C17: CANCIFIER

Predicting cancer tissue of origin based on gene expression profile

TEAM: Danat Yermakovich, Kati Koido, Maksym Zarodniuk

Our Github: https://github.com/plezar/cancifier

Task 2. Business understanding

- Identifying your business goals
  - Background - Cancer of unknown primary (CUP) is defined as metastatic cancer where a primary site of origin cannot be found. CUP accounts for up to 5% of all new cancer cases, with a 5-year survival rate of only 10%. Because knowledge of a patient's primary cancer remains fundamental to their treatment, accurate identification of tissue of origin would allow for directed, personalized therapies to improve clinical outcomes [1, 2].
  - Business goals - Our aim is to identify the site of origin of metastatic cancer for which the tissue of origin is not known using machine learning methods.
  - Business success criteria - Since our business goals are the same as data-mining goals, our success criteria are the same as data mining success criteria. Please see the corresponding paragraph.

- Assessing your situation
  - Inventory of resources
    - three students working on a project
    - support of a few experts in the fields of bioinformatics and machine learning
    - access to High Performance Computing (HPC) and necessary software
    - access to Human Cancer Metastasis Database (HCMDB)
  - Requirements, assumptions, and constraints - We have created a schedule for completion, and the deadline for the project was set to be 14 December (12:00). Requirements for acceptable finished work have been outlined above. Two students have accessed the data repositories and downloaded files, no constraints met.
  - Risks and contingencies - One risk could be that a team member falls ill and therefore the project can't be completed by the deadline. We have regular meetings, have started teamwork early, and have discussed some potential problems that could arise.

- Terminology
  - human tissue - our project studies gene expression data collected from human tissue, no comparative analysis with other species will be carried out;
  - cancer - cancer is an abnormal growth of cells that have no functional purposes and the potential to spread to the adjoining cells or organs or other parts of the body [3];
  - CUP - cancer of unknown primary is defined as metastatic cancer where a primary site of origin cannot be found;
  - gene - gene is a DNA sequence (whose component segments do not necessarily need to be physically contiguous) that specifies one or more sequence-related RNAs/proteins [4];
  - gene expression - gene expression is the process by which information from a gene is used in the synthesis of a functional gene product (Wikipedia);
  - bulk RNA-seq - standard bulk RNA-Seq analysis analyze the expression of RNAs from large populations of cells (Wikipedia);
  - NGS - Next-Generation Sequencing, the high-throughput method for gathering ample information about nucleotide sequences such as DNA;
  - read - one short sequence of DNA that covers some specific area in the human genome, obtained by NGS;
  - RNA microarray - microarray technology allows the simultaneous measurement of tens of thousands of messenger RNA (mRNA) transcripts for gene expression [5];
  - log2 transformation of data - used for gene expression data and results in values that are approximately normally distributed [6];
  - quantile normalization of data - normalization is a process of removing non-biological variation between multiple high density oligonucleotide arrays. The goal of the quantile method is to make the distribution of probe intensities for each array in a set of arrays the same [7];
  - batch effect - batch effect represents the systematic technical differences when samples are processed and measured in different batches and which are unrelated to any biological variation recorded during the microarray gene expression experiment [8];
  - GEO - Gene Expression Omnibus is an international public repository that archives and freely distributes microarray, NGS, and other forms of high-throughput functional genomics data submitted by the research community; https://www.ncbi.nlm.nih.gov/geo/info/overview.html#general;
  - TCGA - The Cancer Genome Atlas is a cancer genomics program; https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga;
  - HCMDB - Human Cancer Metastasis Database is an integrated database designed to store and analyze large scale expression data of cancer metastasis; https://hcmdb.i-sanger.com/;

        - unsupervised ML - unsupervised learning is a type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision (Wikipedia).

    - Costs and benefits - This aspect is not relevant at the moment.

- Defining your data-mining goals
    - Data-mining goals - Our data-mining deliverables will be following: 82 HCMDB and 3 TCGA datasets merged together in which rows correspond to labels (primary site of cancer) and columns correspond to features (genes). A trained model for predicting tissue of origin for CUP. A poster with introduction, methods, results and conclusions. A 3-minute video about our project.
    - Data-mining success criteria - Our evaluation metrics will be:
        - Accuracy: our goal is to make a predictive improvement to an existing method which had an average accuracy of 86.7% [9].
        - Precision
        - Recall
        - F1
        - Specificity
        - AUC

Task 3. Data understanding

- Gathering data
    - Outline data requirements - The data that we would like to gather comes from different experiments and is stored in different databases. Our first step is to download, preprocess, and merge them. The primary source of information about individual datasets comes from the data_information table which contains information about labels and some of the features.
      Requirements:
        - log2 transformed and normalized to remove the batch effect
        - genes are filtered by their mean expression (only top 12000 genes are left in the final dataset)
        - any features or labels should not be missing and the data themselves should not contain any missing values

