Finding Disease-Related Genomic Experiments Within an International Repository: First Steps in Translational Bioinformatics

Atul J Butte, MD, PhD, Rong Chen, PhD
Stanford Medical Informatics, Department of Medicine and Pediatrics,
Stanford University School of Medicine
Stanford, California USA

Abstract

The amount of gene expression data in international repositories has grown exponentially. An important first step in translating the results of genomic experiments into medicine is to relate these genomic experiments to the human diseases they have studied. Unfortunately, repositories for expression data store the crucial annotative details only as free-text, making it manually intractable to link these with human disease. In this study, we sought to find experiments in NCBI GEO that are related to human diseases by making use of annotations relating these experiments with PUBMED identifiers representing the publication in which each experiment was published. In this manner, we find that 35% of PUBMED-associated genomic experiments can be related to a human disease, and that publicly-available data from these genomic experiments can already be related to over 270 human diseases and conditions. This represents an important first step in bridging the world of nucleotides, transcripts and expression with the afflications of us all.

Purpose

Find genomics experiments in the NCBI Gene Expression Omnibus that are related to human diseases in an automated manner.

Background

The obstacle in translating discoveries made using genomic data and technologies to medicine has been difficult to climb, and has been well described.(1-4) To help address this bottleneck, the emerging discipline of translational informatics is focusing on the development of analytic, storage, and interpretive methods to optimize the transformation of increasingly voluminous genomic and biological data into diagnostics and therapeutics for the clinician.

The past 10 years have led to a variety of measurements tools in molecular biology that are high-bandwidth in nature. The premier example of this is the RNA expression detection microarray, which provides quantitative measurements of expression of over 40,000 unique RNAs within cells.(5, 6) Many of the most illuminating experiments involving microarrays are those that have enabled discoveries related to the diagnosis and treatment of medical conditions, including the determination of therapeutic action,(7) development of diagnostic tests,(8) and distinction between disease subtypes.(9, 10)

Corresponding with this success, the amount of gene expression data in international repositories has grown exponentially, because top-tier journals require the public availability of such data.(11) The NCBI Gene Expression Omnibus (GEO) is an international repository for gene expression data, developed and maintained by the National Library of Medicine.(12) As of this writing, GEO holds 67,903 samples (i.e. microarrays) from over 2,900 experiments involving over 120 species, across over 1,600 types of microarrays, with a total of 1,303,250,456 individual gene measurements. More impressively, GEO has been gaining data at 300% per year.

An important first step in translating the results of genomic experiments into medicine is to determine how many genomic experiments are related to the study of human disease, as well as the characteristics of these experiments and the disease they study. Though GEO is already an incredible resource for gene expression measurements, accessing genomic data that is directly or indirectly related to human disease is manually intractable because the crucial annotative details are stored only as free-text.

We recently described the utility of a text-parser in determining the phenotypic, environmental and experimental context from the annotations of genomic experiments.(13) Specifically, our GENOTEXT system processes seven types of GEO annotations and maps these to matching terms from the Unified Medical Language System (UMLS).(14) While GENOTEXT enables searches for genomic experiments related to virtually any biomedical concept, it also enables the relating of genes showing differential expression associated with these concepts, including aging and injury. Despite this success, we found that text-parsing was still an inefficient method to extract the highest value from these associations.

In this study, we sought to find experiments in NCBI GEO that are related to human diseases, as well as generalizable characteristics of these experiments. To do this, we make use of annotations relating GEO series, or the collections of related microarray samples within a single experiment, with PUBMED identifiers representing the publication in which the GEO series was published. These PUBMED identifiers relate to MEDLINE publication records which are manually annotated with Medical Subject Headings (MeSH) by experts. We map these MeSH identifiers back into UMLS and study their semantic types. In this manner, we find that 35% of PUBMED-associated GEO series can be related to a human disease, and that publicly-available data from these genomic experiments can already be related to over 270 human diseases and conditions.

Methods

Gene Expression Omnibus

The Gene Expression Omnibus (GEO) is an international repository for gene expression data, developed and maintained by the National Library of Medicine.(12) We downloaded 3,104 GEO series files on March 1, 2006 and parsed the annotative fields of each GEO series into a relational database.

