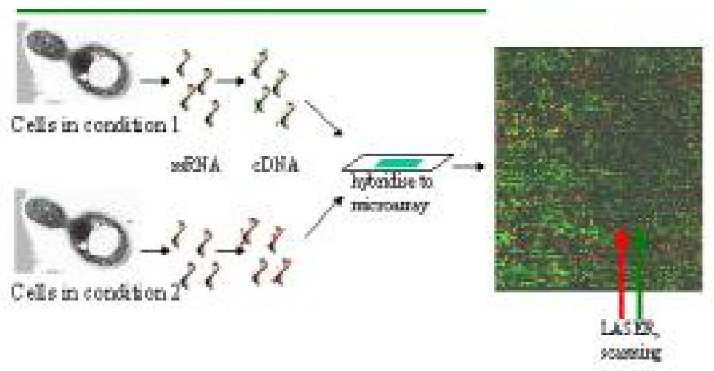
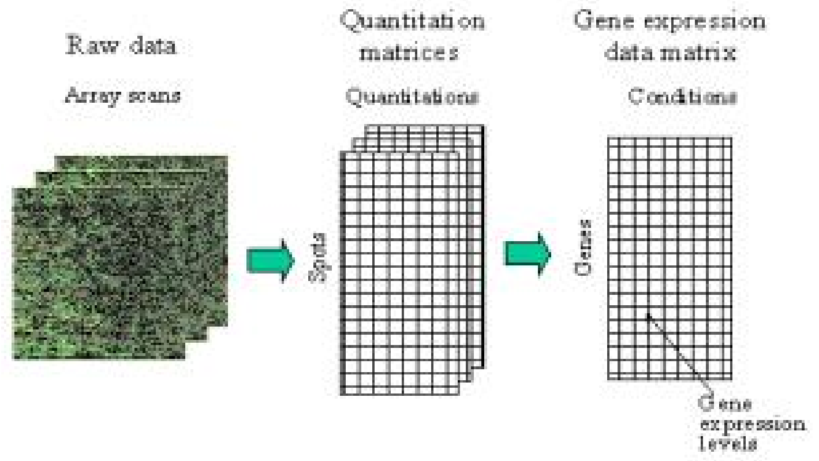
High Performance Clustering and Discrimination in Microarray Bioinformatics

P. Bhamidipati, A. Baras (Georgetown Un.)/B. Frankpitt (AIMS, Inc.), J. Baras

Gene Microarrays

- High-throughput experimental techniques that detect the expression of thousands of genes in a tissue simultaneously.
- DNA complementary to genes of interest are laid out in microscopic quantities at a specific location on an array and hybridized with mRNA from an experiment.
- Presence of DNA is detected by fluorescence following laser excitation.
- Applications:
  - Identification of functions for newly discovered sequences.
  - Drug discovery and toxicology.
  - Mutation or single nucleotide polymorphism(SNP) detection.





