others like them—are necessary to make the conservation and subsequent use of genetic resources as effective and efficient as possible. Geographic information systems may be used to analyze georeferenced data from genetic resource collections, either on their own or in conjunction with the other georeferenced data described earlier.

Geographic Information Systems and Spatial Analysis Tools for Biologists

A GIS may be defined as a database management system that can simultaneously handle digital spatial data (e.g., a map of the countries of the world) and logically attached, nonspatial, attribute data (e.g., the names and populations of the countries) (Guarino *et al*., 2002). In our application, the digital spatial data would be the locations where genetic resource accessions were collected and the attribute data would include the species name, collector, and characterization information associated with each accession. In the past, the adoption of GIS technology required significant investments in hardware, software, and human resources. Nowadays, GIS is within the reach of most interested biologists, given a computer and some good ideas. The tools available include generic GIS software (which will be used in diverse fields ranging from surveying to land planning to mineral exploration), Internet mapping technologies for publishing maps on the web, and specialist software tailored to the spatial analysis of biological phenomena.

Generic GIS software include the Environmental Systems Research Institute's (ESRI's) range of products (such as ArcGIS, ArcInfo, and ArcView 3.2), IDRISI, MapMaker, and the open-source program GRASS. These GIS tools include basic visualization of spatial data, in the form of points (e.g., towns), lines (e.g., roads), polygons (shapes such as country boundaries), and grids (continuous surfaces based on an array of cells, e.g., topography). Once visualized on screen, these software packages provide means of locating specific conditions, analyzing spatial patterns, and combining spatial datasets. Given that the biological researcher has clear

spatially related questions in mind, these generic tools often provide the means to analyze the relevant data and provide a useful answer.

However, there also are various tools for GIS beginners, tailored specifically to biologists, which incorporate established methodologies for the spatial analysis of biological data and facilitate their application. Some tools relevant to PGR conservation and use are presented in the following sections, but many more are available for diverse applications.

Data Checking

In many cases, the locality data are missing or erroneous, especially for older collections, making it important to complete and check the coordinates in the plant collection database before performing any analysis (Hijmans *et al*., 1999). Collecting localities often are distributed nonrandomly in space, showing distinct geographical biases. Hijmans *et al*. (2000) analyzed gene-bank collections of wild potato for bias in their geographic representativeness and detected strong overcollecting along roads and within areas previously identified as hotspots for the gene pool. Herbarium collections focus on diversity at the species level, with a strong taxonomic bias reflecting the specialization of botanists. These biases must be acknowledged in any analysis of PGR data.

Diversity Analysis Tools

Species-level and genetic diversity are not distributed randomly over the surface of the earth, and knowing where they are greatest obviously is a key consideration in targeting conservation efforts. However, diversity is a difficult parameter to map and analyze. Diversity studies usually begin by dividing the target area into a number of smaller zones, for each of which a measure of diversity is then calculated (Jarvis *et al*., 2003; Müller *et al*., 2003). Different geometric, political, or socioeconomic spatial units have been used (Csuti *et al*., 1997), although ideally, areas of equal shape and size (to reduce the area effect on diversity measures), such as square grid cells, are best. For each grid cell, either richness (number of different categories) or an array of diversity indices (combining richness with evenness in different ways) can be calculated, resulting in a diversity surface.

Grid-based mapping of diversity from point data is not a trivial analysis and can be done using various methodologies, all with associated assumptions and caveats. For example, moving the origin of the grid or changing the size of the grid cells can change the final result significantly. Nelson (2004) examined in detail the issue of scale in diversity mapping, identifying

a method for selecting the most appropriate scale of analysis using Monte Carlo simulations and statistical analyses of confidence.

Two GIS tools that can ''map'' diversity using grids in this way are freely available. DIVA-GIS (http://www.diva-gis.org) has a user-friendly interface that permits integrated analysis of PGR data, from mapping of diversity (employing different methods and offering various diversity indices) to understanding environmental adaptations and predicting species distribution (using the DOMAIN and BioClim methods described later in this chapter). DIVA-GIS contains global datasets of climatic variables for both the present and the projected future climate of 2055, as well as land-cover data, topography, and population. DIVA-GIS includes other useful functions for spatial analysis of biological data, many of which are discussed in Hijmans *et al*. (2002). WORLDMAP (http://www.nhm.ac.uk/science/projects/worldmap/) also maps diversity using the grid-based approach, including among other functions a means of mapping diversity weighted for the distinctness of taxonomic units, calculated from a phylogeny based on cladistic analysis (Vane-Wright *et al*., 1991).

