Proceedings of the International Conference on Computational Biosciences and Bio-Informatics (ICCBBI 2011), July 2011, Bhubaneswar, India.

A Software Tool for Information Management and Data Mining of Biological Data for Studying Adaptation of Living Organisms in Amazonia

Pedro Santoro Perez
University of Sao Paulo
Ribeirão Preto, SP, Brazil
pedrosperez@usp.br

Ana Helena Bevilacqua
National Institute for Research in the Amazon
Manaus, AM, Brazil
anahelena.bevilacqua@gmail.com

Andrea Ghelfi
Federal University of Amazonas
Coari, AM, Brazil
andreaghelfi@gmail.com

Alessandra A.Macedo
University of Sao Paulo
Ribeirão Preto, SP, Brazil
ale.alaniz@usp.br

Sérgio R.Nozawa
Nilton Lins University Center
Manaus, AM, Brazil
snozawa@niltonlins.br

José Augusto Baranauskas
University of Sao Paulo
Ribeirão Preto, SP, Brazil
augusto@usp.br

***Abstract*—The Amazon has economic and ecological implications, given the importance of its natural resources. Understanding how organisms are able to adapt to difficult conditions and how organisms take notice of those conditions and migrate are going to be of key importance to improve the effectiveness of saving endangered species, produce effective environmental policies, and produce biological knowledge. The ADAPTA project is held in the Amazon and aims at studying the adaptation of living organisms to changes in environmental conditions, especially considering aquatic ecosystems. Hence, biological information is being collected, integrated and analyzed as never before. The ultimate objective of ADAPTA is to recognize adaptation-related genes and to develop products, processes and environmental policies through data mining strategies. ADAPTA stores information into three integrated information systems, each one being responsible for a major aspect of the project, integrated with and dependent on the others. This study reports the development of the first of such information systems as a software tool that allows management and data mining of biological information in ADAPTA. This root system deals with all the interaction information between organisms and environments gathered by researchers, and it has been developed using open standards.**

Keywords - Biological Information System; Amazon; Data Management and Integration; Data Mining, Machine Learning

I. INTRODUCTION

The significant biological diversity existing in the Amazon basin has been exhaustively documented, and it seems that a definitive number for its dimension is far from being established, considering some vulnerable areas, where environmental changes happen too fast. Only in the Amazon river, for example, there are more different fish species than in entire Europe [1]. Despite all that diversity, current Amazon numbers are thought to be lower than the actual ones [2].

Over the evolutionary process, organisms developed exceptional abilities to interact, reduce and even make use of the effects of extreme natural conditions of the dynamic environments in which they live. It could be foreseen that numerous mechanisms are shared among the organisms of different phylogenetic groups. These are adaptive convergences to the Amazon environment natural variations. There has been also a significant number of anthropogenic interventions in the Amazon with immediate impacts in the aquatic environment such as new hydroelectric power plants, changes in soil use, mining, highways, deforestation and urban impacts. It is still unknown how Amazon organisms, particularly aquatic ones, respond to this kind of environmental changes, to which some species are already adapted. It is necessary to learn what gives them such skills, how these adaptations work and how they are activated [3].

This work is part of a project called ADAPTA (Centre for Studies of Adaptations of Aquatic Biota of the Amazon), hosted by the National Institute for Research in the Amazon[1]. Using modern tools of observation, remote sensing, and biochemical, physiological and ecological aspects, ADAPTA has been studying fish species, aquatic plants, invertebrates, microorganisms and aquatic mammals in natural and impacted environments and under experimental conditions (microcosms). ADAPTA aims to analyze why and how different groups of plants and animals are able to survive or not on extreme environments, and also analyze why those species which would not survive sense the changes in advance and emigrate [3].

These mechanisms are extremely useful for establishing actions for environmental conservation, supporting decisive processes concerning environmental interventions and resulting in products and processes for the improvement of the life quality of the Amazon people. The literature presents isolated descriptions of these fantastic adaptations that occur in all biological levels, from the behavioral to the molecular level, but the dynamics of this adaptive process is not entirely known and there is still much to be understood [3]. Also, these mechanisms constitute important information for the definition of potential biomarkers. In the near future, this information will also be useful for the definition of different species groups that will have the genome (cDNA) analyzed, in order to search for common genes that are up or down regulated. This will be achieved by cataloging the gene expression, sequencing of DNA and subsequent analysis by bioinformatics and data mining. The development of new products and processes considering gene functional characteristics of flora and fauna elements is more likely to be successful than studies on Amazonian biodiversity elements themselves. All results will generate significant support for public policies [4-5], resulting from the analysis of plants and animals chronically exposed into controlled environmental conditions, through specifically built microcosm chambers.

