A Data-Mining Approach To Time-Series Microarray Alignment for Crossing Large-Scale Biomolecular and Literature Information

3rd Workshop on Algorithms in bioinformatics

October 7-9, 2008, Laboratoire J.-V. Poncelet, Moscow

Nicolas Turenne

INRA – Jouy-en-Josas centre

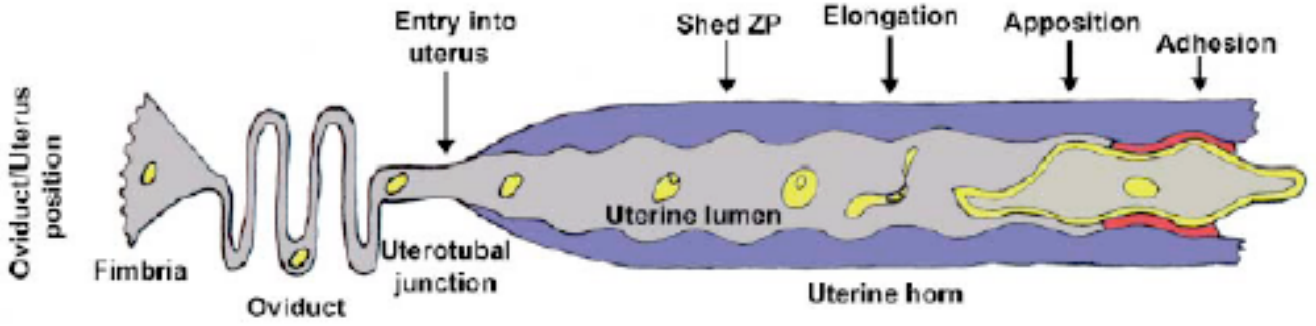
Issue

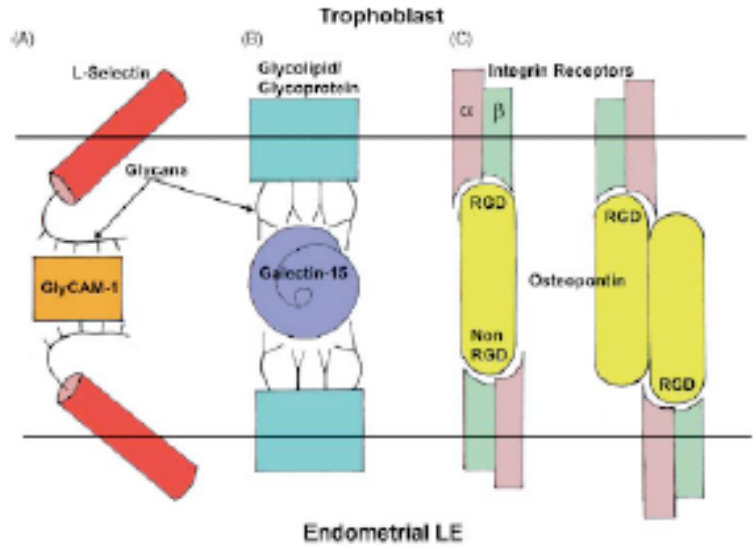
- Part 1 **Project**
- Part 2 **Microarray alignment**

