The Library of Life

A joint program of Bridging the Rift Foundation and Cornell University
Table of Contents

Bridging the Rift Foundation 4
The Library of Life 6
Talk Abstract 8
BTR’s Scientific Programs 10
School of Bioinformatics 12
Opportunities for Participation at Cornell 13
Cornell Units 14
Graduate Degree Programs 16
Library of Life of the Desert 18
Biographies of Key Personnel 37
Bridging the Rift Foundation

Bridging the Rift Foundation is a trilateral project involving Jordan, Israel, and the United States. Launched in 2004, the initiative involves building a joint academic center on the Israeli-Jordanian border in the Wadi Araba/Arava region. In partnership with the leading research centers in Jordan and Israel and with participation of Cornell and Stanford Universities, BTR specializes in collaborative scientific and educational projects focusing on the life sciences. BTR promotes economic development, innovative scientific research, and advanced education opportunities.

Cornell University and Stanford University have joined BTR as major academic partners. Each university has established a BTR Program that will coordinate scientific programs with senior scientists from Jordan and Israel as well as host graduate students, and postdoctoral research associates working in collaborative teams. In partnership with BTR and Stanford and Cornell Universities, the top universities in Jordan and Israel provide leadership and expertise in all of BTR's scientific projects. Leading scientists from both countries oversee the scientific agenda of the BTR Center from deciding which areas of the region to focus study to selecting the best postdoctoral fellows in the region.

The BTR Research Center, to be built on 150 acres of land allotted by both Jordan and Israel and spanning the international border, will become a science and technology hub for the region. The Center will be the base for the laboratory and computer work to be done jointly by teams of Israeli and Jordanian scholars with the collaboration of top professors from Stanford and Cornell. Much of the current research focuses on the biology of the Middle East, ranging from microbiological studies and inventoring microbial biodiversity to ecosystem analyses done with advanced computer technology.
Recognizing the unique potential of Bridging the Rift Foundation, both the leaders of Jordan and Israel have personally committed their continued support. Both counties have signed a Memorandum of Agreement to establish the Bridging the Rift Center showing their approval of the initiative. His majesty King Abdullah II of Jordan has stated that the BTR project “is bigger than both Jordan and Israel.” Former Israeli Prime Minister Ariel Sharon declared that the BTR project is of “first rank strategic importance in the region.”

**The Bridging the Rift Research Center** will become a center of education, research and development for the countries of Jordan and Israel. Once complete, the Center will house the world’s most advanced database, the Library of Life. Plans for the center also include a large auditorium and conference center.
The Library of Life

The sequencing projects that determine the genetic code of numerous life forms (human included), are revolutionizing modern biology, providing sample blueprints of life on Earth. However, the genetic blueprints alone provide an incomplete picture of living organisms. The translation of the blueprints (DNA sequences) into living entities, the interactions of different species with each other and the environment, and their evolution are all parts of the bigger picture of the new Life Sciences. This additional data may be stored electronically in different formats (text, images, sounds), and in different locations. Integrating this information and making it searchable and accessible to computational modeling is the goal of the Library of Life.

This modern approach to Biology is more quantitative than that of a decade ago, and has stronger links to economy and industrial applications through the fields (for example) of biomedical engineering, agriculture, nanobiotechnology and more. These fields rely on rapid access to data and interpolations from genomics to ecosystems. Linking, searching, and modeling for heterogeneous data-rich applications are prime goals of Computer Science. The intellectual contribution of Computer Science to the design and implementation of the Library are therefore profound.

The Middle East, the Dead Sea region, and its nearby desert are ideal places to start such a comprehensive and large-scale project. The limited diversity of life in the desert makes the collection of the relevant and comprehensive data more feasible. Politically and scientifically significant, the Dead Sea is a natural resource shared by both Israel and Jordan and it harbors unique, fascinating and ecologically important forms of life.
We therefore propose to first develop the **Library of Life of the Desert**, focusing on the Dead Sea and the surrounding desert environment. The Computer Science Department at Cornell University has partnered with Bridging the Rift (BTR) to create, design, and implement the Library of Life of the Desert.

Concrete questions that we ask: In what ways do the extreme environment of the Dead Sea affect the life forms of the region? Do the harsh climate conditions of the desert prompt genetic adaptations? How can these specializations be compared and isolated? In the Middle East, the answers to these questions have the potential to influence and revolutionize the industries and economies of the region.

Though the Library of Life project has been initiated in the Dead Sea region between Jordan and Israel, as more Middle Eastern countries and scientists join the Library will expand its scope. We hope that the Library of Life will become a major source of information on life in general and in the Middle East in particular. We also plan to provide data mining and analysis tools to promote discoveries in basic sciences and in relevant industries.
Talk Abstract

The Role of Information, Modeling, and Simulations in Modern Biology

Ron Elber, Cornell University

The general life sciences such as biology, agriculture, and medicine are predicted to advance significantly in the 21st century. They are predicted to impact health, the economy, and the way we approach the process of scientific discovery. While predictions are not always true and must be followed by assessments of outcomes, there are already numerous indicators of profound changes in the way scientific research is conducted in the life sciences. These dramatic changes, which make the life sciences more quantitative and boost discoveries and modeling, are likely to affect all aspects of the life sciences and their applications. In my talk I will discuss why the time is right for this revolution, what exactly this revolution is, and how the Library of Life of the Desert contributes to this effort.

From the perspective of computational modeling, biology presents a new challenge and a change in the paradigm of scientific modeling. Instead of a focus on fundamental general laws that can be described by a small number of parameters, mathematical equations, and intensive simulations (as in physics), biology has become a new data-intensive science. How much information do we need to record to model the function of a cancer cell? How much information do we need to record to model the development of a disease? How much data are required to understand evolutionary pressure? The need for vast data to study biological information and processes is well appreciated in genomics. However, understanding processes in the life sciences require richer data that is also less structured. Even at the cellular level only, there is a great need for more than sequence information. Consider for example protein–protein interactions, gene expression, molecular transport, etc.—all crucial elements of cell function that we wish to study. We also wish to link and mine unstructured data, like text, images, and voice recordings of scientific discussions. At the center of Computer and Information Science one finds studies of information retrieval, data mining and machine learning analyses, making it an ideal partner for the new data-intensive biology. I will discuss concrete examples of biological discoveries led by computational data mining.
Finally I will describe the proposed Library of Life of the Desert project. I will explain how this project fits into the concept of modern biology as a data-intensive science and how it will impact our ability to investigate desert life, to share our findings with the scientific community worldwide, to build better information systems, and to learn from many other shared resources available through the internet.
BTR’s Scientific Programs

