GEOME : A WEB-BASED LANDSCAPE GENOMICS GEOCOMPUTATION PLATFORM

S. Joost [a1], M. Kalbermatten [a]

[a] Laboratory of Geographical Information Systems (LASIG), Ecole Polytechnique Fédérale de Lausanne (EPFL)
(Stephane.Joost, Michael.Kalbermatten)@epfl.ch

Commission VI, WG VI/4

KEY WORDS: WebGIS, geocomputation, landscape genomics, biodiversity conservation, natural selection, High Performance Computing (HPC)

ABSTRACT:

Landscape ecologists and resource conservation managers increasingly need geo-referenced data but are not trained to efficiently use Geographical Information Systems together with appropriate geo-environmental information and spatial analysis approaches. In the present context of rapid global climate change, they show a renewed interest to study adaptation in species of interest (wildlife, livestock, plants) with the help of landscape genomics. This emerging research field is at the interface of genome sciences, environmental resources analysis, bioinformatics, geocomputation and spatial statistics. Their combination permits to assess the level of association between specific genomic regions of living organisms and environmental factors, to better understand the genetic basis of local adaptation. Landscape genomics is thus able to provide objective criteria to prioritize species which are the most worthwhile preserving. In livestock science for instance, husbandry based on adapted breeds is of key importance to emerging countries.
Consequently, to facilitate the use of landscape genomics, the GEOME project proposes a WebGIS-based platform for the integrated analysis of environmental, ecological and molecular data through the implementation of an original set of combined geocomputation, databases, spatial analysis and population genetics tools. The platform will gather two existing software to identify genomic regions under selection (BayeScan and MatSAM) using a WebGIS solution named MapIntera. The latter is a multiple views exploratory visualization interface allowing users to interactively browse and query multiple dynamic layouts (e.g. graphs, maps). It is based on a SVG graphical interface coupled to Javascript interactivity and AJAX technology for the retrieval of information from a spatial database. The platform will be connected to a High Performance Computing infrastructure, able to handle and process very large genome and environmental data sets

INTRODUCTION

Since 2000, we are witnessing a progressive integration of the fields of ecology, evolution and population genetics (Lawry, 2010), and recently also geocomputational tools (Joost et al., 2007). The combination of landscape ecology and population genetics with spatial analysis and GIS led to the advent of landscape genetics whose goal is to understand how geographical and environmental features structure genetic variation in living organisms (Manel, 2003). Recently, landscape genomics (Luikart, 2003; Joost, 2007; Lawry, 2010) emerged as a research field at the interface of genome sciences, environmental resources analysis and spatial statistics. The combination of these fields permits to assess the level of association between specific genomic regions and environmental factors to identify loci responsible for adaptation to different habitats (Lawry, 2010). Henceforth, knowledge of geo-environmental data and skills in GIS are a necessity to develop research or management activities in these landscape sciences. To favor the use of landscape genomics and to facilitate access to necessary geo-environmental data, GIS and spatial analysis tools, the development of a robust and efficient geocomputational infrastructure is required. Therefore, GEOME will constitute a robust, easy-to-use and powerful internet platform based on High Performance Computing (HPC), able to handle and process very large genome and environmental data sets, and offering facilities for the statistical and (geo)visual analysis of results. An initial version has been implemented and can be found here: http://lasigpc8.epfl.ch/geome/.

Forthcoming developments of the platform will include access to free geo-environmental data (database) and provide dedicated GIS tools to scientists and professional users (e.g. landscape ecologists or resource conservation managers) who often do not have a background nor a training in GIScience (Geographic Information Systems and spatial analysis), and do not have time to acquire them. To be able to carry out their investigations, these users need support to i) find appropriate geo-referenced environmental data sets to address their research problematics, ii) integrate their own data sets (e.g. presence/absence of genotypes, of species) with the geo-environmental information mentioned here above, iii) benefit from specific analytical tools implemented within a simple GIS environment.

