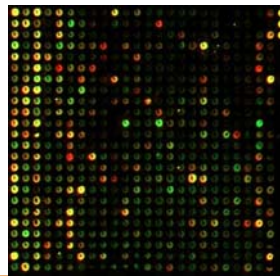
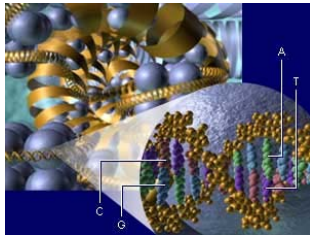


Soft Computing in Bioinformatics

James M. Keller and Mihail Popescu

*Electrical and Computer Engineering Department
Health Management and Informatics Department
University of Missouri-Columbia*

With a lot of help from our friends at MU, Univ of Utah,
Univ. West FL and Indian Statistical Institute



Outline

I. Background

1. Genes and Gene Products
 - i. Sequences
 - ii. Structure
2. Microarrays (expression, hypermethylation)
3. Taxonomies: Gene Ontology and MeSH.

II. Gene Product Similarity Measures

1. Introduction
2. Dot-Plot
3. Smith-Waterman
4. BLAST
5. GO-based measures
 - i. Jaccard, Cosine, Dice
 - ii. Fuzzy measures
 - iii. Choquet Integrals
6. Domain and Motif measures



Outline (Continued)

III. Visualization and Clustering

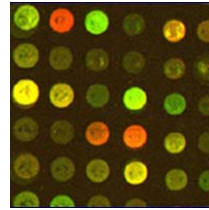
1. Hierarchical clustering
2. Visual Assessment of cluster Tendency
3. FCM and NERFCM
4. Bi-clustering (AKA co-clustering, two-way clustering)

IV. Knowledge Discovery

1. Functional annotation of gene products
2. Functional Clustering of proteins in families
3. Summarization of a set of gene products
4. Hot applications:
 - i. Methylation microarrays
 - ii. Learning biochemical networks from microarray data

I. Background

Introduction



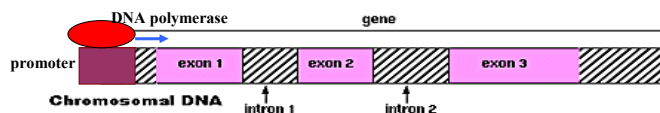
- **Principal features of gene products are**
 - the sequence and expression values following a microarray experiment
- **Sequence comparisons**
 - DNA, Amino Acids, Motifs, Secondary Structure
- **For many gene products, additional functional information comes from**
 - the set of Gene Ontology (GO) annotations and
 - the set of journal abstracts related to the gene (MeSH annotations)
- **For these genes, it is reasonable to include similarity measures based on these terms**

05/22/2005

5



I.1. Gene Product Sequences



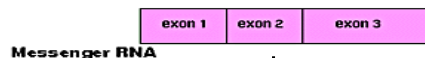
➤ DNA=sequence of nucleotides {A,C,T,G,(N-any)}

Transcription (RNA synthesis)



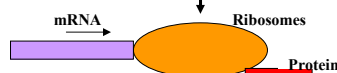
➤ RNA= sequence of nucleotides {A,C,U,G,(N-any)}

RNA Splicing



➤ Protein=sequence of 20 amino acids

Translation



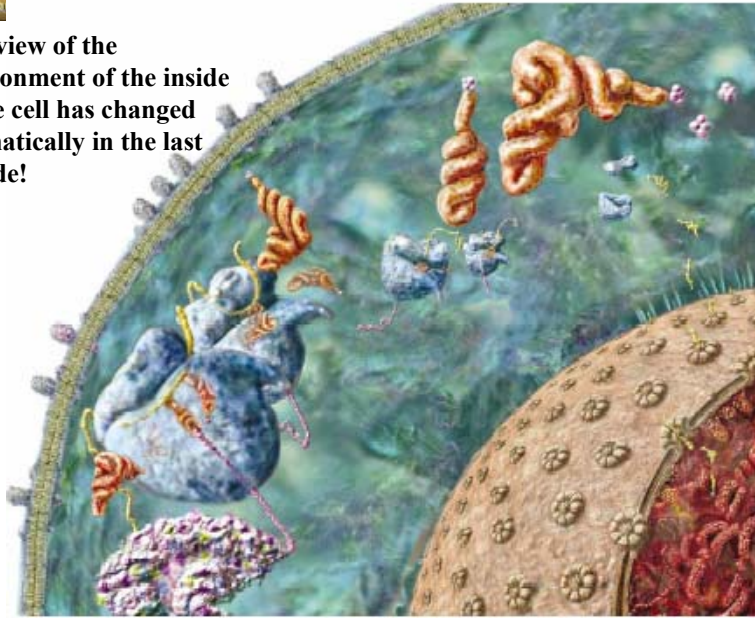
05/22/2005

6





Our view of the environment of the inside of the cell has changed dramatically in the last decade!



Mapping of inner space: proteomics could provide a unique perspective of how cells function

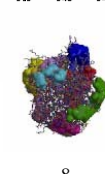
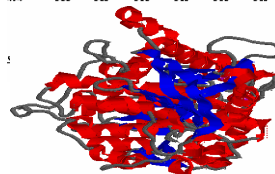
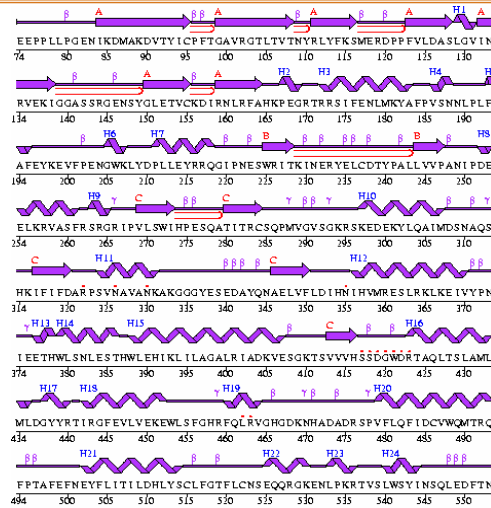
05/22/2005



Protein structure:

- primary (AA sequence)
- secondary (coils, β -sheets, turns)
- tertiary structure
- quaternary structure

Example: MTMR2 (Myotubularin-Related Protein-2)



05/22/2005

8

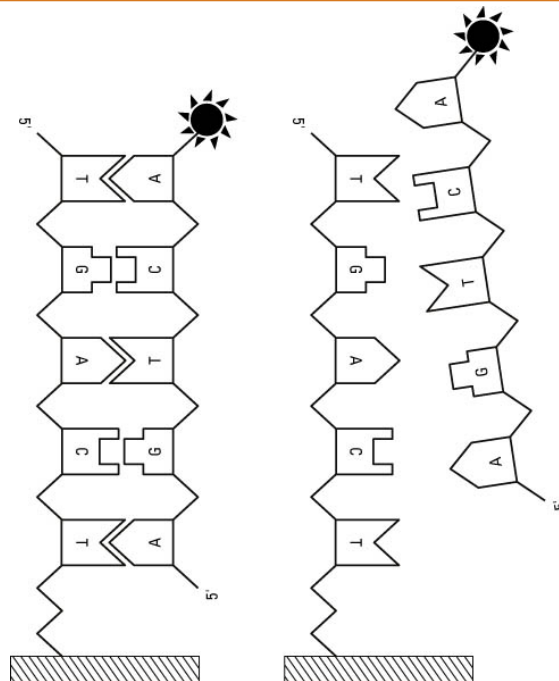


I.2. How Do Microarrays Work?

- **Conceptual description:**
 - Set of targets are immobilized in predetermined positions on a substrate
 - Solution containing tagged molecules capable of binding to the targets is placed over the immobilized targets
 - Binding between targets and tagged molecules occurs
 - Tags allow you to visualize which targets have been bound (and thereby tell you something about the molecules that were present in your solution or about the location of the targets)

05/22/2005

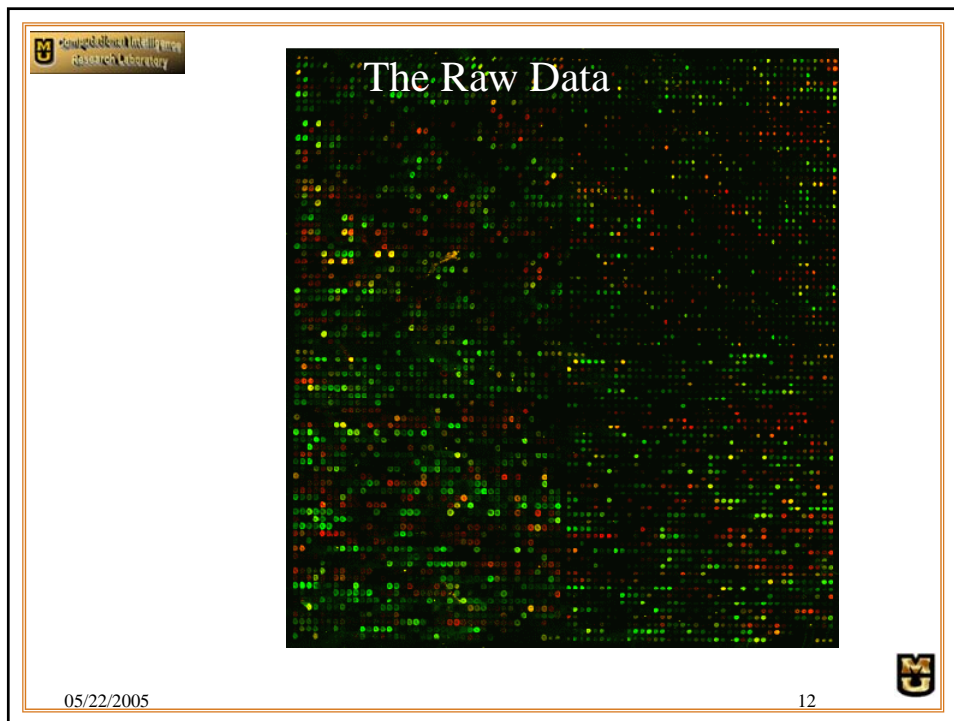
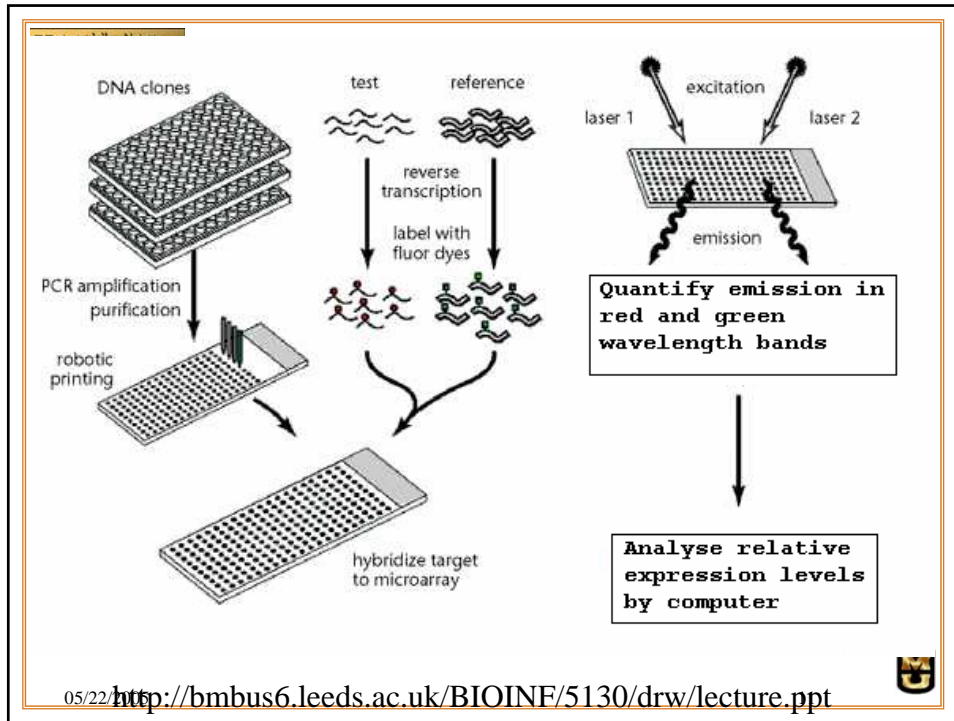
9