Predictive Species Distribution Modeling

Identifying the precise geographic range of a species is often a fundamental step in locating, conserving, and using PGR. Specialist plant collectors use vegetation maps and previous experience to define the geographic range of a species, but this is both subjective and reliant on the availability and quality of these maps. For many species, knowledge is just insufficient to accurately map the geographic distribution. Anderson *et al*. (2002) state that shaded outline maps ranging between and beyond known localities are likely to overestimate species distribution, whereas dot maps of known localities portray species distribution conservatively. Geographic bias in collecting efforts [e.g., along roads (Hijmans *et al*., 2000)] creates further uncertainty in defining species range.

Much effort has gone into the development of methods for predicting the geographic distribution of species and now many of these have been incorporated into user-friendly tools. Typically, these methods use the conditions at points where the species has been found in order to construct a statistical model of the adaptation range of the species, based on a set of user-defined environmental variables. The statistical model then is applied over a wide region to locate other areas where the environmental conditions are potentially suitable for the species in question. These methods have been found to minimize the risk of overestimation and underestimation of geographic range (Franklin, 1995). Although they have been

applied only at the species level, they can be adapted to work at the genetic level if there is good reason to believe that the trait being studied is likely to be distributed nonrandomly with reference to given environmental variables.

Many of these range estimation methods assume that climatic variables are the principal drivers of geographic distribution (Franklin, 1995; Guisan and Zimmerman, 2000; Walker and Cocks, 1991), although other factors also have been used, including soils (Anderson *et al*., 2002), topography (Draper *et al*., 2003), and specific habitat conditions (Reutter *et al*., 2003).

Guisan and Zimmerman (2000) discuss some of the applications of species distribution modeling and the various algorithms that have been applied to the problem. Perhaps the most widely recognized method uses generalized linear models (GLMs), specifically logistic regression, to predict species distribution (Cumming, 2000; Draper *et al*., 2003; Guisan *et al*., 2002; Osborne and Suárez-Seoane, 2002; Pearce and Ferrier, 2000). This method requires not only input points detailing where a species has been found but also points of reported absences. In many cases, especially with PGR databases, these absence data are not available and are difficult to generate. Confirming an absence is also difficult and can often lead to false negatives (Jarvis *et al*., in press). No specific tool exists for performing species distribution modeling with the logistic regression method, but this analysis can be made easily with IDRISI in conjunction with a standard statistical software package (Draper *et al*., 2003).

Another algorithm for predictive species distribution modeling uses principal components analysis (PCA) (Jones *et al*., 1997; Robertson *et al*., 2001). This method involves performing a PCA on the environmental data at the points where a species has been collected and then uses the PC loadings to compute a probability distribution for all other environments in the study area. The result is a map of probabilities of finding the species. This method has been incorporated in the FloraMap software (Jones and Gladkov, 1999) (http://www.floramap-ciat.org/), which has been used in the study of wild crop relatives (Jarvis *et al*., 2003, in press; Segura *et al*., 2003). Further information about FloraMap is available in Jones *et al*. (2002).

Factor analysis also has been adopted for species distribution modeling (Hirzel *et al*., 2002) and is incorporated in BioMapper (http://www.unil.ch/biomapper/), which uses the Ecological Niche Factor Analysis (ENFA) algorithm. Other species distribution modeling tools worthy of mention are BIOM, which combines habitat suitability methods with distance-based calculations (Henning Sommer *et al*., 2003); DOMAIN (http://www.cifor.cgiar.org/scripts/default.asp?ref=research_tools/domain/index.htm), which uses the Gower metric to calculate similarity and distance from the conditions at known points of presence; GARP (http://www.lifemapper.org/

desktopgarp), which uses a neural network, specifically a genetic algorithm, to calculate the fitness of each area based on the calibration dataset (Anderson *et al*., 2002); and BIOCLIM, which uses a bounding box technique to define the environmental envelope that the species inhabits.

