KDBIO

Knowledge Discovery and BIOinformatics Group

Identification of Regulatory Modules in Time-Series Gene Expression Data using Biclustering Algorithms

Sara C. Madeira

INESC-ID

Universidade da Beira Interior (UBI)

Instituto Superior Técnico (IST)

joint work with

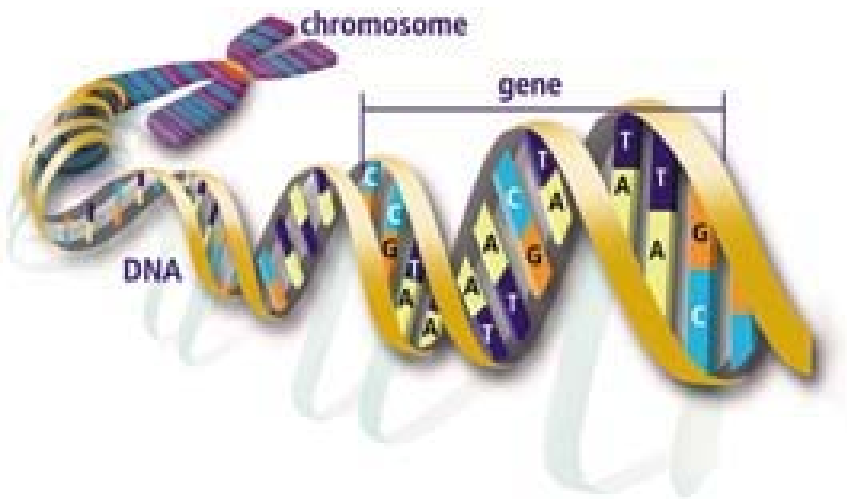
Arlindo L. Oliveira

Outline

- Motivation
- Gene Expression Data
- Biclustering (TCBB'04)
- **Biclustering in Time-Series Expression Data**
  - CCC-Biclustering (WABI'05) **$O(|R||C|)$ linear !**
  - *e*-CCC-Biclustering (APBC'07) **$O(|R|^2|C|^{1+e}|\Sigma|^e)$ polynomial !**
  - Statistical Significance of CCC-Biclusters and Similarity Measure
- **Application to the Identification of Regulatory Modules**

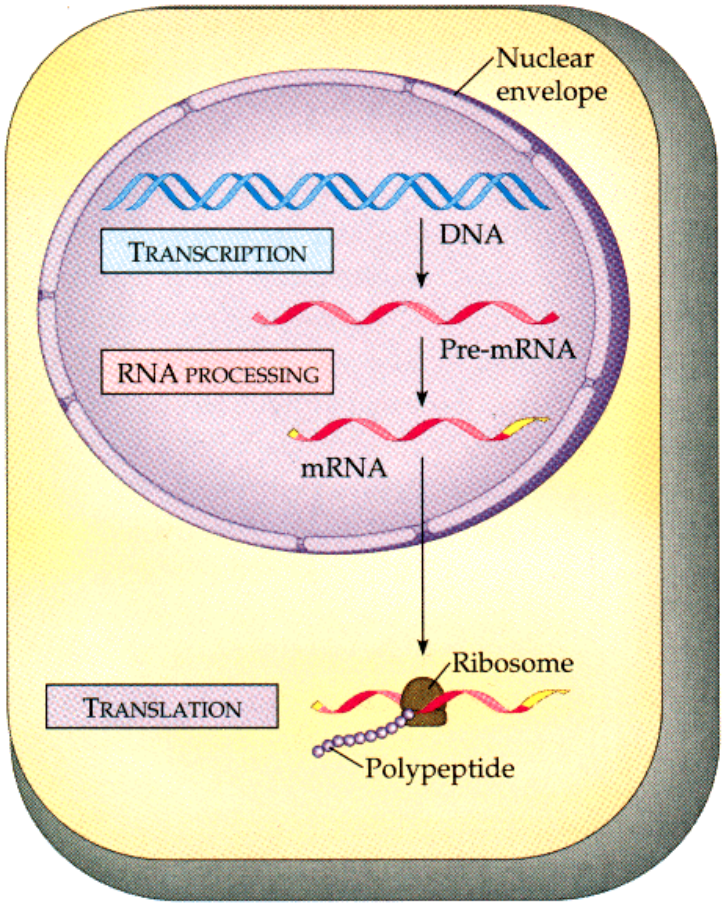
Motivation

- Genes are units in the cells of an organism.



- The same genes are present in every cell in the organism, but they work (express themselves) diferently.
- **Different genes have different expression levels according to their specific function at each time point.**

Motivation



- Genes code for proteins through the intermediary of mRNA.
- mRNA directs the production of cellular proteins (although protein synthesis and activation are not regulated solely at the mRNA level in the cell).
- mRNA measurement can be used to estimate cellular changes in response to external signals or environmental changes.
- Measuring mRNA is critical to the study of gene expression.

The Central Dogma of Molecular Biology

DNA → RNA → Protein

Gene Expression Data

- **Microarrays** measure the expression level of number of genes in different experimental conditions (samples).
- **Gene Expression Data** is arranged in a matrix.
  - **Each gene corresponds to a row and each condition to a column.**
  - **Each element is the expression level of a gene under a specific condition.**
- The **conditions** may correspond to:
  - Different environmental conditions.
  - Different stress conditions.
  - Different tissues / organs / individuals.
  - **Different time-points (Time-Series Expression Data).**

Gene Expression Data Analysis

- **Grouping of genes -** according to their expression under multiple conditions.
- **Grouping of conditions -** based on the expression of a number of genes.
- **Classification of a new gene -** given its expression and that of other genes with known classification.
- **Classification of a new condition (sample) -** given the expression of genes.
- **Identification of Biological Processes.**
  - **Identification of Regulatory Modules - sets of co-regulated genes that share a common biological function.**
  - **Identification of Gene Regulatory Networks.**

What is Biclustering? (TCBB'04)

- **Simultaneous Clustering of both rows and columns of a data matrix.**
  - **Biclustering** - Identifies groups of genes with similar/coherent expression patterns under a **specific subset of the conditions**.
  - **Clustering** - Identifies groups of genes that show similar activity patterns under **all the set of conditions.**
- **|R| by |C| data matrix** *A = (R,C)*
  - $R=\{r_1,\ldots, r_{|R|}\}$ = Set of |R| rows.
  - $C=\{y_1,\ldots, y_{|C|}\}$ = Set of |C| columns.
  - A_{ij}=relation between row *i* and column *j*.
- **Gene expression matrices**
  - ***R* = Set of Genes**
  - **C = Set of Conditions.**
  - A_{ij} **= expression level of gene *i* under condition *j* (quantity of mRNA).**

	Cond 1	…	Cond *j*	…	Cond.*m*
Gene 1	…	…	…	…	…
…	…	…	…	…	…
Gene *i*	…	…	A_{ij}	…	…
…	…	…	…	…	…
Gene *n*	…	…	…	…	…

Biclustering *vs* Clustering

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
G_1	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}	a_{18}	a_{19}	a_{110}
G_2	a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}	a_{27}	a_{28}	a_{29}	a_{210}
G_3	a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a_{36}	a_{37}	a_{38}	a_{39}	a_{310}
G_4	a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a_{46}	a_{47}	a_{48}	a_{49}	a_{410}
G_5	a_{51}	a_{52}	a_{53}	a_{54}	a_{55}	a_{56}	a_{57}	a_{58}	a_{59}	a_{510}
G_6	a_{61}	a_{62}	a_{63}	a_{64}	a_{65}	a_{66}	a_{67}	a_{68}	a_{69}	a_{610}

$R = \{G_1, G_2, G_3, G_4, G_5, G_6\}$

$C = \{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}\}$

$I = \{G_2, G_3, G_4\}$

$J = \{C_4, C_5, C_6\}$

Cluster of Conditions (R,J)

$(R, \{C_4, C_5, C_6\})$

Cluster of Genes (I,C)

$(\{G_2, G_3, G_4\}, C)$

Bicluster (I,J)

$(\{G_2, G_3, G_4\}, \{C_4, C_5, C_6\})$

Why Biclustering and not just Clustering?

Global Model

- When Clustering algorithms are used
  - Each gene in a given gene cluster is defined using all the conditions.
  - Each condition in a condition cluster is characterized by the activity of all the genes.

Local Model

- When Biclustering algorithms are used
  - Each gene in a bicluster is selected using only a subset of the conditions
  - Each condition in a bicluster is selected using only a subset of the genes.

Why Biclustering and not just Clustering?

- Unlike Clustering
  - Biclustering identifies groups of genes that show similar activity patterns under a specific subset of the experimental conditions.
- Biclustering is the key technique to use when
  1. Only a small set of the genes participates in a cellular process of interest.
  2. An interesting cellular process is active only in a subset of the conditions.
  3. A single gene may participate in multiple pathways that may or not be co-active under all conditions.

