École des Mines de Saint-Étienne

Data Science Major

SLIDE

Group 15

Wang Yuteng

Xu Liwei

Ye Wenjing

Zhao Wenxu

A Markov Chain Model-based Method for Cancer Classification[1]

Keyword- Cancer classification; Markov chain; Gene regulation; Gene pathway

[1] Li D , Wang H Q . A Markov chain model-based method for cancer classification[C]// Eighth International Conference on Natural Computation. IEEE, 2012.

Content.

Background
We are facing the huge challenge posed by cancer.

Introduction
Introduce the cancer classification model based on markov chain.

Modeling
Build models, calculate their parameters and evaluate the performance.

Conclusion
Summarize the article and evaluate the model.

We present the article from the four aspects mentioned above.

Let's talk something about cancer.

Background.

In this chapter, we would like to share some brief understanding of cancer and the main challenges facing humanity.

Cancer has always been a powerful enemy of human.

Cancer.

The cell becomes cancerous and it is invading our body.

In medicine, cancer refers to malignant tumors originating from epithelial tissues.

It is the most common type of malignant tumors.

18,100,000+ New

9,700,000+ Died

In 2018.

1/3 of cancers can be prevented, 1/3 of cancers can be cured if diagnosed early ...

- International Anti-Cancer Alliance

Discover and classify cancer early.

Studies have found that early cancer is more likely to be cured.

In the early stage, the patient's immunity is relatively strong. If the cancer is correctly classified at this time, more treatment options can be selected, and the cancer cells are more likely to be completely eliminated..

