

STANFORD UNIVERSITY
DEPARTMENT OF STATISTICS
DEPARTMENTAL SEMINAR

4:15 p.m., Tuesday, August 7, 2007
Sequoia Hall Room 200
(Cookies at 3:45 in 1st Floor Lounge)

Balaji S. Srinivasan
Department of Statistics
Stanford University

Automatic Population of Biomedical Ontologies

Biomedical classification systems like the Gene Ontology (GO) have proven invaluable for converting disordered collections of free text into machine-readable knowledge representations. However, the scalability of these ontologies is currently limited because they are populated manually from the literature at great expense. Here, we present an algorithm which removes this limitation by automatically extracting ontological relationships between objects from a massive corpus of more than 425000 full text biomedical articles.

Given a small training set of biological objects with known relationships such as “`is_a`”, “`localized_to`”, or “`regulates_a`”, our algorithm finds the lexico-syntactic patterns which specify this relationship in plain text. These learned patterns can then be used to find many more examples of objects that satisfy these relationships, thereby automatically populating an ontology. As a case in point, we show the results of applying the algorithm to locate text snippets specifying gene localizations, taxonomic relationships, and anatomical connections. Our methods greatly reduce the amount of manual curator effort, thereby allowing ontological modeling to scale.