The Cattle Model

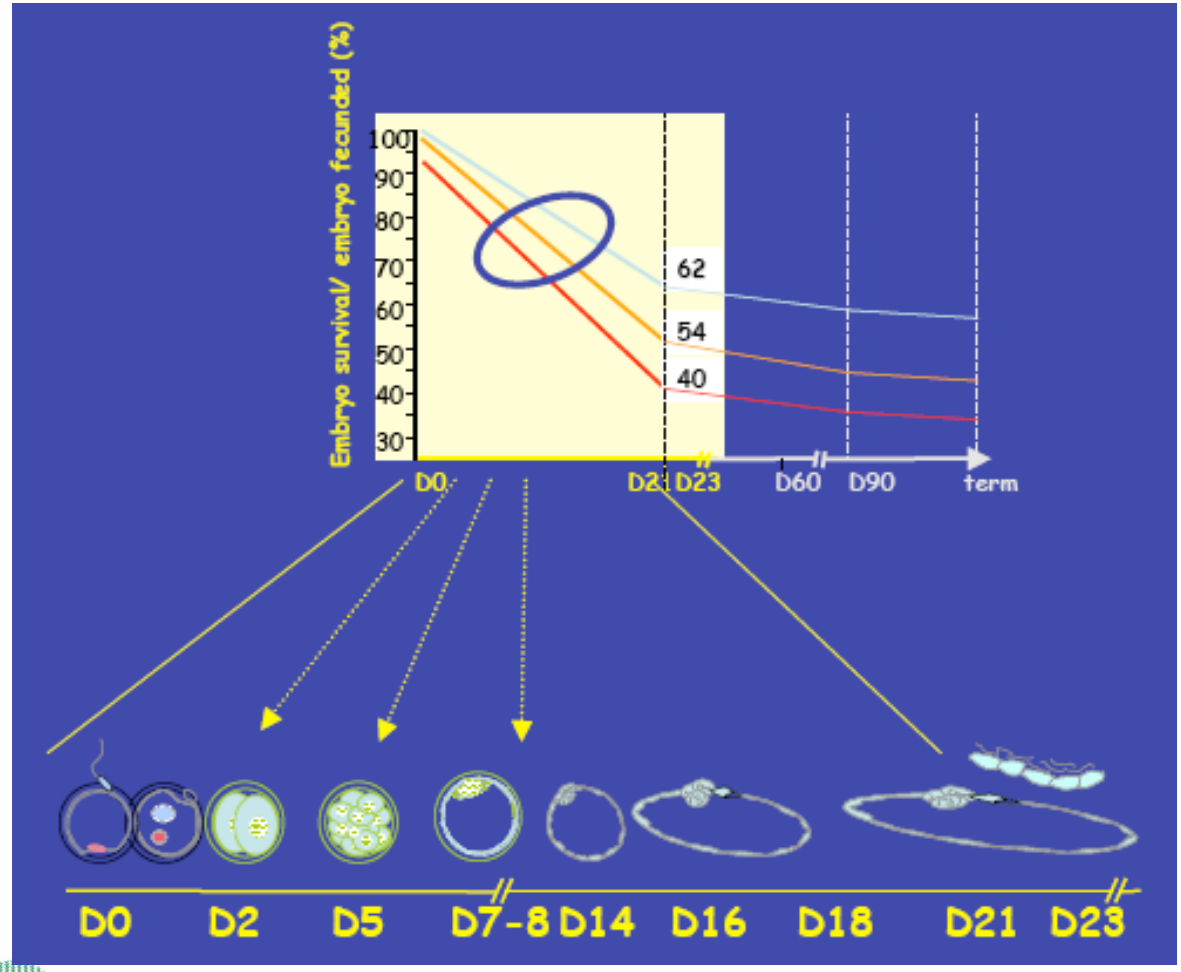
- INRA => french institute of life sciences and food sciences
  - 4000 research scientists, 20 centres, 400 laboratories
- Cattle => Bovine model of interest
  - Perspective for pharmacopea
  - Species to experiment understand life phenomenon as cancer, celullar engineering
- Few data about this species
  - Not enough in Litterature
  - Home microarray about proliferation , on-going published

The Cattle Model : elongation





The Cattle Model : day0-day23



No elongation in human and mouse

No elongation without proliferation
Process known in human and mouse

And without
Embryo development
Process known in mouse

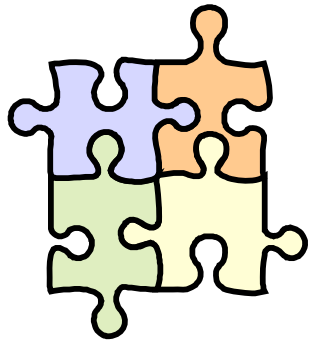
Process not very well known because embryo at this stages develops freely in uterus (no placenta)

Heterogeneous Sources Approach

- Issue : understand which genes of Cattle are related to proliferation and development at embryo stage

- Hypothesis : Inference of knowledge from Standard Model
  species : human, mouse

1- Public-Domain microarrays exist in GEO server about Human and Mouse
  - our goal : data-oriented (time-series) developmental biology



2- Database
  - Genome of Cattle is known 30000 genes, GeneBank Id can be accessible
  - Knowledge Exploration Software, available: Metacore, Ingenuity, David

3- Available Prolific Literature about Human and Mouse (>12 millions documents)

What does we find in Literature ?

