GWU MATHEMATICS COLLOQUIUM
Mathematics and the integration of the sciences
May 4, 2006

Interaction between Mathematics and Molecular Biology

A Collaborative Research Project
Robert Donaldson (Biology), **Akos Vertes** (Chemistry)
Chen Zeng (Physics), **Yongwu Rong** (Mathematics)
On behalf of
the Computational Molecular Biology and Bioinformatics group (CMBB)

In the past decade, there has been a great deal of increased interaction between mathematical and biological sciences. Much of this has been motivated by exciting new advancements in molecular biology, which have raised new challenging problems in mathematics. This is reminiscent of the interaction between mathematics and physics in early 20^{th} century, which led to revolutionary developments in relativity theory and Riemannian geometry. We believe this is the right time for us to work together to face such new challenges and opportunities.

We see two specific directions that we wish to focus on: Complexity and Folding. Details will be discussed below.

1. Complexity – understanding the behavior of systems with a large number of interacting components.

Interactions of Molecules (1000's of different proteins) inside a cell in a living organism, System Biology, is the emerging field in the life sciences. For example, plants can detect and respond to the amount of nitrogen present in the soil (or oxygen in the air). Many different protein molecules (and genes) are responsible for the detection and response. One way this type of problem is confronted is to see what genes are "turned on" (or off) in response to nitrogen. More specifically when there is a change in the amount of nitrogen available certain genes are turned on very quickly and others later. Even though the functions of the genes are not known the hierarchy can provide clues to their interactions in a logical network. Microarray technology tells us what genes are being turned on or off at a particular time. But then we need mathematical and statistical approaches to analyze the information and to model the networks of interactions among the genes.

2. Folding – Understanding the combinatorics and geometry of protein (or RNA) folding.

A protein (or RNA) molecule is initially produced as a relatively long linear string. Somehow the string folds upon itself in a precise way, and usually only one way out of what would appear to be a very large number of possibilities. Chemical forces among different regions of the string, and interactions with the surrounding water molecules, propel the folding to the one most favorable configuration of the string. Modeling and predicting this is very difficult and requires all the mathematical, statistical and computational tools that we can create.

Mathematical Issues.

1) **Combinatorics.** The interactions in a complex system can be described by a graph. The study of folding is also related to graph theory.
2) **Computer Sciences.** Creating algorithms to identify potential genetic information in DNA strings from different organisms.
3) **Dynamics.** Chaos, non-linear dynamics. Physical problems such as the behavior of fluid droplets in an electrical field at the tip of a capillary, as would be encountered in an instrument that is being designed for chemical analyses of (the molecules) in fluid.
4) **Statistics.** Genetics and bioinformatics related to disease, analyses of microarray data showing what genes are active in diseased tissues (e.g. cancer, viral infection), analyses of DNA strings for meaningful information (e.g. what information in the string is directly responsible for the creation of particular protein molecules.
5) **Topology.** The secondary structure of RNA folding can be described by the *Feynman diagram*, introduced by Feynman to study interactions between particles. Surprisingly, they turned out to be powerful tools in studying the *quantum invariants* in contemporary topology. Furthermore, they arise naturally in the study of RNA folding. We expect some methods in topology can be applied to the study of RNA folding.

