



ESF - Science Meeting - Final Report

In Confidence

This form must be signed and returned to the ESF.

APPLICATION DATA

Page 1/2

SCIENCE MEETING

Reference Number : **972**
Report submitted : **18/12/2006 17:19:24**

ESF ACTIVITY

Unit(s) : **LESC**
Activity Title : **Integrating population genetics and conservation biology: Merging theoretical, experimental and applied approaches**
Activity Acronym : **CONGEN**

PROJECT

Science Meeting : **Workshop**
Title of Science Meeting : **DNA Barcode Data Analysis**
Location : **Paris**
Date of Science Meeting : **23/05/2006 - 25/05/2006**
Convenor Name : **Dr. Michel Veuille, Paris, France**

BUDGET

Total estimated Expenditure : **63381 €**
ESF Grant requested : **19587 €**
Co-sponsorship Income : **43793 €**

BUDGET GRANTED

ESF Grant FUNDING : **15000 €**

ACTUAL EXPENDITURE

Travel : **3860 €**
Accommodation : **0 €**
Meals (lunch and dinner) : **1520 €**
Local administrative costs * : **306 €**
TOTAL EXPENDITURE : 5686 €

* includes: administrative and technical assistance, printing, photocopying, telephone, fax, email, etc. Additional support for schools may be considered

Report to ESF
Data analysis of barcode data
Muséum National d'Histoire Naturelle, Paris, 6-8 July 2006

Abstract

The variation within and the divergence among species and subspecies are critical variables in conservation biology, and they can be of legal importance in the designation of threatened and protected species. "DNA barcodes" are standardized molecular markers that are proving useful as diagnostic characters for species-level taxonomy. This project engages young population geneticists, statisticians, and computer scientists in an international effort to develop analytical methods and protocols for the treatment of these data.

Taxonomy is a critical component of conservation biology. The ability to delineate species and subspecies is fundamental to the study of conservation and to regulatory efforts to protect endangered taxa or identify rare species. Morphological criteria have been the traditional way of differentiating species and assigning specimens to them, but genetic data have played an increasingly important role. In the past two years, a series of studies have been published in which "DNA barcoding" was proposed as a tool for differentiating species. Barcoding is based on short gene regions evolve at a rate that produces clear interspecific sequence divergence while retaining low intraspecific sequence variability. The *Cytochrome c oxidase subunit 1* mitochondrial gene ("COI") has emerged as a suitable barcode region for most animal groups. Taxonomists are in the process of identifying appropriate gene regions for barcoding other major groups of eukaryotes. Taxonomic studies of a growing number of taxa have shown that the discontinuity in the levels of barcode sequence divergence (both phenetic and diagnostic) match the species boundaries as delineated by morphological and ecological characters.

The overall goals of this project are:

1. To create a small international research community of population geneticists, statisticians, applied mathematicians, computer scientists and taxonomists that concentrates on DNA barcode data for two years;
2. To support the participation of doctoral students, post-docs and young investigators from Europe and North America in this research community;
3. To identify the types of analytical, interpretive and display tools needed for the optimal treatment of DNA barcode data; and
3. To develop and disseminate the most effective analytical procedures and display tools for barcode data.

DNA barcodes are relevant to conservation genetics for several reasons. DNA barcodes are proving useful as a cost-effective system for assigning specimens to their correct species, especially in cases where morphological characters are insufficient. Barcode data provide a way to easily identify species from a wide spectrum of living organisms in ecological studies, even in the hands of non-professional taxonomists. They can easily be linked to a database giving the biological parameters and the conservation status of each taxon. This approach is being used to identify invasive and pest species, and products made from protected species (e.g., bushmeat).

The Data Analysis Working Group (DAWG) of the Consortium for the Barcode of Life (CBOL) held a 2 ½ day workshop hosted by the Museum National d'Histoire Naturelle (MNHN) in Paris on 6-8 July 2006. Thirty-eight participants from 10 countries attended, the majority of whom were doctoral students, postdoctoral fellows or young researchers.

The overall goal of DAWG is to develop protocols, techniques and software that the barcoding community can use to sample, analyze, interpret and display barcode data. The purpose of the Paris workshop was to allow presenters to describe their preliminary results

and plans for the coming year, and to receive feedback from the other workshop participants. They will continue their work with the goal of presenting finished results at an international conference in June 2007. The final results of their work will be published in a proceedings volume of the June 2007 conference, and their protocols and software will be made available on a Data Portal being developed by CBOL.

