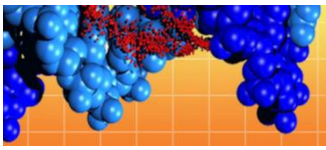

CENTER FOR BIOINFORMATICS & COMPUTATIONAL BIOLOGY

BIOINFORMATICS 2015 SPRING SEMINAR SERIES

Hosted by: Department of Computer and Information Sciences,
Department of Electrical and Computer Engineering &
Center for Bioinformatics and Computational Biology
http://bioinformatics.udel.edu/seminars

MONDAY, February 23, 2015
3:30pm DBI Room 102

Enabling Scalable Data Analysis of Large Computational Structural Biology Datasets on Distributed Memory Systems

Michela Taufer, PhD

David L. and Beverly J.C. Mills Chair of
Computer and Information Sciences University of Delaware
http://gcl.cis.udel.edu/personal/taufer/index.php
http://gcl.cis.udel.edu

ABSTRACT: Today, petascale distributed memory systems perform large-scale simulations and generate massive amounts of data in a distributed fashion at unprecedented rates. This massive amount of data presents new challenges for the scientists analyzing the data to extract scientific meaning. In case of clustering of this data, traditional analysis methods may require the comparison of single records with each other in an iterative process, and therefore involve moving data across nodes of the system. When both the data and the number of nodes increase, clustering methods can increase pressure on the storage and the bandwidth of the system. Thus, the methods become inefficient and do not scale. New methodologies are needed to analyze data when it is distributed across nodes of large distributed memory systems. In general, when analyzing structural biology datasets, we focus on specific properties of the data records such as the molecular geometry or the location of a molecule in a docking pocket. Based on this observation, in this talk we propose a methodology that allows the scalable analysis for large datasets composed of millions of individual structural biology records in a distributed manner on large distributed memory systems. The methodology is based on two general steps. The first step extracts concise properties or features of each data record and represents them as metadata in parallel. The second step performs the clustering on the extracted properties using machine-learning techniques. We apply the methodology to two different computational structural biology datasets to identify geometrical features that can be usedto (1) predict class memberships for structural biology datasets containing ligand conformations from protein-ligand docking simulations and (2) find recurrent folding patterns within and across trajectories (i.e., intra- and inter-trajectory respectively) in multiple trajectories sampled from folding simulations. Our results show that our approach enables scalable clustering analyses for large-scale computational structural biology datasets on large distributed memory systems. In addition, our method achieves better accuracy comparing to traditional analysis approaches.

Bio: Michela Taufer is the David L. and Beverly J.C. Mills Chair of Computer and Information Sciences and an associate professor in the same department at the University of Delaware. She earned her master's degrees in Computer Engineering from the University of Padova (Italy) and her doctoral degree in Computer Science from the Swiss Federal Institute ofTechnology (Switzerland). From 2003 to 2004 she was a La Jolla Interfaces in Science Training Program (LJIS) Postdoctoral Fellow at the University of California San Diego (UCSD) and The Scripps Research Institute (TSRI), where she worked on interdisciplinary projects in computer systems and computational chemistry. From 2005 to 2007, she was an Assistant Professor at the Computer Science Department of the University of Texas at El Paso (UTEP). She joined the University of Delaware in 2007 as an Assistant Professor and was promoted to Associate Professor with tenure in 2012.Taufer's research interests include scientific applications and their advanced programmability in heterogeneous computing (i.e., multi- core and many-core platforms, GPUs); performance analysis, modeling, and optimization of multi-scale applications on heterogeneous computing, cloud computing, and volunteer computing; numerical reproducibility and stability of large-scale simulations on multi- core platforms; big data analytics and MapReduce.