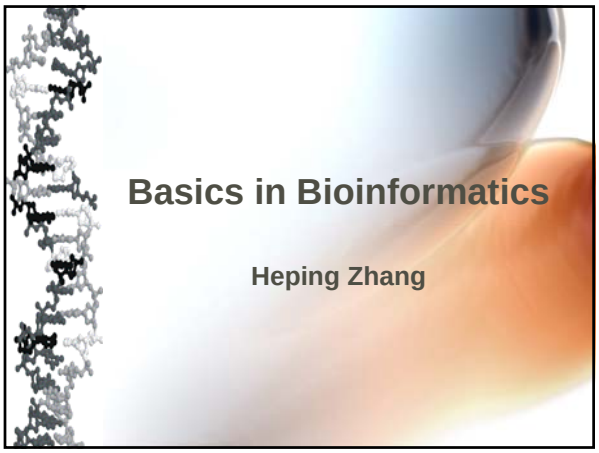


Basics in Bioinformatics

Heping Zhang

What is Bioinformatics?

The use of techniques, including applied mathematics, informatics, statistics, computer science, artificial intelligence, chemistry, and biochemistry, to solve biological problems usually on the molecular level.

Important Topics in Bioinformatics

- Sequence alignment
- Sequencing and sequence assembly
- Microarray data analysis

Sequence Alignment

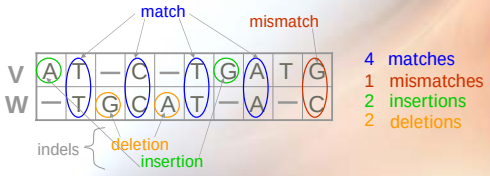
- A sequence alignment is a way of arranging the primary sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences
- Finding sequence similarities with genes of known function is a common approach to infer a newly sequenced gene's function
- In 1984 Russell Doolittle and colleagues found similarities between cancer-causing gene and normal growth factor (PDGF) gene

Alignment of Two Sequences

Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that residues with identical or similar characters are aligned in successive columns.

V = ATCTGATG
W = TGCATAC

n = 8
m = 7



4 matches
1 mismatches
2 insertions
2 deletions

Alignment : 2 * ***k*** matrix (***k*** > ***m***, ***n***)

Global Alignment vs. Local Alignment

- Calculating a **global alignment** is a form of global optimization that "forces" the alignment to span the entire length of all query sequences.
- **Local alignments** identify regions of similarity within long sequences that are often widely divergent overall.

Needleman and Wunsch Algorithm

Step 1: Dot Matrix

	A	B	C	N	J	R	Q	C	L	C	R	P	M
A	1												
J					1								
C			1					1		1			
J					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
B		1											
P												1	

Needleman and Wunsch Algorithm

Step 2: Dynamic Programming

	A	B	C	N	J	R	Q	C	L	C	R	P	M
A	1												
J					1								
C			1					1		1			
J					1								
N				1									
R						1	4	3	3	2	2	0	0
C			1			3	3	4	3	3	1	0	0
K						3	3	3	3	2	1	0	0
C			1			2	2	3	2	3	1	0	0
R						2	1	1	1	1	2	0	0
B		1				1	1	1	1	1	1	0	0
P						0	0	0	0	0	0	1	0

$$M_{ij} = M_{ij} + \max_{k>i,l>j}(M_{kj+1}, M_{i+1l})$$

Needleman and Wunsch Algorithm

Step 3: Trace Back

	A	B	C	N	J	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
J	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	2	0	0
J	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Start from the biggest element in upper row or Left column.

At each step move one row *and* column to the lower right and pick maximum in this row *or* column.

