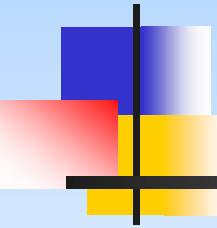
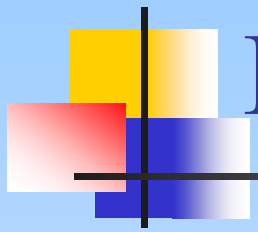
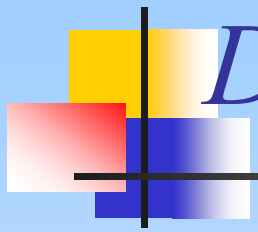


Data Replication for Mobile Computers



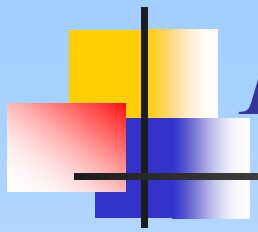
Introduction

- Data: Biomedical data
- Case study: Acute Lymphoblastic Leukaemia (ALL)
- ALL is a heterogenous disease
- More than clinical data required to find the best treatment protocol.



Data

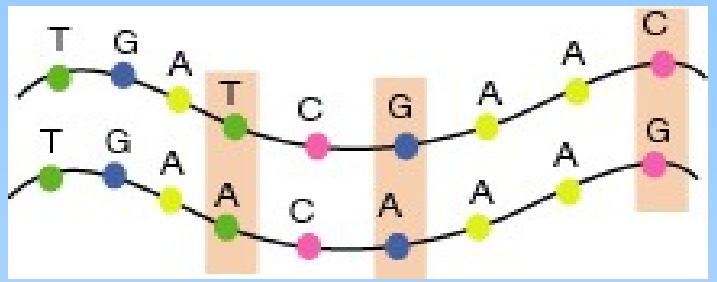
- Clinical data
- Patient outcome data
- Gene expression profile
- Domain ontology data
- Proteomics data
- Single Nucleotide Polymorphism (SNP)



Data

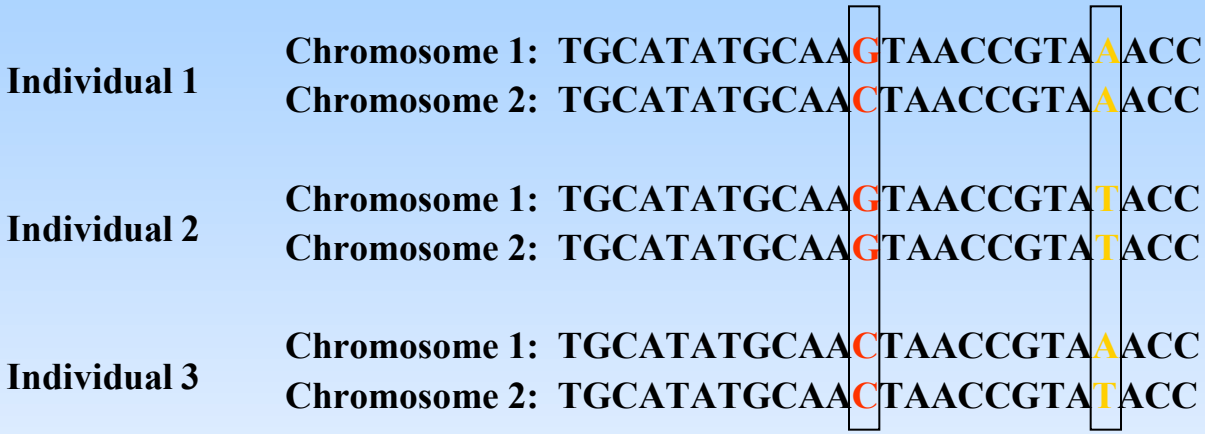
- Clinical data, such as
  - Age, sex, height, weight, white blood cell count, risk category (whether the child died or not) etc….
- Gene expression data
  - There are around 10,000 – 20,000 attributes
  - Several for each gene on the microarray
  - Real-valued, log values of the ratio of the red dye to green dye.

SNP data



- Most of human genetic variations exist in the form of polymorphisms
- SNP are the simplest but most abundant type of genetics variations
- Some of these polymorphisms that occur within coding region produce an amino acid change
- Such SNPs are known to affect the functional efficiency of genes

Single Nucleotide Polymorphisms



	SNP1	SNP2	SNP3	…….
Individual 1	Both	Allele1	Allele1	……
Individual 2	Allele1	Allele2	Both	…..
Individual 3	Allele2	Both	Allele2	……..

Hypothesis

- We hypothesize that the genetic background of childhood ALL patients, as assessed by genome-wide SNP profiles, will be informative of a patient response to therapy and eventual clinical outcome.

