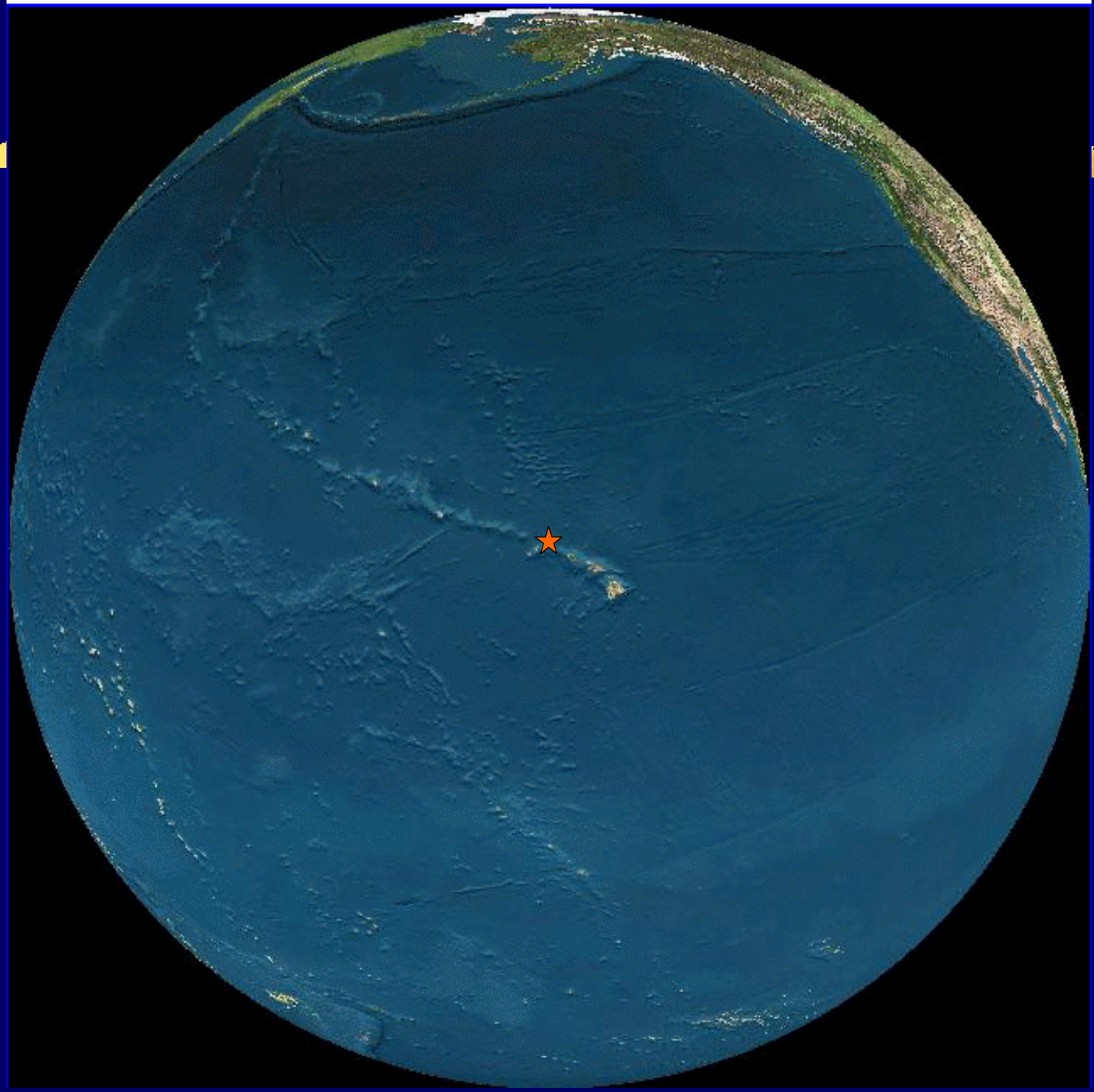
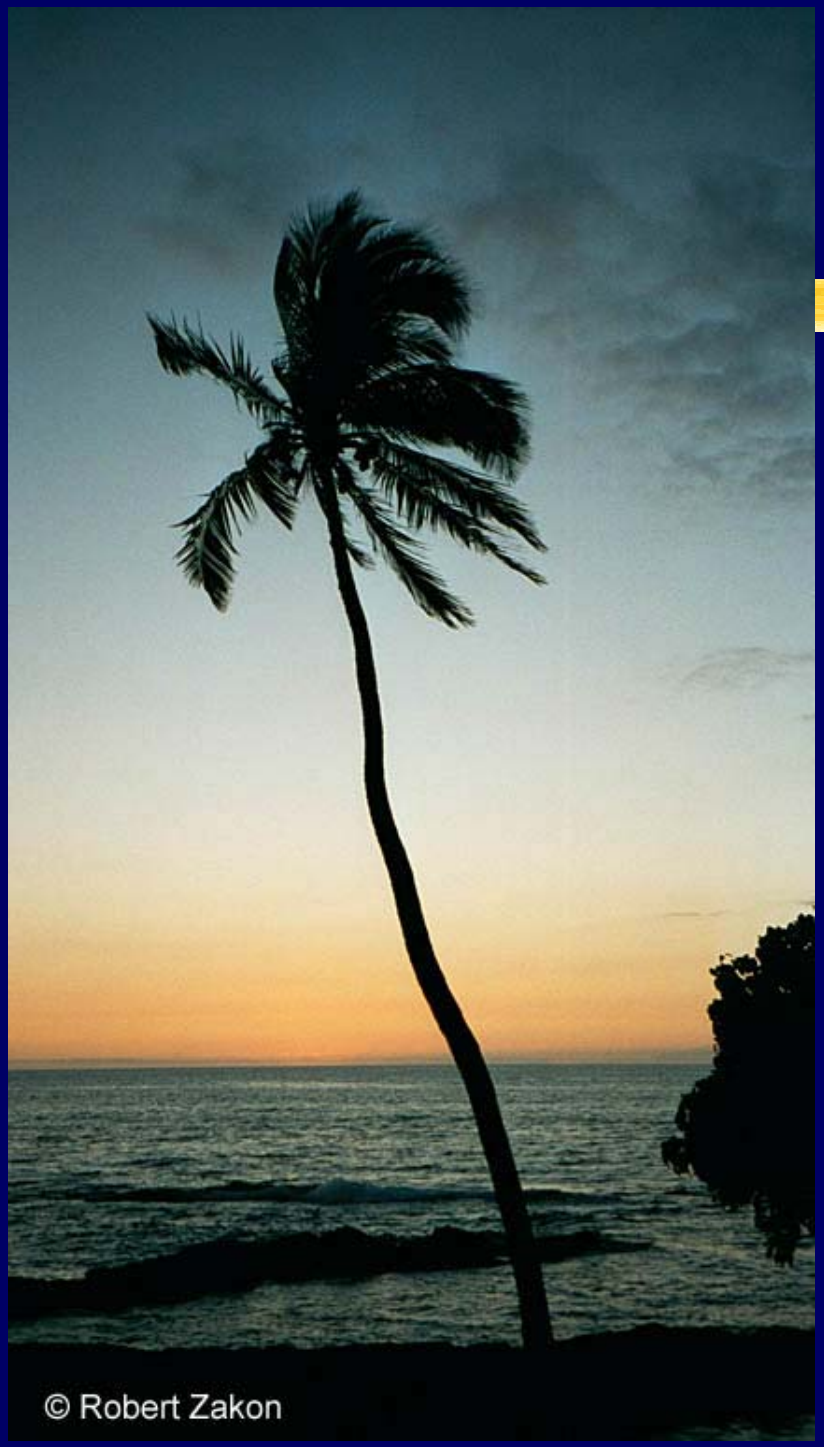
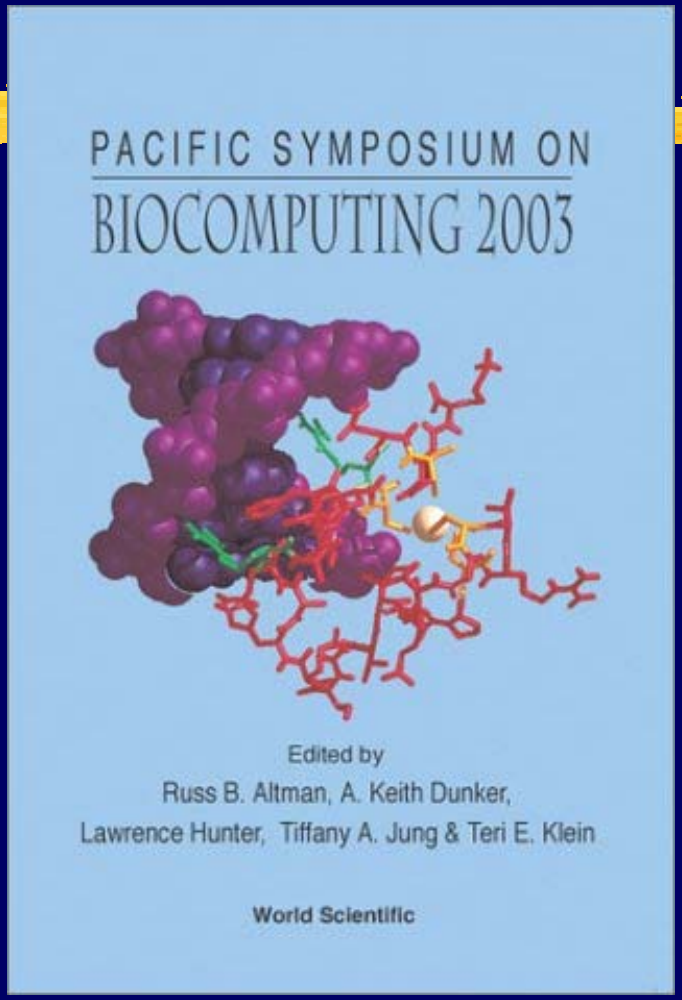
News from PSB 2003

View from 5000 km above 21?59'N 159?21'W




KAUA'I
O'AHU
MOLOKA'I
LANA'I
MAUI
HAWAII


© Robert Zakon


PACIFIC SYMPOSIUM ON
BIOCOMPUTING 2003
Edited by
Russ B. Altman, A. Keith Dunker,
Lawrence Hunter, Tiffany A. Jung & Teri E. Klein
World Scientific

Talks

- Statistically rigorous electronic gene annotation and classification of protein data bank sequences using gene ontology terms
- A Piecewise subtractive quasi-global normalization and gene identification method gives superior results for dna-array analysis
- MULTICLASS CANCER CLASSIFICATION USING GENE EXPRESSION PROFILING AND PROBABILISTIC NEURAL NETWORKS

Statistically rigorous electronic gene annotation and classification of protein data bank sequences using gene ontology terms, Werner G.Krebs, Philip E. Bourne, UCSD

- Allows automatic extension of existing ontologies
- Needs: Cluster of genes based on info given in ontology
- P-value for correlation of cluster with ontology (modelled by hypergeometric distrib)

Ontology based classification

- Bayesian probability for fraction of genes in a cluster having a common GO term
- Third statistic gives confidence interval on Bayesian prob
- find falsely classified genes, help annotate genes, automate process
- PDB: 36000 chains, 23000 a priori classified, 4000 additional with this approach

A Piecewise subtractive quasi-global normalization and gene identification method gives superior results for DNA-array analysis, Yangdagger, Haddaddagger, Tomas, Alsaker, Papoutsakis, NWU

- Array normalization and gene identification method
  - segment entire intensity range in intervals
  - determine mean and SD of ratios for each interval using nearest neighbor nondifferentially expressed genes

Model

- Noise in microarrays:
  - random errors (scanning, spot-to-spot variation) global on array
  - systematic errors (array surface, printing, DNA prep)
- Let x* and y* be the true intensities (no random errors), so x*/y* could be used for normalization

Model

- Consider K non-differentially expressed genes closest to (x*,y*)

$$\log \lambda(x,y) = \log\left(\frac{x^*}{y^*}\right) \approx \frac{1}{K}\sum_{i=1}^{K} \log\left(\frac{x_i^*}{y_i^*}\right) = \frac{1}{K}\sum_{i=1}^{K} \log\left(\frac{x_i - \varepsilon_{x,i}}{y_i - \varepsilon_{y,i}}\right).$$

- if K is large enough

$$\begin{aligned}\log \lambda(x,y) &= E\left(\log\frac{x-\varepsilon_x}{y-\varepsilon_y}\right)\\ &= E\left(\log\left(\frac{x}{y}\right)\right) + E\left(\log\left(\left(1-\frac{\varepsilon_x}{x}\right)\Big/\left(1-\frac{\varepsilon_y}{y}\right)\right)\right),\end{aligned}$$

- normalization: log ý = log y + log λ(x,y)

Normalization

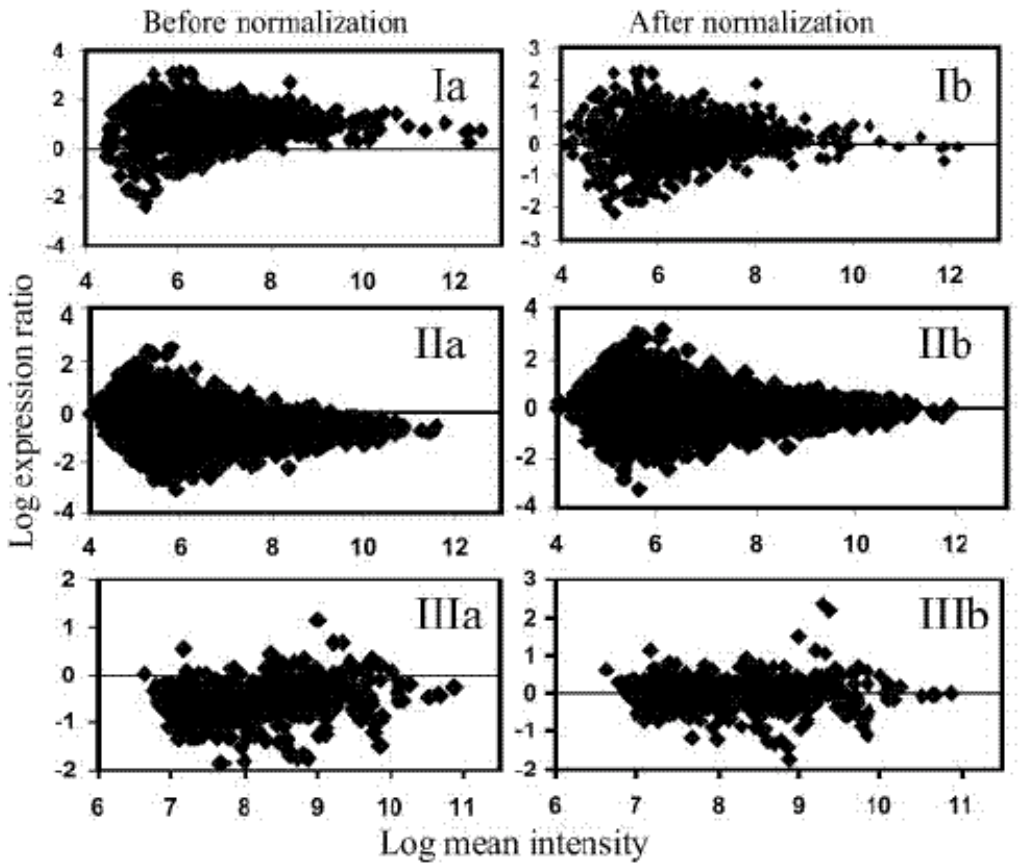


Fig. 1. Normalization results. (*a* and *b*) Original expression ratios (*a*) and normalized expression ratios (*b*) for nylon (*I*), plastic (*II*), and glass (*III*) arrays are shown.

- Random errors in 2 different arrays independent
- wide spread in low intensity

Nondifferential genes

- Remove outliers first
- use increase in stdev as criteria

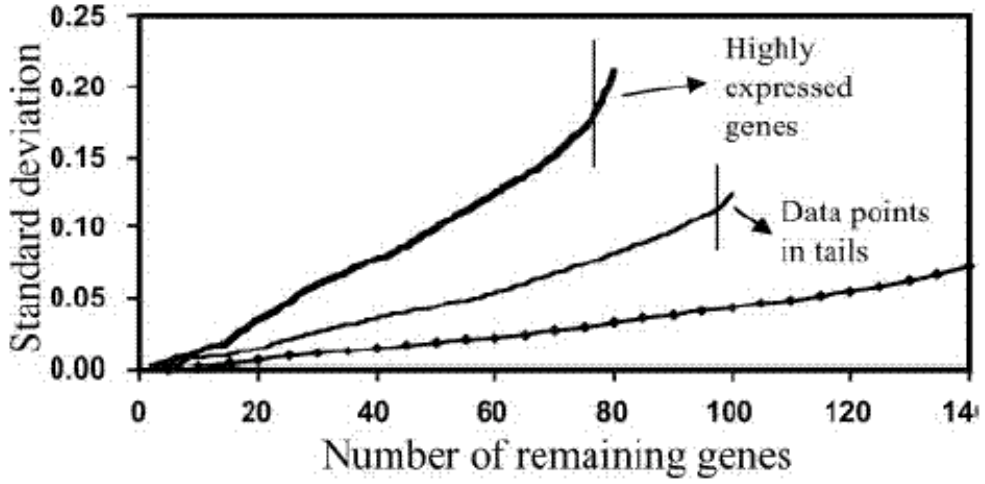


Fig. 2. Identification of strongly expressed genes using the SD profiles as a function of the number of data points from a uniform (diamonds) or normal (thin line) distribution, or an array data set (thick line). Vertical lines separate the highly expressed genes or data points in the tails of a normal distribution from the rest of genes or data points.

Normalization

- Divide whole range of log intensities into M equidistant intervals
- use K nondifferentially expressed genes around the middle of each interval to determine logarithmic expression ratio (LER) mean and its stdev (SD)
- use percentile method to estimate confidence level for each interval

Normalization quality

$$J_{norm\_error} = \frac{1}{p}\sum^{p}\left(\sum_{i=1}^{n}\left(\log\left(\frac{\bar{y}_i}{x_i}\right)\right)^2 \Big/ \sum_{i=1}^{n}\left(\log\left(\frac{y_i}{x_i}\right)\right)^2\right),$$

- n is total number of genes
- p is number of membrane pairs
- ý is normalized y

- the closer to 0 the better
- find optimal M,K for J_{norm_error} (M=20,45,25; K=250,300,200)

Comparison

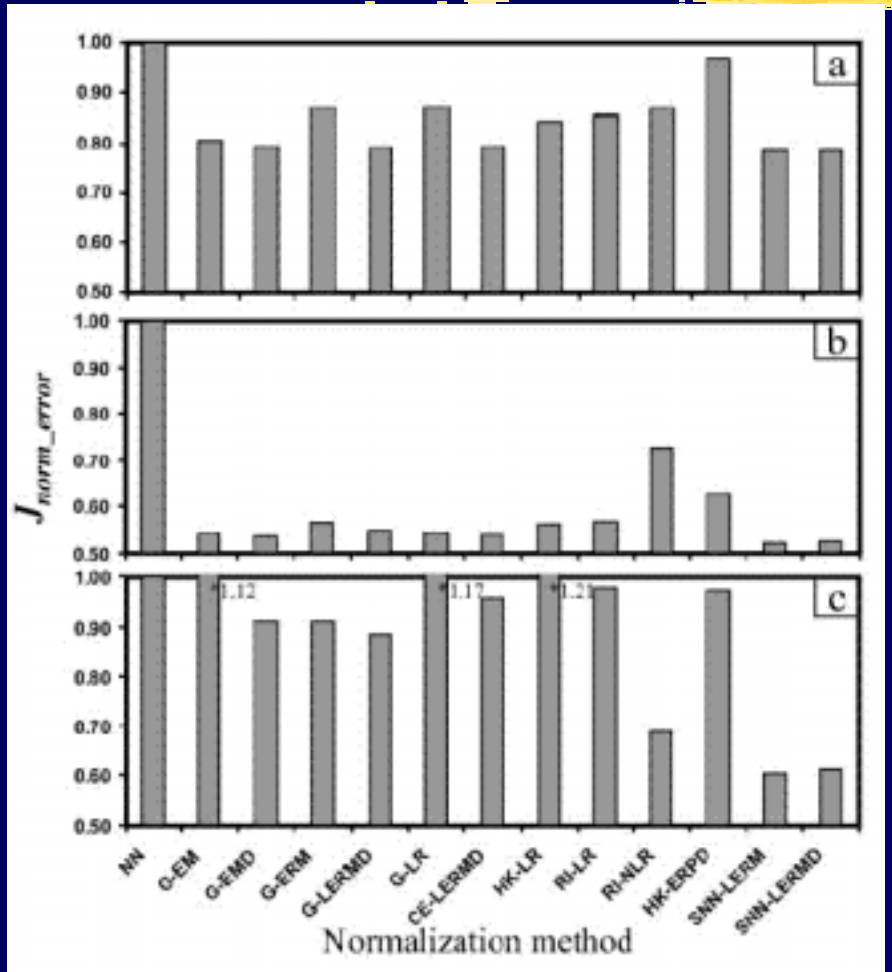


Fig. 3. Comparison of different normalization methods for 15 pairs (30 arrays) of nylon arrays (pairwise normalization) (*a*), 30 nylon arrays (all normalized to the first array) (*b*), and 22 glass arrays (*c*). NN, no normalization.

- NN: no normalization
- G-EM: global expr intsty mean
- G-EMD: global expr inty median
- G-ERM: global expr ratio mean
- G-LERMD: global log ERMD
- G-LR: global log ratio
- CE-LERMD: constantly expressed genes
- HK-LR: house keeping log ratio
- RI-LR: rank invariant log ratio
- RI-NLR: rank invariant nonlinear regression
- HK-ERPD: house keeping expr ratio prob density
- SNN-LERM: segmental nearest neighbor mean of log of expr ratio
- SNN-LERMD: segmental nearest neighbor median of log of expr ratio

Feature Selection

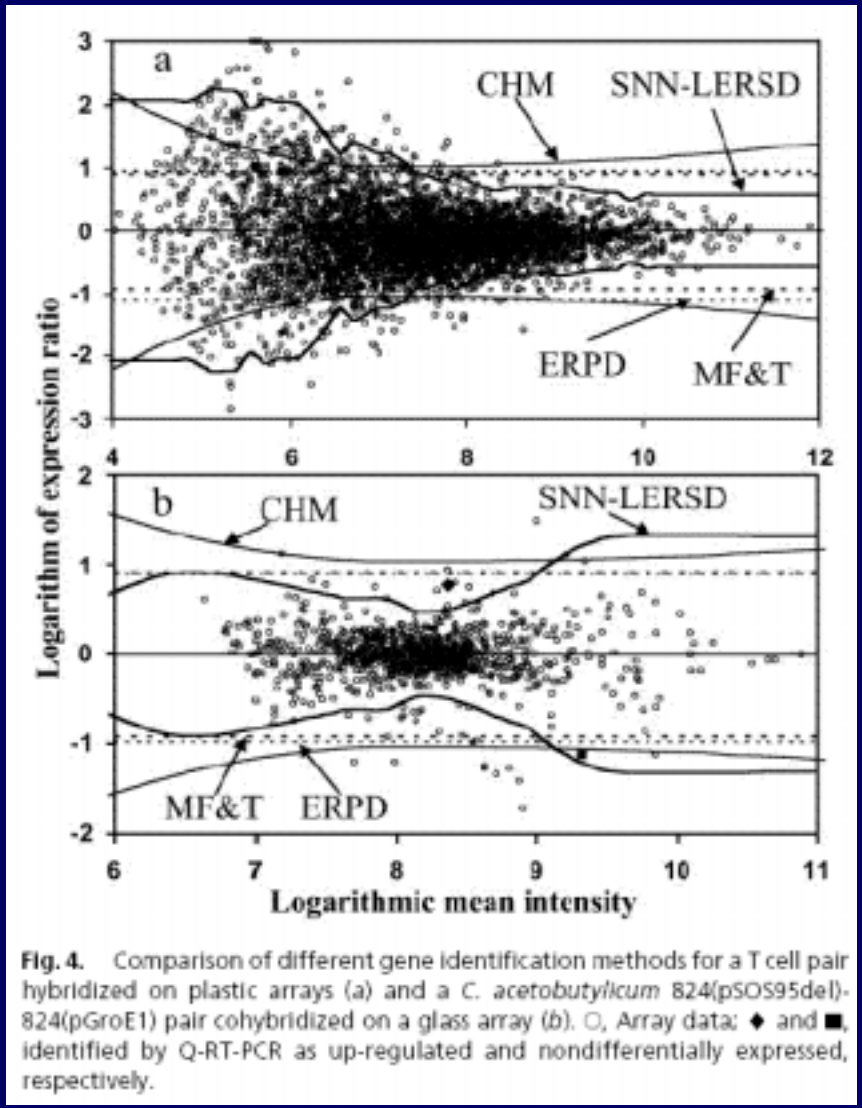


Fig. 4. Comparison of different gene identification methods for a T cell pair hybridized on plastic arrays (*a*) and a *C. acetobutylicum* 824(pSOS95del)-824(pGroE1) pair cohybridized on a glass array (*b*). ○, Array data: ♦ and ■, identified by Q-RT-PCR as up-regulated and nondifferentially expressed, respectively.

- CHM: mask countours (Netwon et al.)
- SNN-LERSD: segmental nearest neighbor log expr ratio std dev
- ERPD: expression ratio probability density
- MF&T: minimal fold change with an intensity threshold

Comparison

Result with Q-RT-PCR	No. of genes identified by array analysis using: ERPD (95%)	MF&T (Mf = 3; Th = 1,000)	MF&T (Mf = 2.2; Th = 500)	CHM (Po = 100:10)	CHM (Po = 100:5)	SNN-LERSD (95%)
Differentially expressed (n_{di} = 34)						
Differentially expressed	14	4	10	9	16	15
Nondifferentially expressed	18	30	23	24	15	18
Oppositely differentially expressed	2	0	1	1	3	1
Nondifferentially expressed (n_{nd} = 114)						
Nondifferentially expressed	99	111	105	108	76	105
Differentially expressed	22	6	11	24	55	11
$J_{iden.error}$	0.36	0.45	0.39	0.39	0.43	0.32

Abbreviations are as in Table 1.

- Assessing accuracy: megaplasmid deficient C. acetobutylicum strain M5
  - up to 178 genes knocked out due to lack of pSOL1 gene
- T cell samples with Q-RT-PCR (148 measurements)

MULTICLASS CANCER CLASSIFICATION USING GENE EXPRESSION PROFILING AND PROBABILISTIC NEURAL NETWORKS, D.P. BERRAR, C. S. DOWNES, W. DUBITZKY

- PNN: RBF neural network
  - Bayes decision strategy
  - Parzen method of density estimation
- PNN advantages:
  - model assymetric classification FN, FP
  - confidence of decision

Building a PNN

- Bayes optimal classifier

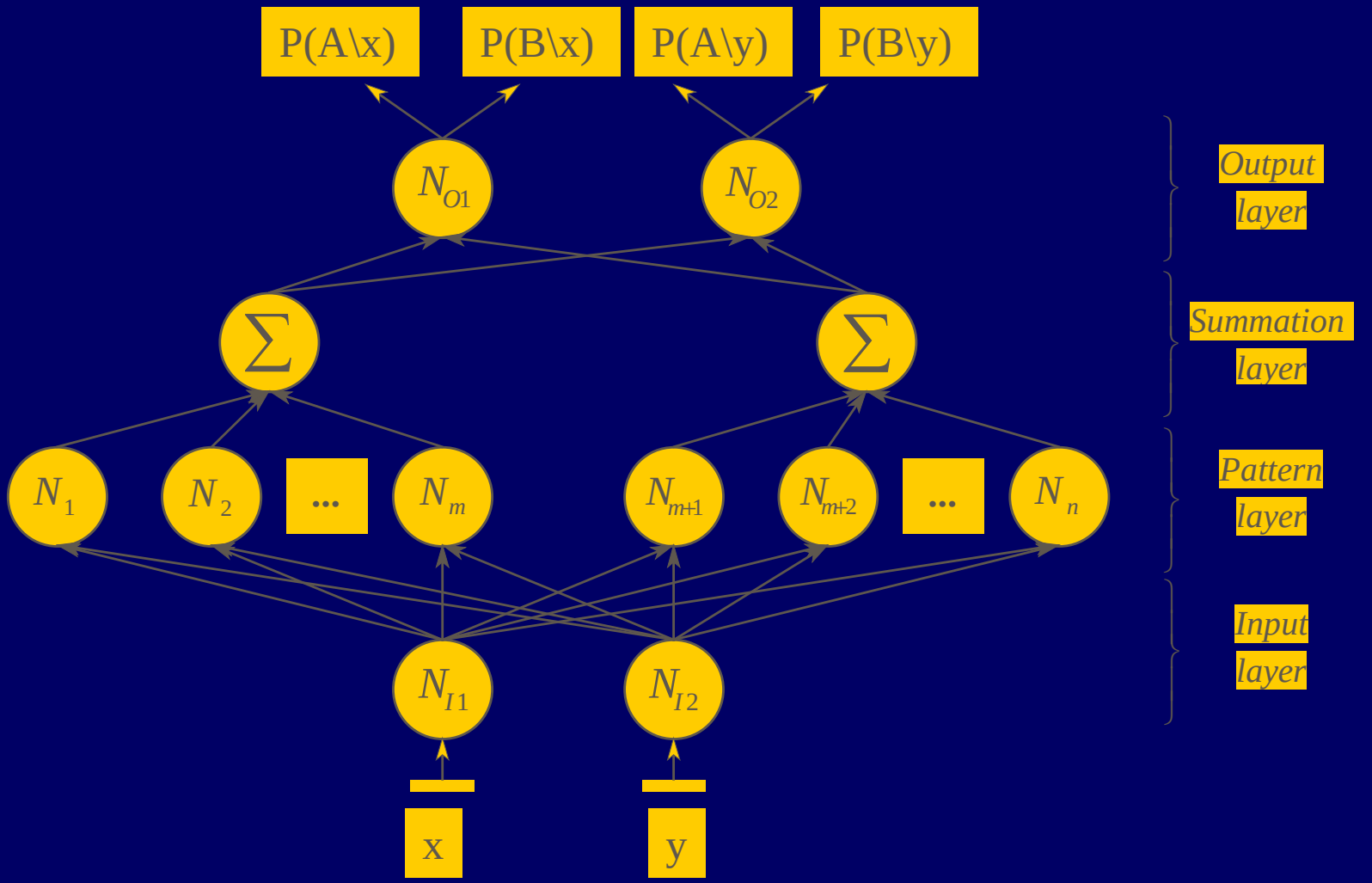
$$h_i \cdot c_i \cdot f_i(x) > h_j \cdot c_j \cdot f_j(x)$$

- Estimator for density function

$$\hat{f}_j(\vec{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^{dim} \sigma^{dim} m_j} \sum_{i=1}^{m_j} exp\left(-\frac{(\vec{x}-\vec{x}_{ij})^T \cdot (\vec{x}-\vec{x}_{ij})}{2\sigma^2}\right) \quad (1)$$

where
- $\hat{f}_j$: estimated density for the j-th class
- $\vec{x}$: test case
- $\vec{x}_{ij}$: i-th training sample of the j-th population / class
- dim : dimensionality of $\vec{x}_{ij}$
- σ: smoothing factor
- T: transpose
- m_j : number of training cases in the j-th class

PNN example


P(A\x)
P(B\x)
P(A\y)
P(B\y)
N_{O1}
N_{O2}
Output layer
Summation layer
Pattern layer
Input layer
N_1
N_2
...
N_m
N_{m+1}
N_{m+2}
...
N_n
N_{I1}
N_{I2}
x
y

Golub

Primary class	*ALL*		*AML*					
Subclass	*B-cell*	*T-cell*	*M1*	*M2*	*M4*	*M5*	*N/a*	Σ
# of cases in training set	19	8	3	5	1	2	0	38
# of cases in validation set	19	1	1	5	3	0	5	34

		Real class							
		M1	*M2*	*M4*	*M5*	*B-cell*	*T-cell*	*N/a*	Σ
Classification	*M1*	1	1	-	-	3	-	1	6
	M2	-	4	1	-	-	-	-	5
	M4	-	-	-	-	-	-	-	-
	M5	-	-	-	-	-	-	2	2
	B-cell	-	-	2	-	16	1	2	21
	T-cell	-	-	-	-	-	-	-	-
	N/a	-	-	-	-	-	-	-	-
	Σ	1	5	3	-	19	1	5	34
	sensitivity	*1.00*	*0.80*	*0.00*	-	*0.84*	*0.00*	*0.00*	
	specificity	*0.85*	*0.97*	*1.00*	*0.94*	*0.67*	*1.00*	*1.00*	

Classification Performance

- Instead of plain accuracy also consider prevalence

$$lift(c_i) = \begin{cases} 0, \text{ if class } c_i \text{ is not predicted} \\ \dfrac{p\left(act(x_j) = c_i \mid prd(x_j) = c_i\right)}{p\left(act(x_j) = c_i\right)} \quad otherwise \end{cases}$$

$$total\ lift = \frac{1}{m} \cdot \sum_{i=1}^{m} lift(c_i)$$

Comparison

- PNN on all data, reduced (PCA)
- PNN vs C5.0 vs. multi-layer feedforward perceptron with back propagation network

- PCA with 23 principal components (>75% variance explained)

NCI60

- 60 cell lines, 1405 genes for 9 cancer classes, Scherf, Weinstein et al

		Class lift of PNN		*Class lift of C5.0*		*Class lift of MLP*	
Class	*Maximum lift*	*All data*	*23 p.c.*	*All data*	*23 p.c.*	*All data*	*23 p.c.*
CNS	10.00	8.33	8.33	1.67	8.33	0.00	2.00
BR	7.50	4.17	3.75	2.14	3.75	1.67	1.25
RE	7.50	5.25	5.83	1.67	3.21	0.00	1.89
LC	6.67	4.17	5.56	2.50	1.03	0.00	1.82
ME	7.50	6.56	6.56	3.75	5.63	1.07	3.75
PR	30.00	0.00	0.00	0.00	0.00	0.00	0.00
OV	10.00	8.00	8.33	0.00	5.56	0.00	1.67
CO	8.57	6.67	7.50	3.43	6.43	1.43	3.43
LE	10.00	10.00	10.00	10.00	8.57	1.00	6.67
Total lift	10.86	6.01	6.21	2.80	4.72	0.57	2.50

- missing values by mean in similar grps

PNN summary

- Artifical neural networks disadv:
  - no precise interpretation of network
  - heuristic parameter estimation
- Probabilistic neural networks disadv:
  - all training data left in memory
  - optimal smoothing parameter needed


PSB 2004
January 6-10, 2004
The Fairmont Orchid, Big Island of Hawaii