Several web-based platforms already exist in the domain of genomics or other genetic topics (population genetics, phylogenetics). An interesting example is the BIOPORTAL platform (http://www.bioportal.uio.no), a web-based service platform developed at University of Oslo for phylogenomic analysis, population genetics and high-throughput sequence analysis. BIOPORTAL is the largest publicly available computer resource with 300 dedicated computational cores in a cluster named TITAN. Currently, applications in chemistry and

[1] Corresponding author.

economics have been integrated, but other applications will be added along the way.
It has to be noted that no existing Web-based platform includes the combination of services to be implemented within GEOME. This combination will permit to address the issue of local adaptation through the complementary implementation of a theoretical approach in population genomics (Foll and Gaggiotti, 2008) and of a spatial analysis approach (Joost et al., 2007) and corresponding software, respectively BayeScan and MatSAM.

LOCAL ADAPTATION

Local adaptation is an important issue in conservation genetics. This process has to be better understood in order to correctly consider the effects of transfers of individuals between populations, which is often a technique proposed to replenish genetic variation and to reduce negative effects of low genetic diversity. The study of local adaptation is also likely to provide objective and unambiguous criteria to characterize conservation areas which are the most worthwhile preserving.

In parallel, development of low impact sustainable agriculture as well as husbandry based on adapted breeds is of priority to most countries in the world, and is of key importance to emerging countries in particular. The genetic basis and the level of adaptation of livestock breeds to their environment has to be investigated, in order to reach better understanding of the relationship between environment and the adaptive fitness of livestock populations, and to favor sustainable production systems based on adapted breeds.

During the last decade, “tremendous advances in genetic and genomic techniques have resulted in the capacity to identify genes involved in adaptive evolution across numerous biological systems” (Lowry, 2010). There is now an important need to provide tools allowing the acceleration of the study of how landscape-level geographical and environmental features are involved in the distribution of functional adaptive genetic variation, as highlighted by recent publications (Lowry, 2010). A few years ago, Holderegger and Wagner (2008) already stated that “Novel approaches linking spatially explicit environmental analysis with molecular genetics could offer effective means to study the spread of adaptive genes across landscapes”. Landscape genomics is one of these approaches, and GEOME will facilitate its use by means of a database providing access to the many geo-environmental data sets available worldwide.

ACCESS TO ENVIRONMENTAL DATA

Joost et al. (2010) provided a non-exhaustive list of environmental data sets available on the Internet. Different initiatives at the regional and global levels influence and promote the creation of Spatial Data Infrastructures (SDIs) to access these data. They also work on the harmonization, standardization, interoperability, and seamless integration of the different GIS layers constituting these data. An example is the Global Earth Observation System of Systems (GEOSS), which is a worldwide effort to connect already existing SDIs and Earth Observation infrastructures. Through its developing GEO portal and related Common Infrastructure, GEOSS is foreseen to act as a gateway between producers of environmental data and end users, with the aim of enhancing the relevance of Earth observations for global issues and to offer public access to comprehensive information and analyses on the environment (GEO secretariat, 2007).

Today's effort on the technical development of SDI components clearly focuses on the exchange of geodata in an interoperable way (Bernard and Craglia, 2005), which is highlighted by the concept of web services and the related Service Oriented Architecture (SOA). Web services constitute a “new paradigm" allowing users to retrieve, manipulate and combine geospatial data from different sources and different formats using HTTP protocol to communicate (Sahin and Gumusay, 2008). Web services enable the possibility to construct web-based application using any platform, object model and programming language. The Open Geospatial Consortium (OGC) has specified a suite of standardized web services. Two of them are of particular interest for data providers and users: the Web Feature Service (WFS) that provides a web interface to access vector geospatial data (like country borders, GPS points or roads) encoded in Geographic Markup Language (GML) and the Web Coverage Service (WCS) that defines a web interface to retrieve raster geospatial data of spatially distributed phenomena such as surface temperature maps or digital elevation models (DEMs). In addition, the Web Map Service (WMS) defines an interface to serve georeferenced map images suitable for displaying purpose based on either vector or raster data.

