Case-based Reasoning for the Analysis of Methylation Data in Oncology

Christopher Bartlett and Isabelle Bichindaritz

CASE-BASED REASONING

- Case-Based Reasoning (CBR) is using previous experiences to understand and solve new problems.

"A case-based reasoner solves new problems by adapting solutions that were used to solve old problems."

Riesbeck & Schank, 1989

R<u>etrieve</u>.
When a new problem is presented, similar cases are retrieved from memory.

R<u>euse</u>.
The solution of the retrieved cases are reused.

R<u>evise</u>.
The solution is revised to fit the new problem.

R<u>etain</u>.
The revised solution is retained for future use.

DNA METHYLATION

- Methylation is the attachment of a methyl group to the DNA molecule.
- Research in methylation investigates the amount of methyl and where it differs in two or more groups
- These differences can be at the position or region level.
- Positions are individual probes on the chip used to detect DNA methylation.
- Regions are clusters of probes that serve a similar functional purpose for gene transcription.

AIMS

- Refine cases to determine a genetic signature for stage 4 breast cancer.
- Use classification to verify the located genetic signature.

CONTRIBUTIONS

- One of the first applications of CBR using methylation data.
- Multi-level case elaboration and refinement which examines biological and statistical differences.

MATERIALS

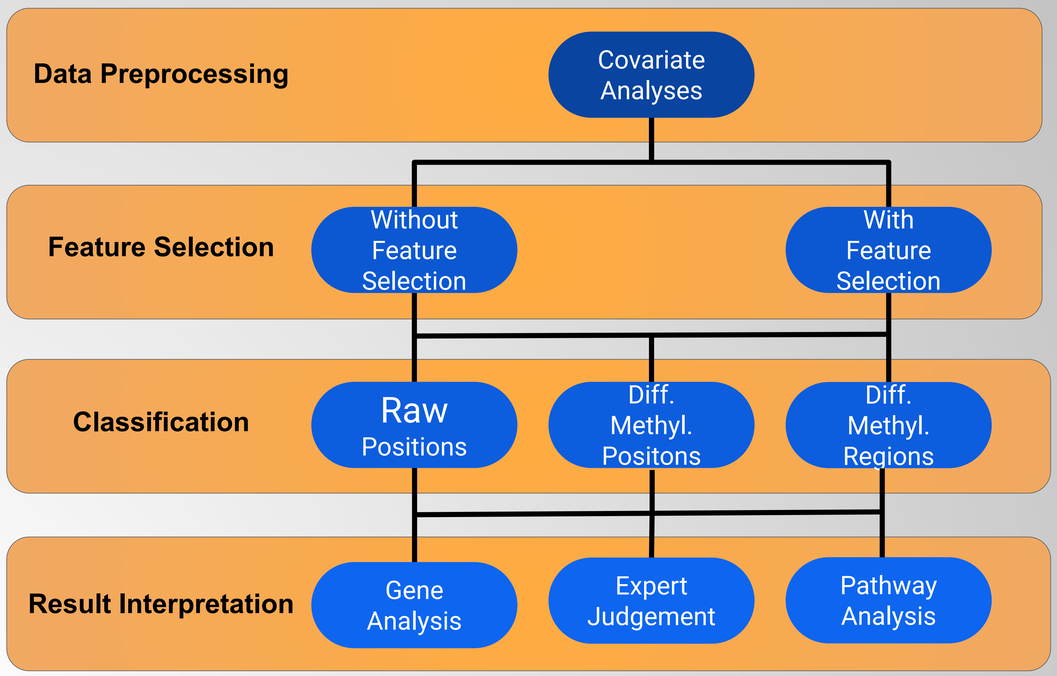
- Methylation data for breast cancer was downloaded from The Cancer Genome Atlas
  - 892 samples
  - 485,577 probes

METHODS

- Extracted 95 solid tissue control samples.
- 10 stage 4, primary solid tumor samples.
- Removed probes significantly associated with a covariate or batch variable were removed.
- 105 samples, 120, 681 sites remained.
- Nearest 1, 2 or 3 cases were retrieved based on similar methylated profiles.
- Used multi-stage feature selection and feature ranking to determine the most relevant features.

Performance Metrics

- Balanced Accuracy
- Area Under the Curve (AUC)



System Architecture

RESULTS AFTER PREPROCESSING

Nearest Cases	BACC	AUC	Correct M Samples
1 case	70%	0.700	4
2 cases	65%	0.750	3
3 cases	75%	0.800	5

RESULTS OF HIGHEST RANKED GENES

Genes	1	1	5	5	10	10	15	15
Classifier	BACC	AUC	BACC	AUC	BACC	AUC	BACC	AUC
1 case	95%	0.950	100%	1.0	100%	1.0	100%	1.0
2 cases	95%	0.950	100%	1.0	100%	1.0	100%	1.0
3 cases	95%	0.949	100%	1.0	100%	1.0	100%	1.0

CONCLUSION

- These experiments display the usefulness of feature selection to improve efficiency and effectiveness of classification on highly dimensional data.
- The methylation signature located that uses a small number of genes will be invaluable for determining a deeper pathophysiological process of the disease.
- This work will be presented at the 33rd International FLAIRS conference in North Miami Beach, Florida.