Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Integrative Gene Selection

Final Presentation

Outline

- Recap of what was said in the intermediate presentation
- Clearly state the problem and its significance
- Explain my approach in technical detail
- How will I evaluate this approach?
- What are my expected results?
- Short outlook
- Q&A

Context of this work

Analysis and classification of diseases on the RNA-Level is an important topic

- come up with new drugs, disease diagnosis …
- identify gene functions, collaborations of genes on modular level, pathways …

Work based on paper: “Towards precise classification of cancers based on robust gene functional expression profiles” [Guo et al, 2004]

⇒ Approach from broad field of feature selection

Feature Selection

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

What: select subset of features from high-dimensional feature space for analysis, classification or similar …

Why: avoid curse of dimensionality, overfitting, noise

How: statistical analysis on dataset, select relevant subset of features

Feature Selection

- **Short, fat data (p >> n)**
  RNAseq data covers p-thousands of genes but only n-hundred samples
- **Gene interactions**
  Genes react to changes of partners in interaction network, interactions have to be taken into account
- **Biological Relevance**
  Statistical approaches concentrate on statistical importance, ignoring biological relevance of genes, like disease driver genes; only showing increased expression profiles of affected pathway genes

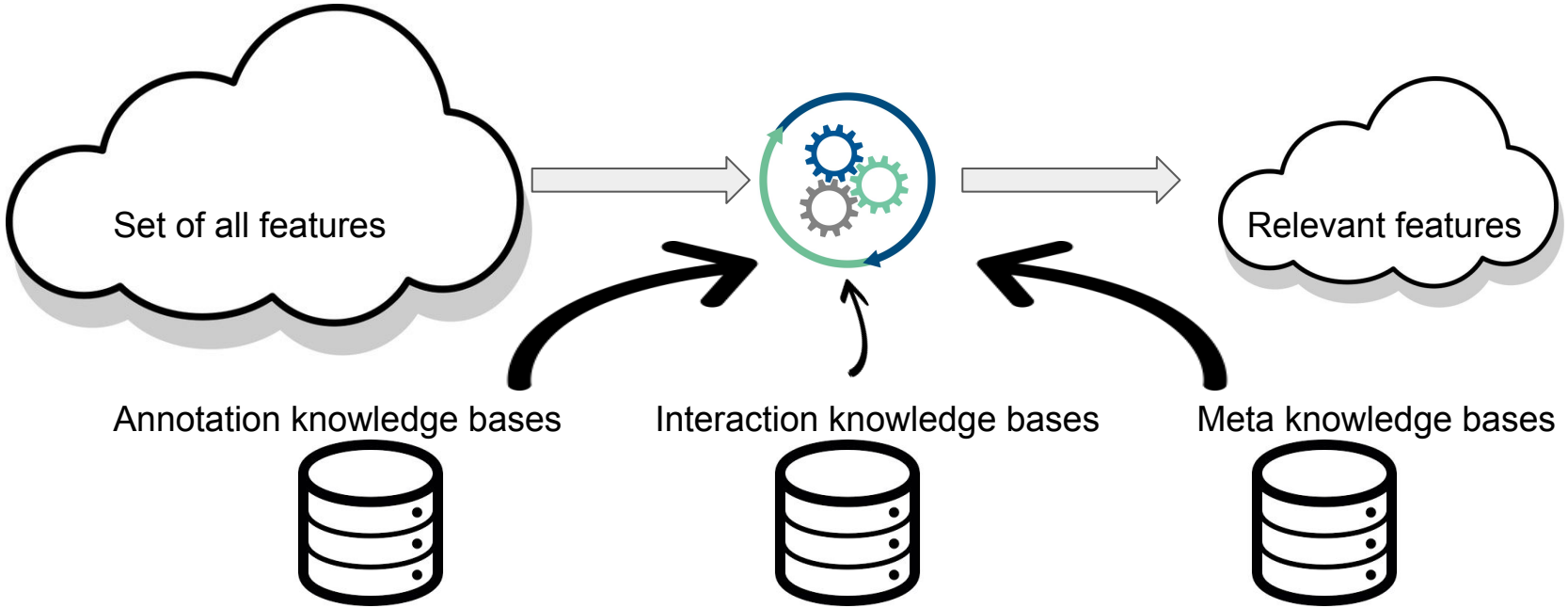
Main problem: robustness!

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Plain statistical approaches are not sufficient enough?

Integrative Gene Selection

Gene selection **not only statistical but also external knowledge**!



