# Characterization of the Global Viral Metagenome

**Jeff Elhai (PI), Virginia Commonwealth University**

**Participating Intitutions:**

(Co-PI's) Hiram College, Univ. of California at Los Angeles

(Others) Beijing Univ. (China), Chesterfield Technical Center, Fudan Univ. (China), Haraldsbogymnasiet (Sweden), Hebrew Univ. (Israel), San Diego State Univ., Univ. of Delaware, Univ. of Warwick (UK)

## Intellectual Merit

Viruses are major cogs in the engine of evolution and critical elements in the web of life on earth. Although there are more viruses than any other biological entity, we know little of the range of their diversity, far less than we know about plants, animals, and single-cell organisms. The project seeks to comprehend the diversity of viruses by mounting a massively parallel sequencing and analysis of viral genomes from different places on earth.

The effort will be driven by high school students and undergraduates around the world. Participating biology classes will collect and process an environmental sample. The samples will be sequenced, and , students will adopt unknown viral genomes and interpret them. The effort will be facilitated by a web-based knowledge/programming environment, BioBIKE, that facilitates flexible analyses by those with no computer programming experience. Insights thus gained will be added by students to a web-accessible database that enables cross-genome comparisons and communal annotation by all participants and the greater scientific community.

Students will in this way experience science not as a body of facts but as a living process. They will:

- Learn by discovery – Through a module of several weeks that introduces students to the concepts of molecular biology and genome analysis by computational experiments
- Learn by doing – Through participation in a project to identify the viral metagenome, requiring each participant to consider a part of the puzzle from multiple perspectives, e.g., molecular biology, ecology, and physiology.
- Learn by teaching – Through peer-to-peer mentoring that pairs students in one locale that have experienced the project with those in another who are just beginning
- Learn by collaborating – Through a research community enabled by a web site that facilitates meeting and discussion by participating students and members of the greater research community.

Faculty at participating undergraduate institutions and high schools will be trained in the working of the web tools and how to help students be successful in research experiences by on-site workshops and continued support from a distance.

## Broader Impacts

The sequences produced and annotated as a result of the project will enormously increase the number of available sequenced viral genomes and provided a valuable resource to aid in our understanding of viruses and evolutionary processes. More than that, however, if successful, the project will provide a scalable model that can be extended to engage a large number of students in a meaningful research experience, widely viewed as the best way to learn what science really is. A common research experience could change the way the citizenry of our country and others approach decisions that rely on an understanding of scientific issues. Students will become important parts of a scientific process that, like science at its best, transcends national boundaries. This may provide inspiration as to how other processes might work better in an ideal world.

**Key terms: student engagement, global network, scalability**

## Characterization of the Global Viral Metagenome
## Senior Participants*

**Cathy Burke (Chesterfield Technical Center): Senior participant**
High school teacher participating in alpha test of program, Spring 2007. Studied molecular biology of tropical parasites in graduate school and beyond.

**Gail Christie (Virginia Commonwealth University): Co-PI**
Studies bacteriophage biology, gene transfer mediated by bacterial viruses, and bacterial phenotypes conferred by resident prophages. Her lab has sequenced and annotated several viruses.

**Jeff Elhai (Virginia Commonwealth University): Co-PI and Principle Administrator**
Director, Bioinformatics and Bioengineering Summer Institute; Studies genome analysis; PI of NSF-funded project to develop BioBIKE as a knowledge-base/programming environment.

**Yuan Gao (Virginia Commonwealth University): Co-PI**
Studies ultrafast sequencing techniques and algorithms for motif discovery

**Brad Goodner (Hiram College): Co-PI**
Pioneer in engaging undergraduates and high school students in genome sequencing projects; Studies infection process by soil bacteria.

**Tom Hanson (University of Delaware): Senior Participant**
PI on NSF-funded project to study microbial diversity through environmental metaproteomics.

***Aaron Kaplan (Hebrew University, Israel): Senior Participant***
*Host of Jerusalem site. Studies relationship between marine and freshwater photosynthetic bacteria and their environments*

**Cheryl Kerfeld (University of California at Los Angeles): Co-PI**
Director, UCLA Undergraduate Genomics Research Initiative; Developer of bioinformatics modules for educators; Studies structural and functional aspects of bacterial microcompartments.