In this study the first software developed for ADAPTA (DB-I) is presented, which handles information about collections, organisms and experiments; analysis of the database, allowing users to set criteria and perform searches based on them; visualization of the collections in a map; and

[1] http://adapta.inpa.gov.br/

application of data mining and machine learning tools, in order to build predictive classification models based on information about collections, organisms, environments and experiments. These tools will help reduce the number of molecular variables to be examined in the future (see Sect. III) in several groups of species, either facing normal or different extreme environmental conditions. This will make it possible to establish bioindicators and biomarkers that will allow the monitoring of environment quality.

II. Related Work

There have been many studies carried out in the Amazon. They all have economic and ecological implications, given the importance of its natural resources. One example is [6], which collected data into a geographical information system in the state of Pará, Brazil. The work of [7] observes that computer software development has contributed to a better understanding of the ecology and functioning of ecosystems and their underlying mechanisms and also that integration of different sorts of biological data is extremely useful for such applications. They developed a system called *SynBioSys*, which integrates different information systems.

In [8] a database framework application was developed to address the problems of data manipulation and analysis in the field of ecology. A difference between their application and the one from this project is that they have based the development on commercial software. The *RAINFOR* database [9] proposes a relational database for recording data about trees in the Amazon. They have also based their development on commercial software. Another example of such systems can be found in [10-11], in which a relational database application makes it possible to store and analyze repeated collection of observational data, analogously in some aspects to the work presented here. An example of using georeferencing resources and applying machine learning algorithms on collected data can be found in [12].

III. Materials and Methods

Information management systems are of key importance because they make it easier to work with huge amounts of data produced nowadays, providing a unified view of data. Intelligent computational systems are able to manage, store, retrieve, share and classify biological information, therefore, providing means to decision support, especially in large projects such as ADAPTA. A step ahead is the integration between artificial intelligence and database systems includes the high availability of shared data for knowledge processing [13]. These systems are supposed to be robust, functioning even in the presence of data omission, modification or substitution [14].

Considering this, a software tool allowing the management and analysis of data and information in ADAPTA has been built in this study. It defines the storage, the user interface and how analysis can be carried out.

A. Storing Components

The ADAPTA project encompasses three research lines: Adaptation to the Biological Environment, Biomarkers and Functional Genomics, and Applied Programs. A repository with collected data is planned for each component, named DB-I, DB-II and DB-III, respectively. It is clear these three databases must be integrated with each other. Fig. 1 shows the structure and functioning of the databases in ADAPTA. Considering it is still under progress DB-II and DB-III are yet to be implemented in the near future; therefore the aim of this communication is to present the development of the first database, DB-I.

B. DB-I Architecture

The DB-I allows ADAPTA's researchers to deal with all the information gathered about the interaction between organisms and their environments. The architecture of DB-I is structured into four layers, as shown in Fig. 2.

The graphical user interface is part of the presentation layer. It is intended to friendly let researchers feed, query and analyze the base. It also allows users to visualize their collections (with latitude and longitude values) in a map. All data collected are stored and maintained by the storage layer.

The query layer is responsible to process and execute queries issued by users. For example, a researcher might want to know which organisms live in an environment where fishing is commercially explored and whose water is at a temperature between 25 and 30^{o} C. That kind of analysis will guide researchers in selecting DB-II contents, pointing out which organisms will potentially have their RNA sequenced.