For the 1,644 GEO series with PUBMED annotations, we downloaded the MEDLINE records via the NCBI Entrez Programming Utilities.(15) We parsed the MeSH terms from these records into a relational database, resulting in a table of 2,889 unique MeSH terms.

We joined this table with the UMLS 2005AA Concept Names and Sources table (MRCONSO), using exact term matching. Of the 2,889 terms, 146 (5%) could not be matched to UMLS concepts. These described terms that were added to MeSH after the release of UMLS 2005AA, such as *Receptors, Notch*, *Wnt Proteins*, and *Dosage Compensation, Genetic*. The remaining terms were successfully joined with MRCONSO.

We then joined these concepts with the UMLS 2005AA Semantic Types table (MRSTY), to determine the semantic type of the concepts. Disease-related concepts were defined as having semantic types of *Injury or Poisoning* (T037), *Disease or Syndrome* (T047), *Mental or Behavioral Dysfunction* (T048), and *Neoplastic Process* (T191).

The UMLS concepts mapped to GEO series in this manner were compared with the UMLS concepts mapped from GEO series title and description annotations found using GENOTEXT, our previously-described text-parsing approach.(13)

TUI	Semantic Type	Count
T116	Amino Acid, Peptide, or Protein	487
T123	Biologically Active Substance	330
T121	Pharmacologic Substance	204
T047	Disease or Syndrome	171
T109	Organic Chemical	150
T126	Enzyme	140
T023	Body Part, Organ, or Organ Component	102
T025	Cell	100
T114	Nucleic Acid, Nucleoside, or Nucleotide	95
T129	Immunologic Factor	93

Table 1: The top ten semantic types of MeSH annotations of publications related to GEO series. The third column is a count of unique MeSH headings under that semantic type in use. More MeSH headings fall into these ten semantic types than in the remaining 107 semantic types.

Results

In mapping GEO series with associated publications, we find that only 1,644 (53%) of 3,104 GEO series are annotated with PUBMED identifiers. Very few (6 of 3,104, or 0.2%) are annotated with two PUBMED identifiers. As GEO series identifiers are assigned serially, we can determine the relative age of the submitted data. Absence of a PUBMED identifier was significantly associated with more recent submissions (unpaired *t*-test $p \leq 1.13$ x 10^{-7}). Based on experience, it is likely that GEO data is typically deposited early in the mansucript submission process, and most PUBMED annotations are likely entered after actual publication (not acceptance) of the manuscript.

We were successful in mapping to UMLS concepts 2,743 MeSH terms annotating the publication related to GEO data sets. We find that these 2,743 UMLS concepts belong to 117 of the 134 available semantic types in UMLS 2005AA. We find that the MeSH annotations for these publications often contain an indication of a protein, chemical, disease, body part, and cell type. The top ten semantic types are listed in Table 1.

MEDLINE records often indicate a primary MeSH annotation as representing the focus of a paper. Only 62% of our MEDLINE records indicated any MeSH heading as being the focus of the publication, and only 251 of 2,743 concepts (9%) were annotated as being a focus. The majority of these were in the semantic types *Genetic Function* (T045), *Gene or Genome* (T028), *Amino Acid, Peptide, or Protein* (T116), *Biologically Active Substance* (T123), and *Nucleic Acid, Nucleoside, or Nucleotide* (T114), likely representing the primary methods used in genomic experimentation. Notably missing from this list were semantic types related to disease.

We find 571 (out of 1,644 or 35%) GEO series related to the four semantic types we considered as disease-related: *Injury or Poisoning* (T037), *Disease or Syndrome* (T047), *Mental or Behavioral Dysfunction* (T048), and *Neoplastic Process* (T191). Within these semantic types, we find 276 unique concepts representing disease annotated the papers related to these GEO series.

The top 30 disease concepts associated with GEO experiments are listed in Table 2. As expected, cancer dominates in terms of relative number of samples in experiments associated with a disease. Microarray studies typically measure genome-wide gene expresion in a sample of tissue; more tissue is likely obtained, saved, and studied in oncology than in any other medical sub-specialty. *Breast neoplasms* itself is associated with 44 GEO series totalling 2,650 samples.

