MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

GenesTrace™: Biological Knowledge Discovery via Structured Terminology

Indra Neil Sarkar, MPhil[†,a], Michael N. Cantor, MD[†,a],
Olivier Bodenreider, MD, PhD[b], Yves A. Lussier, MD[a]

[a]Department of Biomedical Informatics, College of Physicians & Surgeons,
Columbia University, New York, NY, USA
[b]Lister Hill National Center for Biomedical Communications,
National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

[†] These authors contributed equally

Much of biological hypothesis generation follows the proverbial model of trying to discover the "golden needle in the haystack." Finding that 'golden needle' is becoming an even more daunting task, as various genome projects are creating sequence data information at accelerating rates. As a result, it is quickly becoming an intractable problem for the biomedical scientist to stay abreast all of the putative genes that may hold the hidden keys toward the understanding of disease.

One of the great promises in the modern era of molecular biology is the discovery of novel genes and their relationship to the molecular basis of disease. The need for a uniform method of knowledge representation in this endeavor led to the development of the Gene Ontology™ (GO) and its related databases of gene annotation. As the Unified Medical Language System® (UMLS®) is the most comprehensive clinical terminology, mappings between its Metathesaurus® and GO are excellent candidates for the intermediate step linking molecular findings to clinical manifestations of disease.

We have developed *GenesTrace*™, a method that reveals relationships (*traces*) between a disease and a gene using a three-step process. First, diseases are identified that exist in the UMLS® as a concept. Second, relationships are determined between the concept and other UMLS® concepts using both the symbolic relationships and the statistical relationships (co-occurrence information). Finally, putative genes are identified using the related concepts and mapping them to annotated gene lists from a biological terminology such as Gene Ontology. The method is outlined in Figure 1.

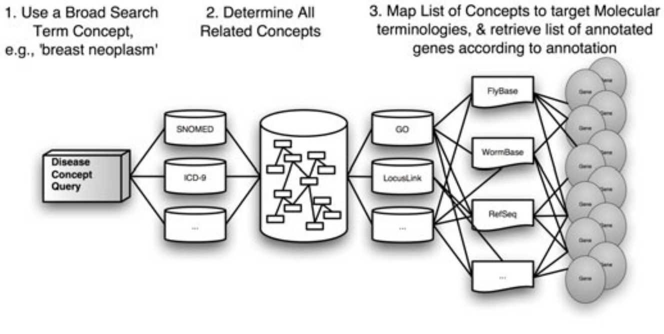
1. Use a Broad Search Term Concept, e.g., 'breast neoplasm'
2. Determine All Related Concepts
3. Map List of Concepts to target Molecular terminologies, & retrieve list of annotated genes according to annotation
Disease Concept Query
SNOMED
ICD-9
...
GO
LocusLink
...
FlyBase
WormBase
RefSeq
...
Gene

Figure 1: *GenesTrace*™ as a three-step process. Some current methods for accessing annotated genes are shown in grey.

To test the feasibility of *GenesTrace*™, we applied it to Breast Cancer (BC) and Alzheimer's Disease (AD). For each trace, we retrieved gene products that have shown to be involved with the disease etiology. Among the list for the BC trace (consisting of 168 GO terms; 10,532 molecular product annotations), we found the murine form of BRCA1. Similarly for the AD trace (resulting in 106 GO terms; 9,526 unique molecular annotations), we found both the murine form of A2M and Amyloid-Beta Precursor Protein.

GenesTrace™ is able to create relevant links between clinical and molecular knowledge. For diseases on which we performed *traces*, we were able to retrieve known molecular determinants. We are working on automating the overall process of doing *traces* on a large scale. We are exploring different views of the data that may be useful to the biomedical scientist. It is our hope that, through the *GenesTrace*™ method, biomedical scientists will be able to identify potential genes that may be of significance to the etiology of disease.

Supplementary material can be found at the following website: http://www.dbmi.columbia.edu/lussier/GenesTrace/

Acknowledgments

INS and MNC are supported by NLM Training Grant LM-07079. Partial support for this work came from NYSTAR-sponsored Center for Advanced Technology at Columbia University Grant C020054 and from the National Institute of Allergy and Infectious Diseases Grant 1U54 AI 57159-01.

Address for correspondence

Yves A. Lussier, MD
622 W. 168th Street
Vanderbilt Clinic, 5th Floor
New York, NY 10032 USA
Yves.lussier@dbmi.columbia.edu