Library of Life of the Desert
At the heart of BTR’s scientific programs is the Library of Life of the Desert (LOLD). The LOLD is an innovative computational infrastructure of scalable software to store and mine diverse and heterogeneous biological data. The Library of Life of the Desert will provide a comprehensive collection of digital records of life forms (e.g. genome sequences), but also pertinent literature, images, voice recordings, etc. It will also include environmental data relevant to the development of desert species, such as global positioning data, climate, ecological and soil conditions. The Library will allow sophisticated data mining and queries about species, their environment and their mutual interactions. The software will integrate diverse and public biological data into one comprehensive database. It will further incorporate concrete information generated by parallel BTR projects that are focused on the Dead Sea desert environment as discussed below.

Discovery in the Dead Sea
The bioinformatics technology developed and applied to the Dead Sea region, will be applicable to other ecosystems and organisms throughout the Middle East. Currently the software is designed by BTR-partnering scientists from Cornell University in collaboration with microbiologists and ecologists from Jordan, Israel and Stanford University.

In partnership with Stanford the Microbiology Project seeks to record and study microscopic life in the Dead Sea region. Samples are taken from both sides of the Jordanian-Israeli border, providing a unique perspective on microbial life that was not possible before. This project will provide the initial set of data in what will become a continuously expanding inventory of the biological life forms and ecosystems in the Middle East desert. An important component is the determination of the genomes of these microbial species, shedding light on molecular origins of salt resistance.
The BTR Ecology and Biodiversity project brings Stanford and Cornell scientists together with experts from Israel and Jordan to study the unique diversity and ecology of the Middle East. Like the Microbiology project, the initial phases of the Ecology project will focus on the life systems of the Dead Sea region, then it will be scaled up to encompass other ecological systems of the Middle East desert.

The primary goal of the research is to investigate the relationships between life forms and life systems throughout the Middle Eastern desert. For the first time, with the help of Bridging the Rift and the cooperation of the Israeli and Jordanian governments, researchers are able to collect samples and do transect surveys that cross the border. The outcome is a greater understanding of this unique environment as well as the ability to guide policymakers how to preserve biodiversity and conserve valuable resources as they plan the future of this land.

The scientific agenda of the Bridging the Rift Center has become a platform for an increasing number of programs and proposals encompassing many different scientific disciplines, all with the goal of bringing people together in a lasting community of cooperation.
School of Bioinformatics

This exciting initiative uses the resources available at Cornell University to introduce scientists to the field of bioinformatics. Once experienced in the latest technologies and applications the students can become part of the team responsible for the design and implementation of the Library of Life. Conceived as a series of intensive workshops, the first will be held at Cornell’s Ithaca campus. As the initiative grows other prestigious institutions will be tapped to host the school.

Computing and information technologies have transformed research and business in the life sciences, facilitating breakthroughs in molecular biology and consequent new applications in biotechnology, pharmaceuticals, health care, the environment, the chemical industry and agriculture.

World leading scientists will lecture in the subjects of software development, engineering, system architecture and modeling. After a morning of lectures from these distinguished professors and scientists, students of the School will spend the afternoons in laboratory sections putting these skills to work.

The attendees will enjoy the elaborate and sophisticated equipment of the university’s laboratories, and personal training will take place in small groups.
Oppunities for Participation at Cornell

The Library of Life of the Desert (LOLD) is a comprehensive research project in modern and quantitative biology that exploits the rapid growth in biological information to promote discovery in molecular, cellular, organismic, and ecological biology. Interested participants may join the Library of Life project as a postdoctoral fellow, M.Eng (Master of Engineering) student in Computer Science, or Ph.D. student in either Computer Science or Computational Biology. Concrete projects with the LODL team under this project are sketched below; depending on qualifications and level of involvement, appropriate academic degrees could be earned. (Mentioned under “knowledge” is the expertise learned under this project.)

A. Integration and overseeing the database implementation: Integration with other bioinformatics sources, integration of different components of the Library of Life. Knowledge of general bioinformatics principles, web programming, C# programming and Perl programming, and bioinformatics modeling.

B. The construction of a structured and unstructured database. Knowledge of database structures, algorithms for data mining, machine learning techniques.


E. Learning from images: creating software to classify and identify desert plants from images. Knowledge: machine learning, computer vision, computational statistics.
Cornell Units

Cornell Facts

Founded in 1865, Cornell University is a member of the Ivy League and a partner of the State University of New York. Cornell is composed of fourteen colleges and schools—seven undergraduate units and four graduate and professional units in Ithaca, New York, two medical graduate and professional units in New York City, New York, and one in Qatar. Cornell’s main campus in Ithaca, New York, currently includes more than 260 major buildings on 745 acres. The Ithaca campus has more than 13,000 undergraduate students, 5,800 graduate students, and 1,500 faculty members. A renowned research institution, Cornell’s 2003-04 research expenditures totaled $537.7 million ($356.1 million of this funding was from federal sources; $181.6 million was nonfederal).

Faculty of Computing & Information Science

To keep pace with the growing need for computing, Cornell University in 2000 committed new resources to creating a campus-wide Faculty of Computing and Information Science (CIS). Led by a dean who reports directly to the provost, CIS does not belong to a single Cornell school or college. Instead it is a “metacollege,” a constellation of individuals, institutes, and programs that connect Cornell computer scientists with colleagues whose work involves computing — or whose work is the study of computing. The Faculty of Computing and Information
Science is founded on the recognition that ideas and technology of computing and information science are relevant to every academic discipline. CIS serves to bring together experts in computing with researchers and scholars in a variety of disciplines. Units within CIS currently include the Department of Computer Science, Department of Statistical Science, Cornell Theory Center, Program of Computer Graphics, Information Science Program, Computational Biology Program, and the Computational Science and Engineering Program.

