Microphenome: Linking Clinical Symptom Clusters via Gene Expression

Radhika Malik

6.872 Final Project

1. Introduction

Most of the work in current genomic research has been focused on mapping genes and proteins to diseases. However, a formal representation of mapping symptoms to gene expression is yet to be created. In this project, we explore one possible approach of testing whether underlying symptoms of diseases have common genetic factors. This project is based on the work done by Neena Parikh, Amin Zollanvari and Gil Alterovitz on analyzing publically available gene expression data for predictive medicine (1). The authors of this paper constructed an automated pipeline involving a closed loop Bayesian network using the Gene Expression Omnibus (GEO) database (2) as their primary data source. In this project, I aimed at utilize a process similar to that described in the paper, but instead of clustering GEO experiments according to diseases, my aim was to cluster them according to symptoms.

2. Materials and Methods

In this section, we discuss the methods used in our approach to explore whether symptoms have common genetic factors.

2.1 The GEO Database

GEO is a publically available repository, which archives high-throughput genomic data submitted by the scientific community. This database organizes genomic data in 4 main categories: Platforms, Samples, Series and DataSets; the first three kinds of data are submitted by researchers while the last is compiled from the others by. For the purpose of this project, we primarily considered Samples and DataSets. A GEO Sample describes the conditions under which a single sample was handled, as well as measurements associated with it. A GEO DataSet (GDS) record represents a collection of biologically and statistically comparable GEO Samples; each GDS record represents an experiment. I obtained a list of all GDS files from the authors of (1).

2.2 Mapping Experiments to Symptoms using MeSH terms

The authors of (1) first obtained a list of all GEO experiments pertaining to a given disease; they did this by downloading all available GEO DataSet files and then mapping each of these experiments to its corresponding PubMed identification

number; each PubMed citation contains a set of Medical Subject Heading (MeSH) (3) terms, from which they were able to map GEO experiments to MeSH terms. They then obtained a mapping from diseases to experiments by comparing each experiment's associated MeSH term with a list of diseases, and matching the experiment with a disease if a match was found. This technique is illustrated in Figure 1.

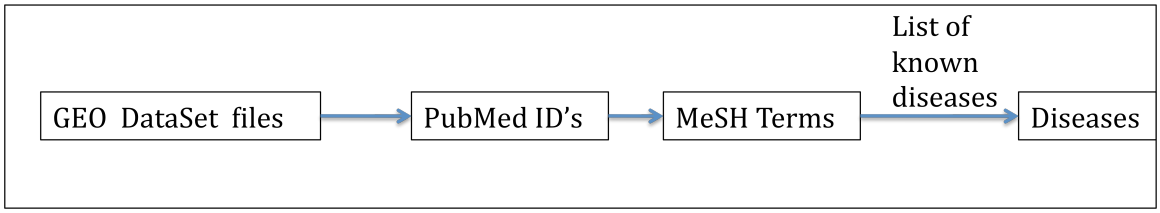

Figure 1: Mapping GEO DataSet Files to Diseases

The first step in my project was to obtain a similar list of all GEO experiments pertaining to a given symptom. My first approach was very similar to the one used in the paper; I took the MeSH terms corresponding to each GEO experiment and compared each MeSH term with a list of symptoms. The list of symptoms included all nodes at the leaves of the ICD-10 (4) knowledge base. Although many symptoms are captured in the leaf nodes of the ICD-10 base, several of these nodes are not actually symptoms. Thus, my approach was to map GEO experiments to all these terms, and then filter out the ones corresponding to symptoms (Figure 2).

