mit primes

THIRD ANNUAL CONFERENCE

MAY 2013

MIT PRIMES: Program for Research In Mathematics, Engineering, and Science for High School Students

Third Annual Conference

Saturday, May 18

Section I. Mathematics

9:00 am Welcoming remarks
Prof. Michael Sipser, Head of the MIT Mathematics Department
Prof. Pavel Etingof, PRIMES Chief Research Advisor
Dr. Slava Gerovitch, PRIMES Program Director

9:15 am Session 1
Jonathan Tidor, *Extremal functions of pattern avoidance in matrices* (mentor Jesse Geneson)
Rohil Prasad, *Investigating GCD in Euclidean domains* (mentor Dr. Tanya Khovanova)
Kavish Gandhi and Noah Golowich, *Inequalities and partition regularity of linear homogenous equations* (mentor Laszlo Lovasz)

10:35 am Session 2
Jin-Woo Bryan Oh, *Towards generalizing thrackles to arbitrary graphs* (mentor Rik Sengupta)
Raj Raina, *Minimum degrees of minimal Ramsey graphs* (mentor Andrey Grinshpun)
Junho Won, *Highly non-convex graph crossing sequences* (mentor Chiheon Kim)

11:50 am Session 3
Leigh Marie Braswell, *The Cookie Monster Problem* (mentor Dr. Tanya Khovanova)
Saarik Kalia and Michael Zanger-Tishler, *Good functions and multivariate polynomials* (mentor Tue Ly)

2:00 pm Session 4
Ying Gao, *Depths of posets ordered by refinement* (mentor Sergei Bernstein)
Vahid Fazel-Rezai, *Equivalence classes of length-changing replacements of size-3 patterns* (mentor Dr. Tanya Khovanova)
William Kuszmaul, *On* q-*enumeration of modular statistics* (mentor Darij Grinberg)

3:15 pm Session 5
Gabriella Studt, *Higher Bruhat order on Weyl groups of Type B* (mentor Daniel Thompson)
Ravi Jagadeesan, *Belyi functions with prescribed monodromy* (mentor Akhil Mathew)
Ritesh Ragavender, q-*analogues of symmetric polynomials and nilHecke algebras* (mentor Alex Ellis)

4:30 pm Session 6
Jeffrey Cai, *Orbits of the symplectic group on partial flag varieties of type A* (mentor Vinoth Nandakumar)
Isaac Xia, *Quotients of lower central series over* Z *with multiple relations* (mentor Dr. Yael Fregier)

Sunday, May 19

Section II. Computer Science

9:00 am <u>Welcoming remarks</u>
Prof. Srini Devadas, MIT Department of Electrical Engineering and Computer Science
Dr. Slava Gerovitch, PRIMES Program Director

9:15 am <u>Session 7</u>
William Wu and Nicolaas Kaashoek, *How to teach a class to grade itself* (mentors Christos Tzamos and Matt Weinberg)
Anish Athalye and Patrick Long, *Performance analysis and optimization of skip lists for modern multi-core architectures* (mentors Austin Clements and Stephen Tu)
Ajay Saini, *Modeling the opinion dynamics of a social network* (mentor Dr. Natasha Markuzon)

10:40 am <u>Session 8</u>
Istvan Chung and Nathan Wolfe, *A collaborative editor in Ur/Web* (mentor Benjamin Barenblat)
Oron Propp and Alex Sekula, *Automating interactive theorem-proving with Coq and Ltac* (mentor Drew Haven)
Nihal Gowravaram, *Avoidance in (2+2)-free posets* (mentor Wuttisak Trongsiriwat)

12:05 pm <u>Session 9</u>
Steven Homberg, *Finding enrichments of functional annotations for disease-associated single-nucleotide polymorphisms* (mentor Dr. Luke Ward)
John Long, *Evidence of purifying selection in humans* (mentor Angela Yen)

2:00 pm <u>Session 10</u>
Dr. Gil Alterovitz, Division of Health Sciences and Technology, Introductory remarks
Ben Zheng, *Removing disorder in drug resistance-related proteins in tuberculosis through hill-climbing algorithms* (mentor Dr. Gil Alterovitz)
Peijin Zhang, *Leveraging disordered-ordered interactions to yield new targets and drugs for tuberculosis* (mentor Dr. Gil Alterovitz)

3:00 pm <u>Session 11</u>
Jonathan Patsenker, *Finding the binding sites of MoRFs on a partner protein* (mentor Dr. Gil Alterovitz)
Yishen Chen, *SMART Genomics API* (mentor Dr. Gil Alterovitz)
Skanda Koppula, *Prediction of disease by pathway-based integrative genomic and demographic analysis* (mentors Dr. Gil Alterovitz and Dr. Amin Zollanvari)