Needleman and Wunsch Algorithm

Results

A	B	C	N	J		R	Q	C	L	C	R		P	M
A	J	C		J	N	R		C	K	C	R	B	P	

A	B	C		N	J	R	Q	C	L	C	R		P	M
A	J	C	J	N		R		C	K	C	R	B	P	

Smith and Waterman Algorithm

Step 1: Dot Matrix

	H	E	A	G	A	W	G	H	E	E
P										
A										
W										
H										
E										
A										
E										

Scoring system
- –Match: 1
- –Mismatch: -1/3
- –Gap penalty: -(1+1/3*k)

Smith and Waterman Algorithm

Step 2: Dynamic Programming

Three possible ways to end an alignment at current cell:
- extend from last stage adding one base from both sequences
- sliding along first sequence causing gap(s) on second sequence
- sliding along second sequence causing gap(s) on first sequence

	H	E	A	G	A	W	G	H	E	E
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
W	0.00	0.00	0.00	0.00	0.00	2.00	0.67	0.33	0.00	0.00
H	0.00	0.00	0.00	0.00	0.00	0.67	1.67	?		
E										
A										
E										

$$H_{ij}=\max\{H_{i-1,\,j-1}+s(a_i,b_j),\ \max\{H_{i-k,j}-P_k\},\ \max\{H_{i,\,j-l}-P_l\},\ 0\}$$

Smith and Waterman Algorithm

Step 3: Trace Back

	H	E	A	G	A	W	G	H	E	E
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
W	0.00	0.00	0.00	0.00	0.00	2.00	0.67	0.33	0.00	0.00
H	0.00	0.00	0.00	0.00	0.33	0.67	1.67	1.67	0.33	0.00
E	0.00	1.00	0.00	0.00	0.00	0.33	0.33	1.33	2.67	1.33
A	0.00	0.00	2.00	0.67	1.00	0.00	0.00	0.00	1.33	2.33
E	0.00	1.00	0.67	1.67	0.33	0.67	0.00	0.00	1.00	2.33

Best local alignment:

```
AWGHE
AW-HE
```


Multiple Alignments

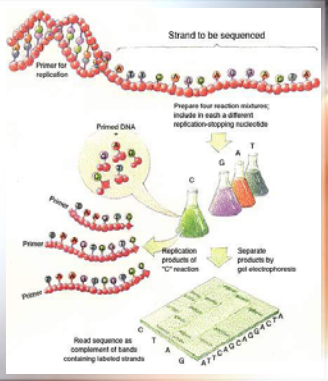
- A **multiple sequence alignment** is a sequence alignment of three or more biological sequences
- Reason to do multiple sequence alignment
  - The sequences *may* share a common origin - a common ancestor sequence. If the similarity is sufficiently convincing or if we have additional evidence for an evolutionary relationship, then we say that the sequences are homologous.
  - The sequences *may* have the same or related structure and function

DNA Sequencing

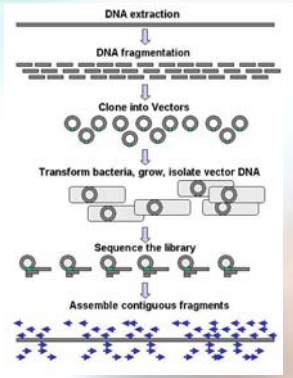
- The term **DNA sequencing** encompasses biochemical methods for determining the order of the nucleotide bases, adenine (A), guanine (G), cytosine (C), and thymine (T), in a DNA oligonucleotide.
- Techniques
  - Chain-termination methods
  - Large scale sequencing strategies
  - Sequencing by hybridization

Sanger Method

1. Start at primer (restriction site)
2. Grow DNA chain
3. Include ddNTPs
4. Stops reaction at all possible points
5. Separate products by length, using gel electrophoresis



Shotgun Sequencing



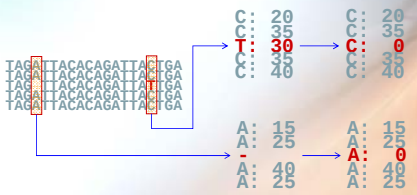
- Suitable for longer sequences
- DNA is broken up randomly into numerous small segments, which are sequenced using the chain termination method to obtain *reads*.
- Multiple overlapping reads for the target DNA are obtained by performing several rounds of this fragmentation and sequencing
- Computer programs then use the overlapping ends of different reads to assemble them into a contiguous sequence.