GEOME’s Web services will provide the geocomputational context with the support of an indispensable High Performance Computing (HPC) infrastructure to enable the processing of associations models between millions of loci and hundreds of environmental parameters (see next sections).

COMPUTATIONAL CHALLENGE

One of the next major steps in evolutionary biology is to determine how landscape-level geographical and environmental features are involved in the distribution of the functional adaptive genetic variation (Lawry, 2010). This challenge will take place in a context where the amount of molecular data to be analyzed will expand very rapidly. Indeed, after the long (13 years) and expensive (3 billion dollars) human genome sequencing project (completed in 2003), the American National Institutes of Health (NIH) proposed in 2004 the challenge to sequence one human genome for $1’000 (Service, 2006). And future generation of sequencers will allow researchers to get to genomic data faster and at a lower cost. For example, the theoretical potential of single-molecule/nanopore sequencing is undeniable (Tersoff, 2001). Based on this technology, with a 100 nanopores in parallel, a mammalian genome could be sequenced in 24 hours with the main cost being the chip itself, probably around $1'000 (Blow, 2008). Several alternative low cost sequencing technologies are under way, even decreasing to $100 in the case of the Pacific Biosciences technology (Eid et al., 2009).

Thus, any research project in landscape genomics will very soon be given the opportunity to investigate the entire genome of sampled individuals, meaning that from now on HPC is required to process these huge quantities of data when compared to eco-climatic parameters. To be ready to handle such volumes of data, the GIS laboratory (LASIG) involved in the development of GEOME is also a partner within the EU FP7 NEXTGEN, the first project in the area of conservation genetics that proposes a comparative analysis of whole genome data at the intraspecific level (http://nextgen.epfl.ch). NEXTGEN will also need GEOME’s expertise in the area of remote sensing, digital elevation models and other environmental data in general.

ENVIRONMENTAL DATA IN A HPC ENVIRONMENT

Environmental sciences are a data-intensive domain in which applications typically produce and analyze a large amount of geospatial data. Moreover, due to the multi-disciplinary nature of environmental sciences (e.g. ecology, climate change, etc.), scientists need to integrate large amount of data distributed all around the world in different data centers. A local cluster is the first mean to upscale computational capacities when the workflow can be distributed into many independent jobs.
When the number of jobs is very large and/or users do not belong to the Institution owning the cluster, grid computing can be an efficient solution. Following Foster et al. (2008), a grid is a parallel processing architecture in which computational resources are shared across a network allowing access to unused CPU and storage space to all participating machines. Resources could be allocated on demand to consumers who wish to obtain computing power. Recent studies have had a successful approach to extend grid technology to the remote sensing community (Muresan et al., 2006), as well as in the field of disaster management (Mazzetti et al., 2009) making OGC web service grid-enabled.

One of the largest scientific grid infrastructures currently in operation is the Enabling Grids for E-sciencE (EGEE) infrastructure bringing together more than 120 organisations to provide scientific computing resources to the European and global research community. The currently 120'000 CPUs available in EGEE are essentially used by the High Energy physics community, but part of EGEE is also opened to other disciplines such as environmental sciences or genetics. A grid environment federates its users through Virtual Organizations (VOs), which are sets of individuals and/or institutions defined by a set of sharing rules (e.g. access to computers, software, data and other resources). One particular VO of interest is named Biomed; it has currently access to 20'000 CPUs and is willing to accept the envisioned GEOME on the Biomed VO.

ARCHITECTURE

The architecture of GEOME is constituted of open-source technologies. It is composed of an Apache server connected to PostgreSQL/PostGIS databases and to an image server (Geoserver). Presently, the latter only includes an Earth Shaded image from Natural Earth.