05/22/2005

10





The Raw Data



- Measuring mRNA expression levels of many genes in a single experiment
- Conceptually: one spot per gene, 10s of thousands spots per array
- Relative mRNA levels between two samples are being measured
- Red (from cy5) sample 1 > sample 2
- Green (from cy3) sample 2 > sample 1
- Yellow (cy3 + cy5) sample 1 = sample 2
- Black nothing in either

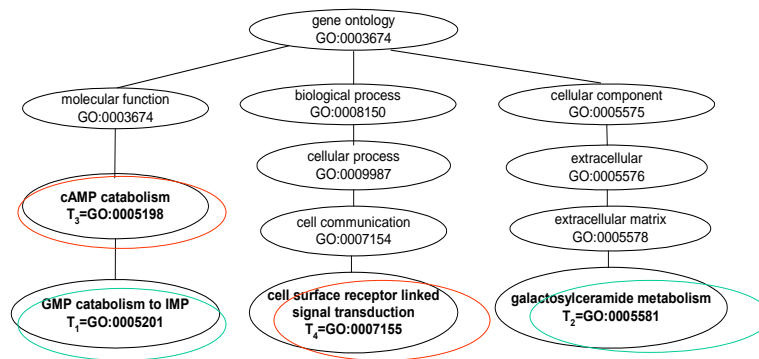
Taken from <http://cmgm.stanford.edu/pbrown/scanner.html>
and <http://bmbus6.leeds.ac.uk/BIOINF/5130/drw/lecture.ppt>

05/22/2005

13



I.3. Taxonomies: Gene Ontology



- Gene ontology (GO) = a controlled terminology
- DAG with “is-a” and “part-of” relationships
- <http://www.geneontology.org>
- COL21A1 : $G_1 = \{T_1, T_2\}$ COL27A1 $G_2 = \{T_3, T_4\}$.

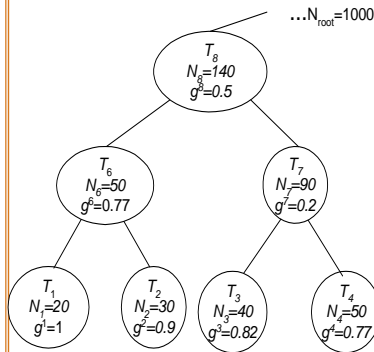
05/22/2005

14



Similarity in an Ontology (DAG)

- Problem: $s(T_1, T_3)=?$
- Many approaches: path-based, depth-based, density-based, information content...
- We use **information content**:
 - Count the occurrence N_i of each term and all children in a corpus (Swiss-Prot)
 - Compute term probability



$$p(T_k) = \left(\frac{N_k}{N_{root}} \right)$$

- Information content of a term is

$$g^k = IC(T_k) = -\ln(p(T_k)) / \max_{T_j \in GO} \{-\ln(p(T_j))\}$$

- Similarity between two terms is (Resnik):

$$s(T_i, T_j) = IC[\text{nearest_ancestor}(T_i, T_j)]$$

$$s(T_1, T_3) = IC(T_8) = 0.5$$

05/22/2005

1. Resnik P., J. of Art. Int. Res. (JAIR), 11, pp. 95-130, 1999.

15



Medical Subject Heading (MeSH)

05/22/2005

16



Bioinformatics Databases (Swiss-Prot, <http://www.ncbi.nlm.nih.gov>, etc)

Literature

GO terms

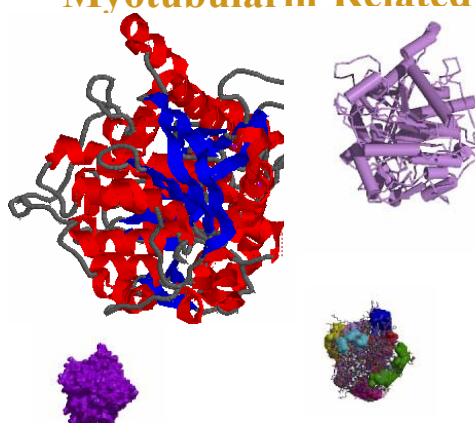

Domains

Sequence

Entry information	
Entry name	ATM_HUMAN
Primary accession number	Q13315
References	
[1]	SEQUENCE FROM NUCLEIC ACID. MEDLINE=96154672; PubMed=8589678; [NCBI, ExPASy, EBI, Israel, Japan] Savitsky K, Sfez S, Tash D.A., Ziv Y., Sarti LA, Collins ES, Strlah Y, Rotman G.; "The complete sequence of the coding region of the ATM gene reveals similarity to cell cycle regulators in different species."; Hum. Mol. Genet. 4:2025-2032(1995).
[58]	VARIANTS MCL LYS 759; LYS 2418 INS; GLY 2423 AND CYS 3008. MEDLINE=20183964; PubMed=10706620; [NCBI, ExPASy, EBI, Israel, Japan] Schaffner C, Eder F, Sillenzbauer S, Doehner H, Lieber P; "Mantle cell lymphoma is characterized by inactivation of the ATM gene."; Proc Natl Acad Sci U S A. 97:2773-2778(2000).
Cross-reference	
GO	GO:0004674; Molecular function: protein serine/threonine kinase activity (traceable author statement) GO:0006281; Biological process: DNA repair (traceable author statement) GO:0007131; Biological process: meiotic recombination (traceable author statement) GO:0000074; Biological process: regulation of cell cycle (traceable author statement) GO:0007165; Biological process: signal transduction (traceable author statement)
Domain	
Pfam	PF02259; FAT1; 1. PF02260; FATC; 1. PF00454; P13_P14_kinase; 1. Pfam graphical view of domain structure
Sequence	
Length: 3056 AA	Molecular weight: 350641 Da
MSVLDLII CCRQLEHRA TERKKEVEK KLRIDGETI KHLDRHSDEK QGKLNWIDAV ...	

05/22/2005
17

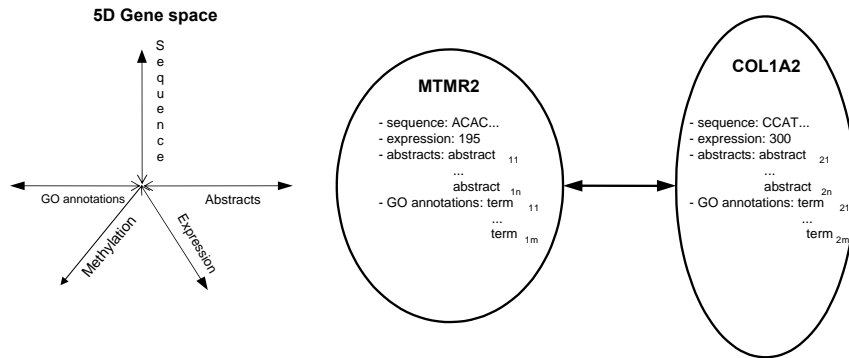
Myotubularin-Related Protein-2 (MTMR2)

Different "looks" at the structure through a study of sequence info

05/22/2005
18

Features for Gene Product Similarity – Our Goal



MTMR2: myotubularin related protein 2
 COL1A2: collagen alpha chain Type 1 Protein 2



Definitions

- **Sequence alignment:**
 - A one-to-one matching of two sequences so that each character in a pair of sequences is associated with a single character of the other sequence or with a null character (gap)

```

-AGGC'TATCACC'TGACC'TCCAGGCCGA--TGCCC---
  || ||||| |||| | || ||| |||||
TAG-C'TATCAC--GACCGC--GGTCGATT'TGCCCGAC
    
```
- **Types of alignment:**
 - Pair-wise vs. multiple
 - Global vs. local
 - Gapped vs ungapped
- **Homologous proteins: share a common ancestor**
 - Orthologous: differ because they are found in different species
 - Paralogous: differ due to a gene duplication event



Scoring the Sequence Similarity

- **Scoring matrices:** each symbol pair is assigned a numerical value based on their biochemical properties
 - DNA scoring matrices
 - Protein scoring matrices: PAM, BLOSUM

- **Gap penalties**
 - allowing gaps can lead to high similarity values for non-homologous sequences
 - Penalizing gaps reduce the number of gaps
 - the cost of a gap is: $C = a + \text{gap_length} * b$



DNA Scoring Matrices

Sequence 1

actaccagttcatttgatacttctcaaa

Sequence 2

taccattaccggtgtaactgaaaggacttaaagact

	A	G	C	T
A	1	0	0	0
G	0	1	0	0
C	0	0	1	0
T	0	0	0	1

Match = 1
Mismatch = 0
Score = 5

- Other choices, e.g., Match = 5, Mismatch = -4: Score = -51



Protein Scoring Matrices

- **Scoring matrices reflect:**
 - # of mutations to convert one to another
 - chemical similarity
 - **observed mutation frequencies**
 - the probability of occurrence of each amino acid
- **Widely used scoring matrices:**
 - PAM [Dayhoff 1978]
 - PAM[1-250]: average change of all amino acid positions
 - BLOSUM [Henikoff 1992],
 - BLOSUM[50-85]: identity between sequences used to build matrix
- **Tips on choosing a scoring matrix:**
 - ✓ For database search the commonly used is BLOSUM62
 - ✓ For closely related proteins use low PAM or high BLOSUM

05/22/2005

25



BLOSUM50

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4
C	-1	-4	-2	-4	13	-3	-3	-3	-2	-2	-3	-2	-2	-2	-4	-1	-1	-5	-3	-1
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3
S	1	-1	1	0	-1	0	-1	0	-1	-3	0	-2	-3	-1	5	2	-4	-2	-2	
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3	
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	-4	-3	-2	-2	2	8	-1
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5

05/22/2005

26



Algorithms for Gene Product Sequence Similarity (Alignment)

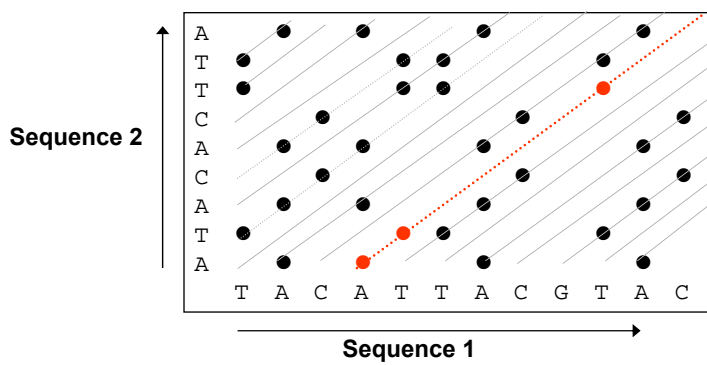
- **Visualization: Dot Plot**
- **Dynamic programming (slow):**
 - Smith-Waterman (local alignment) [Smith 1981]
 - Needleman-Wunsch (global alignment) [Needleman 1970]
- **Heuristic (fast):**
 - Fasta [Pearson 1990]
 - BLAST [Altschul 1990, 1997]

05/22/2005

27



II.2. Dot Plot Similarity Visualization



One possible alignment:

T	A	C	A	T	T	A	C	G	T	A	C
			A	T	A	C	A	C	T	T	A

Window=3, Threshold=2

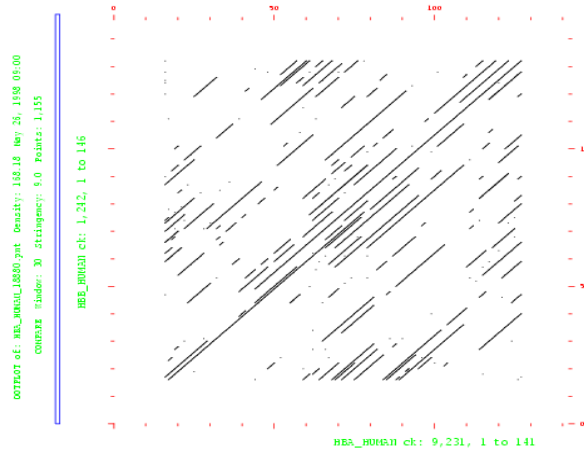
05/22/2005

28



Dot Plot Example

- Hemoglobin α chain (X) vs. Hemoglobin β chain (Y)
 {window =30, stringency (threshold) =9}



II.3. Smith-Waterman

- Recurrence equation:

$$F(i,j) = \max \{ 0, F(i-1, j-1) + s(x_i, y_j), F(i-1,j) - d, F(i, j-1) - d \}$$

- Example: Align HEAGAWGHEE and PAWHEAE

Use BLOSUM 50 for substitution matrix and $d = 8$ for gap penalty

	H	E	A	G	A	W	G	H	E	E
P	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	5	0	0	0	0
W	0	0	0	0	2	0	20	12	0	0
H	0	10	2	0	0	0	12	18	22	14
E	0	2	16	8	0	0	4	10	18	28
A	0	0	8	21	13	5	0	4	10	20
E	0	0	0	13	18	12	4	0	4	16

AWGHE

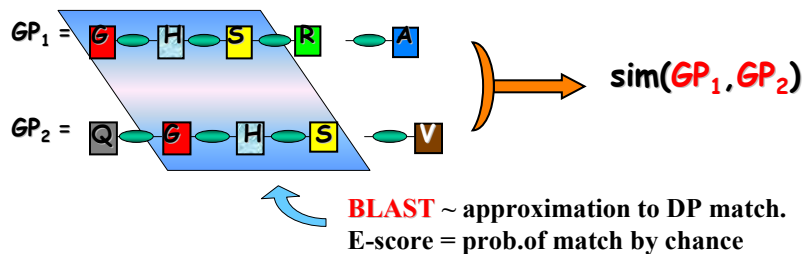
AW-HE



II.4. BLAST

- **Basic Local Alignment Tool** [Altschul et al. 1990, 1997]
- **Designed for searches in large sequence databases**
- **BLAST: is a heuristic that works by finding word-matches between the query and database sequences**
- **BLAST:**
 - Searches for high-scoring local alignments between two sequences
 - Tests for significance of the scores found via P-values.
- **Mathematical basis: random walk** [Ewens et al. 2001, Korf et al. 2003]

Similarity Between Pairs of Sequences



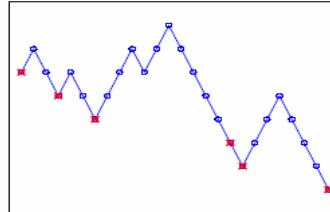
Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped **BLAST** and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Example of a Random Walk

DNA sequence alignment:

```
A G A C T G T A G A C A G C T A A T T A T G C A A
A C G C C C T A G C C A C G A G C G T A T C G C G
```

Score function $S(i, k)$, example: $S(i, k) = 2\delta_{ik} - 1 = \begin{cases} +1 & \text{if } i = k \\ -1 & \text{if } i \neq k \end{cases}$



Ladder points: Points in the walk lower than any previously reached point.

- **Statistic of interest Y_i :** Maximum height of the i th excursion after leaving the i th ladder point and before arriving at the $(i + 1)$ th ladder point.
- **Statistic of interest $Y_{max} = \max\{Y_1, \dots, Y_M\}$**

05/22/2005

33



Steps of BLAST

1. **Filter out low-complexity regions**
 - eliminate statistically significant but biologically uninteresting regions of the query sequence
2. **Create query words of length 3 (for proteins) or 11 (for DNA) from query sequence using a sliding window**

```
MEFPGLGSLGTSEPLPQFVDPALVSS
MEF
EFP
FPG
```

3. **Using a scoring matrix (BLOSUM62 for proteins or +5/-4 for DNA) score all possible words of length $w=3$ (proteins) or $w=11$ (DNA) against each query word**

05/22/2005

34



Steps of BLAST (cont.)

- **4. Select a word threshold (T=14) and keep only the words with score>T (about 50 for each query word)**
 - The total number of high scoring words is about $50 * \text{sequence_length}$
- **5. Scan each database sequence for a match to high scoring-words. Use each match as a seed for an un-gapped alignment**
- **6. Extend each match to the left and right as long as the score increases. This extended matches are called HSP (high-scoring segment pair)**

05/22/2005

35



Steps of BLAST (cont.)

- **7. Determine the statistical significance of each extended match (expect E and p-value) with score > cutoff score S**
 - The expected number of extended matches with score > S expected by chance (E) is:

$$E = kmne^{-\lambda S}$$

m=number of letters in the query

n=number of letters in the database

λ =normalization constant dependent on the scoring matrix

k \approx 0.1, accounts for possible correlation between matches

- The probability of such an alignment (p-value) is:

$$p\text{-value} = 1 - e^{-E}$$

05/22/2005

36



Term-Based Similarity

Given two gene products, G_1 and G_2 , we can consider them as being represented by collections of terms

$$G_1 = \{T_{11}, \dots, T_{1i}, \dots, T_{1n}\} \quad G_2 = \{T_{21}, \dots, T_{2j}, \dots, T_{2m}\}$$

The goal is to define a “natural” similarity: $s(G_1, G_2)$

There are two main approaches

- similarities between pair-wise elements of the two sets are defined and aggregated using a given fusion operator
- the similarity degree can be defined globally for the two entire sets.
 - ❖ In a sense, here the “aggregation” is performed before the similarity is computed.



Gene Product Similarities

Set-based Measures

Jaccard similarity: $s_J(G_1, G_2) = \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|}$

Set Cosine similarity: $s_C(G_1, G_2) = \frac{|G_1 \cap G_2|}{\sqrt{|G_1| |G_2|}}$

Dice similarity: $s_D(G_1, G_2) = \frac{2|G_1 \cap G_2|}{|G_1| + |G_2|}$

Vector Space-based Cosine similarity: $s_V(G_1, G_2) = \frac{v_1 \bullet v_2}{|v_1| |v_2|}$

$v_1 \quad v_2$ are augmented vectors in an augmented space

$$G = G_1 \cup G_2$$



Problems With Existing Approaches

- **“Bag of word” approaches do not account for the information content of the terms**
 - Example 1: $Diet_1 = \{\text{apple, bread}\}$; $Diet_2 = \{\text{pork, bread}\}$. Jaccard: 0.33 (close?)
 - Example 2: $Diet_1 = \{\text{apple}\}$; $Diet_2 = \{\text{orange}\}$. Jaccard: 0 (far?)
- **Existent pair-wise approaches are inconsistent:**
 - Average: $Diet_1 = \{\text{apple, oranges}\}$; $s(Diet_1, Diet_1) < 1$
 - Maximum: $Diet_1 = \{\text{apple, bread}\}$, $Diet_2 = \{\text{fish, bread}\}$; $s(Diet_1, Diet_2) = 1$
- **No approach accounts for uncertainty**
 - $Diet_1 = \{\text{pork(seldom), fish (often), bread (all the time)}\}$,
 $Diet_2 = \{\text{pork(often), fish (seldom), bread (seldom)}\}$;



Building GO Similarities

- **Similarity is computed pair-wise:** $s_{ij}(T_{1i}, T_{2j})$
 - **We Don't** compute similarity directly, but
 - Coefficients of Association
 - Uses Information Theoretic approach
 - **In papers under review (with UWF and ISI)**
 - We fuse with normalized LOS Operators
 - **Here, we look at the sets themselves**



Our New Similarities

- **Based on the concept of Fuzzy Measures**
- **Idea:**
 - Terms describing gene products can be given natural “weights” if they come from taxonomies, like the GO
 - Weights may be based on “information theory” or “depth in tree”
 - Weights might be assigned by experts
 - Fuzzy measures allow the measure of the “whole” to be more (or less) than the “sum of its parts”

05/22/2005

45



Novel Similarities

1. **Fuzzy measure similarity (FMS)⁽¹⁾**
 - Considers the context of a term in a set
2. **Augmented fuzzy measure similarity (AFMS)⁽¹⁾**
 - Addresses the case when there are no common terms
3. **Choquet integral similarity⁽¹⁾**
 - Considers the uncertainty of the objects (annotations)
4. **Linear order statistics similarity (LOS)⁽²⁾**
 - A generalization of the pair-wise maximum and average
5. **PFAM domain similarity⁽³⁾**
 - Uses the distance between two HMM instead of a tree

1. M. Popescu, J.M. Keller, J.A. Mitchell, “Fuzzy Measures on the Gene Ontology for Gene Product Similarity”, *IEEE Trans. Computational Biology and Bioinformatics*, accepted for publication 2005.
2. J.M. Keller, J.C. Bezdek, M. Popescu, N. Pal, J.A. Mitchell, J. Huband, “Gene Ontology-based Knowledge Discovery using GO Similarity Measures based on Linear Order Statistics”, submitted to *Pattern Recognition*, 2005
3. M. Popescu, J. M. Keller, J.A. Mitchell, “Gene Ontology Automatic Annotation Using a Domain Based Gene Product Similarity Measure”, *14th IEEE International Conference on Fuzzy Systems*, Reno, Nevada, May 21-25, 2005.

05/22/2005

46



II.5.i Fuzzy Measure Similarity

- Sources of information in a set G (sensors, features, algorithms, etc.)
 - Here, $G = \{T_1, \dots, T_n\}$, the set of terms describing G
- Worth of sources comes from a **Fuzzy Measure**:
 $g: 2^G \rightarrow [0,1]$ such that
 - $g(\emptyset) = 0$ and $g(G) = 1$
 - $g(A) \leq g(B)$ if $A \subseteq B$
 - If $\{A_i\}$ is an increasing sequence of subsets of G , then

$$\lim_{i \rightarrow \infty} g(A_i) = g\left(\bigcup_{i=1}^{\infty} A_i\right)$$

05/22/2005

47



Fuzzy Measures

- For a fuzzy measure g , let $g^i = g(\{T_i\})$

The mapping $T_i \rightarrow g^i$ is called a fuzzy density function
- The fuzzy density value, g^i , is interpreted as the (possibly subjective) importance of the single information source T_i in determining the similarity of two genes
- General fuzzy measures are broad, but often the densities can be extracted from the problem domain or supplied by experts
- Need fuzzy measures that can be “built” from densities

05/22/2005

48



Fuzzy Measures

- A fuzzy measure g is called a lambda measure (g_λ -fuzzy measure) if additionally:

For all $A, B \subseteq X$ with $A \cap B = \phi$,

$$g(A \cup B) = g(A) + g(B) + \lambda \cdot g(A) \cdot g(B) \text{ for some } \lambda > -1$$

- For any lambda fuzzy measure λ can be uniquely determined for a finite set G by solving

$$1 + \lambda = \prod_{i=1}^n (1 + \lambda g^i)$$

- where $G = \{T_1, \dots, T_n\}$ and $g^i = g(\{T_i\})$ interpreted as the (possibly subjective) importance of the single information source T_i in determining the evaluation of a hypothesis

Construction of Fuzzy Densities

- Collect all the terms (T_j) for all Gene Products in Database
- Compute Information Theoretic Content
 - Use a Corpus (like Swiss-Prot)
 - Certainly other ways to get Term importance

$$p(T_k) = \left(\frac{\text{count}(T_k + \text{children of } T_k \text{ in CORPUS})}{\text{count}(\text{all GO terms in CORPUS})} \right)$$

$$g^k = ic(T_k) = -\log(p(T_k)) / \max_{T_j \in \text{GO}} \{-\log(p(T_j))\}$$

Example GenBank ID AAN03650 (COL24A1 gene)