Section III. Computational and Physical Biology

4:10 pm **Session 12**

Prof. Leonid Mirny, Division of Health Sciences and Technology and Physics Department, Introductory remarks

Boryana Doyle, *Chromatin organization: from polymer loops to topological domains* (mentors Geoffrey Fudenberg and Maxim Imakaev)

Carolyn Lu, *Dynamic folding of chromatin domains by active SMC-mediated loops* (mentors Geoffrey Fudenberg and Maxim Imakaev)

Ashwin Murali, *Lineage-dependent properties of 16S ribosomal RNA nucleotide composition* (mentors Geoffrey Fudenberg and Maxim Imakaev)

Hao Shen, *The impact of gene order on evolution* (mentor Anton Goloborodko)

2013 PRIMES Conference abstracts

Section I: Mathematics

Session 1

Jonathan Tidor

Extremal functions of pattern avoidance in matrices

Mentor Jesse Geneson

Project suggested by Jesse Geneson

A 0-1 matrix is an array of numbers, all of which are either 0 or 1. Such a matrix is said to contain another matrix if the pattern of ones in the smaller matrix can be found in the larger. The weight extremal function $ex(n, M)$ is defined to be the maximum number of one entries in an $n \times n$ matrix that avoids M.

Weight extremal functions have found applications in the convex polygon unit-distance problem, in sequence extremal functions and Davenport-Schinzel sequences, and in the Stanley-Wilf conjecture.

We examine the rectangular weight extremal function, a simple generalization of the normal version to $m \times n$ matrices. We introduce a property known as separability and demonstrate its basic properties including a fundamental connection to the idea of linearity.

Rohil Prasad

Investigating GCD in Euclidean domains

Mentor Dr. Tanya Khovanova

Project suggested by Dr. Ben Hinkle and Dr. Stefan Wehmeier (MathWorks)

We attempt to optimize calculation of the greatest common divisor for elements of either the domain $\mathbb{Z}[\sqrt{2}]$ or $\mathbb{Z}[\sqrt{3}]$. We use three general methods in order to attempt to create an algorithm that is not only asymptotically but also practically fast in its calculation. We compare these methods and conclude for now that the basic Euclidean algorithm is actually the fastest.

Kavish Gandhi and Noah Golowich

Inequalities and partition regularity of linear homogenous equations

Mentor Laszlo Lovasz

Project suggested by Prof. Jacob Fox (MIT)

Rado's Theorem states that given the linear homogenous equation $c_1x_1 + c_2x_2 + \ldots + c_nx_n = 0$, if some nonempty subset of the c_i sum to 0, then the equation is regular over the integers. The problem of adding inequalities of the form $A_{j1}x_1 + A_{j2}x_2 + \ldots + A_{jn}x_n \neq 0$ to linear homogenous equations is investigated in this paper; we prove that adding any finite number of such inequalities does not affect the regularity of a regular equation. Furthermore, we prove that adding a finite number of inequalities does not affect the degree of regularity of the family of equations L_k proven by Alexeev and Tsimerman (2010) to be $k-1$-regular but not k-regular. Finally, we investigate conditions on a linear homogenous equation that guarantee 2-regularity and whether the 2-regularity of these equations is affected by the addition of a finite number of inequalities.

Session 2

Jin-Woo Bryan Oh

Towards generalizing thrackles to arbitrary graphs

Mentor Rik Sengupta

Project suggested by Prof. Jacob Fox (MIT)

In this talk we first review and recapitulate the historical notion of thrackles, and review most of the known results and conjectures about these objects. A thrackle drawing is a graph embedding where no edge crosses itself, but every pair of distinct edges intersects each other exactly once; this point of intersection is allowed to be a common endpoint. A thrackle is a graph that admits a thrackle drawing. Thrackles, however, turn out to be very sparse in the set of all graphs. Hence we explore the question of generalizing the notion of thrackles to arbitrary graphs by defining special graph embeddings called near-thrackle embeddings for any graph G. We then explore some of the properties of these near-thrackle embeddings, including a number of conjectures that seem intuitive and very obviously true.

Raj Raina

Minimal Ramsey graphs

Mentor Andrey Grinshpun

Project suggested by Prof. Jacob Fox (MIT)

For graphs F and H, we write $F \to H$ if every 2-coloring of the edges of F contains a monochromatic copy of H. The graph F is H-minimal if the deletion of any edge or vertex of F results in a new graph F' with $F' \not\to H$. Burr, Erdos, and Lovasz defined $s(H)$ to be the minimum degree of F over all H-minimal graphs for H. Using an extension of a technique developed by Burr, Erdos, and Lovasz we determine that $s(K_t - edge) = (t-2)^2$. This is the first time $s(H)$ has been determined for a very well-connected class of graphs H which are not vertex-transitive. Moreover, we determine a non-trivial bound for $s(K_t - matching)$. We hope that this will be a first step in determining $s(G(n,p))$.