The Apache server produces xHTML pages dynamically and sends them to the client browser. When the user interacts with this content, Javascript sends HTTP requests to upload dynamic content. PostgreSQL/PostGIS servers are called when the user logs into GEOME and when he defines, uploads or deletes his data. Currently, all PostgreSQL/PostGIS databases are active excepted the environmental database and the MatSAM user results database. A new database is defined for each new user in order not to interfere with other users and to ensure privacy.

The mapping system is developed using Scalable Vector Graphics (SVG). Vector enables the possibility to include much more client-side interaction using European Computer Manufacturers Association (ECMA) Javascript. SVG code is produced on the fly using PHP (Figure 1).

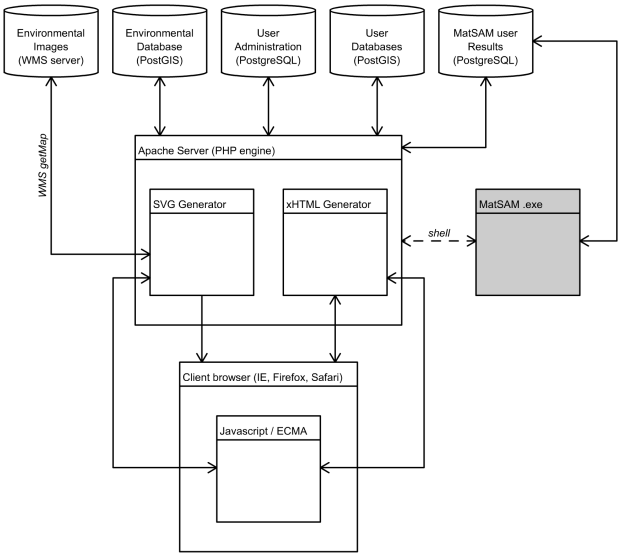
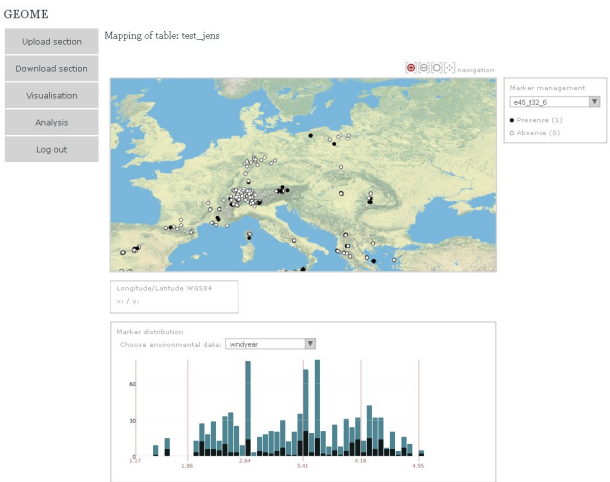


Figure 1. Components of the GEOME WebGIS platform.

Presently, the user can map genetic data according to their geographic coordinates and their presence/absence status (binary value). If the user changes the genetic marker to visualize, the system dynamically loads the values of the markers and modifies the point formatting consequently (Figure 2).

Figure 2. GEOME's interface with the map showing presence/absence of genetic markers and the histogram showing the distribution of environmental variables.

As regards values of environmental variables, for the time being the system only allows to visualize environmental variables the user imported together with his genomic data using a histogram. When the user loads data, a dynamic request checks if all environmental variables are defined. If not, a dynamic interface, loaded using Javascript, asks him to proceed. Currently, two data types can be defined: nominal/ordinal data, and cardinal data. These types are important for the histogram generation and for running MatSAM over the data set. When defining nominal/ordinal data, the user will have to indicate the different (ordered) classes as input. In the second case, data are continuous. In this case, the histogram is generated using 50 different intervals.



Again, the histogram is dynamically computed as the user changes the selected environmental variable. Furthermore a link is created between the geographic representation and the histogram representation. This one permits to identify which point lies within which histogram interval.