$G = \{ T_1 = 5201 \text{ (“extracellular matrix structural component”),}$
 $T_2 = 7155 \text{ (“cell adhesion”), } T_3 = 5581 \text{ (“collagen”) } \}$

$$\{g^k\} = \{0.58, 0.44, 0.65\}$$

$$1 + \lambda = (1 + 0.58\lambda)(1 + 0.44\lambda)(1 + 0.65\lambda) \Rightarrow \lambda = -0.86$$

$$g(\{T_1\}) = g^1 = 0.58, \quad g(\{T_2\}) = g^2 = 0.44, \quad g(\{T_3\}) = g^3 = 0.65,$$

$$g(\{T_1, T_2\}) = g^1 + g^2 + \lambda g^1 g^2 = 0.8.$$

$$g(\{T_1, T_3\}) = 0.9, \quad g(\{T_3, T_2\}) = 0.84, \quad g(G) = 1$$

Fuzzy Measure Similarity

New fuzzy measure similarity between two sets G_1 and G_2 of terms is defined as:

$$s_{FMS}(G_1, G_2) = \frac{g_1(G_1 \cap G_2) + g_2(G_1 \cap G_2)}{2}$$

where g_1 is a fuzzy measure defined on G_1 and g_2 is a fuzzy measure defined on G_2

Example: Two Genes From the Same Family

G_1 : GenBank ID AAH35609 (MTMR4 gene)

G_2 : GenBank ID AAH12399 (MTMR8 gene)

$G_1 = \{T_1=4721(\text{protein phosphatase activity}), T_2=6470(\text{protein amino acid dephosphorylation}), T_3=8270(\text{zinc ion binding})\}$,

$G_2 = \{T_1=4721(\text{protein phosphatase activity}), T_2=6470(\text{protein amino acid dephosphorylation}), T_4=16787(\text{hydrolase activity})\}$.

Densities: $\{g^{1i}\} = \{0.52, 0.57, 0.54\}$; $\{g^{2j}\} = \{0.52, 0.57, 0.33\}$

Here, the set of common terms that supports the similarity of G_1 and G_2 is $\{T_1, T_2\}$

Intra Family Example (continued)

$$s_D = \frac{4}{6} \approx 0.67, \quad s_J = \frac{2}{4} = 0.5.$$

$$v_1 = (0.52 \ 0.57 \ 0.54 \ 0.0), v_2 = (0.52 \ 0.57 \ 0.0 \ 0.33) \Rightarrow$$

$$s_V = \frac{(0.52 \ 0.57 \ 0.54 \ 0.0) \bullet (0.52 \ 0.57 \ 0.0 \ 0.33)}{0.94 * 0.84} \approx 0.75.$$

Lambda measure for G_1 has $\lambda = -0.84$ $g_1(\{T_1, T_2\}) = 0.84$.

Lambda measure for G_2 has $\lambda = -0.72$ $g_2(\{T_1, T_2\}) = 0.88$.

$$s_{FMS}(G_1, G_2) = \frac{g_1(\{T_1, T_2\}) + g_2(\{T_1, T_2\})}{2} = \frac{0.84 + 0.88}{2} = 0.86$$

II.5.ii Augmented Sets

What happens if $G_1 \cap G_2 = \emptyset$?

Suppose that G_1 and G_2 are as before (terms from a taxonomy):

$$G_1 = \{T_{11}, \dots, T_{1i}, \dots, T_{1n}\} \quad G_2 = \{T_{21}, \dots, T_{2j}, \dots, T_{2m}\}$$

Augment each set as: $G'_1 = G_1 \cup \{T_{i,2j}\}$ $G'_2 = G_2 \cup \{T_{i,2j}\}$

$\{T_{i,2j}\}$ is the set of **nearest common ancestors** (NCA) of every pair (T_{1i}, T_{2j})

Then $[G_1 \cap G_2]' = [G'_1 \cap G'_2] = [G_1 \cap G_2] \cup \{T_{i,2j}\}$ and calculate FMS on it

Construction of Augmented Densities

- **Note: Root node (GO) has**
 - Probability 1 and info content 0
- **For each pair of terms (T_i, T_j) in the set of distinct terms**
 - Find the Nearest Common Ancestor node NCA

$$T_{ij} = T(g_{ij}) = T(\text{NCA}(g_i, g_j))$$

- **And set the “augmented density” to**

$$g^k = \text{ic}(T_{ij}) = -\log_2(p(T_{ij}))$$

or

$$g^k = \text{ic}(T_{ij}) = 1 - p(T_{ij})$$

Our Second Approach

- What if pairs of terms have both similarities and “importance” towards determining total gene similarity?
- For example, same or similar annotation terms to generate pair similarity and use “reliability of annotation” to create importance (fuzzy measure)
- Useful (we conjecture) for comparing based on abstracts
 - Keywords build pairwise similarities
 - Impact factors (or source of terms) give importance



II.5.iii Choquet Fuzzy Integral

Suppose that G_1 and G_2 are as before (terms from a taxonomy):

$$G_1 = \{T_{11}, \dots, T_{1i}, \dots, T_{1n}\} \quad G_2 = \{T_{21}, \dots, T_{2j}, \dots, T_{2m}\}$$

Let $X = G_1 \times G_2$ and $s: X \rightarrow [0,1]$

To simplify the notation, we reorder the term pairs and label them by a single subscript so that $X = \{T_1, T_2, \dots, T_{nm}\}$

$T_k = (T_{1i}, T_{2j})$ for some pair (i,j)

Then we compute $s(T_k) = s_{ij}(T_{1i}, T_{2j})$



Choquet Fuzzy Integral

Let g be a fuzzy measure on (finite set) X

Then the Choquet fuzzy integral of s with respect to g is given by

$$C(s) = \sum_{i=1}^{nm} [s(T_{(i)}) - s(T_{(i+1)})] \cdot g(S_i)$$

where the function values are reordered so that

$$s(T_{(1)}) \geq s(T_{(2)}) \geq \dots \geq s(T_{(nm)}) \quad s(T_{(nm+1)}) = 0$$

and

$$S_i = \{T_{(1)}, \dots, T_{(i)}\}$$

Choquet Fuzzy Integral

Define $w_i = g(S_i) - g(S_{i-1}) \quad g(S_0) \equiv 0$

Then the Choquet fuzzy integral can be rewritten as

$$C(s) = \sum_{i=1}^{nm} w_i \cdot s(T_{(i)})$$

Looks linear, but isn't - Depends on the sort

II5.iv Linear Order Statistics (LOS) Similarity

- **Two gene products, G_1 and G_2** , represented by collections of GO terms, journal abstracts or PFAM domains:

$$G_1 = \{T_{11}, \dots, T_{1i}, \dots, T_{1n}\} \quad G_2 = \{T_{21}, \dots, T_{2j}, \dots, T_{2m}\}$$

- The LOS similarity between G_1 and G_2 is:

$$s_{LOS}(G_1, G_2) = \sum_{i=1}^{nm} w_i s(T_{(i)}) \quad s(T_{(1)}) \geq s(T_{(2)}) \geq \dots \geq s(T_{(nm)})$$

where

- $s(T_{(i)})$ = ordered pair-wise similarities
- $T_{(i)}$ = a pair of objects (GO terms, journal abstracts or PFAM domains), (T_{1j}, T_{2k})
- $w = (w_1, \dots, w_{nm})$ is a weight vector

- **OBS: maximum=(1,0,...,0), average=(1/nm,...,1/nm);**

05/22/2005

61



Annotation Reliability

Numeric values chosen for the reliability of the GO annotation

Traceable author statement	Inferred from sequence similarity	Inferred from electronic annotation	Non-traceable author statement	Not documented	Not recorded
TAS	ISS	IEA	NAS	ND	NR
1	0.8	0.6	0.4	0.1	0.1

05/22/2005

62



Example1: Annotation Reliability Example

Earlier Intra family example:

$G_1 = \text{AAH35609 (MTMR4 gene)}$

$G_2 = \text{AAH12399 (MTMR8 gene)}$

$G_1 = \{T_1=4721(\text{TAS}), T_2=6470(\text{IEA}), T_3=8270(\text{NR})\}$

$G_2 = \{T_1=4721(\text{ISS}), T_2=6470(\text{NAS}), T_4=16787(\text{NR})\}$

$$s(T_{1i}, T_{2j}) = \begin{vmatrix} 0.52 & 0.33 & 0 \\ 0.1 & 0.1 & 0 \\ 0 & 0 & 0.57 \end{vmatrix}$$

Annotation Reliability Example

How to Generate Densities?

Reliabilities: $\{c^{1i}\} = \{1, 0.6, 0.1\}$ $\{c^{2i}\} = \{0.8, 0.4, 0.1\}$

Densities: $c^{ij} = c(T_{1i}, T_{2j}) = \begin{vmatrix} 0.8 & 0.4 & 0.1 \\ 0.6 & 0.4 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{vmatrix}$

$$c^{ij} = \min(c(T_{1i}), c(T_{2j}))$$

Annotation Reliability Example

Sorted Similarities and Associated Densities

$$\{s(T_{(i)})\} = \{0.58, 0.52, 0.33, 0.1, 0.1, 0, 0, 0, 0\}$$

$$\{c^{(i)}\} = \{0.1, 0.8, 0.4, 0.4, 0.6, 0.1, 0.1, 0.1, 0.1\}$$

Use Decomposable Measure

$$g(\{c_{(1)}, c_{(2)}\}) = \min(1, g(\{c_{(1)}\}) + g(\{c_{(2)}\}))$$

$$s_{\text{Choquet}} = [0.1(0.58 - 0.52) + 0.9(0.52 - 0.33) + 1(0.33 - 0.1) + 1(0.1 - 0.1) + 1(0.1 - 0)] = 0.5$$



Example 2: LOS Using GO Annotations

- Same 2 gene products (myotubularin family):
 - $G_1 = \text{AAH35609 (MTMR4 gene)} = \{T_{11}=4721, T_{12}=6470, T_{13}=8270\}$
 - $G_2 = \text{AAH12399 (MTMR8 gene)} = \{T_{21}=4721, T_{22}=6470, T_{23}=16787\}$
- Other similarities: FMS=0.86, Blast=0.85, Average=0.28, Maximum=1, Jaccard=0.5

•LOS similarity:

1. Compute pair-wise similarities and order them:

$$\{s(T_{(i)})\} = \{0.58, 0.52, 0.33, 0.1, 0.1, 0, 0, 0, 0\}$$

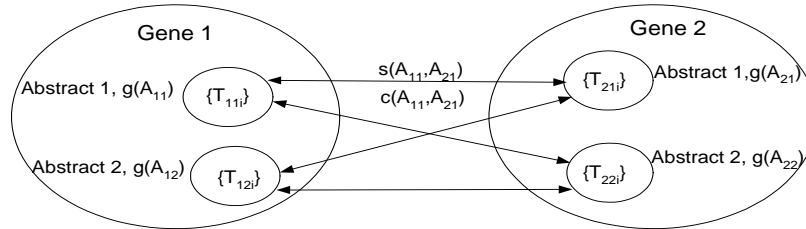
2. Choose the weight vector: $w = (0.4 \ 0.4 \ 0.2 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0)$

3. Compute the LOS similarity:

$$s_{\text{LOS}} = w_1 s(T_{(1)}) + w_2 s(T_{(2)}) + w_3 s(T_{(3)}) = 0.4 * 0.58 + 0.52 * 0.4 + 0.33 * 0.2 = 0.5$$



Example 3: Gene Similarity Based on Abstract Matching



- Use Choquet Fuzzy Integral to fuse!
- What can we discover?:
 - Genes that co-occur in experiments (as reported in journals) are **believed** by the authors to be connected (even if they are not homologues)