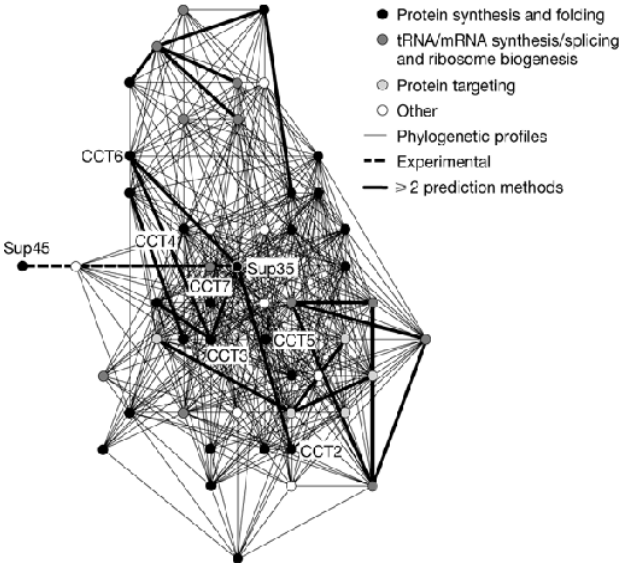

Figure 1. Complexity of a Biological Network

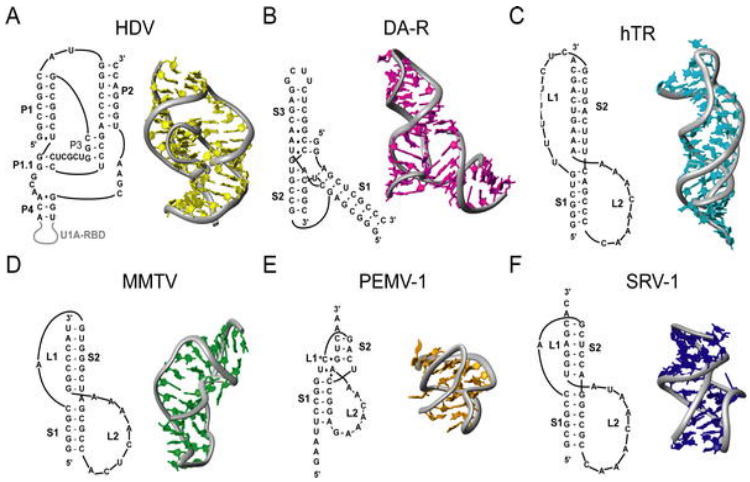

Figure 2. Geometry of RNA folding.

Formal Learning Theory Based on Inductive Inference Machines

A joint research project by
Michele Friend, Philosophy (michele@gwu.edu)
Valentina Harizanov, Mathematics (harizanv@gwu.edu)

"*How comes it that human beings, whose contacts with the world are brief and personal and limited, are nevertheless able to know as much as they do know*?"
Bertrand Russell

Introduction: Problem of induction in the philosophy of science

Inductive inference, the process of hypothesizing a general rule from evidence (examples), is an important component of intelligent behavior. Philosophers have observed since antiquity that if we are interested in general conclusions, the examples typically do not logically entail them. Carnap characterized inductive learning as ampliative, in the sense that the conclusion reaches beyond the information carried in the premises. In contrast, he characterized deduction as non-ampliative; the conclusion of a deductively valid argument is always weaker than, or as strong as, the premises. However, it seems that empirical methods should reliably deliver the truth just as logical methods do. Hence, to find an alternative response to the uncertainty of general conclusions by evidence, we adopted a general framework of inductive inference dating back to the 1960's, from the work of Gold, a mathematician, and Putnam, a philosopher.

In this framework, an inquirer I, who is attempting to correctly describe some unknown rule R, is given larger and larger sequences of examples (data) of R:

(x_0), (x_0, x_1), (x_0, x_1, x_2), …

while generating and successively modifying hypotheses C_0, C_1, C_2, … that are approximating descriptions of R:

(x_0), C_0; (x_0, x_1), C_1; (x_0, x_1, x_2), C_2; …

If after some step $n-1$, the inquirer I stops modifying hypotheses:

$C_0, C_1, C_2, \ldots, C_n, C_n, C_n, C_n, \ldots$

and the hypothesis C_n is a correct description of R, then I correctly identifies (infers) R in the limit, on this sequence of examples. The Inquirer has converged to the correct answer. This does not sit well with Carnap's distinction between ampliative and non-ampliative reasoning.

