Comparing phylogenetic and statistical classification methods for DNA barcoding

Frédéric Austerlitz, Olivier David, Brigitte Schaeffer, Michel Veuille, Catherine Laredo

Workshop Report: Data Analysis Working Group
Consortium for the Barcode of Life
Muséum National d'Histoire Naturelle, Paris, 6-8 July 2006

Michel Veuille[1], Javier Cabrera[2] and David E. Schindel[3]

Introduction

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 Nationale d'Histoire Naturelle (MNHN) in Paris on 6-8 July 2006 (see Appendix 1: Call for Participation). Thirty-eight participants from 10 countries attended (see Appendix 2: List of Participants), 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 for the Barcode of Life Initiative (BOLI).

Acknowledgments

Funding for European participants was provided by a grant from the Conservation Genetics Programme of the European Science Foundation. Support for American participants was provided by the Division of Biological Infrastructure, the Office of International Science and Engineering, and the Division of Information and Intelligent Systems of the National Science Foundation. Financial and in-kind support was also provided by the Muséum National d'Histoire Naturelle, Paris, Ecole pratique des hautes etudes (EPHE), the Center for Discrete Mathematics and Theoretical Computer Sciences (DIMACS), Rutgers University, the Alfred P. Sloan Foundation, and CBOL.

Background

CBOL's Executive Committee created DAWG in late 2004 for the purpose of developing protocols, methods, and software for the analysis, interpretation and display of barcode data. Michel Veuille was asked to chair DAWG and DIMACS was invited to be a principal partner in the Working Group. DAWG met for the first time at the First International Barcode Conference at the Natural History Museum, London, on 9 February 2005. Planning meetings were held at DIMACS on 26 September 2005 and at MNHN on 15 October 2005. A Steering Committee[4] was formed at this second planning meeting.

[1] Chairman, Département de Systématique et Evolution, Muséum National d'Histoire Naturelle, Paris
[2] Department of Statistics and DIMACS, Rutgers University, Piscataway, NJ
[3] Executive Secretary, Consortium for the Barcode of Life, National Museum of Natural History, Smithsonian Institution, Washington, DC
[4] The DAWG Steering Committee includes M. Veuille, Chair (MNHN); Javier Cabrera (DIMACS); Rob DeSalle (American Museum of Natural History); Brian Golding (McMaster University); D. Hickey (Concordia University); and D. Schindel (CBOL);

Based on discussions during the two planning meetings and interactions with CBOL, the Steering Committee formulated a Program of Work[5] whose goals are to catalyze development of the new techniques and tools that will be needed to analyze, interpret and display barcode data. The Program of work included a workshop at which preliminary results would be presented and discussed, and presentation of final results at the Second International Barcode Conference in 2007. The Committee issued a Call for Participation (Appendix 1) to statisticians, computer scientists, taxonomists and others interested in data analysis. The Call included a set of Technical Challenges that were developed by participants in the two planning meetings. The Steering Committee selected 15 abstracts for presentation at the workshop (see Appendix 3, agenda; Appendix 4, abstracts of presentations).

Workshop Structure and Content[6]

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. Table 1 indicates the techniques and datasets used in each study. Appendix 4 contains the abstracts submitted by the presenters and Appendix 5 presents brief summaries of each presentation.

The presentations included five categories of techniques, and many presenters used techniques from several categories and compared their effectiveness (see Table 1). 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

[5] DAWG Program of Work is posted at http://barcoding.si.edu/PDF/Program%20of%20Work%20-%20DAWG%20-%20FINAL.pdf

[6] See meeting agenda, Appendix 4. Presentations linked to the agenda are available at http://www.barcoding.si.edu/DAWG_Paris_Workshop.html

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.

Next Steps

The DAWG Steering Committee agreed to use the NBII Portal as a platform for sharing information and conducting electronic discussions of the issues described above. CBOL will probably call for submission of proposals for sessions at the Second International Barcode Conference around October 2006. The Committee will apply for a half-day session on data analysis. A Call for submission of abstracts will probably be published in December 2006 or January 2007.

Table 1. Classification of presentations according to techniques

	Simulation/ Coalescent Model	Clustering	Character-based	Search algorithms	Bayesian Analysis	Phylogenetic	Datasets
Bautista		X					European fish
Hajibabaei		X	X	X			illustrations with primates and Lepidoptera; no analysis
Hickerson	X				X		simulations and marine snails
Munch					X	X	plants
Bazin							broad data compilation
Pasaniuc		X	X				DIMACS test data, cowries
Austerlitz	X	X	X			X	simulated data, *Litoria*, cowries, *Atraptes*
Sarkar		X	X				*Mopalia*
Barraclough	X						Australian tiger beetles, rotifers, land plants
Abdo	X				X		*Astraptes*, simulated data
Rach		X	X				dragonflies, ND2 and COI
Little		X	X	X			cycads, nuclear, plastid, mitochondrial regions; DAWG training set
Gemeinholzer				X			Asteraceae, ITS region
Hickey		X					fungi, various gene regions
Cabrera (for Ching Ray Yu)		X					DAWG training set
Cabrera-Lo							

APPENDIX 1 : Call for Participation

Data Analysis Challenges Arising from the DNA Barcode Initiative

The Challenge: The Data Analysis Working Group (DAWG) of the Consortium for the Barcode of Life (CBOL) has developed interdisciplinary research challenge problems in statistics and computer science arising from DNA barcoding, a method proposed as a tool for differentiating species. Students, postdocs, and researchers from all over the world are challenged to develop new approaches to these problems. Compelling solutions to these challenges will require collaboration among taxonomists, population geneticists, and evolutionary and systematic biologists, so DAWG encourages the formation of multidisciplinary teams.

Presenting Preliminary Ideas at a Workshop in Paris: Preliminary ideas for approaches to these problems will be discussed at a workshop at the National Museum of Natural History in Paris on 6-8 July 2006 (see http://dimacs.rutgers.edu/Workshops/DNABarcode/). Participation in this workshop will be limited to approximately 40 presenters of preliminary results and attendees who can offer useful feedback to the presenters. Space will therefore be limited and all those wishing to participate in the workshop should register at http://dimacs.rutgers.edu/Workshops/DNABarcode/registnew.html no later than 29 June 2006. However, you are urged to register early as we will close registration when all spaces are filled.