Fragment Assembly

- Assemble individual short fragments (reads) into a single genomic sequence
- The "**OLC Framework**"
  - **Overlap**: find all the overlaps between the reads that satisfy certain quality criteria.
  - **Layout**: given the set of overlap relationships between the reads, determine a consistent layout of the reads, i.e., find a consistent tiling of all the reads that preserves most of the overlap constraints
  - **Consensus:** given a tiling of reads determined in the layout stage, determine the most likely DNA sequence (the consensus sequence) that can be explained by the tiling

Finding Overlapping Reads

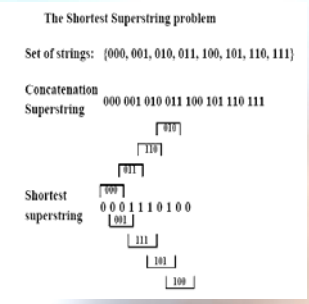
- Correct errors using multiple alignment



- Score alignments
- Accept alignments with good scores

Layout

- A variation of the shortest common superstring problem



- Problem: Given a set of strings, find a shortest string that contains all of them
- Complexity: NP-complete

SSP to TSP (Traveling Salesman Problem)

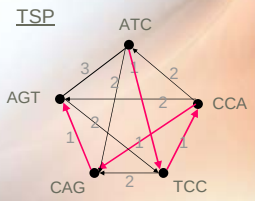
Create an overlap graph *G* where each vertex represents a string. Edge between string s_1 and s_2 will have weight equal to the length of s_1 minus the overlap of s_1 with s_2. (Edge weight is not symmetric) The path visiting all the vertices of minimum total weight defines the shortest common superstring.

S = { ATC, CCA, CAG, TCC, AGT }

SSP

AGT
CCA
ATC
ATCCAGT
TCC
CAG

TSP



ATCCAGT

Derive Consensus Sequence

Derive each consensus base by weighted voting

Sequencing of the Human Genome

- Aims
  - Sequence the entire 3 billion DNA bases
  - Dissect the code of estimated 25,000 genes that determine the physical characters of the human body

The Human Genome Project

- Funded by the National Institutes of Health in the United States, and the UK charity, the Wellcome Trust, and numerous other groups from around the world
  - Total cost $3 billion
  - Started in 1990
- The genome was broken into smaller pieces
  - Approximately 150,000 base pairs in length
    These pieces are called "bacterial artificial chromosomes", or BACs, because they can be inserted into bacteria where they are copied by the bacterial DNA replication machinery
  - Each of these pieces was then sequenced separately as a small "shotgun" project and then assembled
  - The larger, 150,000 base pairs go together to create chromosomes
  - This is known as the "hierarchical shotgun" approach
    The genome is first broken into relatively large chunks, which are then mapped to chromosomes before being selected for sequencing
- Paper published in *Nature*

Celera Genomics

- A similar, privately funded quest was launched by the American researcher Craig Venter and his firm Celera Genomics
  - Total cost: $300 million
  - Started in 1998, intended to proceed at a faster pace and at a fraction of the cost of the publicly funded project.
- Used a riskier technique called whole genome shotgun sequencing

  Whole genome shotgun had been used to sequence bacterial genomes of up to six million base pairs in length, but not for anything nearly as large as the three thousand million base pair human genome
- Draft genome published at the same time as the public effort did
- Paper published in *Science*

Sequencing by Hybridization (SBH)

- **History**
  - **1988:** SBH suggested as an an alternative sequencing method. Nobody believed it would ever work

    *First microarray prototype **(1989)***

    
  - **1991:** Light directed polymer synthesis developed by Steve Fodor and colleagues.

    *First commercial DNA microarray prototype w/16,000 features **(1994)***

    
  - **1994:** Affymetrix develops first 64-kb DNA microarray

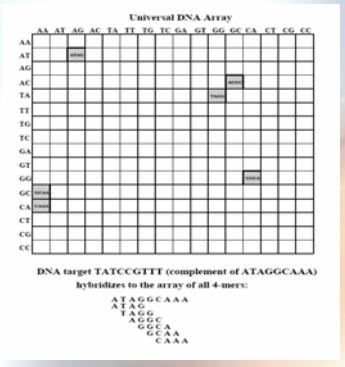
    *500,000 features per chip **(2002)***

How SBH Works

- Attach all possible DNA probes of length to a flat surface, each probe at a distinct and known location. This set of probes is called the DNA array.
- Apply a solution containing fluorescently labeled DNA fragment to the array.
- The DNA fragment hybridizes with those probes that are complementary to substrings of length l of the fragment.
- Using a spectroscopic detector, determine which probes hybridize to the DNA fragment to obtain the *l*–mer composition of the target DNA fragment.
- Apply the combinatorial algorithm (below) to reconstruct the sequence of the target DNA fragment from the *l*–mer composition.