Abstract Term Example

MeSH Terms for L32132:	Tree ID	Density
Amino Acid Sequence	G06.184.603.060	0.2
Animal	check tag	0
Base Sequence	G06.184.603.080	0.2
Carrier Proteins/analysis	D12.776.157	0.18
Carrier Proteins/chemistry	D12.776.157	0.18
Carrier Proteins/genetics*	D12.776.157	0.18
Cattle	B02.649.077.380.271	0.22
Cloning, Molecular	E05.393.220	0.18
DNA, Complementary/analysis	D13.444.308.497.220	0.22
Human	check tag	0
Liver/metabolism	A03.620	0.16
Male	check tag	0
Molecular Sequence Data	L01.453.245.667	0.2
RNA, Messenger/analysis	D13.444.735.544	0.2
Rabbits	B02.649.521.700	0.2
Rats	B02.649.865.635.560	0.22
Rats, Sprague-Dawley	B02.649.865.635.560.670	0.23
Sequence Homology, Amino Acid	G06.184.842.200	0.2
Support, Non-U.S. Gov't	check tag	0
Support, U.S. Gov't, P.H.S.	check tag	0



Matching by Abstract

- $s(\text{ATM}, \text{STK11})=?$ Expert: Should be “Medium” Similar
- Algorithm:
 - Retrieve PubMed abstracts for ATM, STK11
 - Calculate all the pair-wise distances based on the MeSH indexing
 - Keep the 4 best-matching pairs
 - Find the impact factor for each journal: $g(A_i), i=1\dots 8$

ATM	12917635- Oncogene (6.737)	12970738- Oncogene (6.737)	14500819-Nucleic Acids Res. (6.373)	14499692-Science (23.329)
STK11	12183403 – Cancer Res (8.30)	12234250 – Biochem J (4.326)	12805220 - EMBO J. (12.459)	11853558- Biochem J (4.326)

Abstract Similarity Example

Calculate the confidence of the pair (use IF, here)
 $g^{ij}=g(A_1, A_2) = \text{IF}(A_1) * \text{IF}(A_2)$ and normalize to max

$$\{g^{ij}\} = \begin{vmatrix} 0.19 & 0.10 & 0.29 & 0.10 \\ 0.19 & 0.10 & 0.29 & 0.10 \\ 0.18 & 0.09 & 0.27 & 0.09 \\ 0.67 & 0.35 & 1.00 & 0.35 \end{vmatrix}$$

Abstract Similarity Example

Abstract Pairwise Similarity by FMS

$$s(A_k)_{\text{FMS}} = \begin{vmatrix} 0.44 & 0.0 & 0.00 & 0.00 \\ 0.07 & 0.29 & 0.1 & 0.11 \\ 0.00 & 0.13 & 0.26 & 0.32 \\ 0.00 & 0.20 & 0.16 & 0.24 \end{vmatrix}$$

Weighted Average:
 $s_a(\text{ATM}, \text{STK11}) = 0.37$

Choquet Integral
 $s_{\text{Choquet}}(\text{ATM}, \text{STK11}) = 0.53$



II.5.v: Domain-Based Similarity

- Two gene products described by sets of PFAM (<http://pfam.wustl.edu>) domains $G_1 = \{M_{11}, \dots, M_{1N}\}$, $G_2 = \{M_{21}, \dots, M_{2K}\}$ where M_{ij} is the number of PFAM domains λ_j contained in gene product i
- Define a **PFAM domain pair-wise similarity** using ⁽¹⁾:

$$D_s(\lambda_1, \lambda_2) = \frac{D(\lambda_1, \lambda_2) + D(\lambda_2, \lambda_1)}{2}$$

$$D(\lambda_1, \lambda_2) = \frac{1}{T} [\log P(O_1 | \lambda_1) - \log P(O_1 | \lambda_2)]$$

where O_1 is a sequence of length T generated with λ_1

- **Assumption:**
 - neglect the order of the domains
 - To account for the domain order we use dynamic programming together with the above HMM distance



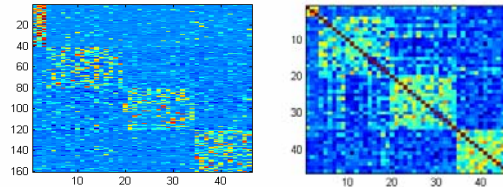
PFAM Similarity Between ATM & STKB

- $G_1 = \text{ATM_HUMAN} = \{M11=1(\text{PI3_PI4_KINASE})\}$
 $G_2 = \text{STKB_HUMAN} = \{M21=1(\text{PKINASE})\}$
- Smith-Waterman=0.04, Jaccard=0
- The HMM similarity between the 2 domains= 0.17 → the gene similarity is $s(G_1, G_2)=0.17$
- OBS: in the general case we:
 - Use $M_{i1} * M_{2j} * s(\lambda_i, \lambda_j)$ for measures such as FMS, Jaccard, etc
 - Integrate $M_{i1} * M_{2j}$ w.r.t. the HMM similarity $s(\lambda_i, \lambda_j)$ for the Choquet similarity

III. Visualization and Clustering

Two Types of Data

- **Relational:**
 - Obtained by computing the similarities between a set of objects
 - Examples: patient-patient in microarray experiments, gene-gene in family classification.
 - Algorithms: hierarchical, VAT, FCM, NERFCM
- **Object data**
 - Examples: Patient-genes in microarray experiments, gene-domains, gene-GO terms
 - C-means algorithms (hard, fuzzy, possibilistic) do not usually work due to the high dimensionality of the data (8000-30000 dimensions).
 - Algorithms: bi-clustering (co-clustering)



05/22/2005

75



Our Experimental Design

- **Extract “families” of Gene Products**
 - Sequence ID
- **Get Sequence data**
 - Compute sequence-based similarities
- **Get GO annotations**
 - Construct similarities from sets of annotating terms
 - We’ll use set-based methods (like fuzzy measures)
- **Visual Comparisons**
- **Clustering and Knowledge Discovery**

05/22/2005

76



Construction of GPD194_{12.10.03}

- 194 human gene products clustered into three protein families using the Markov clustering algorithm (Enright 2002)
- From ENSEMBL Genome Browser: www.ensembl.org

Characteristics of the GPD194_{12.10.03} data set

Ensembl ID	N _i = Number of Human Gene Products	F _i = Protein Family	No. of genes
ENSF00000000339	21	myotubularin	7
ENSF00000000073	87	receptor precursor	7
ENSF00000000042	86	collagen alpha chain	13

05/22/2005

77



Sequence Comparison

- The 194 DNA sequences are submitted to
 - the Smith-Waterman routine and
 - the BLAST procedure to obtain
- Sets of pairwise numerical similarities
 - $\{s_{ij} : s_{ij} \in [0, 1] ; 1 \leq i, j \leq 194\}$ and $\{b_{ij} : b_{ij} \in [0, 1] ; 1 \leq i, j \leq 194\}$

$$s_{ij} = \frac{\text{alignment_length}(\text{gene_product}_i, \text{gene_product}_j)}{\min\{\text{length}(\text{gene_product}_i), \text{length}(\text{gene_product}_j)\}}$$

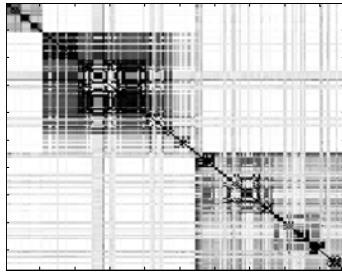
$$b_{ij} = \begin{cases} 0 & \text{if } -\log(E\text{-score}) < 0 \\ 1 & \text{if } -\log(E\text{-score}) > 100 \\ -\log(E\text{-score})/100 & \text{else} \end{cases}$$

05/22/2005

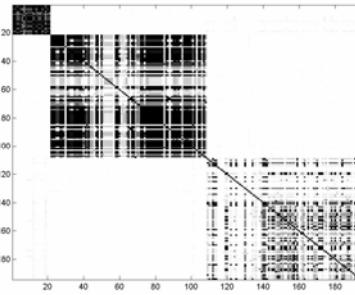
78



Sequence (Dis)similarity Images



Smith-Waterman



Blast

Pretty Binary!

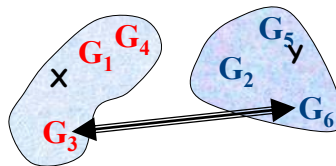


III.1 Hierarchical Clustering $\Rightarrow U_{\text{crisp}}$

Most used clustering in microarray studies

Different linkage types: complete (max), single (min), average

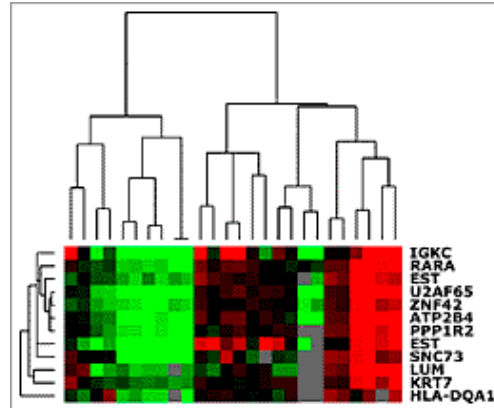
Input $D_{n \times n}$	$D_{ij} \geq 0$ ∇ $D_{ii} = 0$ ∇ $D_{ij} = R_{ji}$
Pick	A metric δ on <i>pairs of sets</i>
From : $c = n$ To : $c = 1$	Merge most similar clusters $c \Rightarrow c-1$ ∇ $D_n \Rightarrow D_{n-1}$



$$d_{\text{CompleteLinkage}}(X, Y) = \max_{\substack{x \in X \\ y \in Y}} \{d(x, y)\} = \max_{\substack{1 \leq i, j \leq n \\ i \neq j}} \{d_{ij}\}$$



Example of Hierarchical Clustering



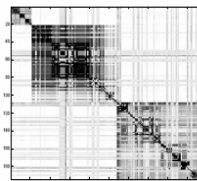
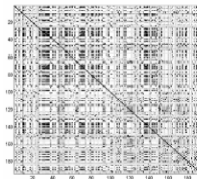
05/22/2005

81



III.2 Visual Assessment of cluster Tendency (VAT)

- Based on Minimum Spanning Tree (Prim's algorithm)
- Input: Dissimilarity matrix



VAT Ordering and Display Algorithm

- Step 1** Set $K = \{1, 2, \dots, n\}$; $I = J = \emptyset$; $P[0] = (0, \dots, 0)$.
- Step 2** Select $(i, j) \in \arg \max_{p \in K, q \in K} \{R_{pq}\}$.
Set $P(1) = i$; $I = \{i\}$; and $J = K - \{i\}$.
- Step 3** For $r = 2, \dots, n$:
Select $(i, j) \in \arg \min_{p \in I, q \in J} \{R_{pq}\}$.
Set $P(r) = j$; Replace $I \leftarrow I \cup \{j\}$ and $J \leftarrow J - \{j\}$.
Next r .
- Step 4** Obtain the ordered dissimilarity matrix \tilde{R} using the ordering array P as: $\tilde{R}_{ij} = R_{P(i)P(j)}$, for $1 \leq i, j \leq n$.
- Step 5** Display the reordered matrix \tilde{R} as the ODI \tilde{I} using the conventions given above.

