Clustering and Diagnosing Patients with Rare Genetic Disorders

Jialin Song and Jonathan Zung
Computational Biology Lab, University of Toronto

Introduction

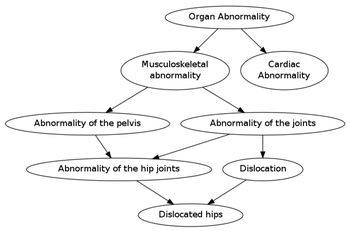
Rare genetic disorders are caused by abnormalities in the human genome. Due to different levels of gene expression and influences from the environment, even patients with a same underlying disorder may exhibit varying symptoms, which makes accurate diagnosis challenging. By taking into account statistics on the prevalence of different phenotypes as well as the relationships between phenotypes provided by ontologies, we can design more informed algorithms for analyzing patient data.

Objectives

- **Diagnose patients**: We want to help clinicians find the correct disorder from among thousands of possibilities.
- **Cluster patients with similar phenotypes**: We want to present clinicians with similar cases so that knowledge about treatment and underlying mechanisms may be shared between them.

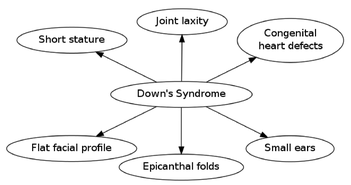
Resources

HPO



The Human Phenotype Ontology (HPO) organizes standardized terms describing abnormal human phenotypes as a directed acyclic graph. More specific terms are children of more general terms. For example, "dislocation" is a child of "abnormality of the joints".

OMIM



Online Mendelian Inheritance in Man (OMIM) is a large database of human genetic disorders. It associates each disorder with a list of standard human phenotypes and the probabilities with which they occur in patients having the disorder. For example, it indicates that a patient with Down's syndrome has a probability of 0.5 of having a heart defect.

PhenoTips

PHENOTIPS

PhenoTips is a web application which helps clinicians to record patients' phenotypic profiles using standard HPO terms. Standard terminology eliminates ambiguities in symptom descriptions.

Previous Work

Phenomizer (Robinson et al., 2008 *AM J HUM GENET 83*, 610 - 615)
Phenomizer computes the similarity between two sets of phenotypes S and T.
Let $p(t)$ be the probability of a phenotype t in the general population.
The similarity between two phenotypes s and t is defined as

$$sim(s,t) = \max_{a \text{ a common ancestor of } s \text{ and } t} -\log p(a)$$

The similarity between two sets of phenotypes S and T is defined by averaging certain pairwise similarities:

$$sim(S,T) = avg[\sum_{s \in S} \max_{t \in T} sim(s,t)] + avg[\sum_{t \in T} \max_{s \in S} sim(t,s)]$$

This method can be applied both to measure the similarity between a pair of patients and between a patient and a disorder.

Clustering Methods

Patient-Patient similarity metrics

We can define a patient-patient similarity metric, and then apply a standard clustering algorithm (in our case, spectral clustering).
Examples of similarity metrics:

- Euclidean Metric: Count the number of phenotypes shared between two patients.
- Information Metric: Here we count shared phenotypes weighted by their log probabilities in the general population, so that patients sharing rarer phenotypes are more similar. The similarities are normalized by the analagous weighted count of the total number of phenotypes for the two patients.

Mixture Models

First we formulate a probabilistic model for patient phenotypes with a single latent variable d representing the underlying disorder. We can use the EM algorithm to fit the model, and then cluster by the inferred values of d.
Examples of mixture models:

- Independent Mixture: Phenotypes are independent given d.
- Conditional Mixture: Phenotypes are conditionally independent given d and their ancestors in HPO. In order to have a phenotype, a patient must also have its ancestors in HPO.

