Fighting HIV with Machine Learning and HPC

David Heckerman
Distinguished Scientist
Senior Director of e Science
Microsoft Research
Fighting HIV with Machine Learning and High Performance Computing

David Heckerman

eScience Group, Microsoft Research
Really? Why?
The convergence of computer science and biology

- DNA is a programming language and a computation device
The convergence of computer science and biology

- Drinking from the fire hose or ...
The convergence of computer science and biology

- Striking similarities in concepts that can be shared both ways
Fighting HIV with machine learning (aka statistics) and high performance computing

- HIV & immunology 101
- PhyloD.net: A tool for studying HIV
- Important discoveries toward a vaccine and possible treatments
HIV is the virus that causes AIDS

- AIDS kills 5,000 people every day
- Drugs work fairly well but are expensive and need to be taken regularly
- Vaccine is perhaps the best hope for developing countries
HIV Lifecycle

- HIV particle
- HIV binds to host cell
- New virion particles
- gp120
- CCR5
- Infected cell
- HIV particle budding from cell
- CD4
- Reverse transcription
- HIV proteins
- Protease
- RNA genomes
- DNA copy of HIV RNA
- Integrase
- DNA integrates into host genome
Our immune system fights viral infections

- Innate (e.g., natural killer cells)
- Adaptive
  - Antibodies (humoral arm)
  - T cells (cellular arm)

Vaccines pretrain the adaptive response thereby generating a stronger response that prevents infection or at least keeps the virus under control
T-cells

Different viral protein fragments

Host Cell

Epitope

HLA
T-cells

Epitope

Different viral protein fragments

Host Cell
T-cells

- Naive Killer T-cell
- Epitope
- Different viral protein fragments

Host Cell
T-cells

- Trained Killer T-cell
- Epitope
- Different viral protein fragments

Host Cell
T-cells

Different viral protein fragments

Host Cell

Epitope

HLA

Trained Killer T-cell
Progression to AIDS
Slide from Richard Harrigan
Mechanisms of T-cell Escape

Killer T-cell

Epitope

HLA

Different viral protein fragments

Host Cell
STOPPING SPAM

By Joshua Goodman, David Heckerman, & Robert Rounthwaite

In 1978 one man spam-e-mails—a spam from a marketing representative at Digital Equipment Corporation for the new DEC- 
Controller—was dropped to about 50 people, so the spam was not considered dangerous. Today a spam e-mail can 
reach hundreds of thousands of people, and consequently carries a much greater threat of viruses and 
malware. The typical spam e-mail, for example, contains more than just a single byte of executable code that 
injects a virus, but also a link that might lead to a web page containing a virus or malware.

SPAM FILTERING

Goodman, Heckerman, & Rounthwaite
Scientific American, April 2005
From spam filters to vaccine design

- Immune system attacks HIV: filter blocks spam
- Problem: Spammers “mutate” their emails
- Solution: Spam mail can’t be arbitrary – they are trying to sell you something
  - Look for disguised product name (e.g., “V1AGRA”)
  - Follow the money

- What is HIV’s Achilles’ heel?
Hypothesis: Certain parts of HIV are critical to its function

If HIV mutates within these epitopes, it becomes less or non-functional

Suggests a vaccine design...
A design for an HIV vaccine

Left to its own devices, our immune system attacks at random epitopes.
A design for an HIV vaccine

A “whole protein” vaccine does little to help the situation (explains failure of Merck vaccine)
A design for an HIV vaccine

A focused vaccine can show immune system where to attack
A design for an HIV vaccine

- Accumulating evidence for hypothesis and identifying these “protective epitopes”
  - Brute force testing of known epitopes (Walker and Pereyra)

- Continuing to search for these protective epitopes with more clever methods
Fighting HIV with machine learning (aka statistics) and high performance computing

- HIV & immunology 101
- PhyloD.net: A tool for studying HIV
- Important discoveries toward cures/vaccines
HLA variability

- Each person has up to 6 different HLA types: (2 ‘A’, 2 ‘B’, 2 ‘C’)

- HLA region is most variable region of DNA--rare for two people to have the same HLA types
Epitope variability

HLA Molecule

Epitope
Use this variability to search for epitopes and the HLAs that attack them

Example:

[Diagram showing a comparison of HIV protein sequences with HLA B57 binding sites.]
Straightforward approach

- Sequence someone’s HIV when they first get infected and re-sequence every month or so

- Expensive, difficult to find subjects early infection (lucky to find 100 in a year)
Our approach: PhyloD.net

- Take a single snapshot of a person’s HIV
- Use the phylogeny of sequences among individuals to infer the infecting sequence
- Requires machine learning algorithms and high-performance computing
- Bottom line: much less expensive and can get data from thousands of subjects
PhyloD.Net: Basic idea

Phylogeny of HIV sequences

Individuals with similar sequences

Individuals with not so similar sequences
PhyloD.Net: Basic idea

Focus on single position

Likely that this individual was infected with lys, which then mutated to arg due to HLA=B57
Multiple positions, multiple HLAs

Science 2007
Mutations are not independent
Covariation Effects

Figure 12
Gag phylogenetic dependency network for combined HOMER and Contract cohorts.
High performance computing a must

Fortunately, the computations are pleasantly parallel
Fighting HIV with machine learning (aka statistics) and high performance computing

- HIV & immunology 101
- PhyloD.net: A tool for studying HIV
- Important discoveries toward cures/vaccines
PhyloD.net publications


What do they tell us?

- Identifying more normal epitopes (11)
- Identifying novel class of epitope targets (2)
- Identifying novel immune responses (1)
Central Dogma of Molecular Biology: Eukaryotic Model
Major errors in translation
Translation
Alternate reading frames lead to gibberish
This gibberish produces a lot of epitopes targeted by the immune system.
First evidence that innate arm of immune system drives HIV evolution

Points of attack by natural killer cells
Summary and next steps

- HIV is not invulnerable
- We can use machine learning and HPC to find HIV’s Achilles’ heel(s)
- Test the vaccine (with Jim Mullins)
PhyloD.net is part of Microsoft Biology Foundation

The Microsoft Biology Foundation

The bioinformatics community has developed a strong tradition of open development, code sharing, and cross-platform support, and a number of language-specific bioinformatics toolkits are now available. These toolkits serve as valuable nucleation points for the community, promoting the sharing of code and establishing de facto standards.

The Microsoft Biology Foundation (MBF) is a language-neutral bioinformatics toolkit built as an extension to the Microsoft .NET Framework. Currently it implements a range of parsers for common bioinformatics file formats; a range of algorithms for manipulating DNA, RNA, and protein sequences; and a set of connectors to biological Web services such as NCBI BLAST. MBF is available under an open source license, and executables, source code, demo applications, and documentation are freely downloadable.