Biclustering Time-Series Gene Expression Data

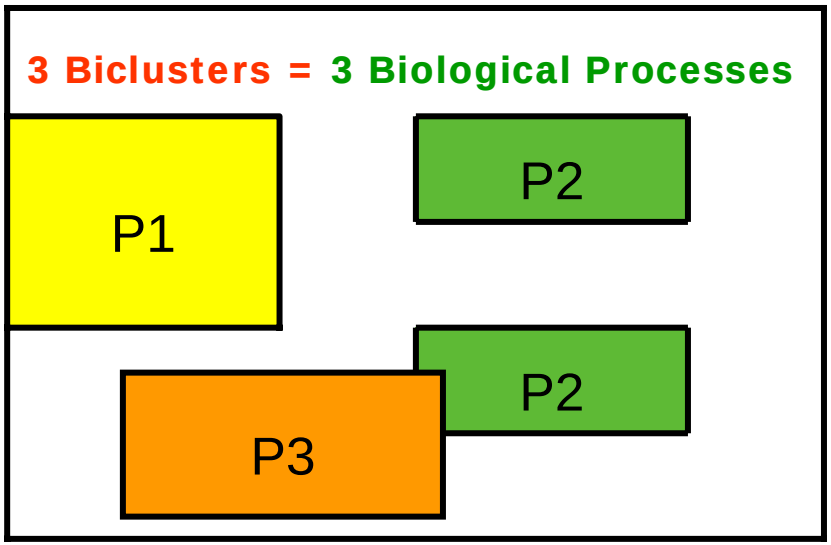
- **Biclustering Problem: NP-complete [Peeters, 2003].**

- **In this case, it is reasonable to restrict to biclusters with contiguous columns.**

  =>Tractable problem!

- **Assumption:** The activation of a set of genes under specific conditions corresponds to the activation of a particular biological process.

- As time goes on, **biological processes** start and finish, leading to increased (or decreased) activity of genes that can be identified because they **form biclusters with contiguous columns.**



- We believe that the identification of biological processes that lead to the creation of the biclusters is crucial for the **identification of gene regulatory networks.**

Discretizing Time-Series Gene Expression Data

Case of Interest: gene expression levels can be *discretized* to a **set of symbols ∑** (set of distinct activation levels)

- **∑ = {D, N, U}** = {Down-Regulated, No-Change, Up-Regulated}

Matrix A'	C1	C2	C3	C4	C5
Gene 1	0.07	0.73	-0.54	0.45	0.25
Gene 2	-0.34	0.46	-0.38	0.76	-0.44
Gene 3	0.22	0.17	-0.11	0.44	-0.11
Gene 4	0.70	0.71	-0.41	0.33	0.35

Gene Expression Matrix

Matrix A	C1	C2	C3	C4	C5
Gene 1	N	U	D	U	N
Gene 2	D	U	D	U	D
Gene 3	N	N	N	U	N
Gene 4	U	U	D	U	U

Discretized Expression Matrix

CCC-Biclusters (WABI'05)

- A **Bicluster** is a subset of rows $I = \{i_1,\ldots,i_k\}$ and a subset of columns $J=\{j_1,\ldots,j_s\}$ from matrix A, such that it can be defined as a ***k* by *s* sub-matrix of matrix *A***.

- A **Trivial Bicluster** is a Bicluster with only one row or only one column.

- A **CC-Bicluster** (**Coherent Column Bicluster**) is a subset of rows $I = \{i_1,\ldots,i_k\}$ and a subset of columns $J=\{j_1,\ldots,j_l\}$ from matrix *A* such that $\boldsymbol{A_{ij} = A_{lj}}$**, for all** $\boldsymbol{i \in I}$ **and** $\boldsymbol{j \in J}$ ***(constant columns).***

- A **CCC-Bicluster** (**Contiguous Column Coherent Bicluster**) is a subset of rows $I = \{i_1,\ldots,i_k\}$ and a **contiguous** subset of columns $J=\{j_r, j_{r+1} \ldots, j_{s-1}, j_s\}$ from matrix A such that $A_{ij} = A_{lj}$ for all $i \in I$ and $j \in J$ ***(contiguous constant columns)***.

Each CCC-Bicluster defines a **string *S*** that corresponds to an **Expression Pattern** common to every row in the CCC-Bicluster (between columns *r* and *s* of matrix A).

Maximal CCC-Biclusters

- A **CCC-Bicluster is Row-Maximal** if no more rows can be added to its set of rows *I* while maintaining the coherence property.

- A **CCC-Bicluster is Right-Maximal** if its expression pattern *S* cannot be extended to the right by adding one more symbol at its end (the column contiguous to its last column of cannot be added to *J* without removing genes from *I*).

- A **CCC-Bicluster is Left-Maximal** if its expression pattern *S* cannot be extended to the left by adding one more symbol at its beginning (the column contiguous to its first column of cannot be added to *J* without removing genes from *I*).

- A **CCC-Bicluster is Maximal** if it is Row-Maximal, Left-Maximal and Right-Maximal.

→ **NO** other CCC-Bicluster exists that properly contains it, that is, if for all other CCC-biclusters *(L,M)*, $I \subseteq L$ *and* $J \subseteq M \Rightarrow I = L \wedge J = M$.

Maximal Non-Trivial CCC-Biclusters

Each CCC-Bicluster defines a String corresponding to an Expression Pattern common to every row in the CCC-Bicluster.

Matrix A'	C1	C2	C3	C4	C5
Gene 1	0.07	0.73	-0.54	0.45	0.25
Gene 2	-0.34	0.46	-0.38	0.76	-0.44
Gene 3	0.22	0.17	-0.11	0.44	-0.11
Gene 4	0.70	0.71	-0.41	0.33	0.35

Matrix A	C1	C2	C3	C4	C5
Gene 1	N	U	D	U	N
Gene 2	D	U	D	U	D
Gene 3	N	N	N	U	N
Gene 4	U	U	D	U	U

B1 =({G1,G2,G4},{C2,C3,C4}, [UDU])

B2 = ({G1,G3},{C4,C5}, [UN])

After Alphabet Transformation …

Each CCC-Bicluster defines a String corresponding to an Expression Pattern common to every row in the CCC-Bicluster.

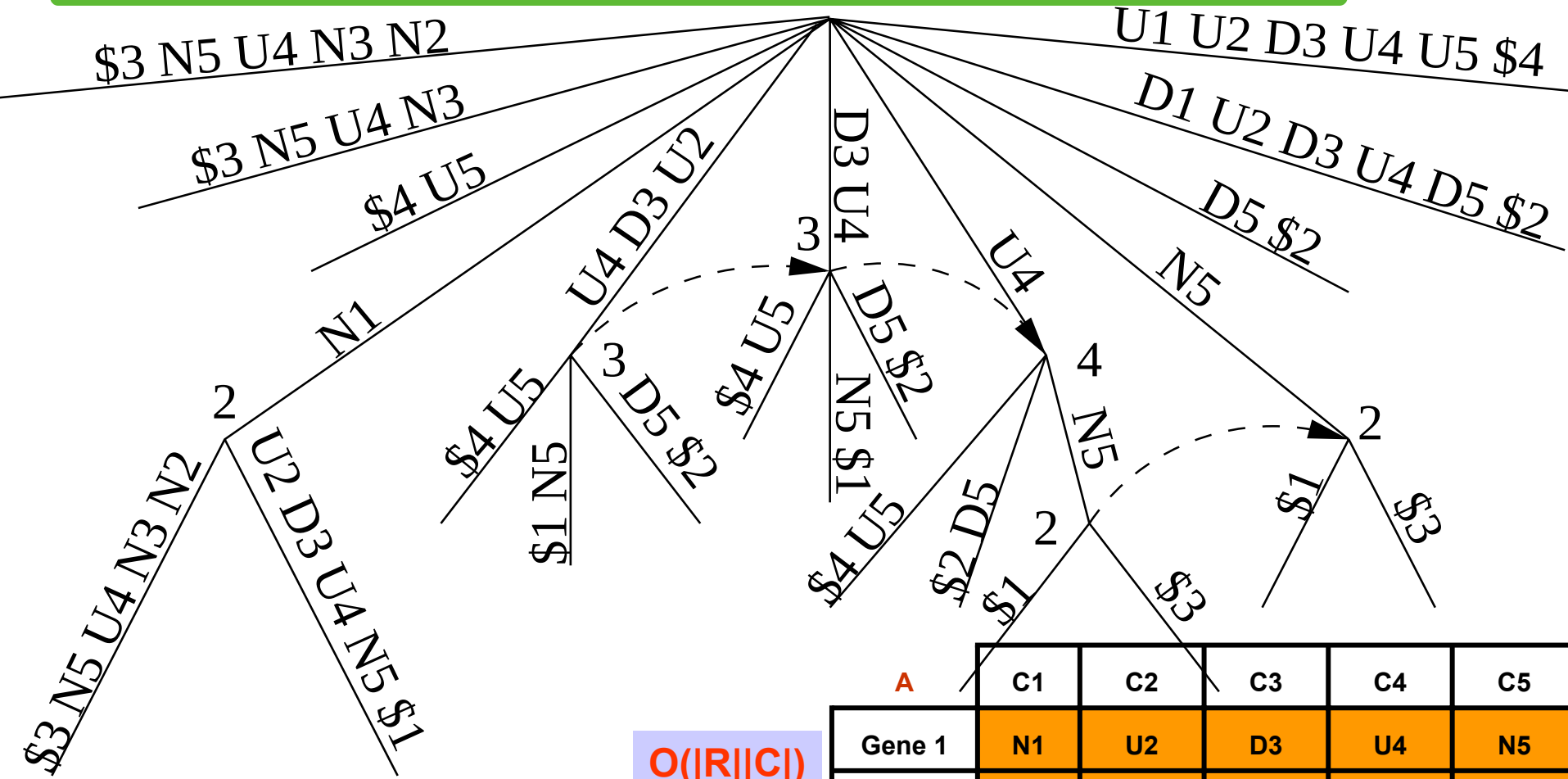
Matrix A	C1	C2	C3	C4	C5
Gene 1	N	U	D	U	N
Gene 2	D	U	D	U	D
Gene 3	N	N	N	U	N
Gene 4	U	U	D	U	U

