ABSTRACT

MEDLINE®/PubMed® is a richly annotated resource of over 21 million article citations, growing at a modern rate of over 600,000 citations annually. One grand challenge of bioinformatics is analysing the extensive literature for a biomedical entity such as a gene or disease. This thesis explores using over-representation to extract pertinent biomedical annotation from the research articles for an entity. The quantitative profiles generated are compared to predict novel associations between entities.

Medical Subject Heading Over-Representation Profiles (MeSHOPs) are constructed from the primary literature of an entity of interest. Medical subject annotations for each article are extracted. Statistical tests evaluate the significance of each term's frequency across the set of articles, compared against an appropriate background set. The resulting MeSHOP is composed of each term and corresponding enrichment p-value.

MeSHOPs can be computed for any entity with an associated bibliography of PubMed articles. We evaluate the predictive performance of quantitatively comparing MeSHOPs to discover novel associations between gene and disease entities, achieving up to 16% improvement in accuracy compared to gene or disease baseline features (measured as increased Receiver Operating Characteristic Area Under the Curve). Strong literature annotation level bias on the predictive performance for future gene-disease association was seen. We observe similar results in a parallel analysis of associations between drugs and disease.

Efficiently identifying authors with similar research interests is a challenge in science. During the peer review process, authors seek scientists with similar expertise. MeSHOPs are generated for individual authors, identifying their research foci. Extending the methods to allow comparison across large sets of entities, overlapping research interests between researchers were identified. The predictive performance was evaluated for capacity to identify authors working in the same research domains.

Biomedical annotation analysis of primary literature provides insight into the areas of research focus, and is demonstrated to link entities through similarities in their MeSHOPs. We quantitatively confirm the trend where well-studied genes, diseases and drugs are more likely to be the focus of further research. MeSHOP analysis demonstrates that knowledge in the annotated primary literature can be efficiently mined, and the untapped knowledge can be discovered computationally.

BIOGRAPHICAL NOTES

Born:	June 18, 1979, Hong Kong
Academic Studies:	B.Sc., University of British Columbia, 2002 M.Sc., University of British Columbia, 2005

GRADUATE STUDIES

Field of Study:	Biomedical Literature Over-Representation Annotation

Selected Courses		**Instructors**
CPSC 545	Algorithms of Bioinformatics	Dr. H. Hoos
MATH 561	Mathematical Biology II	Dr. Y. Li
MEDG 520	Advances in Human Molecular Genetics	Dr. M. Lorincz
STAT 540	Statistical Methods for High Dimensional Biology	Dr. R. Gottardo
CMPT 880	Medical Imaging (SFU)	Dr. G. Hamarneh
MBB 841	Bioinformatics (SFU)	Dr. F. Brinkman
MBB 821	Nucleic Acids (SFU)	Dr. D. Sen

SELECTED AWARDS

Micheal Smith Foundation for Health Research Senior Graduate Studentship
National Science and Engineering Research Council Postgraduate Scholarship D3

SELECTED PUBLICATIONS

W. Cheung, G. Hamarneh. n-SIFT: n-Dimensional Scale Invariant Feature Transform. *IEEE Transactions on Image Processing*. 18, no. 9 (2009):2012 - 2021.
Warren Cheung and William Evans. Pursuit-Evasion Voronoi Diagrams in L1. In Proceedings of the *4th International Symposium on Voronoi Diagrams in Science and Engineering (ISVD) 2007*. pp.58-65.

SELECTED PRESENTATIONS

W. A. Cheung, BF F. Ouellette, W. W. Wasserman. Medical Subject Heading Over-representation profiles: Bibliographic Analysis of Over-represented Medical Subjects. Poster presented at CIHR National Health Research Poster Presentation, Canadian Student Health Research Forum 2011. June 7-9, 2011, in Winnipeg, Canada.
W. A. Cheung, BF F. Ouellette, W. W. Wasserman. Predicting Gene-Disease Relationships via Gene Characteristic Profiles Constructed from Text Analysis. Poster presented at 14th International Conference on Research in Computational Molecular Biology. August 12, 2010 in Lisbon, Portugal.
Co-Chair of the Canadian Student Conference on Biomedical Computing, 2009..

SUPERVISORY COMMITTEE

Angela Brooks-Wilson (Medical Genetics)
Jennifer Bryan (Statistics)
Kendall Ho (Emergency Medicine)
Francis Ouellette (co-supervisor, Cell and Systems Biology, University of Toronto)
Wyeth W. Wasserman (co-supervisor, Medical Genetics)

Faculty of
Graduate Studies

UBC

THE UNIVERSITY OF BRITISH COLUMBIA

PROGRAMME

The Final Oral Examination
For the Degree of

DOCTOR OF PHILOSOPHY
(Bioinformatics)

WARREN A. CHEUNG

B.Sc., University of British Columbia, 2002
M.Sc., University of British Columbia, 2005

Friday, July 27, 2012, 12:30 pm
Room 101, Michael Smith Laboratories
Latecomers will not be admitted

"Inferring Novel Relationships through Over-Representation Analysis of Medical Subjects in Biomedical Bibliographies"

EXAMINING COMMITTEE

Chair:
Dr. Wayne Riggs (Pharmaceutical Sciences)

Supervisory Committee:
Dr. Wyeth W. Wasserman, Research Co-Supervisor (Medical Genetics)
Dr. B.F. Francis Ouellette, Research Co-Supervisor (Cell and Systems Biology, University of Toronto)
Dr. Jennifer Bryan (Statistics)

University Examiners:
Dr. Daniel Goldowitz (Medical Genetics)
Dr. Frederic Pio (Molecular Biology and Biochemistry, Simon Fraser University)

External Examiner:
Dr. Chris Upton *(Attending)*
Department of Biochemistry and Microbiology
University of Victoria
Victoria, British Columbia