Hybridization on DNA Array



l-mer composition

- ***Spectrum (s, l)*** - *unordered* multiset of all possible *(n – l + 1)* *l*-mers in a string *s* of length *n*
- The order of individual elements in *Spectrum (s, l)* does not matter
- For *s* = TATGGTGC all of the following are equivalent representations of *Spectrum (s, 3):*

  {TAT, ATG, TGG, GGT, GTG, TGC}

  {ATG, GGT, GTG, TAT, TGC, TGG}

  {TGG, TGC, TAT, GTG, GGT, ATG}

  We usually choose the lexicographically maximal representation as the canonical one.

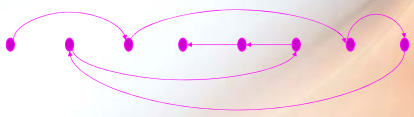
The SBH Problem

- Goal: Reconstruct a string from its *l*-mer composition
- Input: A set *S*, representing all *l*-mers from an (unknown) string *s*
- Output: String *s* such that *Spectrum (s,l) = S*

SBH: Hamiltonian Path Approach

S = { ATG AGG TGC TCC GTC GGT GCA CAG }

H ATG AGG TGC TCC GTC GGT GCA CAG



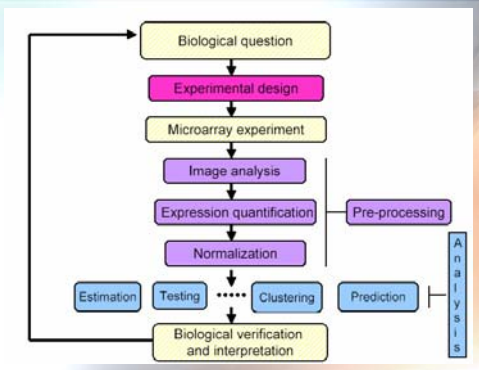
ATGCAGGTCC

Path visited every VERTEX once

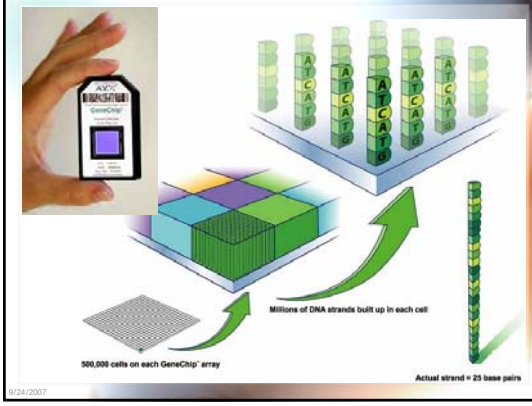
Some Difficulties with SBH

- **Fidelity of Hybridization:** difficult to detect differences between probes hybridized with perfect matches and 1 or 2 mismatches
- **Array Size:** Effect of low fidelity can be decreased with longer *l*-mers, but array size increases exponentially in *l*. Array size is limited with current technology.
- **Alternative:** SBH is not competitive due to the above reason
  - Positional SBH
  - Shotgun SBH