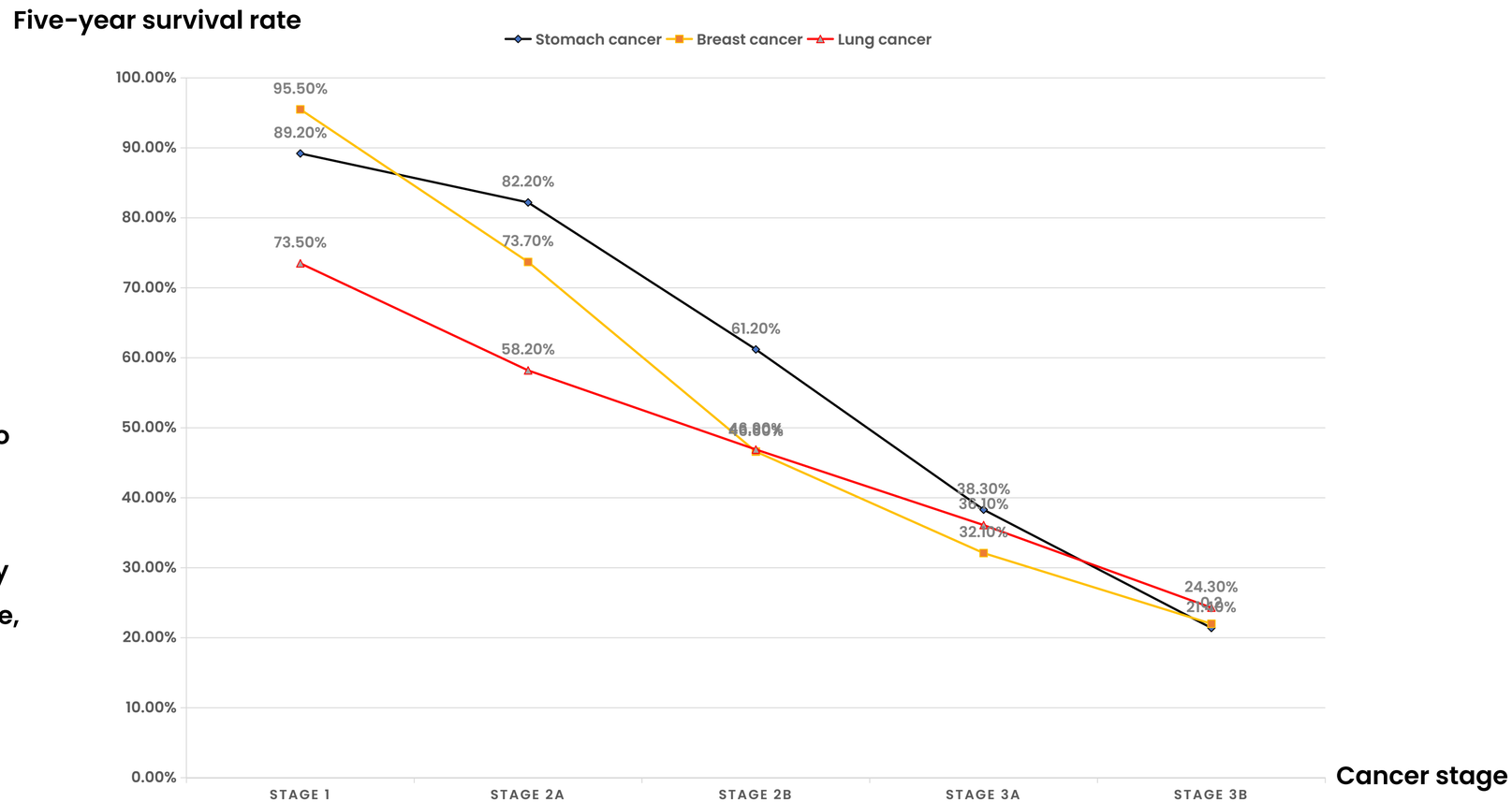


Figure: Five-year survival rate varies with cancer stage

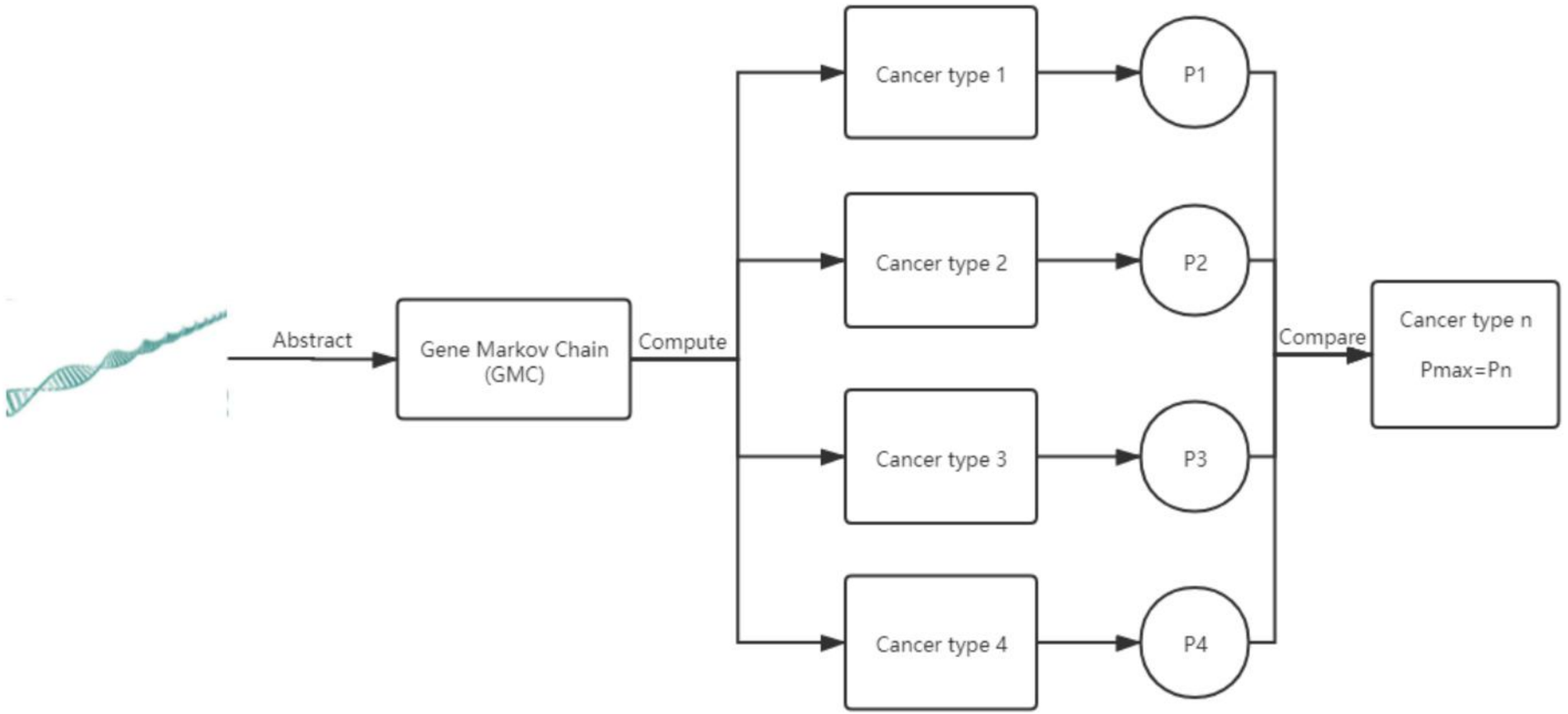
What are we going to do for this?