More free environmental data will be inserted in a new database and spatial overlay will enable the user to retrieve adequate data to characterize the location where genetic data were sampled. These data will be stored as raster matrices in a PostgreSQL database, but also as ready-to-use formatted images on a WMS (Web Map Service) server for mapping purposes.

At last, and once the users have defined all the parameters regarding their data, an interface dedicated to MatSAM will also be developed. This one will give the opportunity to run MatSAM on the server side (via dynamic HTTP requests) and the results will be stored on a database as long as the user has not downloaded them.

CONCLUSION

No software tool presently exists to answer challenges of developing better approaches for linking ecologically relevant data sets to specific loci or genes. Moreover, no software tool can currently integrate geo-environmental variables and molecular data within a WebGIS environment, with analysis modules included i) to detect loci under selection according to two complementary approaches, ii) to analyze and characterize the spatial distribution of these loci, and iii) to produce predictive habitat maps derived from them.

Thanks to a robust HPC infrastructure, GEOME's existing and future Web services will be useful to a very large number of users, and will possibly contribute in understanding how landscape-level geographical and environmental features are involved in the distribution of the functional adaptive genetic variation.

REFERENCES

Bernard, L., and Craglia, M., 2005. SDI - From Spatial Data Infrastructure to Service Driven Infrastructure. *First Research Workshop on Cross-learning on Spatial Data Infrastructures (SDI) and Information Incrastructures (II)*, Enschede, The Netherlands.

Blow, N., 2008. DNA sequencing: generation next-next. *Nature Methods*, 3: 267- +.

Eid, J. et al., 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 5910: 133-138.

Foster, I, Zhao, Y., Raicu, I., Lu, S., 2008. Cloud Computing and Grid Computing 360-degree compared. *IEEE Grid Computing Environments* (GCE08), pp. 1–10.

GEO secretariat, 2007. Strategic Guidance for Current and Potential Contributors to GEOSS. *Fourth Plenary Session of GEO (GEO-IV)*, Cape Town, 4 p. http://www.earthobservations.org/docs_od_ple.shtml

Holderegger, R. and Wagner, H.H. (2008) Landscape genetics. *Bioscience*, 58: 199-207.

Joost, S., Bonin, A., Bruford, M.W., Després, L., Conord, C., Erhardt, G., Taberlet, P., 2007. A Spatial Analysis Method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology*, 18: 3955–3969.

Joost, S., Colli, L., Baret, P.V., Garcia, J.F., Boettcher, P.J., Tixier-Boichard, M. , Ajmone-Marsan, P. & the GLOBALDIV Consortium, 2010. Integrating geo-referenced multiscale and multidisciplinary data for the management of biodiversity in livestock genetic resources. *Animal Genetics*, 41:47-63.

Manel, S., Schwartz, M., Luikart, G., Taberlet, P., 2003. Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, 18: 189-197.

Mazzetti, P., S. Nativi, Angelini, V. Verlato, M., Fiorucci, P., 2009. A Grid platform for the European Civil Protection e-Infrastructure: the Forest Fires use scenario. *Earth Science Informatics*, 2: 53-62.

Muresan, O., Pop, F., Gorgan, D., Cristea, V., 2006. Satellite Image Processing Applications in MedioGRID. *Proceedings of The Fifth International Symposium on Parallel and Distributed Computing (ISPDC'06)*, 6-9 July 2006, Timisoara, pp.253-262.

Lowry, D.B., 2010. Landscape evolutionary genomics. *Biology Letters*, DOI: 10.1098/rsbl.2009.0969.

Luikart, G., England, P.R., Tallmon, D., Jordan, S., Taberlet, P., 2003. The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics*, 4: 981-994.

Sahin, K. and Gumusay, M. U., 2008. Service oriented architecture (SOA) based web services for geographic information systems. *XXIst ISPRS Congress*, Beijing, pp. 625-630.

Service, R.F., 2006. The race for the 1000$ genome. *Science*, 311: 1544-1546.

Tersoff, J., 2001. Nanotechnology: Less is more. *Nature*, 412: 135-136.