bhi
biomedical
and health
informatics

UNIVERSITY OF WASHINGTON

Biomedical and Health Informatics Lecture Series

Tuesday, April 27, 2010
12:00 - 12:50 p.m., Room RR-134

Andrew M. Simms, BS, MS

Biomedical and Health Informatics Graduate Student in PhD Program
Daggett Research Group, University of Washington, Seattle

"Mining Mountains of Data: Organizing All Atom Molecular Dynamics Protein Simulation Data into SQL and OLAP Cubes"

A significant portion of my research effort has involved organizing all atom molecular dynamics protein simulation data into a form that is both manageable and is conducive to analysis. These simulation data consist of multi-gigabyte collections of four-dimensional atomic coordinates (*x*, *y*, *z* and *time*) as well as a rich set of secondary analyses that are derived from the coordinate trajectories. I designed a relational schema to house these data for my master's work, and have continuously extended and enhanced the design to accommodate longer simulations, additional analyses, as well as a variety of mining queries.

The relational model has worked well for managing our data in general. However, the ideal representation of large matrices, such as those that represent distances between individual atoms in frames of a simulation has been elusive. In this talk, I will discuss modeling simulation data, including atomic contact matrices, as OLAP cubes and some early results of analyses based on this model.

Andrew M. Simms has a bachelor's degree in Computer Science and worked in the software industry for 15 years. He joined the BHI program as a student in 2005 and completed his Master's degree in May 2007. He started his NLM traineeship in July 2007 and is continuing to build on his Master's thesis work, a very large scale data warehouse for protein simulation data produced in the Daggett Lab. The warehouse design includes a scalable and extensible SQL schema, ETL utilities, and a web interface available at http://www.dynameomics.org. He currently is designing analytics methods and tools to study how single nucleotide polymorphisms (SNPs) alter protein conformation and function. His interests include large scale database design; relational and multi-dimensional modeling of biological data; causes and mechanisms of rheumatic diseases; and personal health tools for managing chronic conditions.

* *

NOTE: Podcasts from MEBI 590 Lecture Series talks for this quarter are available at
http://courses.washington.edu/mebi590/schedule.htm

Podcasts from previous quarters are available at
http://courses.washington.edu/mebi590/past.lecture.schedules.html