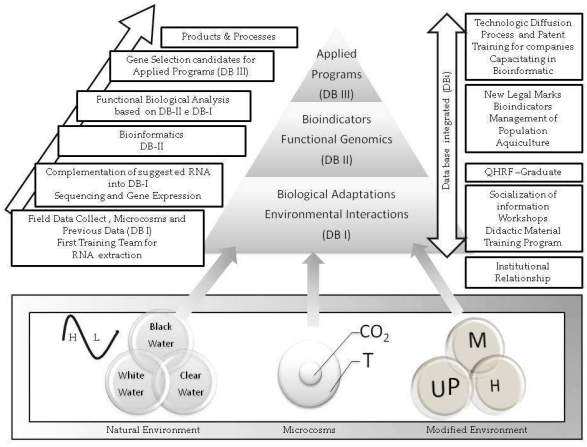

Figure 1. The pyramid shows the three levels of activity, encompassing the three research lines. The bottom of the figure indicates that environments and microcosms will be studied and where the data for DB-I is being generated. The left side of the picture presents the actions which will be developed in order to obtain information to build the components. The right side of the figure lists the main products and their developments, which will integrate all actions, compiled in the integrated database, composed of the three repositories, with different kinds of information: DB-I stores all interaction data collected (in the field, laboratory and literature) by researchers and has been used in different analyses; DB-II will store and analyze information about RNA extraction, gene expression and RNA sequencing of those organisms stored at DB-I; at last, DB-III will serve for comparison of gene products. Legend: H (natural environment) - high water season, L - low water season, T - temperature, CO_2 - carbon dioxide; UP - urban pollution; M - mining (oil and gold); H (modified environment) - hydroelectric plants; QHRF - qualified human resources training. This figure was developed and granted by Vera Val, 2009 and has already been published in [3] and we were allowed to reproduce it here.

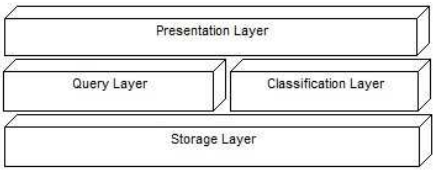

Figure 2. DB-I four layers: (i) the presentation layer is composed of graphical user interfaces responsible for user interactions; (ii) the query layer is responsible for answering researchers' questions; (iii) the classification layer is a machine learning framework that uses decision trees and provides comprehensible models explaining data; and (iv) the storage layer, which manipulates data from an abstraction of the biological application.

The classification layer offers the possibility of building classification models based on data collection. It takes attributes and classes from the database DB-I and builds predictive tree-like models [15]. Induction of decision trees is a machine learning approach that has been applied on several tasks.

Decision trees are well-suited for large real world tasks, since they scale well and can represent complex concepts by constructing simple, yet robust logic-based classifiers amenable to direct expert interpretation. Top-down induction of decision tree algorithms [16] in general choose a feature that partitions the training data according to some evaluation function. Partitions are then recursively split until some stopping criterion is reached. After that, the decision tree is pruned in order to avoid overfitting [17].

Decision trees are interesting from the biological point of view, since they preserve the logical connections among attributes, enabling the induction of human-comprehensible classifiers, expressed as rules [18-19]. In this case, the aim is to build classifiers composed of rules with few conditions (typically 2 to 5 attributes). It is important to stress that these simple classifiers may have lower predictive accuracy than more complex classifiers [20], but they explicitly emphasize the importance of the correlation among attributes, including expressed and/or non-expressed genes [21].

DB-I was designed with help from biologists of the project. Since there are 21 Brazilian associated laboratories involved in ADAPTA, besides associated laboratories from other countries, with different characteristics and needs, intense work was done to find out what would be the relevant information to the project as a whole. The main entities (and their attributes) were then defined. Fig. 3 shows the main classes and some of their methods. *Person* represents people involved in ADAPTA, e.g., collectors and laboratory coordinators. A *Person* object belongs to one or more *ResearchGroup* objects. *ResearchGroup* represents the laboratories involved in the project. A *ResearchGroup* object has one *Person* object as its coordinator. A *ResearchGroup* object, along with its coordinator, is responsible for zero or more *Collection* objects. *Collection*, which represents data collected from several sources; *Environment*, which is where data are collected; *Individual*, collected organisms along with all data about the environments and their conditions; and *Experiment*, to hold data about laboratory procedures performed with the collected organisms.

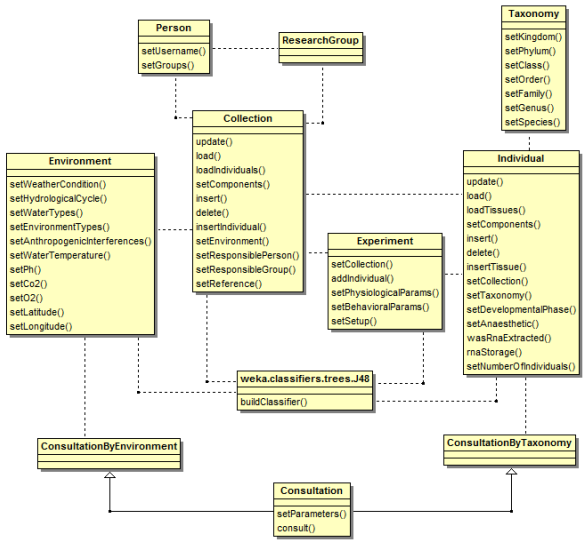

Figure 3. DB-I class diagram. Only main classes were shown.