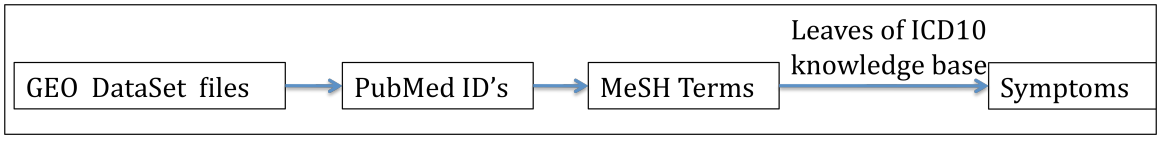

Figure 2: Mapping GEO DataSet Files to Symptoms

The technique that had been used to map GDS records to diseases did not give work as well to map GDS records to symptoms. When I compared each MeSH term exactly with an ICD-10 leaf term, only 1 of the 859 experiments was associated with an ICD-10 term (this term was “Glycosuria”). Even after relaxing the code from only considering exact matching of MeSH terms with ICD-10 terms to now matching an experiment with an ICD-10 term if the term was contained in the MeSH term (for example, the experiment with a MeSH term "Immunoglobulin Heavy Chains” was now also matched with the ICD-10 term "Immunoglobulin"), only 20 experiments had an associated ICD-10 term. These unsatisfactory results indicated that an alternative approach to obtain the experiment to symptoms mapping was required.

2.3 Mapping Experiments to Symptoms using MeSH terms and UMLS concepts

The second approach I employed was to map each experiment to UMLS (Unified Medical Language System) (5) concepts along with MeSH terms, and then use this combination of UMLS concepts and MeSH terms to map experiments to symptoms. Since UMLS consists of terms from many more ontologies, each experiment has a significantly greater number of UMLS concepts than MeSH terms associated with it. To obtain UMLS concepts corresponding to each disease, I used the free text description associated with each DataSet in GEO. Using the GEOQuery package (6), I obtained the description for each experiment and then derived a list

of UMLS concepts from each description using the MetaMap (7) software, a program to map biomedical text to UMLS concepts (Figure 3).

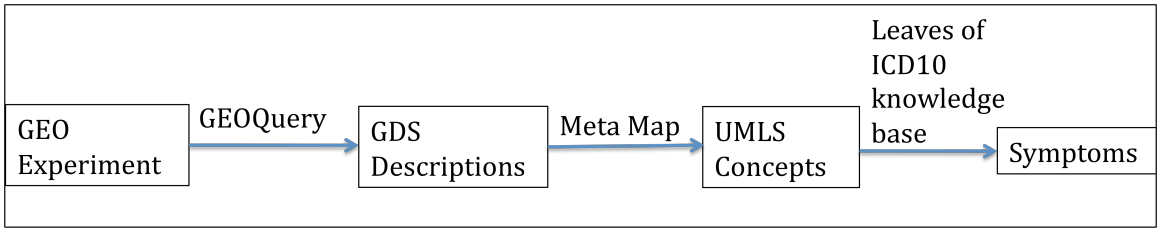

Figure 3: An Alternative Approach to Mapping GEO DataSet Files to Symptoms

Each experiment, from only being mapped to 20-30 MeSH terms on average was now mapped to more than 100 UMLS concepts and MeSH terms. Using this enhanced mapping, I used the ICD10 leaf terms again to compare the concepts and terms associated with each experiment to a list of symptoms. With this approach, 68 experiments had UMLS concepts that matched 25 ICD-10 leaf terms. Using the same relaxation from before of also including symptoms contained in terms, 94 GEO experiments matched with 32 ICD-10 leaf terms matched GEO experiments.

2.4 Approximate String Matching to Enhance Experiments to Symptoms Mapping

To further improve the matches obtained, I explored the possibility of using approximate string matching while comparing the UMLS concepts and MeSH terms with the list of symptoms. The approach I used was to match an experiment with a symptom either if the symptom was contained in, or approximately matched with

the experiment's UMLS concepts/MeSH terms. A variety of approximate string matching techniques have researched such as Levenshtein distance, Smith-Waterman algorithm, Jaro-Wrinkler distance, Needleman–Wunsch algorithm, Jaccard similarity index, QGramsDistance, Monge Elkan distance and Cosine Similarity; a few including Smith-Waterman and Needleman-Wunsch have been applied in bioinformatics (9). I used the Simmetrics (8) open source library to experiment with the various matching methods and to estimate the best metric for this given matching.

Most of the algorithms resulted in a large number of false positives. For example, using the Jaro-Winkler distance and the Smith-Waterman algorithm, "malignant neoplasm of ovary" was continually matched with "malignant neoplasm of prostrate". The nature of the problem was such that terms should be matched only if they differ by a few letters. Thus, a baseline test for each metric was that it should rank the matching between terms with only 2 or 3 letters differing higher than a match between strings that are lexically quite apart such as "malignant neoplasm of ovary" and "malignant neoplasm of prostrate". The final metric I settled on was Levenshtein distance, which is defined as the minimum number of edits required to transform one string into the other; an allowable edit operation is an insertion, deletion, or substitution of a single character.