Junho Won

Highly non-convex graph crossing sequences

Mentor Chiheon Kim

Project suggested by Prof. Jacob Fox (MIT)

Graphs are abstract mathematical objects defined as a collection of a set of vertices and a set of edges that connect the vertices. In order to visualize graphs, we can draw them on a surface, corresponding vertices with points and edges with curves. When drawn on a plane, planar graphs can be drawn without crossings (curve intersections), whereas non-planar graphs cannot. In this talk I will explain how the notion of non-planarity can be generalized to drawings on different (orientable) surfaces, and present crossing sequences which provide an interesting generalized measurement of non-planarity of graphs. Specifically, with the goal of characterizing crossing sequences, I will construct a new example of a highly non-convex crossing sequence of arbitrary length, for which there are only few examples known.

Session 3

Leigh Marie Braswell

The Cookie Monster Problem

Mentor Dr. Tanya Khovanova

Project suggested by Dr. Tanya Khovanova

In 2002, the $Inquisitive\ Problem\ Solver$ created a hungry cookie monster who wants to empty a set of m jars filled with various numbers of cookies, recorded in a sequence $S = (s_1, ..., s_m)$. On each of his moves, he may choose any subset of jars and take (or devour immediately) the same number of cookies from each of those jars. The cookie monster number of S, $CM(S)$, is the minimum number of moves the cookie monster must use to empty all of the jars.

We will discuss the three generally useful algorithms Michael Cavers suggested as cookie monster strategies. They are the $Empty\ the\ Most\ Jars\ Algorithm$, in which the monster empties as many jars as he can for each move; the $Take\ the\ Most\ Cookies\ Algorithm$, in which the monster takes as many cookies as possible for each move; and the $Binary\ Algorithm$, in which the cookie monster takes 2^k cookies from all jars that contain at least 2^k cookies for k as large as possible. None of these algorithms are optimal in all cases. We will discuss bounds of $CM(S)$ for all S. It is easy to show that for all S, $\lceil \log_2(m+1) \rceil \leq CM(S) \leq m$. We will also discuss for which S these bounds may be improved.

We also will present our cookie monster with interesting sequences of cookies in his jars. For example, we challenge our monster to empty a set of jars containing cookies in the Fibonacci sequence, and prove that when $S = \{F_2, \ldots, F_m\}$, then $CM(S) = \lceil \frac{m}{2} \rceil$. This may be generalized for n-naci sequences to show that jars containing the first distinct $m-(n-1)$ n-naci numbers may be emptied in $\lceil \frac{n-1}{n} m \rceil - (n-2)$ moves.

The proof of the n-nacci claim suggests that if the initial distribution of cookies satisfies certain inequalities, we may find a closer lower bound for $CM(S)$. We analyze the result of several of these inequalities. We will also discuss how the growth of interesting sequences is related to the bound on $CM(S)$.

Saarik Kalia and Michael Zanger-Tishler

Good functions and multivariate polynomials

Mentor Tue Ly

Project suggested by Prof. Dmitry Kleinbock (Brandeis University)

The (C, α)-good property of functions was introduced by Dmitry Kleinbock and Grigory Margulis in 1998 and during the last 15 years showed up in many problems related to dynamics and Diophantine approximation. The optimal values for C and α have been proven for single variable polynomials, but have yet to been proven for multivariable polynomials. We work towards finding the optimal values of C and α for these multivariable polynomials.

Session 4

Ying Gao

Depths of posets ordered by refinement

Mentor Sergei Bernstein

Project suggested by Prof. Richard Stanley (MIT)

In 2008, Herzog et al. found a connection between the concept of Stanley depth in algebra and partitions of certain partially-ordered sets. In doing so, they motivated a pair of papers which defined and found a quantity called depth for posets. Biro et al. found the depth of the poset of subsets of a set of size n ordered by inclusion, and Wang found the depth of a product of chains of equal length. We study the depth of posets constructed from the partitions of a set ordered by refinement. Specifically, we study a sequence we call G_i of depths of refinement-ordered posets of the partitions of sets of size i. We have found the value of G_i for small i, and have found bounds on the value of G_i as i increases.

Vahid Fazel-Rezai

Equivalence classes of length-changing replacements of size-3 patterns

Mentor Dr. Tanya Khovanova

Project suggested by Prof. James Propp (University of Massachusetts at Lowell)

Much work has been done in examining the replacement of patterns in permutations. In this research, we study a new type of pattern replacement that changes the length of the permutation. In particular, we focus primarily on replacements between 123 and another pattern with only two integer elements. For each replacement, we partition the set of all permutations of any length into equivalence classes, consisting of permutations reachable from one another through a series of replacements. We break the eighteen replacements of this type into four cases by structure of equivalence classes, and fully characterize the equivalence classes of each replacement. Exactly half of these replacements have only five classes while the remaining half have infinitely many classes. In characterizing infinitely many classes we often make use of a primitive permutation, a unique permutation of shortest length equivalent to some given permutation. In some replacements, the process of finding a primitive permutation involved showing that equivalence under a length-changing replacement implies equivalence under a length-preserving replacement, providing a link to previous research. This work establishes methods and results that can be used in other short replacements or generalizations.