Philosophers now have some work to do. We begin with further enquiry into the notion of "correctly identifying." For, in learning theory we can even allow that C_n, C_{n+1}, C_{n+2}, C_{n+3}, … are not the same, but equivalent, in the sense that they correctly describe R. A philosophical forerunner of this idea is Peirce's notion of finding truth "in the limit of inquiry." Although successful inductive inquiry arrives at the truth after finitely many steps, an inquirer can be in possession of the truth without being certain that he is. For this reason, the inductive process continues *ad infinitum*. Specifically, the inquirer I cannot determine whether he has converged to a correct hypothesis since new examples may or may not conflict with the current conjecture. In philosophy, such idea of uncertainty surfaces as a cluster of problems: Hume's problem of induction, Wittgenstein's "Rule Following Considerations" and Goodman's "Grue Paradox."

Fundamentals of formal learning theory

Gold's inductive inference overlaps with learning from examples, although learning is a much broader and more complex act of gaining knowledge. It also relates to the formal study of language acquisition in linguistics. In fact, Gold's theory of inductive inference is also called "formal learning theory." The exact methods used in this theory come mainly from computability theory (recursive function theory). We simplify the study of language learning by using Gödel's coding by natural numbers of syntactic objects, including sentences, grammars, and programs. Since in most cases the learner is algorithmic, we can identify him with a Turing machine, or via coding, with a computable function on natural numbers.

In formal (i.e., machine) languages, a sentence is a finite string of symbols drawn from some fixed, finite alphabet. A formal language in Chomsky's hierarchy is defined to be the set of well-formed sentences obtained from some given set of symbols by application of production rules—the grammar of the language. Learning a language is equivalent to finding a program for a grammar that generates (enumerates) the language. Grammars are *equivalent* if they generate the same language. Not every Chomsky language is decidable. That is, the existence of a generative algorithmic grammar does not guarantee the existence of a recognition algorithm for the language. However, all context-sensitive, and hence all context-free and regular languages are decidable.

An outside source (often called 'teacher') feeds data to the learner from the language to be learned. Positive data are the correct sentences, and negative data are the incorrect sentences in proper vocabulary. The learner's hypotheses are grammars (coded by natural numbers). For the learning process to be successful, the learner's sequence of hypothesized grammars has to converge to correct grammars for the language being learned.

There are two main notions of convergence in this context. In *explanatory* (syntactic) learning, *EX*-learning, a learner after finitely many steps, always outputs a hypothesis for the *same* grammar generating the language to be learned. In *behaviorally correct* (semantic) learning, *BC*-learning, a learner, after finitely many steps, outputs hypotheses for possibly different grammars, but all generating the language to be learned. So, the hypotheses have not converged onto a unique grammar, but onto equivalent and correct grammars. While, *EX*-learnability clearly implies *BC*-learnability, it can be shown that the converse does not hold.

The teacher and the learner follow a specified protocol, depending on the type of learning. There are several protocols considered. Learning from *text* is learning from positive data only, and eventually (in the limit) all positive data are fed to the learner. In the case of learning from an *informant*, all data, both positive and negative, are eventually fed to the learner. Learning from text is much more restrictive than learning from an informant.

The interplay between philosophers and mathematicians over problems of induction dates back to the 1960's. However, the exchange has been scant. Friend and Harizanov have collaborated to make the exchange more robust by contributing to, and editing, a volume of papers written by philosophers and mathematicians on the problems of inductive inference. The manuscript for the volume will be sent to Springer. We hope to foster further collaboration.

Future research problems

We focus on general theoretical properties and models of inference methods, rather than on finding specific, practical inference methods.

(1) *How can we develop algorithmic learning theory for languages more complicated than Chomsky languages, in particular, ones closer to natural language?*

We know that Chomsky languages exactly correspond to those enumerated by Turing machines—these can be coded by computably enumerable sets. Gödel showed that computably enumerable sets are exactly those that can be defined by existential formulae in arithmetic, where arithmetic is the standard model of natural numbers with addition and multiplication. These languages are at the first level in the arithmetical hierarchy. At the 0^{th} level are the decidable languages, which are exactly those definable by quantifier-free formulae. The hierarchy continues higher and higher (and even transfinitely) giving more and more complicated languages. The pleasing feature is that there is exact correspondence between computability and definability, where the complexity of definability is measured by the number of quantifier alterations.