Travel awards for a limited number of Europeans who would like to give presentations at this workshop will be available through funding from the Conservation Genetics Programme of the European Science Foundation. Travel awards for US presenters will also be available, pending funding agency approval. Travel support will focus primarily on increasing the participation of students, postdocs and junior faculty.

Presenting More Advanced Results at a Conference in Southeast Asia: The preliminary workshop will be followed by an international conference in southeast Asia in February 2007, during which the most promising approaches to these challenge problems will be presented. Travel awards will also be available (pending funding agency approval).

For the full Call for Participation, including the statement of the research challenges, see: http://dimacs.rutgers.edu/Workshops/BarcodeResearchChallenges/.

For instructions on how to submit an abstract for the Paris workshop, see http://dimacs.rutgers.edu/Workshops/DNABarcode/abstractsubmissionform.html.

To apply for travel funds to give a presentation at the Paris workshop, see http://dimacs.rutgers.edu/Workshops/DNABarcode/travelsupport.html

To register for the workshop, see http://dimacs.rutgers.edu/Workshops/DNABarcode/registnew.html

For information about the DNA Barcode Initiative, see: http://dimacs.rutgers.edu/Workshops/DNAInitiative/.

Important dates:

Deadline for submission of abstracts: 2 June 2006

Deadline for submission of requests for travel support: 2 June 2006

Deadline for registration: 29 June 2006

Announcement of final agenda of presenters, awards of travel support: as early as possible after 2 June 2006

APPENDIX 2: PARTICIPANT LIST

Name	Title of Abstract (if submitting)	Institution	Country	Phone	Email
Zaid Abdo	A step towards barcoding life I: A new method to assign genes to preexisting species groups	Department of Biology, McMaster University	CANADA		zabdo@uidaho.edu
Marcella Attimonelli		Dipartimento di Biochimica e Biologia Molecolare	ITALY	Phone: 390805442399	m.attimonelli@biologia.uniba.it
Frederic Austerlitz	Comparing phylogenetic and statistical classification methods for DNA barcoding	CNRS - Universite Paris-Sud, laboratoire Ecologie, Systematique et Evolution	FRANCE	Phone: +33169157720	Frederic.Austerlitz@ese.u-psud.fr
Tim Barraclough	Biological inferences from barcoding data	Imperial College London, Biology	UK	Phone: 020 7594 2247	t.barraclough@imperial.ac.uk
José M. Bautista	Fish barcoding from the FishTrace database: the control gene, the data validation analysis and the backup reference biological data	Universidad Complutense de Madrid, Biochemistry and Molecular Biology IV	SPAIN	Phone: +34913943823	jmbau@vet.ucm.es
Eric Bazin	MtDNA variation and effective population size	University of Montpelier 2	France		bazin@univ-montp2.fr
Kuke R. Bijlsma	The Conservation Genetics Program, European Science Foundation	Evolutionary Genetics Group, University of Groningen	Netherlands		
Javier F Cabrera		Rutgers University, Statistics Department	USA	Phone: 732-4852537	cabrera@rci.rutgers.edu
Gerard Delvare		CIRAD	FRANCE	Phone: 33 (0) 4 67 59 31 20	Gerard.delvare@cirad.fr
Birgit Gemeinholzer	Possibilities and limitations of sequence similarity and homology search tools implemented in molecular nucleotide databases for organism identification	Botanic Garden and Botanical Museum Berlin-Dahlem	GERMANY		b.gemeinholzer@bgbm.org
Sylvain Glémin		University of Montpelier 2	France	+33 (0) 4 67 14 46 84	glemin@univ-montp2.fr
G. Brian Golding		McMaster University, Biology	CANADA	Phone: 905-525-9140	Golding@McMaster.CA
Heike Hadrys		TiHo Hannover ITZ Ecology & Evolution	Germany	Phone +49 511 953 8880 Fax: +49 511 953 8584	heike.hadrys@ecolevol.de

Mehrdad Hajibabaei	Google Gene: searching for DNA barcode sequences using Google search engine	University of Guelph, Integrative Biology	CANADA	Phone: (519) 824-4120-ext 56393	mhajibab@uoguelph.ca
M. Angeles Hernández		University of Navarra, Zoology and Ecology	SPAIN	Phone: 34 948425600	mahermin@unav.es
Michael James Hickerson	Quantifying uncertainty in species discovery with approximate Bayesian computation (ABC): single samples and recent radiations	University of California-Berkeley, Museum of Vertebrate Zoology, University of California	USA		mhick@berkeley.edu
Donal Hickey	DNA Barcoding of Fungi: a Feasibility Analysis	Concordia University, Biology	CANADA	Phone: (514)848-2424	dhickey@alcor.concordia.ca
Karen Elizabeth James		The Natural History Museum, Botany	UK	Phone: 44 2079425161	karj@nhm.ac.uk
Catherine Laredo		INRA	FRANCE	Phone: 0134652226	Catherine.laredo@jouy.inra.fr
Raphael Leblois		Musée de l'Homme, MNHN	FRANCE	Phone: +33 (0)1 44 05 73 43	leblois@mnhn.fr
Jere H. Lipps		University of California, Berkeley, Dept of Integrative Biology #3640	USA	Phone: 510-642-9006	jlipps@berkeley.edu
Damon P. Little	A comparison of algorithms for identification of specimens using DNA barcodes: examples from gymnosperms	The New York Botanical Garden, Lewis B. and Dorothy Cullman Program for Molecular Systematic Studies	USA	Phone: 718-817-8130	dlittle@nybg.org
Albert Lo		Univ. Science and Technology, Honk Kong	China		imaylo@ust.hk
Volker Loeschcke	The Conservation Genetics Program, European Science Foundation	University of Aarhus	Denmark		volker.loeschcke@biology.au.dk
Ion Mandoiu		University of Connecticut	USA	Phone: 1-860-486-3784	ion@engr.uconn.edu
Kasper Munch	Bayesian DNA barcoding	University of Copenhagen	DENMARK		rasmus@binf.ku.dk
Bogdan Pasaniuc	DNA Barcode Data Analysis: Boosting Assignment Accuracy by Combining Distance- and Character-Based Classifiers	University of Connecticut	USA		bogdan@engr.uconn.edu
Nicolas Puillandre		MNHN	FRANCE	Phone: 0140 79 37 52	puillandre@mnhn.fr