05/22/2005

82



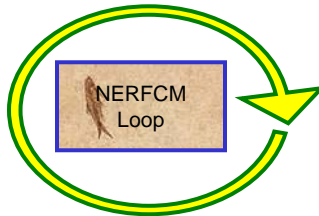
III.3 Non-Euclidean Relational Fuzzy C-means (NERFCM)

input: Dissimilarity matrix, D ; output: Fuzzy memberships, U_{fuzzy}

Input $D_{ij} \geq 0$ $D_{ij} = D_{ji}$ $D_{ii} = 0$ $D \circ \alpha([1]_n - I_n)$

Pick $2 \leq c < n$ $m > 1$ $\epsilon > 0$

Initialize $r = 0$ $\beta = 0$ $D_\beta = D + \beta[1]_n$ $U^{(0)} \in M_{fcn}$



Outputs $U^* \in M_{fcn}$ $\{v_1, \dots, v_c\}$

Repeat Until $\|U^{(r)} - U^{(r-1)}\| \leq \epsilon$

$$v_i^{(r)} = ((U_{i1}^{(r)})^m, (U_{i2}^{(r)})^m, \dots, (U_{im}^{(r)})^m) / \sum_{j=1}^n ((U_{ij}^{(r)})^m)$$

$$d_{ik} = (D_\beta v_i)_k - (v_i^T D_\beta v_i) / 2$$

IF $d_{ik} < 0$ for any i and k **% Adjust β**

$$\Delta\beta = \max_{i,k} \left\{ -2d_{ik} / \|v_i - e_k\|^2 \right\}$$

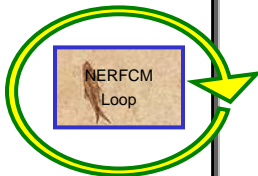
% $e_k = (0, \dots, \frac{1}{k}, \dots, 0)^T \in \mathfrak{R}^n$

$$d_{ik} \leftarrow d_{ik} + (\Delta\beta / 2) \cdot \|v_i - e_k\|^2$$

$$\beta \leftarrow \beta + \Delta\beta$$

IF $d_{ik} > 0$, $i \neq 1$ to c **THEN** $U_{ik} = \left(\sum_{j=1}^c d_{ik} / d_{jk} \right)^{\frac{-1}{(m-1)}}$

ELSE $U_{ik} = 0$ if $d_{ik} = 0$, $U_{ik} \in [0, 1]$ s.t. $\sum U_{ik} = 1$
 $r = r + 1$



III.4 Co-Clustering

- AKA simultaneous clustering, two-way clustering, biclustering
- Applied mainly in two fields: text (web) mining and bioinformatics (microarrays)
- Text mining: each column represents a key word, each row represent a document
- Microarray: each column represents a patient and each row represent a gene
- Idea: cluster patient (documents) and genes (key words) *simultaneously*



Why Co-Clustering?

	P ₁	P ₂	P ₃	P ₄	P ₅	P ₆
g ₁	1	1	1	0	1	0
g ₂	0	1	1	0	1	1
g ₃	0	0	1	1	1	0
g ₄	1	0	1	1	0	0
g ₅	1	1	0	0	1	1
g ₆	1	1	0	0	1	1
g ₇	0	1	0	0	1	0

- Problem reduces to finding dense submatrices
- Exact solution is impractical
- Only patients {p₁, p₂, p₃} have the genes {g₁, g₂} expressed
- Microarray significance: find subgroups of given cancer (leukemia) patients that respond different to different treatments, that is, {p₁, p₂, p₃} respond to drug A while {p₄, p₅, p₆} not.
- Web significance: documents {p₁, p₂} can be summarized by words {g₁, g₂, g₃}; If {g₁, g₂, g₃} can be in turn summarized by {G} (using an ontology)=> G can link to {p₁, p₂}



Application Algorithms

- **Web mining:**
 - Marker propagation, ping-pong: Oyanagi 2001
 - Fuzzy co-clustering, FCCM: Oh 2001
 - Fuzzy co-clustering, FSKWIC, Frigui 2002
 - Fuzzy co-clustering, CoDoK, Kummamuru 2003
- **Bioinformatics:**
 - Residue minimization biclustering: Cheng & Church 2001
 - Spectral graph approach: Cho & Dhillon 2001
 - Coupled two way clustering (CTWC): Getz 2000

III.5 Other Clustering Algorithms Used in Bioinformatics

- **Markov Clustering**
 - Used to cluster Swiss-Prot gene products (~150000) in families
 - Used Blast similarity (E-score)
 - Results: Ensembl browser (www.ensembl.org)
- **Minimum spanning trees (MST)**
 - Used for gene expression data
- **Super paramagnetic clustering (SPC)**
 - Used in CTWC (Getz 2000)
 - Uses paramagnetic spin propagation to define a local similarity measure

Clustering References

- **Hierarchical**
 - Claverie J.-M., Human Molecular Genetics, No. 8, pp. 1821-1183, 1999.
 - S. Raychaudhuri, PD Suthphin, JT Chang, RB Altman, "Basic microarray analysis: grouping and feature reduction", Trends in Biotechnology, Vol. 19, No5, May 2001.
 - Mei-Ling Ting Lee, Analysis of microarray gene expression data, Kluwer AP, Boston, MA, 2004.
- **VAT**
 - Bezdek, J.C.; Hathaway, R.J.; VAT: a tool for visual assessment of (cluster) tendency, Neural Networks, 2002. Proceedings, IJCNN '02, Volume 3, May, 2002, pp. 2225-2230.
- **NERFCM**
 - R. J. Hathaway and J. C. Bezdek, "NERF C-Means: Non-Euclidean relational fuzzy clustering", Pattern Recognition, vol. 27, No. 3, pp. 429-437, 1994.
- **FCM**
 - Claverie J.-M., Human Molecular Genetics, No. 8, pp. 1821-1183, 1999.
- **Bi-clustering (Co-clustering)**
 - Y. Cheng, G. M. Church, Biclustering of Expression Data, Proceedings of the Eighth International Conference on ISMB, 2000, Pages: 93 - 103
 - G Getz, E Levine and E Domany, Coupled two-way clustering analysis of gene microarray data. Proc Natl Acad Sci U S A 2000, 97:12079-12084
 - H. Cho, I. S. Dhillon, Y. Guan, and S. Sra, Minimum Sum-Squared Residue Co-clustering of Gene Expression Data, Proc. of the 4th SIAM International Conference on Data Mining, pages 114-125, April 2004
 - Kummamuru, K., Dhawale, A.K., Krishnapuram, R.: Fuzzy co-clustering of documents and keywords. In: Proc. of FUZZIEEE, St. Louis, USA (2003)
 - Oh, C.H., Honda, K., Ichihashi, H.: Fuzzy clustering for categorical multivariate data. In: Proc. of IFSA/NAFIPS, Vancouver (2001) 2154-2159
 - Oyanagi, S., Kubota, K., Nakase, A.: Application of matrix clustering to web log analysis and access prediction. In: Proceedings of WEBKDD, San Francisco (2001)
 - Frigui, H., Nasraoui, O.: Simultaneous categorization of text documents and identification of cluster-dependent keywords. In: Proceedings of FUZZIEEE, Honolulu, (2002) 158-163
 - Raghuram Krishnapuram, Introduction to Knowledge Management and Text Mining, Tutorial FUZZIEEE 2003, St Louis, MO.
- **Other**
 - MCL: Enright A.J., Van Dongen S., Ouzounis C.A., *Nucleic Acids Res.*, vol. 30, no. 7, 2002.
 - MST: Ying Xu, Victor Olman, Dong Xu, Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees Bioinformatics, Vol. 18 no. 4 2002, Pages 536-545



IV. Knowledge Discovery



Knowledge Discovery in Bioinformatics

- 1. Clustering (and cluster validity) of gene products in families;
- 2. Automatic annotation (GO, Domains, etc) of gene products (verification of the existent ones)
- 3. Functional summarization of gene products (what are the main functions of a set of genes?)
- 4. Other bioinformatics applications
 - a. Phylogenetic trees
 - b. Secondary structure prediction
 - c. Learning biochemical networks from microarray data

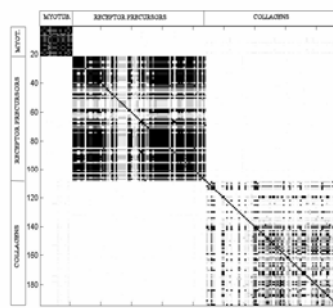
05/22/2005

91

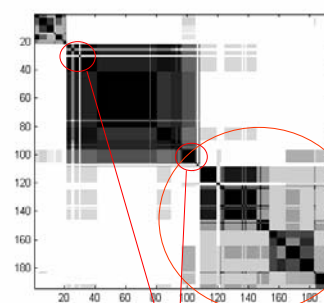


IV.1 Clustering of Gene Products in Families

- BLAST and FMS similarity matrix among the 194 gene products
- Cluster the similarity matrix using FCM⁽¹⁾
- Collagen superfamily substructure was later confirmed by biologists⁽²⁾



BLAST



FMS

COL1A2,
COL21A1,
COL24A1,
COL27A1,
COL2A1,
COL3A1,
COL4A1,
COL4A2,
COL4A3,
COL4A6,
COL5A3,
COL9A1,
COL9A2

Annotation errors!

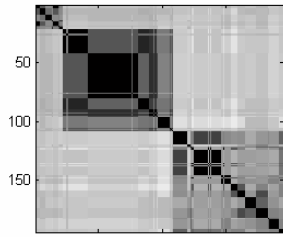
1. Claverie J.-M., Human Molecular Genetics, No. 8, pp. 1821-1183, 1999.
2. Myllyharju J, Kivirikko K.I., Trends in Genetics 2004; 20(1), pp. 33-43.

05/22/2005

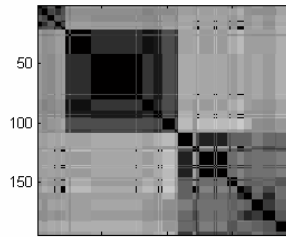
92



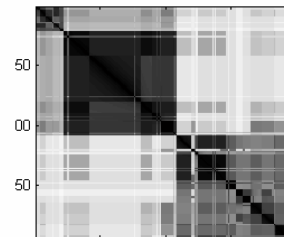
Visual Inspection: Augmented Sets



Jaccard

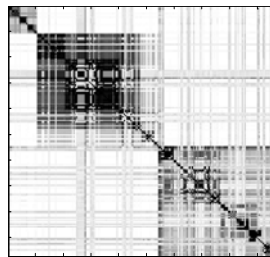


Cosine



FMS

- Raises all similarities
- Somewhat stronger within family values



Smith-Waterman

05/22/2005

93



Anything More Than Just a “Pretty Face”?

Pearson’s coefficient between similarities and BLAST and “Ideal”

GO similarity	FMS	AFMS	Jaccard	Average	Maximum
Person’s Coefficient (vs. BLAST)	0.52	0.54	0.44	0.44	0.47
Person’s Coefficient (vs. Ideal)	0.9	0.86	0.72	0.82	0.84

05/22/2005

94



Pearson's coefficient for the measures using the information reliability

Similarity Measure/ Comparison target	Reliability Weighted Jaccard	Choquet
Pearson coefficient (BLAST)	0.41	0.49
Pearson coefficient (Ideal case 1-0 similarity)	0.65	0.85

Simple Clustering Example

Number of mismatches between three gene families from MCL (Ensembl) and respective similarity type using complete linkage in Hierarchical Clustering

	Jaccard	Cosine	FMS	Blast
Nonaugmented	105	105	35	85
Augmented	11	27	0	---

Simple Clustering Example

Number of mismatches between three gene families from MCL and respective similarity type using **single linkage** in Hierarchical Clustering

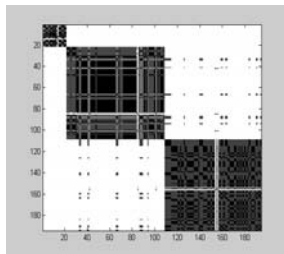
	Jaccard	Cosine	FMS	Blast
Nonaugmented	0	0	84	105
Augmented	0	106	0	---

05/22/2005

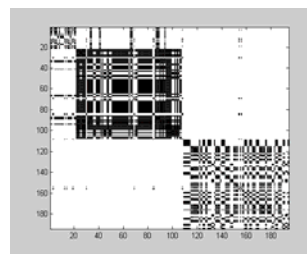
97



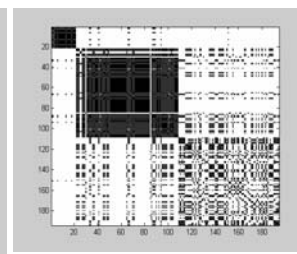
Can Look at Sub-Taxonomies, mf,cc,bp



mf-molecular function



cc-cellular component



bp-biological process

Actually we have used the MF branch for functional summarization

05/22/2005

98



IV.2 Automatic Functional Annotation of Gene Products

- **GO similarity measures work for known genes**
 - Where annotation terms are known
- **What is wrong with BLAST?**
 - The match might not be related to the function
 - Score accounts for the largest match => tends to be binary
- **Represent gene products using DOMAINS**
 - A **DOMAIN** is a structurally compact, independently folding unit that forms a stable 3D structure and shows a certain level of conservation
- We use the **hidden Markov model (HMM)** of a domain as found in the PFAM database <http://pfam.wustl.edu/>

05/22/2005

99



GO Functional Annotation (cont.)

