Nature Precedings : doi:10.1038/npre.2009.3569.1 : Posted 6 Aug 2009

Annotation-based Meta-analysis of Microarray Experiments

Jie Zheng, Junmin Liu, Elisabetta Manduchi, Christian J. Stoeckert, Jr.
Center for Bioinformatics, Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, USA

Provided by Nature Precedings

CORE

ABSTRACT

We are developing software applications to perform meta-analysis of microarray experiments based on standardized experiment annotations aiming to identify similar experiments and cluster experiments. The applications were tested on files obtained from the ArrayExpress public repository. Annotation terms covering the biological intent and context of experiments were used to compute dissimilarity measures between experiments. Our applications will categorize the experiments resulting from keyword search based on experimental annotation information. These applications may motivate efforts of bench biologists to better annotate experiments.

Meta-analysis software applications contain three components. The first software component extracts annotations from appropriate fields in MAGE-TAB files. The second module computes dissimilarity measures between pairs of experiments using extracted annotations. Given a set of keywords, the third software module will categorize the experiments containing given keywords using the dissimilarity matrix generated by the second software module.

Our goal is to develop software applications that can identify similar experiments to a query experiment and group the experiments resulting from keyword search into various categories based on experimental annotations.

Methods

- Source Of Experiments

  Experiments in ArrayExpress[1] which are in MAGE-TAB format[2] and typically contain some annotations using the MGED ontology (MO)[3]

- Annotation Components Used And Scores
  - **StudyName** (free-text)
  - **ExperimentDesignType** (MO terms) (scored 1 if annotated, otherwise scored 0)
  - **ExperimentFactorType** (MO terms) (scored 1 if annotated, otherwise scored 0)
  - **ExperimentFactorValue** (free-text or measurement or ontology terms)
  - **Organism** (ontology terms) (scored 1 if annotated, otherwise scored 0)
  - **BiomaterialCharacteristics** (types are ontology terms, values are free-text or ontology terms) (scored as # of characteristic types annotated)
  - **ProtocolType** (MO terms) (score 1 if annotated, otherwise scored 0)

  How well an experiment annotated was evaluated by the sum of the scores of the various annotation components. The higher scores indicate richer annotation.

- Dissimilarity Measures Using Jaccard Or Kulczynski Distance

  For each annotation component, let A and B be the sets of annotation terms for the two experiments, respectively.

  Jaccard Distance = 1 - |A ∩ B| / |A U B|

  Kulczynski Distance = 1-1/2 (|A ∩ B| /A + |A ∩ B| /B)

  The dissimilarity between two experiments was defined as the weighted average of their component-wise distances.

- Experiments Clustering

  Based on dissimilarity matrix, automatically generate a hierarchical tree.
  - Generate clusters of varying sizes *n*
  - Assess clusters using silhouette scores and select the best cluster size

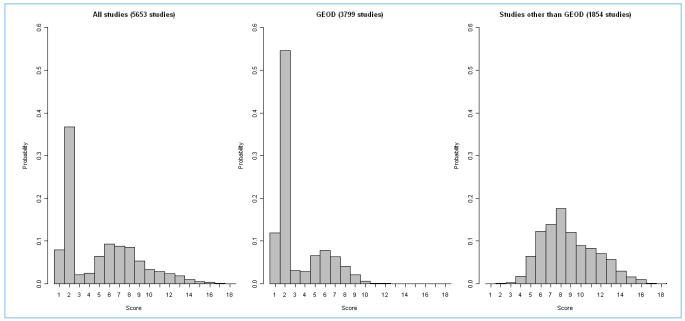
- Gold Standards For Meta-analysis
  - A list of experiments about glucose responsive genes and insulin secretion in islets
  - A list of experiments about genes involved in organismal lifespan alteration generated based on keyword searches in ArrayExpress

- Term Harmonization

  Annotation terms were harmonized using Unified Medical Language System (UMLS)

Annotation Information Of Experiments

The first software module to retrieve annotations from MAGE-ML or MAGE-TAB files has been developed. Experimental annotations were successfully extracted from 5632 experiments. According to the scores of annotation components, around 50% of the experiments were not well annotated (scores ≤ 2), most coming from GEOD. We focused on a total of 2435 experiments with all considered components annotated in the following meta-analyses.



Experimental Annotation-based Clusters (2)

Cluster ③

ExperimentDesign
genetic_modification
_design

ExperimentFactor
genetic_modification

BioCharacteristics
GeneticModification
Age, OrganismPart
Taxon
Mus musculus
ProtocolType
labeling, purify
nucleic_acid_extraction
linear_amplification

8 Experiments

Experiments:

E-CBIL-13	Beta cell specific ablation of Foxa2 (HNF-3b) in mice (glucose homeostasis)
E-CBIL-6	HNF4alpha in beta cell function (insulin secretion and glucose tolerance)
E-CBIL-37	foxA1 and beta cell function (glucose homeostasis)
E-CBIL-39	Foxa2 controls vesicle docking and insulin secretion in mature beta-cells
E-CBIL-21	Beta Cell Growth in Tcf-1 Deficient Mice (pancreatic insulin content)
E-CBIL-23	Fat and Normal adipocytes from insulin receptor knockout mice sorted into small and large cells (glucose intolerance)
E-CBIL-19	Skeletal Muscle Insulin Receptor Knockout - Control, Streptozotocin Diabetic and Insulin Treated
E-CBIL-35	GSIS Study of Rat INS1 cell lines (glucose responsiveness)

*: Experimental IDs in red are the experiments in the list of gold standard (total 8 experiments).

- Most of the experiments in the gold standard list were categorized in one group for two different sets of keyword searches.
- We obtained similar results by using either the Jaccard or the Kulczynski distance to compute dissimilarity measures between pairs of experiments.

Experimental Annotation-based Clusters (1)

Two sets of keywords searches were performed
- keywords: insulin or glucose found in 76 experiments
- keywords: aging, longevity, or lifespan found in 88 experiments

Clustering of 76 Experiments (containing "insulin" or "glucose")

①	②	③	④	⑤
ExperimentDesign co-expression_design **ExperimentFactor** genotype **BioCharacteristics** Genotype **Taxon** *Mus musculus* *S. cerevisiae* **ProtocolType** labeling, purify nucleic_acid_extraction	**ExperimentDesign** compound_treatment _design **ExperimentFactor** compound **BioCharacteristics** CellType, CellLine OrganismPart **Taxon** *Mus musculus* *Homo sapiens* **ProtocolType** labeling, purify nucleic_acid_extraction	**ExperimentDesign** genetic_modification _design **ExperimentFactor** genetic_modification **BioCharacteristics** GeneticModification Age, OrganismPart **Taxon** *Mus musculus* **ProtocolType** labeling, purify nucleic_acid_extraction linear_amplification	**ExperimentDesign** strain_or_line_design **ExperimentFactor** strain_or_line **BioCharacteristics:** StrainOrLine **Taxon** no common term **ProtocolType** labeling, purify nucleic_acid_extraction	**ExperimentDesign** co-expression_design growth_condition_design disease_state_design **ExperimentFactor** growth_condition disease_state **BioCharacteristics** StrainOrLine,DiseaseState Age, OrganismPart **Taxon** *Mus musculus* *Homo sapiens* **ProtocolType** labeling, purify nucleic_acid_extraction
16 Experiments	26 Experiments	8 Experiments	5 Experiments	21 Experiments

Conclusion and Future Directions

- Meta-analyses based on annotation can help to identify closely related experiments. The richly annotated experiments gave better results.
- To further improve the comparison of annotation terms, we will apply ontological relationships to refine the dissimilarity measures.
- In the future, we will expand keyword search results based on dissimilarity matrix if necessary.

Funding: NHGRI grant R21 HG004521

References:

1. Parkinson *et al.*, Nucleic Acids Res. 2007;35 (Database issue):D747.
2. Rayner *et al.*, BMC Bioinformatics 7:489.
3. Whetzel *et al.*, Bioinformatics. 2006; 22(7): 866.

Metadata, citation and similar papers at core.ac.uk