Hai-Son P. Le

Curriculum Vitae

September 2011

Address: 600 S. Negley #A3 Pittsburgh, PA 15232
Phone: +1 650-336-3418
Email: haisonle@gmail.com
WWW: www.cs.cmu.edu/~hple

Education

2013(expected)	Ph.D. in Machine Learning	Carnegie Mellon University
2011	M.S. in Machine Learning	Carnegie Mellon University
2006	B.S. in Computer Science	University of Texas at Austin

Research

2008-Present · Graduate Research Assistant · Machine Learning Department, Carnegie Mellon University
Thesis adviser: Ziv Bar-Joseph

Error correction for RNA-Seq reads

Inferring Interaction Networks: Applications to microRNA target prediction

Determining interactions between entities, the overall organization and clustering of nodes in networks is amajor challenge when analyzing biological and scocail network data. We develop a model to integrate noisy interaction scores with properties of individual entites for inferring interaction networks and clustering nodes within these networks. We focus on applications to study how microRNAs regulate mRNAs in cells.

Cross-Species Expression Analysis

Inferring Ortholog and clustering of genes
Recent studies compare gene expression data across species to identify core and species specific genes in biological systems. To perform such comparisons researchers need to match genes across species. This is a challenging task since the correct matches (orthologs) are not known for most genes. Here we develop a new method that can utilize soft matches (given as priors) to infer both, unique and similar expression patterns across species and a matching for the genes in both species. Our method uses a Dirichlet process mixture model which includes a latent data matching variable.
Large gene expression databases query
Expression databases, including the Gene Expression Omnibus and ArrayExpress, have experienced significant growth over the past decade and now hold hundreds of thousands of arrays from multiple species. However, while several methods exist for finding co-expressed genes in the same species as a query gene, looking at co-expression of homologs or arbitrary genes in other species is challenging. Thus, to carry out cross-species analysis using these databases, we need methods that can match experiments in one species with experiments in another species.

2005-2006 · Undergraduate Research Assistant · University of Texas at Austin

Cache-oblivious algorithms and Hirschbergs space reduction technique. Experimental evaluation of existing implementations and new algorithms. NP-harness proof for the problem of finding the maximum agreement subtree of k area claedograms

Honors and awards

2010 Nominated by MLD for IBM fellowship
2008 Faithful Steward Endowed Fellowship in CSE, University of Washington (declined)
2008 Microsoft Endowed Recruitment Award, University of Washington (declined)
2008 Google's EMG Award for contribution to Gmail
2007 Department of Computer Science - Undergraduate Honors Thesis Award
2006 Honorable Mention, Computing Research Association (CRA) Outstanding Undergraduate Award
2006 College of Natural Sciences - Deans Honored Graduate (top 1%)
2006 Graduated with High Honor, Turing Scholars Honors, Special Honor in CS, GPA: 3.93 (Major: 4.00)
2006 Phi Beta Kappa Society
2006 Angus G. and Erna Pearson Endowed Undergraduate scholarship

Industry experience

Summer 2009 · Software Engineer · Google Inc. Pittsburgh, PA

- Researched methods to extract relationship between products through web crawl data
- Developed a novel method combining hierarchical clustering and matrix operations
- Finished a prototype and a pipeline for automatic extraction of related product clusters

2007-2008 · Software Engineer · Google Inc. Mountain View, CA

- Worked in the Gmail and Video infrastructure Automation team
- Received the EMG Award for contribution in Gmail
- Developed internal applications for tracking bugs and code review
- Analyzed audio and video track of video files to detect audio/video sync issues

Summer 2005,2006 · Software Design Engineer · Microsoft Corp. Redmond, WA

- Worked in the Windows Presentation Foundation everywhere (WPF/e) / SilverLight team
- Researched and implemented a prototype for the integration to post-Vista applications
- Factored and designed code structure for easier future development and maintenance
- Improved the Contact Management feature of Outlook 2007

Summer 2004 · Software Engineer · National Instruments Austin, TX

- Worked Designed and implemented the Intermediate Language Interpreter, written in Assembly and C++ to create a cross-platform runtime environment for LabVIEW

Publications

Journals

1. Ganapathy, G., B. Goodson, R. Jansen, H.-S. Le, V. Ramachandran, and T. Warnow (2006). Pattern identification in biogeography. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 334–346.
2. Chowdhury, R., H.-S. Le, and V. Ramachandran (2010). Cache-oblivious dynamic programming for bioinformatics. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* **7**(3), 495–510.
3. Le, H.-S., Z. Oltvai, and Z. Bar-Joseph (2010). Cross-species queries of large gene expression databases. *Bioinformatics* **26**(19), 2416.
4. Gupta, A., P. Nagilla, H. Le, C. Bunney, C. Zych, A. Thalamuthu, Z. Bar-Joseph, S. Mathavan, and V. Ayyavoo (2011). Comparative Expression Profile of miRNA and mRNA in Primary Peripheral Blood Mononuclear Cells Infected with Human Immunodeficiency Virus (HIV-1). *PloS one* **6**(7), e22730.

Papers in conference proceedings

1. Le, H.-S. and Z. Bar-Joseph (2010). "Cross Species Expression Analysis using a Dirichlet Process Mixture Model with Latent Matchings". In: *Advances in Neural Information Processing Systems 23*. Ed. by J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, pp.1270–1278.
2. Le, H.-S. and Z. Bar-Joseph (2011). "Inferring Interaction Networks using the IBP applied to microRNA Target Prediction". In: *Advances in Neural Information Processing Systems 24*.

Unpublished working papers

1. Le, H.-S. (2006). Algorithms for Identification of Patterns in Biogeography and Median Alignment of Three Sequences in Bioinformatics. *Undergraduate Honors Thesis*, *Department of Computer Sciences*, *University of Texas at Austin*, *CS-TR-06-29*.

Skills

- Languages: R, C/C++, Java, Perl, PHP, Python, SQL, Scheme, ML, Bash shell, AWK
- Tools: lex, yacc, GTK++, Eclipse SWT/JFace, MATLAB, XCode
- Languages: fluent in Vietnamese, proficient in French
- Sports: soccer, tennis, swimming