Integrative Gene Selection

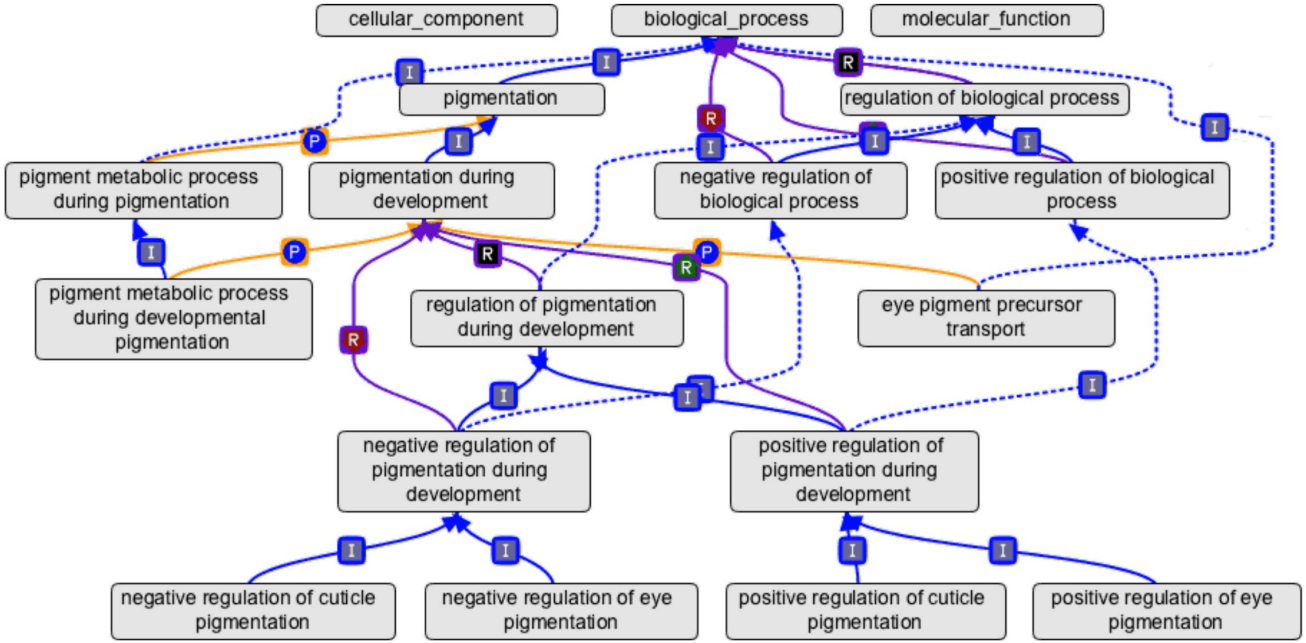
- What: feature selection integrating external knowledge
  - Accounting information about gene modules and their activities
  - Called functional expression profiles in the paper

“Towards precise classification of cancers based on robust gene functional expression profiles” [Guo et al, 2004]

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Main Section

Modules, Ontologies and Network



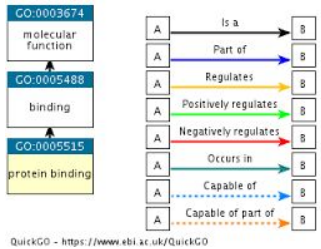
- Network-approach: integrate knowledge about pathways and interactions

Modules, Ontologies and Network

- Problems…
  - Gene expression profiles (GEP) likely to change
    - Differences and error rates in sequencing and alignment processes
  - GEP correlate by chance but low probability that modules correlate by chance
- … less relevant by using networks instead of raw-counts

- Genes work on activities together ⇒ feature space drastically reduced using modules instead of genes

- Guo et al mapped genes to functional GO-categories

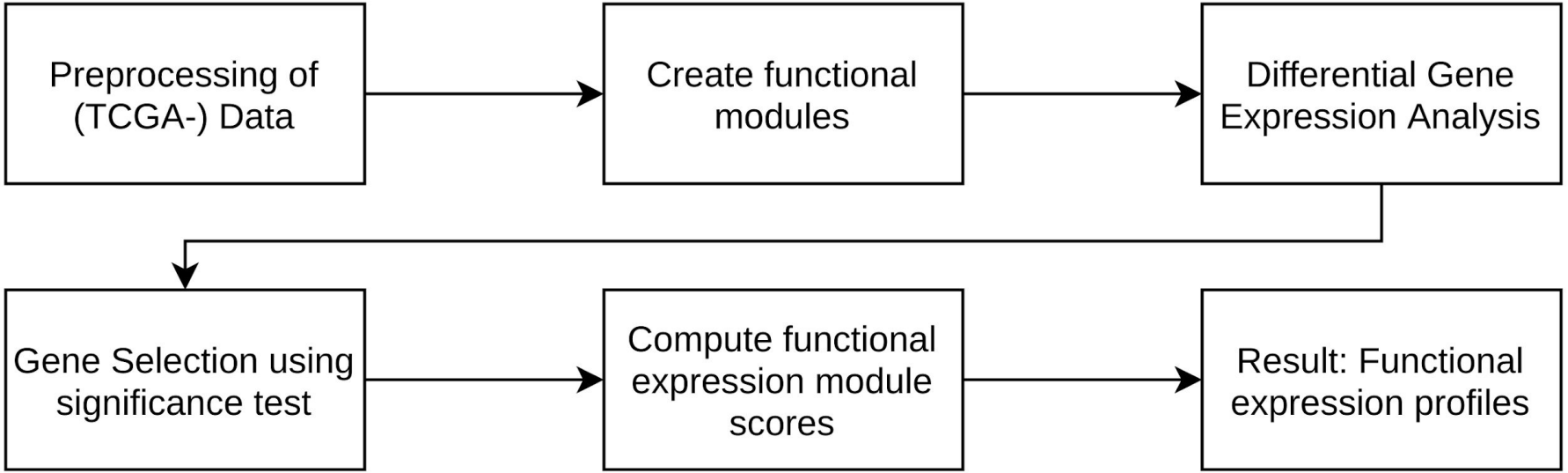


Main Section Overview

- Approach-Speciality: Compute summary measure(s) for the normalized expression values of the annotated genes
  ⇒ Capture the overall activity level of the module