Taxonomy represents information about the taxonomic classification of organisms. There are also secondary entities, which support the main ones such as *Species, Anthropogenic interference, Hydrological cycle*, among others (not shown).

The database was implemented in PostgreSQL, a robust and powerful open source object-relational database management system. The secondary entities were mapped into auxiliary tables in the database. Those tables were fed with data provided by researchers of ADAPTA.

C. *Funcionalities*

In order to allow researchers to view all data in a more friendly way and to facilitate the performance of customized analyses, a Java application was built. Each user is given a username and a password for authentication in the system. Typical users have access permission only to data related to the laboratories of which they are part. The products of DB-I have been divided into five sets of functionalities.

The first set of functionalities is responsible for managing collections. It involves everything related to the storage, manipulation and visualization of data resulting from what biologists collect. The second one deals with the management of the experiments. It involves everything related to the storage, manipulation and visualization of the data resulting from experiments biologists perform. The next set allows researchers to query the database and analyze data about collections and experiments. The fourth set allows classification of data, using decision trees. Visualization, in a map, of the locations where collections were performed is delivered by the fifth set.

Some aspects of the system are context-sensitive, to make it more intuitive to its users. This non-functional requirement states the system should show and require from users only information really relevant to the task being accomplished. More details will be given in Sect. IV.

Part of the information about environments in which organisms are collected deals with geographical references (latitude and longitude). We used an open map server called

OpenStreetMap [2], whose license is *CC-BY-SA* [3]. The classification functionality was implemented using Weka [22] classes.

IV. Results and Discussion

Among the vast amount of characteristics, a descriptive statistics of data deposited at DB-I until April 30, showed some important distributions. Data on DB-I were represented in three major Fila, Chordata (63.7%), Arthropoda (32.6%) and Gimnosperms (3.7%). The samples' distribution showed that almost 90.0% were represented by two phylogenetic Classes: 57.7% Actinopterygii and 32.1% Insecta (Fig. 4). These Classes were sampled in seven environmental types, landscape influenced by rivers (*igarapé*), river, natural lake, inundation forests (*várzea*), transition land between dry land and *igarapé* (*igapó*), interfluvial field (*campo interfluvial*) and natural breeding (*criadouro natural*). However the first two constituted more than 75.0% of all samples (Fig. 5). Another relevant datum at DB-I is the water type, which showed the following distribution: black-water (*água preta*) 66.9%, clear-water (*água clara*) 18.3% and white water (*água branca*) 14.8%. All samplings were performed in four Brazilian States and more than 80.0% were sampled in five cities in the North region: Cantá (Roraima state) with 26.6% of total samples, followed by Santarém (Pará state) with 15.8%, Manaus (Amazonas state) with 14.6%, Barcelos (Amazonas state) with 13.9%, and Boa Vista (Roraima state) with 9.8% (Fig. 6). These data show a great collection of various species, from different locations and sorts of water.

Some environmental characteristics were assessed, such as pH, temperature and dissolved oxygen concentration (DO). According to the histograms an initial analysis can be done. The pH (Fig. 7) showed a range of 3.0 to 8.0, with 72.0% of all samples falling between pH 5.8 and 7.0. The water temperature (Fig. 8) varied between 16 and 34ºC, with 97.0% between 26.0 and 34.0º C. Finally the DO (Fig. 9) ranged from 0.0 to 7.5 ppm, with 68.0% between 5.0 and 7.5 ppm.

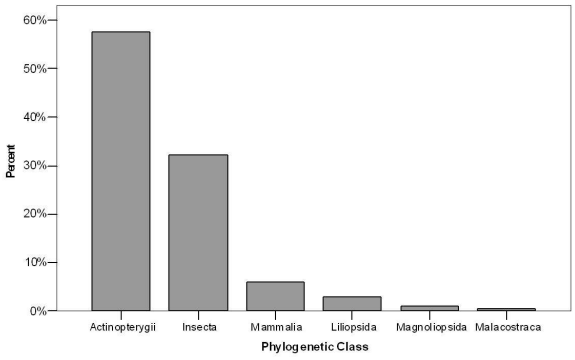

Figure 4. Phylogenetic distribution at class level for samples deposited at DB-I (percentage).