***Songgang Li (Beijing University): Senior Participant***
*Host of Beijing site. Has studied systems ecology, now focusing on integrating multiple sources of information for the analysis of genomic data.*

***Nick Mann (University of Warwick, UK): Senior Participant***
*Host of Coventry site. Studies ecology, community structure, and evolution of marine viruses.*

**Shozo Ozaki (Virginia Commonwealth University): Senior participant**
Protozoan parasite genomics. Developed project in metagenomics for high school students to characterize environmental microbial diversity, through a university and high school partnership

***Per Paulsrud (Haraldsbogymnasium, Sweden): Senior participant***
*High school teacher participating in alpha test of program, Spring 2007; PhD thesis in symbiotic associations by soil cyanobacteria*

**Forest Rohwer (San Diego State University): Senior participant**
Group leader of several projects to sequence viral metagenomes and PI on NSF-funded project to assess the diversity of aquatic bacterial viruses.

***Dave Scanlan (University of Warwick, UK): Senior Participant***
*Host of Coventry site. Studies molecular ecology and genomics of marine photosynthetic bacteria.*

**Ryan Templeton (Clover Hill High School and Virginia Commonwealth Univ.): Senior participant**
Director, VCU's Governor's School for Life Sciences and Medicine; High School Biology teacher; Graduate student in Dept. of Education; Studies methods of engaging students in research.

**Eric Wommack (University of Deleware): Senior participant**
Director of NSF-funded Microbial Observatory for Virioplankton Ecology. Studies roles of viruses in a variety of microbial communities.

***Jun Xie (Fudan Univesity, China): Senior participant***
*Host of Shanghai site. Studies protein self-splicing.*

---

* International participants listed in red and italicized

# Characterization of the Global Viral Metagenome

## I. Introduction

### I.A.Engagement of students in the scientific process

In olden times, two of us were presenting to a sophomore-level course in genetics the well-worn paradigm of how *E. coli* regulates the expression of its *lac* operon in response to the levels of lactose and glucose. The students had read in their textbook how cyclic AMP mediated the effect of glucose, and now they were hearing in class about an experiment that tested that idea. They were carefully brought through the structure of the experiment and what each possible result would signify. By the end of the class period, the results of the experiment were displayed, and it was clear: the textbook was wrong!

There was a silence in the room. You recognize that silence – the momentary stoppage of time when the framework that kept certain loose ends from connecting suddenly breaks apart, and its pieces reform in a new way that brings those ends together.

But that's ***our*** silence. In that class on that day, the silence was broken by groans and notebooks snapping shut. "*Why did we have to go through **that**?*" "*If it's not true, then why didn't you just say so?*" "*If the book's wrong, then how are we supposed to know what's right?*"

The premise of this proposal is that science education in the United States is in need of reform,[1,2] leading the great majority of students to view science, particularly biology, as a body of facts to be learned rather than a process of understanding to be applied. It does no good to ***teach*** that science is a process – the very act of teaching this view subverts the message! Students must discover discovery, by involving themselves in the process.

It isn't difficult to devise an environment where students are drawn into the process of discovery: we just invite them to join us in playing the same game we find so satisfying. This works fine to affect the next generation of researchers, but the model requires considerable individual attention – as it should – and each of us can mentor only a small number. In the end, the model can be applied only to an elite few, leaving the great majority with the prevailing authority-based view of science that has served us so poorly in critical societal decisions.

The goal of this proposal is to develop a different model, one that can enable a large number of students to absorb how science works by making them a contributing part of the scientific process.

### I.B. What viruses populate our planet?

Extrapolations based on direct measurements of viral particles in environmental samples indicate that the global viral population is on the order of $10^{31}$ particles,[3,4] about $10^{9}$ per gram of soil[5] and as many as $10^{8}$ per ml of in aquatic samples,[3] making them by far the most numerous class of biological entities on Earth. And they're more than flu bugs. There is a growing appreciation that viruses can alter the composition of microbial populations[6] and are a major driving force in evolution, shuttling DNA between different organisms[7] and different environments,[8] serving as pool of rapidly changing genes.[9,10]

It is surprising, then, how little we know about viral DNA. Those viruses whose DNA sequences have been determined represent a minute fraction of the population and are heavily biased towards those related to disease processes. Recently, sequences from hundreds of DNA viruses taken from seawater[11,12] marine sediment,[13] or fecal samples[14,15] were obtained at once, and in all cases, the great majority of genes observed were unlike anything seen before.[16] Viruses -- unlike plants, animals, and bacteria -- remain largely unknown, a major gap in our understanding of biological entities on our planet.