    - Verify data availability - Two students have accessed the data repositories and downloaded files, no constraints met.
    - Define selection criteria - All that could be processed. In GEO datasets we extract the data from the fields “VALUE”, “GB_ACC”, and “ID_REF.” We also use the “dataset_information.xlsx” spreadsheet with a summary of the entire database. For TCGA datasets we obtain raw read counts (htseq tool) per sample according to the mentioned in the spreadsheet datasets.
- Describing data
    - sources of the data: individual studies

    - formats: .soft and .series GEO data; raw reads count (two-columns tables: gened_id read_counts) from TCGA
    - there’s some preprocessing done on the individual GEO datasets (log2 transformation)
- Exploring data
    - performing PCA/tSNE and clustering on the final dataset
- Verifying data quality
    - planning to check for outliers, missing values, correct for technical noise in the data etc
    - since the data are mostly dispersed through different datasets, the batch effect should be taken into account as well as normalize all data within the processing and in the end.

Task 4. Planning your project

Task	Danat	Kati	Maksym
Access databases and download data files	8	1	9
Data preprocessing and quality control	10	1	12
Machine learning	6	10	6
Interpreting the results	6	7	4
Presentation of results	4	8	3

Hours each team member is going to contribute to each task

1. Access databases and download data files

We have fetched 82 GEO datasets and 15 TCGA datasets from the corresponding databases.

2. Data preprocessing and quality control

We have applied quantile normalization to the obtained individual datasets, merged them, and then normalized the resulting megaset to correct for the batch effect. We have also filtered out lowly expressed genes by keeping only the top 15000 expressed genes. In the quality control step, we are planning to first of all check for missing values, and then use principal component analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and hierarchical clustering to better understand the structure of our data. This will also allow us to detect outliers.

3. Machine learning

In this step we are planning to use several models. We are planning to train a random forest classifier and a non-linear SVM, probably using recursive feature elimination (RFE) [10]. Since our dataset will be imbalanced, we are planning to assign appropriate weights to different classes to avoid undersampling. To evaluate our models, we will use V-fold cross-validation in order to retain the original size of the training dataset.

4. Interpreting the results

We are planning to use the aforementioned (see task 2, data-mining goals) list of metrics to evaluate our models. If there is enough time, we will try to optimize some parameters or use different models in case of a poor performance of our models.

5. Presentation of results

We will make a poster as well as a video to present and explain the results of our project.

References

[1] Wei I H, Shi Y, Jiang H, Kumar-Sinha C, and Chinnaiyan A M (2014). RNA-Seq Accurately Identifies Cancer Biomarker Signatures to Distinguish Tissue of Origin. Neoplasia 16, 918–927

[2] Zhao Y, Pan Z, Namburi S, Pattison A, Posner A, Balachander S, Paisie C A , Reddi H V, Rueter J, Gill A J, Fox S, Raghav K P S, Flynn W F, Tothill R W, Li S, Karuturi R K M, George J (2020). CUP-AI-Dx: A tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. EBioMedicine 61, 103030

[3] Wang J-J, Lei K-F, Han F (2018). Tumor microenvironment: recent advances in various cancer treatments. Eur Rev Med Pharmacol Sci 22, 3855-3864

[4] Portin P and Wilkins A (2017). The Evolving Definition of the Term “Gene”. Genetics 205, 1353–1364

[5] Sealfon S C, Chu T T (2011). RNA and DNA Microarrays. In: Khademhosseini A., Suh KY., Zourob M. (eds) Biological Microarrays. Methods in Molecular Biology (Methods and Protocols), vol 671. Humana Press, Totowa, NJ

[6] Transformation and Normalization. In: Analysis of Microarray Gene Expression Data. (2004). Springer, Boston, MA. https://doi.org/10.1007/1-4020-7788-2_6

[7] Bolstad B M, Irizarry R A, Astrand M, Speed T P (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 22, 185-93

[8] Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, Weiss-Solís D Y, Duque R, Bersini H, Nowé A (2012). Batch effect removal methods for microarray gene expression data integration: a survey. Brief Bioinform 14, 469–490

[9] Liu X, Li L, Peng L, Wang B, Lang J, Lu Q, Zhang X, Sun Y, Tian G, Zhang H and Zhou L (2020). Predicting Cancer Tissue-of-Origin by a Machine Learning Method Using DNA Somatic Mutation Data. Front Genet 11:674

[10] Pirooznia M, Yang J Y, Yang M Q, Deng Y (2008). A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics 2008, 9(Suppl 1):S13