An important issue with species distribution modeling is validation of the results. Evaluations of species distribution models typically use presence/absence data to test how well the prediction fits with reality (Fielding and Bell, 1997). However, sampling issues complicate this, because absence is difficult to confirm, especially if the study covers a large area (Jarvis *et al*., in press). Both the kappa statistic and the area under the curve (AUC) (derived from the threshold Receiver Operating Characteristic [ROC]) have been used in the literature to validate presence/absence distributions (Cumming, 2000; Osborne and Suárez-Seoane, 2002; Pearce and Ferrier, 2000; Robertson *et al*., 2001). Manel *et al*. (2001) conclude that Cohen's kappa provides an appropriate statistical evaluation, benefiting from its simplicity to calculate and interpret.

As can be seen, many methodologies and tools are available for predicting species distributions. Manel *et al*. (1999) compare different methodologies, concluding that model performance differs only marginally and that the choice of method should depend more on the research questions being asked and the type of data that are available. Some of the criteria that might be used to decide among methods include the following:

- Whether the user needs to provide presence and absence data or presence only
- Whether the environmental variables can be categorical and continuous
- The degree of explanation that the method provides in terms of the environmental adaptation of the species
- Ease of use of the software and the inclusion of built-in datasets

Table I provides a brief review of four of the most common methods and tools, with some critical analyses of their relative advantages and disadvantages.

Genetic Diversity Models

A large body of work also endeavors to understand genetic structure of populations through simulation modeling of ecological processes at the genetic level. These models often include a spatial component that takes into account the effect of distance and population distribution on ecological and genetic processes. Typically, they require large calibration datasets to run simulations on real-life biological populations.

TABLE I
ADVANTAGES AND DISADVANTAGES OF FOUR COMMON METHODS AND TOOLS FOR PREDICTING SPECIES DISTRIBUTION

Method	Tool	Ease of use	Type of input biological data	Associated environmental data	Validation	Explanatory power
Logistic regression	No tool performs logistic regression; requires generic geographic information systems (GIS) and statistical software	Use of statistical package and generic GIS tool requires some basic knowledge	Presence and absence	Continuous data only, not necessarily independent variables; user must select, find, and provide associated variables for this analysis through use of a generic GIS tool	Split-sample validation can be made using the presence/absence data within a statistical package	Detailed analysis of statistical coefficients permits the user to understand the environmental factors most important in defining the species distribution
Principal components analysis (PCA)	FloraMap	User-friendly tool with example dataset and manual	Presence only	Continuous data only, not necessarily independent variables; FloraMap uses built-in datasets of 36 climate variables (monthly rainfall, temperature, and diurnal range in temperature), near-global coverage; users cannot use other datasets	No validation method incorporated in FloraMap; validation is difficult in FloraMap because of the difficulty in confirming an absence given the large cell size of the climate data (1–18 km)	Principal component scores and loadings allow the user to assess which climatic factors are most important in defining the species distribution

Ecological niche factor analysis (ENFA)	BioMapper	User-friendly tool with example dataset and manual; importing data can be difficult without some GIS experience	Presence only	Continuous data only, not necessarily independent variables; BioMapper does not include any data, so data must be imported, requiring some basic GIS skills	BioMapper contains a detailed set of validation techniques, including Cohen's kappa and area-under-the-curve (AUC) plots	Results of the factor analysis permit the user to examine which environmental variables are most important in defining the distribution
Genetic algorithm	Desktop GARP	User-friendly tool	Presence only	Desktop GARP contains built-in global datasets of climate and topography but also offers users the ability to import their own data (requires some GIS experience)	Desktop GARP offers automated split-sample validation techniques with statistical reports	Result is presence/absence (0/1), and provides no evidence for goodness of fit; model is black box and provides little insight into the contributing variables, although multiple runs can be made using jackknifing to define the most contributing variables

An example of this is the simulation model, ECO-GENE (http://www.ecogene21.org/), developed to study temporal and spatial dynamics of the genetic structure of tree populations (Degen *et al*., 1996). It is a distance-dependent model that combines elements of population genetics, demographical dynamics, forest growth, and management models. Overlapping or separate generations can be created, and different processes such as gene flow, mating systems, flowering phenology, selection, random drift, and competition can be simulated. It has been applied to study the impact of different silvicultural practices and the effect of air pollution on the genetic structure of tree populations (Degen *et al*., 1997, 2002; Takahashi *et al*., 2000). In its current form, ECO-GENE only deals with neutral traits, but the results provide a means of testing the possible effects of different management options.