Matrix A	C1	C2	C3	C4	C5
Gene 1	N1	U2	D3	U4	N5
Gene 2	D1	U2	D3	U4	D5
Gene 3	N1	N2	N3	U4	N5
Gene 4	U1	U2	D3	U4	U5

B1 =({G1,G2,G4},{C2,C3,C4}, **[UDU]**)

B2 = ({G1,G3},{C4,C5}, **[UN]**)

CCC-Biclustering and Generalized Suffix Trees (WABI'05)

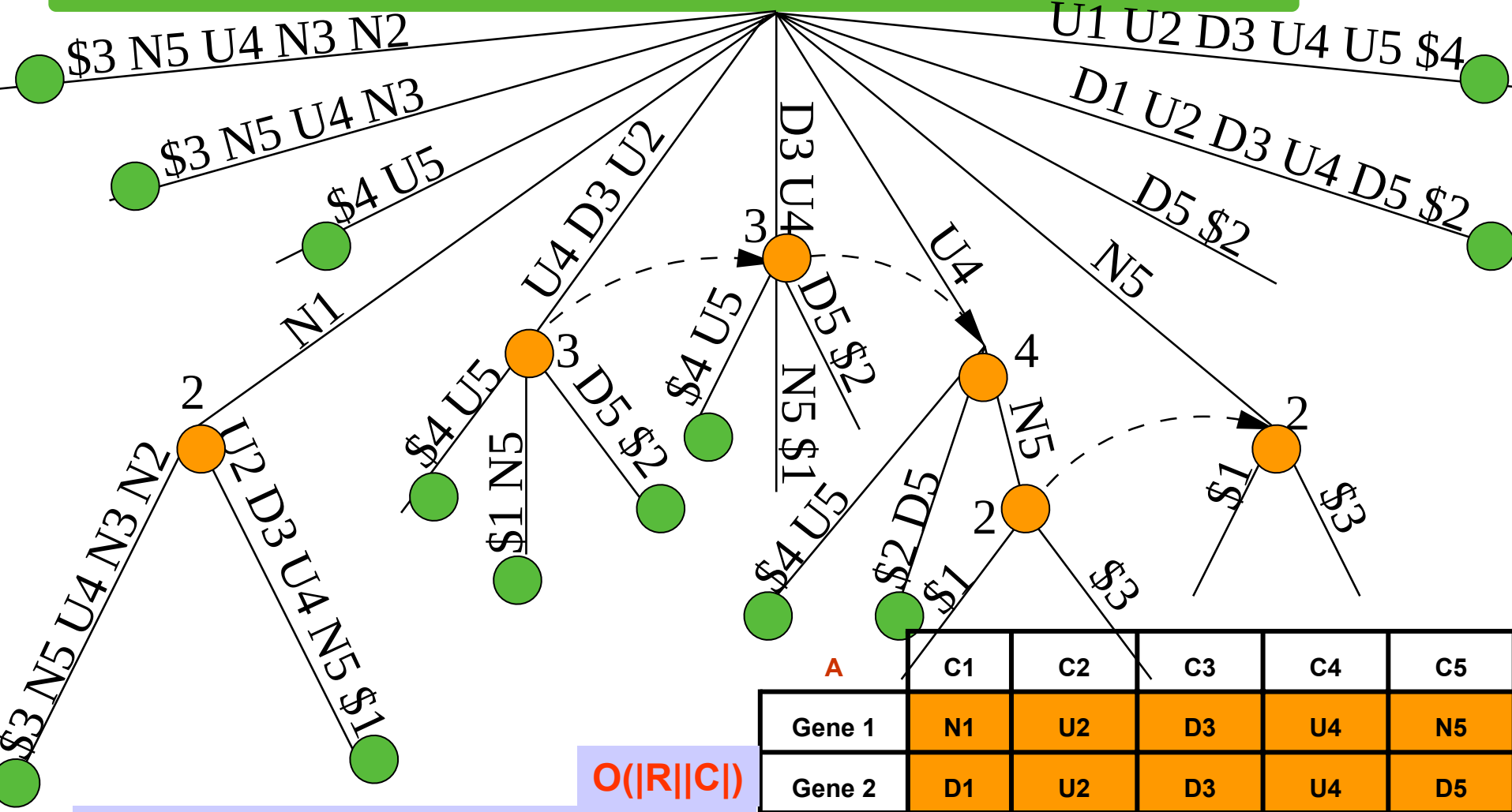


A	C1	C2	C3	C4	C5
Gene 1	N1	U2	D3	U4	N5
Gene 2	D1	U2	D3	U4	D5
Gene 3	N1	N2	N3	U4	N5
Gene 4	U1	U2	D3	U4	U5

O(|R||C|)

Build suffix tree with suffix links [Ukkonen, 1995]
Compute number of leaves of each node

CCC-Biclusters in the Generalized Suffix Tree

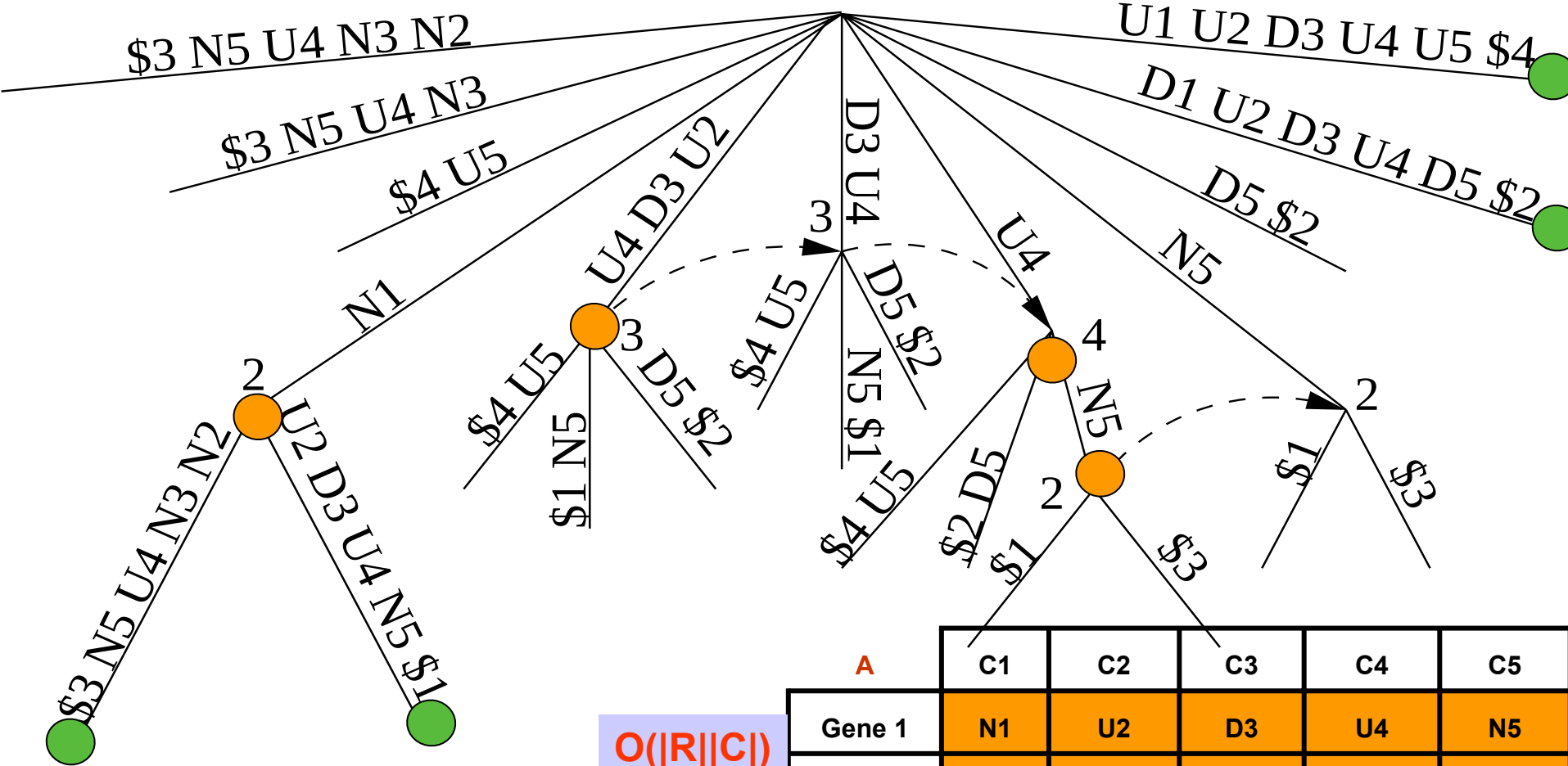


A	C1	C2	C3	C4	C5
Gene 1	N1	U2	D3	U4	N5
Gene 2	D1	U2	D3	U4	D5
Gene 3	N1	N2	N3	U4	N5
Gene 4	U1	U2	D3	U4	U5