Introduction.

In this part, a Markov chain model(MCM) based method is proposed for cancer classification. Cancers can be classified into different types according to the the form of regulation state transition in gene pathway.

Prognose cancer type on the basis of their probability scores.

Introduction.

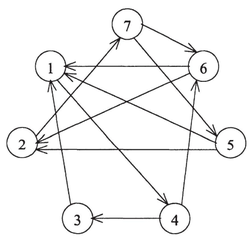


Schematic diagram

Introduction.

We view a gene chain extracted from a gene pathway map as a gene Markov chain (GMC).

Gene Chain



Markov Chain

Genetic signal can sequentially flows from the first gene to the last one .

- $S_g = \{-1,0,1\}$

A 3-state regulatory state space , where -1,0 and 1 represent down-regulation, non-regulation and up-regulation respectively.

- $G = \{g_1, g_2, g_3 \cdots\cdots g_n\}$

G represent a gene chain of length n.

- $\{X(g), g \in G\}$

It represents a GMC ,where X(g) represents the regulation state of the gth gene.

Modeling Method.

For an observation (sample), the probability that it comes from a given GMC model can be easily obtained by

$$\mathrm{P}(x) = p_0(x1)\prod_{t=1}^{n-1} p(X(t+1) = x_{t+1} \mid X(t) = x_t)$$

$$= p_0(x_1)\prod_{g=1}^{n-1} p(x_{g+1}, x_g, g)$$

where $X=\{x_1, x_2, \ldots, x_n\}$ is an observation with n observed states
$p_0(x_1)$: initial statedistribution
$p(x_{g+1}, x_g, g)$: state transition probability

Homogeneous Markov Chain

Non-Homogeneous Markov Chain

HMC

NMC

$$p_0(x) = \frac{1}{w}\sum_{k=1}^{w} I(x_{k,1} = x)$$

same k-step transition probability matrix at any time

$$p(x,y) = \frac{\sum_{k=1}^{w}\sum_{g=1}^{n-1} I(x_{k,g} = x \,\&\, x_{k,g+1} = y)}{\sum_{k=1}^{w}\sum_{g=1}^{n-1}\sum_{v\in S_g} I(x_{k,g} = x \,\&\, x_{k,g+1} = v)}$$

k-step transition probability matrix change with time

$$p_{ij}^{g} = p(x,y,g) = \frac{\sum_{k=1}^{w} I(x_{k,g} = x \,\&\, x_{k,g+1} = y)}{\sum_{k=1}^{w}\sum_{v\in S_g} I(x_{k,g} = x \,\&\, x_{k,g+1} = v)}$$

where $x_{k,1}$ represents the state of the first gene of the GMC in sample k.

$x_{k,g}$ and $x_{k,g+1}$ represent the states of gene g and $g+1$ in sample k.

I represents an indicator function whose value is equal to 1 if true and 0 otherwise.

How do we practice it?

Modeling.

In this chapter, we will apply the model to a real cancer classification problem.

