Build Systems to Reverse Engineer The Most Amazing Operating System on the Planet

Gill Bejerano

http://bejerano.stanford.edu
Today’s Agenda

You should rotate.

You should rotate in Genomics.

You should rotate in Genomics with the Bejerano lab.

http://bejerano.stanford.edu/
Why Rotate?

Welcome to one of the best CS depts in the world!

Our CS dept. encourages you to undertake three rotations, over three quarters before aligning for your PhD.

So you try new things before picking your field.

So you are happier by graduation with your choice of PhD, often with the choice of a career for life.

Don’t rush.

http://bejerano.stanford.edu
Why Rotate Away from your Undergrad Field?

To get into Stanford you must have done well as undergrad in some field of CS.

However, researchers, at all levels, often do their best work when they cross fields, taking ideas from one field to another.

Moreover, regardless of your choice, in 3-4 years, when you start pondering life after grad school, you will be happier knowing you tried more options.

Don’t rush.

http://bejerano.stanford.edu
Why Rotate in a Young Field?

The most exciting science, the most opportunities to change the world, the most faculty positions, the most industry / start-up opportunities, lie in young fields.

In particular, most opportunities lie at the intersection of established fields.

You’re young. You should try…

http://bejerano.stanford.edu
Why Rotate in Genomics?

Genomics is at the intersection of CS & BioMedicine. Genomics is fundamentally a computational science.

This century is owned by Genomics.

“There is gold in them thar hills” – lots to build & discover!

Many traditional CS fields impacted by Genomics.

“I don’t know any biology..”

Neither did many of us when we started!
(I hadn’t taken a biology class 10th grade through undergrad;
The human genome was first assembled by a graphics programmer).

http://bejerano.stanford.edu
Genomics is affecting multiple fields of CS

Storage  
Crypto  
Architecture  
Databases  
ML  
etc.

Even if you do not want to be a genomicist, some of the most exciting challenges in your field will be at the interface with Genomics.

Everybody wants a piece of the action.
Genomics is affecting multiple fields of CS

Some current & (potential) collaborations here in CS:

- Dan Boneh
- Bill Dally
- Chris Re
- (Greg Valiant)
- (Jure Leskovec)

Look for this star for 5 current project offerings, in later slides. (OCT '16)
Shoot! I should learn some Genomics

http://cs273a.stanford.edu

A Computational Tour of The Human Genome

Fall quarter
MW 1:30-2:50p
Beckman B302

You can still audit :)
Okay, Fine, What is Genomics?

The Most Amazing Operating System on the Planet: The Human Genome

hOS
The Human Genome

Genome = The Operating System that runs every cell in our body
3*10^9 letters long, over the DNA alphabet = \{A,C,G,T\}

http://bejerano.stanford.edu
The Biggest Challenge in Genomics…

… is **computational**: 

How does this

```
3CTAGATGCGCTGGTG
3CTTTGC GCCCGTCAAF
3TCTTTGAAGGCTGTGA
TCAAGCTTTCTGGC GAI
3CGTTTGACC GGA GC
3TTGCAATGAGTTCC
3AGCTGTCTATATGA
3CAACAAATAGGCAAI
```

Program  

Output

This “coding” question has **profound** implications for our lives
The Biggest Challenge in Genomics…

… is computational:

How does this encode this

Program

Forks & re-merges

Where did we come from? How are we different from each other?
The Biggest Challenge in Genomics…

… is computational:

How does this program suite of related products encode this?

What in our genomes make us different from other species?

http://bejerano.stanford.edu
The Biggest Challenge in Genomics…

… is computational:

How does this encode this?

Program

Bugs

Output

What genomic mutations predispose us to disease?
The Biggest Challenge in Genomics…

… is computational:

How does this encode this

Program

Bugs

What genomic mutations determine our drug response?

http://bejerano.stanford.edu
The Biggest Challenge in Genomics…

… is computational:

How does this encode this

We can eliminate suffering by not “booting” “buggy” embryos
The Biggest Challenge in Genomics…

… is computational:

How does this encode this

Program

Bugs

We can eliminate suffering by fixing people’s “buggy” genomes

http://bejerano.stanford.edu
 Literally Save Lives From Your Keyboard!

http://bejerano.stanford.edu
Gene Therapy: We Design the Cure

How to Cure a Bubble Boy

Thanks to gene therapy, a boy born without an immune system can now play in the yard.

Life saving "code injection"

With Matt Porteus
They experiment.
We help interpret.
Real Stanford patients.

http://bejerano.stanford.edu
Biomedicine is facing a phase transition

From an obsession with the interpreter,

To an obsession with the code.

(When your code has a bug what do you fix?..)
Okay.. what do we do in Genomics?

We build systems to
Reverse Engineer the Most Amazing
Operating System on the Planet:
The Human Genome

Projects in our group include:

http://bejerano.stanford.edu
1 in 33 babies suffers from some form of rare disease.

Discovering the “bug” in their code: step 1 for intervention, cure, eradication.

We have genomic data and medical records for hundreds of real patients.