O(|R||C|)

Mark nodes as **"valid"** CCC-Biclusters

Maximal CCC-Biclusters in the Generalized Suffix Tree

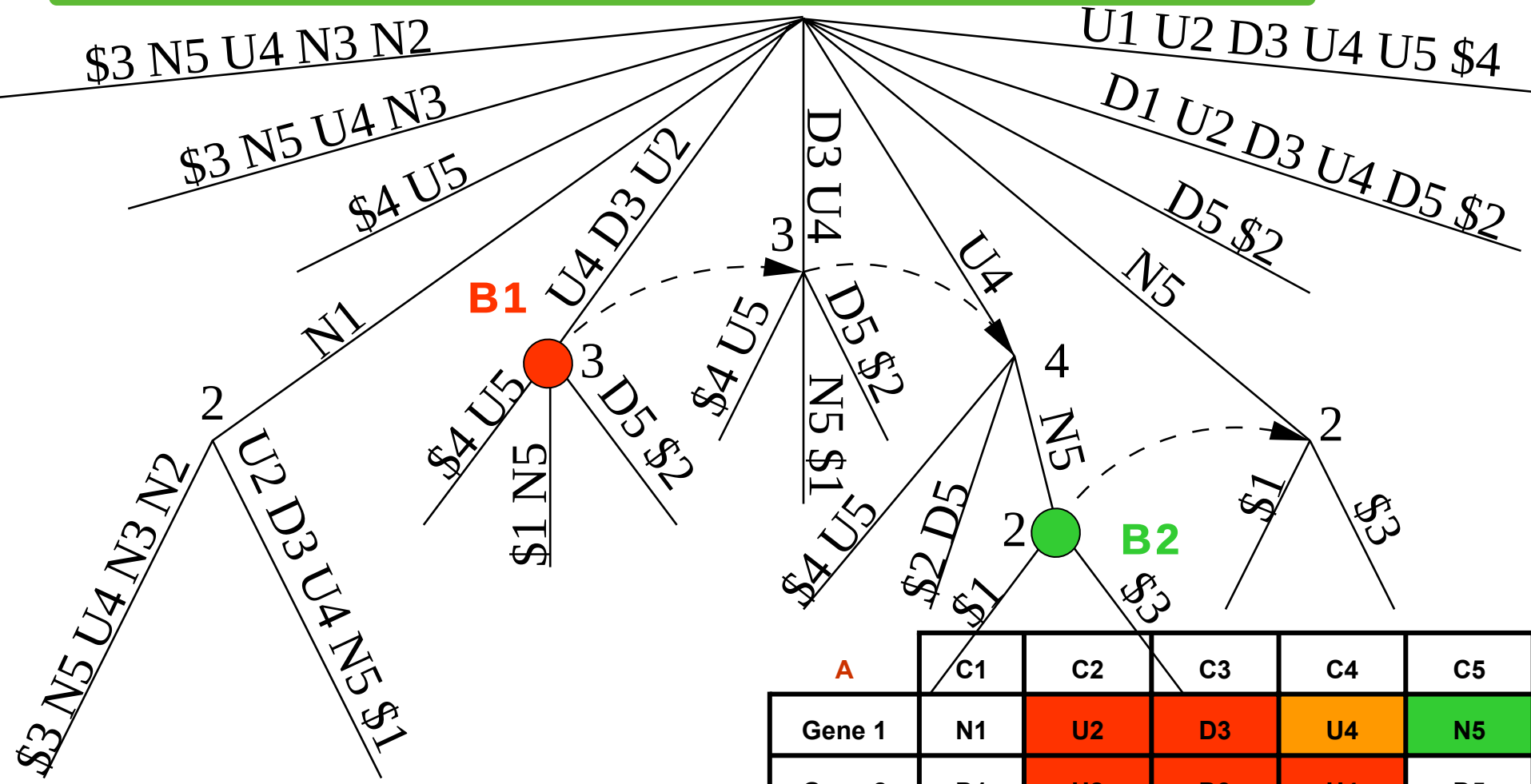


A	C1	C2	C3	C4	C5
Gene 1	N1	U2	D3	U4	N5
Gene 2	D1	U2	D3	U4	D5
Gene 3	N1	N2	N3	U4	N5
Gene 4	U1	U2	D3	U4	U5

O(|R||C|)

Mark nodes as **"invalid"** CCC-Biclusters
Report maximal CCC-Biclusters ("Valid")

Maximal Non-Trivial CCC-Biclusters in the Suffix Tree



A	C1	C2	C3	C4	C5
Gene 1	N1	U2	D3	U4	N5
Gene 2	D1	U2	D3	U4	D5
Gene 3	N1	N2	N3	U4	N5
Gene 4	U1	U2	D3	U4	U5

e-CCC-Biclustering (APBC'07)

- **Drawback of CCC-Biclustering**
  - All genes in the CCC-Bicluster have the same **(perfect) expression pattern**.
  - Efficiency can be reduced due to **noise** and **discretization errors**.

$O(|R||C|)$ *vs* $O(|R|^2|C|^{1+e}|\Sigma|^e)$

Linear *vs* polynomial !

- **Advantages of *e*-CCC-Biclustering**
  - Considers an **expression pattern** **(MOTIF)** for the *e*-CCC-Bicluster.
  - Genes have **approximate expression patterns in the *e*-Neighborhood of the motif.**
  - **General Errors -** Substitution of a symbol in the expression pattern by other symbols in the alphabet **(measurement errors).**
  - **Restricted Errors -** Substitution of a symbol in the expression pattern by its *k* lexicographically closer discretization symbols **(discretization errors)**.

e-CCC-Biclusters and Maximal *e*-CCC-Biclusters

- An **e-CCC-Bicluster** **(Contiguous Column Coherent Bicluster allowing *e* errors)**

  CCC-Bicluster where all the strings S_i defining the expression patterns of each of the genes in the bicluster are in the ***e*-Neighborhood** of an expression pattern *S* ***(Motif)*** definining the *e*-CCC-Bicluster.

- A *0*-CCC-Bicluster is equivalent to a CCC-Bicluster.

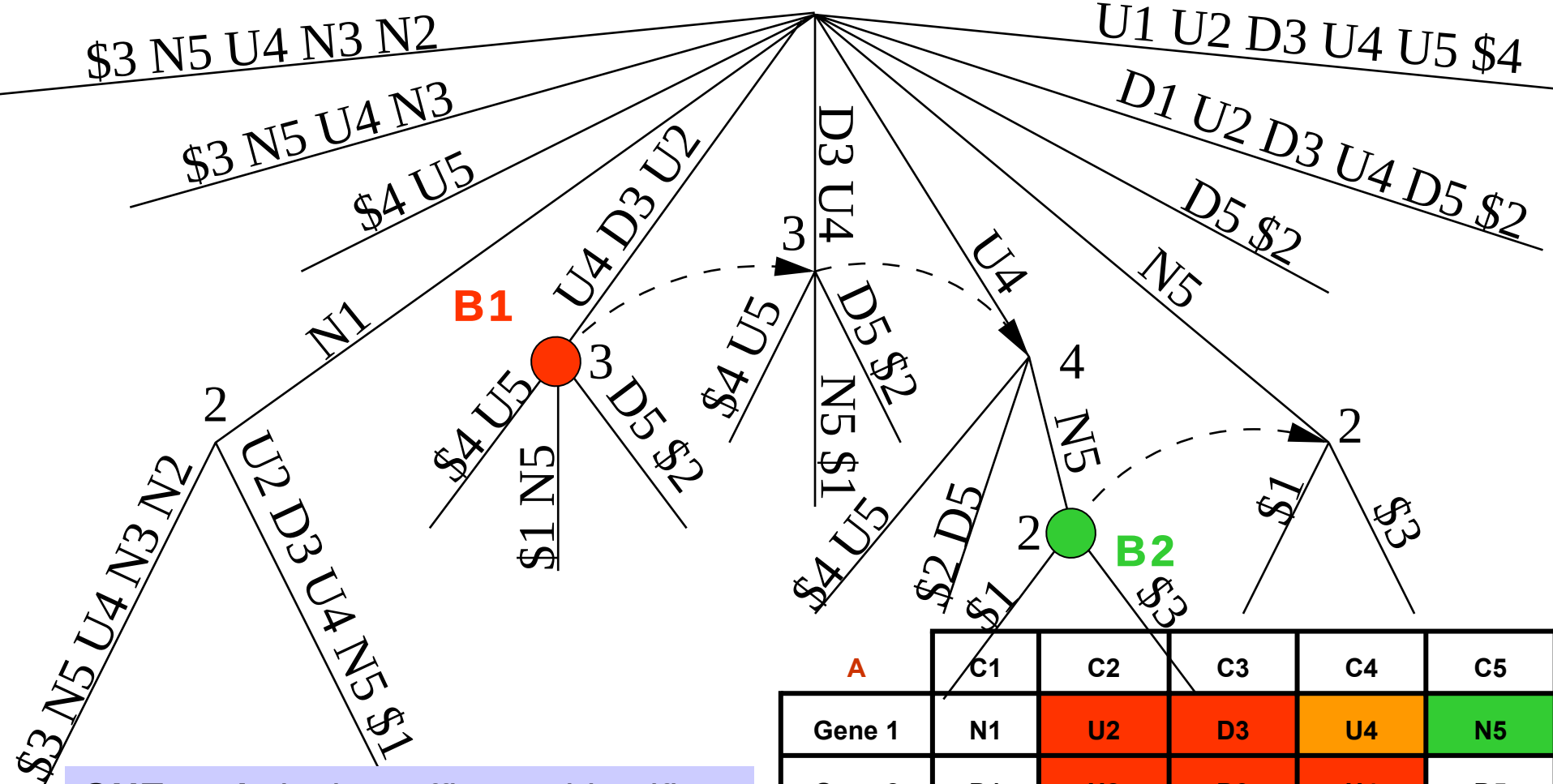
- An **e-CCC-Bicluster is Maximal**

  IF it is **Row-Maximal**, **Left-Maximal** and **Right-Maximal**.