**Department of Computer Science**

The Department of Computer Science at Cornell University, which was organized in 1965, is one of the oldest departments of its kind in the country. It has a full-time faculty of 43, approximately 110 resident Ph.D. students, 100 M.Eng students, and the undergraduate program graduates about 200 C.S. majors each year. The department is typically ranked as one of the top five in the country.
Graduate Degree Programs

**Ph.D. in Computer Science**

Cornell offers one of the leading Computer Science Ph.D. programs in the U.S. It is distinguished by the excellence of the faculty, by a long tradition of pioneering research, and by the breadth of its Ph.D. program. Each year, about 20 to 25 new Ph.D. students join the department. The department and Cornell University as a whole make it easy to share ideas across subfields of computer science and across academic disciplines. With forty-three professors and senior research associates doing research in at least ten areas that cover nearly every topic within computer science today, the department is diverse. It also is very well integrated. As a result, interesting ideas are shared — a fact that contributes to both collegiality and productivity. Faculty members are readily available to graduate students and one another. Professors’ offices are interspersed with graduate students’ offices, and most work with open doors.

**M.Eng in Computer Science**

The Master of Engineering (M.Eng) program in Computer Science is a one-year program that is designed to enhance professional skills in practical computer science. The program is especially well-suited to students seeking advanced credentials for employment in industry. The program requires taking several advanced courses and completion of a faculty-supervised project.
Ph.D. in Computational Biology

Now in its second year, the graduate field of Computational Biology offers Ph.D. degrees in the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological systems. The field provides interdisciplinary training and research opportunities in a range of subareas of computational biology involving topics such as DNA and protein databases, protein structure and function, computational neuroscience, biomechanics, population genetics, and management of natural and agricultural systems.

Computational Biology (CB) at Cornell is an interdisciplinary program that links the computational and mathematical sciences with the life sciences. The Computational Biology faculty come from 16 departments in a spectrum of endowed and contract colleges, including the Weill Cornell Medical College. As might be expected, these faculty also represent the related fields such as Computer Science, Biophysics, Applied Mathematics, and Biometry.

For Additional Information

Cornell University: www.cornell.edu
Computing & Information Science: www.cis.cornell.edu
Department of Computer Science: www.cs.cornell.edu
Computer Science Ph.D./M.Eng programs: www.cs.cornell.edu/degreeprogs/grad/
Computational Biology Ph.D. program: www.cb.cornell.edu/
Library of Life of the Desert

Introduction

Focusing on the promise of the new life sciences, this proposal suggests a new type of database and data structure for biological information: The Library of Life. Ongoing sequencing projects (including Homo sapiens) underline the need for tools and innovative means to analyze the rapidly growing genomic data and its supplements, the variations between and within species, and to connect the sub-cellular information to phenotypic observations such as biological resistance to extreme climate conditions. The Library of Life and its prototype, the Library of Life of the Desert, will be an essential tool for research in academia and industry, facilitating discoveries with no geographical boundaries.

To understand the goals and the challenges facing the Library it is useful to briefly compare biology to other natural sciences (physics and chemistry) to put this proposal in perspective.

Natural sciences (physics and chemistry) have made a profound impact on our intellectual heritage, the economy, and the length and quality of human life. Their prime goal is the quantitative description and modeling of natural phenomena. Historically, the search for new knowledge in these sciences followed three steps: (i) data acquisition, (ii) data mining, and (iii) modeling and simulations. For example, the positions and motions of stars were carefully recorded (step (i)) before an analysis of the trajectories suggested the existence of a force inversely proportional to the square of the distance between a pair of stars (step (ii)). Based on the discovery of this beautiful and simple law, simulations are conducted today that predict the behavior of entire galaxies and star systems (step (iii)), making it possible to better understand the past (and the future) structure of the universe.

Chemists followed a remarkably similar path to scientific discovery during the 20th century when simulations of molecular systems played a crucial role in the academic and industrial search for materials with special properties. The properties of materials that we seek vary from macroscopic strength and flexibility to microscopic health-related drug activities. As the complexity of systems increases, computer simulations play an ever-increasing role in developing intuition for the problem at hand and making predictions. Compared to physics, the chemical sciences put a stronger emphasis on data collection and data mining since exact
simulations, following fundamental laws, were proven difficult in the general case. Nevertheless, the search for fundamental laws, and approximations based on fundamental principles drive many of the chemical discoveries, similarly to physics.

Studies in the life sciences are different since the prime focus so far has been on data collection and classification (step (i)). We do not know if general guiding principles (conceptually similar to gravitation) could be found at all or if we have to shift our approach and learn to comprehend more complex and concrete models. Obviously many successful modeling and simulations (steps (ii) and (iii)) were carried out in biology, and we gained considerable insight into specific systems. Nevertheless, we do not have a comprehensive view of life on earth, or in the other extreme, even enough data to model the basic living entity – the cell. To understand cells quantitatively the highly visible sequencing projects (such as the sequencing of the human genome) must be supplemented with information on regulation, expression levels, interaction networks, cellular locations of gene products, and more, data which is only partially available. More complex multi-cellular life forms are obviously more difficult to record and analyze than individual cells. To understand life on earth, we need data at levels higher than the cell, or individual life forms. We need to understand how different species interact with each other, with their environment, and how the environment “responds.” This information includes images, global positioning data, time courses, and climate conditions, data that are more heterogeneous, less structured, and are distributed in many laboratories across the world with limited organization. Amusingly, not only the raw data but also the scientific literature in the life sciences is orders of magnitude more vast and scattered than the physical or chemical literature, presenting another challenge for assembling and integrating information.

The enormous complexity of life makes current biological studies “data-centric” rather than “compute-centric.” Researchers mine for interesting patterns in existing databases and drive discoveries in medicine, agriculture, bio-nanotechnology, and more. In this research it is not possible to use laboratory notebooks or linear files to effectively store and access this heterogeneous information. Sophisticated (computer science) approaches are required to enable the mining and modeling of the data by the research community. While significant progress has been made in the representation of structured information at the sub-cellular level,
considerably less was done to represent less structured data at the cellular and multi-cellular levels. *The emergence of the new biology as a data-intensive discipline is fundamentally different from the evolution of physics and chemistry, and it makes Computer Science a natural partner of modern biology.*

We propose to develop such a biological network of information that will enable easy population of the databases as well as mining through heterogeneous and unstructured data. We call this system the Library of Life. As a prototype of the Library of Life we will develop the Library of Life of the Desert, a comprehensive database that includes information describing the life in the Dead Sea and the nearby desert.