Modeling Gene Pathway. 01

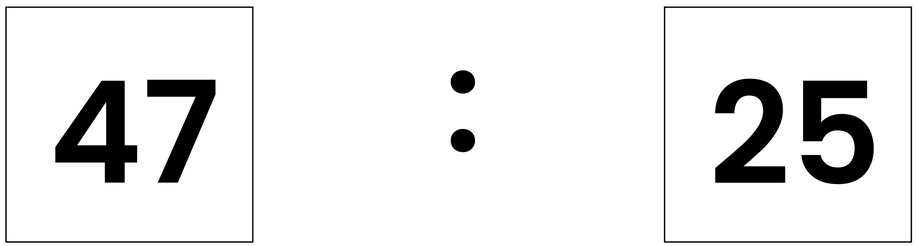
Take Acute Lymphoblastic Leukemia (ALL) and Acute Myelogenous Leukemia (AML) as examples for practical research.

Result 02

Evaluate the performance by comparing KNN, SVM, HMM and NMC models in various aspects.

Datasets.

The dataset contains 72 samples: 47 are ALL and 25 are AML, Each sample consists of the expression levels of 7129 genes.



Two human pathways related to leukemia : hsa04510 (focal adhesion) and has04010 (MAPK Signaling Pathway) .

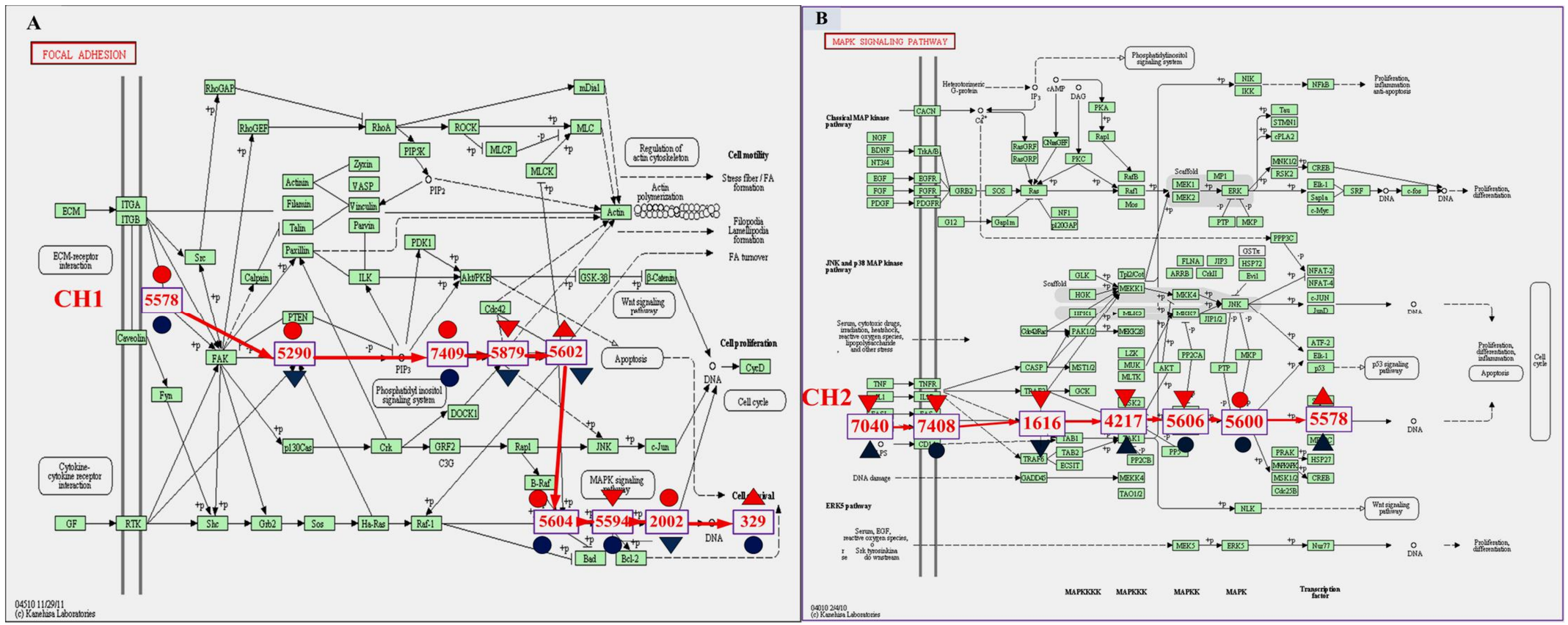
Pathway ID	Pathway name	Description
has04510	Focal adhesion	Cell-matrix adhesions
has04010	MAPK signaling pathway	mitogen-activated protein kinase (MAPK) cascade

Human pathway dataset from KEGG pathway database [http://www.genome.jp/kegg/]

Pathway.