  →No more rows or contiguous columns can be added to it while maintaining the coherence property (all expression patterns S_i are in the *e*-Neighborhood of motif S).

REMEMBER: When NO errors are allowed…

CCC-Biclustering and Generalized Suffix Tree



A	C1	C2	C3	C4	C5
Gene 1	N1	U2	D3	U4	N5
Gene 2	D1	U2	D3	U4	D5
Gene 3	N1	N2	N3	U4	N5
Gene 4	U1	U2	D3	U4	U5

<u>ONE</u> node in the suffix tree identifies **<u>ONE</u> CCC-Bicluster**, maximal or not !

Two Examples of Maximal Non-Trivial *1*-CCC-Biclusters

MOTIF = [D1 U2 D3 U4 N5]

Matrix A	C1	C2	C3	C4	C5
Gene 1	N1	U2	D3	U4	N5
Gene 2	D1	U2	D3	U4	D5
Gene 3	N1	N2	N3	U4	N5
Gene 4	U1	U2	D3	U4	U5

B1 =({G1,G2},{C1,C2,C3,C4,C5}, **[DUDUN]**)

MOTIF = [U2 D3 U4 D5]

Matrix A	C1	C2	C3	C4	C5
Gene 1	N1	U2	D3	U4	N5
Gene 2	D1	U2	D3	U4	D5
Gene 3	N1	N2	N3	U4	N5
Gene 4	U1	U2	D3	U4	U5

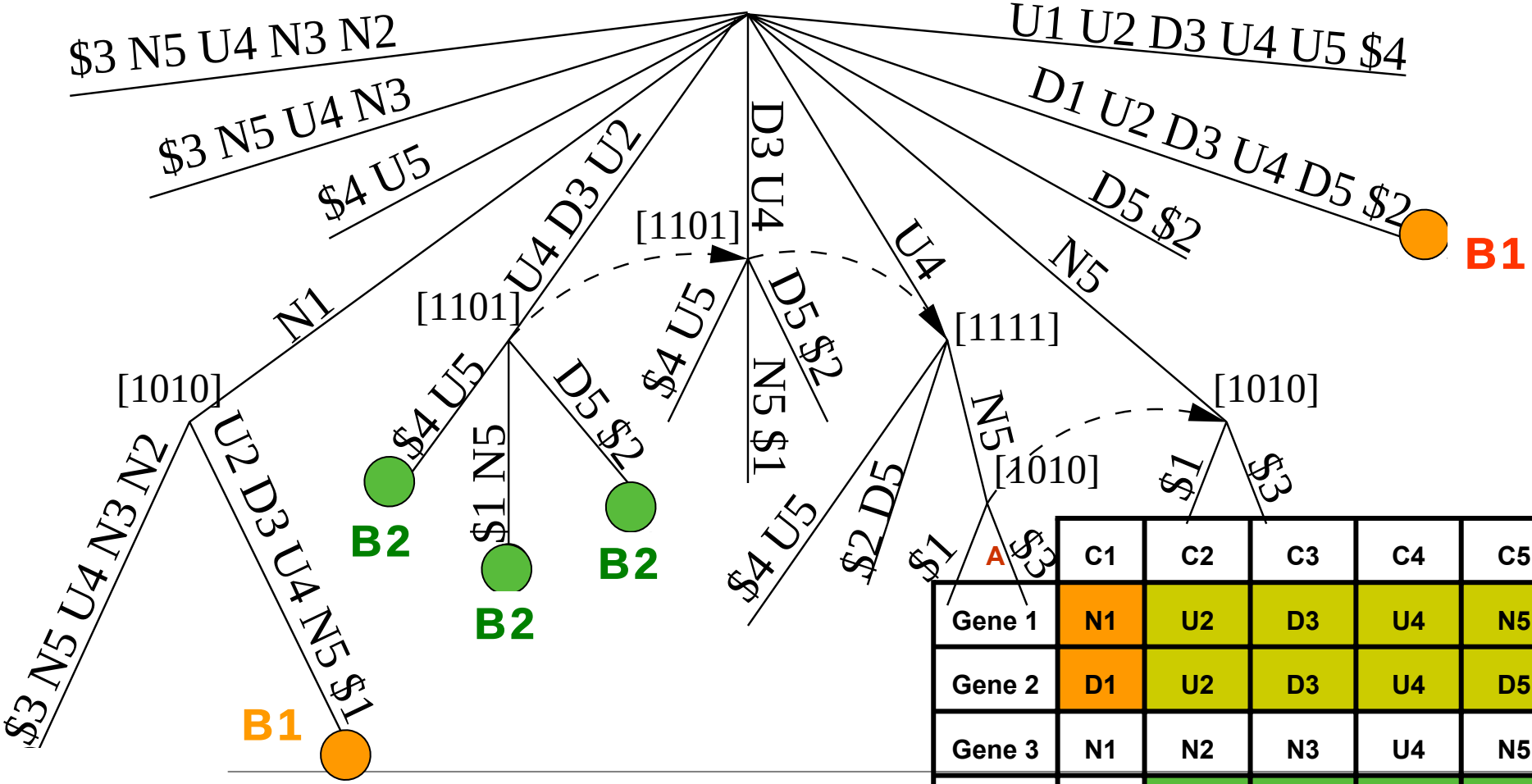
B2 =({G1,G2,G4},{C2,C3,C4,C5}, **[UDUD]**)

ONE ***e*-CCC-Bicluster** can be identified by **SEVERAL nodes** in the suffix tree!

When *e* errors are allowed…

e-CCC-Biclustering and Generalized Suffix Tree with Colors

STORE all models (motif + node-occurrences) with no more than *e* errors corresponding to **Right-Maximal *e*-CCC-Biclusters**: ADAPT **SPELLER** [Sagot, 98]



A	C1	C2	C3	C4	C5
Gene 1	N1	U2	D3	U4	N5
Gene 2	D1	U2	D3	U4	D5
Gene 3	N1	N2	N3	U4	N5
Gene 4	U1	U2	D3	U4	U5

Right-Maximal *e*-CCC-Biclusters can be NON Left-Maximal

REMOVE from the stored models those corresponding to **NON-Left-Maximal *e*-CCC-Biclusters**: USE A **TRIE** built with reversed motifs and store #genes sharing the motif in its nodes → a node with a child with as many genes as itself is not left-maximal !

MOTIF = [U2 D3 U4]

Matrix A	C1	C2	C3	C4	C5
Gene 1	N1	U2	D3	U4	N5
Gene 2	D1	U2	D3	U4	D5
Gene 3	N1	N2	N3	U4	N5
Gene 4	U1	U2	D3	U4	U5

B=({G1,G3,G4},{C1,C2,C3}, [UDU])

Non-Left Maximal

MOTIF = [D1 U2 D3 U4]

Matrix A	C1	C2	C3	C4	C5
Gene 1	N1	U2	D3	U4	N5
Gene 2	D1	U2	D3	U4	D5
Gene 3	N1	N2	N3	U4	N5
Gene 4	U1	U2	D3	U4	U5

B=({G1,G2,G4},{C1,C2,C3,C4}, [DUDU])

Left-Maximal

Different Motifs can Identify the Same *e*-CCC-Bicluster

REMOVE from the stored models those identifying the **SAME *e*-CCC-Bicluster:**
USE AN **HASH TABLE** → all models with the same first and last conditions and same genes identify the same *e*-CCC-Bicluster

Matrix A	C1	C2	C3	C4	C5
Gene 1	N1	U2	D3	U4	N5
Gene 2	D1	U2	D3	U4	D5
Gene 3	N1	N2	N3	U4	N5
Gene 4	U1	U2	D3	U4	U5

MOTIF = [D1 U2 D3 U4 N5]

Genes = {G1, G2}

Conditions = C1-C5

Matrix A	C1	C2	C3	C4	C5
Gene 1	N1	U2	D3	U4	N5
Gene 2	D1	U2	D3	U4	D5
Gene 3	N1	N2	N3	U4	N5
Gene 4	U1	U2	D3	U4	U5

MOTIF = [N1 U2 D3 U4 N5]

Genes = {G1, G2}

Conditions = C1-C5

Statistical Significance of CCC-Biclusters

- Given a CCC-Bicluster *B* of size |*I*|x|*J*| and expression pattern p_B
- Measure the statistical significance of *B* against the null hypothesis
  - The null hypothesis is the probability of B occurring by chance in a randomly generated expression matrix with |G| genes and |C| time-points.
- Compute this value using the ***tail of the binomial distribution***, P
  - P is the probability of an event with probability *p* occurring *k* or more times in *n* independent trials
- The statistical significance of *B* is the ***p*-value(*B*)**
  - Computed using the *tail of the binomial distribution* by obtaining the probability of a **random occurrence of the expression pattern p_B, k=|I|-1 times in n=|G|-1** independent trials

Statistical Significance of CCC-Biclusters

- Simplifying Assumption
  - The probability of occurrence of a specific expression pattern p_B is adequately modeled by a **first order Markov Chain**, with state transition probabilities obtained from the corresponding columns in the matrix

Example:

IF **B=({G1,G2,G4\},{C2,C3,C4},[UDU])**

THEN $\mathbf{p_B}$**=P(U2D3U4) =P(U2)P(D3|U2) P(U4|D3)**

WHERE **P(U2)= |U2|/|G|**

P(D3|U2)= P(U2D3)/P(U2) = |U2D3|/|U2|

P(U4|D3)= P(D3U4)/P(D3) = |D3U4|/|D3|

Similarity Measure between CCC-Biclusters

- Given two CCC-Biclusters, $B_1=(I_1,J_1)$ and $B_2=(I_2,J_2)$
- Computed using the **Jaccard Index: Size of the submatrix in the intersection between the submatrices representing B_1 and B_2**

$$J(B_1,B_2) = J((I_1,J_1), (I_2,J_2)) = |B_1 \cap B_2| / |B_1 \cup B_2|$$

  - $|B_1| = |I_1| \times |J_1|$
  - $|B_2| = |I_2| \times |J_2|$
  - $|B_1 \cap B_2| = |I_1 \cup I_2| \times |J_1 \cup J_2|$
  - $|B_1 \cup B_2| = |B_1| + |B_2| - |B_1 \cap B_2|$

Application to the Identification of Regulatory Modules

- ***Saccharomyces Cerevisiae* response to heat stress** (Gash'2000)
  - 5 Time-Points along the 1st hour of exposure to 37º (0', 5', 15', 30',60')
  - 1st time-point is an average of 3 replicates of time zero
- Discretized matrix using **three symbols {D, N, U}** and variations between successive time-points
- Found **167** maximal non-trivial CCC-Biclusters
- Sorted using *p*-value of statistical significance (***only* 25 passed the statistical test at 1% level after Bonferroni correction**)
- Filtered **highly overlapping** CCC-Biclusters (similarity score above 25%)
- **Analyzed the remaining 16 CCC-Biclusters**

Application to the Identification of Regulatory Modules

SUMMARY OF THE CCC-BICLUSTERS PASSING THE STATISTICAL TEST AT THE 1% LEVEL AFTER BONFERRONI CORRECTION (AFTER FILTERING CCC-BICLUSTERS WITH SIMILARITY ABOVE 25%)

ID	*Variation Pattern*	*#Time-Points (first-last)*	*#Genes*	*Sorting P-value*	*#p-values <0.01*	*# p-values 0.01≤<0.05*	*Best p-value (Level>2)*	*Dataset Frequency*
124	DNU	4(2-5)	904	2.56E-84	40	8	8.23E-63(7)	18.52
14	UND	4(2-5)	1091	1.64E-58	62	12	2,79E-24(5)	10.78
27	UUND	5(1-5)	290	3.69E-44	7	6	3.28E-08(3)	21.59
39	UNND	5(1-5)	258	8.65E-42	0	0	1.65E-04(3)	8.18
151	DNNU	5(1-5)	232	3.99E-31	12	2	3.19E-14(3)	93.26
48	UDUD	5(1-5)	182	1.35E-26	0	1	6.98E-05(3)	90.11
142	DUDU	5(1-5)	248	2.84E-24	8	19	4.37E-09(4)	41.27
43	UNDD	5(1-5)	109	6.56E-24	0	0	1.97E-04(11)	4.62
147	DNUU	5(1-5)	144	6.03E-21	0	3	4.50E-05(3)	87.07
83	NUNN	5(1-5)	224	1.90E-16	2	4	1.41E-05(6)	10.13
42	UNDN	5(1-5)	131	3.30E-11	2	1	6.85E-06(9)	4.44
148	DNUN	5(1-5)	192	6.00E-11	4	4	7.68E-07(3)	88.08
159	DDUU	5(1-5)	56	1.37E-07	0	0	1.14E-03(6)	13.64
79	NUUN	5(1-5)	97	4.41E-07	2	3	2.46E-06(3)	20.00
92	NNUN	5(1-5)	52	3.88E-05	2	0	1.64E-06(4)	27.27
99	NNDN	5(1-5)	39	4.79E-05	1	0	2.13E-05(6)	13.79

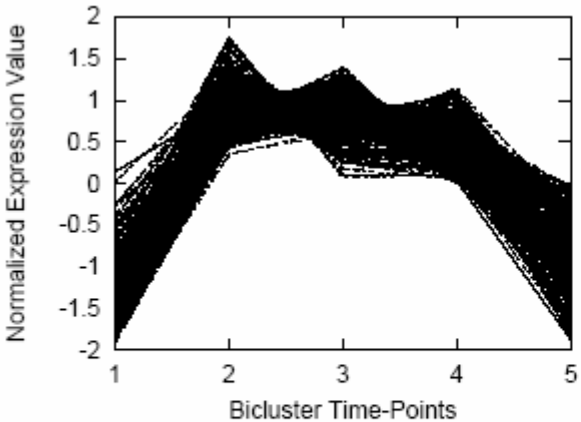
Application to the Identification of Regulatory Modules

(Joint work with Miguel C.Teixeira and Isabel Sá Correia)

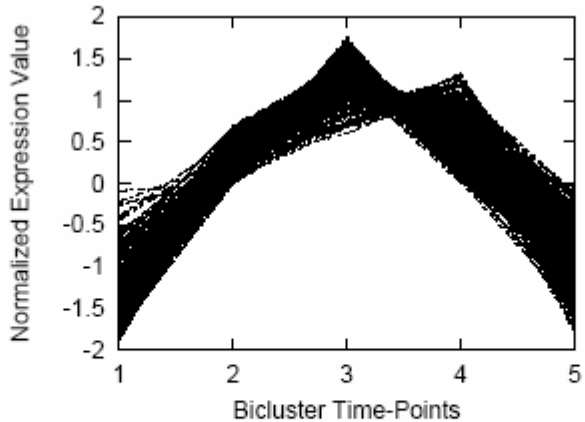
- **GOAL:** Biologists are able to identify Regulatory Modules
  - CCC-Biclusters describing transcriptional Up-Regulation patterns
  - CCC-Biclusters describing transcriptional Down-Regulation patterns
- **HOW:** Analyze the Biological relevance of CCC-Biclusters
  - Gene Ontology Annotations (**GOToolBox:** *http://crfb.univ-mrs.fr/GOToolBox/*)
  - Trancriptional Regulations (**YEASTRACT Database**:*www.yeastract.com*)

Co-regulation
(genes regulated by the same TFs)
+
Co-expression
(genes with coherent expression patterns)

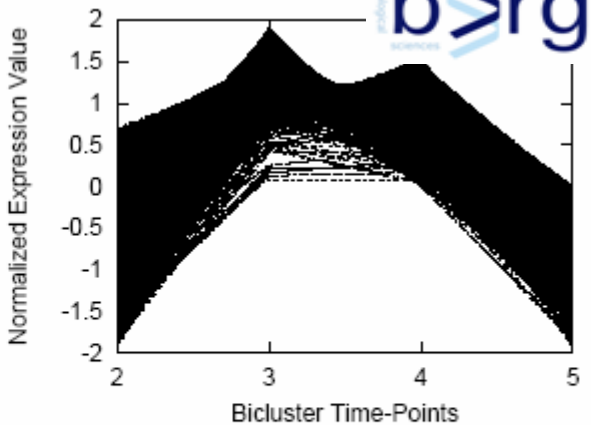
Transcriptional Up-Regulation and Down Regulation Patterns