- Implications
  - Feature space becomes reduced
  - Disease-driving modules can be identified
  - Mapping genes to modules is m:1, or m:n, mapping
    ⇒ Some uncertainty about which specific genes led to some disease

Approach in detail from a technical view

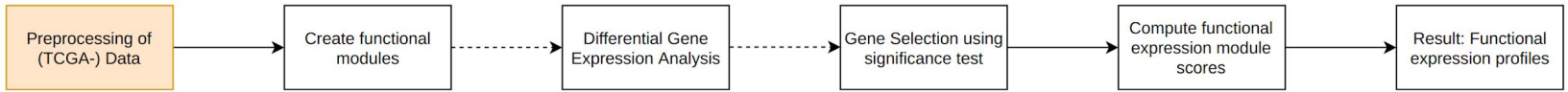


Not part of the approach, but necessary:

Evaluation: Apply proper classification algorithm (e.g. C4.5), train it and evaluate!

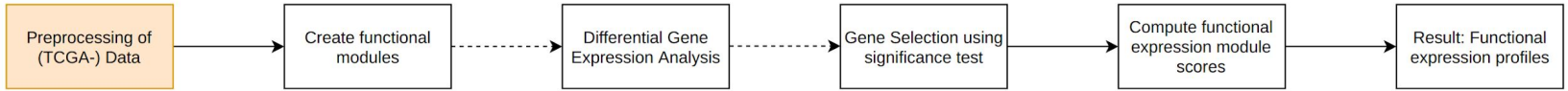
Normalization

- Changing values without losing information in the range of values
  ⇒ Tackle the problem of hard to combine raw count values differing drastically in expressed amount
- Log-transformation, widely used in biomedical research for skewed data
- x = (a + (x - A) * (b - a)) / (B - A)
- Reason: Some methods, like the t-Test, assume normal population distributions and not skewed distributions



Preprocessing of *(TCGA-)*Data

- What: Genes have varying IDs
  - TCGA-Data: ENSG IDs as columns (features), labeled samples (cancer subtypes) as rows
  - GO has its own IDs
- Why: ENSG-IDs can not be directly interpreted by GO
  - To use the same approach as the authors of the paper: map ENSG-IDs to GO IDs
- How:
  - Combine Mappings from ENSG-IDs to GO categories:
    - Biological Processes, Functional Modules and Cellular Components
  - Remove duplicates

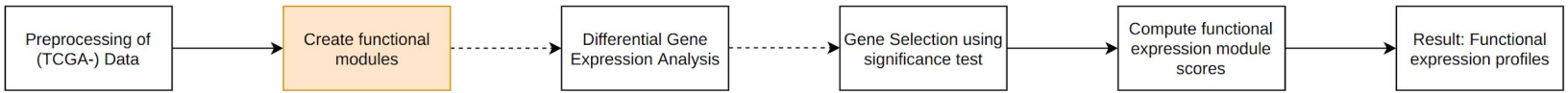


Creation of functional expression modules

- Start with list of gene modules obtained by previous step

- Implied by the ontology structure: one gene annotated to a category is also within the ancestor categories on the same path
  - Therefore, if all descendants represented, remove ancestor module from feature space

- To use not all modules left, perform dgea and significance test to further reduce the feature space

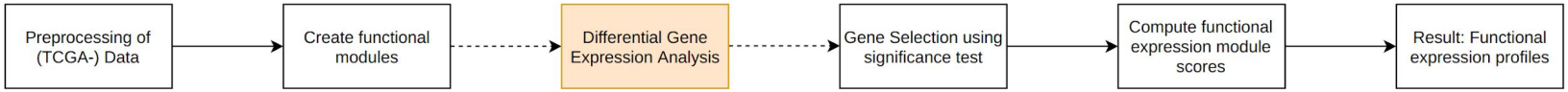


Differential Gene Expression Analysis

- What: Search for differentially expressed genes

- Why: Interested in differences normal tissues vs diseased tissues, not raw count

- How: Standard procedure, like parametric two-sample t-Test
  - or Welch t-Test, assuming the data to be tested is normally distributed
  - or perform permutation t-Test or Mann-Whitney U (Wilcoxon rank sum) test; no assumptions



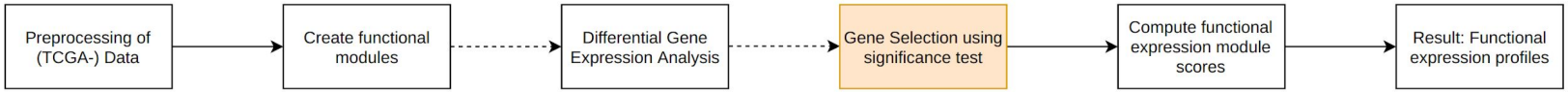
Significance Test

- Significant number of differentially expressed genes compared to random
  - Standard procedure

- Abbreviations:
  - N = total set of genes annotated in GO
  - C = total set of genes differentially expressed; subset of N
  - Testing for module X,
    - n genes out of N
    - k genes out of C

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{C}{i}\binom{N-C}{n-i}}{\binom{N}{n}}$$

- Sampling without replacement; a gene can be selected only once
- Categories containing significantly ($p \leq 0.05$) larger number kept for analysis

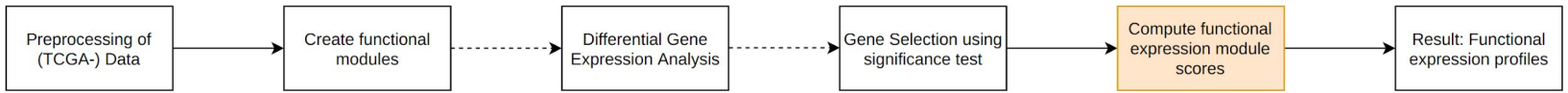


Score-Computation: Mean or Median?