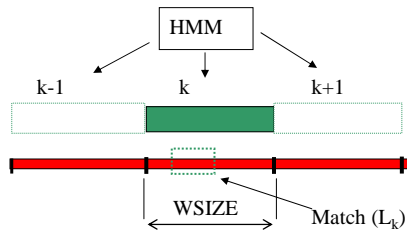
- **Problem statement:** find the functions an unknown gene product P_{unk} using a set of N genes $\{P_k\}_{k=1,N}$ with known functions
- **Approach:**
 - Use **Gene Ontology** annotations
 - Use HMM representation of protein domains (PFAM) to compute the similarity
 - Use **fuzzy K-nearest neighbor** to find k-most similar gene products to the unknown one
 - Score the annotation algorithm using a receiver-operator characteristic (**ROC**) curve.

05/22/2005

100



How Do We Extract Domain Features?



$$M_i = \frac{1}{\text{WSIZE}} \sum_{k=1}^{\lfloor L/\text{WSIZE} \rfloor + 1} L_k$$

- Use each HMM as a feature extractor (as implemented in [hmmer-http://hmmer.wustl.edu/](http://hmmer.wustl.edu/))
- Use a sliding window: $\text{WSIZE} = \text{length}(\text{HMM})$
- For each window k, record the length of the match L_k that has a $\text{log-likelihood} > \text{THRESHOLD}$
- M_i = the amount of match ($\in \mathbb{R}$) of domain i in the sequence
- A unknown sequence P:
 $P = \{M_1, \dots, M_N\}$

Computation of the Similarity

- $P_1 = \{M_{11}, \dots, M_{1N}\}, P_2 = \{M_{21}, \dots, M_{2M}\}$
- Could use set-based similarity measures if we consider only $M_{ij} > \text{THRESHOLD}$.
- If we use vector representation ($M_{ij} \geq 0$, hence $N=M=ND$), the domain similarity $s_{\text{DOM}}(P_1, P_2)$ is:

$$s_{\text{DOM}}(P_1, P_2) = \frac{\sum_{i=1}^{ND} \min(M_{1i}, M_{2i})}{\sum_{i=1}^{ND} \max(M_{1i}, M_{2i})}$$

Example of Domain Similarity

- Two collagen genes, COL1A2 (collagen 1 alpha 2) and COL21A1 (collagen 21 alpha 1), contain ND=3 PFAM domains, namely:
 - COLLAGEN(“Collagen triple helix repeat”)
 - COLFI(“Fibrillar collagen C-terminal domain”)
 - VWA(“von Willebrand factor type A domain”).
- The domain representation is $P_1 = \text{COL1A2} = (18,1,0)$ and $P_2 = \text{COL21A1} = (6,0,1)$, the above similarity is:

$$s(P_1, P_2) = \frac{(\min(18,6) + \min(1,0) + \min(0,1))}{(\max(18,6) + \max(1,0) + \max(0,1))}$$

$$= 0.3$$
- This low value is “good” since they are not in the same family

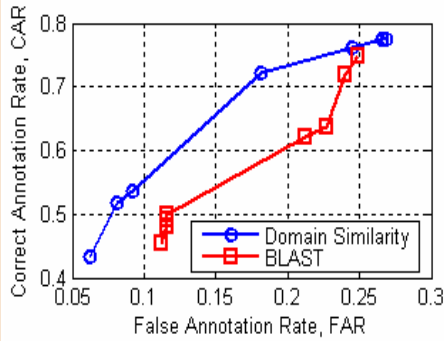
Automatic GO Annotation Algorithm

- Want to annotate an unknown gene product P_{Unk} .
- Algorithm:
 - Compute the ND features of the unknown sequence P_{Unk}
 - Compute the similarities between P_{Unk} and all N annotated gene products $\{s_{\text{Unk},k} = s_{\text{DOM}}(P_{\text{Unk}}, P_k)\}_{k=1,N}$
 - Pick K most similar gene products $\{P_k\}_{k=1,K}$
 - Use the similarities as fuzzy memberships in **fuzzy K-NN**

$$w(i) = \sum_{k=1}^K \frac{s_{\text{Unk},k} \delta_{ik}}{K}$$

- Annotate P_{Unk} with terms i for which $w(i) > \text{THRESHOLD}$

Results



- Use the previous data set (194 known gene products, containing 13 domains)
- Use a leave-one-out scheme
- Compute CAR, FAR for THRESHOLD=0.1...0.9
- **OBS: we can reach much lower FAR than BLAST for the same CAR**

$$CAR = \frac{|\{t \mid t \in T_{True} \cap T_{Computed}\}|}{|T_{True}|}$$

$$FAR = \frac{|\{t \mid t \in T_{Computed} - T_{True}\}|}{|T_{Computed}|}$$

05/22/2005

105



IV.3 Functional Summarization Using the GO

- Given a group of N gene products, find M < N Gene Ontology terms that describes them (microarray experiments)
- **Algorithm:**
 1. Compute the similarity matrix between the N gene products
 2. Cluster the gene products in M clusters (M could be determined using a cluster validity measure)
 3. Represent each cluster using i ∈ [1, M] the most frequent term found in cluster i.

05/22/2005

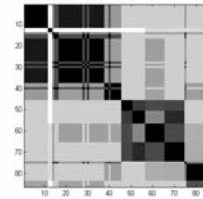
106



Functional Summarization Using the GO: Example

COL1A2, COL21A1, COL24A1, COL27A1, COL2A1, COL3A1,
COL4A1, COL4A2, COL4A3, COL4A6, COL5A3, COL9A1, COL9A2

Cluster	1	2	3
FMS	5581/1 (collagen)	5587/1 (collagen type 4)	5594/1 (collagen type 9)
BLAST	16740/1 (transferase activity)	5201/0.95 (extracellular matrix structural constituent)	5201/1 (extracellular matrix structural constituent)



FMS-based Clusters Produce More Specific Summaries

IV.4 Hot Applications

- i. Methylation microarrays
- ii. Learning biochemical networks from microarrays

IV.4.i Epigenetic Alterations in Cancer Hot Off the Press (For Us)

“A study of heritable changes that modulate chromatin organization and gene expression without changes in DNA sequences”

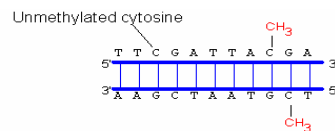
05/22/2005

109



Differential Gene Methylation Microarray Data

- **Gene expression** = whether the gene was transcribed or not
- **Methylation** = method of controlling gene expression
- **Mechanism:** an enzyme (methyltransferase) “tags” cytosine with a methyl group
- **Outcome:** If the promoter region (rich in CG) of a gene is heavily methylated, the gene is not expressed
- **Reason:** Not every gene should be expressed in every cell of our bodies (don’t want our brain cells to make hemoglobin, the protein required to carry oxygen around in our blood)

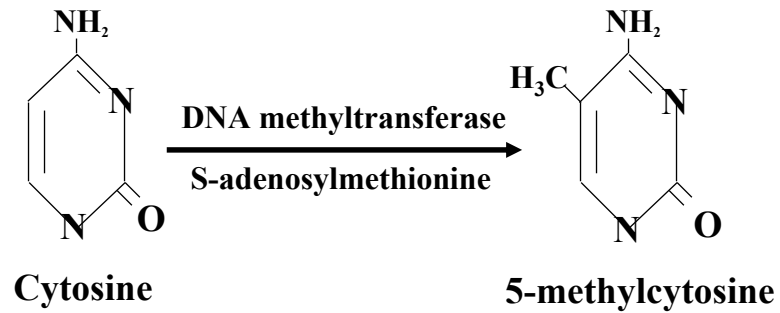


05/22/2005

110



CpG Methylation

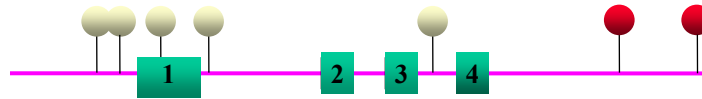


Some Facts Related to CpG Islands

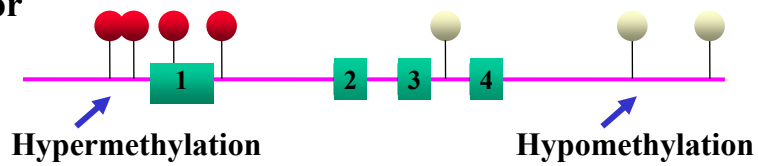
- ~28,529 CpG islands per haploid genome in humans.
(Based on Celera Sequencing data)
- 40-50% of all genes are associated with CpG islands.
- Other CpG islands are located in regions containing no genes.
 - Most CpG islands are unmethylated in normal cells
 - Exception
 - Genes on the inactive X chromosome
 - Imprinted genes
 - Repeated sequences or transposable elements

Aberrant DNA Methylation in Cancer

Normal



Tumor



05/22/2005

113



Hypothesis

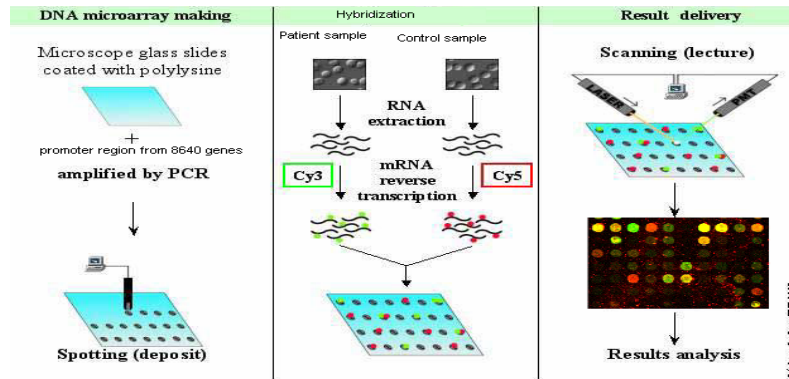
- CpG island hypermethylation is heritable in tumor cells
- Multiple methylated loci are progressively accumulated during tumorigenesis
- As a result, tumor cells can generate unique epigenetic signatures that are associated with specific cancer subtypes

05/22/2005

114



Methylation DNA Microarray



- Sample (patient) and control genes (normal) are marked with different fluorescent dyes (Cy5-red, Cy3-green)
- Use a scanner to obtain for each spot 3 values: R, G, B
- Advantage vs. Chip microarray: can select your own genes
- Disadvantage: noisier data due to quality control problems

05/22/2005

115



Lymphoma Experiment

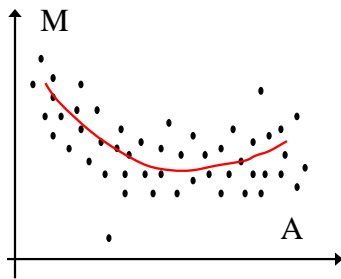
- **4 groups of patients:**
 - Hyperplasia, normal (HP)-3 patients
 - 3 types of lymphoma
 - chronic lymphocytic leukemia (CLL) –16 patients
 - Follicular lymphoma (FL) – 15 patients
 - Mantle cell lymphoma (MCL)-12 patients
- **The number of genes investigated is 8640.**
- **Goals:**
 - Improve the accuracy of lymphoma classification
 - Find differentially methylated genes
- **Questions:**
 - What are the genes that are uniquely methylated in each group?
 - What are the genes that are uniquely unmethylated in each group?
 - Can we cluster the patients such that we match the conventional pathologically determined lymphoma diagnoses?