Jessica Rach	Character-based DNA barcoding for identifying conservation units in Odonata	TiHo Hannover, ITZ Ecology & Evolution	GERMANY		jessica.rach@ecolevol.de
Cecilia Saccone		CNR - Istituto di Tecnologie Biomediche, Sede di Bari	ITALY	Phone: 080.5929661	cecilia.saccone@ba.itb.cnr.it
Sarah Samadi		Muséum National d'Histoire Naturelle	FRANCE		sarah@mhhn.fr
Indra Neil Sarkar	Automated Barcoding Using the Characteristic Attribute Organization System	American Museum of Natural History	USA		sarkar@amnh.org
Bernd Schierwater		TiHo Hannover ITZ Ecology & Evolution	Germany	Phone +49 511 953 8880 Fax: +49 511 953 8584	bernd.schierwater@ecolevol.de
David Schindel		Consortium for the Barcode of Life	USA	Phone: 202-633-0812	SchindelD@si.edu
Mila Tommaseo-Ponzetta		Dipartimento di Zoologia	ITALY	Phone: 390805443361	m.tommaseo@biologia.uniba.it
Michel Veuille		Muséum National d'Histoire Naturelle	FRANCE	Phone: 33 [0]1 4079-4804	veuille@mnhn.fr
Haile Frederick Yancy		FDA/CVM, FDA	USA	Phone: 301-210-4096	hyancy@cvm.fda.gov
Sisi Ye		INRA	FRANCE		sisi.ye@jouy.inra.fr

Registered but unable to attend

John Olayinka Atoyebi	National Centre for Genetic Resources and Biotechnology, Moor Plantation	NIGERIA	Phone: 00234-8033824752	johnyinka@yahoo.fr
Stephen L. Clifford	Dalhousie University	CANADA	Phone: 902-494-1398	Stephen.clifford@dal.ca
Joseph Hughes	University of Glasgow	UK	Phone: 01413305346	j.hughes@bio.gla.ac.uk
Renaud Lahaye	University of Johannesburg, Botany and Plant Biotechnology	SOUTH AFRICA	Phone: +27 11 489 3477	lahaye@cict.fr
Olivier Guillaume Maurin	University of Johannesburg, Botany and Plant Biotechnology	SOUTH AFRICA	Phone: +27 11 489 3477	olive.maurin@gmail.com
Stefano Mona	University of Bari	ITALY	Phone: +390805443361	Stifano1@yahoo.it
Dirk Steinke	Biodiversity Insitute of Ontario - University of Guelph, Guelph Centre for DNA Barcoding	CANADA	Phone: 519-824-4120 ext. 56393	dsteinke@uoguelph.ca
Michelle Van der Bank	University of Johannesburg, Botany and Plant Biotechnology	SOUTH AFRICA	Phone: +27 11 489 2495	mvdb@na.rau.ac.za
Roxana Yockteng	MNHN, Systématique et Evolution	FRANCE	Phone: 33(0)1.44.79.53.80	yockteng@mnhn.fr
Ching-Ray Yu	Rutgers University, Statistics Department	USA	Phone: 732-445-2641	chingray@eden.rutgers.edu
Phoebe Zhang	Rutgers University, Institute of Marine and Coastal Sciences	USA	Phone: 732-932-6555	phoebe@imcs.rutgers.edu

APPENDIX 3: Program of the meeting
Data Analysis Working Group, MNHN -Paris - 6-8 July 2006

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	**Tim BARRACLOUGH** – Imperial College London, Silwood Park Campus	Biological inferences from barcoding data (optional)
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	**Zaid ABDO -** Deparment of Biology, McMaster University, Canada	A step towards barcoding life I: A new method to assign genes to preexisting species groups
10:45	**Jessica RACH -** TiHo Hannover, ITZ Ecology & Evolution, Germany	Character-based DNA barcoding for identifying conservation units in Odonata
1130	**Coffee break**	
11:45	**Damon LITTLE -** The New York Botanical Garden, USA	A comparison of algorithms for identification of specimens using DNA barcodes: examples from gymnosperms
12:30	**Lunch**	
13:30	**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
14:15	**Donal HICKEY** - Concordia University, Canada	DNA Barcoding of Fungi: a Feasibility Analysis
15:00	**Coffee break and discussion**	
15:15	**Javier CABRERA** – Department of statistics, Rutgers University, USA	An MLE-based clustering method on DNA barcode
16:00	**Organization and agenda of the DAWG** – closure at 17:00	

APPENDIX 4: WORKSHOP ABSTRACTS

A step towards barcoding life I: A new method to assign genes to preexisting species groups

Zaid Abdo[1,2] and G. Brian Golding[1]

A major part of the barcoding of life problem is to be able to assign newly sequenced or sampled individuals to existing groups that are pre-identified externally (by a taxonomist, for example). This problem involves evaluating the statistical evidence towards associating a new individual with a group or another. The main concern of our current research is to perform this task in a fast and accurate manner. To accomplish this we developed a model based, decision theoretic framework based on the coalescent theory. Under this framework we utilize both distance and the posterior probability of a group given the data and the newly sampled individual to assign this new individual. We believe that this approach maximizes the use of the available information in the data. Our preliminary results indicate that this approach is superior to using a simple measure of distance for assignment.

[1] Deparment of Biology, McMaster University, Canada
[2] Departments of Mathematics and Statistics, University of Idaho, USA

Comparing phylogenetic and statistical classification methods for DNA barcoding

Frederic Austerlitz[1], Olivier David[2], Brigitte Schaeffer[2], Michel Veuille[3], & Catherine Laredo[2]

Intraspecific variation in mtDNA results from a genealogical process. We can therefore address similar questions as were raised when cladistics methods were introduced in species systematics: should our classification methods account for the characteristics of the evolutionary process in producing patterns of taxonomic diversity? Should we use either phylogenetic methods, or statistical methods that are not based on evolutionary biology models, or both? The difference is that the old cladistic-classification debate was concerned with distantly related taxons, whereas the main issue for the barcode is to correctly interpret individual differences around the speciation threshold. The aim of our ongoing research is to compare the efficiency of genealogical methods (neighbour-joining and maximum likelihood methods) and statistical methods (supervised classification). To this end, we use both empirical and simulated data following given scenarios. We partition the data into two subsets: (1) some individuals make up a reference sample of known specific status, having each a sequence and a species name; (2) some individuals are known only from their DNA sequence, and the method is used to assign them to the right species. We compare the efficiency of the two series of methods in interpreting the dataset. This can be assessed exactly for the simulated sets. We will present the design and the preliminary results of this investigation on the empirical and simulated data sets.