We propose to engage students throughout the world to sample their local environments and through the collaborative analyses of viral genome sequences, characterize what is the diversity of viruses on our planet.

## II. Project Description

It is difficult enough to engage even experienced researchers in a collaborative project, never mind those with no research experience. To be successful, we need to construct a community environment that invites students to join and to present them with a series of well-defined tasks that lead them to the process of discovery. All of this must fit readily into existing educational structures and not add significantly to the net burden instructors must bear. Finally, the project itself must yield a living product of obvious worth to the scientific community, enticing researchers to add their own creative input.

Outline of Student Experience

A. Introductory curriculum to prepare students for the research experience
B. International effort to characterize the global viral metagenome
    1. Sample collection
    2. Sequencing and assembly
    3. Analysis of viral sequences and formation of hypotheses
    4. Experimental testing of hypotheses
C. Integration of students into an international research community
D. Training and assessment

### II.A. Introductory curriculum to prepare students for the research experience

Last year, high school teacher Cathy Burke and Co-PI Jeff Elhai began developing a curriculum to introduce high school students to the concepts of molecular biology through computational experiment. Underlying the effort was BioBIKE,[17†] an integrated knowledge and programming environment developed to enable research biologists without programming experience to manipulate genomic, metabolic, and experimental data in novel ways. We propose to develop a complete, self-contained web-based curriculum that might fit as a 3-week unit within a larger biology course.

The curriculum will be based on discovery modules, each one exploring a concept of molecular biology in the manner an extraterrestrial might explore life on earth. Earthling cells contain DNA composed of millions of letters. They seem to be organized into functional entities we'll call "genes", but physically, the genes have no obvious demarcations. How do the cells determine where a gene begins? Textbooks begin by presenting students with the concept of "start codon". Instead, students will be led through a series of steps (enabled by BioBIKE) in which they gather the beginnings of hundreds of genes and look for patterns within this set. Their observations suggest questions that can be addressed by computational experiment. How many genes begin with "ATG" (the canonical start codon)? (see **Fig. 1** for BioBIKE's rendition) Is the more general pattern "NTG" ***sufficient*** to mark the start of the gene? What genes ***don't*** begin with the more general pattern "NTG"? And so forth. Through this investigation, students come to discover what a start codon is (and isn't), but more importantly, they discover that they ***can*** discover and need not rely on the faulty generalities found in their textbooks.[‡]

The modules begin with explicit directives but midway through ask the student to extend what they have already done, based on more fuzzy directives. By the end, students are given an open-ended task: *Use the tools that helped you find signals marking the beginnings of genes to find signals marking the* ***end*** *of genes.* In this way, they learn first by example and then by extension how to use computer programming and quantitation to solve scientific problems, but the focus is always on the science.

We found that most high school students needed considerable help absorbing the new concepts of computation and quantitation and the notion that they might learn about nature by examining it. The help was given in the form of one-to-one mentoring provided by students in a Virginia Commonwealth University (VCU) bioinformatics class who had themselves recently completed the same modules. These pairings proved to be of great use to both mentor and mentee.

---

† Formerly called BioLingua; funded by NSF grant DBI-0516378.

‡ The example describes an existing module, accessible at http://ramsites.net/~biobike

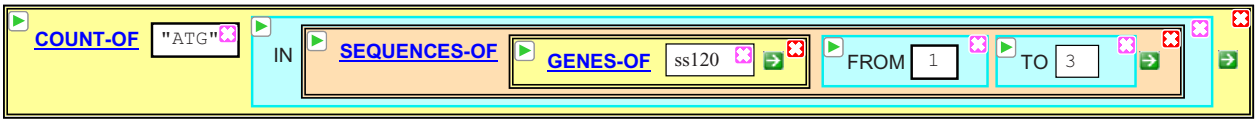


**Figure 1: Example of BioBIKE graphical language.** The statement, constructed part by part, asks for the number of "ATG" trinucleotides in the first three nucleotides of each gene of the cyanobacterium *Prochlorococcus marinus* SS120. The identities and sequences of the genes are built into the system. For an example of the language solving a problem of biological interest, see http://ramsites.net/~biobike (link at bottom of page)

We propose to modify BioBIKE to enable students to mentor and collaborate with each other at a distance and to provide the kind of on-line help they will need. Ryan Templeton and Co-PIs Jeff Elhai and Cheryl Kerfeld will develop a full curriculum, greatly extending the few modules already developed and translating them into the new graphical BioBIKE language. By the end of the first year we intend to have an array of modules that can be incorporated into existing biology courses. Along with Per Paulsrud (Sweden), we will continue testing the ideas described here on a small scale this coming Spring.