William Kuszmaul

On q-enumeration of modular statistics

Mentor Darij Grinberg

Project suggested by William Kuszmaul

We study equivalence classes of n-tuples of integers under the following relation. Consider two n-tuples equivalent if they differ by a sum of nontrivially periodic n-tuples. Using properties of cyclotomic integers, we find an exhaustive set of invariants of n-tuples up to this equivalence.

This set of invariants leads to a new approach to q-enumeration of modular statistics. When a set of objects M exhibits certain properties with regard to function $f : M \to \mathbb{Z}$, we find a formula which solves for the number of $m \in M$ with $f(m) \equiv i \pmod{n}$ for a given i and n. There are several applications for this result, including a solution for the number of Dyck paths with major index $\equiv i \pmod{n}$ on a $j \times j$ grid when $n|2j$.

Session 5

Gabriella Studt

Higher Bruhat order on Weyl groups of Type B

Mentor Daniel Thompson

Project suggested by Dr. Benjamin Elias (MIT)

Consider the Weyl Group of a Type B_n root system, whose simple roots are the short root $\alpha_1 = e_1$ and the long roots $\alpha_i = e_i - e_{i-1}$ for $2 \leq i \leq n$. The group is generated by the reflections s_{α_i} with respect to these simple roots. Reflection through the hyperplane orthogonal to e_k is given by the conjugation of s_{e_1} by the reflections with respect to the long roots $\alpha_2, ..., \alpha_k$. The longest element of the group is obtained by reflecting with respect to the standard basis to obtain $-Id$. In other words, $w_0 = \Pi_{i=1}^{n} s_{e_i} = -Id$. This series of reflections, when expressed in terms of the simple reflections, constitutes the lexicographic total order ρ on the inversions $J \in C(n, 2)$. Similarly, we can obtain the lexicographic total order on $C(n, k)$ for each subsequent value of k. Thus we are able to iteratively construct the Higher Bruhat Order on the Weyl Group of Type B. This order, and its accompanying properties, are the focus of this research.

Ravi Jagadeesan

Belyi functions with prescribed monodromy

Mentor Akhil Mathew

Project suggested by Prof. Noam Elkies (Harvard University)

We study the action of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ on the category of Belyi functions (finite covers of $\mathbb{P}^1_{\mathbb{C}}$ unbranched outside $\{0, 1, \infty\}$). In particular, we consider Galois orbits of Belyi functions whose monodromy generators have fixed cycle types. We prove a lower bound on the number of such Galois orbits. As a corollary, we obtain that for all N, there is an $n \leq N$ such that a monodromy cycle type class of degree n Belyi functions splits into at least $\frac{1}{16} 2^{\sqrt{\frac{2N}{3}}}$ Galois orbits.

Ritesh Ragavender

q-analogues of symmetric polynomials and nilHecke algebras

Mentor Alex Ellis

Project suggested by Alex Ellis (Columbia University)

Odd symmetric functions and the odd nilHecke algebra were first introduced by Ellis, Lauda, and Khovanov to study the quantum case of noncommutative symmetric functions. We extend the construction of odd symmetric functions from $q = -1$ to more general q and describe the resulting q-bialgebra structure. By studying a diagrammatic interpretation for the q-analogue of the standard bilinear form on symmetric functions, we find some relations involving generators of the q-bialgebra. We then use q-divided difference operators to develop a q-nilHecke algebra and detail its various properties.

Session 6

Jeffrey Cai

Orbits of the symplectic group on partial flag varieties of type A

Mentor Vinoth Nandakumar

Project suggested by Vinoth Nandakumar (MIT)

Orbits of $G = \mathrm{GL}_n$ on $V \times G/B \times G/B$, where $V = \mathbb{C}^n$ and G/B is the complete flag variety, are considered and orbit representatives found, and the number of orbits is determined. Orbits of $K = \mathrm{Sp}_{2n}$ on $V \times G/P$ for certain choices of the generalized flag variety G/P are considered and orbit representatives found.

Isaac Xia

Quotients of lower central series over Z with multiple relations

Mentor Dr. Yael Fregier

Project suggested by Prof. Pavel Etingof (MIT)

Classical physics is usually described by means of commutative algebra, while quantum physics belongs to the world of non-commutative algebras. Finitely generated non-commutative algebras are quotients of A_n so we concentrate on this algebra and its most elementary quotients. We study the lower central series filtration ($L_1 := A$ and $L_{i+1} := [L_i, L_1]$) of associative algebras of the form $A = A_2/\langle x^m, y^n \rangle$. We present sample calculations and a conjecture of P. Etingof on the divisibility of dimensions of the quotients $N_k := M_k/M_{k+1}$ by primes (where M_k is the ideal generated by L_k) along with a few of its corollaries.