Data mining and knowledge discovery

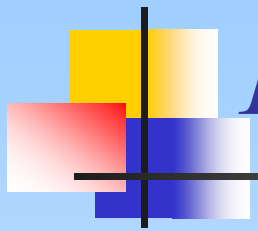
The aims of the project are:

i. Construct a model based on SNP data. How to deal with high-dimensionality problem induced by this data?

ii. Integrate SNP data with other datasets in order to have a better understanding of the problem

iii. Patient-to-patient comparison based on genome-wide SNP data and integrated dataset

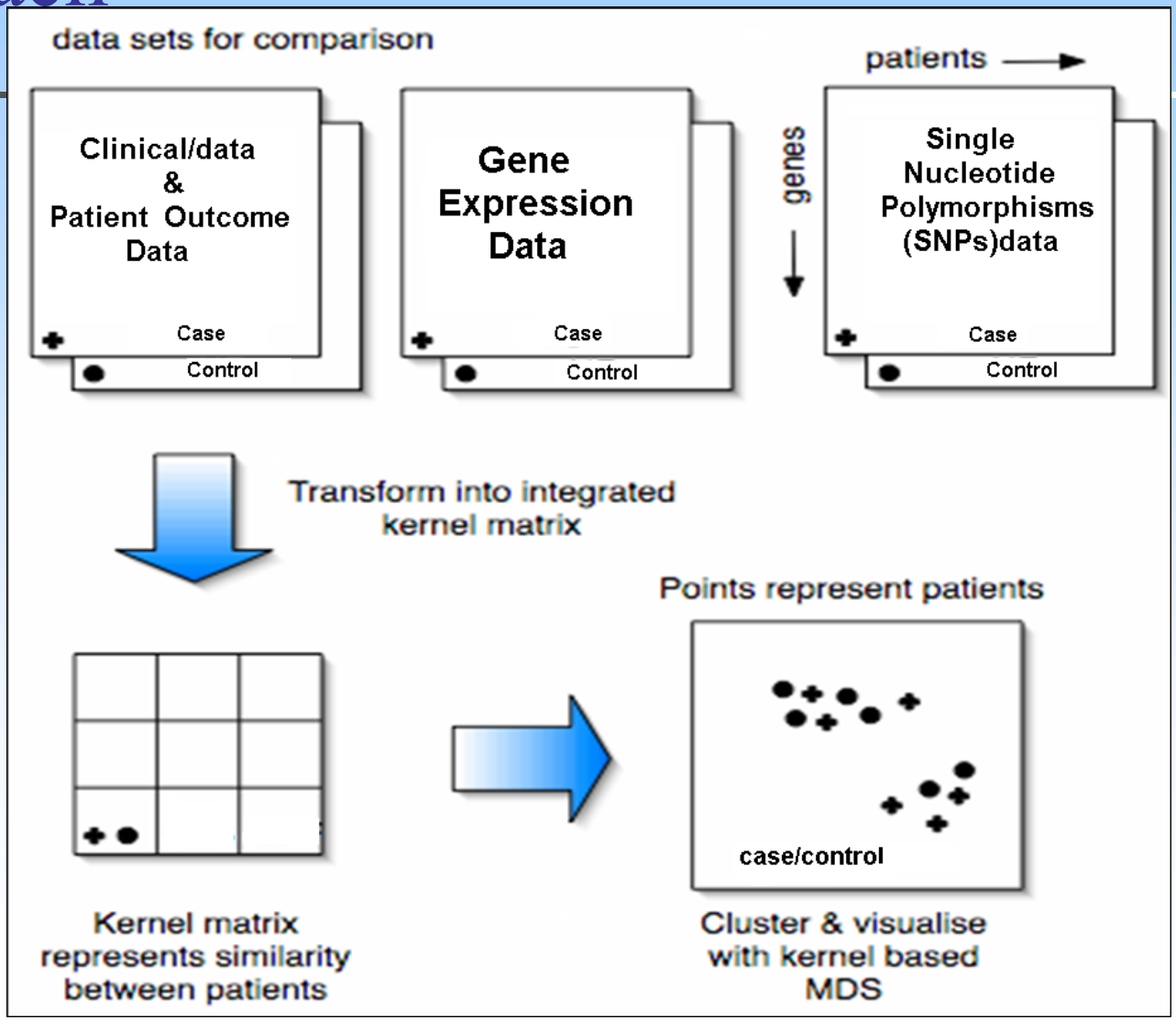
iv. Generation of knowledge – try to identify the genetic markers which correlate with poor patient response to therapy



Data mining approach

- Pattern recognition problems in system biology are characterized by high dimensionality and noisy data, limited sample size, etc.
- Affected by the curse of dimensionality.
- Focus on clustering and visualization of patients in the space of low-dimensional projection of the original data.
- Find a low dimensional (3-D) projection of the integrated datasets so that distance between points in the projection is similar to the distance in the kernel-induced feature space.
- Using kernel-based methods such as KPCA and Laplacian eigenmaps.
- Kernel methods can deal with high-dimensional data.

Approach


data sets for comparison
patients
genes
Clinical/data & Patient Outcome Data
Case
Control
Gene Expression Data
Case
Control
Single Nucleotide Polymorphisms (SNPs)data
Case
Control
Transform into integrated kernel matrix
Points represent patients
case/control
Kernel matrix represents similarity between patients
Cluster & visualise with kernel based MDS

Thank you

Questions ?