[1]Laboratoire Ecologie, Systématique et Evolution, U.M.R. C.N.R.S./U.P.S./E.N.G.R.E.F. 8079, Université Paris-Sud, Bâtiment 360, F-91405 Orsay cedex, France.
[2]INRA, Laboratoire de Biométrie, Centre de Recherches de Jouy-en-Josas, 78352 JOUY-EN-JOSAS, France.
[3]Muséum National d'Histoire Naturelle, 16 rue Buffon, 75005 Paris, France.

Biological inferences from barcoding data

Tim Barraclough

Problem: Analysis of barcode data requires combined consideration of population and between-species processes. The challenge is to devise a statistical method for analyses of such data that is robust to inherent uncertainties (e.g. were population sizes constant or growing, is there a complete sample of species or not), while reflecting the underlying biological processes that generated barcode diversity and allow inferences on those processes.
Relevance: key requirement is statistical framework for barcode related questions; one criticism of barcodes has been the lack of biological relevance, but clearly they contain vast source for biological studies, perhaps in combination with additional minimal data.
Approach: We are developing statistical models for DNA trees from combined population and phylogenetic samples. The approach does not assume explicit population and speciation parameters, which can be clunky with large datasets or restrictive to particular models such as the neutral coalescent. Instead, we're devising generic models that incorporate parameters summarizing typical departures from strict neutral assumptions, such as increasing or decrease population size or incomplete samples of species. To our knowledge, no similar approach is published at present.
Preliminary results: We have applied the approach to the question of species delimitation from barcodes in tiger beetles from arid Australia (468 individuals from around 48 species). The paper is in press with Systematic Biology. We have also applied it to a dataset of over 500 individuals of bdelloid rotifers. It will soon be applied to demonstrate accuracy of plant barcoding markers under development by a Moore-Sloan project led by Robyn Cowan and Mark Chase at Kew Gardens.
Proposed deliverables: research level software with the option for more easily applied front-ends; exemplar papers demonstrating how uncertainties in barcode projects can be treated.

Fish barcoding from the FishTrace database: the control gene, the data validation analysis and the backup reference biological data.

Authors: The FishTrace consortium*

Presented by: José M. Bautista

This European initiative has catalysed the pooling of biological material and sequence data corresponding to more than 220 European marine fish species of commercial, ecological and zoological interest. These species have been *ad hoc* sampled from most European sea areas as well as from some extra-European areas. Overlapping species sampling from different geographical areas allows the morphological and genetic comparison of specimens from widespread species.
FishTrace database provides information on the nucleotide sequences of the mitochondrial cytochrome *b* gene (complete sequence) and the nuclear rhodopsin gene (partial sequence) from the target species. These molecular data form the basis for the validation of taxonomic data and for the development of practical tools for species diagnosis. Given the possible subtle genetic variation in populations at the mtDNA level, a second genetic marker is used. The nuclear gene coding for rhodopsin shows minimal population variation in fish and is intron-less in all teleost species. The sequence of this gene is also easily obtained and amendable of analysis which serve in our database as an internal quality control to confirm sequence analysis from mitochondrial sequences and to confer upon them the degree of reliability required to quantify the level of divergence among species, while maintaining homogeneity in the same species. The supply of sequences from two genes with different evolution rate in different species from the same phylum guarantees its application for development of phylogenetic tools for the precise ascribing of a given DNA sample. Moreover, the given independent variation rate for each gene will allow to track basal phylogenetic relationships and identify any rare case of heteroplasmy, paraphyly or hybridization between close species.
The online database at www.fishtrace.org contains standardised information on taxonomy, DNA sequences and reference-collections designed to directly confront the problem of reliable fish species identification and/or the differentiation between closely related species. Fishtrace database ensures the highest standards for marine fish identification through the accurate validation of the information compiled in the database. Online molecular and morphological identification tools are also available from the web interface.
In addition the FishTrace network holds backup biological reference collections including DNA, tissue, voucher specimens, and otoliths from the taxonomically and genetically validated fish species. These collections, deposited in four European natural history museums, constitute a reference infrastructure, unique in Europe, with important applications in fish species authenticity and related biological research.

*The FishTrace consortium is formed by 53 members from the following institutions: University Complutense of Madrid; Joint Research Centre of the European Commission; Swedish Museum of Natural History; Canarian Institute of Marine Sciences; French Research Institute for the Exploitation of the Sea; Netherlands Institute for Fisheries Research; Natural History Museum of Funchal; Natural History Museum of Tenerife; Fisheries Research Institute of Kavala; and National Natural History Museum of Paris.

Possibilities and limitations of sequence similarity and homology search tools implemented in molecular nucleotide databases for organism identification

Birgit Gemeinholzer

As a contribution to establish taxonomic identification methods using DNA sequence data, we screened newly determined ITS 1 sequences from the Asteraceae (tribes Lactuceae and Anthemideae) against the molecular nucleotide databases where sequences from that exact species, from other species of the same genus or only from other genera of the same family have already been published. This was done to evaluate to which extent the present set-up of the nucleotide databases (NCBI, EBI, GenomeNet) allows to use them for reliable routine plant identification applying the implemented sequence similarity and homology search tools. Four different sequence similarity and homology search algorithms for comparison of nucleotide-nucleotide sequences were compared (WU-Blast2, Fasta algorithm, MEGABLAST, BLASTN) and the accuracy of the sequence similarity and homology search algorithms was evaluated. We not only evaluated the effects of the default settings of the algorithms but also examined optimization criteria such as word sizes [blastn and MEGABLAST], changing the sensitivity of the search algorithm [WU BLAST2], the gap penalty [FASTA and blastn] and the filter option [blastn]. Even the sequence similarity and homology search tools were not created to serve for species identification the results in parts were quite satisfying. However, evaluating the weak points for taxon identification by sequence comparison with the available tools, we also discovered the limits of the BLAST® algorithms currently implemented in the nucleotide databases. Optimization criteria are recommended. As our main interest is based on the practical application of DNA-Barcoding in plants and diatoms, also biological mechanisms and evolutionary constrains are presented, which might always hamper the success of similarity and homology search tools. Future work will focus on the establishment of further DNA-Barcoding-datasets and algorithm exploration.