Case and Royer are developing a learning theory for certain languages in the second level of the arithmetical hierarchy. These languages, whose grammars they call correction grammars, are the *differences* of Chomsky languages. The difference languages are those with two Chomsky-type grammars: one adds the sentences, and the other one occasionally subtracts some of these

sentences. There is a corresponding difference hierarchy with more and more complicated languages, where complexity is measured by how many times the same sentences can be added and subtracted.

Another related study has recently been initiated by Stephan, Ventsov and Harizanov by generalizing learning within mathematics. The generalization is from learning languages coded by sets, which are collections of distinct elements that are mutually independent, to learning mathematical structures in which various elements are interrelated. Thus, by learning one element, the learner automatically learns all dependent elements.

(2) *How can we formally define and study certain learning strategies?*
There have already been attempts to formally capture some important concepts. For example, *consistent* learning has been introduced, as well as *Popperian* learning, and also *confident* learning, *reliable* learning, *decisive* learning, etc.

An example of ongoing research in this area is the work of Case et al. in U-*shaped* learning, which is motivated by recent discoveries in developmental and cognitive psychology. Psycho-linguists studying how children learn verbs have observed that they usually conjugate even irregular verbs correctly when they are first learning to speak a language. They then go through a phase of treating irregular verbs as regular, and then they start to conjugate verbs correctly again. Mathematical models of learners have been developed that mimic this learning-unlearning-relearning pattern.

(3) *What is the significance of negative versus positive information in the learning process*? While learning from text provides no negative information to the learner, learning from an informant gives all negative information, together with all positive information. There are also intermediate learning protocols with regard to the amount of negative information available to the learner. Jain, Stephan and Harizanov have recently studied learning by *switching*, using an analogy with limit-computable functions in computability theory. In this protocol, the learner can request positive or negative data from the teacher, but it is only allowed a finite number of switches, from positive to negative, or *vice versa*. It can be shown that learning from switching provides more learning power than learning from text, but it is weaker than learning from an informant.

(4) *What are good formal frameworks that unify deduction and induction*?
Martin, Sharma and Stephan recently proposed such a framework by using parametric logic. Their approach is model theoretic, based on the Tarskian notion of logical consequence. The difference between deductive and inductive consequences lies in the process of deriving a consequence from the premises. Parametric logic combines fundamental notions from formal learning theory with notions from logic and topology. Again, the philosophical importance of distinction between induction and deduction, as it is characterized in parametric logic, is a delicate matter, since it is not clear that differences in the process of drawing inferences are philosophically significant. This is an open problem.

A Bayesian Approach to Lattice QCD Data

A joint research project by
Frank Lee, Associate Professor of Physics (fxlee@gwu.edu)
Nozer Singpurwalla, Professor of Statistics (nozer@gwu.edu)
Cornelius Bennhold, Professor of Physics (bennhold@gwu.edu)

The Context

Quantum Chromodynamics (QCD) is the fundamental theory of the strong nuclear force, the force that binds quarks and gluons inside protons and neutrons which make up the nucleus at the heart of an atom. The strong force is one of the four fundamental forces in nature, besides electromagnetism, gravity and the weak force responsible for radioactive decays. Unraveling the structure of matter at its deepest level as governed by QCD is key to our understanding of the physical world, and presents one of the most challenging tasks in contemporary nuclear and particle physics. Although QCD is simple to write down, it is notoriously difficult to solve because of the unique structure of the strong interactions among quarks and gluons. At present, the only known way to solve QCD is by large-scale supercomputer simulations; a method that goes by the name of *lattice QCD.*