Composed of gene

Principal State Transition (PST) : formed by the states with the highest transition probability in each step for each class

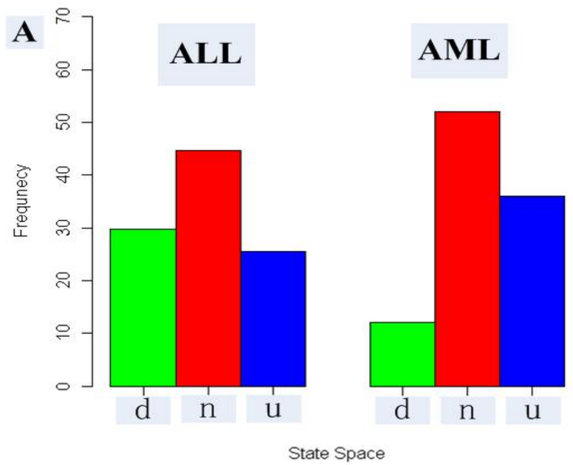


Maps of the two pathways used, has04510 (A) and has04010 (B), and PST of the two GMC, CH1 and CH2,for the two cancer classes, ALL (red) and AML (darkblue).
Solid circle, down- and up-triangle represent non-regulated,down-regulated and up-regulated state, respectively.
The numbers in red are the gene IDs of the corresponding genes.

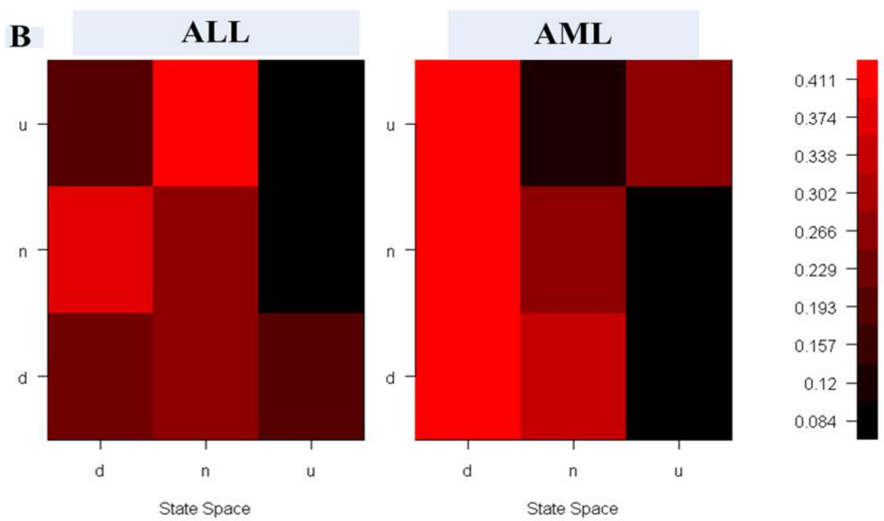
Human pathway dataset from KEGG pathway database [http://www.genome.jp/kegg/]

Model Parameters.

