Report of Planning Meeting at DIMACS, Rutgers University
New Approaches to the Analysis, Interpretation and Visualization of DNA Barcode Data

Held under the auspices of the Data Analysis Working Group
Consortium for the Barcode of Life
26 September 2005

Background. The Consortium for the Barcode of Life (CBOL; www.barcoding.si.edu) is an international initiative supported by the Sloan Foundation and hosted by the Smithsonian Institution. CBOL's mission is to explore and promote the development of "DNA barcoding", a new technique for identifying species that uses a short gene sequence from a standardized position in the genome. CBOL has created four Working Groups that address technical challenges to DNA barcoding, one of which is devoted to data analysis. CBOL asked Dr. Michel Veuille, a population geneticist and Chairman of the Department of Systematics and Evolution at the National Museum of Natural History, Paris, to chair the Data Analysis Working Group (DAWG). CBOL also approached the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) at Rutgers University to assess its interest in the activities of DAWG. DIMACS subsequently requested support from CBOL for a one-day planning/brainstorming meeting on the analysis of DNA barcode data. The meeting was held at DIMACS on Monday, 26 September 2005 and was attended by 26 participants (see participant list), including Dr. Veuille from Paris, a representative of the Canadian Barcode Network, and a CBOL representative. Other participants were from the statistics, computer science, bioinformatics and machine learning communities in the area. A similar meeting will be held at the Paris Museum on 15 October, at which the majority of participants will be population geneticists. A DIMACS representative will attend that meeting.

Meeting Content. The meeting was divided into three components (see meeting agenda). The first third was devoted to presentations by participants that provided background on DNA barcoding, barcode data, and their use in taxonomy and systematic biology. The second third included a presentation and demonstration by Rebecka Jornstein (Statistics Department, Rutgers University) on DNA barcode data that were available in public databases and analytical tools that are currently being used on barcode data. Participants discussed the technical details of barcode data during this part of the agenda. The final third of the day was devoted to a roundtable discussion of possible directions that DAWG could take in its efforts to improve the treatment of DNA barcode data.

Meeting Outcomes. Participants in the meeting agreed that DNA barcode data present some significant and interesting research opportunities at the biology-mathematics interface. DNA barcoding is clearly a fast-growing research area at an early stage. It is in the early phase of generating a very large body of standardized data that will be of interest to statisticians and others. In addition, barcoding has diverse areas of application (agriculture, public health, species conservation) with interested colleagues in taxonomy and consumer groups. Participants agreed that there is significant potential for collaboration among taxonomists, statisticians, machine learning specialists and other computer scientists. Further, they agreed that DIMACS could be an important agent for catalyzing these partnerships.

Some participants felt that DNA barcode data analysis would require only existing approaches, while others thought there were several areas in which new techniques were needed. At this early stage, barcode data appear very clean and complete, which may require less technique development. Future datasets will be much less clean and complete, and will be highly variable in length and quality. This will be especially true of older, degraded museum specimens.

The roundtable discussion produced the following list of technical challenges arising out of the DNA barcode initiative:

- One use of DNA barcoding will be to look for potential new species in collections of unidentified organisms. Some technique development has begun on novelty detection, and DAWG could explore the relevance of this work on novelty detection and on classification in other applied areas;
- The other main use of DNA barcoding will be assigning unidentified specimens to known species, using reference barcode sequences taken from specimens that have been assigned to species with a high degree of confidence. DAWG could explore the sensitivity of class membership to clustering algorithm and data format (sequence similarity versus site-specific character data);
- Differences in DNA barcodes can be minor and subtle among recently formed species, or among species that are still hybridizing to a degree. DAWG could develop new techniques or modify existing techniques for novelty detection and classification that can identify subtle differences from barcode data;
- Current analytical methods rely on trees to display group membership, coherence, and inter-group separation. DAWG could promote the creation of new tools for analysis and visualization that will maximize intragroup coherence and intergroup separation. Ideally, these techniques should address the effects of varying sample size per speices, variable lengths of sequence data per specimen, and variable certainty of nucleotide base-pair calls within a sequence;
- Taxonomists will want to know the level of confidence associated with the assignment of unknown individuals to known categories, and for the validity of new groups proposed on the basis of barcode data;
- Obtaining samples of organisms and their DNA barcode sequences can be labor- and cost-intensive. Strategies for minimizing sample sizes are needed, as well as maximizing confidence measures with limited sample sizes;
- High quality DNA will be difficult to obtain in many cases, especially when the only representatives will be museum specimens. They will produce short, non-contiguous barcode sequence data. DAWG could develop techniques that minimize the sequence length needed to assign unknown specimens to known species;
- Techniques will be needed for the treatment of ‘semi-missing data’, i.e., the analysis of specimens for which DNA barcode sequences include different numbers of nucleotide sites and sites whose data are of varying certainty. Not every specimen produces a DNA barcode sequence of exactly the same length; some nucleotide sites are not “called” (assigned to A, G, T or C), and some sites are called but with some level of uncertainty;
- There are two controversies surrounding the source and form of DNA barcode data:

  - Is it better to use overall similarity measures (percent sequence divergence) or discrete discrete character data (A, G, C or T at each nucleotide site)? Can the use of each form of data be optimized to analyze data at different scales of resolution?
  - Is it better to use shorter sequences from more than one gene region or longer sequences from one gene region?

The meeting participants discussed the possible activities that could be included in the DAWG's programs of work, and they suggested the following:

- Periodic workshops at DIMACS (or elsewhere), each with a different technical focus;
- Hold multiple competitions, each with a specific technical problem and objective criteria for selecting winner; e.g., separate competitions/challenges to maximize accuracy of assignments using reference records (samples with high-certainty species assignments) and unlabelled data (specimens from poorly known groups with uncertain or no species boundaries). Each competition would culminate in a workshop at which approaches and results would be presented and compared. The organizers might offer participants the opportunity to submit preliminary results and receive intermediate-stage feedback;
- Hold one competition with multiple problems, no objective criteria for winners, selection of winners by committee. Each competition would culminate in a workshop at which approaches and results would be presented and compared. The organizers might offer participants the opportunity to submit preliminary results and receive intermediate-stage feedback;
- Create incentives in the form of travel to meeting, opportunity to publish in a themed journal volume;
- Offer graduate student dissertation enhancement awards; and
- Provide a partnering service that would introduce taxonomists with barcode projects and data to statisticians, machine learning specialists, and computer scientists;

Relatively little time was spent discussing how linkages could be built to groups of researchers working in relevant areas of population genetics, beyond saying that it would be valuable and important to have interactions among taxonomists, population geneticists, statisticians, and computer scientists.

Participants agreed that it is important to build on the momentum established by the meeting at DIMACS. They concluded that there should be informal follow-up as soon as possible in the form of the following next steps:

- Provide a brief report of this meeting as input to the Paris planning meeting;
- Circulate the meeting report to its participants as soon as possible, and solicit additional ideas and reactions to the meeting from them;
- Hold another brainstorming session for local statisticians, bioinformaticists, machine learning specialists and population geneticists at DIMACS;
- Create a DAWG website at DIMACS on which presentations from the meeting, a summary of the meeting, background publications and pilot datasets can be posted;
- Explore possible interest in publishing a themed Barcode issue in Bioinformatics, other journals at math/biology boundary;

- Explore possibility of holding competitions through the Genetic Analysis Workshop; and
- Ask the CBOL Scientific Advisory Board for suggestions of experimental datasets of various types that could be used for DAWG activities: e.g., poorly and well-known groups; well-defined populations versus hybrid zones; labeled and unlabeled with species names; small and large sample sizes per species; short and long sequences; broad well-mixed populations versus local populations with limited gene flow; time-series from population studies.