Non-cancerous diseases are also well-represented in GEO series. *Spinal cord injuries* was associated with GEO experiments with titles such as "Chronic contusion spinal cord injury in rats" and "CNS Regeneration." *Obesity* was associated with GEO experiments with titles including "Diet induced changes in mouse liver" and "Dog heart and high fat diet". *Pharyngitis* was associated with a GEO experiment with the title "Longitudinal analysis of the group A Streptococcus transcriptome".

The majority (953 out of 1,644 or 58%) of disease-related GEO series incorporate samples from *Homo sapiens*; these could either be from primary human samples or from cell lines. Another 11% incorporate samples from *Mus musculus* and 4% include samples from *Rattus norvegicus*. However, GEO series with samples from a total of 25 species are associated with disease. Many of these other species are represented microarray samples obtained from pathogens, including *Mycobacterium tuberculosis* (2 GEO series), *Borrelia burgdorferi* (2), *Coxiella burnetii* (2), *Chlamydia trachomatis* (2), *Staphylococcus aureus* (2), *Vibrio cholerae* (2), *Streptococcus pneumoniae* (2), *Campylobacter jejuni* (2), *Plasmodium falciparum* (1), *Medicago truncatula* (1), *Oncorhynchus mykiss* (1), and *Paracoccidioides brasiliensis* (1).

We compared the MeSH concepts associated with GEO series to those UMLS concepts found by a text-parsing method previously applied to GEO series titles and descriptions.(13) A total of 284 GEO series

CUI	Disease Term	Count GSM
C0006149	Breast Neoplasms	2650
C0027651	Neoplasms	1628
C0001314	Acute Disease	1049
C0024121	Lung Neoplasms	882
C0027627	Neoplasm Metastasis	875
C0001418	Adenocarcinoma	693
C0019204	Carcinoma, Hepatocellular	649
C0037929	Spinal Cord Injuries	571
C0028754	Obesity	570
C0038356	Stomach Neoplasms	565
C0022665	Kidney Neoplasms	564
C0026975	Myelitis	544
C0024232	Lymphatic Metastasis	538
C0919267	Ovarian Neoplasms	537
C0025202	Melanoma	511
C0023418	Leukemia	468
C0024299	Lymphoma	439
C0037286	Skin Neoplasms	362
C0023449	Leukemia, Lymphocytic, Acute	361
C0033578	Prostatic Neoplasms	336
C0007621	Cell Transformation, Neoplastic	327
C0013264	Muscular Dystrophy, Duchenne	307
C0023452	Leukemia, Lymphocytic, Acute, L1	282
C0079487	Helicobacter Infections	278
C0026764	Multiple Myeloma	274
C0025362	Mental Retardation	272
C0031350	Pharyngitis	259
C0007137	Carcinoma, Squamous Cell	236
C0006118	Brain Neoplasms	233
C1261473	Sarcoma	227

Table 2: The top 30 disease-related MeSH headings associated with GEO series. The third column indicates the total number of GEO samples across the GEO series associated with the disease, out of a potential 44,650 GEO samples associated with any PUBMED-relatable GEO series.

were studied in the previous approach and contain PUBMED relations allowing mapping to MeSH headings through the current approach. Text-parsing found 6,254 mappings from these 284 GEO series to UMLS concepts, while use of the PUBMED relations provided 4,332 mappings to MeSH headings. Only 425 mappings were in common between the two approaches. If we view the MeSH headings as a gold-standard, we find that our previously described text-parsing approach was only 10% sensitive in finding the MeSH headings as mapped concepts.

Discussion

International repositories for genomic data are rapidly approaching nearly 100,000 publicly-available microarray samples as they triple in size yearly. Unfortunately, the crucial details regarding experimental processes and conditions studied in these experiments are currently represented in free-text. It is only through these annotative details that translational links between genomic measurements and disease can be made. Here, we describe the first use of an important translational "short-cut," by representing each genomic experiment in terms of the MeSH headings used to annotate the paper in which each genomic experiment was published.