The proposal below outlines the first three years of the Library of Life, involving the development of core software and infrastructure. The development of the infrastructure will follow two major directions. The first aim is the development of data collection, storage and data mining techniques appropriate for the vast and diverse information of the Library of Life. The second aim is the development and aggregation of analysis and software tools to convert information into knowledge. In the first three years we will focus on the following projects (the contact persons in Cornell are listed in the parentheses):

A. Integration and overseeing the implementation (Dr. Jaroslaw Pillardy)

B. The construction of a structured and unstructured database (Prof. Johannes Gehrke)

C. Analysis of genomics data (Profs. Ron Elber and Uri Keich)

D. Automated searching in scientific literature (Prof. Thorsten Joachims)

E. Learning from images (Profs. Richard Caruana and Daniel Huttenlocher)
A. The design and implementation of the Library of Life database

Functional Components

Figure 1 shows the functional components of our proposed bioinformatics architecture. We now discuss these components in more detail.

![Figure 1: System Architecture](image-url)
A.1 Data Collection

The Data Collection component deals with collecting and uploading data from the field. There are several issues that need to be addressed in this context. First, the researchers in the field should be provided with appropriate tools so that data can be collected easily. Second, the data should be formatted appropriately so that it can be easily uploaded into the database system. Finally, techniques for effectively uploading large volumes of collected data need to be developed.

We propose to develop a set of data collection software tools that can be used by researchers to capture and enter data from the field. Researchers can enter the data using a desktop or laptop computer. We also propose to develop flexible yet expressive formats for representing the collected data, which include various facets of the collected information, such as the observed parameters, GPS location, and annotated text. Finally, we propose to develop web service interfaces that can be used to upload the collected data. By relying on a generic web service architecture, the data collection component will work from anywhere in the world and function across multiple platforms.

A.2 Data Storage and Organization

The Data Storage and Organization component deals with efficiently storing and organizing the collected data. There are two main challenges associated with the design of this component. First, we need to design a data store that can scale to the hundreds of terabytes of data that will be collected over the course of the project. Second, we need to handle highly diverse unstructured data since the Library of Life will be a catalog for highly disparate yet related data sources, such as genome data, phenotype data, various species data, and environmental data.

We propose to use a relational database to store and organize the collected data. The benefits of using a relational database are (a) it can scale to large data sets, (b) it supports a high-level query interface that can be used to query the data, and (c) it is a proven and robust technology that is well supported by commercial vendors. However, one of the challenges in using a relational database for a data collection as diverse as the Library of Life is dealing with the vast diversity of interconnected data. For instance, a bacteria will likely have different properties than, say, mammals.
or environmental factors, although all of them may be related by the same ecology. A common way to store diverse data in relational databases is to store them in separate relations (or tables), each of which stores properties specific to the organism stored in that relation. However, a downside of this approach is that it limits querying across organisms -- such as, finding all organisms related to a given DNA sequence -- since the information is split across multiple tables.

We thus propose to develop highly flexible relational schemas that can capture the diversity of the data while still allowing users to efficiently query across the different organisms. Our proposed solution is to model the biological taxonomy as an inheritance relationship and map this to relations. In this way, common properties of organisms (such as DNA sequences) are abstracted away in a single relation (which can be queried across organisms), while specific properties of individual organisms are stored in separate tables. While there has been some work on modeling single inheritance (where each organism can have exactly one superclass) in relational databases, this model is too restrictive for the Library of Life where we can have a general taxonomy that is a DAG (Directed Acyclic Graph) instead of a simple tree. This requires multiple inheritances, and we will develop new techniques for efficiently modeling multiple inheritances in relational databases.

A.3 Search and Browsing

The Search and Browsing component enables users to explore and query data stored in the database. There are three main challenges that need to be addressed in this context. First, users should be provided with personalized views of the Library of Life, since the entire database will be very large and users may be interested in exploring specific parts of the database in more detail. Second, users should be able to download the desired data in bulk if they wish to do some offline processing. Finally, users should be able to issue exploratory queries to discover patterns that they previously did not know existed, and be able to browse these results.

We propose to develop a personalized web interface for users that will track user preferences and present the information most relevant to the user. This personalization will also use collaborative filtering techniques to notify users of related data in their domain of interest, and also notify them of newly added data. We also propose to develop a web service in-
interface that will enable users to download the data for offline processing.

### A.5 Data Integration

There is already a multitude of existing data sources available, especially in molecular biology and biochemistry. For example the NCBI, the Protein Data Bank, and PUBMED. For the Library of Life to be successful, it will have to integrate existing data in order to permit queries that span local data that is hosted by the Library of Life and external data sources.

Integration of external data sources requires that we first answer the question of how we will actually access external data. One possibility would be to copy external data regularly into the Library of Life; this would require us to find out what part of the data has changed and has been updated, and to handle possible inconsistencies between copies taken at different times. Another possibility is that other data providers will expose Web-Service interfaces to their system and that we can thus easily interface the Library of Life with their systems. Interoperability through Web Services is currently one of the major movements in industrial system integration, and other scientific communities (for example, the astronomy community) are moving towards this modus of interaction. We do not believe that there is a silver bullet that solves the data integration problem. As a first step, we will enable both ways of data integration for the Library of Life, and time will tell whether integration through Web Services will prevail.

Data integration also brings about hard problems in data cleaning. Different external data sources could describe the same concept, but carry contradictory information. Or different external data sources could describe the same concept, but use very different language or identifiers to describe it. We need to develop novel data cleaning methods to resolve these questions; first thoughts lead us to methods that take the reputation of the data source into account and methods that use redundancy across several data sources to disambiguate fact from fiction.
B. Data mining

B.1 Queries Over Structured and Unstructured Data.

Note that data in the Library of Life of the Desert will come in many different structural representations. There will be structured data such as data about experiments, data that associates attributes with sample specimens found in the desert, and other structured data. There will be unstructured data such as digital lab notes from biologists out in the field or transcripts of conversations or meetings about the ongoing research. There will be semi-structured data for example documents from the scientific literature that have some structured fields (such as the name of the author, the date the paper was submitted, its title, etc.) and associated unstructured data (such as the actual content of the research paper). There will be video and audio data from researchers out in the desert who will upload their data via their PDAs. Thus we need a data management infrastructure that will have to be multimodal – rather than working only over structured (for example, relational data or data encoded in XML) or unstructured data (for example, document collections), it will work across a variety of data formats, not only in terms of storage, but also in terms of query semantics. Thus we will need to unite both structured and unstructured data as well as other types of data for example images, video, and audio. What are some of the requirements of this system? First, it will have to return ranked results similar to Google queries; thus it has to support search capabilities. Second, users need to be able to adjust the relevance function that is used to ascertain the importance of results. Third, users need to be able to refine their query results using biological taxonomies, and in general, need to be able to incorporate background knowledge.