I first tried to match UMLS concepts/MeSH terms with symptoms if they were a Levenshtein distance of 2 apart. Using this metric, 165 experiments either

had a symptom contained in their terms or had terms approximately matching a symptom. However, upon closely scrutinizing the matched terms with symptoms, I noticed some false positives, such as diseases with "Stenosis" as a UMLS concept being mapped to the ICD-10 term "Stannosis". Consequently, I decreased the distance threshold for the approximate string match to be only 1 edit apart. Now, 148 experiments were matched with 46 symptoms. It is important to note, however, that approaches with approximate string matching took significantly longer than matching with basic lexical search.

2.5 Identifying Case versus Control for Symptoms

The existing automated pipeline developed to merge the GEO experiments and obtain all relevant gene expression data associated with a disease takes as input all experiments corresponding to a disease and filters experiments that include GEO Samples with two states for a disease (control state versus disease state). I considered the list of the mapped ICD-10 terms and manually curated the terms that corresponded to symptoms; these terms were usually of semantic type "Sign or Symptom", "Finding", or "Acquired Abnormality". I considered a few that could potentially be symptoms, and passed in their corresponding experiments through the GEO pipeline. From the experiments that were selected by the pipeline, I would then check whether the disease state (case/control) could be extended to the symptom state.

3. Results and Findings

In this section, we summarize the results and findings of the approach we use to establish whether symptoms have genetic factors.

3.1 Nature of the matched ICD-10 terms

Most of the matched ICD-10 leaf terms were not symptom terms. Many of these terms either correspond to disease/syndromes or neoplastic processes. These include lymphocytic choriomeningitis, chlamydial pneumonia, tuberous sclerosis, eosinophilia, primary open-angle glaucoma, valproic acid, tuberculoid leprosy, ischaemic cardiomyopathy, malignant neoplasm of ovary, psoriasis vulgaris, immunoglobulin, refractory anaemia with ringed sideroblasts, multiple myeloma, exposure to ionizing radiation, living alone, borderline tuberculoid leprosy, pneumonia due to klebsiella pneumoniae, actinic keratosis, exposure to radiation, tetracyclines, sulfonamides, juvenile dermatomyositis, borderline leprosy, acne vulgaris, acute myelomonocytic leukaemia, lifestyle-related condition, cryptosporidiosis, malignant neoplasm of prostate, burkitt lymphoma, drug use, isolation, osteolysis, pterygium, severe pre-eclampsia, multiple sclerosis, bcg vaccine and muscular dystrophy.

3.2 Pipeline Experiments with Matched Symptoms

For terms that could potentially be characterized as symptoms, running their corresponding GEO DataSet ID's through the pipeline yields the following results.

3.2.1 Dyspnoea

The first symptom I considered was Dyspnoea. Its corresponding GDS records were GDS1505, GDS1568 and GDS3071. The pipeline discards experiments whose samples either have the keyword "time" in a column except those titled "sample" and "description", or that do not have an identifiable case/control state for the samples. For Dyspnoea, the pipeline discarded all experiments.

3.2.2 Nervousness

The corresponding GDS records for nervousness were GDS2652, GDS1979, GDS532, GDS916, GDS1347, GDS2470 and GDS1112. The pipeline filtered these experiments to yield GDS1112 and GDS2470. However, for these experiments, the disease state (control/case) of the DataSet samples did not extend to the symptom state.

3.2.3 Congestive Heart Failure

The corresponding GDS records for nervousness were GDS2205, GDS2206 and GDS1362. The pipeline rejected all three experiments.

3.2.4 Muscle strain

The corresponding GDS records for nervousness were GDS3041, GDS3434, GDS3260, GDS1478, GDS1530, GDS177, GDS858, GDS3333 and GDS1676. The pipeline filtered these experiments to yield GDS3041 and GDS1478. For these experiments, the disease state (control/case) of the DataSet samples did not extend to the symptom state.

3.2.5 Senility

The corresponding GDS records for nervousness were GDS3041 and GDS3434. The pipeline rejected both experiments.