Through our method, we find 35% of genomic experiments can be mapped to over 270 diseases and conditions. Most of these diseases involve cancers, but others include many other common disorders, including obesity, type 2 diabetes mellitus, HIV infection, malaria, and other common infectious diseases.

We have previously described some success in using text-parsing to acquire the context of these genomic experiments. With only 10% overlap, we find that the MeSH headings mapped to genomic experiments through PUBMED associations are essentially different than the ones found by parsing the annotations of genomic experiments. Though we acknowledge our previous efforts were not tuned for the annotations of genomic experiments, we are finding a text-parsing rate in this domain similar to, if not worse than, those seen in attempts by others to find MeSH headings from MEDLINE abstracts.(16) Because of this success (or failure) rate, we feel that obtaining the annotations of genomic experiments though their associated MEDLINE records will become the premier manner in which to determine the context of these experiments in an automated method.

Future Directions

Only half of GEO experiments are annotated with a PUBMED identifier. An important next step is to use machine-learning methods to predict which of those unannotated experiments might represent a disease, given learned characteristics of disease-related experiments similar to the ones described here.

With additional verification, the MeSH headings mapped to GEO series might be usable as a gold-standard against which future text-parsers applied to GEO annotations can be tested.

Finally, the critical next step is to relate these diseases to the actual genes significantly differentially expressed in the associated experiments. While representing 270 diseases by their genome-wide changes in gene expression will bring us one step closer to a objective definition of disease, it will certainly be the next important step in bridging the world of nucleotides, transcripts and expression with the afflications of us all.

Acknowledgements

The work was supported by grants from the National Library of Medicine (K22 LM008261), the Lucille Packard Foundation for Children's Health, and the Pharmaceutical Research and Manufacturers of America Foundation.

References

1. Snyderman R. The clinical researcher--an "emerging" species. Jama. 2004 Feb 18;291(7):882-3.
2. Lenfant C. Shattuck lecture--clinical research to clinical practice--lost in translation? N Engl J Med. 2003 Aug 28;349(9):868-74.
3. Rees J. Complex disease and the new clinical sciences. Science. 2002 Apr 26;296(5568):698-700.
4. Schwartz K, Vilquin JT. Building the translational highway: toward new partnerships between academia and the private sector. Nat Med. 2003 May;9(5):493-5.
5. Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, et al. Accessing genetic information with high-density DNA arrays. Science. 1996;274(5287):610-4.
6. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet. 1996;14(4):457-60.
7. Carson JP, Zhang N, Frampton GM, Gerry NP, Lenburg ME, Christman MF. Pharmacogenomic identification of targets for adjuvant therapy with the topoisomerase poison camptothecin. Cancer Res. 2004 Mar 15;64(6):2096-104.

8. Zhukov TA, Johanson RA, Cantor AB, Clark RA, Tockman MS. Discovery of distinct protein profiles specific for lung tumors and pre-malignant lung lesions by SELDI mass spectrometry. Lung Cancer. 2003 Jun;40(3):267-79.

9. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A. 2001 Nov 20;98(24):13790-5.

10. Yanagisawa K, Shyr Y, Xu BJ, Massion PP, Larsen PH, White BC, et al. Proteomic patterns of tumour subsets in non-small-cell lung cancer. Lancet. 2003 Aug 9;362(9382):433-9.

11. Microarray standards at last. Nature. 2002 Sep 26;419(6905):323.

12. Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, et al. Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res. 2004 Jan 1;32 Database issue:D35-40.

13. Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. Nat Biotechnol. 2006 Jan;24(1):55-62.

14. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32 Database issue:D267-70.

15. Sayers E, Wheeler D. Building Customized Data Pipelines Using the Entrez Programming Utilities (eUtils). NCBI Short Courses. Bethesda, MD: National Library of Medicine; 2004.

16. Srinivasan S, Rindflesch TC, Hole WT, Aronson AR, Mork JG. Finding UMLS Metathesaurus concepts in MEDLINE. Proc AMIA Symp. 2002:727-31.