- Computation of “FEP”-scores: Paper stated it made a difference for precision and recall using mean or median: Mean != Median
  - Mean: Average one is used to, add up all numbers and divide by # of numbers
  - Median: ”Middle” value in the sorted list of all values

- The computation of “FEP”-scores in the paper mostly used the median for computation. This is due to the fact that the median is not affected by “outlier” values as strong as the mean is.

- Example: (4,5,1,2,240,3,7,9,15) → Mean: 31,777… ; Median: 5

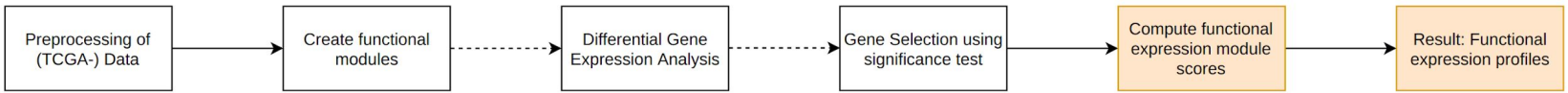


Computation of score for module

- Compute two measurement scores for every module of all expression values
  ⇒ to capture the activity of this module
  - Mean: use of data is gaussian or symmetrically distributed
  - Median: use in case of outliers

- With the dynamic decision on which measurement score to use the approach tries to fit the “real” distribution of raw gene expression values best.

- Save Set of functional expression profiles



Current State

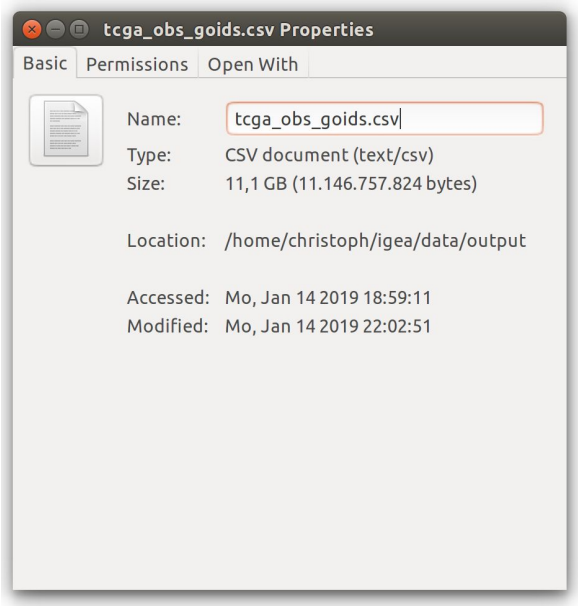
Current state of implementation

- Load already preprocessed TCGA-Dataset
- Map ENSG-IDs to GO-IDs, achieved using R-Package
  - “org.Hs.eg.db” which provides HomoSapiens Gene Identifiers
- Remove duplicate ENSG to GO ID mappings
  - formerly introduced by different GO Ontologies
  - Biological Processes, Functional Modules and Cellular Components

Current state of implementation

Translated every Gene to GO-ID, copying expression profile, save …

~11.1 GB of data… … in contrast to formerly 2.1 GB (TCGA)



Future state of implementation

- Save gene's expression profiles
- Translate every Gene to GO-ID
  - Obtain list of active modules
  - Save the mapping ENSG-to-GO-Module
- Perform selection of only important gene modules
  - Remove modules for which all the child-modules are present
- Perform DGE-Analysis on the remaining genes
  - Using R's package DESeq2

Future state of implementation

- Save list of differentially expressed modules
- Perform significance test
  - Remove modules not significantly differentially expressed
- Obtain list of modules separating different cancer subtypes and healthy tissue
- Integrate into the existing IGEA-Framework

Evaluation

Evaluation

- What: To evaluate findings, test performance on data classifier has not seen but correct class label is known
  - For example with k-fold cross-validation
- Why: Using known data, the class label is likely always correctly classified
  - But we are interested in correctly classifying unseen data!
- Compute evaluation metrics used for comparison in medicine:
  - Accuracy
  - Precision
  - Recall

Evaluation

- Abbreviations:
  - TP = True Positive
  - TN = True Negative
  - FP = False Positive
  - FN = False Negative

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

- The authors of the paper used k-fold cross-validation on four different datasets; k being set to five

- Still to do: Split data into five datasets of approx. same size, train on four subsets, evaluate selection using created decision tree on remaining subset

Evaluation using C4.5

- What: Paper used a decision tree for classification purposes: C4.5
  - A classification decision tree, kind of flowchart, ultimately assigning a label to an input

- Test understanding of diseases and influence of identified modules
- Be able to classify new inputs in order to assist in disease classification

Nice benefit would be: Come up with a decision tree that contains genes in its pathways which are not yet considered to be involved in a disease or a disease subtype

Evaluation - Questions for Experiments

- Is the process robust?
- How are the process measurement scores?
- How does the process perform on other datasets?
- How good is this approach compared to “plain statistical approaches”?

