Biology 644

Old Title: Bioinformatics for Molecular Biologists

Potential New Title: Integrated Bioinformatics Using R for Both Wet and Dry Scientists

Outline

- Cancer Review
- Golub, et al. 1999 Science Paper
- R Lab
  - Chapter 1 Supplemental
  - Chapter 1: Sections 1.5 – 1.10
  - Chapter 1: Section 1.11 Exercises

Cancer

- Known medically as a malignant neoplasm
- A broad group of diseases involving unregulated cell growth.
- Cells divide and grow uncontrollably, forming a tumor
- Not all tumors are cancerous; benign tumors do not invade neighboring tissues and do not spread throughout the body.
- Malignant tumors invade nearby parts of the body.
- Metastasis is the spread of a malignant cancer to more distant parts of the body through the lymphatic system or bloodstream.
- 0ver 200 different known cancers that affect humans.
- Causes of cancer are diverse, complex, and only partially understood. Many things are known to increase the risk of cancer, including tobacco use, dietary factors, certain infections, exposure to radiation, lack of physical activity, obesity, and environmental pollutants. These factors can directly damage genes or combine with existing genetic faults within cells to cause cancerous mutations.
- Approximately 5–10% of cancers can be traced directly to inherited genetic defects.

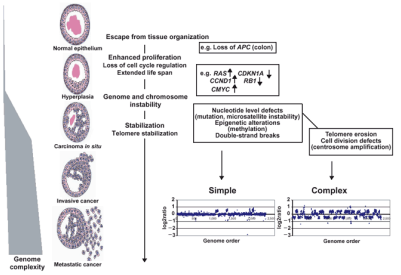
Cancer II

- Cancer can be detected in a number of ways, including the presence of certain signs and symptoms, screening tests, or medical imaging.
- Once a possible cancer is detected it is diagnosed by microscopic examination of a tissue sample by a pathologist.
- Cancer is usually treated with chemotherapy, radiation therapy and/or surgery.
- The chances of surviving the disease vary greatly by the type and location of the cancer and the extent of disease at the start of treatment. Early diagnosis is key to successful treatment.
- While cancer can affect people of all ages, and a few types of cancer are more common in children, the risk of developing cancer generally increases with age.
- In 2007, cancer caused about 13% of all human deaths worldwide (7.9 million).
- Rates are rising as more people live to an old age and as mass lifestyle changes occur in the developing world.
- Generally, microtumors accumulate in our bodies as we age, and the chances of one of them becoming cancerous increases with age

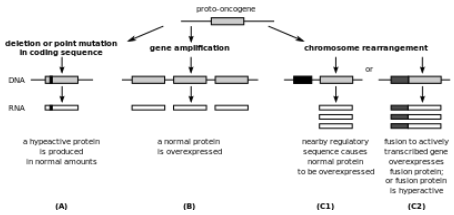
Cancer III

- As tumors grow they evolve with increased mutations, tumor heterogeneity, chromosomal instability, copy number variation, deletions, and translocations
- An oncogene is a gene that has the potential to cause cancer.
- Activated oncogenes can cause those cells designated for apoptosis to survive and proliferate instead.
- A proto-oncogene is a normal gene that can become an oncogene due to mutations or increased expression
- Highly aggressive and metastatic tumors often have acquired a handful of key mutations or translocations that have:
  - deactivated the cell cycle arrest and apoptosis pathways
  - activated at least a proto-oncogene
- 50% of all tumors have a p53 mutation
- Examples of proto-oncogenes include RAS, WNT, MYC, ERK, and TRK
- Chemotherapy and radiation therapy aim to push the unstable cancerous cells down the apoptosis pathway (while not killing too many of the healthy cells.)