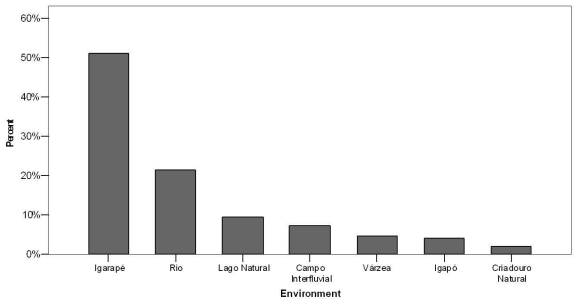

Figure 5. Samples distribution for each Amazonian environmental type. *Igarapé* (landscape influenced by rivers), *Rio* (river), *Lago Natural* (natural lake), *Campo Interfluvial* (interfluvial field), *Várzea* (inundation forests), *Igapó* (transition land), and *Criadouro Natural* (natural breeding).

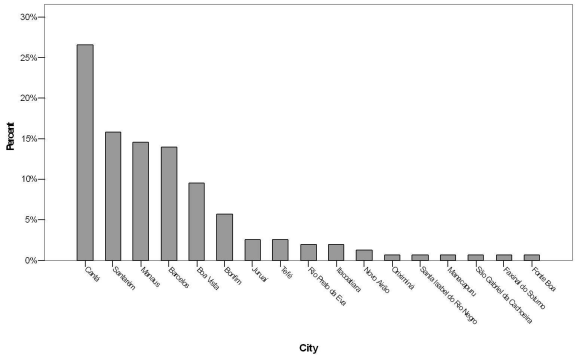

Figure 6. Samples distribution for Brasilian cities in four States. Barcelos, Fonte Boa, Itacoatiara, Juruá, Manacapuru, Manaus, Novo Airão, Rio Preto da Eva, Santa Isabel do Rio Negro, São Gabriel da Cachoeira, Tefé (Amazonas state), Oriximiná, Santarém (Pará), Faxinal do Soturno (Rio Grande do Sul), Boa Vista, Bonfim, Cantá (Roraima).

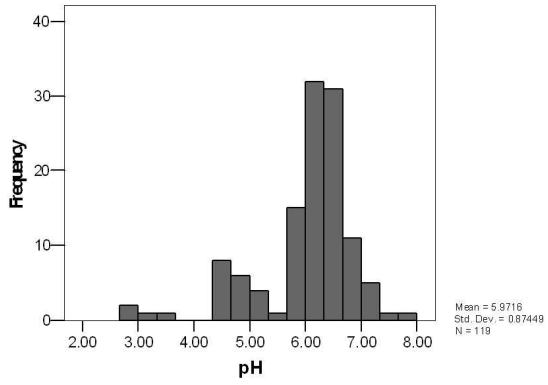

Figure 7. Histogram depicting the frequency of pH measurements from 119 samples of DB-I. Mean equals 5.97 and standard deviation equals 0.87.

2 http://www.openstreetmap.org/

3 http://creativecommons.org/licenses/by-sa/2.0/

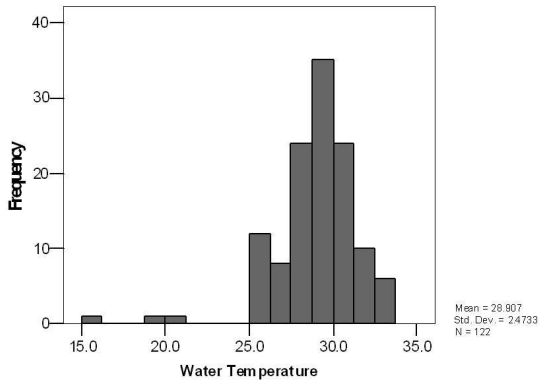

Figure 8. Histogram of water temperature (Celsius degrees) from 122 samples of DB-I. Mean equals 28.91°C and standard deviation equals 2.47°C.

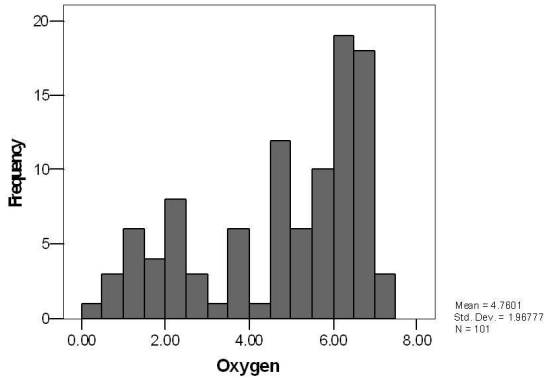

Figure 9. Histogram for percentage of dissolved oxygen levels from 101 samples of DB-I. Mean equals 4.76 and standard deviation equals 1.97.