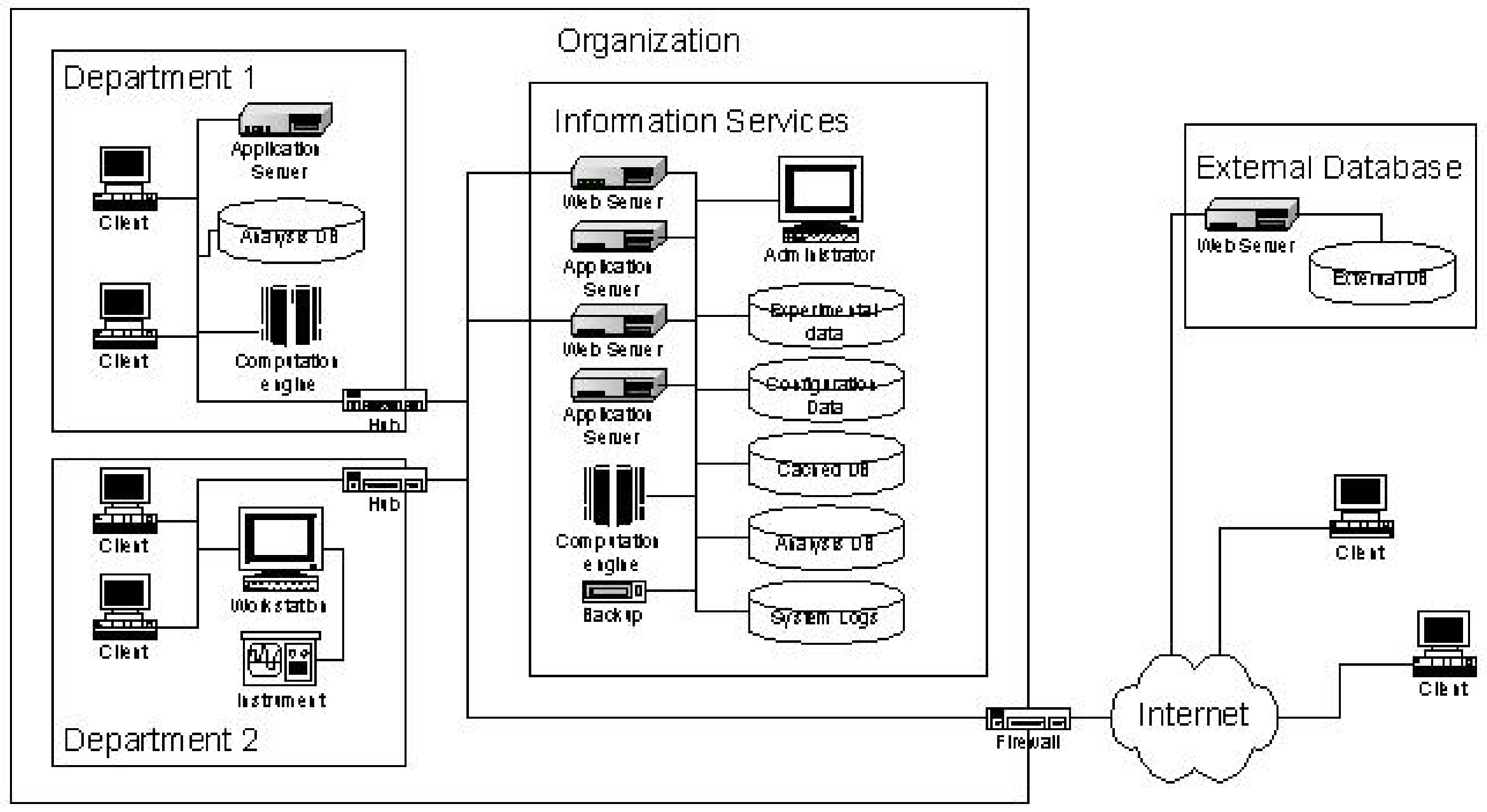
Expression Analysis

- Objectives:
  - Compare supervised classification with LBG/LVQ and SVMs (CLUSTER).
  - Compare the performance of the new unnormalized distance (AIMS patent pending) and log-Euclidean distances in ***similarity-based*** techniques with oligonucleotide arrays (**FIND_SIMILAR**).
  - Evaluate the performance of ***similarity measures*** in retrieving informative instances (**FIND_DISCRIMINATING**).
- Two types of experiments are studied:
  - Gene functional classification
  - Tissue-type/Phenotype classification
- Similarity Measures:
  - Log-Euclidean distance
  - New unnormalized distance (AIMS patent pending)