Workshop structure and content

The workshop began with three introductory presentations: M. Veuille greeted participants; D. Schindel described the workshop's goals; and V. Loeschke and K. Bijlsma described the European Science Foundation's Conservation Genetics Programme. The balance of the workshop was devoted to presentations of preliminary results by 15 participants who had submitted abstracts. Each presentation lasted for 30 minutes, after which all participants engaged in open discussion.

The presentations included five categories of techniques, and many presenters used techniques from several categories and compared their effectiveness. The five categories are:

1. Character-based classifications. A number of techniques and of computer programs are available for classifying objects, in a way that is not limited to biological species. They generally rely on ways to partition sets into subsets based on shared properties (Classification and Regression Trees, CART, is one such approach presented at the workshop). In systematics, so-called "informative characters", as used in cladistics, belong to this category. Since the barcode is not concerned with phylogeny, a simplified form of this approach is used by Character Attribute Organization System (CAOS, also presented at the workshop). However, homoplasy and the segregation of ancestral polymorphism limit the use of this approach in closely related species, which is the level of differentiation that matter the most in barcoding.

Phylogenetic analysis also uses gene sequence data analyzed as a series of discrete attributes. CBOL has stressed that barcode data, by themselves, are inadequate bases on which to reconstruct phylogenetic relationships. However, phylogenetic methods can be used to determine affinities among specimens and between specimens and known taxonomic categories (at the species level and higher in the taxonomic hierarchy). These methods use a variety of parsimony algorithms to build trees.

2. Distance-based clustering methods. When there is no simple discriminating character between species, distance based clustering methods can be used. The most popular method in the barcode community appears to be neighbor-joining (NJ), an algorithm starting from the most closely related clusters of sequences, and then proceeding stepwise to the rest of the sample. It is generally calculated using the K2P distance (Kimura 2-parameter model), the simplest way to deal with nucleotide change when there are very different mutation rates in transitions and transversions, as is the case in mtDNA. The accuracy of these methods matters only for recent nodes, since barcoding is mostly interested in identifying species. This method of "clustering" sequences does not provide a tree of species, but a tree of genes.

3. Coalescent theory. Coalescent theory provides a tool for studying the ancestry of a sample of sequences by looking backwards in time. Contrary to phylogenetic methods, which are based on parsimony principles or on assumptions of the constancy of evolutionary rates (the "molecular clock"), the coalescent theory is based on our present understanding of the actual mechanism of evolutionary change within species. Models based on the Coalescent theory include parameters that represent forces such as random drift and natural selection. Coalescent theory lends itself easily to computer simulations, allowing one to run a series of simulations (classically between 1,000 and 10,000) to assess the probability of an assumption leading to the observed state of the dataset. Its applications are not limited to the classical mutation-drift

equilibrium neutral model. It is thus possible to explore the parameter space along individual axes (e.g., panmixia vs. population structuring, changes vs. constancy in population size). When there is no diagnostic character that separates species, it may be counterintuitive to obtain a result in the form of a probability of an accession belonging to some species. However, such outputs may be useful in further research. For instance, they may also allow one to estimate the optimum sample size, based on prior information and assuming some population model. Applications of coalescent theory may thus be intervening steps in a research protocol.

4. Bayesian statistics and maximum likelihood. These are statistical methods that can be used in a wide range of statistical applications, including in applications referred to above, such as coalescent theory. They are very powerful, but their use assumes some preliminary knowledge on the model being applied (Maximum Likelihood), or of the distribution of one of the parameters given some knowledge on another one (Bayesian). The main difficulty with these methods is their high computation time. A minor problem is that it is generally difficult to say what character is the cause of the distinction between two species, which is always counterintuitive. ABC methods (referred to in the meeting) are much less demanding in computer time.

5. Miscellaneous points. As the barcode dataset grows larger, it may be difficult to identify the reference sequences closest to a query sequence. This question was addressed at the meeting by the proposal to use the Google search engine, and by another aiming to identify the sister-clade of some query at the appropriate taxonomic level. Two groups (working with CART and the coalescent respectively) have identified an error in the *Astraptes* dataset.