B.2 Topology Queries.

Heterogeneous biological data such as the Library of Life of the Desert will describe different aspects of the same or related biological systems from molecules to cells and organisms. There are complex relationships between these structures, and understanding and discovering such relationships is an important area in computational biology. We propose to take the next step – to investigate how such relationships can be queried efficiently, how to represent query answers that are paths or in general
topologies of relationships between entities, and to concisely summarize these topologies in a way the investigating scientists understands. We believe that this is a very open research area that involves novel techniques from data management, machine learning, and statistics. Our work could lead towards a general study of relationships between entities and compact representations of these relationships.

**B.3 General Impact**

While we will use the Library of Life of the Desert as a prototype, we believe that our research will have much broader applicability. The processing of queries across structured and unstructured data is an important topic in the general area of search – spanning applications from electronic commerce to web- and intranet search. Queries over graphs are equally important especially with the move towards a graph-structured data model in parts of the scientific community and in managing the graph structure of the WWW. On the educational side, we will develop a new set of courses that integrate topics from data management and information retrieval to prepare students with the right concepts for building the next generation information management infrastructure.

**C. Analysis of genomic sequences**

**C.1 Introduction**

Clearly DNA sequences of a variety of species are one of the most promising approaches to quantitative biology and will be an essential component of the Library of Life. We will initiate the genomics databases of the Library of Life of the Desert with microbial life forms that lived near or in the Dead Sea under extreme salination conditions (halophilic). It is planned to experimentally determine the sequences of microbial genomes from the Dead Sea and to conduct detailed analysis of their genomes to search for clues for their salt resistance. While analyses of this type were performed in the past we expect that (a) the use of multiple genomes will enable comparisons of subtle differences, and (b) the use of advanced annotation schemes, including modeling of protein structures and their interactions models will expand the scope and accuracy of past annotations.

In the context of the Library of Life the detailed analysis of molecular
biology data is of considerable interest for a number of reasons:

- It is unlikely that the Library of Life will have complete and comprehensive data sets on all living species on Earth. An obvious example is of genomic sequences. Even with the reduction in sequencing cost, it is unlikely that complete genomes of all life forms will be available anytime soon. Considerable interpolations, extrapolation, and model building will be required to create from the partial data set a complete picture of life on earth, the desired outcome of the library.

- Working with less than whole genomes poses interesting challenges. For example, fragments of sequences (so-called EST) are relatively easy to obtain but make it necessary to predict properties based on partial and noisy data.

- Evaluations of predictability of biological properties from partial sequence data are essential steps in further planning the Library of Life. This analysis will help us determine what needs to be collected, at what density, and what can be predicted with such high accuracy that field sampling or “wet” experiments could be deferred.

In the first three years the Library of Life will focus on developing and applying tools to convert genomic data into biological, biochemical, and evolutionary information.

The analysis will use two types of tools:

a. Existing tools developed by the Library of Life team and elsewhere. The analysis will use either remote computing facilities, or freely available software installed in the Library of Life computational resources.

b. The creation of new tools and enhancement of existing approaches, as we see fit and necessary.

The two approaches are necessary to add value to the information stored in the Library of Life databases, making the data an important research and educational tool. We comment that the current and widely used analysis tools are still not satisfactory. For most genomes only 30 percent of the genes are annotated in detail. Our experience suggests that by taking integrated approaches including both sequence- and structure-level
analysis the fraction of accurate annotations is likely to increase to 50 percent. Adding expression data, evolutionary information, and network analyses are expected to enhance annotation accuracy even further.

C.2 Building an annotation system

During the first two years we will build on the extensive experience of the CBSU of the Cornell New Life Science Initiative (Computational Biology Service Unit) in the application of a wide range of software tools to timely problems in genomics and proteomics. Examples of some of the widely accessible tools that the Unit maintains for use by the Cornell research community include BLAST, CLUSTALW, PFAM, MODELLER and more (http://cbsu.tc.cornell.edu). We will rely on the Unit expertise in the application of publicly available codes to study the targeted genomes. In particular we will perform a large-scale annotation in which function, structure, variation, and interactions of genes will be predicted.

We will define vectors of features for each genome that will include (for example) sequence, evolutionary diversity, structural environment, secondary structure, exposed surface area, and more. The analysis will assist in the identification and classification of individual genes and their relationships. These vectors of features will be used to assess the quality of predictions of independent publicly available programs, and to combine the predictions in a more reliable way. For each of the proteins with statistically significant prediction of a template we will construct an atomically detailed model. The structures will offer useful starting points for further biochemical and biophysical studies, and will also serve to define alternative and more accurate evolutionary distances between proteins.

The structures, functional annotations, likely protein interactions, and evolutionary analysis will be kept in a SQL database making the comparison, mining, integration, and further computational analysis more straightforward.

We will also add unique annotation and classification tools that the team of the Library of Life will develop to further assess the predictions and to produce higher quality structural models. (Here we can build on our experience in building automated servers for large scale protein modeling -- http://ser-loopp.tc.cornell.edu/cbsu/loopp.htm.)
C.2 Developing a k-mer measure

A novel approach, based on word statistics will be explored and discussed below. Using a linguistic analogy the basic idea is that given only the word count of an unknown book we can readily find its title (provided it is in our database). Indeed, it stands to reason that the word count, or how many times each word appears should provide sufficient information to uniquely identify the title in all but very few cases. Of course, not too many publishers would accept just the word count from an author... Carrying this analogy one step further, using the same basic technique we can identify most of the books that deal with, say Moby Dick: words such as Ahab, Moby, Dick, whale, etc. would be significantly over-represented.