The Problem

Let us assume we want to determine the mass of the proton and its excited states (resonances) in lattice QCD. An example of the kind of quantity needed for such a calculation and generated by supercomputers in lattice QCD is the so-called correlation function which generally takes the form of $G(t) = \sum_{n=1}^{\infty} A_n e^{-E_n t}$. Here, the E_n's are the masses of the proton and its resonances under study, and the A_n's are amplitudes which quantify the probability of a particle being produced. They are the physics quantities we are after. The energies are ordered, $E_1 < E_2 < E_3 < \ldots$, and the amplitudes can be normalized between 0 and 1. The number of correlation functions in a calculation is finite since they come from a finite number of Monte-Carlo "measurements" on the computer. The conventional way is limited to looking at this function at large time where the ground state (the first term in the sum) becomes dominant. The problem is to fit an infinite number of E_n's and A_n's using only a finite number of computer-generated values of G(t). It is well known that simply including two or more exponential terms in a conventional χ^2-fitting algorithm does not work because of large systematic errors. Thus, new algorithms that go beyond the conventional are needed, calling for sophisticated statistical methods.

The Research Plan

We are investigating a new method that is rooted in Bayesian statistical inference to tackle the problem. It is based on the introduction of Bayesian priors. In this case, the prior knowledge is from our knowledge of nature: the fact that the masses are ordered (proton resonances are always heavier than the proton itself), the amplitudes are normalized (the probability of producing a proton or one of its resonances can never be larger than 1), or even the fact that they are positive numbers. The method goes by the name of constrained curve-fitting with Bayesian priors. Among its many benefits are that it uses all data points in the correlation function, not just the few slices at large time where the signal becomes noisier. It can use multiple terms in the fitting model, not just the first term, thus providing access to excited proton states in a systematic manner. In practice, the number of terms to include is determined by the quality of convergence of the fit. The major issue with this method is how to obtain the Bayesian priors. Prof. Lee has done previous work that tried to seek information on the priors from a subset of the data to be analyzed (sometimes referred to as an "empirical" Bayesian method,) then use that information on the rest of the data set. In order to develop a truly Bayesian approach a collaboration ensued between statistics and physics, with Prof. Singpurwalla proposing a new method to deal with the problem. This new algorithm is completely Bayesian and self-contained; and if successful, would represent an important breakthrough in applying Bayesian statistics to a challenging problem in physics.

Furthermore, the algorithm may find applications in other fields that have a similar structure to the problem analyzed here.

Other Mathematical/Statistical Issues in Lattice QCD

6) Maximum Entropy Method (MEM)

It has the same foundation in Bayesian statistical inference, plus uses the principle of "Maximum Entropy". MEM has been widely used in other fields, such as condensed-matter physics, astrophysics, image reconstruction and so on. MEM has three important features: 1) it makes no *a priori* assumptions on the shape of the spectral functions, 2) for a given data set, a unique solution is obtained if it exists, and 3) the statistical significance of the solution can be quantitatively analyzed. Prof. Lee has successfully implemented the algorithm, and has begun `blind' tests to reproduce artificial data. The challenge now is to apply the algorithm to real lattice QCD data.

7) Jackknife and Bootstrap

Data generated with Monte-Carlo methods are correlated. The two methods can give realistic estimates of the statistical errors in such data. How can these errors be incorporated into Bayesian methods for a reliable error analysis?

8) Markov Chain Monte-Carlo (MCMC)

This is the method that enables the study any Quantum Field Theory on the computer. The method generates new configurations of fields from old configurations by an update procedure that uses a random number generator and that yields a distribution proportional to the exponential of the action. The expectation value of an observable is computed by averaging over the observable evaluated over the ensemble of configurations. How to improve the update procedure in dynamical lattice QCD is a major challenge.

9) Matrix Inversion Algorithms

The basic mathematical problem is to solve the sparse linear system M*x=b where M is a sparse non-Hermitian matrix for which the eigenvalues are complex. The matrix is typically on the order of a few million by a few million. Better and more efficient algorithms to invert such a matrix are much in demand.