Google Gene: searching for DNA barcode sequences using Google search engine

Mehrdad Hajibabaei, Gregory Singer, Donal Hickey

Large scale DNA barcoding projects are producing a massive number of barcode sequences from thousands of species. These barcode records provide the library against which an unknown query sequence will be searched. Accurate and fast search methods for barcode data are therefore critical for the use of barcodes in routine specimen identification. This project aims at using the popular and powerful search engine of Google for searching for barcode sequences. We developed a character (word) based algorithm to divide a barcode sequence into words and then used these word patterns for searching the library of sequences using different character lengths and formats. We implemented this approach in a computer program that can connect the user to Google Desktop Search (GDS). Initial tests shows that this approach can successfully identify barcode sequences stored in different file formats in a desktop computer. More tests are underway to benchmark this search method against conventional approaches such as BLAST or distance based clustering methods. Using GDS indexing plug-ins like

Könguló, we can allow the user's GDS engine to index DNA barcode sequences stored on different servers. In this way, the user is not limited to searching for sequences stored on his or her own computer. We are investigating the possibility of incorporating tools such as GoogleBase and Google Co-op into the project with the help of Google R&D group.

Title: Quantifying uncertainty in species discovery with approximate Bayesian computation (ABC): single samples and recent radiations

Michael Hickerson

The Problem and its Relevance: Apart from the challenges facing both "species discovery" and "species-identification" via DNA barcoding, the former endeavor poses additional statistical and biological challenges. First, deployment of DNA-barcodes for species discovery must use statistical methods that work well when there is only a single specimen from a potential new species because "new" species are likely to be rare. Secondly, the methods must work given very recent speciation because this is the scenario where undiscovered species are most likely to be found, especially in cases where recent strong natural selection has driven rapid speciation without sufficient time for "pruning" (extinction).
Approach: Therefore, I focus on "species discover" given these two likely challenges. The overall goal is to provide a quick and usable method that will minimize false negatives while identifying where the inherent limits to a single-gene approach are. Minimizing false negatives comes with the cost of increasing false positives, yet the latter type of error is less of a problem when a motif is to identify sub-specific classification units (i.e. ecologically significant unit's; ESU's). Our method uses approximate Bayesian computation (ABC) under a family of standard population genetic coalescent models.
The speed and flexibility of ABC maker it appropriate for quickly identifying problematic parameter space, optimizing experimental design, and choosing suitable parameter-detection thresholds. The basis of ABC-based estimation is to calculate Bayesian posterior probabilities using summary statistics from the data rather than using the data itself. Although there is some information sacrificed, the consequent benefits of flexibility and speed often outweigh this cost in idiosyncratic systems with no general models (such as "species" delineation). Instead of using a summary statistic threshold as a means to "discover" species (i.e. 10X or reciprocal monophyly), ABC could use parameter thresholds (i.e. minimum time since isolation and zero migration). This allows a natural way to express statistical confidence in scoring "new" species with Bayes factors. For example, instead of a getting a simple yes or no given a new specimen (and its mtDNA), the researcher gets a numerical level of support for there being a potentially new species to investigate in more detail. The flipside of this method is that it can also inform species identification decisions because non-discoveries can be seen as close matches with known species in the reference database.
To test our method and find the optimal conditions, I use three types of data: 1.) mtDNA phylogenies simulated under a Yule birth/death model of speciation and extinction; 2.) a real mtDNA phylogeny of gastropods; and 3.) the data provided by DIMACS.
Preliminary results:
When speciation is approximated by a Yule model of equal speciation/extinction probabilities, successful detection is optimized when there are at least 5 individuals in the reference database from the most similar known species. Deployment of more sensitive thresholds comes with the cost of over-detection of new species (false positives). Therefore, the reference phylogeny should statistically inform a suitable threshold for each species discovery test. I have already used this coalescent simulation-based framework to explore how well various DNA barcode thresholds detect new species across a range of divergence times and effective population sizes when reproductive isolation operates under a simple Bateson-Dobzhansky-Muller model (*Systematic Biology; in press*). Although I will not focus on the dynamics of the speciation process here, the potentially huge amount of variance in speciation times should be accounted for in any method. More details of the ABC method can be viewed in the application for travel support.
Proposed deliverables:
Method for quick quantification of species discovery probability using msBAYES (author's program). Using pre-simulated prior distributions for a particular taxonomic group (i.e. birds), quantification of species discovery via Bayes factors can be accomplished in under a minute per specimen. This work will continue after the meeting, be presented again at the February 2007 conference (after improvements), and also be submitted to Molecular Ecology, a forum that has been recently designated for Barcode-related research.

DNA Barcoding of Fungi: a Feasibility Analysis

Donal Hickey

DNA Barcoding works very efficiently in animals, but it is not yet known if will be equally applicable to other groups of organisms, such as fungi. For instance, the standard models of population genetics are based on the assumption of outbreeding and sexual reproduction, but many fungi violate these assumptions. In addition, the use of mitochondrial sequences as indices of molecular evolution has recently been called into question (Bazin et al, Science 2006). We have analyzed the patterns of variation among fungal mitochondrial sequences to evaluate the applicability of barcoding to these organisms. We have also analyzed the patterns of mitochondrial inheritance in different organisms, in order to understand the population genetics implications of using mitochondrial rather than nuclear barcodes. Our results indicate that mitochondrial barcodes should be applicable to fungi as they are to animals.

A comparison of algorithms for identification of specimens using DNA barcodes: examples from gymnosperms

Authors: Damon P. Little and Dennis Wm. Stevenson

In order to use DNA sequences for specimen identification (e.g., barcoding, fingerprinting) an algorithm to compare query sequences to a reference database is needed. Precision and accuracy of query sequence identification was estimated for hierarchical clustering (parsimony and neighbor joining), similarity methods (BLAST, BLAT, and megaBLAST), combined clustering/similarity methods (BLAST/parsimony and BLAST/neighbor joining), diagnostic methods (DNA-BAR), and two novel

methods (DOME ID, ATIM) using publicly available gymnosperm nrITS 2 and plastid matK sequences as test data sets. We offer two novel alignment-free algorithmic solutions (DOME ID and ATIM) to identify query sequences for the purposes of DNA barcoding. On the test datasets, clustering methods performed the worst (perhaps due to alignment issues). Similarity methods, ATIM, DNA-BAR, and DOME ID all performed at approximately the same level. Almost all of the methods were able to accurately identify sequences to genus, however no method was able to accurately identify query sequences to species at a frequency that would be considered useful for routine specimen identification. Given the relative precision of the algorithms (median = 67% unambiguous), the low accuracy of species level identification observed could be ascribed to the lack of correspondence between patterns of allelic similarity and species delimitations.