Primary sequences of proteins are significantly shorter than books, and their words are not as clearly defined but the idea is the same. As a surrogate for words we use k-mers of a certain length, say, dimers. While the conventional way to measure the similarity between primary sequences is by computing the score of the optimal global or local alignment, the word based approach would have us compare the dimers frequency tables of both sequences. This approach has a natural extension for comparing a sequence against a family and a family against another family. The underlying assumption is that each family tends to use its own “vocabulary”.

Since the size of a frequency table of k-tuples of amino acids (k-mers) is $20^k$ it is far too large already for $k=4$ for obtaining reliable estimates of all the parameters. An intuitively appealing solution is to reduce the size of the table by reducing the size of the alphabet. More precisely one can partition the 20 letter alphabet into classes such that all the letters in a cluster would be aliased, or replaced by a new symbol representing that class. We propose applying multiple clustering schemes to k-mer frequency tables as a novel approach to the problem of identification. The obtained reduced frequency tables, will be stored and used to compare against a query sequence whose identification is sought.

D. Augmenting the Library of Life from the Literature

As a library in the original sense, the Library of Life of the Desert will have to include all existing and relevant literature to be a comprehensive resource. This requires selecting relevant literature, as well as indexing
and annotating it. However, bringing literature content into more accessible form via manual collection, curation, and annotation is costly and slow. At the end of 2004, the MEDLINE database contained over 15.3 million articles and it was growing at a rate of more than 600,000 records every year. Without automated tools, building and maintaining curated databases from a document collection of this scale and dynamic is beyond the resources available to most researchers. This bottleneck severely limits innovation and efficiency of research in biology and genomics. To overcome this problem, we will develop and provide text-mining methods that

1. Support human annotators and make them more efficient.

2. Completely automate annotation. The gain in coverage and timeliness resulting from fully automated annotation can be beneficial, even if annotation accuracy is lower than for manual annotation.

The proposed project will deliver software tools for automated annotation and curation support. These tools will be made available to the research community, and they will be integrated into the Library of Life to facilitate and expedite the creation of the database. Building on our expertise in curation systems for biological literature, machine learning, and text-mining, some annotation tasks are already solvable with current techniques. For other annotation tasks, we will conduct research on novel methods to improve annotation accuracy and to be able to address annotation problems that are currently not feasible. Our team of researchers combines the necessary expertise to have substantial impact on basic research in biology, machine learning, and text-mining, and on the infrastructure that can facilitate more effective innovation throughout many fields of science.

**D.1 Goals and Tasks**

Our goal is to develop accurate and efficient text-mining methods that make it feasible to build curated databases that allow structured access to the information in the body of biological literature. These text-mining methods will automate parts of the curation process that currently require human labor, as well as assist human curators to achieve better coverage and make their work more efficient. The overall task can be broken into five sub-tasks:
Filtering Irrelevant Articles: Only a relatively small fraction of the 15.3 million articles will contain information relevant to Library of Life. In the first step of processing, we will use automated text classification techniques to filter out irrelevant articles. It is important to achieve high recall on this task, meaning that only a few documents are falsely rejected by the filter. This task is a type of text classification problems that is likely to be solved with high accuracy.

Taxonomic Classification: The relevant articles will be classified into a hierarchy of classes that describe how they are relevant to the Library of Life and what type of information they contain (e.g. information about protein interactions, experimental vs. computational evidence, etc.). This will aid assigning articles to the most appropriate curator. Furthermore, these semantic tags will allow hierarchical browsing for articles and augment keyword search. Again, this is a text classification problem which has been shown to be tractable.

Entity Recognition: The first step of extracting facts from relevant documents lies in recognizing the entities of interest discussed in the text (e.g. gene names, protein names). We will generate automatic extraction rules that recognize entities based on patterns in how their names are composed (e.g. “ease”) and how they are used in text (e.g. “D3241 transcribes ...”). It was shown that this problem is tractable in principle, but that further increases in accuracy are necessary.

Entity Resolution: Entities names are not used consistently between articles. For example, “caspase-3” and “CASP3” refer to the same entity. The task of entity resolution is recognizing that two names refer to the same entity. Resolution can be done based on textual similarity (e.g. sequence alignment), based on contextual use (e.g. co-occurrence), or based on explicit statement (e.g. “caspase-3 (CASP3)” ). Entity resolution is substantially more challenging than entity recognition and further research is likely to lead to substantial improvements in accuracy.

Role Labeling: The final step is to extract stated relationships between entities (e.g. P1 regulates P2). The types of relationships (e.g.}
is_located_in(protein, location), regulates(protein, protein)

that are relevant for the Library of Life are defined in an ontology. Depending on the type of role that is to be extracted, this is again a challenging task where further research is necessary to push the envelope of what is tractable.

Depending on their accuracy, the automated techniques for the five sub-tasks will be used to assist the curators by suggesting annotations, filtering annotations that can be rejected with high confidence and making automatic annotations when the confidence of the automated methods is sufficiently high.

D.2 Approach and Methods

We will build upon our expertise in statistical natural language processing (NLP) tasks like Text Classification, Information Extraction, Tagging and Entity Resolution.

We will approach these statistical NLP tasks using machine learning techniques. Machine learning methods have shown to provide excellent accuracy and have become the methods of choice for Text Classification and many other statistical NLP tasks. In particular, the method of Support Vector Machines (SVM) has become the state-of-the-art for Text Classification. Support Vector Machines have been used extensively for text classification of biomedical text producing state-of-the-art results also in this particular domain. We will build upon our experience with SVM Text Classification and on our software for SVM training (SVM-light) for the tasks of Filtering Irrelevant Articles and Taxonomic Classification.

The other three tasks, however, cannot be phrased as classification problems in a straightforward way. The Entity Recognition task, for example, is more naturally phrased as a tagging problem, where one needs to predict a sequence of tags that indicate a type for each word in the sentence. For example, the sentence “... we show that apoptosis-related cysteine protease is regulated ...” should be tagged as “... *** P P P *** ...” where “P” indicates that the word is part of a protein name, and “*” indicates that the word does not describe any entity. This means that we need learning algorithms that can predict sequences of tags. Note that independently classifying each word is not a viable solution, since it ignores
dependencies between tags. Similarly, for extracting roles between entities, one needs to predict trees to capture the linguistic dependency structure within sentences. And for Entity Resolution one needs to predict an equivalence relation over all objects.