Section II: Computer Science

Session 7

William Wu and Nicolaas Kaashoek

How to teach a class to grade itself

Mentors Christos Tzamos and Matt Weinberg

Project suggested by Prof. Costis Daskalakis (MIT)

This project focuses on the development of an efficient system for peer grading in online courses. It explains the motivation behind solving an issue such as this, and explains the tools needed to do so, primarily Game Theory and mechanism design. The project further focuses on what it takes to design an efficient mechanism in Game Theory, and the equations that were created to test mechanisms such as this. Also explored are the mechanisms created so far, the reasoning behind them, and why they work or don't work.

Anish Athalye and Patrick Long

Performance analysis and optimization of skip lists for modern multi-core architectures

Mentors Austin Clements and Stephen Tu

Project suggested by Prof. Costis Daskalakis (MIT)

With the widespread adoption of modern multi-core systems, it is necessary to design and develop high-performance concurrent data structures and parallel algorithms to take advantage of the hardware. We implement a skip list, a data structure implementing an ordered set, and analyze its performance with both read-only and read-write workloads on a modern 80-core machine. We implement the skip list in Java and C++ and analyze the scalability of various workloads over different VM and memory allocator configurations. We find that with a wait-free implementation of read-only operations, read-only benchmarks scale linearly. With read-write workloads, scalability varies with choice of language and C++ memory allocator, but scalability is not affected by choice of Java VM. We observe that the C++ implementation scales and performs better than the Java implementation. We find that the choice of C++ memory allocator has a significant effect on scalability, with performance differing by up to a factor of 40.

Ajay Saini

Modeling the opinion dynamics of a social network

Mentor Dr. Natasha Markuzon

Project suggested by Dr. Natasha Markuzon (Draper Laboratory)

Social networks have been extensively studied in recent years with the aim of understanding how the connectedness of different societies and their subgroups influences the spread of innovations and opinions through human networks. However, most existing studies model hypothetical aspects of society and few attempt to simulate the dynamics of real social networks or changes in the connectivity of such networks over time. We propose a data driven modeling approach where interactions between subjects account for changes in opinions and observations of a real community validate changes in connections.

In particular, we use the Social Evolution dataset of the MIT Human Dynamics Lab, which contains observations of a student community over a period of 9 months and gives each student's opinion and friendship information at four data points in this period of time. The proposed model aims to reflect the overall dynamical changes in opinion and connectivity in the observed student group, which can then be validated through numerous simulations.

We demonstrate that under certain conditions, the average opinion in the proposed network has a tendency to stabilize within a bound and that if not specifically restricted, the interconnectedness of the network increases with time. We also find that allowing for dynamical connection change significantly impacts the long-term stability of the network. Lastly, we see that the rate of opinion convergence is crucial in determining the long-term connectivity of the network.

Session 8

Istvan Chung and Nathan Wolfe

A collaborative editor in Ur/Web

Mentor Benjamin Barenblat

Project suggested by Prof. Adam Chlipala (MIT)

Computer programs are plagued with errors and security flaws resulting from the incorrect storage of executable code as static data. These issues are manifest in cross-site scripting, cross-site request forgery, and SQL injection attacks, among others. For this reason, we use the Ur/Web programming language to create robust, secure applications. Our project, a collaborative editor, takes advantage of Ur/Web's security and type-safety features to allow users to view and edit documents together, seeing each other's changes in real time. Furthermore, this project is part of a larger course management system being written in Ur/Web, allowing a cleaner, easier way for students and professors to manage courses, assignments, and documents. In our presentation, we introduce Ur/Web and its benefits, explain the goals of our research project, and discuss the results achieved so far, a SQL-backed server coordinating communication between users, and the client-side user interface enabling users to easily edit documents together.

Oron Propp and Alex Sekula

Automating interactive theorem-proving with Coq and Ltac

Mentor Drew Haven

Project suggested by Prof. Adam Chlipala (MIT)

Mathematical proofs have been a longstanding method of conclusively rationalizing assertions. While the field of mathematics has been completely transfigured through the utilization of this tool, up until recently, proofs been verified solely by attaining unanimous approval of their reasoning. However, with the relatively recent invention of the computer, the possibility of utilizing the unfailing computational abilities of these machines to verify mathematical proofs has been instantiated. Coq is one such programming language that implements these capabilities. Using Coq, mathematical proofs can be codified into language that a computer can then process and validate. Despite being extremely high-level, proofs written in Coq generally do not resemble their mathematical roots, nor are they succinct, involving many tactics and manipulations to prove the simplest of statements. To remedy this issue, higher-level ”tactics” were built which resulted in automated proofs with a greater level of correspondence to their mathematical counterparts. With the help of this automation, Coq has the potential to become an educational tool for university students to learn how to write formal proofs and confirm that their mathematical proofs are valid.