Measurement Scores and Robustness

- Measurement Scores: use (k=5) evaluation data of already preprocessed TCGA-Data to obtain measurement scores

- To check for robustness, take more TCGA-Data
  - Same features (genes)
- Classify new data in batches
  - Using the trained classifier
- If the approach is robust
  ⇒ Expect similar measurement scores for every batch of unseen data

Performance on other datasets

- Search for labeled, publicly available and approved datasets
- Classify new data
  - Using the trained classifier
- Compute the measurement scores
- Comparison
  - Compare achieved scores to TCGA-Dataset (k=5)
  - Compare achieved scores to other approaches used on the same dataset

This will give some hints on robustness once again…

Comparison with statistical approaches

- Run statistical approaches from the IGEA-Framework…
  - VB-FS
  - InfoGain
  - ReliefF
  - SVM-RFE
- … on same TCGA-Dataset used for creation of functional modules

- For each statistical approach:
  - Train the same classifier on output of statistical approach
  - Use (k=5) evaluation data of already preprocessed TCGA-Data to obtain measurement scores

- Compare the measurement scores of statistical approaches and integrative

Expected Results

Expected results

- Find a decision tree classifying disease types from the TCGA-Dataset

- Not expecting to find modules newly separating data; though it might happen
  - Paper studied is from 2004
  - I did not find many papers applying decision trees on functional module-level since then
  - GO grew constantly since 2004

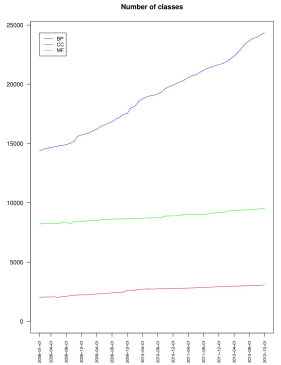
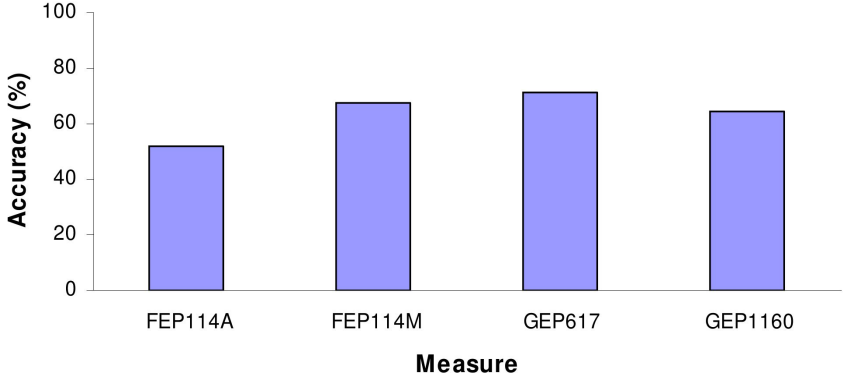
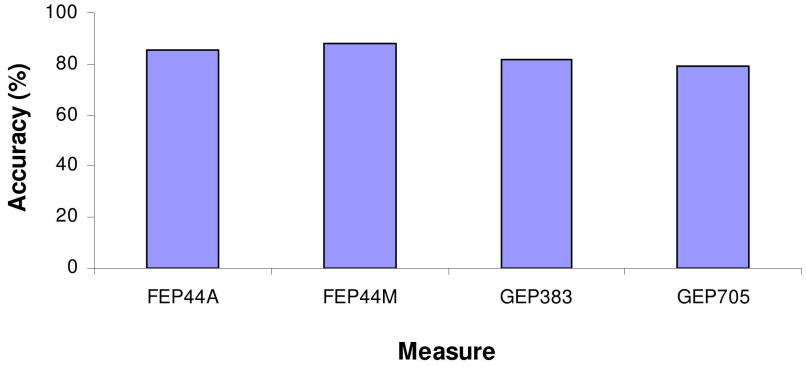


Figure 1. Evolution of the number of classes of the three branches of the Gene Ontology. Biological process (BP), Cellular component (CC) and Molecular function (MF).
doi:10.1371/journal.pone.0079993.g001



Other Cancer-types



Lymphoma

- FEP114A / FEP44A = Functional Expression Profile … Mean
- FEP114M / FEP44M = Functional Expression Profile … Median

- With that said - let us jump right into our short time for Q&A!

Related Work / Interesting Papers read

- Paper von Cindy
- Yuan et al, A Sparse Regulatory Network of Copy-Number Driven Gene Expression Reveals Putative Breast Cancer Oncogenes, 2012
- Park et al, Interaction-Based Feature Selection for Uncovering Cancer Driver Genes Through Copy Number-Driven Expression Level, 2016
- Liu et al, Feature selection of gene expression data for Cancer classification using double RBF-kernels, 2018
- Meng et al, Dimension reduction techniques for the integrative analysis of multi-omics data, 2016
- Huang et al, Systematic Evaluation of Molecular Networks for Discovery of Disease Genes, 2018
- Quanz et al, Biological Pathways as Features for Microarray Data Classification, 2008
- Chen et al, Integrating Biological Knowledge with Gene Expression Profiles for Survival Prediction of Cancer, 2009
- Zhang et al, Module-based breast cancer classification, 2013
- Chuang et al, Network-based classification of breast cancer metastasis, 2007
- Gu et al, Multiclass classification of sarcomas using pathway based feature selection method, 2014
- Guo et al, Towards precise classification of cancers based on robust gene functional expression profiles, 2005
- Payán-Gómez et al, Integrative analysis of global gene expression identifies opposite patterns of reactive astrogliosis in aged human prefrontal cortex, not yet peer-reviewed, 2018
- Khatri P, Draghici S, Ostermeier GC, Krawetz SA: Profiling gene expression using onto-express. Genomics 2002, 79:266-270.
- Al-Shahrour F, Diaz-Uriarte R, Dopazo J: FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 2004, 20:578-580.
- Al-Shahrour F, Diaz-Uriarte R, Dopazo J: FatiGO: Data mining with Gene Ontology. [http://fatigo.bioinfo.cnio.es/].

