Implementing a Student Allele Database via the World Wide Web

(Extended Abstract)

Lee A. Newberg
Biological Sciences Division, The University of Chicago
Chicago, IL 60637-5415, USA
ℓ-newberg@uchicago.edu, http://http.bsd.uchicago.edu/∼ ℓ-newberg/

Jerome J. Jahnke
Biological Sciences Division, The University of Chicago
Chicago, IL 60637-5415, USA
j-jahnke@uchicago.edu, http://http.bsd.uchicago.edu/∼jahnke/

John A. Kruper
Biological Sciences Division, The University of Chicago
Chicago, IL 60637-5415, USA
j-kruper@uchicago.edu

David Micklos
DNA Learning Center, Cold Spring Harbor Laboratory
Cold Spring Harbor, NY 11724-1400, USA
micklos@cshl.org

AACE ED-MEDIA Presentation: June 22, 1996

Abstract: We describe the design and implementation of the Human Genome Diversity Project's Student Allele Database Facility and its interface via the World-Wide Web. The electronic lab bench is 80% complete at the time of this writing and can be found at [http://http.bsd.uchicago.edu/hgd-sad/].

1 Overview of the Project

The focus of this paper is the design and implementation of the Student Allele Database Facility and its interface via the World-Wide Web. The implementation of the electronic lab bench is 80% complete at the time of this writing. The curious are invited to explore [Database, 1996]. The Student Allele Database Facility is part of the larger Human Genome Diversity project which we describe in this section.

The Human Genome Diversity—Student Allele Database (HGD-SAD) will involve high school students from around the country in a long-term research project that illustrates many facets of the Human Genome Project. The project is centered around a hands-on laboratory that enables a student to produce a personal "DNA fingerprint" of the TPA-25 polymorphism on chromosome 8. In the schools, students isolate their own DNA from cheek cells obtained using a safe and simple mouth wash procedure. The crude DNA samples are then analyzed on site or passed to a partner genome research center for PCR amplification and separation of allele polymorphisms by agarose electrophoresis. If analysis is off site, photographs of the electrophoresis results are returned to the school.

Students determine their own genotypes and have the option of submitting their genotypes to a *Student Allele Database* maintained at The University of Chicago. Via the world-wide web, students can perform Hardy-Weinberg calculations and statistical tests to compare their allelic frequencies with those in the growing database. The addition of student populations from throughout America, Europe and, hopefully, other parts of the world will allow students to compare allelic frequencies in divergent populations and, perhaps, see evidence for genetic drift and evolutionary patterns.

HGD-SAD is a model for leveraging precollege biology instruction into the world of contemporary research. The project provides a cost-effective means to directly link students with human genome research and relegates appropriate team participation roles to high school teachers, genome researchers, database managers, and companies. HGD-SAD provides mechanisms to train teachers and research partners, as well as laboratory and computer infrastructures to facilitate participation by large numbers of students throughout the United States. The project extends the DNA Learning Center's expertise in developing laboratory curricula in molecular genetics and administering teacher-training workshops throughout the United States. It builds upon strong collaborations between the computational biology groups at Cold Spring Harbor Laboratory, The University of Chicago's Biological Sciences Division, and Washington University (St. Louis); and the company laboratories of Roche Molecular Systems, Carolina Biological Supply Company, and the Porto Conte Research and Training Laboratories (Sardinia, Italy).

In brief the project aims to:

1. Develop an educational analog of the human genome project that "personalizes" gene technology and provides unique learning opportunities for high school students.
2. Provide an appropriate mechanism for increased collaboration between genome research centers and local schools.
3. Support a Student Allele Reference Laboratory at the Cold Spring Harbor Laboratory.
4. Develop a Student Allele Database Facility at The University of Chicago.
5. Conduct training experiences to introduce teachers and genome researchers to their roles in the project.
6. Develop curriculum materials to support the evolving project.
7. Provide *bona fide* scientific data on allele frequencies.

2 Student Allele Database Design Considerations

In designing the Student Allele Database our considerations fell into several categories. We implemented the Student Allele Database while paying close attention to such issues as hardware dependence; ease-of-use; novice vs. expert access modes; clear presentation of sophisticated analyses; the ability to use subsystems stand-alone; and ease of cooperation among students, teachers, and researchers.

- We wished that access to the software be as universal as possible. Many schools have only Macintosh computers while other schools have only IBM PC clone computers, thus we did not feel comfortable developing software that runs on just one of these platforms. We chose to implement the Student Allele Database via the World-Wide Web so that it would be available across many platforms — World-Wide Web browsers are available on Macintoshes, IBM PC clones, UNIX computers, and through many on-line service providers.
- Although many schools now have access to the Internet and the World-Wide Web, not all of them have high speed connections. We therefore implemented the interface so that it did not require a high data-rate connection. Although we have made use of graphs, charts, and artistic graphics, we have taken care not to use too many. Furthermore, wherever possible the interface makes sense even if the students use a text-based browser, forgoing all images. We were also aware that technology progresses more quickly than school budgets. We chose to freeze the web standard at HTML 3 (see [HTML, 1996]) and HTTP/1.0 (see [HTTP, 1996]). These are available from many web browsers, including Netscape 1.1 which is available free (see [Netscape, 1996]) to educational institutions.
- The differences between typical World-Wide Web applications and standard applications running on Macintoshes, Microsoft Windows, or the X Window System can be considerable. Often the way in which, for instance, menus are presented on the World-Wide Web is not intuitive to the users of these traditional platforms. We chose intuition over gimmickry to facilitate access to the Student Allele Database.

- We took into consideration the great variety among users of our software. In several places we have provided both a path for novices and a path for experts so that either a novice or an expert can take advantage of as much power as he or she can handle. This approach can be seen in both the data submission and data analysis sections of the software.

- Conceptually, simulations of data do not require the existence of tools for submitting and analyzing actual data. We therefore implemented the Student Allele Database to reflect this independence. Each of our data simulation tools can be used in a stand-alone mode. This allows use of the Student Allele Database simulation software in many situations and circumstances, even if they are barely related to Human Genome Diversity Project or the Student Allele Database.

- Learning does not take place in a vacuum. The Student Allele Database design reflects this by allowing communication between students, teachers, and researchers through its bulletin board. Through forms-capable World-Wide Web browsers the bulletin board is designed to allow participants to pose, debate, and answer questions on many topics. The designed bulletin-board interface allows participants to cite each other, particular data sets, or arbitrary web pages, easily and intuitively.

- Because we know that students and teachers may want to know more, we think it is necessary and have implemented the Student Allele Database with hyperlinks to genome resources around the world.

In the following sections we describe the Student Allele Database facility. The description cannot accurately reflect the actual interface and we strongly recommend that you explore it yourself. See [Database, 1996].

3 Data Submission and Editing

A classroom of students with their TPA-25 genotypes will want to put their data into the Student Allele Database. If they haven't already done so, they first create a "group" in which their data will be placed. The group mechanism is convenient in that the students can later supply their group's tag when running statistical tests if they wish to isolate those tests to their group only.

The group creation web page is a simple form that requests the country, state or province, and city of the group. It also requests the longitude and latitude of the group. These values can be used in later data analysis to restrict the data set being analyzed. The form includes a hyperlink to U.S. Census data so that groups in the United States that do not know the longitude and latitude of their city can look it up.

Once a group has been created the individuals' data must be entered. This can be accomplished using an easy-to-use form that allows the addition, modification, or deletion of one individual's data at a time. The instructions accompanying the short form are simple and explicit and even those who have little experience with computers will be able to enter their own data. Along with the genotype of an individual, the form requests information about race and gender. This information allows later data analysis based upon these categories. One of the goals of the Human Genome Diversity Project is the maintenance of the privacy of the participants and the entry form does not request any further identifying information.

If many additions or modifications need to be made to the group's data, the bulk entry form can be used. It provides the entire data set for the group in one large edit-able area. The data may be modified in any way and resubmitted to the database. The associated instructions encourage the user to copy and paste the data from the edit-able area into his or her favorite word processing program. From there edits can be made with ease. The final data is pasted into the web browser and submitted to the database.

With these two input formats we have achieved our twin goals of providing both a simple but also a powerful way to enter data into the Student Allele Database. The simpler way can be used individually by each student as he or she enters personal data and later to modify that data if needed. The bulk entry method can be used for later corrections on a large scale, or for initial data entry if so desired.

4 Data Analysis

The data analysis tools allow any student or teacher to ask questions about the data — and get answers. The tools can be broken into two categories according to the number of populations analyzed. Several tools take a population of the whole database or some subset and measure it or compare it to predictions. Other tools allow the student or teacher to compare two distinct populations to measure their similarities and differences.

After the user has decided how many populations will be analyzed, he or she is asked to describe the population or populations. A single web page allows the specification of one or both populations. The form allows the user to define the population by placing restrictions according to zero or more of race, gender, group tag, longitude, and latitude. The user leaves blank those fields for which no restrictions are desired but may enter values to specify ranges (in the case of latitude or longitude), or selects particular attributes (such as Asian, Black, Hispanic, Native American, and/or White for race). For ranges, two inputs boxes are supplied, one for the lower limit and one for the upper limit. For selections among multiple discrete possibilities a selection list-box is provided that allows the user to choose each selection by the click of a mouse.