Meeting results

In addition to providing the presenters with feedback on their preliminary results, the workshop participants agreed on the need to:

- Develop standard methods for comparing results of competing techniques (e.g., common sample sizes, effective population sizes, mutation rates, other population genetics parameters). Javier Cabrera agreed to develop a draft standard for comment by the workshop participants. Provide additional online datasets with different characteristics and smaller minimum sample sizes.
- Develop consensus recommendations to the barcoding community concerning: - Adequate sample sizes. Many presenters had recommendations on sample sizes and DAWG will need a mechanism to compile them, promote comparison, and facilitate discussion leading to a consensus. - Standard treatment and presentation of cluster diagrams. Many presenters showed cluster diagrams with a variety of filters on branch nodes based on bootstrapping. DAWG could provide a valuable service by developing recommendations to the barcoding community on standard presentations. - Standard vocabulary and usage of statistical terms in discussions of barcode data (e.g., accuracy, precision, error rates, false positives/negatives).
- Identify and engage specialists in data visualizations and display. Several participants mentioned software programs that might be applicable to barcode data, and visualization specialists who might be interested.
- Determine the best way to disseminate the results of the DAWG initiative. In addition to posting software and protocols on the BOLI Data Portal being developed by CBOL, there will be a proceedings volume based on the Second International Barcode Conference. Participants discussed whether it would be best to publish data analysis papers in the proceedings volume or in another journal, such as *Systematic Biology*. The Steering Committee needs to facilitate this discussion and promote a consensus.

| Thursday 6 July 2006 | | |
|--|---|--|
| Opening session – Chair : Brian Golding | | |
| 14:00 | David SCHINDEL - Secretary of the CBOL | Welcoming address |
| 14 :15 | Michel VEUILLE - Chair of the DAWG | Opening of the meeting |
| 14:30 | Voelker LOESCHKE - ESF | The CON-GEN program |
| 15:00 | José M. BAUTISTA - FishTrace consortium / Complutense University of Madrid, Spain | Fish barcoding from the FishTrace database: the control gene, the data validation analysis and the backup reference biological data |
| 15:45 | Coffee break | |
| 16:15 | Mehrdad HAJIBABAEI - University of Guelph, Canada | Google Gene: searching for DNA barcode sequences using Google search engine |
| 17:00 | Group visit of the vertebrate collections | |
| Friday 7 July 2006 | | |
| Chair : Donal Hickey | | |
| 10:00 | Michael J. HICKERSON - University of California, Berkeley, Museum of Vertebrate Zoology, USA | Quantifying uncertainty in species discovery with approximate Bayesian computation (ABC): single samples and recent radiations |
| 10:45 | Kasper MUNCH - University of Copenhagen, Denmark | Bayesian DNA barcoding |
| 11:30 | Coffee break | |
| 12:00 | Eric BAZIIN - University of Montpellier II, France | MtDNA variation and effective population size |
| 12: 45 | Lunch | |
| 13:45 | Bogdan PASANIUC - University of Connecticut, USA | DNA Barcode Data Analysis: Boosting Assignment Accuracy by Combining Distance- and Character-Based Classifiers |
| 14:30 | Frederic AUSTERLITZ - Ecologie, Systématique et Evolution, Orsay, France | Comparing phylogenetic and statistical classification methods for DNA barcoding |
| 15:15 | Coffee break | |
| 15:30 | Indra Neil SARKAR - American Museum of Natural History, USA | Automated Barcoding Using the Characteristic Attribute Organization System |
| 16:15 | Zaid ABDO - Department of Biology, McMaster University, Canada | A step towards barcoding life I: A new method to assign genes to preexisting species groups |
| 17:00 | Group visit of the arthropod and insect collections of the MNHN with the curators | |
| Saturday 8 July 2006 | | |
| Chair : David Schindel | | |
| 10:00 | Jessica RACH - TiHo Hannover, ITZ Ecology & Evolution, Germany | Character-based DNA barcoding for identifying conservation units in Odonata |
| 10:45 | Damon LITTLE - The New York Botanical Garden, USA | A comparison of algorithms for identification of specimens using DNA barcodes: examples from gymnosperms |
| 1130 | Coffee break | |
| 11:45 | Birgit GEMEINHOLZER - Botanic Garden and Botanical Museum Berlin-Dahlem, Germany | Possibilities and limitations of sequence similarity and homology search tools implemented in molecular nucleotide databases for organism identification |
| 12:30 | Lunch | |
| 13:30 | Donal HICKEY - Concordia University, Canada | DNA Barcoding of Fungi: a Feasibility Analysis |
| 14:15 | Javier CABRERA – Department of statistics, Rutgers University, USA | An MLE-based clustering method on DNA barcode |
| 15:00 | Discussion | |
| 15:45 | Coffee break | |
| 16:15 | Organization and agenda of the DAWG – closure at 17:00 | |

Each talk is 30 min long, plus 15 minutes of discussion