05/22/2005

116



Normalization of Methylation DNA Microarray



- Many normalization procedures
- Use a goal-driven approach to select best normalization: select the normalization that produces 4 clusters of patients that match best the pathologically determined lymphoma diagnoses
- Intensity-dependent normalization
 - $M = \log R - \log G$
 - $A = \frac{1}{2}[\log R + \log G]$
 - Fit a curve (LOWESS, loess)¹: $L(A)$
 - Normalize: $M - L(A)$

1.Y.H. Yang *et al*, Nucl. Acid. Res. 30 (2002)

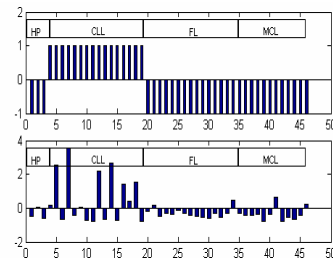
05/22/2005

117



Lymphoma Patient Clustering

	P ₁	P ₂	P ₃	P ₄	P ₅	P ₆
gs ₁	5	6	1	0	1	0
gs ₂	8	7	1	0	1	1
gs ₃	0	0	5	7	1	0
gs ₄	1	0	6	9	0	0
gs ₅	1	1	0	0	6	8
gs ₆	1	1	0	0	4	6
gs ₇	0	1	0	0	1	0



- Select gene uniquely hypermethylated in one group
- Use a modified “idealized expression pattern” algorithm (Golub, 1999) *: compute the correlation between a gene profile and the “idealized” profile
- Use again a goal-driven approach
- Select 40 genes in each group=>each patient has 160 features
- Compute cross-correlation between patients

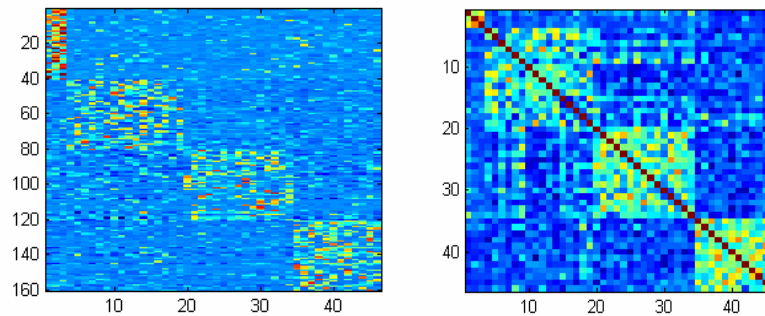
*Golub, TR, et al., Science, vol 286, 531:537, 1999

05/22/2005

118



Lymphoma Patient Clustering-Results



- Clustering was performed using **FCM** (Claverie 1999)
- The clustering of the patients based on the selected 160 genes was able to **match perfectly** the pathologically determined lymphoma classes.
- Initial evaluation indicates that the identified genes are indeed involved in essential cellular processes including **apoptosis**, and **proliferation**

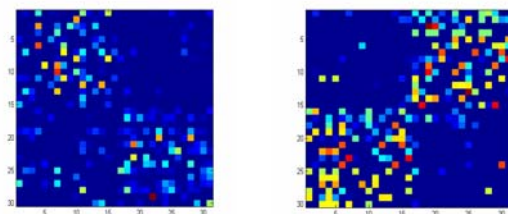
05/22/2005

119



Simultaneous Gene Selection From Methylation and Expression Microarrays

- Data set: 31 expression microarray and 31 methylation microarray from two types of lymphoma: CLL and FL
- Question: select genes that are not expressed but methylated for each type of lymphoma



Results: Genes exclusively methylated and not expressed in FL :
PSMB4, LRP1B, TSPY1/2, EIF4EBP1, MYOD1, MNAT1

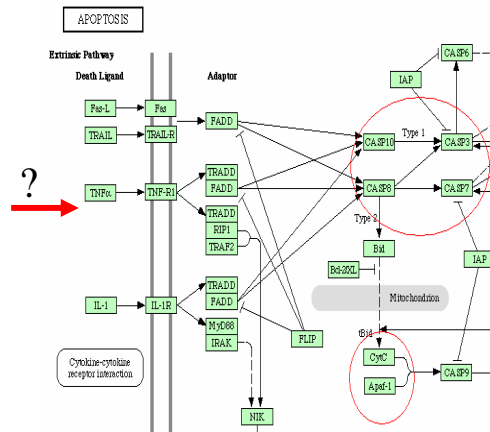
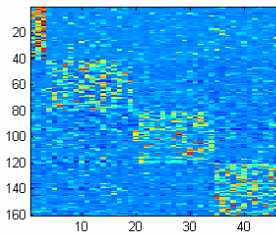
05/22/2005

120



IV.4.ii Learning Signaling Pathways From Microarray Data

- For each lymphoma, find the unique gene hypermethylation pattern of signaling pathways such as apoptosis and cell proliferation



05/22/2005

121



Knowledge Discovery References

- Protein clustering**
 - Enright, A.J., Van Dongen, S. and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families, *Nucleic Acids Research*, 30(7).
 - Hanisch, D., Zien, A., Zimmer, R. and Lengauer, T. (2002). Co-clustering of biological networks and gene expression data, *Bioinformatics*, 18, Suppl. 1.
 - Raychadhuri, S. and Altman, R.B. (2003). A literature-based method for assessing the functional coherence of a gene group, *Bioinformatics*, 19(3).
- Functional annotation**
 - Renner, A., and Aszodi, A. (2000). "High-throughput functional annotation of novel gene products using document clustering", In Proc. 6th Pacific Symposium on Biocomputing (PSB 2000).
 - TR Hvidsten, J. Komoroski, AK Sandvick, A. Legreid, "Predicting gene function from gene expression and ontologies", <http://www.smi.stanford.edu/projects/helix/psb01/hvidsten.pdf>.
 - MA. Andrade, NP Brown, C Leroy, S. Hoersch, et al., "Automated genome sequence analysis and annotation", *Bioinformatics*, vol. 15, no. 5, 1999.
 - S. Moller, W Fleischmann, R Apweiler, "EDITtoTrEMBL: a distributed approach to high-quality automated protein sequence annotation", *Bioinformatics*, vol. 15, no. 3, 1999.
 - TR Hugh, MJ Marton, AR Jones, CJ Roberts, et al., "Functional discovery via a compendium of expression profiles", *Cell*, vol. 102, July 7, 2000.
 - E. Kretschmann, W Fleischmann, R Apweiler, "Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied to SWISS-PROT", *Bioinformatics*, vol. 17, no. 10, 2001.
 - ALC Bazan, PM Engel, LF Schroder, SC da Silva, "Automated annotation of keywords for proteins related to mycoplastataceae using machine learning techniques", *Bioinformatics*, vol. 18, Suppl. 2, 2002.
 - AJ Perez, A. Rodriguez, G Thode, "A computational strategy for protein function assignment which addresses the multidomain problem", *Comparative and Functional Genomics*, vol. 3, 2002.
 - S. Khan, G. Situ, K. Decker, CJ Schmidt, "GoFigure: Automated Gene Ontology annotation", *Bioinformatics*, vol. 19, no. 18, 2003.
 - AJ Perez, G. Thode, O Trelles, "AnaGram: protein function assignment", *Bioinformatics*, vol. 20, no.2, 2004.
 - Y Huang, Y. Li, "Prediction of protein locations using fuzzy k-NN method", *Bioinformatics*, vol. 20, no. 1, 2004.
 - A. Prlc, FS Domingues, P Lackner, MJ Sippl, "WILMA-automated annotation of protein sequences", *Bioinformatics*, vol. 20, no. 1, 2004.

05/22/2005

122



Knowledge Discovery References (Cont.)

- **Functional summarization**
 - L.Y. Lee, J.M. Ho, and W.C. Lin. "An algorithm for generating representative functional annotations based on Gene Ontology", *Proceedings, DEXA'03, Prague, Czech Republic, Sept 2003*.
 - C.A. Joslyn, S.M. Mniszewski, A. Fulmer, and A. Heaton. "The Gene Ontology Categorizer", *Bioinformatics*, vol. 20 Suppl. 1 2004, pp 69–77.
- **Microarrays**
 - Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., and Golub, T.R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96, 2907–2912.
 - Dor, A., Shamir, R., and Yakhini, Z. (1999). Clustering gene expression patterns. *J. Comp. Biol.* 6(3–4), 281–297.
 - Zhou, X., Wang, X., Dougherty, E.R., Russ, D. and Suh, E. (2004). Gene Clustering Based on Clusterwise Mutual Information, *J. Comp. Biol.* 11(1), 147–161.
 - Resson H, Reynolds R, Varghese RS. Increasing the efficiency of fuzzy logic-based gene expression data analysis, *Physiol Genomics*. 2003 Apr 16;13(2):107–17. Review.
 - Woolf PJ, Wang Y. A fuzzy logic approach to analyzing gene expression data. *Physiol Genomics*. 2000 Jun 29;3(1):9–15
 - Ando T, Suguro M, Hanai T, Kobayashi T, Honda H, Seto M. Fuzzy neural network applied to gene expression profiling for predicting the prognosis of diffuse large B-cell lymphoma., *Jpn J Cancer Res*. 2002 Nov;93(11):1207–12.
 - Wang J, Bo TH, Jonassen I, Myklebost O, Hovig E. Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data., *BMC Bioinformatics*. 2003 Dec 02;4(1):60.
 - Futschik ME, Reeve A, Kasabov N. Evolving connectionist systems for knowledge discovery from gene expression data of cancer tissue., *Artif Intell Med*. 2003 Jun;28(2):165–89.
- **Learning pathways**
 - Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M. (1999). Systematic determination of genetic network architecture. *Nature Genet.* 22, 281–285.
 - Sokhansanj BA, Fitch JP, Quong JN, Quong AA. Linear fuzzy gene network models obtained from microarray data by exhaustive search. *BMC Bioinformatics*. 2004 Aug 10;5(1):108.
 - Pickert L., Reuter I., Klawonn F., Wingender E., Transcription regulatory region analysis using signal detection and fuzzy clustering. *Bioinformatics*, vol 14, no 3, 1998, pp.244–251
 - Creighton C., Hanash S., Mining gene expression databases for association rules. *Bioinformatics* vol 19, no 1, 2003, pp. 79–86
 - D'haeseleer P., Liang S., Somogyi R., Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, vol 16, no 8, 2000, pp. 707–726
 - Halfon, Marc S., and Alan M. Michelson. Exploring genetic regulatory networks in metazoan development: methods and models. *Physiol. Genomics* 10: 131–143, 2002;
 - Pettinen A., et al., Simulation tools for biochemical networks: evaluation of performance and usability, *Bioinformatics* vol 21, no 3, 2005, pp 357–363

05/22/2005

123



Conclusions

- **Introduced Soft computing methods to determine gene product similarity from taxonomy terms**
 - Use fuzzy measures on (augmented) term intersection set
 - Have fuzzy integrals to fuse confidence and “worth” (very general)
 - Investigating other combination schemes
- **Results can (should) be combined with sequence information, e.g., domains and motifs, and expression values for robust similarity**
- **Next steps**
 - Apply to new database of hand curated (RefSeq) proteins (~9000 proteins/~6000 Annotated)
 - Clustering and classification on microarray data
Expression and Hyper/Hypo Methylation
- **Knowledge Discovery**
 - Do the clusters found exhibit linguistic similarity?
 - Unknown gene product maps into cluster by sequence: share the linguistic properties?



05/22/2005

• **You should
always thank
your friends:**



- **National Library of Medicine Biomedical and Health Informatics Research Training grant 2-T15-LM07089-11 supporting M. Popescu**
- **And all of you!**

05/22/2005

125

