http://www.bioinform.com/issues/12_16/features/146391-1.html

bioINFORM
The integrated informatics news source

Microsoft Research and Partners Develop Tool to Refine Low-Res HLA-Typing Data

April 18, 2008 | Vol. 12 No. 16

By Vivien Marx

Microsoft Research and several collaborators have developed an in silico method to probabilistically refine low-resolution human leukocyte antigen-typing data that could potentially aid vaccine design, transplantation medicine, and immunology research, according to a recent study.

In a paper published in the February *PLoS Computational Biology*, the research team, which included scientists from Microsoft, Massachusetts General Hospital, the National Cancer Institute, and the University of Oxford, reported that the new method outperformed the multinomial-based approach that the HLA community currently uses to refine data.

A software package based on the approach, the HLA Completion tool, is available here.

HLA genes encode for proteins that play a key role in immune response, and scientists have so far identified hundreds of alleles for various classes of these genes. HLA typing is commonly used for a number of immunology applications, including identifying immunogenetic disease risk factors and selecting stem cell transplant donors for unrelated recipients.

However, high-resolution HLA typing using sequencing-based platforms is a costly, laborious process, David Heckerman and Jennifer Listgarten, members of Microsoft's eScience Research Group, told *BioInform* in a joint interview. As a compromise, researchers often resort to using low-resolution data from antibody-based serological tests, yielding ambiguous results that can hinder immunology research.

The researchers developed a statistical approach that uses existing high-quality HLA data to reduce the ambiguity in lower-quality data sets. The authors note in the *PLoS Computational Biology* paper that the probabilistic approach improves upon the standard method of HLA typing refinement, which relies on multinomial parameterization trained with an expectation-maximization algorithm.

Instead, the new method infers HLA haplotypes via parsimonious parameterization and smoothing the maximum likelihood solution, making it "possible to scale the refinement to a larger number of alleles and loci in a more computationally efficient and stable manner," the scientists wrote in their paper.

HLA genes, which make up the major histocompatiblity complex on chromosome 6, regulate the immune response by producing antigen-presenting proteins that tag tell-tale pathogen parts and present them to the immune system, which signals other immune cells to destroy invaders with these tags.

Genetically, this chromosomal region is tricky. “The reason it is difficult to type is because the gene region is very polymorphic,” Heckerman explained. The allele distribution also varies across ethnic and geographic populations, he said.

As typing methods have moved from low-resolution antibody-based serological tests to more precise DNA-based methods, typing resolution has increased but still shows ambiguity due to the ever-increasing number of alleles being discovered at these loci.

For example, genes can be at the same locus on each of a chromosomal pair, but heterozygous HLA typing results can reflect several combinations of different alleles residing at both loci.

Because of their close proximity, the HLA genes are inherited together and therefore considered a haplotype; if specific alleles show this linkage disequilibrium, the researchers can statistically disambiguate HLA types, they wrote. The idea is to be able infer and predict the haplotype structure of HLA alleles across multiple loci.

As Listgarten explained, the current method for building haplotype models relies on an expectation-maximization algorithm in which haplotype probability is parameterized with a multinomial table and delivers the probability of every haplotype.

While this approach is used effectively in SNP analysis, it presents difficulties in HLA typing because it “scales badly with the number of loci and with the number of alleles at each locus,” the researchers wrote.

“Most of the work has been done in the SNP domain where you have many, many genes and few [allele] possibilities at each gene,” said Listgarten. “In our HLA domain you have just a few genes … but hundreds of possible [alleles] at each one. This poses different challenges.”

“Essentially, this is a way to do a probabilistic re-typing in silico.”

In model fitting, when there are many parameters and relatively little data, you “take the data and learn particular parameters based on the data, [but] you don’t have enough data to do that well,” she said. As a result, the accuracy of the result suffers.

The researchers chose to use a regression function that allowed them to reduce the number of parameters in the model in an “intelligent way,” Listgarten said. “That lets us get a better statistical estimate and then we can show it is a better statistical estimate in our experiments, [an aspect that] … nobody has actually looked at."

The statistical refinement methods they chose, said Listgarten, are common approaches to tackling general problems in machine learning or applied statistics.

"These are things that people pick out of their toolbox and for us are familiar from any number of domains, including computational biology," she said. "We know there is always a problem if you don't have enough data relative to the number of parameters."

Model Try-Out

The team trained its model with data from two high-resolution data sets: One set of 13,270 individuals provided by study co-author Mary Carrington, who heads the HLA Typing section of the National Cancer Institute's Laboratory of Experimental Immunology. The second set of 2,400 individuals was drawn from the publicly available dbMHC database.

The researchers then used this trained model to resolve ambiguity in a lower-resolution data set. For example, if a patient's allele at a given locus could be one of several different possibilities, the statistical model draws from the training data to assign a probability to each of those possible alleles.

"For the high-resolution data we have from Mary Carrington, we essentially pretend for some of it that we don't have high-resolution data and we use our technique to fill it in and see how well did we do," said Listgarten.

The "good news" in the paper, said Heckerman, is that this method can "produce good extensions" of HLA typing in studies with the kinds of differences between populations one might expect.

The team compared its approach to the currently used expectation-maximization method and found that "our method statistically significantly outperforms the sort of standard approach," Listgarten said. "Essentially this is a way to do a probabilistic re-typing in silico."

Rescuing Old Data

This method could potentially find use in many applications, according to the researchers. Very-high resolution HLA-typing data is essential for immunology applications such as determining matches for bone marrow transplants or developing vaccines.

"The more refined, the more accurate we are," Carrington wrote in an e-mail to *BioInform*. "The more accurate we are, the less likely we are to make clinically relevant mistakes, like mismatching in transplant settings."

Carrington noted that the high variability of the HLA region is a challenge for researchers. "When you have lots of variants in a sequence, unless you know something about the likelihood of one variant being found on the same chromosome (same allele) as a second variant, it can be ambiguous," she explained.

She said old, lower-resolution data can often be quite informative for projects studying special disease cohorts, or a rare population with individuals who can no longer be typed, or when the samples are simply no longer available for re-typing.

“We can ignore it or try to estimate its full typing, which is useful if the method for doing so is robust,” she said.

Carrington said that she performs high-resolution typing for her research, so the new tool does not have immediate application to her work, but she said there are many labs that only have access to low-resolution typed data, and in those cases, the method “definitely could be applied to lots of old transplant typing lab data.”

Heckerman said that refining HLA-typing techniques could help develop an AIDS vaccine, a field that has seen recent setbacks. Research has shown that the way a human body reacts to HIV depends on very specific variations in HLA types, he said.

Only a difference of a few nucleotides in the HLA gene “makes a big difference in how you respond to HIV,” said Heckerman. And because HIV responds in this precise fashion, “we need [high-resolution] HLA typing to map out how HIV responds to our immune system.”

While Microsoft Research is first and foremost “about advancing the state of the art in computer science,” he said, these advances increasingly help to solve problems in many areas including biology.

“When someone at Microsoft Research creates some innovation and it turns out it has an implication for society, we’re encouraged to follow through, even though it might not have a direct impact on Microsoft’s main business,” he said.