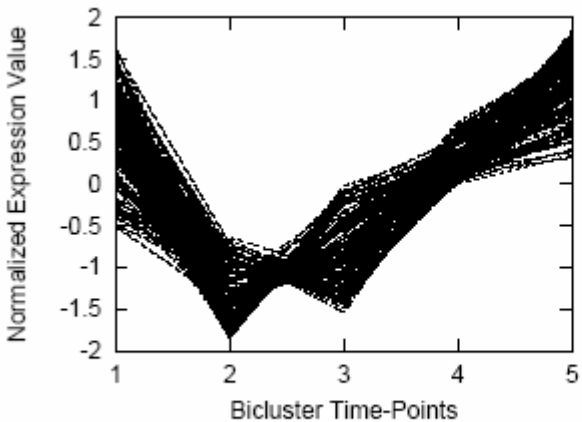
(a) CCC-Bicluster 39



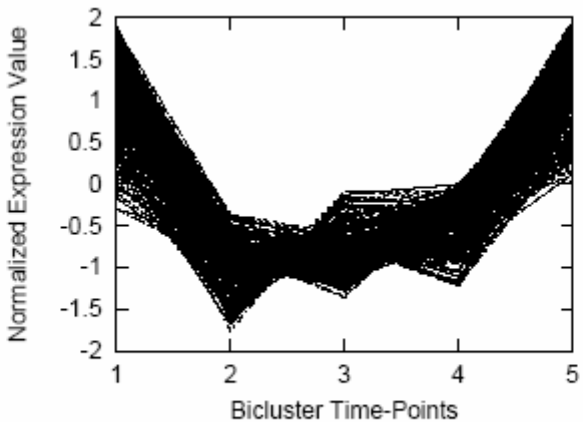
(b) CCC-Bicluster 27



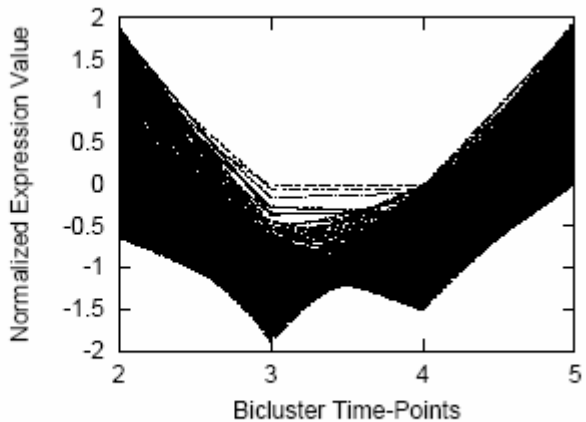
(c) CCC-Bicluster 14



(a) CCC-Bicluster 147

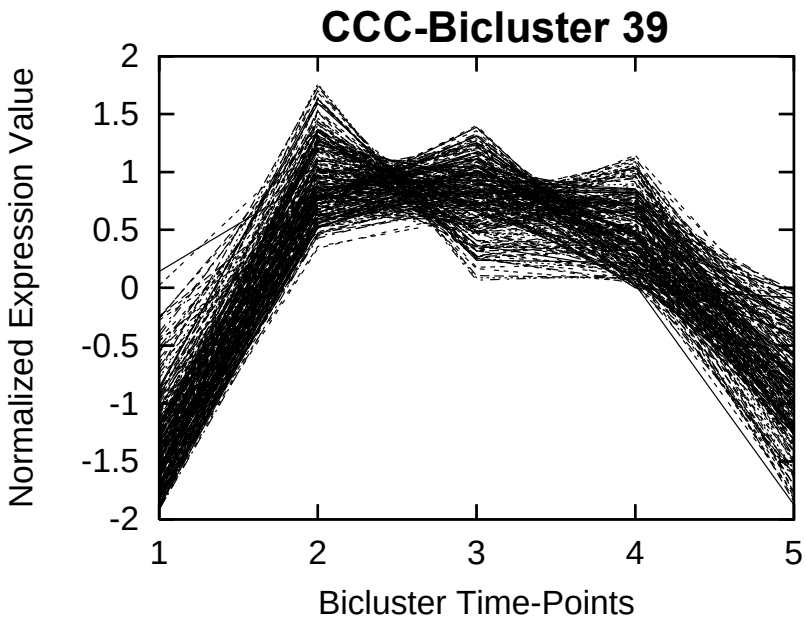


(b) CCC-Bicluster 151



(c) CCC-Bicluster 124

Early Drastic UP-Regulation

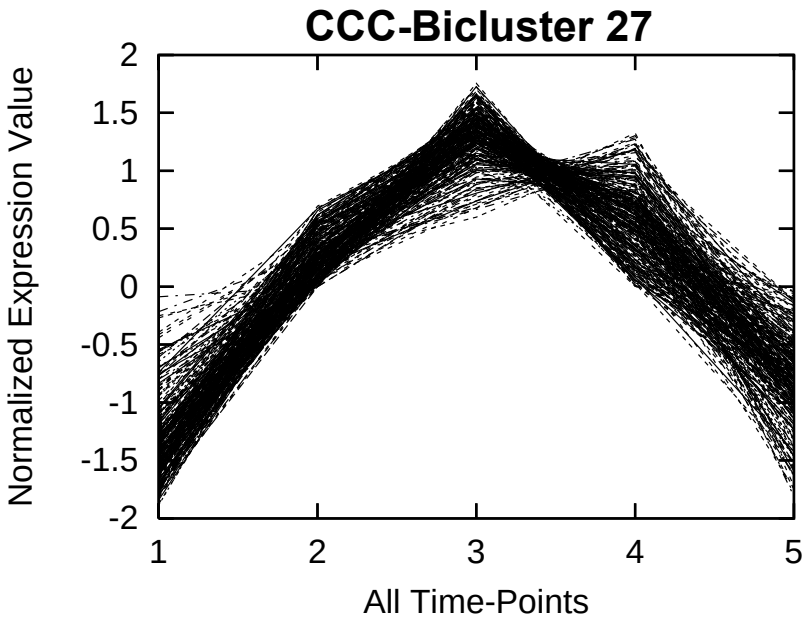


Transcription Factors	%
Sok2p	23,89%
Arr1p	16,37%
Hsf1p	15,93%
Msn2p	14,16%
Rpn4p	14,16%

GO term	Frequency	*p*-value
signal transduction	8,18%	1,65E-04
regulation of transcription from RNA polymerase II promoter	8,18%	2,80E-02

Not significant after Bonferroni correction: p-value > 0.05

Early Slow UP-Regulation



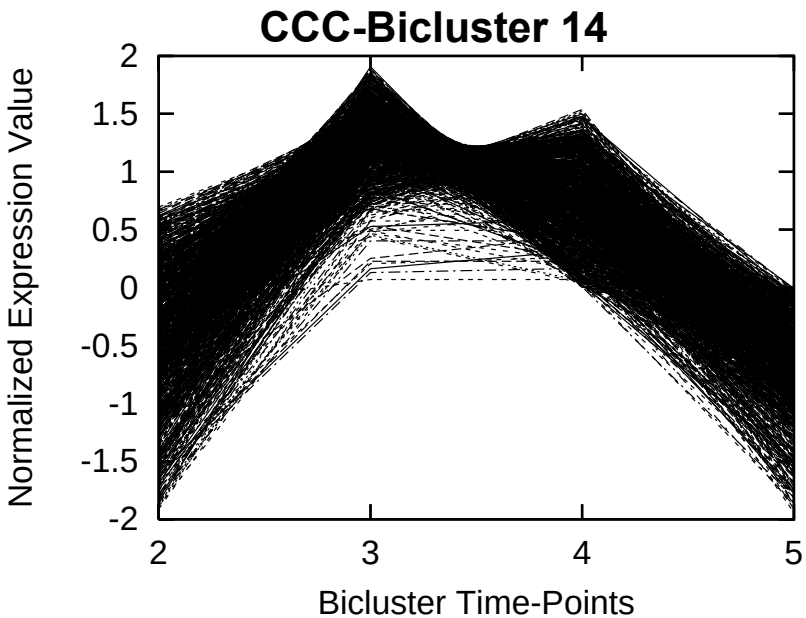
Transcription Factors	%
Hsf1p	23,62%
Sok2p	22,14%
Msn2p	20,66%
Rpn4p	18,45%
Msn4p	17,71%

Early Slow UP-Regulation

GO term	Frequency	*p*-value
response to stimulus	21,59%	3,28E-08
carbohydrate metabolism	1,25%	7,33E-08
generation of precursor metabolites and energy	11,36%	1,86E-06
energy derivation by oxidation of organic compounds	10,23%	3,60E-06
response to stress	15,34%	4,88E-06
carbohydrate biosynthesis	5,68%	5,63E-06
response to stimulus	21,59%	3,28E-08

Highly Significant after Bonferroni correction: p-value < 0.01

Middle UP-Regulation



Transcription Factors	%
Hsf1p	23,62%
Sok2p	22,14%
Msn2p	20,66%
Rpn4p	18,45%
Msn4p	17,71%

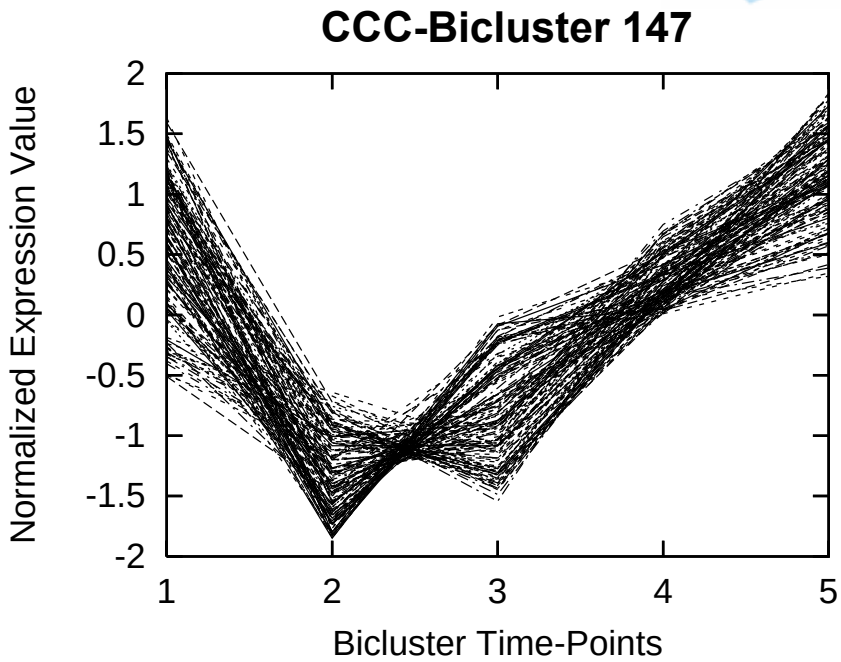
Middle UP

GO term	Frequency	*p*-value
generation of precursor metabolites and energy	10,76 %	2,79E-24
carbohydrate metabolism	10,01 %	4,87E-21
energy derivation by oxidation of organic compounds	9,14 %	3,92E-20
cellular carbohydrate metabolism	8,64 %	1,24E-16
response to stimulus	16,77 %	1,51E-16
response to stress	13,14 %	1,02E-15