Examples of GIS Use in Genetic Resources Conservation and Use

As has been shown, genetic diversity is in some part shaped by the environment, and in many cases, adaptations to local environments are of most interest to gene banks and germplasm users. Several concrete studies have shown how spatial analysis might prioritize conservation intervention, optimizing genetic conservation.

Jones *et al*. (1997) used the FloraMap program to predict the geographic distribution of wild bean (*Phaseolus vulgaris* L.) based on the distribution of germplasm and herbarium specimens. The results correctly predicted areas where wild bean had not been collected but was reported to occur in the literature. Given the success of this research, Segura *et al*. (2003) also used FloraMap to predict the distribution of five species of the genus *Passiflora* in the northern Andes. The results fitted closely with areas of known distribution of the species and identified two separate climatic adaptations within one species. Isoenzyme studies identified different zymotypes that were closely related to the two climatic clusters within the species *Passiflora tripartita* var. *mollissima* (Kunth) Holm-Niels. and P. Jorg. The study also identified collection gaps where *ex situ* germplasm collection should be focused.

Jarvis *et al*. (in press) also used FloraMap, this time in combination with land-cover maps, to locate potential collection sites for the rare wild pepper, *Capsicum flexuosum* Sendtn. In a controlled experiment, plant collectors visited 10 predefined sites where the species was predicted to be present and 10 sites where its absence was predicted. This methodology aimed to allow a formal validation of the method. Six new populations were found, representing a significant improvement over two previous collecting missions for the species in the same region, undertaken without

the use of GIS targeting. *C. flexuosum* was found at five of seven points predicted to harbor the species and not found at four of five points predicted not to harbor the species. Genetic analyses of this species are now being used to examine genetic diversity of collections from different climatic regions.

Draper *et al*. (2003) used cluster analysis of a suite of sites to define ecogeographic units to stratify germplasm collections of various species within a region. This method was applied to ensure that the germplasm collection covered the full environmental gradient, with the aim of conserving the greatest genetic diversity. This type of analysis could lead to the generation of sampling strategies to conserve the greatest intraspecific diversity in the least number of accessions.

Also at the species level, Draper *et al*. (2001) used logistic regression to select translocation sites for the rare species *Narcissus cavanillesii* Barra and G. López in Portugal. The natural habitat of the species was under threat from the construction of a dam, so spatial analysis was used to identify a region where the species might survive translocation. Climatic and ecological variables were used to find the optimum sites, and the survivorship of the species in their new habitat is being monitored.

A number of ecogeographic studies of crop wild relatives have been made using GIS-based approaches, with the aim of describing the biogeography of the gene pool and prioritizing potential conservation programs. Hijmans and Spooner (2001) constructed a database of more than 6000 collections of wild potatoes (*Solanaceae sect*. Petota) and analyzed the distribution of each species, locating the continental hotspots of species diversity. They used the grid-based diversity mapping method in DIVA-GIS to locate the areas with the greatest number of species, as well as Rebelo's (1994) complementarity algorithms to select the least number of grid cells to capture all species. Analyses of species distribution were also made, quantifying the spatial area that each species occupied and defining the maximum distance between collection points. Careful examination of these distribution characteristics permitted the identification of species that were undercollected relative to their geographic range.

Maxted *et al*. (2004) used this work as a model to analyze the biogeography of wild *Vigna* species in Africa. In this case, conservation priorities were assessed through comparing the actual species richness of germplasm and herbarium collection with the potential species richness calculated through predictive species distribution modeling. The grid-based diversity mapping method was used to identify the currently known hotspots of *Vigna* diversity, based on the existing collections. Then FloraMap was used to map the potential distribution of each of the 70 species. If the probability of finding the species was more than 0.5, presence was assumed and the

results of each species were combined to create a map of potential species richness. Comparing the ''potential'' with the ''actual'' species richness permitted the authors to identify new areas for germplasm collection or areas already visited but that were identified as potentially containing more species.