Nowadays, DB-I is being used by researchers of ADAPTA. The first set of functionalities enables users to list, add, delete and edit a collection. Examples of data to be entered for a collection are: date and time, location, environmental conditions, etc. The system also allows users to list, add, delete and edit experiments. An experiment is always associated with one individual (or group of individuals, if all share the same characteristics) of one and only one collection.

The third set of functionalities, perhaps the most important one, because it will be the link between DB-I and DB-II, was implemented in order to allow users to perform the following analyses: which individuals were collected under certain environment conditions (water temperature, O_2 concentration, etc.)?; which collections contain specific individuals, based on taxonomy? The user enters the criteria and receives the result, from which it is possible to view and edit the individual or collection.

Users must authenticate in the system before they can use it. This helps the system control which information to show to whom, i.e., users are allowed to see the information from the laboratories of which they are part. This behavior also allows researchers of the same laboratory to work in a collaborative way.

As mentioned previously, the software shows context-specific behavior. For example, suppose a user is part of three different laboratories. When that user is filling in a collection form, he or she will have to choose to which laboratory that collection is related. If no laboratory has been provided yet, the fields related to the person responsible for the collection are disabled. When the user wants to fill in the associated laboratory, only his or her laboratories will be shown as options. Once the user provides the associated laboratory, the fields related to the responsible person become available. When the user tries to fill in the latter fields, only those people who are part of the chosen laboratory will be shown as options.

Another example of context-specific behavior is related to the fields shown to the user. For example, for a given collection, the environment can be natural or captive. If the associated environment is natural, then only fields relevant to that kind of environment are displayed; the same holds to captive environments. There is a third kind of environment (microcosm) that is actually represented by an experiment (experimental environment).

When a user wants to add an organism to the collection, that organism will be in one of the several possible groups of organisms. Again, only relevant fields are shown, depending on the group chosen. One example would be the possible species from which to choose. Another example: several groups of organisms have anesthetic as a relevant field (fish, for example), but they might use different kinds of anesthetics. Only those kinds of anesthetics that are relevant for the chosen group are shown as options. On the other hand, some organisms, like plants, have no fields related to anesthetics, since these don't apply to plants.

When adding a collection, users are also able to associate a bibliographic reference to it. If a certain collection is part of the logbook of some researcher, anyone who wants to consult that collection and to know more about it can get the logbook reference present in the form and look for more information in the actual logbook.

Another important piece of information that may be provided when adding an organism to a collection is whether RNA from the organism was collected or not and if so, where the samples are stored.

Organisms may be added to a collection in groups, i.e., if organisms share the same characteristics, users may add them all at the same time by just providing the number of organisms associated with that entry. Each organism must be associated with a species. However, since sometimes researchers are uncertain of which species a specific organism is, or that information is not always relevant, users may associate an organism with one of the following levels in the taxonomic hierarchy: class, order, family, genus, or species.

The application built also provides some statistics of the activity of the laboratories. For example, project administrators are able to check how many collections each of the laboratories has and how many organisms it has collected.

Concerning the fourth set of functionalities, if required, the system reads information from collections stored in the

database, defines a set of examples based on that information and builds a tree classifier by a call to *J48*'s *buildClassifier()* method. Since ADAPTA is still under progress, we are not allowed to publish classifiers from actual data. However, for sake of completeness we show examples of output using fictitious data in Figs. 10-12. In Fig. 10 the class is the hydrological cycle of the environment of a collection at the time it was performed, and the attributes are: water temperature, CO_2 concentration, O_2 concentration, pH, conductivity and weather condition. In the example, the system builds a classification model to predict the hydrological cycle based on those attributes. In Fig. 11, the class is the presence of anthropogenic interventions (true or false) in the environment of the collection and the attributes are also physico-chemical measures. The resulting model predicts if there are anthropogenic interventions in the environment based on the attributes. In Fig. 12, the class is size of a certain fish species (≤ 9cm or > 9cm) and the model bases its decision on the presence of anthropogenic interventions (T for True or F for False) and physico-chemical measures. Database tables store parameters used to generate the models, the models themselves and some metrics of performance assessment [23].

The implementation of the fifth functionality allows biologists to visualize all collections (relevant to the current user of the system) as points in a map, since the system reads the latitude and longitude information of the collections from the database (see Fig. 13 for an example).

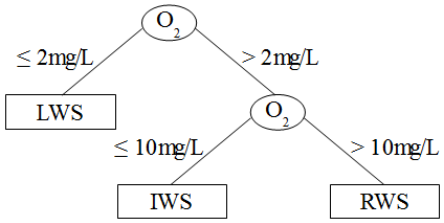

Figure 10. Tree model induced by J48 inducer. O_2 represents water oxygen concentration and LWS, IWS and RWS represent the following values for hydrological cycle, respectively: Low water season, Increasing water level season and Receding water season. This is just a fictitious example.