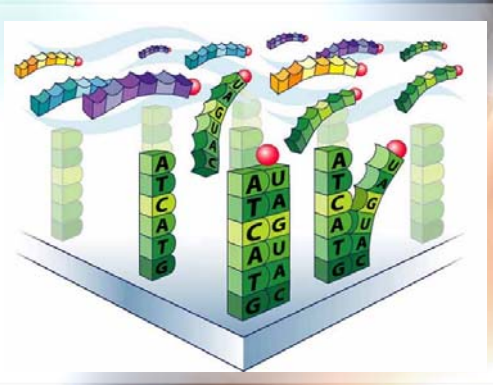
Microarray Data Analysis



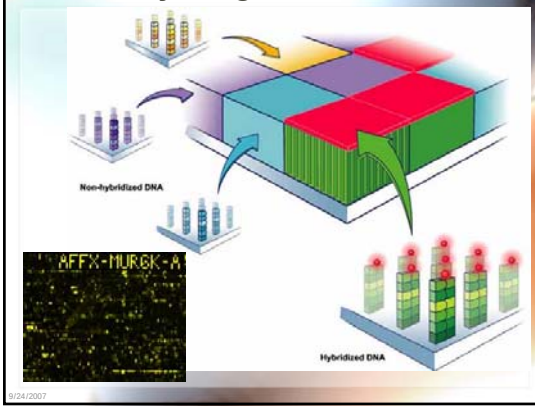
Affymetrix GeneChip



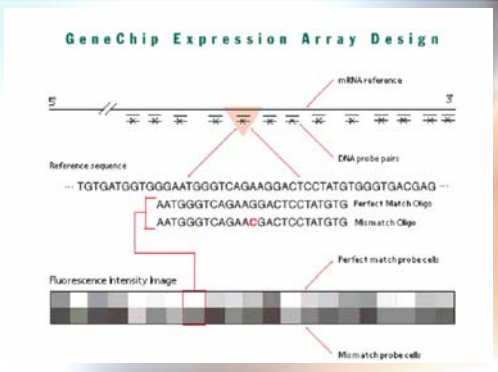
Hybridization



Microarray Images



Perfect Match (PM) vs MisMatch (MM)



http://www-microarrays.u-strasbg.fr/base.php?page=affyExpressionE.php

Quality Assessment and Control

- RNA sample quality control (e.g., RNA degradation plot)
- Array hybridization quality control
  - Probe array image inspection
  - Control genes
  - PM/MM present calls
  - Scaling Factors
- Statistical diagnostics plots

RNA Degradation

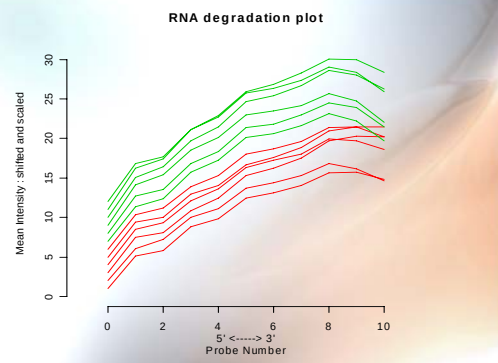
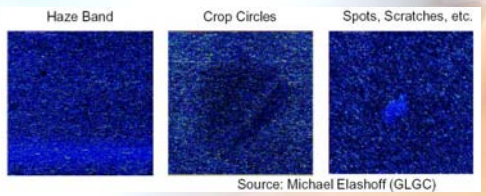
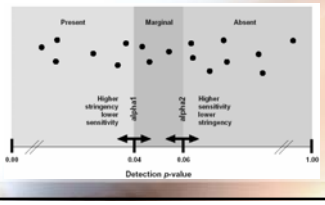


Image Inspection



Detection Calls

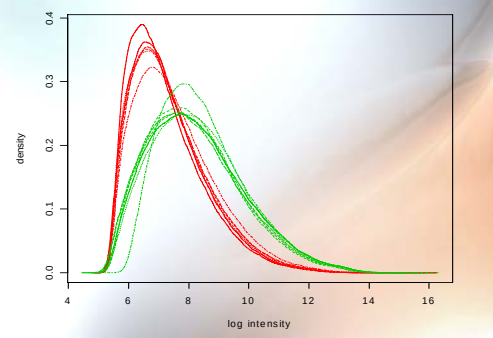
- Calculating probe-specific detection *p*-value
  - Discrimination score
    R = (PM – MM) / (PM + MM)
  - User-definable threshold τ (default 0.015)
    R ~ τ, all probe pairs for a given probe set
  - One-sided Wilcoxon's Signed Rank test assigns each probe pair a rank based on how far the probe pair Discrimination score is from τ
- Present call based on the *p*-value