### II.B. International effort to characterize the global viral metagenome

High school students are often encouraged to view research as individualistic and competitive (think of a good idea, do the experiment, win a prize!). This view of science is far from reality and arguably has no staying power as an incentive for students to pursue science or adopt scientific thought in their routine affairs. We propose a collaborative research project in which the incentives are the thrill of exploring the edge of what is known, the satisfaction of sharing insights within a community of those with like interests, and the conviction that the end result makes a difference to the world. And it's a lot more fun.

#### II.B.1. Isolation, processing, and sequencing of viruses from environmental samples

Institutions participating in the project will generally come as pairs: a university and a local high school. During the first operational year of the project, approximately 10 participating institution pairs will obtain environmental samples at local sites and process them as described by Fig. 2. Selecting the site and performing the sampling is the first step of many that create a link between students and the project. Some operations may need to take place at the university.

Viruses will be isolated from environmental samples as previously described[13,16] and with the advice of participating experts (Nick Mann, Forest Rohwer, and Eric Wommack) and counted by fluorescence microscopy by students, who will then send the viral particles to be sequenced by VCU (coordinated by Shozo Ozaki) and the Department of Energy's Joint Genome Institute (JGI),[§] using a combination of Sanger (ABI), Solexa, and 454 pyrosequencing methods.[**] We estimate that an environmental sample will contain between 1000 and 10,000 viral genome types, i.e., ~50 to 500 Mb.[16] At today's prices (rapidly dropping), the cost of obtaining 1x (partial) coverage of genomes within an environmental sample using a GS100 pyrosequencer is an estimated $8,000 to $80,000. The JGI contribution will be at no cost. In parallel, students in the UCLA Undergraduate Genomics Research Initiative (http://www.lsic.ucla.edu/ugri/) will characterize microbial biodiversity by sequencing and analyzing

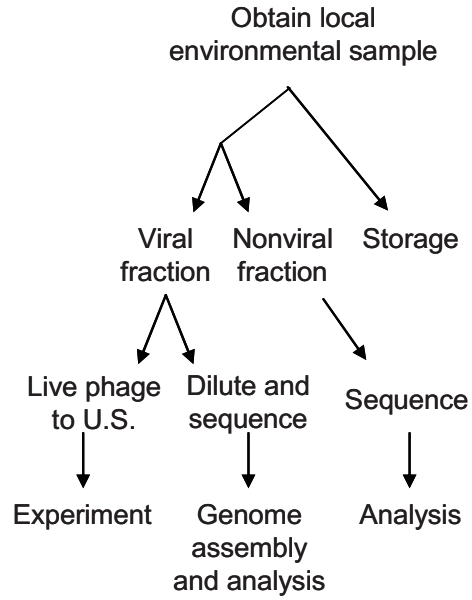


**Figure 2: Flow chart of fate of environmental samples.** Sequencing and assembly will take place at one of a few sites.

[§] We have received a verbal from JGI for a substantial amount of the sequencing.
[**] The university will purchase one or the other, but the decision hasn't been made which yet.

16S RNA genes in the nonviral fractions of selected environmental samples, directed by Co-PI Cheryl Kerfeld with the advice of microbial diversity experts Tom Hanson and Dave Scanlan.

The assembly of sequences obtained from mixed environmental sample present special challenges,[16] and Co-PI Yuan Gao will develop web-based sequence-assembly software that takes into account the sequence polymorphisms expected from an environmental sample and that brings students into the process. Sequences that have been assembled and roughly annotated will be made available to students (and the world) in a dedicated BioBIKE site containing these and all other available viral sequences. BioBIKE precomputes protein similarity and facilitates analysis by means besides sequence similarity that may be more appropriate to comparisons amongst genes of viruses and their hosts.[18]

### II.B.2. Analysis of viral sequences

Each student will be given a large contiguous sequence from a viral genome to analyze and will be aided in this effort by their peer mentor (an experienced student somewhere in the world), the student's teacher (trained as described in Section II.D), the project's help desk composed of students at the lead institutions, supplemented with students anywhere in the world who wish to continue their connection with the project, and online project descriptions coordinated by virus expert Co-PI Gail Christie. Students will analyze their sequences with at least the following questions in mind:

- Is the provisional annotation of the gene correct? (finding errors in what seems like an authoritive statement is often a liberating experience for students)
- What previously identified viruses (if any) are similar to the one in question?
- What genes does the virus possess? What are they similar to? Are there any that are surprising?
- Is there evidence of mosaicism?
- What is the global distribution of viruses similar to the one in question?
- Is there a correlation between viruses similar to the one in question and properties of the isolation site, e.g. environmental characteristics and microbes found to be present?