Image Sources

- Log2-Graph: https://en.wikipedia.org/wiki/Binary_logarithm
- Database: https://www.flaticon.com/free-icon/database_4426
- Cloud: https://all-free-download.com/free-vector/download/nuage-cloud_116075.html
- Process: https://positiveenergy.pro/process/
- Curved Arrow: https://thenounproject.com/term/curved-arrow/69289/
- GO-Structure: http://geneontology.org/page/ontology-structure
- Module_ancestors: https://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0005515
- Distribution: https://commons.wikimedia.org/wiki/File:Normal_Distribution_NIST.gif
- Calculator: https://pixabay.com/de/rechner-zahlen-0-1-2-3-4-5-6-2374442/
- Table: https://commons.wikimedia.org/wiki/File:Table_of_the_Value_and_Weight_of_Coins,_1815.png
- Network: https://commons.wikimedia.org/wiki/File:Ego_network.png
- GO_Growth: https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0075993&type=printable
- Others (Precision, Recall, Accuracy and Decision-Tree: Paper: Towards precise classification of cancers based on robust gene functional expression profiles, Guo et al, 2004)

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

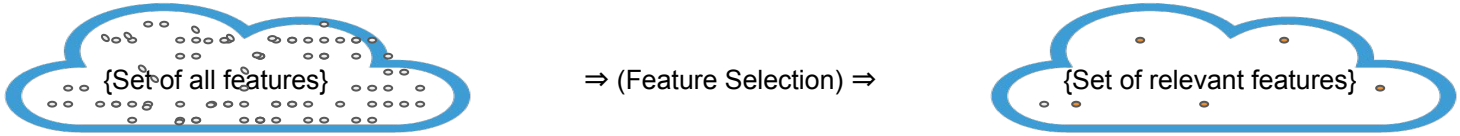
Q&A

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Thank you for your attention!

Backup-Slides

Feature Selection



- **Short, fat data (p >> n)**

  RNAseq data typically covers p-thousands of genes but only few n-hundred samples

- **Incomplete view on cell process**

  RNAseq data covers a complete snapshot of the gene activity of a cell; only a single point in time - not more, not less...

- **Gene interactions**

  Genes react to changes of partners in the interaction network, interactions have to be taken into account

- **Biological Relevance**

  Statistical approaches only concentrate on statistical relevance, ignoring the biological relevance of genes, like disease driver genes, while only showing the respective behaviour of increased gene expression profiles of affected pathway genes

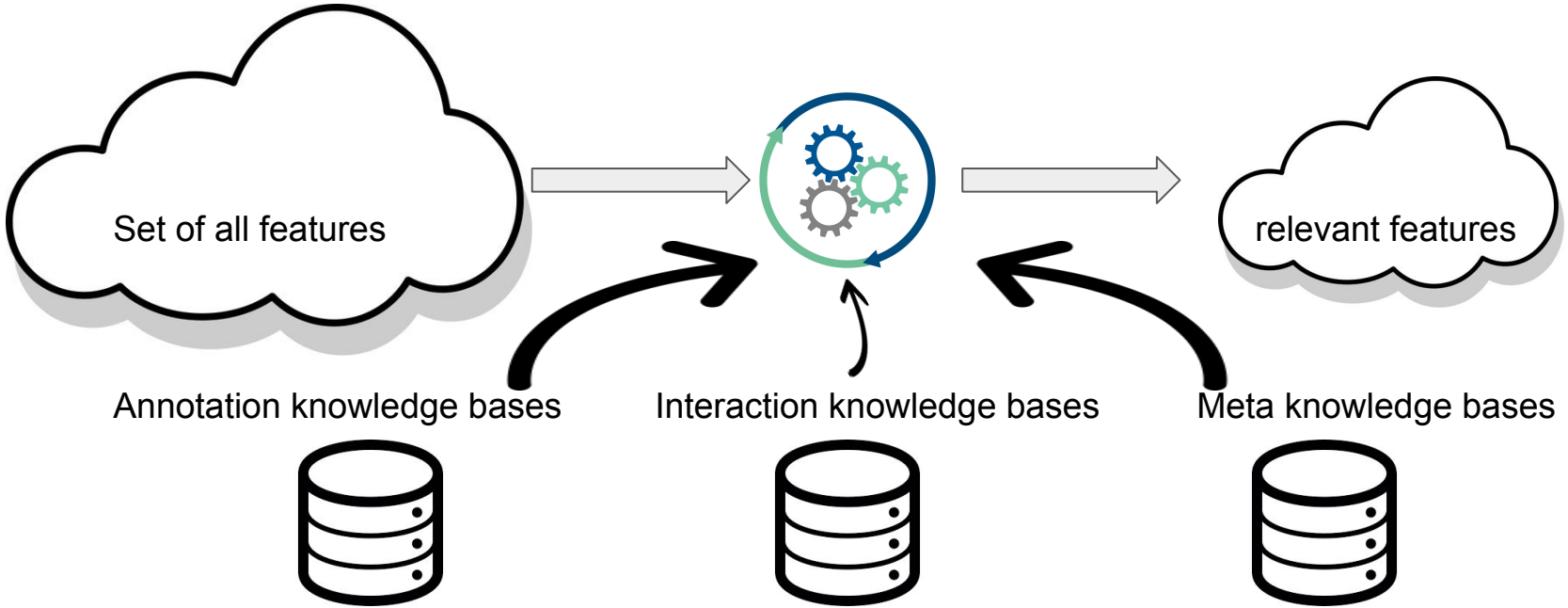
Gene Selection (a.k.a. Feature Selection 4 RNAseq-Data)