Nihal Gowravaram

Avoidance in (2+2)-free posets

Mentor Wuttisak Trongsiriwat

Project suggested by Prof. Alex Postnikov (MIT)

We investigate avoidance in (2+2)-free partially ordered sets. In particular, we extend on the bijection by Bousquet-Melou et al. from (2+2)-free posets to ascent sequences, drawing connections between avoidance in ascent sequence and avoidance in (2+2)-free posets, partially resolving the question for size three posets. In addition, we conduct enumerations on (2+2)-free posets, computing the number of posets avoiding (2+2) and other posets p of size 4. Lastly, we consider the idea of Wilf-Equivalences, determining such equivalences in (2+2)-free posets for a number of pairs of patterns.

Session 9

Steven Homberg

Finding enrichments of functional annotations for disease-associated single-nucleotide polymorphisms

Mentor Dr. Luke Ward

Project suggested by Prof. Manolis Kellis (MIT)

The advancement of genome-wide association studies, which establish associations between single-nucleotide polymorphisms (SNPs) and diseases, as well as that of genomic annotation, which ascribes role or function to SNPs in the genome, permit the establishment of associations between the annotated functions and roles of SNPs with the diseases with which they are associated through computational analysis. This connection provides potentially causal mechanisms by which genetic variants act. Results of our analysis indicate several significant associations between common diseases and genomic annotations, providing groundwork for further biological investigation of the potentially causal link between SNPs, noncoding regulatory elements, and disease.

John Long

Evidence of purifying selection in humans

Mentor Angela Yen

Project suggested by Prof. Manolis Kellis (MIT)

The Human Genome Project completed in 2003 gave us a reference genome for the human species. Before the project was completed, it was believed that the primary function of DNA was to code for protein. However, it was discovered that only 2% of the genome consists of regions that code for proteins. The remaining regions of the genome are either functional regions that regulate the coding regions or junk DNA regions that do nothing. The distinction between the two types of regions is not completely clear. Evidence of purifying selection, the decrease in frequency of deleterious alleles, is likely a sign that a region is functional. The goal of this project was to find evidence of purifying selection in 3 newly acquired regions in the human genome that are hypothesized to be functional: 5′ UTRs, Exonic Splicing Enhancer regions, and miRNA binding sites. The mean Derived Allele Frequency of SNPs in the featured regions was compared to control regions. A lower mean DAF value was used as evidence of purifying selection.

Session 10

Ben Zheng

Removing disorder in drug resistance-related proteins in tuberculosis through hill-climbing algorithms

Mentor Dr. Gil Alterovitz

Project suggested by Dr. Gil Alterovitz

The objective of this project is to find more efficient methods to eliminate intrinsically disordered regions (IDRs) in proteins by experimenting on particularly disordered proteins — primarily those associated with drug resistance in tuberculosis (TB). To achieve results, we wrote a hill-climbing algorithm to be used in finding the most effective changes that could be made to a protein's amino acid sequence to restore order to the IDRs. We hypothesized that our method would be more efficient than the current method for eliminating disorder. In order to verify the results of our program, we found a protein associated with drug resistance in TB using the TBDReaM Database and not listed in the Protein Data Bank to change and visualize. Previous research had found that internally disordered proteins in TB tended to be associated with drug resistance, suggesting a new possible target for antibiotics. By using Cn3D to project a simulation of our developed protein, we were able to see if the new protein was truly more ordered as well as statistically analyze the results using TraDES. Although our results suggest that there has been no significant improvement in the disorder of the protein, testing with different algorithms and more proteins may yield positive results.

Peijin Zhang

Leveraging disordered-ordered interactions to yield new targets and drugs for tuberculosis

Mentor Dr. Gil Alterovitz

Project suggested by Dr. Gil Alterovitz

Despite the development of powerful antibiotics in the twentieth century, the increasing prevalence of multi-drug resistant and extensively drug resistant TB (MDR/XDR-TB) has caused TB to return to the forefront of public health concerns. TB is one of the deadliest infections diseases in the modern world, and new drugs are desperately needed to combat the growing problem of drug resistant tuberculosis.

This study relied on analysis of previously identified single nucleotide polymorphisms (SNPs) related to TB drug resistance in conjunction with intrinsically disordered proteins (IDPs) to better understand and target the mechanisms behind the evolution of MDR/XDR-TB. IDPs were hypothesized to play crucial roles in biological processes within TB that allow for the expression of drug resistance. Disordered regions were identified through cross-referencing empirically validated protein structures across various databases and through utilizing artificial neural-network-based prediction methods.