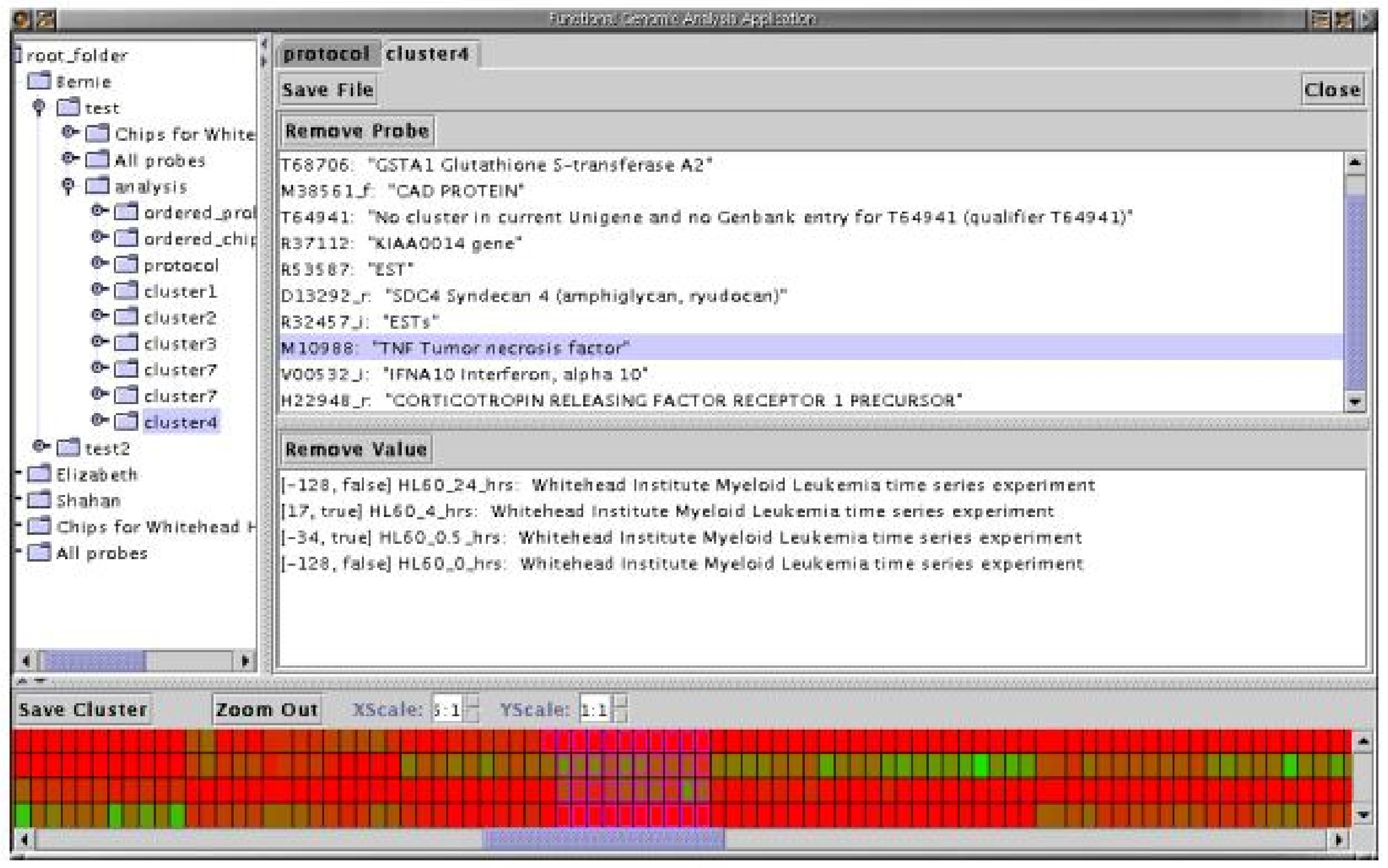
Modern Drug Discovery

- Four main phases
  - Target identification and validation
  - Compound identification and validation
  - Pre-clinical testing
  - Clinical Trials
- **Impact of genomic technologies**
  - Knowledge of gene sequence ➔Increased number of feasible targets: receptors to genes in the transcriptional mechanism in cancerous cells.
  - Genome-wide experiments ➔ Reduced turn-around time, *in vitro* toxicity testing.
  - Genotype analysis – correlation between chromosomal regions and traits.

System Architecture (AIMS)



User Interface (AIMS)



T68706: "GSTA1 Glutathione S-transferase A2"
M38561_f: "CAD PROTEIN"
T64941: "No cluster in current Unigene and no Genbank entry for T64941 (qualifier T64941)"
R37112: "KIAA0014 gene"
R53587: "EST"
D13292_r: "SDC4 Syndecan 4 (amphiglycan, ryudocan)"
R32457_i: "ESTs"
M10988: "TNF Tumor necrosis factor"
V00532_i: "IFNA10 Interferon, alpha 10"
H22948_r: "CORTICOTROPIN RELEASING FACTOR RECEPTOR 1 PRECURSOR"

[-128, false] HL60_24_hrs: Whitehead Institute Myeloid Leukemia time series experiment
[17, true] HL60_4_hrs: Whitehead Institute Myeloid Leukemia time series experiment
[-34, true] HL60_0.5_hrs: Whitehead Institute Myeloid Leukemia time series experiment
[-128, false] HL60_0_hrs: Whitehead Institute Myeloid Leukemia time series experiment

Project Goal:

Develop IT Infrastructure for biologists working with gene expression microarray data in large organizations.