- Goals:
  - shorten training times
  - avoid the curse of dimensionality (overfitting)
- Approach:
  - Use statistical models to select interesting “genes” for future analysis
  - Currently use only a bit of existing knowledge of underlying biological processes

Different approaches exist, depending on where and how they integrate external knowledge!

Integrative Gene Selection

Gene Selection using **not only statistical** methods **but also external knowledge bases**!



Problem and its significance

Disease Driver Gene(s) Identification is a hard problem

- Problems stated at feature selection
- Genes can be involved in multiple pathways (**pathway overlaps**)
- **Conventional gap** between static networks and dynamic cell process

We do not know a whole lot about cancer and it’s subtypes, how they begin to develop and what genes are interchangeably active within these activities

Problem and its significance

Robust and efficient methods for analyzing and interpreting high dimensional gene expression data are needed, because, …

Systems used to extract genes are not 100% accurate and thereby result in some variance within the expressed data (e.g. measurement/alignment of RNA is often not absolutely accurate! - refer to DMfDH)

To overcome this very important problem, we can use integrative gene selection to come up with an interesting list of genes to further analyze

We do this in the context of modules, as it can be assumed that genes express and perform their functions in a modular fashion within the cells

Main Section Overview

- Approach: Some knowledge bases, like Gene Ontology (GO), map genes via ID and functionality as well as modules that share some behavioural aspects. One can think of these as a network of genes.
- GO works with its own IDs and three ontologies: cellular component, molecular function and biological process; terms forming an directed acyclic graph.
- A network-approach is an approach that is integrating biological knowledge about pathways and gene interactions. Guo et al mapped Genes to their functional categories defined by GeneOntology (GO).

Main Section Overview

Very likely that gene expression profiles are correlated by chance, simply because of the high amount of genes (features)

By taking the selection process to another level, the modular level, problems stated beforehand vanish, as it can be assumed that genes express and perform their functions in a modular fashion within the cells

It is very unlikely that modules correlate by chance

Plenty of genes work on the same activities together, so the feature space is drastically reduced by using modules instead of using genes directly

Normalization (what)

Normalization is the process of changing the values of numeric columns without losing information or distorting differences in the range of values.

Raw count values, as obtained by RNAseq, are hard to combine, because they may differ drastically in their expressed amount.

Normalization tackles this problem by creating new values that maintain the general distribution and ratio in the source data, but scaling values within a specific scale applied across all numeric columns used.

Normalization (what)

Log-transformation is widely used in biomedical and psychosocial research to deal with skewed data, believed that it decreases the variability of data and making data conform more closely to the normal distribution [“Log-transformation and its implications for data analysis”]

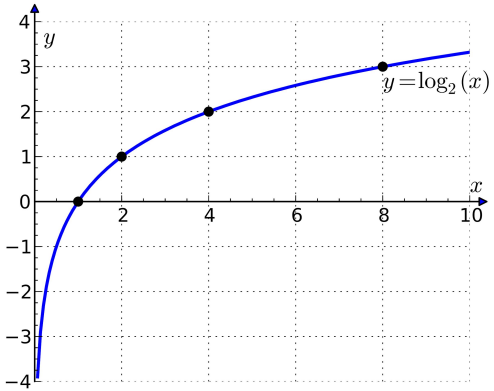
Data used within this project is of level (1,2,3,...) from TCGA, meaning it is, (normalized and log-transformed, normalized, post-normalized, …)??

Normalization (how and why)

Normalization is basically taking the minimum (A) and maximum (B) value of the original data, define minimum (a) and maximum (b) value for normalized data and then computing the normalized values (x) as follows:

$x = (a + (x - A) * (b - a)) / (B - A)$

Log-Normalization is simply taking the logarithm of values



Reason: Some Methods, like the t-Test,
assume normal population distributions and not skewed distributions

Problems to state

- While using integrative gene selection, it might be that we do not take into account small changes at the modular level, because there may be very complex, interchanging module-activities carried out

Differential Gene Expression Analysis (how ctd.)

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Other approaches may be:

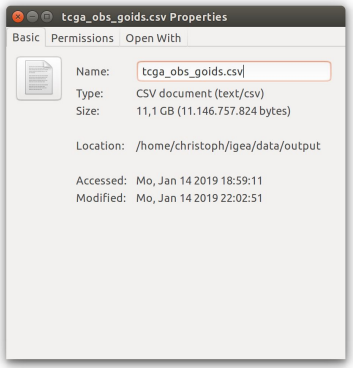
Compare t-value with standard t-table

When testing so many genes for differential expression, one will likely falsely call many genes “differentially expressed” ~ How to tackle this problem?