Jarvis *et al*. (2002, 2003) made a similar analysis of the biogeography of wild peanuts in Latin America, also prioritizing areas for *ex situ* and *in situ* conservation. Using the same dataset of wild peanuts, Ferguson *et al*. (2005) analyzed the climatic adaptation of each species through the extraction of climate data for each collection point in the database. Multivariate statistics permitted the authors to identify clustering of species adaptations and provided insights into the potential evolution of the cultivated peanut (*Arachis hypogaea* L.) from its wild relatives. This supported molecular evidence indicating that the species *Arachis duranensis* Krapov and W. C. Gregory and *A. ipaënsis* Krapov and W. C. Gregory are the wild progenitors of the cultivated peanut. Analysis of climatic adaptations like this provides key information to improve the use of genetic resources and can feed into crop improvement programs.

A limited amount of literature of diversity mapping is available at the genetic level. Hoffmann *et al*. (2003) used molecular data of the number of variable positions in the alignments and the distribution of recombinant sequence blocks to map genetic level diversity of *Arabidopsis thaliana* (L.) Heynh. In this case, the method of Kriging (a form of spatial interpolation) was applied to accessions with 13 sequenced loci to identify areas of greater diversity. The Atlantic Coast in Europe, from the western Iberian Peninsula to southern Great Britain, was found to have the greatest genetic variability. Although this type of analysis provides no insight into the processes creating the pattern of genetic diversity, it did detect spatial patterns in diversity that had not been identified in the data before geographic analysis.

Concluding Remarks

This chapter has shown how georeferenced biological information, analyzed with georeferenced environmental and socioeconomic data, can be used to understand the processes that generate genetic diversity. Such knowledge is necessary to answer the ''where'' questions that PGR researchers and users must address to be able to target their interventions most effectively and efficiently. Most of the practical examples of spatial analysis in PGR research have been made at the species or gene-pool level, locating areas of high species diversity, or using species distribution

modeling. These are important contributions to PGR work, but more examples at the genetic level are needed now.

Despite an established body of theory about how adaptive genetic structure may vary over space and along environmental gradients, experiments have tended to focus on simple and controlled examples to test the theories. As such, few examples translate this into general methods for prediction and analysis in complex environments more typical of most plant species. In part, this can be explained by the limitations of analytical models in simultaneously accounting for various spatially varying evolutionary forces. GIS-based analysis is one means of overcoming this limitation, because maps can inherently represent spatial processes, greatly simplifying the models required for analysis. Likely as a result of the difficulty of simultaneously accounting for varying selection pressure and gene flow, most genetic analyses in PGR have until now used neutral markers. Though useful for understanding gene flow and population demographics, neutral markers are of little use in assessing the adaptive traits that are of importance in conservation and PGR. Coupled with the improvement of methods for identifying adaptive traits (e.g., QTLs and SNPs), GIS-based spatial analysis will enable the rapid assessment of genetic diversity without costly field-based sampling and laboratory-based genetic analyses. As knowledge of adaptive genetics increases, tools such as FloraMap and DIVA-GIS will be invaluable in revealing spatial patterns, as has been done at the species level, thus guiding efforts to conserve and use this diversity. In order for genetic analyses to benefit from spatial analysis, it is important that germplasm collections are accurately georeferenced and that the geneticist considers spatial variation from the point of defining the sampling strategy through to analysis and interpretation of results.

References

Anderson, R., Gomez-Laverde, M., and Peterson, A. (2002). Geographical distributions of spiny pocket mice in South America: Insights from predictive models. *Global Ecol. Biogeogr.* **11,** 131–141.

Arnaud, J.-F. (2003). Metapopulation genetic structure and migration pathways in the land snail *Helix aspera*: Influence of landscape heterogeneity. *Landscape Ecol.* **18,** 333–346.

Bekessy, S. A., Ennos, R. A., Burgman, M. A., Newton, A. C., and Ades, P. K. (2003). Neutral DNA markers fail to detect genetic divergence in an ecologically important trait. *Biol. Conserv.* **110,** 267–275.