Gene Expression Microarrays:

- Measures the concentration of mRNA molecules in small tissue samples.
- Expression levels for ~10^4 genes in a single assay.
- Biologists infer hypotheses about the regulation and structure of cell biochemistry from patterns in microarray data.

Tissue-type Classification

- Lymphoma Tissue-type Classification
  - Cancer Tissue Data Distribution (96 arrays & 4026 genes)

	DLBCL	GCB	NLT	APB	RAT	TCL	FL	RPB	CLL
# Arrays	46	2	2	10	6	6	9	4	11
Cancerous Tissue	Y	N	N	N	N	Y	Y	N	Y

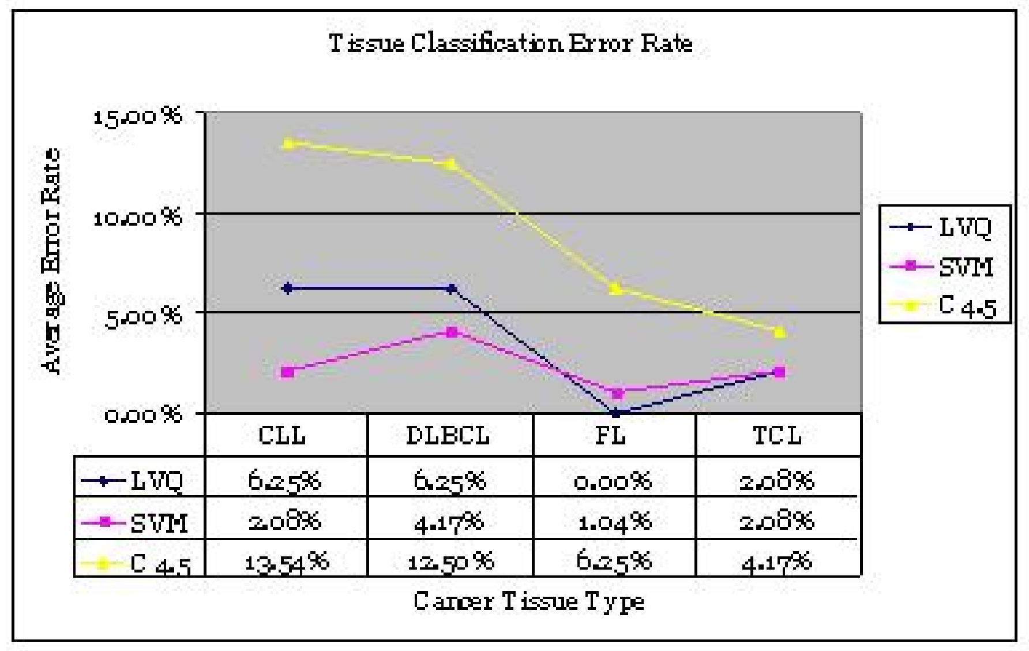
  - Methods compared with 10-fold cross-validation:
    - C4.5 (Quinlan) Decision Tree – univariate implementation, continuous-valued attributes.
    - SVM with a linear kernel.
    - LBG/LVQ with log-Euclidean Distance.
  - Tests run:
    - Binary Classification (Cancer Detection)
    - Cancer Tissue Classification
  - Performance measure = $\frac{fp + fn}{N_s}$

Results – Lymphoma Data

- Removing 218 known lymphoma genes does not affect the algorithm performance.
- The C4.5 decision tree performance, indicating that the cancer mechanism, which causes mutations in several genes can not be clustered with simple univariate, attribute-based rules.
- SVM outperforms log-Euclidean LVQ in binary phenotype classification, and in two types of cancer tissues.

Binary Classification	LVQ	SVM	C4.5
Average Error Rate	8.333%	1.000%	18.333%

Tissue Classification	LVQ	SVM	C4.5
Average Error Rate	9.72%	4.17%	29.17%



	CLL	DLBCL	FL	TCL
LVQ	6.25%	6.25%	0.00%	2.08%
SVM	2.08%	4.17%	1.04%	2.08%
C 4.5	13.54%	12.50%	6.25%	4.17%