Chromosomal evolution in human solid tumor progression



From proto-oncogene to oncogene



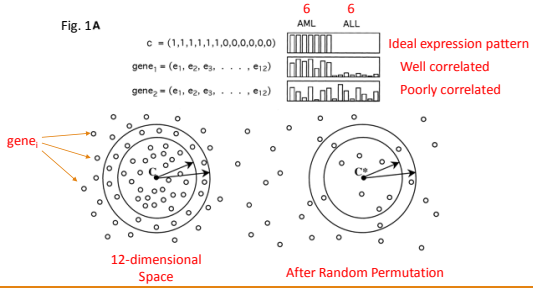
"Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" – Golub et al., Science 1999

- Training Set = 27 ALL and 11 AML gene expression sets
- Test Set = 24 bone marrow and 10 peripheral blood samples
- RNA prepared from bone marrow mononuclear cells
- Affymetrix microarray containing probes for 6,817 genes
- Supervised Learning to classify ALL vs. AML
  - Using a voting scheme with 50 "informative genes" (biomarkers)
  - 25 highly expressed in ALL compared to AML
  - 25 highly expressed in AML compared to ALL
  - Prediction strength used to provide prediction confidence and a threshold for "certainty"
- Unsupervised Learning to discover T- and B- lineage cancer subtypes
  - 2-Cluster SOM (Self-Organizing Map) properly clustered ALL vs AML
  - 4-Cluster SOM also "discovered" 2 subtypes: T- vs B-lineages ALL

"Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" – Golub et al., Science 1999

"neighborhood analysis"

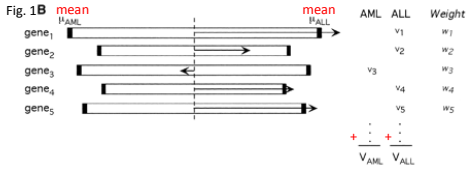
Fig. 1A



"Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" – Golub et al., Science 1999

Supervised Learning
50 Biomarkers ("informative genes")

Fig. 1B

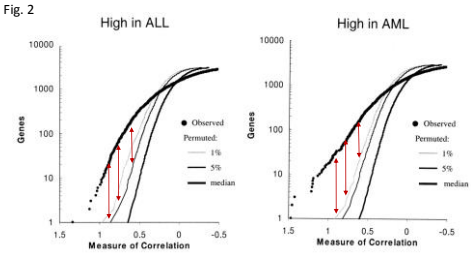


$v_i = |x_i - (\mu_{AML} + \mu_{ALL})/2|$ reflects the deviation of the expression level in the sample from the average of μ_{AML} and μ_{ALL}

w_i is a weighting factor that reflects how well the gene is correlated with the class distinction

"Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" – Golub et al., Science 1999

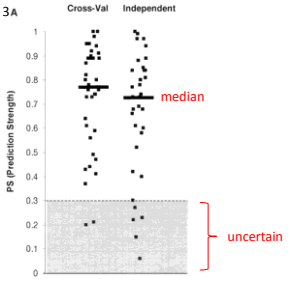
Fig. 2



"Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" – Golub et al., Science 1999

Supervised Learning

Fig. 3A



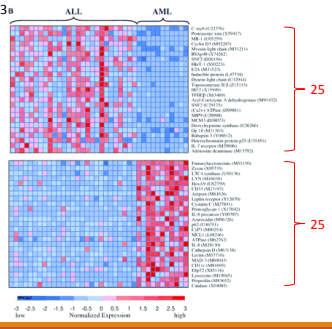
Prediction Strength reflects the margin of victory and is defined as $(V_{win} - V_{lose})/(V_{win} + V_{lose})$, where V_{win} and V_{lose} are the respective vote totals for the winning and losing classes

Leave-one-out (jack knife) cross-validation scheme

"Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" – Golub et al., Science 1999

Fig. 3B

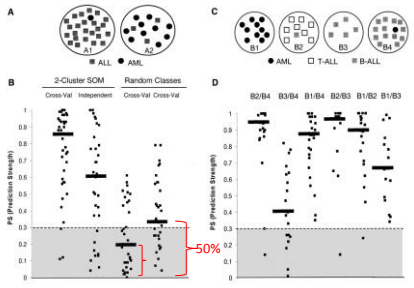
Red-Blue Heatmap



"Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" – Golub et al., Science 1999

Fig. 4

Unsupervised Learning



Biomarker discovery aims to properly classify the type, subtypes, stage, or therapeutic response of a cancer