Highly Significant after Bonferroni correction: p-value < 0.01

Early Drastic DOWN-Regulation, followed by Rapid UP-Regulation

Transcription Factors	%
Ste12p	16,67%
Rap1n	15,83%
Swi4p	15,00%
Rpn4p	13,33%
Ino4p	11,67%

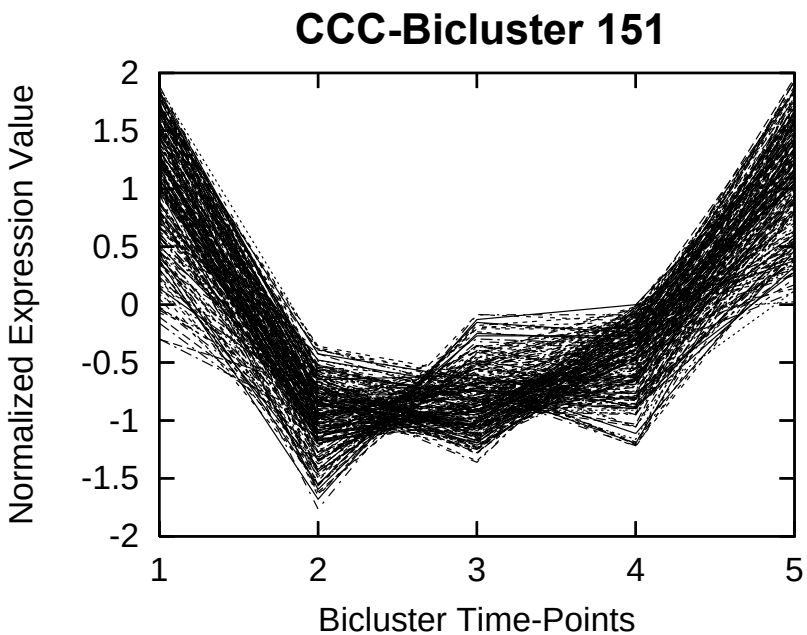


Early Drastic DOWN-Regulation, followed by Rapid UP-Regulation

GO term	Frequency	*p*-value
regulation of progression through mitotic cell cycle	2,56%	5.60E-05
cellular physiological process	86,32%	9.34E-05
biopolymer glycosylation	5,17%	1.03E-03
protein amino acid glycosylation	5,17%	1.03E-03
steroid metabolism	4,31%	1.05E-03
protein targeting to ER	3,45%	1.33E-03
glycoprotein biosynthesis	5,17%	1.48E-03

Significant after Bonferroni correction: 0.01 <=p-value < 0.05

Early drastic DOWN-Regulation, followed by late UP-Regulation



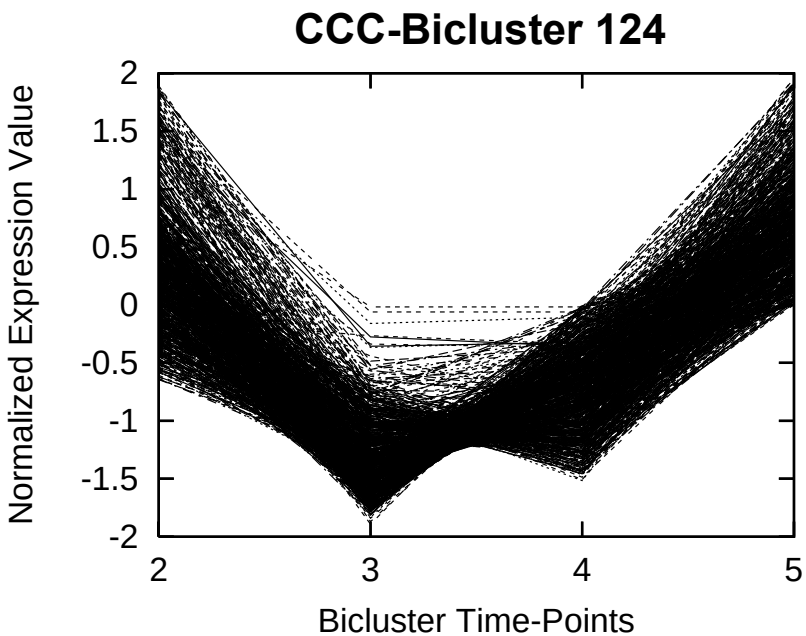
Transcription Factors	%
Mbp1p	12,37%
Arr1p	11,34%
Rpn4p	9,79%
Ino4p	8,76%

Early drastic DOWN-Regulation, followed by late UP-Regulation

GO term	Frequency	*p*-value
cell organization and biogenesis	39,38 %	6,95E-08
cell cycle	16,58 %	1,30E-07
nucleobase, nucleoside, nucleotide and nucleic acid metabolism	38,34 %	7,04E-07
biopolymer metabolism	3,99 %	8,41E-07
mitotic cell cycle	10,88 %	3,32E-06
regulation of biological process	19,17 %	7,17E-06
regulation of physiological process	18,65 %	1,08E-05

Highly Significant after Bonferroni correction: p-value < 0.01

Delayed DOWN-Regulation



Transcription Factors	%
Sfp1p	33,00%
Rap1p	20,89%
Rpn4p	18,91%
Arr1p	16,19%
Fhl1p	12,36%

Delayed DOWN-Regulation

GO term	Frequency	*p*-value
ribosome biogenesis	18,52%	8,23E-63
ribosome biogenesis and assembly	19,97 %	8,38E-62
cytoplasm organization and biogenesis	19,97 %	8,38E-62
rRNA processing	14,81 %	1,42E-48
RNA metabolism	2,48 %	3,36E-40
RNA processing	18,84 %	6,65E-38
rRNA metabolism	15,46 %	1,58E-35
organelle organization and biogenesis	33,01 %	6,49E-30
nucleobase, nucleoside, nucleotide and nucleic acid metabolism	42,51 %	1,20E-29
cell organization and biogenesis	39,61 %	1,79E-23

Highly Significant after Bonferroni correction: p-value < 0.01

Future Work

- ***e*-CCC-Biclustering (extensions)**
  - Restricted / Weigted errors
  - Coherent Sign Changes (Anti-Correlation)
  - Time-Lagged Expression Patterns (Activation/Inibition Relationships)
  - Coherent Expression Patterns with Gaps
  - Multiple Related Gene Expression Time-Series
- **Genomic Data Fusion:** gene expression , ChIP-chip and motif data.

References

- Sara C. Madeira and Arlindo L. Oliveira. ***Biclustering Algorithms for Biological Data Analysis: A Survey.*** *IEEE/ACM Transactions on Computational Biology and Bioinformatics,VOL 1, NO. 1, pp. 24-45, January-March 2004.*

- Sara C. Madeira and Arlindo L. Oliveira. ***A Linear Time Biclustering Algorithm for Time Series Expression Data.*** *Workshop on Algorithms in Bioinformatics, WABI'05 , pp 39-52, Spain, October, 2005. (Springer, LNCS/LNBI 3692)*

- *Sara C. Madeira and Arlindo L. Oliveira,* ***An Evaluation of Discretization Methods for Non-Supervised Analysis of Time-Series Gene Expression Data****, Technical Report 42/2005, INESC-ID, December 2005.*

- *Sara C. Madeira and Arlindo L. Oliveira,* ***"Discovering Modules in Time-Series Gene Expression Data using Biclustering (Abstract) "****, IFCS 2006 Conference: Data Science and Classification (IFCS'06),* ***Invited Session*** *on "Clustering and Classification of Microarray Gene Expression Data", Slovenia, July, 2006.*

- *Sara C. Madeira,* ***An Overview on Mixtures and Hidden Markov Models of Gene Expression in Time-Series****, Technical Report , IST, Learning Theory, December, 2006.*

- Sara C. Madeira and Arlindo L. Oliveira. ***An Efficient Biclustering Algorithm for Finding Genes with Similar Patterns in Time-Series Expression Data****. 5th Asia Pacific Bioinformatics Conference (APBC'07), Imperial College Press, Hong-Kong, January, 2007.*

- Marie-France Sagot. ***Spelling approximate repeated or common motifs using a suffix tree****. Latin'98*, pp 111–127, 1998.

- *YEASTRACT database: http://www.yeastract.com*

- Esko Ukkonen. ***On-line construction of suffix trees.*** Algorithmica, VOL 14, pp. 249-260, 1995.

- René Peeters. ***The maximum edge biclique problem is NP-Complete.*** Journal of Discrete Applied Mathematics, VOL 131, NO. 3, pp. 651-654, 2003.