Subsequent statistical analysis on the prevalence of protein residues affected by drug resistance related SNPs within disordered regions of IDPs confirmed the original hypothesis of the importance of IDPs in the development of drug resistance. Thus, proteins that contain residues modified by drug resistance related SNPs are likely to have crucial functions relating to TB drug resistance, and would therefore be ideal drug targets.

Based on this finding, virtual drug screening was performed on target IDPs to discover new compounds capable of targeting MDR/XDR-TB. TB growth inhibition assays validated the inhibitory effects of compounds identified in this study. Future research on these results could lead to the creation of direly needed anti-MDR-TB drugs.

Session 11

Jonathan Patsenker

Finding the binding sites of MoRFs on a partner protein

Mentor Dr. Gil Alterovitz

Project suggested by Dr. Gil Alterovitz

A type of protein with regions containing high amounts of structural disorder is called a MoRF, and has certain binding tendencies that are important to take notice of. One specific property that is important to be examined, is the MoRF's binding location on a rigid, ordered, partner protein, and what causes the MoRF to bind there. So far, there have been a few predictors that predict where on the MoRF the binding site is located, and whether these two proteins will bind, but a predictor that can figure out where on the protein the MoRF will bind has not been created. To create this predictor, we set up a Bayes Net model, that then compares certain pieces of the protein through many attributes, in a windowed approach method, and figures out more likely bonding sites. This is all done by a program implemented in Knime, using Weka tools to set up a Bayes Net. The location of bonding on rigid proteins of MoRFs, and the attributes that cause certain MoRFs and rigid proteins to behave as they do when bonding impacts our understand of the way many pathogens function, and can help lead to more effective medicines and treatments.

Yishen Chen

SMART Genomics API

Mentor Dr. Gil Alterovitz

Project suggested by Dr. Gil Alterovitz

With the influx of data supporting personalized genetic medicine, a need arises to accommodate the use of this information in the electronic medical record, by point-of-care providers. SMART Genomics API is a means to classify and package genomic information for the use in the clinical realm. Since a clinical API has been defined and supported by the SMART clinical initiative, it is natural to model the use of genomic data in a way similar to this established method. We demonstrate this work via a ”Genomic Advisor” app that integrates clinical and genomic information for medical decision support.

Skanda Koppula

Prediction of disease by pathway-based integrative genomic and demographic analysis

Mentors Dr. Gil Alterovitz and Dr. Amin Zollanvari

Project suggested by Dr. Gil Alterovitz

To better understand the origins of addiction, we developed a method to analyze and model clinically-ascertainable demographic and genetic factors at the cellular pathways level by introducing a SNP to gene to pathway mapping. This approach allows us to obtain biological meaning from the constructed predictive models of alcohol dependence. In its analysis of a cohort of 2,762 subjects, our model achieves a predictive performance (AUROC) of 90% on a separate dataset, and is the first predictor of alcoholism. The model is significantly superior to models with only genetic variables (83%) or only demographic and environmental variables (61%). This lead was not simply due to an increase in the number of variables, but rather due to a synergy between two types of variables: demographics and genetics. Furthermore, the predictive model reveals 17 biological pathways associated with the development of alcoholism worth for future investigations. Our method is able to analyze a variety of data modalities and factor types, and future investigations include the study of other complex diseases.

Section III. Computational and Physical Biology

Session 12

Boryana Doyle

Chromatin organization: from polymer loops to topological domains

Mentors Geoffrey Fudenberg and Maxim Imakaev

Project suggested by Prof. Leonid Mirny (MIT)

The classic model of eukaryotic gene expression requires spatial contact between a distal enhancer and a proximal promoter. Various models of mediating enhancer-promoter interactions have been proposed, including the topological model, in which two or more control elements in the vicinity of an enhancer and the promoter interact to form loops. An outstanding question in the field is whether the topological model is effective at mediating enhancer-promoter interactions.

Here we use a polymer model of chromatinized DNA and Langevin dynamics simulations to study the topological model of insulation. We consider one- and two-loop topological elements and assess whether spatial contacts between various regions of DNA are induced or suppressed by the topological element. We find that sequestration of an enhancer (or promoter) within a loop acts as an insulator. In addition to confirming the insulating effect of topological elements, we find that forming a loop in the region between an enhancer and a promoter facilitates enhancer-promoter interactions. Furthermore, we consider various parameters including enhancer-promoter genomic distance, density, topological constraints, two-loop elements, and length scale. As enhancer-promoter distance increases, the facilitation effect decreases, while insulation remains unchanged. Both observed effects are more dramatic at low density than at high density and with two-loop elements rather than one-loop elements. Lastly, our results hold both with and without topological constraints and at different length scales.