Strategy: Define a false positive rate - how many of the genes one calls “differentially expressed” are actually not differentially expressed. One method: Benjamini Hochberg Method

Preprocessing of TCGA-Data (how)

- Map ENSG-IDs to GO-IDs because TCGA-data had ENSG-IDs
  - Achieved using R-Package “org.Hs.eg.db” which provides HomoSapiens Gene Identifiers
- Remove duplicate ENSG to GO ID mappings
  - Formerly introduced by different GO Ontologies
- Combine Mappings to GO-IDs from Biological Processes as well as Functional Modules and remove duplicates
- Translate every Gene to GO-ID, while copying its expression profile, save these
  - ~11.1 GB of data…
    … in contrast to
    formerly 2.1 GB (TCGA)



Future state of implementation

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Using R's package DESeq2 calling "DESeq" results in:

Estimate size factors to account for differences in sequencing depth, showing factors of variation within the data, using the "median ratio method" developed by Anders and Huber in 2010

Estimate dispersion of data for each gene, maximizing the Cox Reid-adjusted profile likelihood developed by McCarthy et al in 2012

Tests for significance of coefficients in a Negative Binomial GLM, using previously calculated values on the data

Take a look at result tables, interpreting computed log2 fold changes and p-values

Differential Gene Expression Analysis (how ctd.)

Using R's package DESeq2

Estimate size factors to account for differences in sequencing depth, showing factors of variation within the data, using the "median ratio method" developed by Anders and Huber in 2010

Estimate dispersion of data for each gene, maximizing the Cox Reid-adjusted profile likelihood developed by McCarthy et al in 2012

Tests for significance of coefficients in a Negative Binomial GLM, using previously calculated values on the data

Take a look at result tables, interpreting computed log2 fold changes and p-values

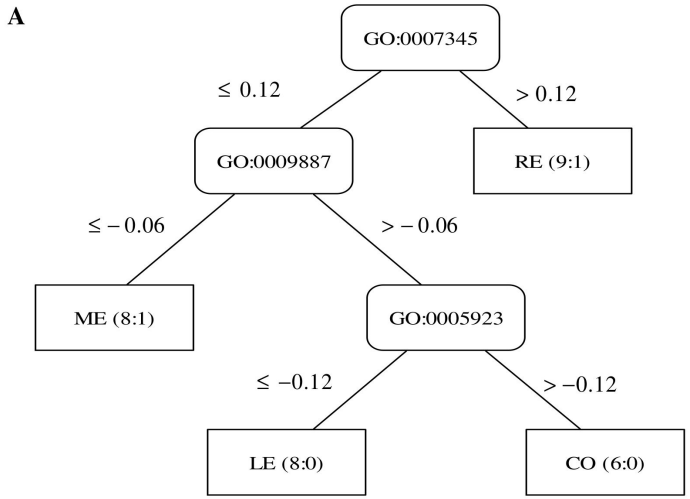
Evaluation using C4.5

General: Decision trees generated by C4.5 can be used for classification, which is why this algorithm is often referred to as a statistical classifier. Authors of Weka use C4.5.

How:

1. Obtain a list of classes (labels) and a matrix of values
2. For each attribute (gene) *attr*
   a. Find the normalized information gain ratio from splitting on *attr*
   b. Split the data on the attribute *attr_best* where the information gain ratio is highest
   c. Create a decision node that splits the data at *attr*
3. Continue on the sublists obtained from splitting the data at *attr*

Evaluation using C4.5



Authors of the paper were successful with selection process and decision tree, finding one distinguishing DLBCL with 98% precision from other cancer types

Evaluation

Technical: How will this approach be evaluated, using which classification algorithm(s)? With which algorithm will I start? To which other, mainly statistical, approaches will i compare the approach? On which datasets will be compared? Are there already solutions using the TCGA-Dataset?

What are the results I think I will see, why do I think so? What does the paper state? Can one include Graphics from the paper, if so, explain what I think this explanation will answer.

Write one experiment as an example, with setting, data, ...





Q&A

- How does the process of getting to functional expression profiles from differentially expressed genes work?
- How does C4.5 work internally?
- Can this approach potentially be used in clinics? What do you think?
- What is the difference in using mean vs. using median and why does one use this or that, depending on this or that distribution?
- What exactly is integrative gene selection in this context?
- How does your approach behave? Filter, Wrapper, ((extending or sth. alike))?
- Where does org.Hs.eg.db come from? Is it trustworthy?

Questions: Good / Bad? Robustness?

To compare this approach to other approaches I will have to compute the measurement scores for this approach and check for other gene selection approaches and the respective measurement scores

To check for robustness I will take more TCGA-Data that has the same features (genes) and will simply run in batches through the trained classifier

If the approach is robust, then I expect to achieve similar measurement scores for every batch of unseen data that follows the same structure as seen with the test-data (k=5)

Question: Other datasets? Vs. Statistical?

To see how good this approach performs on other datasets I will search for labeled, publicly available and approved datasets, run the approach on this dataset too, compute the measurement scores and compare it to other approaches used on the same dataset

To answer how good this approach is compared to “plain statistical” approaches, I will compare it to statistical approaches from the IGEA-Framework or even implement a statistical approach myself and compare the measurement scores