Brouat, C., Sennedot, F., Audiot, P., Leblois, R., and Rasplus, J. (2003). Fine-scale genetic structure of two *carabid* species with contrasted levels of habitat specialization. *Mol. Ecol.* **12,** 1731–1745.

Bucci, G., and Menozzi, P. (2002). Spatial autocorrelation and linkage of mendelian RAPD markers in a population of *Picea abies* Karst. *Mol. Ecol.* **11,** 305–315.

Csuti, B., Polasky, S., Williams, P. H., Pressey, R. L., Camm, J. D., Kershaw, M., Kiester, A. R., Downs, B., Hamilton, R., Huso, M., and Sahr, K. (1997). A comparison of reserve selection algorithms using data on terrestrial vertebrates in Oregon. *Biol. Conserv.* **80,** 83–97.

Cumming, G. (2000). Using between-model comparisons to fine-tune linear models of species ranges. *J. Biogeogr.* **27,** 441–455.

Degen, B., Gregorius, H.-R., and Scholz, F. (1996). ECO-GENE, a model for simulation studies on the spatial and temporal dynamics of genetic structures of tree populations. *Silvae Genet.* **45,** 323–329.

Degen, B., Roubik, D. W., and Loveless, M. D. (2002). Impact of selective logging and forest fragmentation on the seed cohorts of an insect-pollinated tree: A simulation study. *In* ''Modelling and Experimental Research on Genetic Processes in Tropical and Temperate Forests'' (B. Degen, M. D. Loveless, and A. Kremer, eds.), pp. 108–119. Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), Amazonia Oriental, Belém, Brazil.

Degen, B., Streiff, R., Scholz, F., and Kremer, A. (1997). Analyzing the effects of regeneration regime on genetic diversity and inbreeding in oak populations by use of the simulation model ECO-GENE. *In* ''Diversity and Adaption in Oak Species'' (K. C. Steiner, ed.), pp. 9–21. Pennstate College of Agricultural Sciences, Pennsylvania.

Draper, D., Rossello-Graell, A., and Iriondo, J. M. (2001). A translocation action in Portugal: Selecting a new location for *Narcissus cavanillesii*. A. Barra and G. López. Poster presented at the third Planta Europa Conference, 23–28 June 2001, Pruhonice, Czech Republic. http://www.plantaeuropa.org/html/conference_2001/conference_poster_pres.htm.

Draper, D., Rossello-Graell, A., Garcia, C., Gomes, C., and Sergia, C. (2003). Application of GIS in plant conservation programmes in Portugal. *Biol. Conserv.* **113,** 337–349.

Ennos, R. A. (1994). Estimating the relative rates of pollen and seed migration among plant populations. *Heredity* **72,** 250–259.

Escudero, A., Iriondo, J. M., and Torres, M. E. (2003). Spatial analysis of genetic diversity as a tool for plant conservation. *Biol. Conserv.* **113,** 351–365.

Ferguson, M. E., Jarvis, A., Stalker, H. T., Valls, J. F. M., Pittman, R. N., Simpson, C. E., Bramel, P., Williams, D., Guarino, L. (2005). Biogeography of wild *Arachis*: Distribution and environmental characterization. *Biodivers. Conserv.*

Fielding, A., and Bell, J. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv.* **24,** 38–49.

Franklin, J. (1995). Predictive vegetation mapping: Geographic modelling of biospatial patterns in relation to environmental gradients. *Prog. Phys. Geogr.* **19,** 474–499.

Galen, C., Stanton, M. L., Shore, J. S., and Sherry, R. A. (1997). Source-sink dynamics and the effect of an environmental gradient on gene flow and genetic substructure of the alpine buttercup *Ranunculus adoneus*. *Opera Bot.* **132,** 179–188.

Guarino, L., Jarvis, A., Hijmans, R. J., and Maxted, N. (2002). Geographic information systems (GIS) and the conservation and use of plant genetic resources. *In* ''Managing Plant Genetic Diversity'' (J. E. A. Engels, ed.), pp. 387–404. CAB International, Wallingford.

Guisan, A., and Zimmermann, N. (2000). Predictive habitat distribution models in ecology. *Ecol. Model* **135,** 147–186.

