genomic markers
delivered

Multi-objective optimization for concurrent mining of disparate genomic datasets

Cole Harris & Peter Hraber

Exagen Diagnostics, Inc.

Albuquerque, New Mexico

Introduction - Background

Exagen Diagnostics develops medical tests for patient prognosis and drug response from discriminative patterns discovered in genomic data.

The human genome consists of approximately 30,000 genes. With currently available technology, each of these genes, or the products of these genes, can be measured simultaneously.

However it is difficult to acquire human samples to perform such experiments.

Introduction - Data characteristics

A typical experiment might obtain genome-wide data on the order of 10-100 patient samples. Thus, a typical dataset will consist of

~30,000 features (genes) X ~100 examples (patients)

"The curse of dimensionality"

Introduction - Practical tests

Low-dimensional tests

Currently, all FDA approved tests are low-dimensional. The majority of tests measure a single analyte, but more complex tests are under development by us and others.

However, as the number of distinct measurements increases, so does the cost and complexity of the test, and thus marketplace resistance.

Exagen is developing tests that consist of a small number (typically 3-5) of measurements.

Introduction - Feature Selection

Even for a small number of discriminative features (3-5), when selected from a very large pool of potential features (~30,000), there are an extremely large number of potential combinations.

We employ a genetic algorithm for feature selection, coupled with objective functions derived from low-complexity classification/regression algorithms, to efficiently search this space.

Single objective function

To data, several groups have developed techniques for identifying discriminative patterns in single genome-wide datasets. Often the application of these methods will either not find any patterns, or more often identify chance patterns.

Approaches to combat this problem:

- Evaluate the statistical significance of the identified patterns
- Split data into training and test sets
- Acquire additional data

Significance testing

We employ a class label/response permutation-based approach, using the objective function value for the putative pattern as the test statistic. First, the class labels (response values) are randomly permutated, and second, this modified data is searched for discriminative patterns. This process is repeated many (~100) times. The result is an estimate of the null distribution for the test statistic, from which the null hypothesis

H_0: There is no association between the feature combinations and class label (response)

may be tested.

Training/Test split

Identify patterns with only training data, and then test patterns on test data.

Often this approach is of limited utility, as the number of samples in a typical dataset is small. In the worst case, any split of the data may either hinder the ability to find true patterns, result in insufficient power to evaluate patterns in test data, or both.

More Data?

Ideally, additional data could be acquired. In practice this is difficult, as the experiments are expensive, and biological samples may not be available.

Alternative to More Data

Often multiple datasets are available that

- are collected from ‘similar’ biological systems:
  - Tumors
  - ‘treated’ cell lines
  - Man/mouse
- consist of measurements of ‘similar’ biological quantities:
  - DNA content
  - gene expression
  - protein expression.

The goal: combine similar datasets.

Combining datasets

Other groups have developed methods for directly combining datasets via a mapping approach. This is somewhat similar to the normalization techniques used to minimize sample variability within a dataset, and may be too simplistic for combining datasets collected using different instrumentation, or datasets collected from different types of samples, or under different conditions.

Combining datasets

In our approach, the datasets are not modified. The only requirement for combining datasets is that the datasets share ‘similar’ mappable features.

Example 1

- Dataset 1: gene expression for genes A,B,C,…
- Dataset 2: protein expression for genes A,B,C,…

Example 2

- Dataset 1: gene expression collected using Affymetrix platform
- Dataset 2: gene expression collected using cDNA platform

Example 3

- Dataset 1: gene expression for cell lines treated with drug A
- Dataset 2: gene expression for cell lines treated with drug B

Scoring patterns in multiple datasets

Given a set of features, and a relationship between these features and the class labels (response), various single dataset objective functions can be defined:

- Accuracy
- Maximum likelihood
- AIC, BIC, etc.

A multiple dataset objective function can be defined from these in various ways, and work in this area continues. Thus far we have investigated arithmetic and geometric means of accuracy, and likelihood sum.

Example analysis:

Taxane response in tumor cell lines

Taxanes are a class of chemotherapeutic agents used in the treatment of a broad range of cancers. However individual taxanes may differ in their efficacy in a particular cancer. The goal of this analysis is to identify patterns in gene expression data that are predictive of taxane response across the spectrum of cancers. In this example, we seek to find linear predictors of response.

Example analysis:

Taxane response in tumor cell lines

Available gene expression data:

Staunton, et. al. (affymetrix)	6817 probes X 60 cell lines
Scherf, et. al. (cDNA)	9703 probes X 60 cell lines

Drug response data from NCI:

taxotere (docetaxel)	GI50 X 60 cell lines
taxol (paclitaxel)	GI50 X 60 cell lines

Example analysis:

Taxane response in tumor cell lines

Preprocessing steps

- •QC, normalize, log transform individual datasets
- •Identify common set of features between datasets (nontrivial!)
  - •2344 common genes
- •Identify common cell lines between GI50, gene expression
  - •taxol: 58 cell lines
  - •taxotere: 57 cell lines

Example analysis: Taxane response in tumor cell lines

Multi-objective function definition

GI50 response	Gene expression 2344 genes
taxol 58 cell lines	Affymetrix
taxol 57 cell lines	cDNA
taxotere 57 cell lines	Affymetrix
taxotere 57 cell lines	cDNA

For a particular set of features: x, y, z

Maximum likelihood

→ a

→ b

→ c

→ + d

score for (x,y,z)

Example analysis: Taxane response in tumor cell lines

Genetic algorithm search (plain vanilla)

- Representation: integer (1:2344)
- 'Elite' proportion: 50%
- Population size: 5000
- 20 iterations
- Dimensionality: 3 features
- Objective: linear regression ML

Example analysis: Taxane response in tumor cell lines

Preliminary results

- In the highest ranked solutions, several genes appearing frequently in different combinations may be relevant to taxane response, in that they have been associated with the cytoskeleton – the target of taxanes – or that they have been associated with tumor biology.
- In the highest ranked solutions, the coefficients defining the predictors for the individual datasets, while highly variable across datasets, are sign-consistent across datasets.

Example analysis: Taxane response in tumor cell lines

Pending internal review, additional results will be presented at the workshop.

Further research

- Investigate alternative multi-objective function definitions
- Investigate multi-objective functions for unlabeled data
- Examine sensitivity of results to GA parameters for concurrent mining.