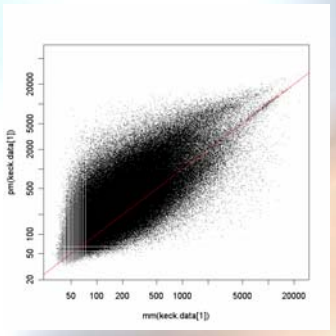
Diagnostics plots

- Image plot
- PM/MM intensity plot
- Histogram plot
- Density plot
- Scatter Plot
- Boxplot

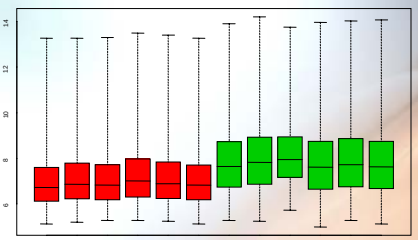
Density plot



Scatter plot



Boxplot



Boxplot of the log2-scaled PM intensity plot :

red – GE, green – KECK

Pre-analysis

- Background correction
- Normalization
- Artifacts (outliers) detection and management

Background correction

- Why: to eliminate the low levels of noise that are present on any microarray
- Different methods:
  - Local neighborhood detection
  - Negative control
  - Mismatch probes
  - Model-based: RMA and MBEI

Normalization

- Why normalization?
  - Minimizing non-biological factor
  - Reducing unwanted variance across chips
- Different methods:
  - All genes on a slide (global normalization)
  - Constantly expressed genes (Housekeeping)
  - Set of control genes

    e.g., Mouse430 chip, 1415670_at to 1415769_at
  - Rank-invariant gene
  - Quantile: giving each chip the same empirical distribution; reducing variance w/o introducing drastic bias effects

Artifacts & Outliers Detection

- Different methods apply different algorithms to detect outliers (artifacts) and take different actions;
- For example, Li-Wong model-based (dChip) method: identify extreme residuals, remove them, re-fit, …, converge.

Low-level analysis

- Generally, refer to **probe-level analysis** for Affymetrix Chips: how to extract gene expressions from probe data
- Three common approaches:
  - **MAS 5.0** (MicroArray Suite Version 5)
  - **Li-Wong model-based** analysis (dChip)
  - **RMA** (Robust multi-chip analysis)

Identifying differentially expressed genes

- Fold change detection;
- Student's *t* test;
- Mann-Whitney U test;
- Multiple comparisons adjusted P-values and confidence intervals.

Microarray High-level Analysis

- **Clustering** (unsupervised learning)
  Grouping objects based on their similarity in feature space, e.g., identifying groups of co-regulated genes;
- **Classification** (supervised learning)
  Training machine and assigning new cases into known classes, i.e., differentiating tumor / normal cells;
- **Network analysis**
  Inferring and building regulatory networks.

Clustering

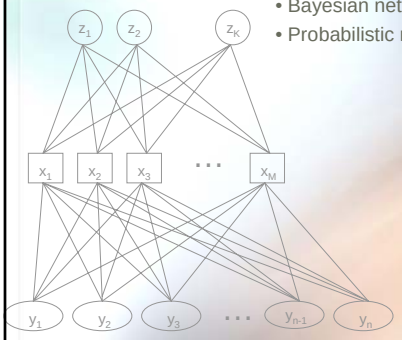
- Two categories
  - Hierarchical methods
    - Divisive: successively splitting larger clusters, top-down
    - Agglomerative: successively merging smaller clusters, bottom-up
  - Partitional methods
    - Determines all clusters at once
    - e.g., K-means, SOM, etc.

Classification

- k-nearest neighbors (KNN)
- Classification Tree
- Linear discriminant analysis (LDA)
- Bayesian Regression
- Support vector machines (SVM)
- Artificial neural networks (ANN)

Pathway analysis

- Bayesian networks
- Probabilistic rational model



References

Many slides are from www.bioalgorithms.info

•Simons, Robert W. *advanced Molecular Genetics Course*, UCLa (a00a). http://www.mimg.ucla.edu/bobs/Ca59/Presentations/Benzer.pdf

Batzoglou, S. *Computational Genomics Course*, Stanford University (a004). http://www.stanford.edu/class/csa6a/handouts.html