Guisan, A., Edwards, T. C., Jr., and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distribution: Setting the scene. *Ecol. Model* **157,** 89–100.

Hedrick, P. W. (1986). Genetic polymorphism in heterogeneous environments: A decade later. *Annu. Rev. Ecol. Syst.* **17,** 535–566.

Hedrick, P. W., Ginevan, M. E., and Ewing, E. P. (1976). Genetic polymorphism in heterogeneous environments. *Annu. Rev. Ecol. Syst.* **7,** 1–32.

Henning Sommer, J., Nowicki, C., Rios, L., Barthlott, W., and Ibisch, P. L. (2003). Extrapolating species range and biodiversity in data-poor countries: The computerized model BIOM. *Rev. Soc. Boliv. Bot.* **4,** 171–190.

Heywood, J. S. (1991). Spatial analysis of genetic variation in plant populations. *Annu. Rev. Ecol. Syst.* **22,** 335–355.

Hijmans, R. J., and Spooner, D. (2001). Geographic distribution of wild potato species. *Am. J. Bot.* **88,** 2101–2112.

Hijmans, R. J., Schreuder, M., De la Cruz, J., and Guarino, L. (1999). Using GIS to check coordinates of gene bank accessions. *Genet. Resour. Crop Evol.* **46,** 291–296.

Hijmans, R. J., Garrett, K., Huaman, Z., Zhang, D., Schreuder, M., and Bonierbale, M. (2000). Assessing the geographic representativeness of gene bank collections: The case of Bolivian wild potatoes. *Conserv. Biol.* **14,** 1755–1765.

Hijmans, R. J., Guarino, L., Cruz, M., and Rojas, E. (2002). Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS. *Plant Genet. Res. Newsl.* **127,** 15–19.

Hirzel, A., Hausser, J., Hessel, D. C., and Perrin, N. (2002). Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology* **83,** 2027–2036.

Hoffman, M. H., Glas, A. S., Tomiuk, J., Schmuths, H., Fritsch, R. M., and Bachmann, K. (2003). Analysis of molecular data of *Arabidopsis thaliana* (L.) Heynh. (Brassicaceae) with geographical information systems (GIS). *Mol. Ecol.* **12,** 1007–1019.

Jarvis, A., Guarino, L., Williams, D., Williams, K., and Hyman, G. (2002). Spatial analysis of wild peanut distributions and the implications for plant genetic resource conservation. *Plant Genet. Res. Newsl.* **131,** 29–35.

Jarvis, A., Ferguson, M., Williams, D., Guarino, L., Jones, P., Stalker, H., Valls, J., Pittman, R., Simpson, C., and Bramel, P. (2003). Biogeography of wild *Arachis*: Assessing conservation status and setting future priorities. *Crop Sci.* **43,** 1100–1108.

Jarvis, A., Williams, K., Williams, D., Guarino, L., Caballero, P., Mottram, G. (In press). Use of GIS for optimizing a collecting mission for a rare wild pepper (*Capsicum flexuosum* Sentn.) in Paraguay. *Genet. Resour. Crop Evol.*

Jin, Y. M., He, T., and Lu, B. R. (2003). Fine scale genetic structure in a wild soybean (*Glycine soja*) population and the implications for conservation. *New Phytol.* **159,** 513–519.

Jones, P., and Gladkov, A. (1999). ''FloraMap: A Computer Tool for the Distribution of Plants and Other Organisms in the Wild.'' Centro Internacional de Agricultura Tropical (CIAT), Cali, Colombia.

Jones, P., Beebe, S., Tohme, J., and Galwey, N. (1997). The use of geographical information systems in biodiversity exploration and conservation. *Biodivers. Conserv.* **6,** 947–958.

Jones, P., Guarino, L., and Jarvis, A. (2002). Computer tools for spatial analysis of plant genetic resources data: 2. FloraMap. *Plant Genet. Res. Newsl.* **130,** 1–6.

Lande, R. (1991). Isolation by distance in a quantitative trait. *Genetics* **128,** 443–452.

Linhart, Y. B., and Grant, M. C. (1996). Evolutionary significance of local genetic differentiation in plants. *Annu. Rev. Ecol. Syst.* **27,** 237–277.