Once the population or populations have been selected, there are several measures and/or comparisons to choose from. The tests and measures include those of Polymorphism Information Content (PIC), Entropy, Genetic Drift, and Chi-Square computations. When only a single population is analyzed comparisons of its genotype distribution to that predicted by Hardy-Weinberg Equilibrium are allowed.

The entropy and PIC measures are separate, though similar, methods for measuring diversity within the selected population. Each is a measure of how easy it is to determine whether two DNA samples come from the same person. The web page which presents the results of either measurement describes what the computed numerical value means. The discussion includes information to help place the numerical value in context; high measured values show that, on average, it is easy to determine that two samples come from different individuals but that low measured values mean the opposite. The discussion also gives the Entropy or PIC for the population under the assumption that the population follows the Hardy-Weinberg Equilibrium and compares that measurement to the actual measurement.

The Genetic Drift and Chi-Square comparisons compute the similarity or difference between two selected populations. The Chi-Square comparison computes the chi-square value and the p-value of the null hypothesis that the two populations are examples of subsets drawn from a single larger population. A high chi-square value and corresponding low p-value indicates that the two populations are dissimilar and that it is unlikely that their differences are merely statistical. The student is shown the calculation of chi-square and the implications of its actual value are discussed.

Genetic Drift is a measure of the number of generations that, according to a common theoretical model, one has to go back to find the common ancestor population. Similar populations will have low Genetic Distances while a pair of dissimilar populations will have a large Genetic Distance. The computed distance is displayed and its implications are discussed.

In all cases, the emphasis is on the qualitative implications of the measured value. Through questions, students are encouraged to discuss these implications with their teacher and classmates. They are encouraged to examine whether the particular measurement accurately reflects their class. The questions are presented along with the results and are tailored to be interesting in the context of the actual computed values.

5 Data Simulators

The data simulators provide a mechanism for students to generate fictitious populations and test them. Because this ability is useful even when a student is not part of a "real" population these simulators are designed to be stand-alone. They can be used by a student even if he or she knows nothing about the remainder of the Student Allele Database.

The simulators correspond to the measures and comparisons described previously. Each generates a user-specified number of sample populations according to user-specified parameters. These are measured or compared, as appropriate, and the data is plotted in a histogram. Using a simulator the students can get a real feel for the diversity among populations. They can see *directly* how normal a particular population is. They can see *directly* the extent that two populations are more dissimilar than other pairs of populations.

Each simulator is implemented as two web pages. The first is for user input and the second is for the simulator output. The input form allows various parameters to be specified. The set of parameters depends on the simulation to be run but always includes, for instance the size of a trial population, and the number of such trial populations.

The output page shows the input parameters and the histogram of the results. It is designed to be compact so that it can be printed on a single page and turned in as part of a homework assignment. The Hardy-Weinberg Simulator (see [Hardy-Weinberg, 1996]) is now used as part of a Population Genetics laboratory at the University of Chicago even though that class does not otherwise interact with the Student Allele Database.

6 Genome Resources

It is hoped that students will take their experience with the Student Allele Database and the questions that it provokes and seek out further information. To ease the search hyperlinks to genome resources are available from the main Student Allele Database web page and also appear on other web pages in relevant contexts. We provide dozens of links to resources which fall into three primary categories: research genome resources, educational genome resources, and professional science organizations.

7 Bulletin Boards

The bulletin boards provide a place for discourse among students, teachers, and researchers. Discussions will be separated by topic and participation in one or more conversations is easily accessible through menu options. The full paper will discuss the bulletin board part of the Student Allele Database in more detail.

8 Conclusions

We expect full implementation of the Student Allele Database Facility, well in advance of the 1996-1997 school year. We are excited about the project, both as a tool for teaching biology as well as an example of how the World-Wide Web can be used to great advantage to bring cross-platform, computing power and database access to students throughout the country and the rest of the world. We strongly recommend that you take a look at the Student Allele Database web pages themselves [Database, 1996] to see what we are excited about!

9 Acknowledgments

This activity was funded in part by NSF grant 9254626. The Student Allele Database software was implemented by Lee Newberg, to whom correspondence should be sent. Jerry Jahnke was a contributor to the Hardy-Weinberg Equilibrium simulator. Graphic design is by Jilla Rafizadeh.

References

[Database, 1996] Student allele database. [http://http.bsd.uchicago.edu/hgd-sad/].

[Hardy-Weinberg, 1996] Hardy-Weinberg equilibrium simulation. [http://http.bsd.uchicago.edu/hgd-sad/HWSimulator/sim.cgi].

[HTML, 1996] Hypertext markup language specification version 3.0. [http://www.w3.org/pub/WWW/MarkUp/html3/CoverPage.html].

[HTTP, 1996] Overview of HTTP. [http://www.w3.org/pub/WWW/Protocols/Overview.html].

[Netscape, 1996] Netscape navigator. [http://www.netscape.com/comprod/products/navigator/].