Our polymer simulations confirm that loop-forming topological elements are capable of mediating enhancer-promoter interactions. We note that the biological model of insulator activity around H19 and Igf2 genes in mouse liver cells presented by Kurukuti et al. (PNAS 2006), is a specific realization of the two-loop topological insulator discussed here. Finally, our simulations demonstrate that accounting for the polymeric nature of chromatin is essential for understanding how enhancer-promoter interactions are modulated.

Carolyn Lu

Dynamic folding of chromatin domains by active SMC-mediated loops

Mentors Geoffrey Fudenberg and Maxim Imakaev

Project suggested by Prof. Leonid Mirny (MIT)

The recently developed Hi-C method has begun to reveal aspects of chromosomal organization at unprecedented detail by measuring direct spatial contacts in 3D between all pairs of genomic loci. In particular, high-resolution Hi-C experiments have found that the genome is organized into small highly-interactive Topologically Associated Domains (TADs), ranging from 300kb to 1Mb in size. However, the organization of TADs in three-dimensional space, and their relation to nuclear and genomic features remains unclear. Here we investigate a set of polymer models to test potential biological mechanisms for the formation of TADs. First, we investigated if a change in local chromatin structure could introduce a boundary between TADs, and therefore prevent contacts between neighboring TADs. We studied models of chromatin fibers with variable stiffness and thickness. We also investigated whether RNAs attached to DNA at highly transcribed genes could create a TAD boundary. Most of the studied models could lead to the formation of TAD-like structures. However, boundaries between TADs in these models were not sufficiently strong at biologically-relevant parameter values. Finally, we tested whether the formation of chromatin loops by Structural Maintenance of Chromosome (SMC) proteins could, under certain assumptions, lead to formation of TADs. It was recently proposed that the SMC complex can bind to DNA and move along it, extruding a chromatin loop in the process. Our hypothesis was that boundaries between TADs have the ability to unload SMC complexes from DNA, which would prevent the formation of chromatin loops connecting neighboring TADs. This model successfully formed TAD-like structures and reconstructed experimentally observed features of a Hi-C map around a TAD. Additionally, we show that assigning different efficiencies to the SMC-unloading elements could create nested TAD inside a TAD structures, similar to those observed in the experimentally measured Hi-C contact maps.

Ashwin Murali

Lineage-dependent properties of 16S ribosomal RNA nucleotide composition

Mentors Geoffrey Fudenberg and Maxim Imakaev

Project suggested by Prof. Leonid Mirny (MIT)

The 16s region of rRNA (ribosomal RNA) in bacteria is well known to be essential to the identification and definition of bacterial species. The Ribosomal Database Project provides an extensive catalog of over a million sequences of the 16s region. It is previously known that the GC (guanine + cytosine) content of bacteria is highly varying, ranging from 20% to 80%. We find that purine and GU (guanine + uracil) content is relatively more conserved than GC content. Moreover, we find that Actinobacteria, a major phyla, do not follow this overall trend; in Actinobacteria, GU content varies similarly to GC content. In Actinobacteria, guanine and uracil are the least correlated pair of nucleotides. This lack of correlation may explain the high variation we see in Actinobacteria GU content, which is unobserved in other major phyla. This trend of low GU correlation is not only present in the larger Actinobacteria phyla but also in the smaller clades of Actinobacteria. Further investigation must be conducted to figure out the cause of this deviation from the overall trend of purine and GU conservation.

Hao Shen

The impact of gene order on evolution

Mentor Anton Goloborodko

Project suggested by Prof. Leonid Mirny (MIT)

It has been experimentally shown that the order of bacterial and eukaryotic genes is non-random and affects the viability of the organism. However, the proposed explanations of this effect lack the support of any rigorous simulations or quantitative studies.

In this project, we investigate the effects of the gene order on the evolutionary dynamics of populations using computer simulations. Traditionally, populations are simulated as a group of reproducing organisms with individual genomes. In the classical Wright-Fisher (WF) model, genes additively contribute to the rate of the organism's reproduction, or fitness; their order in the genome therefore does not play any role. Here we modify the WF model to show that a combination of two well-known biological phenomena, epistasis and recombination, makes the population sensitive to the arrangement of genes in the genome. The former, epistasis, denotes the synergistic effect of gene combinations on overall fitness. The latter, recombination, is a process of genetic exchange where two parents swap pieces of their DNA to produce the genome of their progeny. In the populations where genes with strong epistasis are located close on the genome, recombination rarely break their beneficial combinations, increasing the average viability of the progeny. This project uses computer simulations to analyze how recombining populations develop when various genes are strongly epistatically interacting. In particular, we study how the dynamics of the population and its final steady state are affected by the order of genes. Implications of this project can be useful for analysis of available genomic data. Notably, the observed relationships can be used to interpret the observed history of gene rearrangement in evolving populations.