Until recently, these structural prediction problems were either tackled with hand-coded rules, or with generative methods such as Hidden Markov Models (HMMs). Thus, these prediction problems are in a state of development analogous to Text Classification before 1998. At that time, Naïve Bayes was among the methods of choice and the introduction of discriminatively trained regularized methods like SVMs made it possible to substantially improve prediction accuracy. HMMs are the direct analog of Naïve Bayes for sequence prediction and tagging, and machine learning research is currently discovering how to do discriminative training for the prediction of sequences, trees and rankings. In particular, we have recently developed an SVM method that can be used to predict sequences, trees, alignments and equivalence relations, and for which we have empirical evidence that it provides similar benefits over HMMs for these structured prediction tasks as did regular SVM over Naïve Bayes. As part of this project we will conduct research on discriminative training techniques for structured prediction to further advance the accuracy of Entity Recognition, Entity Resolution and Role Labeling.

E. Field characterization of living species and analysis of images: Application to plants

E.1 Overview

A task of the bioinformatic component of the Library of Life of the Desert is to develop enabling technology to classify and catalog living species in the field (the Arava desert), and to add the digital information to the library. Another important task is the ability to analyze images and to extract relevant information from them. Below we describe a machine learning and computer visualization project that addresses the two tasks and is initiated with a focus on plants. We propose to develop a multi-national multi-disciplinary collaboration that will include the following: (i) field work on palmtop computers that will collect digital images, global positioning data, season information, and replies to targeted questions, (ii) initial analysis of the data on the palmtop and data transmission to a central computing facility, (iii) further analysis and data interpretation in the
central facility, and (iv) transmission of the information back to the field to assist ongoing experiments or observations.

This project has a number of technological and scientific goals, as well as challenges for a successful interdisciplinary collaboration.

The first goal is the design of a palmtop device for field work. This device would be used to collect information (such as global positioning data and season information) and digital images, perform initial analysis, submit essential data to the central facility, and receive and display the results.

The focus of the research project here is on mobile computing and wireless databases and communication.

The second goal is to use the data collected in the field. The data are forwarded to the central computing facility for analysis and detailed recording. We are interested in both classification and modeling. While our initial focus will be on classification, we do expect that in the long run, modeling of environmental factors and their interactions with the plant population will be essential. This will require the use of machine learning, computer vision, and data mining techniques, as well as techniques of mathematical modeling.

The third goal is conducting fieldwork and interactions with experts from the life sciences. The collection of the data will require coordination and planning, since the needs of the computational models (see below) are complex and the problems at hand are far from trivial. Of course the classification models that we will develop will have to be tested by experts to obtain further input on the characteristics of the computational models, and their success in modeling biological behavior.

The current proposal is focused on the second and the third goals (the first goal is better completed under the umbrella of the general infrastructure for the Library project).

E.2 Research plan

Learning from images is a challenging task in computer vision and machine learning. We wish to extract maximum information directly from the image and assist (if necessary) the classification scheme by a few concrete questions, and by timing and global positioning data. The project as such presents an intriguing and non-trivial machine learning task, which
is likely to have an impact on the computer science community in addition to biology. The problem identifying objects in pictures is that the conditions in which the image was taken can be highly diverse. Even if the objects appear identical, the image background can be different depending on the time of the day, the season, and the geographical location. It may also depend on the camera, the angle in which the picture was taken, and (can be a major factor!) on the photographer. Of course the objects of interest (here we have plants in mind) are far from identical, which does not make the problem simpler.

Clearly we cannot control all these variables in field conditions. The approach to the problem that we will follow is to take the diversity into account and sample the distributions of the variables to the greatest extent possible. This requirement will lead to a rather intense plant photography experience. For effective learning we will need pictures at different times of the day, in different weather conditions, and at different times of the year. We will need pictures of different angles and different orientations (an early tip is to minimize the background effects and to take the pictures from above), and we will need close-up pictures of leaves and fruits as well as pictures that capture the plant as a whole. Another idea of how to deal with the background variation, is to take pictures in black and white as well as in color. Of course we will also need pictures of different samples of the same plant species. The long list of requests implies that 1,000 pictures of single plant species are not too many for the problem at hand. Based on initial discussions the number of plant species for the desert is expected to be less than one thousand. Obviously, we will need to form a team of bio-photographers to go to the field to provide these pictures. For exploration purposes it will be useful to have an early set of (say) about 50 images of each species.

We wish to start collecting the data before the palmtop technology is developed in full. Rapid data collection is important to facilitate the development of the (highly complex) learning algorithm. It is our hope that once the procedure of learning images is established, the protocol could be extended to other areas and other living organisms. We envision using standard digital cameras with resolution of two to three megapixels. With this resolution in mind 2GB of storage will make it possible to store about 1,000 pictures on a single camera. Besides the pictures we will need to record the global positioning data, the date and time, and answer query questions prepared by biologists.
We will use the GPS, time information, and the image in machine learning techniques. The data will be processed to compute numerical characteristics of the system (features). The features that we will learn are not always the obvious human way of recognizing an object, but they are effective approaches that were tested in the past. For example, a useful feature in learning images is the (non-intuitive) histogram of pixel colors. We will compute a series of features for each plant species from each of the thousands images produced during the learning process. These features will be combined in a neural network, a support vector machine, or other machine learning model to suggest the most likely plant classification. The computational model, which will reside at the central computing facility, will have wireless access from the field. The set of features will be computed for a new plant image and will be passed through the classification scheme with the additional information (GPS and date) to provide a result to the field workers. The result may be the final classification or a few more queries to pinpoint the result (see below).

As mentioned earlier, learning from images is a challenging ongoing research topic in computer vision and machine learning. It is therefore possible that our success in classifying the plants according to time, place, and image only will not be sufficiently accurate or conclusive. We will therefore have plan B that will be partially based on probe questions in addition to the image and the geographical data (e.g. is the leaf rough or smooth?). Of course, our goal is to reduce overhead to a minimum and to use the smallest number of queries. The queries will be analyzed in conjunction with the digital information collected automatically to provide conclusive and accurate classification.
Biographies of Key Personnel

Ron Elber

Professor
Director, Ithaca campus of the Tri-institutional Program in Computational Biology
Director, Computational BiologyService Unit
ron@cs.cornell.edu
http://www.cs.cornell.edu/ron/

Ron Elber obtained a bachelor’s degree in chemistry and physics in 1981, and a Ph.D. in theoretical chemistry in 1984 at the Hebrew University of Jerusalem. He was a postdoctoral fellow in theoretical biophysics from 1984–1987 at Harvard University. Ron was on the chemistry faculty of the University of Illinois (1987–1992) and on the chemistry and biology faculty at Hebrew University (1992–1999). Since 1999 he has been on the CS faculty at Cornell where he is currently a full professor.

Ron’s research is in computational biology and bioinformatics. His group is developing novel tools (like MOIL) to simulate dynamics of biological-macromolecules. His current research focuses on algorithms to extend the time scales of simulations, and to study complex processes such as the kinetics of protein folding. Ron’s techniques for path following and enhanced sampling are in wide use and motivated the development of related algorithms. His bioinformatics investigations focus on protein annotation, using sequence-to-structure matches (LOOPP). LOOPP linked a gene that influences the size of the tomato fruit with a human protein that controls cell growth and may cause cancer.
Johannes Gehrke
Associate Professor
Faculty Associate Director, Cornell Theory Center
johannes@cs.cornell.edu
http://www.cs.cornell.edu/johannes/

Johannes Gehrke obtained his Ph.D. in computer science at the University of Wisconsin at Madison in 1999, and he has been a Professor in CS since then.

Johannes’s research interests are in the areas of data mining and database systems. With Professor Al Demers, he is working on distributed data management for wireless sensor networks. With Professor Jayavel Shanmugasundaram, he is building a peer-to-peer database system that scales to thousands of nodes. He is also interested in techniques for processing high-speed data streams and in data privacy. His data-mining research includes privacy-preserving data mining, theoretical foundations of data mining, and applications of data mining to problems in the sciences. His group has developed some of the fastest known algorithms for several important data-mining tasks.
Uri Keich
Assistant Professor
keich@cs.cornell.edu
http://www.cs.cornell.edu/~keich/

Uri Keich received his Ph.D. in mathematics from the Courant Institute in New York City in 1996, and his M.Sc. in mathematics from Technion in Israel in 1991.

Before coming to CS at Cornell, he was a project scientist at the Department of Computer Science and Engineering of the University of California at San Diego, and assistant professor at the Department of Mathematics of the University of California at Riverside until 2000. He was also a Von Karman Instructor at the Applied Mathematics Department of the California Institute of Technology.

Keich’s research interests include statistical and algorithmic problems that arise in areas of bioinformatics such as motif finding, seed design for similarity search, and sequence assembly.
Dan Huttenlocher

Professor
dph@cs.cornell.edu
http://www.cs.cornell.edu/~dph/

Dan Huttenlocher received a dual degree in computer science and experimental psychology from the University of Michigan in 1980, and master’s and Ph.D. degrees in computer science from M.I.T. in 1984 and 1988, respectively. He has been on the CS faculty since 1988. He holds a joint appointment with the Johnson Graduate School of Management at Cornell.

Huttenlocher’s research interests are in computer vision, computational geometry, electronic-collaboration tools, financial-trading systems, and the principles of software development. In addition to teaching and research, Dan has considerable experience managing software-development efforts in corporate and academic settings. He has been chief technical officer of Intelligent Markets, a leading provider of advanced trading systems. He also spent more than ten years at the Xerox PARC, directing work that led to the ISO JBIG2 image-compression standard, and serving as part of the senior management team.

Huttenlocher has been recognized on several occasions for his teaching and research, including being named a Presidential Young Investigator by the NSF in 1990, the New York State CASE Professor of the Year in 1993, and a Stephen H. Weiss Fellow by Cornell in 1996. He holds twenty-four U.S. patents and has published more than fifty technical papers, primarily in the areas of computer vision and computational geometry.
Rich Caruana

Assistant Professor
caruana@cs.cornell.edu
http://www.cs.cornell.edu/~caruana/

Rich Caruana obtained his Ph.D. in computer science from Carnegie Mellon University in 1998. Currently he is an assistant professor in CS, where he does research in machine learning and data mining. His current focus is on ensemble learning, inductive transfer, adaptive clustering, learning rankings, and applications of these learning methods to problems in medical decision making and bioinformatics.

Learning rankings is an exciting area in machine learning where the goal is not to predict a classification or value for an item, but to predict an ordering for a set of items. Caruana is developing algorithms that learn rankings for problems in medical decision-making where it is difficult to assess absolute risk for any given patient, but easier to learn to order patients by relative risk. One method he developed to learn rankings outperformed a dozen other machine learning methods in a large multi-institutional pneumonia-risk prediction project. In 2000–2001 Caruana led a team of researchers that developed the first automated system for the early detection of bioterrorist releases of anthrax. The system applies data mining to consumer purchases in supermarkets to look for unexplained increases in the sales of products such as cough syrup that may signal the onset of symptoms from a recent attack. Because consumers tend to self-medicate using easily available products such as cough syrup and throat lozenges before consulting physicians, the system can detect the onset of flu-like symptoms twenty-four to forty-eight hours before these can be detected by visits to hospitals and doctors offices.

A theme that runs through all of Professor Caruana’s work is the importance of developing methods that are effective on real-world problems. He likes to mix algorithm development with applications work to insure that the methods he develops are useful in practice.
Thorsten Joachims
Assistant Professor
tj@cs.cornell.edu
http://www.cs.cornell.edu/People/tj/

Thorsten Joachims is an Assistant Professor in the Department of Computer Science at Cornell University. In 2001, he finished his dissertation with the title “The Maximum-Margin Approach to Learning Text Classifiers: Methods, Theory, and Algorithms,” advised by Prof. Katharina Morik at the University of Dortmund. His research interests center on a synthesis of theory and system building in the field of machine learning, with a focus on Support Vector Machines and machine learning with text. He authored the SVM-light algorithm and software for support vector learning.

Professor Joachims’s research interests center on a synthesis of theory and system building in the field of machine learning, with a focus on support-vector machines, text mining, and machine learning in information access. In particular, he has worked on WebWatcher, an adaptive browsing assistant for the Web. He has developed and authored the SVM-Light algorithm and software for inductive and transductive support-vector learning. His most recent work is on learning from click-through data in search engines, and on discriminative training for predicting complex multivariate objects.
Library of Life

Bridging the Rift Foundation
60 Columbus Circle
New York, NY 10023
Phone: 212-801-3745
Fax: 212-801-1056
Email: info@bridgingtherift.org