Only 25% of cases are solved. We bring new data and new approaches.

http://bejerano.stanford.edu
Mendelian Diseases

- Proband variants
- Unaffected variants
- Gene-disease associations

Causal mutation
Example 2: Solving Singletons

intellectual disability with impaired speech development and aggressive behavior

83 candidate genes in her exome with rare variants

AC018470.1, ACAP3, ADAP1, AMPD1, ASPM, ASXL2, BAZ1B, BHLHE22, BTBD9, C17orf104, C17orf74, C19orf26, C1orf87, C2orf81, CCNL2, CDH10, CHD6, CNOT3, COL6A5, DCHS2, DEAF1, DNM1, FAM216B, FAM73B, FAM83H, FAM84B, FAT3, FBXO25, FCRLB, FLJ00104, FRS2, GRK7, HEPHL1, HOXD11, IL19, INSRR, IQCC, KIAA0825, LAMA5, LAMC3, LGR6, MAST4, MBD6, MBLAC2, MCM10, MDH2, METRN, MSL2, N4BP3, NCKAP5, NUP50, NYNRIN, ORC3, PDCD2L, PDXP, PLEKHG1, PLIN2, POU3F2, PXMP2, RAB11FIP1, RASSF1, RIMS1, RTKN2, SASS6, SERPINA3, SH3BP1, SHB, SLC2A9, SLC38A8, SON, SP8, SPTBN5, SRRM2, TAAR1, TARSL2, TET2, TRIM72, TSPAN15, TSPYL4, WDR20, XPNPEP1, ZFYVE16, ZNF469, ZSCAN29

Sooo… guess which gene is at fault?

http://bejerano.stanford.edu
2) Link Patient Mutations to Patient Disease

We curate gene – phenotype association databases.

We teach the computer to read the current literature.

The computer helps rank the match between patient mutations and patient phenotypes.

http://bejerano.stanford.edu/
Teach the Computer to Solve Patients

• Novel facts not relevant to our patient:

  Mutations altering the coding sequence of **KRT16** cause **Pachyonychia Congenita** …

  Recessive mutations in **LTBP2** (…) have been identified as a cause of early-onset **glaucoma** …

  Mutations in **QARS**, encoding Glutaminyl-tRNA synthetase, cause progressive **microcephaly** …

• Bingo!

  Mutations affecting the SAND domain of **DEAF1** cause **intellectual disability** …

http://bejerano.stanford.edu
Bingo Moment: Patient Solved

• Literature paper teaches us:

Mutations affecting the SAND domain of **DEAF1** cause **intellectual disability** with severe **speech impairment** and **behavioral problems**.

• Patient: **intellectual disability with impaired speech development and aggressive behavior**

AC018470.1, ACAP3, ADAP1, AMPD1, ASPL, ASXL2, BAZ1B, BHLHE22, BTBD9, C17orf104, C17orf74, C19orf26, C1orf87, C2orf81, CCNL2, CDH10, CHD6, CNOT3, COL6A5, DCHS2, DNM1, FAM216B, FAM73B, FAM83H, FAM84B, FAT3, FBXO25, FCRLB, FLJ00104, FRS2, GRK7, HEPHL1, HOXD11, IL19, INSRR, IQCC, KIAA0825, LAMA5, LAMC3, LGR6, MAST4, MBD6, MBLAC2, MCM10, MDH2, METRN, MSL2, N4BP3, NCKAP5, NUP50, NYNRIN, ORC3, PDCD2L, PDXP, PLEKHG1, PLIN2, POU3F2, PXMP2, RAB11FIP1, RASSF1, RIMS1, RTKN2, SASS6, SERPINA3, SH3BP1, SHB, SLC2A9, SLC38A8, SON, SP8, SPTBN5, SRRM2, TAAR1, TARSL2, TET2, TRIM72, TSPAN15, TSPYL4, WDR20, XPNPEP1, ZFYVE16, ZNF469, ZSCAN29

http://bejerano.stanford.edu
Our family thanks you so much for repeating those “normal” genetics tests and never giving up the search for answers, even when we all but had.
Network Analysis

Networks

genomic data

patient

medical record

genes
complexes
pathways
processes
model org.
phenotypes
signs & symptoms
human disease
The problem:

- Our genomes are best understood in light of each other.
- But our genomes tell so much about us, sharing may lead to discrimination.

Goal:

Find ways to share genomes without revealing genomes

Design distributed computations that analyse, while revealing nothing else about genomes

http://bejerano.stanford.edu
Love Animals?

http://bejerano.stanford.edu/
The “patients” are all toothed whales. How are they still alive?

No other sequenced mammal has dared lose both Mx genes...

Quarter rotation turned 1st author paper:

(Braun et al, PNAS, 2015)

http://bejerano.stanford.edu/
3) Crack the genome code

Goal: combine to find many novel genome → trait mappings.
We have 100 mammalian genomes from one community.
We have 5,000 mammalian traits from a different community.
4) Obama’s BRAIN initiative

Funded as part of The President’s BRAIN initiative to map cell types across (the coolest part of) the mammalian brain.
We are in charge of all genomic analysis in our group.

Joint w 3 other institutes.

Plus related work with Sue McConnell.

http://bejerano.stanford.edu
5) Programmable Hardware for Genomics

Accelerate hard sequence searches in hardware (using FPGAs).
Make amazing discoveries (e.g. genome instructions we share w/ ticks) using these tools. A “Hubble Telescope” for genomics.

http://bejerano.stanford.edu

Joint w Bill Dally (CS/EE).
Our people / our projects

Patient genome solver

Better sequence homology search

Large genomic knowledgebase

Map animal genomes to animal traits

BRAIN initiative

Firestone Medal for Excellence in Undergraduate Research

SGF
Bio-X
NSF

STANFORD SCHOOL OF MEDICINE | Pediatrics

STANFORD COMPUTER SCIENCE

SiGF
SNI

STANFORD UNIVERSITY

PhRMA
CEHG

Rotation project → 1st author PNAS paper

New faculty (K99/R00 award)

I’m outta here!

http://bejerano.stanford.edu
Interested?

Ping me for a chat.
Plenty of ways to get started.

Come to my open office hours
(posted on lab website)
OR
bejerano@stanford.edu

Questions, please?