HMC transition matrix visualization

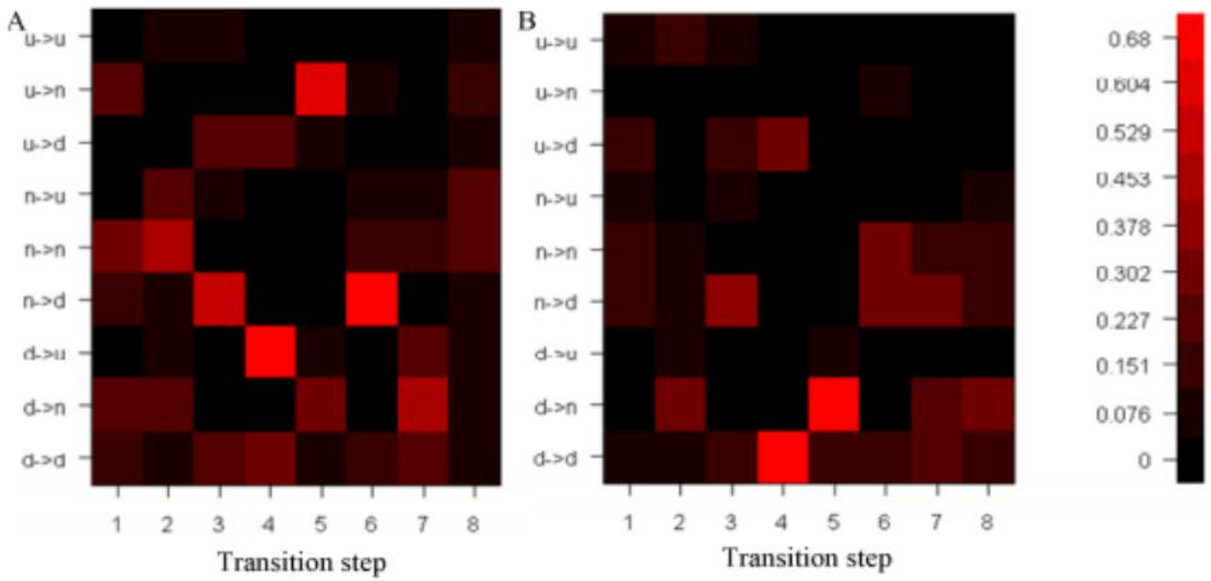


(A) Histograms of the initial state probability distributions



(B) heatmaps of the state transition probability matrices of HMCM on CH1 for ALL and AML.

NMC transition matrix visualization



Heatmaps of the state transition probability matrices of NMCM on CH1 for ALL (A) and AML (B).

* Symbols, u, d and n, represent up-, down- and non-regulation states.

Evaluation

Data obtained through training.

Methods	CH1						CH2					
	ACC	*std*	*min ACC*	*max ACC*	*SPE*	*SEN*	*ACC*	*Std*	*max ACC*	*min ACC*	*SPE*	*SEN*
HMCM	74.54	0.82	50	90.91	76.4	70.57	71.68	0.67	90.91	50	74.47	65.71
NMCM	**80.55**	0.57	59.09	**95.45**	84.73	71.57	**74.36**	0.67	**90.91**	**54.55**	79.4	63.57
KNN (k =1)	73.59	0.54	54.55	90.91	80.87	58	71.27	0.62	**90.91**	50	81.07	50.29
KNN (k =2)	74.23	0.66	59.09	**95.45**	81.07	59.57	71.72	0.62	86.36	50	81.13	51.57
SVM(*Gaussian Kernel*)	79.36	0.40	63.62	90.91	85.8	65.57	73.41	0.67	**90.91**	45.45	88.2	41.71

ACC-mean accuracy, minACC-min accuracy, maxACC-max accuracy, std-standard deivation of accuracies accuracy, SEN-sensitivity and SPE-specificity

ACC : Accuracy
SEN : Sensitivity
SPE : Specificity

Compared with the results by these previous methods, our method, HMCM or NMCM, achieved a more balanced values of specificity and sensitivity

What did we do?

Conclusion.

We hope to summarize experience and determine future work.

"In this part, we review what has been done in the article."

In the end.

Conclusion.

Proposed two models based on Markov chain, HMCM and NMCM, for cancer classification.They calculate the transition probability matrix and implement cancer classification.

Evaluated the two models on leukemia datasets and compared them with several previous methods including KNN and SVM. Experiment results show the effectiveness and efficiency of the MCMs, especially NMCM, for cancer classification.

“ Future work will focus on how to obtain larger data sets and improve models.”

École des Mines de Saint-Étienne

Thank you

Presentation about

A Markov Chain Model-based Method for Cancer Classification