Bayesian DNA barcoding

Authors: Rasmus Nielsen, Wouter Boomsma, Kasper Munch and Eske Willerslev.

We present a new Bayesian method for assigning individuals to taxonomic groups. The method uses a purely tree-based criterion to assign probabilities of membership of taxonomic groups, and explicitly integrates over uncertainty regarding the tree and other parameters. However, it does not directly address the population genetic issues associated with DNA barcoding. We illustrate the method using an application on aDNA from permafrost soil samples, and show how the method can be used to reconstruct ancient ecological communities.

DNA Barcode Data Analysis: Boosting Assignment Accuracy by Combining Distance- and Character-Based Classifiers

Bogdan Pasaniuc, Sotirios Kentros, and Ion Mandoiu
Computer Science & Engineering Department, University of Connecticut

Our work addresses three of the DAWG I research challenges: (a) assignment to known species, (b) character-based approaches to barcode data analysis, and (c) detection of possible new species. To ensure high assignment accuracy while avoiding overfitting the training data, our approach is to combine multiple simple classifiers. The current implementation combines a variety of both distance-based and character-based classifiers, as follows:

- *Hamming distance.* The Hamming distance between two aligned barcodes is defined as the number of positions where the two sequences have different nucleotides. The new sequence is assigned to the species of the closest sequence (MIN-HD) or to the specie with minimum average distance (AVG-HD).
- *Aminoacid similarity.* After translating barcodes to aminoacid sequences, we compute pairwise similarity scores using the Blosum62 matrix. A new barcode is assigned to the species containing the highest similarity sequence (MAX-AA-SIM) or with maximum average similarity (AVG-AA-SIM).
- *Convex-score similarity.* The similarity score between two aligned barcode sequences is determined from the positions where the two sequences have matching nucleotides by summing the contributions of consecutive runs of matches, where the contribution of a run is convexly increasing with its length. A new sequence is assigned to the species containing the highest scoring sequence (MAX-CS-SIM).
- *Trinucleotide frequency.* For each species we compute the vector of trinucleotide frequencies, and the new sequence is assigned to the species whose frequency vector is closest (MIN-3FREQ).
- *Positional weight matrix.* For each species we compute a positional weight matrix (PWM). For each new sequence we compute the probability of being generated according to the PWM of each species, and select the species that gives the highest probability (MAX-PWM).
- *Character-based pairwise species discrimination.* For each pair of species we pick the *k* most discriminating characters, and, based on these characters only, we decide to which of the two species is the new barcode more likely to belong. The new barcode is assigned to the species that is preferred in the largest number of pairwise comparisons (*k*-BEST).

The combination of methods is done by a simple voting scheme. The following table gives the average identification accuracy of proposed methods in experiments on the provided 1623 barcodes with 10-50% of the barcodes in each species used for test and the remaining ones used for training. The combined method consistently outperforms individual classifiers, with an average accuracy between 98-99.4%.

Classifier	Percentage of barcodes removed from each species and used for testing				
	10%	20%	30%	40%	50%
MIN-HD	98.8	98.0	97.8	97.2	96.0
AVG-HD	97.2	97.2	96.6	96.2	95.6
MAX-AA-SIM	99.0	99.0	99.2	98.4	96.8
AVG-AA-SIM	94.6	94.2	94.8	94.2	93.0
MAX-CS-SIM	98.2	98.2	98.6	97.6	97.4
MIN-3FREQ	94.6	93.8	94.2	92.0	92.4
MAX-PWM	98.0	98.6	97.8	95.4	94.6
10-BEST	98.6	97.0	97.6	96.2	96.2
COMBINED	99.4	99.4	99.6	98.6	98.0

We have also extended our methods to detecting the case when new barcodes do not belong to any of the known species. In experiments in which entire species are removed and used for testing, the combined method correctly detects barcodes not belonging to known species 94.5% of the time, while maintaining an accuracy of over 97% for known species barcode classification.

These and further methods will be implemented and released as open source packages under the support of an NSF grant on "Bioinformatics Tools Enabling Large-Scale DNA Barcoding" awarded to IM for the following 3 years.

Character-based DNA barcoding for identifying conservation units in Odonata

J. Rach[1], R. DeSalle[3], I.N. Sarkar[3], B. Schierwater[1,2] & H. Hadrys[1,2]

DNA barcoding has become popular as a rapid and general method for the identification of organisms. Researchers are yet exploring suitable procedures that make DNA barcoding simple and reliable. Currently, phenetic approaches and tree building methods have been used to define species boundaries and discover cryptic species. These approaches highlight a central question: How can a species be defined? Obviously, a universal threshold of genetic distance values to distinguish taxonomic groups cannot be determined. A new, promising alternative for DNA barcoding incorporates a "character-based" approach. In this method, species are delimited through the presence or absence of discrete characters within a DNA sequence. Here, we demonstrate the potential of character-based DNA barcoding for 842 Odonate specimens belonging to 64 species. A total of 57 species can be reliably discriminated through unique combinations of character states within the sequences of the mitochondrial ND1 (NADH dehydrogenase 1) and CO1 (cytochrome *c* oxidase subunit 1) gene regions. Furthermore, character-based DNA barcodes were also successfully generated on a population level for ten populations that are associated with four species. The ladder is particularly important for conservation management. Odonates are excellent indicators for a variety of ecosystems due to substantial differences in habitat specificity and their complex life cycles.
Our data suggest that character-based DNA barcoding can deliver an identification system that achieves (i) the assignment of odonate specimens to taxonomic groups and (ii) the discovery of conservation units rapidly and accurately. We thus show that character based DNA barcoding is a powerful alternative and harbors several advantages compared to phenetic approaches currently being used.