Manel, S., Dias, J. M., Buckton, S. T., and Ormerod, S. J. (1999). Alternative methods for predicting species distribution: An illustration with Himalayan river birds. *J. Appl. Ecol.* **36,** 734–747.

Manel, S., Williams, H., and Ormerod, S. (2001). Evaluating presence-absence models in ecology: The need to account for prevalence. *J. Appl. Ecol.* **38,** 921–931.

Manel, S., Schwartz, M. K., Luikart, G., and Taberlet, P. (2003). Landscape genetics: Combining landscape ecology and population genetics. *Trends Ecol. Evol.* **18,** 189–197.

Maxted, N., Mabuza-Dlamini, P., Moss, H., Padulosi, S., Jarvis, A. Guarino, L. (2004). African *Vigna*: An ecogeographic study. International Plant Genetic Resources Institute (IPGRI), Italy.

Müller, R. T., Nowicki, C., Barthlott, W., and Ibisch, P. L. (2003). Biodiversity and endemism mapping as a tool for regional conservation planning—case study of the Pleurothallidinae (Orchidaceae) of the Andean rain forests in Bolivia. *Biodivers. Conserv.* **12,** 2005–2024.

Nagylaki, T. (1994). Geographical variation in a quantitative character. *Genetics* **136,** 361–381.

Nelson, A. (2004). ''The Spatial Analysis of Socio-Economic and Agricultural Data Across Geographic Scales: Examples and Applications in Honduras and Elsewhere.'' Ph.D. Thesis, School of Geography, University of Leeds, Leeds, UK.

Nevo, E. (2001). The evolution of genome–phenome diversity under environmental stress. *Proc. Natl. Acad. Sci.* **98,** 6233–6240.

Osborne, P., and Suárez-Seoane, S. (2002). Should data be partitioned spatially before building large-scale distribution models? *Ecol. Model.* **157,** 249–259.

Pearce, J., and Ferrier, S. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecol. Model.* **133,** 225–245.

Pfenninger, M. (2002). Relationship between microspatial population genetic structure and habitat heterogeneity in *Pomatias elegans* (O. F. Müller, 1774) (Caenogastropoda, Pomatiasidae). *Biol. J. Linn. Soc.* **76,** 565–575.

Rebelo, A. G. (1994). Iterative selection procedures: Centres of endemism and optimal placement of reserves. *Strelitzia* **1,** 231–257.

Reutter, B. A., Helfer, V., Hirzel, A. H., and Vogel, P. (2003). Modelling habitat-suitability using museum collections: An example with three sympatric *Apodemus* species from the Alps. *J. Biogeogr.* **30,** 581–590.

Robertson, M., Caithness, N., and Villet, M. (2001). A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Divers Distrib.* **7,** 15–27.

Segura, S., Coppens d'Eeckenbrugge, G., Lopez, L., Grum, M., and Guarino, L. (2003). Mapping the potential distribution on five species of *Passiflora* in Andean countries. *Genet. Res. Crop Evol.* **50,** 555–566.

Takahashi, M., Mukouda, M., and Koono, K. (2000). Differences in genetic structure between two Japanese beech (*Fagus crenata* Blume) stands. *Heredity* **84,** 103–115.

Vane-Wright, R. I., Humphries, C. J., and Williams, P. H. (1991). What to protect? Systematics and the agony of choice. *Biol. Conserv.* **55,** 235–254.

Volis, S. (2004). The influence of space in genetic-environmental relationships when environmental heterogeneity and seed dispersal occur at similar scale. *Am. Nat.* **163,** 312–327.

Walker, P., and Cocks, K. (1991). Habitat: A procedure for modelling a disjoint environmental envelope for a plant or animal species. *Global. Ecol. Biogeogr. Lett* **1,** 108–118.

Williams, D. (2001). New directions for collecting and conserving peanut genetic diversity. *Peanut Sci.* **28,** 135–140.

Wright, S. (1943). Isolation by distance. *Genetics* **28,** 114–138.

Wu, L., and Antonovics, J. (1976). Experimental ecological genetics in *Plantago*. II. Lead tolerance in *Plantago lanceolata* and *Cynodon dactylon* from a roadside. *Ecology* **57,** 205–208.