3.2.6 Ganglion

There was only 1 GDS record for Ganglion: GDS2892. Although the pipeline accepted this experiment, the disease states of the samples did not extend to symptom states.

3.2.7 Hypothermia

There was only 1 GDS record for hypothermia (GDS1886) and it was rejected by the pipeline.

3.2.8 Albinism

There was only 1 GDS record for hypothermia (GDS2723) and it was rejected by the pipeline.

4. Analysis

As discussed in the previous section, most of the ICD-10 terms matched with experiments were not actually symptoms. Furthermore, although the leaves of the ICD-10 knowledge base include several symptoms, they also miss many others. Even a casual browse through a medical encyclopedia such as PubMed Health (11) can brings out many common symptoms that the ICD-10 knowledge base misses. Thus, to map experiments to symptoms, an alternative ontology must be used to generate a list of symptoms to compare UMLS concepts/MeSH terms associated with experiments.

For some of the symptoms, including muscle strain, nervousness and ganglion, the disease states (case/control) did not extend to the symptom states. These seem to be primarily because of a large number of terms generated by MetaMap. When run with default options, MetaMap maps each GDS description to a large number of UMLS concepts; these concepts, although directly related to the terms in the GDS descriptions, do not apply in the context of the experiment. For example, the term “muscle strain” has been matched as a UMLS concept of several experiments, which have the term “strain” in their description. However, the term “strain” in the experiment description refers to its meaning in the biological context and not muscular strain. The inherent risk with curtailing the number of matching UMLS concepts returned by MetaMap is that the number of symptoms associated with diseases will decrease, while throughout Section 2, we have been discussing

methods to increase the number of GEO to symptom mappings. With the use of an alternative ontology to ICD-10, this issue might be able to be mitigated.

5. Future Work

I envision a number of possible extensions to this project. One possible approach, as mentioned in the previous section, is to restrict the UMLS concepts returned by MetaMap and to derive a list of symptoms from an alternate source instead of the ICD-10 knowledge base. Another approach is to directly map GDS descriptions to symptoms. MetaMap has the option to restrict the UMLS concepts it returns to specific semantic types. If we consider all ontologies under UMLS but restrict the semantic types to "Sign or Syndrome", we may be able to map experiment descriptions directly to UMLS concepts representing symptoms.

6. Acknowledgements

Firstly, I would like to acknowledge the help of Gil Alterovitz and Amin Zollanvari. Gil introduced me to the project, helped me get started and also provided advice whenever I was stuck. Amin guided me through all stages of the project. Finally, I would like to thank Peter Szolovits for helping me decide on this project, as well as for all the guidance and teaching throughout 6.872.

References

1. Neena Parikh, Amin Zollanvari and Gil Alterovitz, An Automated Bayesian Framework for Integrative Gene Expression Analysis and Predictive Medicine
2. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A. NCBI GEO: archive for functional genomics data sets—10 years on Nucleic Acids Res. 2011 Jan;39(Database issue):D1005-10
3. Medical subject headings, Bull Med Libr Assoc. 1963 Jan;51:114-6. Available at: http://www.nlm.nih.gov/mesh/meshhome.html
4. Organization, W., (2004). *International Statistical Classification of Diseases and Health Related Problems*. Geneva: World Health Organization.
5. Unified Medical Language System (UMLS) U.S. National Library of Medicine (NLM).
6. Davis, S. and Meltzer, P. S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics, 2007, 14, 1846-1847
7. MetaMap 2011. Software available at http://metamap.nlm.nih.gov/
8. SimMetrics, Similarity Metric Library, UK Sheffield University, Funded by (AKT) an IRC sponsored by EPSRC, grant number GR/N15764/01. Software available at: http://sourceforge.net/projects/simmetrics/
9. Navarro G (2001). "A guided tour to approximate string matching". *ACM Computing Surveys* **33** (1): 31–88.

10. Smith, Temple F.; and Waterman, Michael S. (1981). "Identification of Common Molecular Subsequences". Journal of Molecular Biology 147: 195–197.
11. PubMed Health [Internet]. Bethesda (MD): National Library of Medicine (US); [updated 2011 Jan 1; cited 2011 Jan 6]. Available from: http://www.ncbi.nlm.nih.gov/pubmedhealth/