Diagnosis Methods

Naive Bayes

- For a patient p, we want to find the disorder d with the highest conditional probability $Prob(d \mid p)$.
- By Bayes' Theorem,

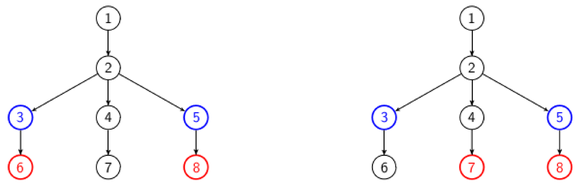
$$Prob(d \mid p) = \frac{Prob(p \mid d) \times Prob(d)}{Prob(p)} \propto Prob(p \mid d) \times Prob(d)$$

- Naive Bayes assumes that each phenotype is independent once a disorder is given

$$Prob(d \mid p) \propto Prob(d) \times \prod_i Prob(phenotype_i \mid d)$$

Phenotype Matching

- The objective is to measure how close a patient is to the canonical form of a disorder.
- For each of a patient's phenotypes, compute the closest distance to each phenotype annotation of a disorder. Construct a distance matrix from the calculated data.
- Use the cost matrix to determine the matching between patient's phenotypes and disorder annotations that minimizes the total distance.
- The minimimum distance required to transform from the real patient to the canonical form of a disorder is used as the measure for closeness.



A Toy Example

A patient's phenotypes are marked in blue and the annotations of a disorder are marked in red. In the example on the right, the real patient can transform into the canonical form of a disorder via $3 \to 2 \to 4 \to 7$ and $5 \to 8$, resulting in a total cost of 4. In the left example, we can achieve a total cost of 2 by $3 \to 6$ and $5 \to 8$.

Datasets

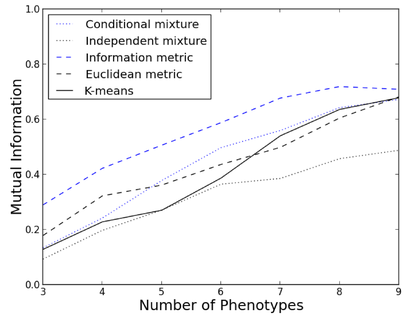
25 real patients from published studies

- 13 with Floating Harbor syndrome
- 6 with Rubinstein-Taybi syndrome
- 6 with Opitz-Kaveggia syndrome

1498 simulated patients

- A patient is simulated by selecting a disorder at random, and then assigning phenotypes to the patient according to the probabilities recorded on OMIM. Noise is then added by appending some of the 100 most common phenotypes and then performing a random walk on the HPO graph.

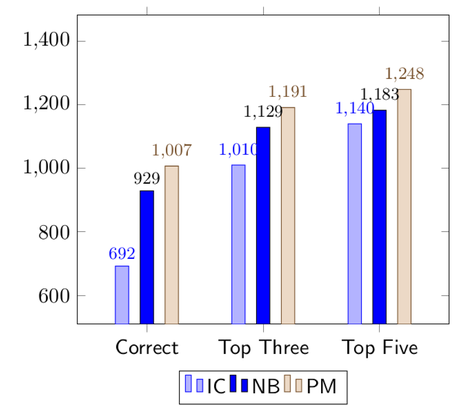
Clustering Results



Clustering Results for 25 Real Patients

The figure above compares clustering results for five algorithms on the 25 real patients. To make the clustering problem more difficult, the algorithms were only provided with a random subset of between 3 and 9 of each patient's phenotypes. The vertical axis shows the adjusted mutual information of the computed clustering with the true clustering.
The information metric outperformed the Euclidean metric by giving rare symptoms more weight. The conditional mixture model outperformed the independent mixture model by modelling the restriction that a phenotype may only be present when its ancestors are.

Diagnosis Results



Diagnosis Results for 1498 Simulated Patients

Naive Bayes made nearly 35% more correct diagnoses than Phenomizer by taking into account the incidence rates of abnormal phenotypes, while phenotype matching outperformed naive Bayes by utilizing structural information in HPO.