Students will post their analysis and annotation to a publicly accessible website modeled after CyOrf,[††] a site supported by GenomeNet/KEGG[19] that facilitates the community annotation of all available cyano-bacterial genomes. The site but will be available to all and modifiable by registered users consisting of participating students and anyone from the research community. At the same time, students at some institutions (initially Hiram College) will attempt to piece together partial viral genomes, using methods successfully employed by Co-PI Brad Goodner's group to close gaps in microbial genomes.[20]

### II.B.3. Experimental testing of hypotheses

It is important for students to experience how their analysis of genome sequences connects with laboratory experiments. During the pilot phase of this project, there will be an interaction between students at the international sites contributing sequence data and students at Hiram College, guided by Co-PI Brad Goodner. In succeeding years, we will recruit other institutions to join in.

A single environmental sample provides a snapshot of the viral population at a single time and place. To go beyond these restrictions, students will identify unique sequences from within putative viral genomes, synthesize and label them with different fluors, and hybridize them with samples from the original viral fractions for analysis by confocal fluorescence microscopy (available at Hiram College). The probes can be used to assess viral populations in the original environmental sample as well as other related samples.

In addition, students at the international sites will suggest possible microbial hosts on the basis of their analysis of viral sequences and the results of microbial 16S analysis. These hypotheses will be tested in the laboratory by plating viral fractions on a battery of potential hosts, both lab-adapted and wild. Viral DNA will be isolated from plaques for sequence identification.

---

[††] http://cyano.genome.jp (see http://ramsites.net/~biobike, bottom of page for CyOrf under development)

**II.C. Integration of students into a focused international research community**

A research topic that is interesting and productive is an important attractant to students, but equally important is a community environment in which students feel they share the journey and rewards with others with common interests. We will devote much effort to bring such a community into existence:

a. Students will collaborate in the analysis of their genomes with peer mentors and later mentor others.
b. Each participating class will focus on a single environmental sample, with an effort made to integrate individual findings.
c. The web interface will encourage comparisons between different environmental samples analyzed by different groups.
d. We will set up specialty centers. Sites will send samples to one site for characterization of the microbial population and to another for analysis by electron microscopy, but students at the original site will coordinate the integration of information about their particular sequences.
e. Students in English-speaking countries will collaborate with students in non-English speaking countries to translate the system's documentation.
f. Finally, students will have input from the existing scientific research community interested in viral diversity, including participants Nick Mann, Forest Rohwer, and Eric Wommack.

Many of these efforts will be facilitated by a central web site modeled after CyanoBIKE[‡‡] that currently brings together the cyanobacteriological research community. The site will make it easier for distant participants to meet each other and will support and encourage public discussion groups on topics that arise.

**II.D. Training and assessment**

Planning the curriculum and research project will draw on the extensive experience of the Co-PIs in engaging students in research both within and outside of classroom environments. The project will also draw heavily on participant Ryan Templeton, recipient of the R.E.B. Foundation Award for Teaching Excellence to support his research on how to engage students in research and how to train teachers to do the same in their classrooms. Ryan and others in the VCU Department of Education will also develop a plan to assess the effectiveness of the project in changing students' attitudes towards research and their abilities to analyze scientific problems.

Before a site begins the curriculum and research project, a team from one of the lead institutions will visit the site to put on a workshop intended for those at the university and local high schools interested in exploring the project. The team will include a student from a lead institution who has gone through the curriculum and research experience and is interested in the country to be visited. Since workshops will be conducted in the local language (when appropriate), the student will speak the language at least to some rudimentary extent. VCU students going to China will be guided by participant John Herman, Director of VCU's Study Abroad in China program.

