

Open Source at Microsoft

Advancing AIDS vaccine research through open source approaches

Assemble a medical research team. Give them open source medical tools, high performance computing capability, and an application formerly known for its ability to fight spam. Let them create. And what you get is another step closer to an AIDS vaccine.

It is also a truism that in today's heterogeneous software ecosystem, operating systems and applications must work together as smoothly as possible. That's why Microsoft has in recent years been building bridges with other software providers, including those who offer open source products.

The Machine Learning and Applied Statistics (MLAS) group (part of Microsoft Research) is a part of this effort, and in June 2007 the team released the source code for a set of four software tools developed to advance AIDS research. By sharing the code openly and at no charge at its CodePlex project hosting Web site, Microsoft hopes to help speed the worldwide AIDS-research community toward a vaccine.

To understand how Microsoft got to this point, you have to step back 15 years to Stanford University, where David Heckerman had just added a PhD in biomedical research to go with his medical degree. Microsoft recruited him when they realized that the statistical tools he'd developed to help physicians diagnose patients could help computers diagnose themselves.

After joining Microsoft, he spent the next decade applying the diagnostic work he'd done in a medical context to efforts such as developing data mining tools for Microsoft SQL Server and building spam filters.

Eventually, he realized that his work had implications for biomedicine, his original field of interest. Heckerman, now the senior researcher for the MLAS, took these lessons and returned to the fold of biomedical research—in the search for an AIDS vaccine.

Applying high performance computing and open source to vaccine research

What makes AIDS pernicious is the speed of its mutations. "Normally with a virus, bits of protein, called epitopes, appear on the exterior of the cell, and white blood cells come along and destroy them," Heckerman says. "That's what *should* happen. But HIV mutates too fast, so the immune system can't recognize it when it changes."

The goal of Heckerman's team is to identify a set of epitopes common to most people and then apply them to a vaccine. It is that process of identification that lends itself to high performance computing (HPC) and open source solutions. With the AIDS work, researchers "have to do a lot of random calculations," Heckerman says. A typical analysis would take an average CPU a year, but with their 320-

processor cluster, the team can do the work in a day. “HPC clusters are the way things are headed in biology.”

And open source is already established as a successful approach through which distributed researchers can share and collectively understand data. “Open source is very common in biomedicine,” Heckerman says. “When people develop tools, they routinely make those tools available for researchers. Typically, though, biologists won’t trust software until they can see it. They will be skeptical about it.”

Tracking mutations across large patient populations

There are usually about six people on Heckerman’s team, with dozens of biomedical researchers collaborating. Their approach—open source combined with high-performance computing—has already had some success. As Heckerman describes it, the team is trying to determine whether HIV mutates randomly or does so because it is trying to escape the actions of the immune system.

Random mutation would make the virus much harder to fight. But with a software tool called PhyloD, Heckerman’s team found that HIV is in fact mutating in response to the immune system activities.

PhyloD – so named for its ability to incorporate phylogeny (evolution) into its pattern analysis –

looks for correlations between human leukocyte antigen (HLA) and HIV. Those correlations can be significant, since HLA proteins can recognize epitopes on infected cells and alert the immune system to their presence.

HIV survives by mutating faster than HLA proteins can recognize their viral DNA and sound the alarm. Furthermore, this evolution starts anew in each person, since everyone has a different set of HLA proteins. The challenge, then, is to identify those HIV mutations across large patient populations. Using PhyloD, researchers are beginning to address this challenge and understand the complex rules of HIV mutations.

Four tools available on CodePlex

PhyloD is one of four AIDS-related tools that the team has developed (all are available via the [CodePlex](#) site). Researchers have also released the source code for an Epitope Prediction tool, which uses a machine-learning method related to Microsoft spam-filtering technology to scan proteins for likely epitopes in people with any HLA type. The vaccine work combines technologies such as graphical models and other machine-learning techniques to comb through thousands of strains of HIV to find the genetic patterns necessary to train a patient's immune system to fight the virus.

A third software tool developed by Microsoft Research, called an HLA Assignment tool, aims to find epitopes more efficiently. Whereas the Epitope Prediction tool takes a pure machine-learning approach to identifying epitopes, the HLA Assignment tool also takes external biological evidence into account.

The fourth tool, HLA Completion, is designed to help scientists get more research out of the same dollar by addressing the hierarchy of the immune system's HLA types.

Research has implications for other diseases

Researchers who access the four tools at CodePlex have two choices. They can download pre-compiled programs and run those programs on their own computers, an option that gives them complete control, lets them use all of their own computing resources, and gives them access to the full functionality of the programs. Or, they can download the source code and compile the applications themselves, an option that allows them to modify and build on the code so they can further optimize the tools for their own needs in vaccine work.

From there, progress is a series of recursive steps: listen, refine, and share. "Biologists are telling us what they'd like to see," Heckerman says. The team will take that feedback and use it to improve the tools they have and develop new tools.

The implications of the research are just beginning to be explored. While tools are tailored to AIDS research, they can be applied to vaccine design for other fast-mutating diseases as well, such as malaria and hepatitis C. With the ongoing leadership of Heckerman's team and the Microsoft commitment to community and learning, amazing results are possible.

Copyright

All other trademarks are property of their respective owners.

Information in this document, including URL and other Internet Web site references, is subject to change without notice and is provided for informational purposes only. The entire risk of the use or results from the use of this document remains with the user, and Microsoft Corporation makes no warranties, either express or implied. Unless otherwise noted, the companies, organizations, products, domain names, e-mail addresses, logos, people, places, and events depicted in examples herein are fictitious. No association with any real company, organization, product, domain name, e-mail address, logo, person, place, or event is intended or should be inferred. Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Microsoft Corporation.

© 2007 Microsoft Corporation. All Rights Reserved.

Microsoft may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Microsoft, the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property.

Microsoft, Windows, Windows XP, Windows Server, and Windows Vista are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries.