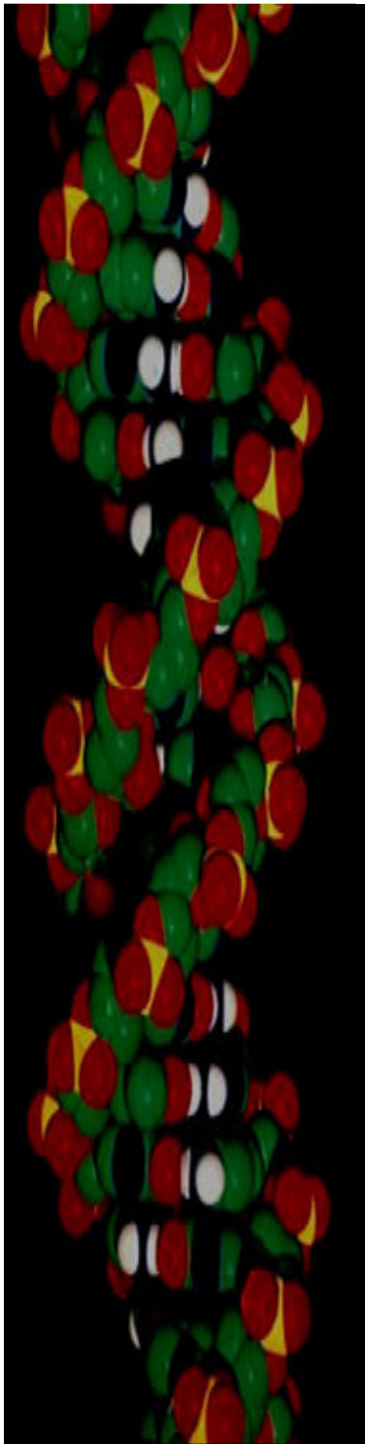


U-curve Search for Biological States Characterization and Genetic Network Design

Marcelo Ris – *Universidade de São Paulo – Instituto de Matemática e Estatística*
Junior Barrera – *Universidade de São Paulo – Instituto de Matemática e Estatística*
Helena Brentani - *Hospital do Câncer, Fundação Antônio Prudente*

FAPESP CAPES GENSIPS2005 NATIONAL INSTITUTES OF HEALTH USP - Universidade de São Paulo IME - Instituto de Matemática e Estatística

Outline

- Introduction
- Feature selection problem
- U-curve search algorithm
- Characterization of biological states
- Genetic network design
- Application

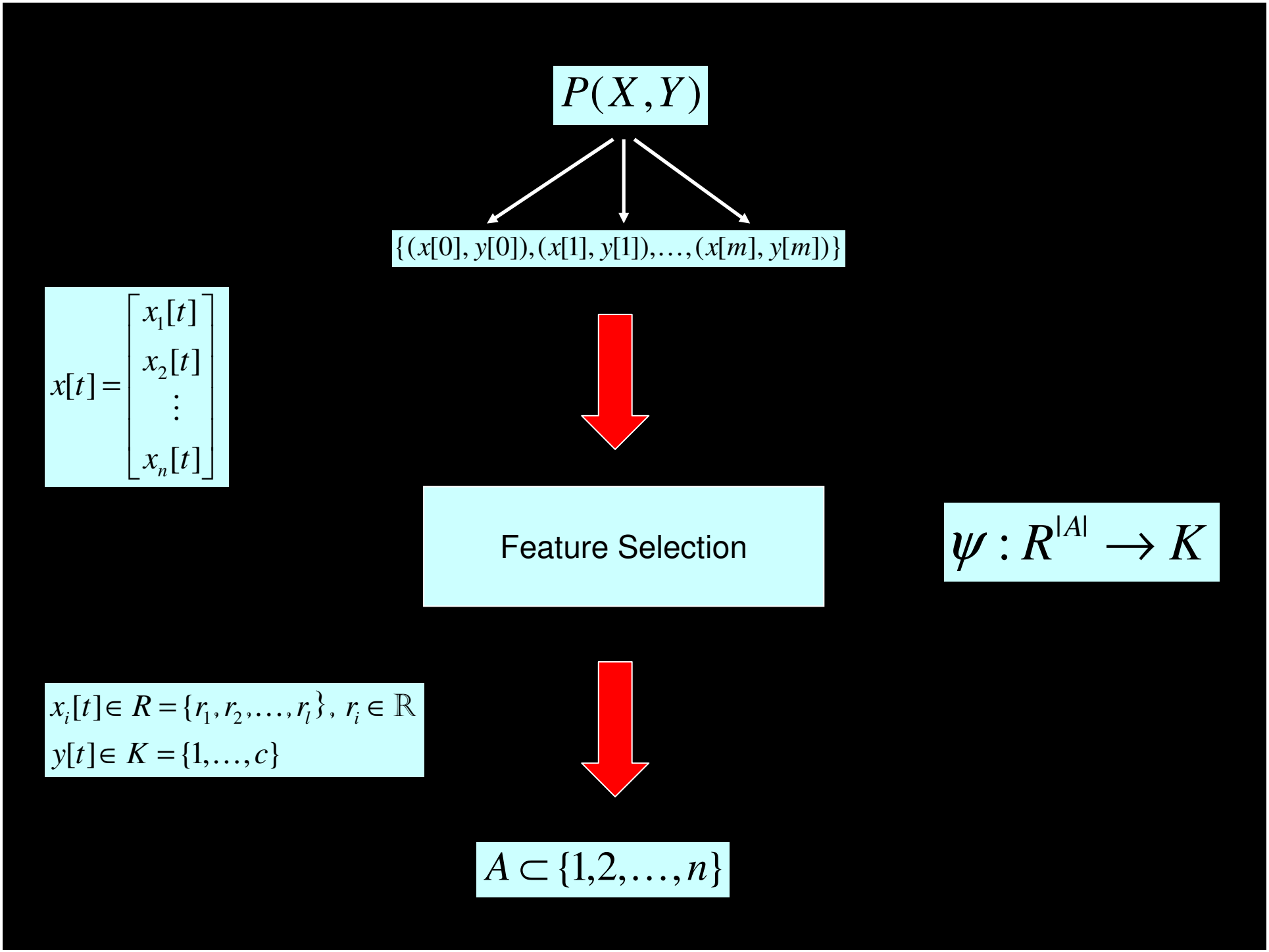
- Introduction
- Feature selection problem
- U-curve search algorithm
- Characterization of biological states
- Genetic network design
- Application

- Biological Problems
  - P1. Biological states characterization
  - P2. Genetic Network Design

- Gene expression data
  - P1. States samples
  - P2. Time-course samples

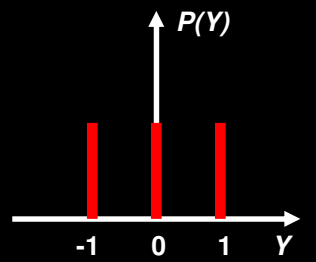
- Mathematical approach
  - Feature Selection Problem
  - State of the art: heuristic optimizations
  - U-curve algorithm

- Introduction
- Feature selection problem
- U-curve search algorithm
- Characterization of biological states
- Genetic network design
- Application


$P(X,Y)$
$\{(x[0], y[0]), (x[1], y[1]), \ldots, (x[m], y[m])\}$
$x[t] = \begin{bmatrix} x_1[t] \\ x_2[t] \\ \vdots \\ x_n[t] \end{bmatrix}$
Feature Selection
$\psi : R^{|A|} \rightarrow K$
$x_i[t] \in R = \{r_1, r_2, \ldots, r_l\},\ r_i \in \mathbb{R}$
$y[t] \in K = \{1, \ldots, c\}$
$A \subset \{1,2,\ldots,n\}$

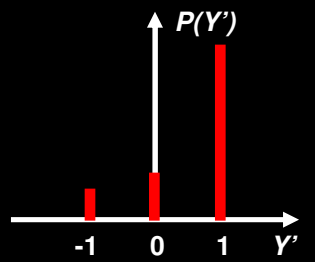
Y Distribution

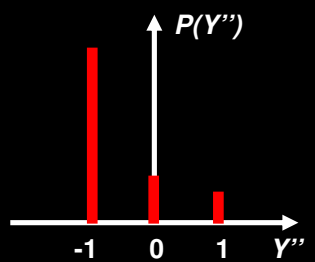
$$P:\{-1,0,1\}\rightarrow[0,1]$$
$$\sum_{y\in\{-1,0,1\}} P(y)=1$$



Y Entropy

$$H(Y)=-\sum_{y\in\{-1,0,1\}} P(y)\log P(y)$$
$$H(Y)>H(Y')$$
$$H(Y')=H(Y'')$$





Mutual Information

$$I(X,Y)=H(Y)-H(Y\mid X)\geq 0$$

Mean Conditional Entropy

$$E[H(Y \mid X)] = \sum_{x \in \{-1,0,1\}} P_X(x) \sum_{y \in \{-1,0,1\}} P_{Y|X}(y \mid x) \log P_{Y|X}(y \mid x)$$

Estimation

$$\widehat{E}[H(Y \mid X)] = \sum_{x \in \{-1,0,1\}} \widehat{P}_X(x) \sum_{y \in \{-1,0,1\}} \widehat{P}_{Y|X}(y \mid x) \log \widehat{P}_{Y|X}(y \mid x)$$

Mean Mutual Information

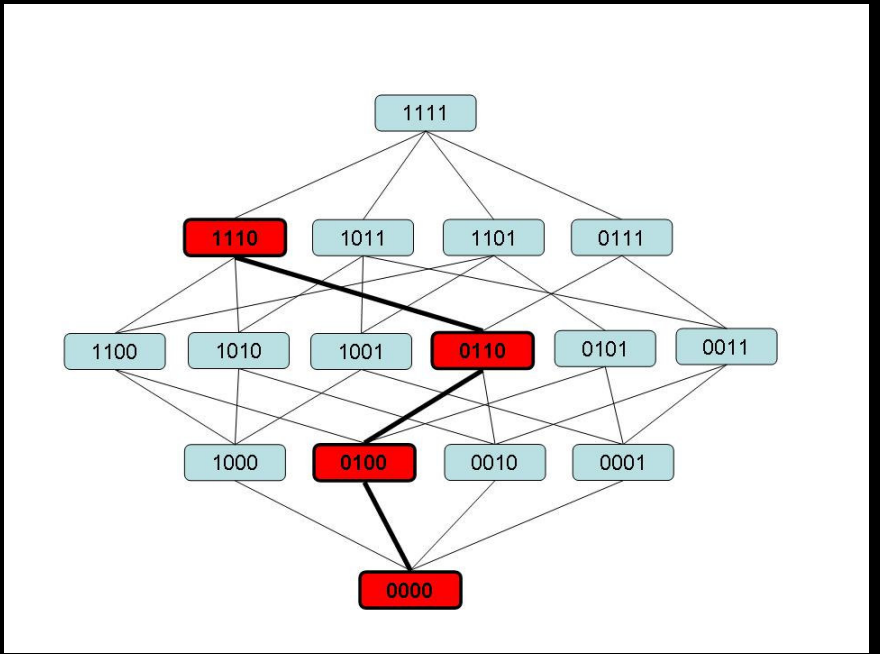
$$E[I(X,Y)] = H(Y) - E[H(Y \mid X)]$$

Estimation

$$\widehat{E}[I(X,Y)] = \widehat{H}(Y) - \widehat{E}[H(Y \mid X)]$$

- Problem
  - find the subset ***A*** that optimizes the cost function
  - Ex: mean conditional entropy minimization (cost function)
  - Exponential

- Search Space
  - Complete boolean lattice of order ***n***
  - Each node represents a possible candidate ***A***
  - Cost function: estimated for each node
  - Find the node with the minimum cost

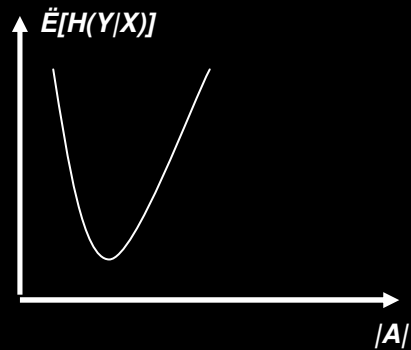


Boolean Lattice of order 4
4-element chain is emphasized

- Heuristics: SFS, SFFS
  - Incremental
  - Does not search all the candidates space
  - Could not obtain the “best” result
    - Ex: 2 elements alone turns the result worse, but together improves it a lot

- Introduction
- Feature selection problem
- U-curve search algorithm
- Characterization of biological states
- Genetic network design
- Application

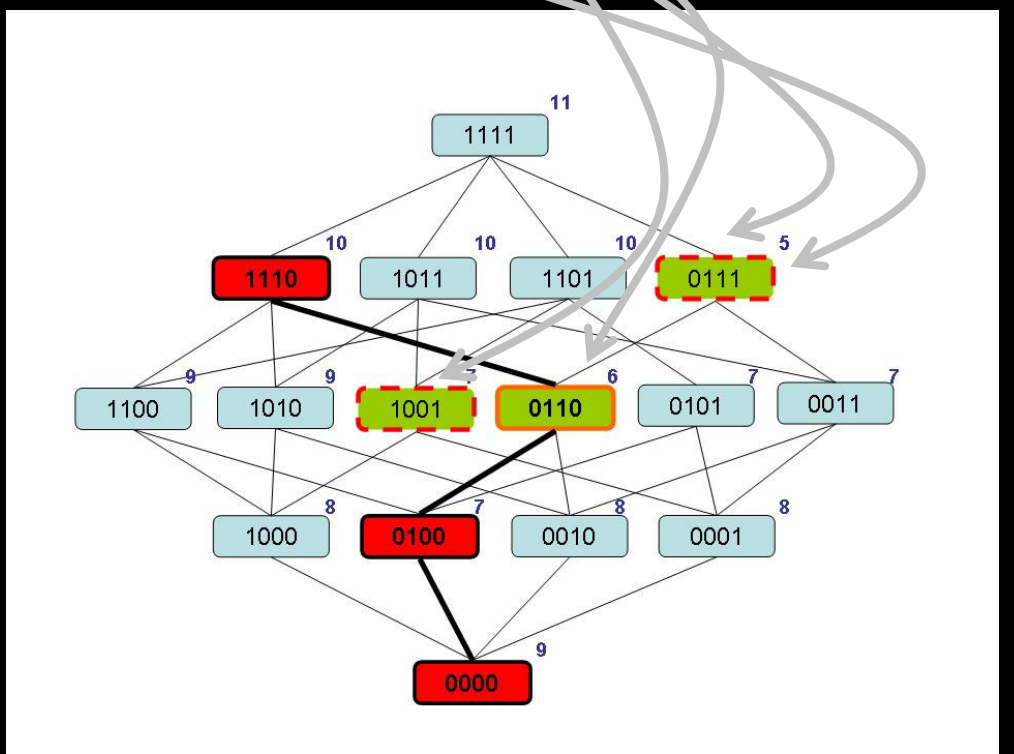
• U-curve property of $\hat{E}[H(Y|X)]$



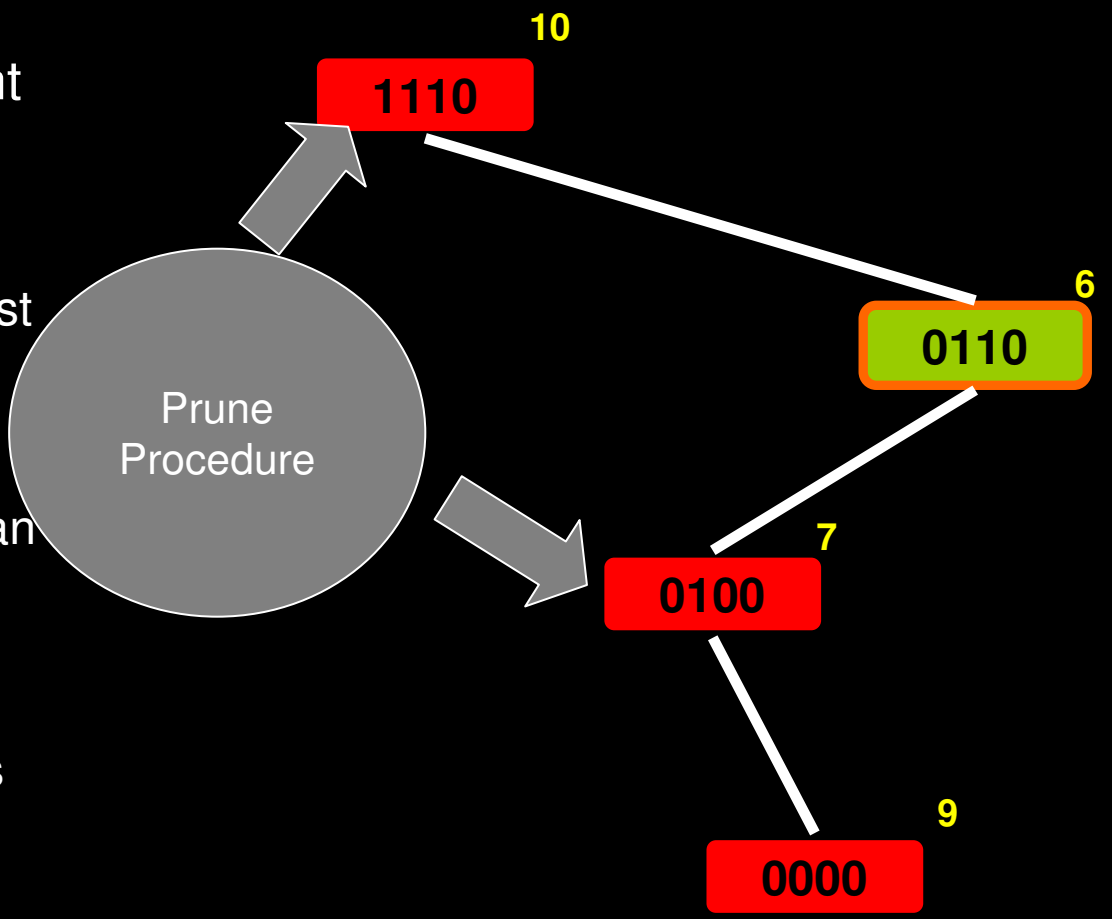
- For a fixed number of samples
- For any chain of the search space
- $\hat{E}[H(Y|X)]$ forms an U-curve

- Why ?
- Estimation composed by:
  - Real measure – decreases from $H(Y)$ to the real value $E[H(Y|X)]$
  - Estimation error – increases as more attributes are added to X

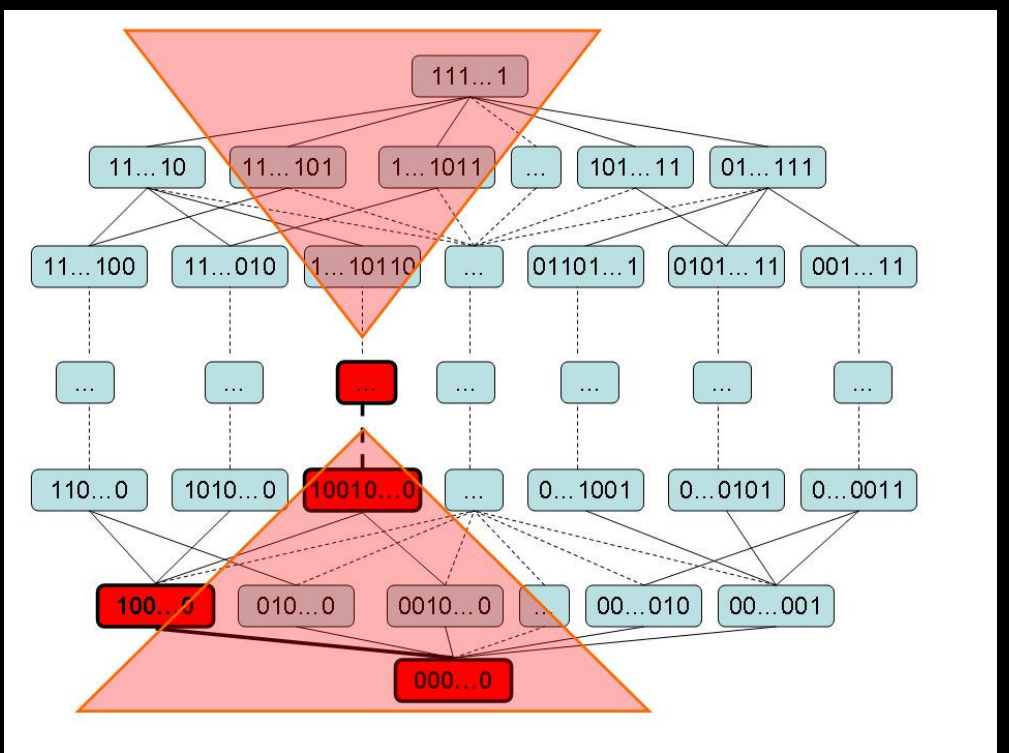
- Features of the algorithm
  - *Branch-and-Bound:* go through the whole space without having to visit all the candidates
  - Stochastic
  - Some definitions:
    - U-cost Boolean Lattice
    - Local minimum
    - Exhausted minimum
    - Global minimum



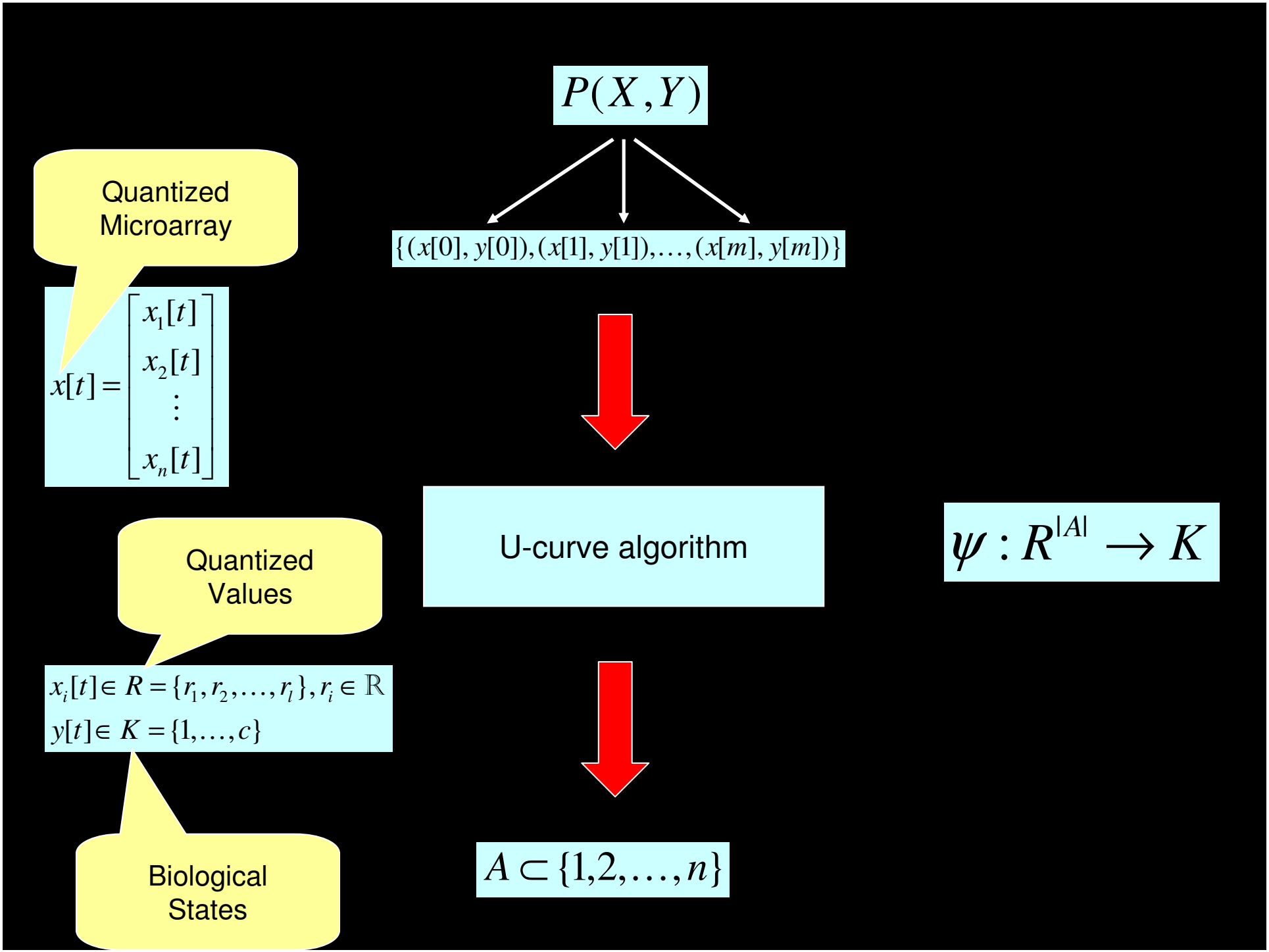
- Search space characterized by:
  - Upper Bound List
  - Lower Bound List
- An element is *reachable* if there is a chain from an upper or lower list element
- At each step:
  - Select with some probability a beginning list
  - Select an aleatory element from this list
  - Build a chain iteratively:
    - Inserts to the chain an aleatory reachable adjacent to the last one
    - Stop, when the cost of the last element is greater than the last one



- Additional Procedures
  - Minimum exhausting
    - Avoid more than one visit to the same candidate
    - Using a stack
  - Pruning elements from an element ***E***
    - Upper bound list – remove elements *U's* that contain *E*, and inserts elemets reachable from *U* that not contain *E*
    - Lower bound list – remove elements *L's* that are contained in *E*, and inserts elemets reachable from *L* that is not contained in *E*



- Introduction
- Feature selection problem
- U-curve search algorithm
- Characterization of biological states
- Genetic network design
- Application


$P(X,Y)$
$\{(x[0], y[0]), (x[1], y[1]), \ldots, (x[m], y[m])\}$
Quantized Microarray
$x[t] = \begin{bmatrix} x_1[t] \\ x_2[t] \\ \vdots \\ x_n[t] \end{bmatrix}$
U-curve algorithm
$\psi : R^{|A|} \rightarrow K$
Quantized Values
$x_i[t] \in R = \{r_1, r_2, \ldots, r_l\}, r_i \in \mathbb{R}$
$y[t] \in K = \{1, \ldots, c\}$
Biological States
$A \subset \{1,2,\ldots,n\}$

- Introduction
- Feature selection problem
- U-curve search algorithm
- Characterization of biological states
- Genetic network design
- Application

- Dynamical Systems
  - State: vector x
  - Transition function Φ
  - $x[t+1] = \Phi(x[t])$

- Stochastic Process
  - Stochastic transition function
    - Next State – aleatory vector realization
  - Ex: Markov Chain ($\pi_{X|Y}$, π_0)
    - Time-discrete, finite-size vector, finite domain
    - Aleatory state sequence

$$\pi_0 = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ \vdots \\ p_{|R|^n} \end{bmatrix} \qquad \pi_{Y|X} = \begin{bmatrix} p_{1|1} & p_{2|1} & p_{3|1} & \cdots & p_{|R|^n|1} \\ p_{1|2} & p_{2|2} & p_{3|2} & \cdots & p_{|R|^n|2} \\ p_{1|3} & p_{2|3} & p_{3|3} & \cdots & p_{|R|^n|3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{1||R|^n} & p_{2||R|^n} & p_{3||R|^n} & \cdots & p_{|R|^n||R|^n} \end{bmatrix}$$

- Probabilistic Genetic Networks - PGN

- Markov Chain $(\pi_{Y|X}, \pi_0)$ with the following axioms :

a. $\pi_{Y|X}$ is homogeneous, $p_{y|x}$ independs on t,

b. $p_{y|x} > 0, \forall\ x, y \in R^n$

c. $\pi_{Y|X}$ é condicionally independent, that is,

$$\forall\ x, y \in R^n,\ p_{y|x} = \prod_{i=1}^{n} p(y_i \mid x),$$

d. $\pi_{Y|X}$ almost - deterministic, that is, $\forall\ x \in R^n$ e

$i \in N = \{1,..,n\}$, there is $r \in R \mid p_{y_i = r|x} \approx 1$,

e. $\forall\ x \in R^n, \forall\ i \in N$, there is a sub - space of dimension j, $j << n$, such as : $p_{y_i|x'} = p_{y_i|x}$,

where x' is the projection of x on this sub - space

- Markov Chain

$$\pi_{Y|X} = \begin{bmatrix} p_{1|1} & p_{2|1} & p_{3|1} \cdots & p_{3^n|1} \\ p_{1|2} & p_{2|2} & p_{3|2} \cdots & p_{3^n|2} \\ p_{1|3} & p_{2|3} & p_{3|3} \cdots & p_{3^n|3} \\ \vdots & \vdots & \vdots \ddots & \vdots \\ p_{1|3^n} & p_{2|3^n} & p_{3|3^n} \cdots & p_{3^n|3^n} \end{bmatrix}$$

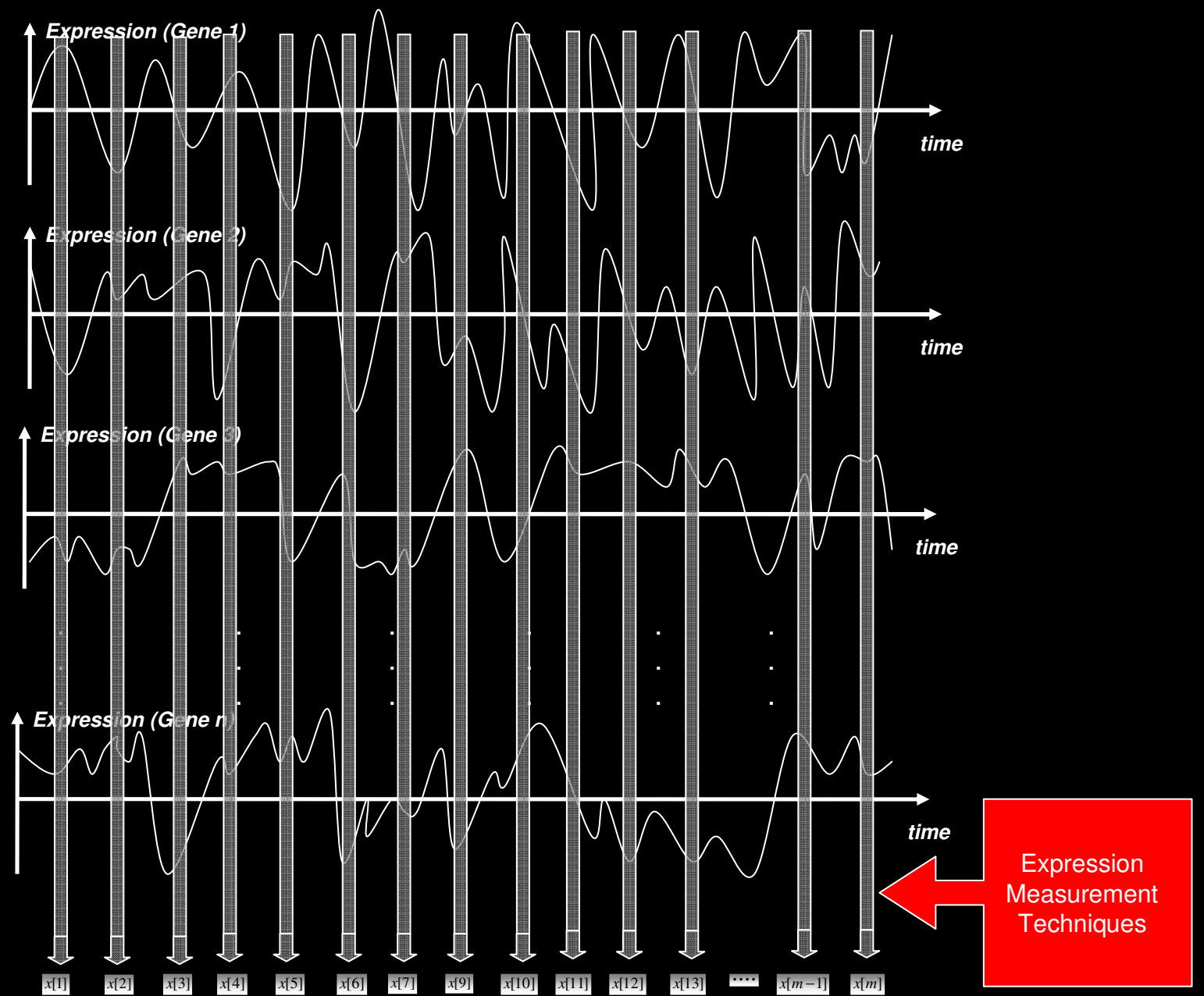
- Probabilistic Genetic Networks - PGN

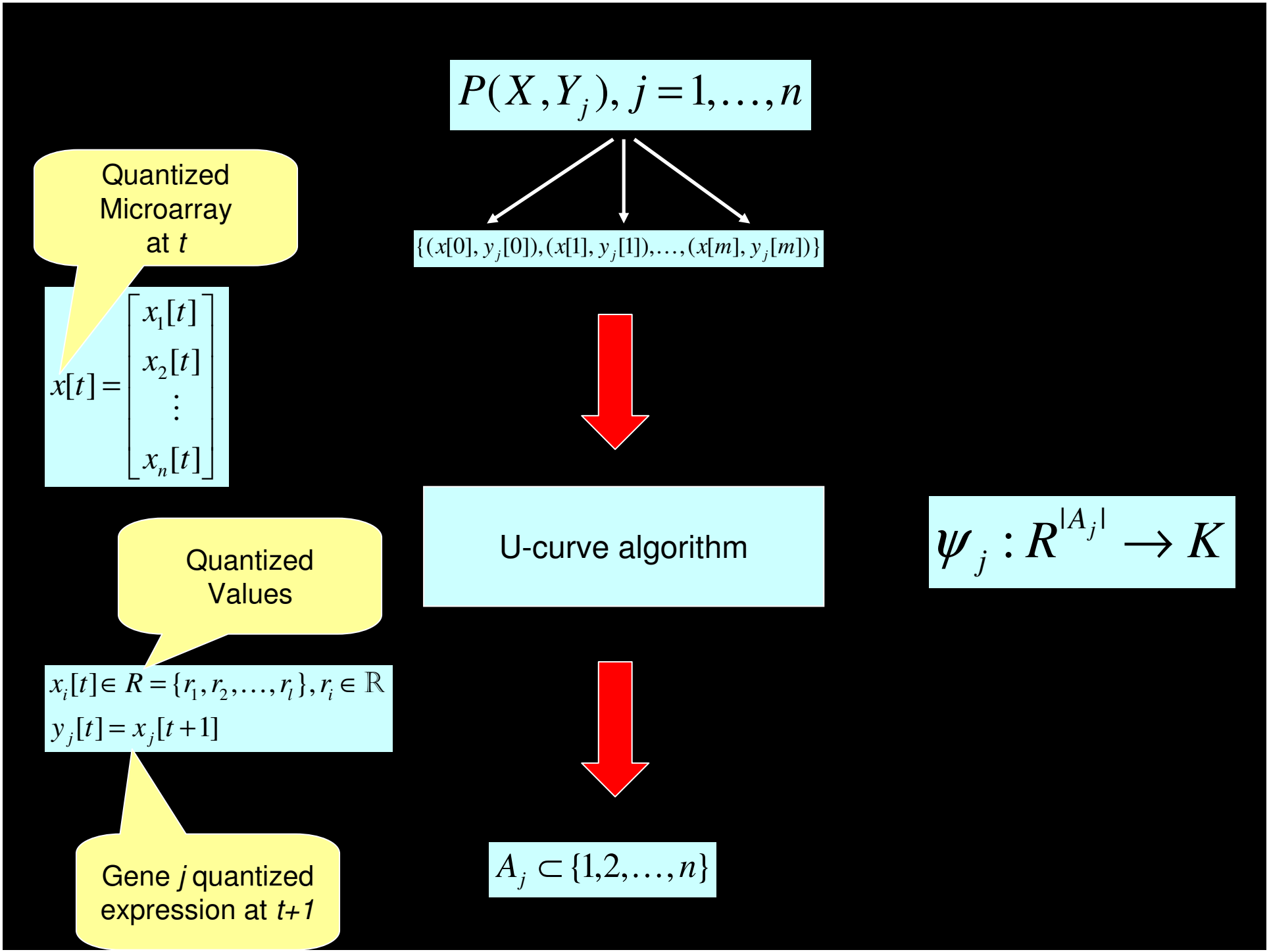
$$P_{X_1|X}, P_{X_2|X}, \ldots, P_{X_n|X}$$

$$P_{X_i|X} = \begin{bmatrix} p_{r_1|1} & p_{r_2|1} & p_{r_3|1} \cdots & p_{r_l|1} \\ p_{r_1|2} & p_{r_2|2} & p_{r_3|2} \cdots & p_{r_l|2} \\ p_{r_1|3} & p_{r_2|3} & p_{r_3|3} \cdots & p_{r_l|3} \\ \vdots & \vdots & \vdots \ddots & \vdots \\ p_{r_1|R|^n} & p_{r_2|R|^n} & p_{r_3|R|^n} \cdots & p_{r_l|R|^n} \end{bmatrix}$$

Almost Deterministic

Time-Course Gene Expression Data

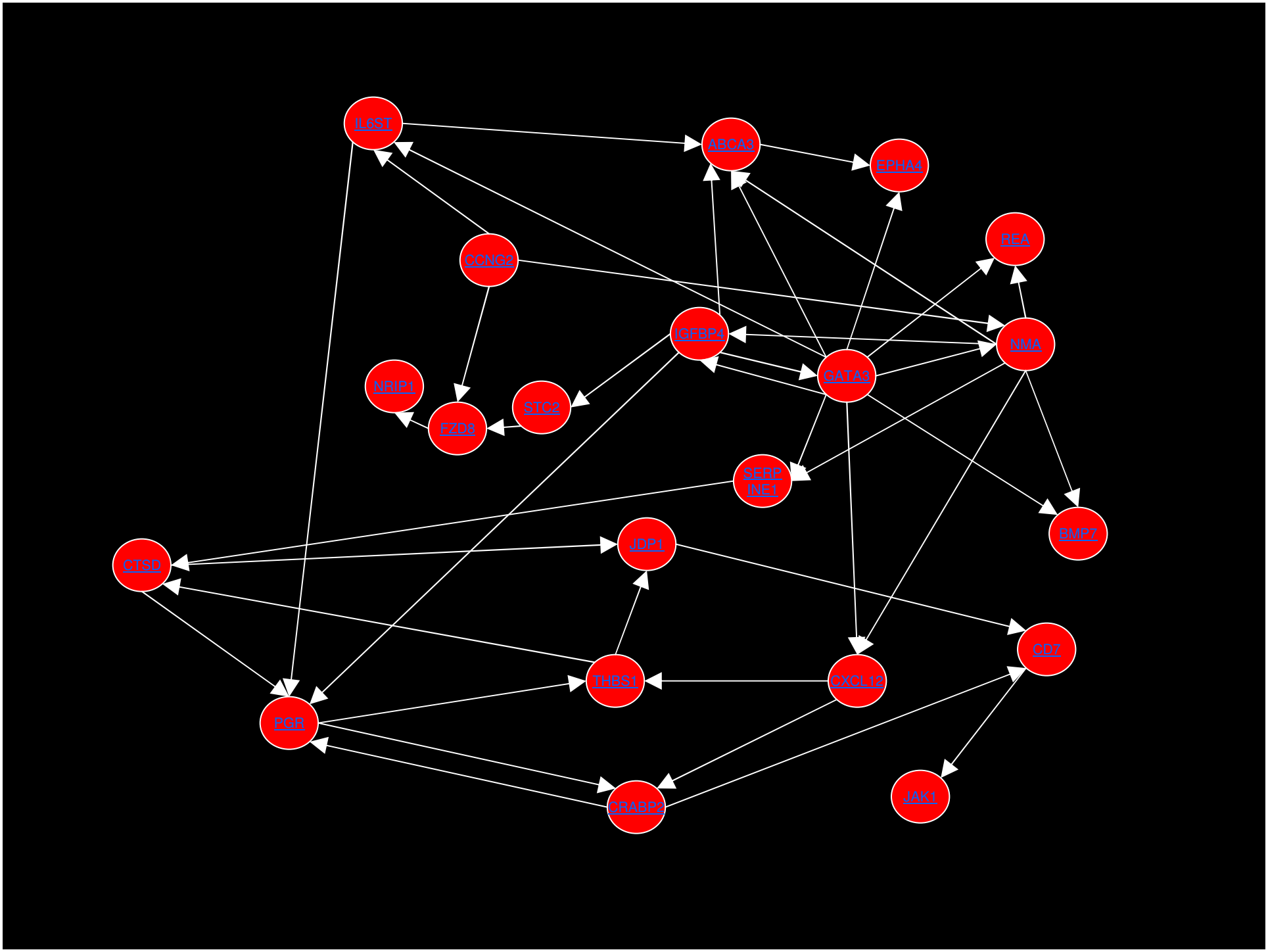


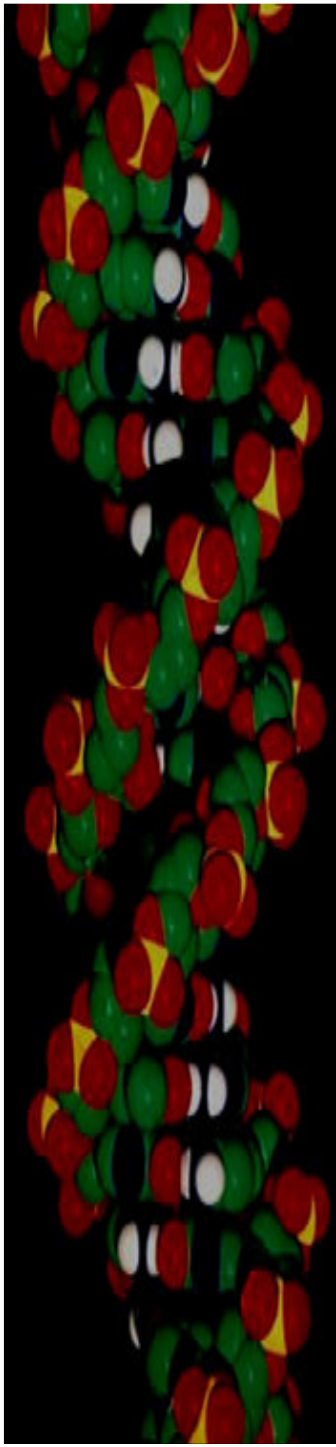

$P(X, Y_j), j = 1, \ldots, n$
$\{(x[0], y_j[0]), (x[1], y_j[1]), \ldots, (x[m], y_j[m])\}$
Quantized Microarray at t
$x[t] = \begin{bmatrix} x_1[t] \\ x_2[t] \\ \vdots \\ x_n[t] \end{bmatrix}$
U-curve algorithm
$\psi_j : R^{|A_j|} \rightarrow K$
Quantized Values
$x_i[t] \in R = \{r_1, r_2, \ldots, r_l\}, r_i \in \mathbb{R}$
$y_j[t] = x_j[t+1]$
Gene j quantized expression at $t+1$
$A_j \subset \{1, 2, \ldots, n\}$

- Introduction
- Feature selection problem
- U-curve search algorithm
- Characterization of biological states
- Genetic network design
- Application

- Application to design a estrogen regulated network
  - 21 target genes estrogen regulated
  - Time-course gene expression (Ed Liu et al.)
    - *MCF-7* cells treated with estrogen
    - 16 microarray experiments in 24 hours:
      - each hour: 8 first hours
      - each 2 hours: 16 last hours
  - Quantization (Barrera et al.)
    - 3 levels {-1, 0, 1}
  - For each target gene
    - 15x21 matrix
      - 20 first columns – gene expressions on the first 15 experiments
      - Last column – gene target expression on the last 15 experiments
  - The algorithm returns between the 20 genes the subset that best predict the target

Gene Alvo	Predictores	Entropia
THBS1 (up-regulated)	CXCL12;PGR	0.547628
IGFBP4 (up-regulated)	NMA;GATA3	0.463504
ABCA3 (up-regulated)	IL6ST;IGFBP4 or NMA;GATA3	0.431752
STC2 (up-regulated)	IGFBP4	0.484124
CXCL12 (up-regulated)	NMA;GATA3	0.547628
FZD8 (up-regulated)	CCNG2;STC2	0.431752
JDP1 (up-regulated)	CTSD;THBS1	0.315876
PGR (up-regulated)	CTSD;CRABP2 or IL6ST;IGFBP4	0.431752
IL6ST (up-regulated)	CCNC2;GATA3	0.480961
NRIP1 (up-regulated)	FZD8	0.506504
CTSD (up-regulated)	SERPINE1;THBS1	0.449209
JAK1 (up-regulated)	CD7	0.383581
SERPINE1 (down-regulated)	NMA;GATA3	0.431752
EPHA4 (down-regulated)	CCNG2;STC2	0.431752
CCNG2 (down-regulated)	ABCA3;GATA3	0.534371
CD7 (down-regulated)	CRABP2;JDP1	0.547628
REA (down-regulated)	NMA;GATA3	0.347628
CRABP2 (down-regulated)	CXCL12;PGR	0.663504
BMP7 (down-regulated)	NMA;GATA3	0.431752
NMA (down-regulated)	CCNG2;GATA3	0.365085
GATA3 (down-regulated)	IGFBP4	0.484124


IL6ST
ABCA3
EPHA4
REA
CCNG2
NMA
IGFBP4
GATA3
NRIP1
STC2
FZD8
SERPINE1
BMP7
JDP1
CTSD
CD7
THBS1
CXCL12
PGR
CRABP2
JAK1



Thanks!!!