**III. Greater Impact**

The model we describe presents students in the U.S and around the world with an enticing entry into collaborative science, where their individual efforts make a measurable contribution to solving a problem important to a scientific community and their collective efforts produce results whose significance they can immediately appreciate. Quantitation, computation, ecology, and other areas emerge not as separate bodies of facts to be learned (and forgotten) but rather as tools useful in solving a problem at hand. We anticipate that sequencing costs will soon drop sufficiently so that many institutions can participate in the model we will develop. It is difficult to project the impact a large number of students engaged actively in science as a process might have on a society that increasingly relies on wise science-based decisions.

Even the pilot implementation that is projected for the first year will greatly increase the number of viral genomes that have been sequenced and analyzed by a human. The resulting database and the site that makes it publicly available will be a great resource to those interested in evolutionary processes and the nature of viral communities.

---

[‡‡] http://ixion.csbc.vcu.edu (see http://ramsites.net/~biobike, bottom of page for guided tour of site)

## Characterization of the Global Viral Metagenome
## REFERENCES


1. Bio 2010 National Research Council, *BIO 2010: Transforming Undergraduate Education for Future Research Biologists* (National Academies Press, Washington DC, 2003).
2. Handelsman J, Ebert-May D, Beichner R, Bruns P, Chang A, DeHaan R, Gentile J, Lauffer S, Stewart J, Tilghman SM, Wood WB. (2004). Scientific Teaching, Science 304:521-522.
3. Wommack KE, Colwell RR (2000). Virioplankton: Viruses in aquatic ecosystems. Mirobiol Molec Biol Rev 64:69-114.
4. Wittman WB, Coleman DC, Wiebe WJ (1998). Prokaryotes: The unseen majority. Proc Natl Acad Sci USA 95:6578-6583.
5. Williamson KE, Radosevich M, Wommack KE (2005). Abundance and diversity of viruses in six Delaware soils. Appl Environ Microbiol 71:3119-3125.
6. Mühling M, Fuller NJ, Millard A, Somerfield PJ, Marie D, Wilson WH, Scanlan DJ, Post AF, Joint I, Mann NH (2005). Genetic diversity of marine *Synechococcus* and co-occuring cyanophage communities: evidence for viral control of phytoplankton.
7. Boyd EF, Davis BM, Hochhut B (2001). Bacteriophage–bacteriophage interactions in the evolution of pathogenic bacteria. Trends Microbiol 3:137-144.
8. Sano E, Carlson S, Wegley L, Rohwer F (2004). Movement of viruses between biomes. Appl Environ Microbiol. 70:5842-5846.
9. Lerat E, Daubin V, Ochman H, Moran NA (2005). Evolutionary origins of genomic repertoires in bacteria. PLoS Biol 3:e130.
10. Breitbart, M, P Salamon, B Andresen, J Mahaffy, A Segall, D Mead, F Azam, F Rohwer (2002) Genomic analysis of uncultured marine viral communities. Proceedings of the National Academy USA. 99:14250-14255.
11. Microbial Observatory for Virioplankton Ecology. University of Delaware, Eric Wommack. http://www.virusecology.org/
12. Breitbart, M, Felts, B, Kelley, S, Mahaffy, JM, Nulton, J, Salamon, P, and F Rohwer (2004) Diversity and population structure of a nearshore marine sediment viral community. Proceedings of the Royal Society B. 271. 565-574.
13. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, Rohwer F (2003). Metagenomic analysis of an uncultured viral community from human feces. J Bacteriol 185:6220-6223.
14. Cann AJ, Fandrich SE, Heaphy S (2005). Analysis of the virus population in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. Virus Genes 30:151-156.
15. Edwards RA, Rohwer F (2005). Viral metagenomics. Nature Rev Microbiol 3:504-510.
16. Massar JP, Travers M, Elhai J, and Shrager J (2005). BioLingua: a programmable knowledge environment for biologists. Bioinformatics 21:199-207. (*http://dx.doi.org/10.1093/bioinformatics/bth465)*
17. Blaisdell BE, Campbell AM, karlin S (1996). Similarities and dissimilarities of phage genomes. Proc Natl Acad Sci USA 93:5854-5859.
18. Kanehisa M, Goto S, Kawashima S, Nakaya A (2002). The KEGG databases at GenomeNet. Nucleic Acids Res 30:42-46.
19. Ewing A, Chaney L, Kadoi R, Guercio A, Goodner B (2004). Bioinformatics effort at Hiram College. 25th Annual Crown Gall Conference, Cornell University, August 2004.