[1]ITZ, Ecology & Evolution, TiHo Hannover, Bünteweg 17d, D-30559 Hannover, Germany
[2]Yale University, Dept. of Ecology and Evolutionary Biology, New Haven, CT 06520-8104, USA
[3]American Museum of Natural History, Division of Invertebrate Zoology, New York, NY, USA

Automated Barcoding Using the Characteristic Attribute Organization System

Indra Neil Sarkar, Ryan P. Kelly, and Rob DeSalle
American Museum of Natural History, New York, NY USA

Many barcoding methods rely on distance-based, or 'phenetic,' approaches to create classification tools for species discrimination. These phenetic methods gain much of their computational tractability from the derivation of vector distances (i.e., 'similarity scores'), which are based on overall similarity metrics. However, the utilization of distance-based vectors may obscure potentially useful diagnostic characters that could be informative within a barcoding framework. In contrast, cladistic methods aim to consider comprehensive character evolution when organizing sequence information. While phenetic methods will assuredly recover the correct larger taxonomic groupings for well-studied groups of organisms, they may not always be reliable for recovering taxonomic groups for closer related species. Because biodiversity information is often sparse in taxonomic coverage, the use of phenetic methods for some data sets may therefore lead to erroneous conclusions. Cladistic methods are generally considered more reliable for phylogenetic examinations in cases where taxonomic sampling may result in closer related species. A significant limitation that is perceived with regards to cladistic methods is that the computational effort required to generate phylogenetic classifications greatly hinders the ability to develop rapid classification techniques that are character-based. We have been developing the Characteristic Attribute Organization System (CAOS) as a computationally efficient heuristic method that approximates cladistic classification. For datasets that have been organized based on a cladistic analyses, new sequences can be classified within the same framework using CAOS-based classifiers without performing a tree search. Because CAOS-based diagnostics are phylogenetically significant, they can be used as a first *in silico* step towards developing oligo-nucleotide probes for gene expression microarrays.
In this presentation, we explore a COI dataset consisting of nearly 130 Mopalia. We first compare and contrast the phenetic and cladistic tree topologies, demonstrating that the methods disagree for closer related species (i.e., the 'tips' of the tree). Next, we will showcase a new set of applications that use CAOS to develop a character-based barcode, which can subsequently be used for classification of new species. Using a 50% resampling technique, we show that the CAOS-based barcode is generally more reliable for distinguishing between closer related species. We will conclude with an exposition of how the CAOS-based barcode can be used for subsequent development of oligonucleotides.

An MLE-based clustering method on DNA barcode

Ching-Ray Yu

Species clustering problem is a very important issue on barcode project. So, I focus on clustering species problem and developing evolution trees and propose an MLE-based clustering method.
Problem: MEGA3 is the integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment by S. Kumar, K. Tamura, and M. Nei (2004). MEGA3 contains facilities for automatic and manual sequence alignment, web-based mining of databases, inference of the phylogenetic trees, estimation of evolutionary distances and testing evolutionary hypotheses. When applying parts of training datasets which are from the DIMACS website, the clusters of species from MEGA3 is shuffled. So, It is very interesting if I can propose a statistical clustering method, which is better than the existing methods in MEGA3.
Relevance: A good method to cluster new biological specimens to their proper species is very important. This is also one of the goals of DNA Barcode Data Analysis Initiative.
Approach: I propose a very simple statistical method, which is Maximum Likelihood Estimator (MLE) method. Suppose $X_{ij} \in \{A,T,C,G\}$ are identical independent random variables which are multinomial distributed with parameter

$\Theta_j = (\theta_j^A, \theta_j^T, \theta_j^C, \theta_j^G)$, where $i = 1,...,N$ stands for the i^{th} DNA (COI) sequence and $j = 1,...,n_i$ indicates for the j^{th} position in the i^{th} sequence and $\theta_j^A + \theta_j^T + \theta_j^C + \theta_j^G = 1$. Then we write down the likelihood function and get the MLE of the parameters. Suppose $\hat{\Theta}_{jk} = (\hat{\theta}_{jk}^A, \hat{\theta}_{jk}^T, \hat{\theta}_{jk}^C, \hat{\theta}_{jk}^G)$ are the MLE estimators of Θ_{jk} in species *k*. Then the clustering rule is follows:

A new unidentified specimens $Y = (Y_1,...,Y_n)$, where $Y_i \in \{A,T,C,G\}$, is assigned to species *k* if $k = \arg\max_r \sum_{i=1}^{n} \hat{\theta}_{ir}^{y_i}$.

Furthermore, we can calculate the distance between species *n* and *m* using correlation matrix between $\hat{\Theta}_{jm}$ and $\hat{\Theta}_{jn}$.

Preliminary results: There are 150 species with 1623 COI sequences in the training datasets provided by DIMACS. In order to performance the misclassification rate of the MLE-based clustering method, I randomly select 200 COI sequences as testing set. The MLE of parameters is estimated from the rest 1423 COI sequences. The misclassification rate is about 26%. This result is not so good, but improvable. Applying to the training dataset with 346 COI sequences, the clustering result is in the appendix. This preliminary result of this simple model is not good enough, but the misclassification rate could be improved if the more complicated models are considered in the future (e.g. Kimura-two-parameter (K2P) model (1980) considers the biological diversity with different evolution rate). Further more, the correlation between species can be calculated by MLE and the tree can be constructed based on the correlations.

Proposed deliverables: If the MLE-based clustering method works well on barcode sequences in the future work, it can be written as a paper with R-code for public users of barcode data.

Reference:

M. Kimura (1980) *A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences.* J. Mol. Evol. 16: 111-120.

Sudhir Kumar, Koichior Tamura and Masatoshi Nei (2004) *MEGA 3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment*. Briefings in Bioinformatics. Vol 5. No 2. 150-163.

APPENDIX 5: SUMMARIES OF WORKSHOP PRESENTATIONS

Voelker Loeschke (University of Aarhus) and **Kuke R. Bijlsma** (University of Groningen) presented and overview of the supported by the European Science Foundation. They explored the degree to which and the ways in which DNA barcoding would contribute to the population-level research of interest to conservation geneticists.

José M. Bautista (Complutense University of Madrid) described the FishTrace project supported under FP5 by the European Commission's Directorate General for Research, and compared it to CBOL's FISH-BOL initiative. Both initiatives aim to construct reference databases that can be used to identify unknowns. FishTrace uses sequence data from the control gene rhodopsin and a mitochondrial gene (*cyt b*) focuses on species ocurring in European coastal waters (9 areas, 3200 specimens representing 221 species. FISH-BOL uses a difference mitochondrial region (COI) and plans to obtain at least 5 representative specimens from all 30,000 species of marine and freshwater fish. Both initiatives use Neighbor Joining (NJ) cluster methods based on overall genetic distance (K2P weighted distance measure). Some cluster diagrams show bootstrap values (percent of resampled cases supporting each node).