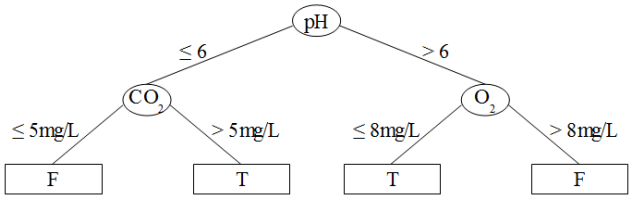

Figure 11. This tree was built based on physico-chemical attributes (water temperature, pH, O_2 concentration, CO_2 concentration, etc.) of collections and predicts if the environment in which the collections were performed presented anthropogenic interventions (T means True) or not (F means False). The model verifies if the physico-chemical measures change when anthropogenic interventions are present.

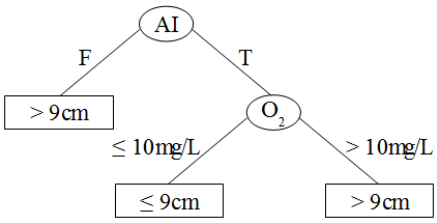

Figure 12. This tree was built based on the presence of anthropogenic interventions (T for True and F for False) and physico-chemical attributes (water temperature, pH, O_2 concentration, CO_2 concentration, etc.) of collections and predicts if the size of a certain fish species is less than or equal to 9cm or greater than 9cm. The model verifies if those attributes influence the size of the fish.

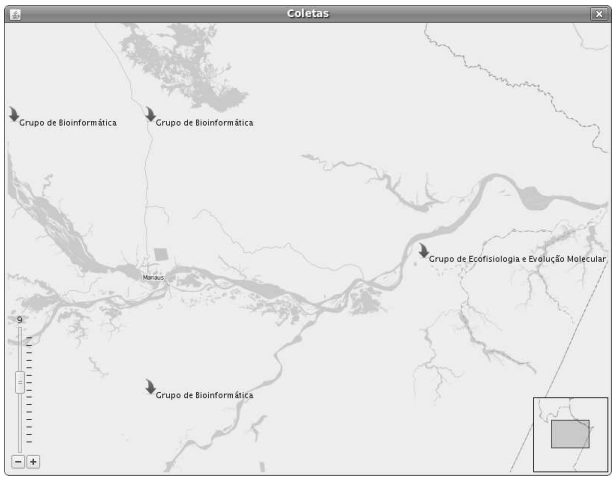

Figure 13. Map showing the places where collections were performed.

V. CONCLUSIONS

The Amazon provides humans and nature in general with a rich set of resources, a great part of which hasn't been discovered yet. ADAPTA is part of the initiatives created to help Brazil's government make good use of its piece of the Amazon, and it will probably be of great value to all of those involved with the forest, no matter where they are from. Understanding how organisms are able to adapt to difficult conditions or how organisms take notice of those conditions and migrate is going to be of key importance to improve effectiveness of saving endangered species, produce effective environmental policies, and produce biological knowledge.

Building a database tool for ADAPTA has not been so trivial, given the complexity of the system and the amount of data and information produced by the project. Nonetheless, manipulation and analysis of all that data need the assistance of computer tools. As shown in Sect. II, there are many studies working with similar objectives, but they all have very specific needs. This project, although dealing with integration of information systems and data mining, is also dealing with specific needs of ADAPTA. It has been a challenge too to put together needs from all the several laboratories involved in the project and to choose only those needs that are relevant to the project as a whole.

One good aspect of this implementation is that it has been making use of open standards. Although this is not necessarily a better approach in terms of the quality of the results, because commercial and open technology can both offer very nice solutions, it will probably facilitate continuation of the development of tools in the future, at least considering the financial cost of the implementation.

The system implemented presents some nice features such as a context-specific behavior, friendly and simple graphical user interface, whereas access is controlled by user profiles. The data mining and georeferencing features are very useful extra tools for the researchers of the project.

ACKNOWLEDGMENT

This work was funded by a joint grant between National Research Council of Brazil (CNPq) and Amazon State Research Foundation (FAPEAM) through the Program National Institutes of Science and Technology (INCT) to the ADAPTA Project (Centre for Studies of Adaptations of Aquatic Biota of the Amazon) coordinated by Adalberto Luis Val. This work was also funded by Sao Paulo Research Foundation (FAPESP).

REFERENCES

[1] R. Mcconnell and R. H. Lowe-Mcconnell, Ecological studies in tropical fish communities, Cambridge Univ Pr, 1987.

[2] C. Araujo-Lima and M. Goulding, So fruitful a fish: Ecology, conservation, and aquaculture of the Amazon's tambaqui, Columbia Univ Pr, 1997.

[3] A. L. Val et al. National Institute of Science and Technology of the Adaptations of Aquatic Biota of the Amazon, 2008.

[4] R. Frankham, D. D. Ballou and D. A. Briscoe, Introduction to Conservation Genetics. Cambridge UK: Cambridge University Press, 2002.

[5] F. W. Allendorf and G. Luikart, Conservation and the genetics of populations. Wiley-Blackwell, 2007.

[6] C. Uhl et al., "Natural resource management in the Brazilian Amazon," Bioscience, vol. 47, n. 3, pp. 160-168, 1997.

[7] J. H. J. Schaminée, S. M. Hennekens and W. A. Ozinga, "Use of the ecological information system SynBioSys for the analysis of large datasets," Journal of Vegetation Science, vol. 18, n. 4, pp. 463-470, 2007.

[8] M. G. Le Duc, L. Yang and R. H. Marrs, "A database application for long-term ecological field experiments," Journal of Vegetation Science, vol. 18, n. 4, pp. 509-516, 2007.

[9] J. Peacock, T. R. Baker, S. L. Lewis, G. Lopez-Gonzalez and O. L. Phillips, "The RAINFOR database: monitoring forest biomass and dynamics," Journal of Vegetation Science, vol. 18, n. 4, pp. 535-542, 2007.

[10] H. I. I. Lee et al., "Integrated monitoring and information systems for managing aquatic invasive species in a changing climate," Conservation Biology, vol. 22, n. 3, pp. 575-584, 2008.

[11] Y. Sucaet, J. Van Hemert, B. Tucker and L. Bartholomay, "A web-based relational database for monitoring and analyzing mosquito population dynamics," Journal of Medical Entomology, vol. 45, n. 4, pp. 775-784, 2008.

[12] J. T. Pollettini et al., "A software architecture-based framework supporting suggestion of medical surveillance level from classification of electronic patient records," The 12th IEEE International Conference on Computational Science and Engineering. 2009. 8p.

[13] D. P. Mckay, T. W. Finin and A. O'hare, "The intelligent database interface: Integrating AI and database systems," Proceedings of the 1990 National Conference on Artificial Intelligence. 1990. pp. 677-684.

[14] J. G. Wolff, "Towards an intelligent database system founded on the SP theory of computing and cognition," Data & Knowledge Engineering, vol. 60, n. 3, pp. 596-624, 2007.

[15] C. Kingsford and S. Salzberg, "What are decision trees?," Nature Biotechnology, vol. 26, n. 9, pp. 1011-1013, 2008.

[16] J. R. Quinlan, C4.5: Programs for machine learning. Morgan Kaufmann, 1993.

[17] P. Domingos, "MetaCost: A general method for making classifiers cost-sensitive," Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1999. p. 155-164.

[18] D. Gamberger, N. Lavrač, F. Zelezný and J. Tolar, "Induction of comprehensible models for gene expression datasets by subgroup discovery methodology," Journal of Biomedical Informatics, vol. 37, n. 4, pp. 269-284, 2004.

[19] N. Rosenfeld et al., "MicroRNAs accurately identify cancer tissue origin," Nature biotechnology, vol. 26, n. 4, pp. 462-469, 2008.

[20] J. A. Baranauskas and M. C. Monard, "Combining symbolic classifiers from multiple inducers," Knowledge-Based Systems, vol. 16, n. 3, pp. 129-136, 2003.

[21] O. P. Netto, S. R. Nozawa, R. A. R. Mitrowsky, A. A. Macedo and J. A. Baranauskas, "Applying decision trees to gene expression data from DNA microarrays: a leukemia case study," In XXX Congress of the Brazilian Computer Society, X Workshop on Medical Informatics, 10p., Belo Horizonte, MG, 2010, ISSN 2175-2761.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The Weka Data Mining Software: an update," SIGKDD Explorations, vol. 11, n. 1, pp. 10-18, 2009.

[23] J. Demšar, "Statistical comparison of classifiers over multiple data sets," Journal of Machine Learning Research, vol. 7, n. 1, pp. 1-30, 2006.