Mehrdad Hajibabaei (University of Guelph) reviewed the laboratory protocols used to obtain DNA barcodes from biological tissue, and the standard data analysis leading to an NJ cluster diagram. He showed that many branch nodes collapse into polychotamous branch points when bootstrap confidence limits are applied. These "collapsed trees" are easier to interpret and are more realistic portrayals of genetic distance. Mehrdad went on to show how searches could be done on barcode sequences using search engines like Google. Long gene sequences could be broken up into words of specific character length and degree of matching would reflect sequence similarity. Google's Co-op initiative would allow barcode projects to form a shared database that could be searched efficiently.

Michael Hickerson (University of California Berkeley) addressed the process of "species discovery" – deciding whether a specimen belongs to a known species (with some documented range of variability) or belongs to a new and previously unknown species. To date, researchers have proposed a purely empirical approach that relies on a threshold ratio (interspecific divergence / intraspecific variation) for this decision. Another approach to the problem is to model of population-level change leading to species divergence, and then use the resulting patterns to interpret real-world barcode (and other) data. False positives result when barcode data incorrectly suggest the presence of a distinct species where none exists ("over-splitting"). False negatives result when barcode data assign representatives of separate species in the same species ("over-lumping"). Mutation rate, gene exchange, duration of isolation and other factors affect the likelihood of false positives and negatives. Bayesian methods use the available data distributions to assign probabilities to splitting/lumping decisions, rather than reaching yes/no conclusions. Initial results suggest that within-species samples of five individuals are optimal for minimizing bad decisions.

Kasper Munch (University of Copenhagen) noted that we often lack adequate knowledge of intraspecific variability on which to base confident assignment of specimens to species. An alternative approach is to use phylogenetic analysis techniques to assign specimens to the lowest possible taxonomic group supported by the data. Bayesian methods can be used to assign probabilities to alternate trees (constructed from the reference barcode database and the barcode sequence of the unknown) with different taxonomic assignments for the unknown specimen. The taxonomic hierarchy would be the one used to annotate the reference database.

Eric Bazin (University of Montpelier 2) recently published an article in Science comparing allozyme data (from 849 species) with nuclear (162 species) and mitochondrial DNA sequences (1629 species). This compilation suggested that mitochondrial sequence variation is higher than but invariant with nuclear sequence variation and allozyme polymorphism. Bazin inferred that mitochondrial variation does not reflect effective population size, and is limited by the selective advantage of certain variants, rather than by selection against deleterious variants. Mitochondrial variation should therefore remain stable and low within species. This finding supports the use of mitochondrial sequences as diagnostic species markers.

Bogdan Pasaniuc (University of Connecticut) used a variety of distance measures: minimum versus average distance similarity measures, longest runs ("convex score similarity"), amino acid sequences, trinucleotide frequencies, and character-based similarity to assign specimens to known species, and to determine the presence of new species. Assignment to known species is extremely high for all distance measures (above 94% when 10% of the records are used as query samples against the remaining 90% authority file). New species detection was tested by omitting one entire species from the authority file, and accuracy of some methods fell to 80%, while several measures remained as high as 98% accurate.

Frederic Austerlitz (CNRS and Univ. Paris-Sud) compared the results of a distance-based Neighbor Joining, a maximum likelihood phylogenetic method, and a character-based that selected the most informative diagnostic character state at each node (Classification and Regression Trees, CART). Analytical methods were tested on simulated data with varying levels of mutation rate, sample size, and divergence time. The phylogenetic method was more accurate with simulated data for all sample sizes. CART was consistently more accurate than phylogenetic methods for cowries and simulated data with low mutation rates, but phylogenetic methods worked better for *Litorina*, *Astraptes*, and simulated data with high mutation rates.

Neil Sarkar (American Museum of Natural History) presented the Character Attribute Organization System (CAOS) that detects diagnostic characters within datasets of aligned gene sequences. CAOS and NJ were both 100% accurate in identifying all members of a small dataset, and CAOS performed better with incomplete datasets.

Tim Baraclough (Imperial College) compared several datasets with predictions from the neutral coalescent model. Different datasets showed strong agreement with theoretical models, though some datasets diverged from the model in 5-10% of cases.

Zaid Abdo (McMaster University/University of Idaho) is attempting to formulate a taxonomic decision system that takes into account population genetics information on from coalescent theory while minimizing the risk of misassignment based on Bayesian posterior probabilities. The performance of the decision system was compared to NJ analysis of skipper butterfly data. The "coalescent assigner" method performed better than distance-based systems under almost all conditions.

Jessica Rach (TiHo Hanover) demonstrated the use of character-based data from two mitochondrial regions (ND2 and COI) from dragonflies using CAOS. NJ methods failed to separate some species that could be distinguished using character data. Combining COI and ND2 data allowed separation of local conservation units.

Damon Little (New York Botanical Garden) tested a variety of techniques (cluster methods, search algorithms, character-based techniques) on several plant gene regions (nuclear genes, spacer regions, plastid sequences). In assessing the success of the methods, he distinguished precision (ability of a method to correctly assign a specimen to its own species) from

accuracy (ability to tell sibling species apart). All methods showed high precision but accuracy at the species level was variable.

Birgit Gemeinholzer (Berlin Botanical Garden) compared the effectiveness of different search systems in the BLAST family for identification in plants, where sequence lengths vary and alignment is a problem.

Donal Hickey (Concordia University) used primate data to illustrate the importance of benchmarking the results of NJ output through bootstrapping. Bootstrapping collapses many unsupported supra-specific nodes but generally does not lump species together. He then explored the gene regions that have been proposed for barcoding fungi.

Javier Cabrera, for Ching Ray Yu (Rutgers University) explored the use of a Maximum Likelihood Estimator for clustering. He used the DIMACS pilot dataset, reserved 200 of 1623 records for testing, and obtained a misclassification rate of 16%.

Javier Cabrera (Rutgers University) and **Albert Lo** (University of Science and Technology, Hong Kong) described the potential application of the Weighted Chinese Restaurant Process (WCRP) to DNA barcode data. WCRP describes a process whereby partitions (analogous to species) are defined as tables shared by customers in a restaurant. With each iteration of the process, customers move to a different table or sit at an unoccupied table (analogous to forming a new species). This process creates a distribution of partitions (species) that contain customers (specimens). Bayesian posterior probabilities can then be used to select the most likely arrangement of specimens into species. Cabrera suggested several ways of presenting the resulting data using techniques from multivariate data analysis and Projection Pursuit.