- Rough query on Medline server (http://www.ncbi.nlm.nih.gov/pubmed/)
  - bovine and (embryo or placenta) -> 14000 documents
  - human and (embryo or placenta) -> 185000 documents
  - mouse and (embryo or placenta) -> 57000 documents

More concretly in Literature, two corpus

- 77333 documents 06 Aug 2007

#req1 OR #req2 OR #req3 OR #req4

#req4	human AND embryo Field: Title/Abstract, Limits: Humans
#req3	human AND embryo Field: MeSH Terms , Limits: Humans
#req2	human AND placenta AND cancer Field: Title/Abstract, Limits: Humans
#req1	human AND placenta AND cancer Field: MeSH Terms , Limits: Humans

- 34529 documents 06 Aug 2007

#req1 OR #req2

#req1	mouse AND embryo Field: Mesh Terms, Limits: Animals
#req2	mouse AND embryo Field: Title/Abstract, Limits: Animals

Named Entities Extraction Tools

- Since 1998 more than 50 tools of named entities tools has been developped
  - Gene name extraction
  - Network reconstruction
- LingPipe [Carpenter, 2004]
  – sentence segmentation

CorpusH -> 515500 sentences
CorpusM -> 276100 sentences

PMID - 15556029
DP - 2004 Dec
TI - Sporulation of Bacillus subtilis.
AB - Differentiation of vegetative Bacillus subtilis into heat resistant spores is initiated by the activation of the key transcription regulator Spo0A through the phosphorelay. Subsequent events depend on the cell compartment-specific action of a series of RNA polymerase sigma factors. Analysis of genes in the Spo0A regulon has helped delineate the mechanisms of axial chromatin formation and asymmetric division. There have been considerable advances in our understanding of critical controls that act to regulate the phosphorelay and to activate the sigma factors.
AD - Department of Microbiology and Immunology, Temple University School of Medicine. 3400N. Broad St., Philadelphia, Pennsylvania 19140, USA.
FAU - Piggot, Patrick J
AU - Piggot PJ
FAU - Hilbert, David W
AU - Hilbert DW
SO - Curr Opin Microbiol 2004 Dec;7(6):579-86.

Sporulation of Bacillus subtilis.
Differentiation of vegetative Bacillus subtilis into heat resistant spores is initiated by the activation of the key transcription regulator Spo0A through the phosphorelay.
Subsequent events depend on the cell compartment-specific action of a series of RNA polymerase sigma factors.
Analysis of genes in the Spo0A regulon has helped delineate the mechanisms of axial chromatin formation and asymmetric division.
There have been considerable advances in our understanding of critical controls that act to regulate the phosphorelay and to activate the sigma factors.

Genes names extraction

abner

[Settles, 2005]
Training annotated corpus
Conditional random fields Models
Uses regular expression formalism
No explicit syntactic and semantic rules

60611 nouns phrases (CorpusM)
82903 nouns phrases (CorpusH)

genia

[Tsuruoka *et al*, 2005]
Training annotated corpus
Part-of-speech tagging with cyclic dependency network
Maximum Entropy Classifier
No explicit syntactic and semantic rules

37607 nouns phrases (CorpusM)
48909 nouns phrases (CorpusH)

lingpipe

[Carpenter, 2004]
Training annotated corpus
Bayesian Generative Model and Maximum Likelihood
Viterbi decoder
No explicit syntactic and semantic rules

80308 nouns phrases (CorpusM)
93673 nouns phrases (CorpusH)

nlprot

[Mika *et al*, 2004]
Training corpus
Syntactic-Rules and Support Vector Machine classifiers
Use of biology name dictionaries
No explicit semantic rules.

42427 nouns phrases (CorpusM)
48086 nouns phrases (CorpusH)

Expert Extraction Software : *Metacore, Ingenuity, David*

Ingenuity

http://www.ingenuity.com/ Ingenuity Systems, Inc. (California, USA)
IPA - ingenuity pathway analysis software (liccnce = 6000€/year; 25000 users)

- 1.7 millions « biological findings »
- Own ontology (knowledge base)

- Since 1997
- Knowledge base (ontology) build upon criteria :
  - 300 reviews (full papers)
  - manual extraction (1000 documentalists)
  - 5 years
  - update each 3-month , 80000 new findings
  - optimized rules for manual scan (less people required)

- Link with Gene Ontology (GO)
- Available Synonyms and homonyms names (« ingenuity facets »)
- Grabbed information from NCBI, Swissprott and Kegg
- 12 branches in the global ontology (only 3 in GO)

Crossing Information Sources

Ingenuity / Information Extraction Tools

Database ⇔ Literature

Why ?

- expert extraction interpretation-dependent
- multipe-interpretation in documents
- merging results from automatic extraction and expert extraction can be more riched if hypothese-oriented

Crossing Information Sources

Ingenuity / Information Extraction Tools

Database ⇔ Literature

Gene Lists extracted from Ingenuity about *development*

	Tissue + development (A)	Connective + tissue (B)	Cellular + development (C)	A ∩ B ∩ C	proliferation + development (D)
From Ingenuity	615	532	482	52	
From GO					204

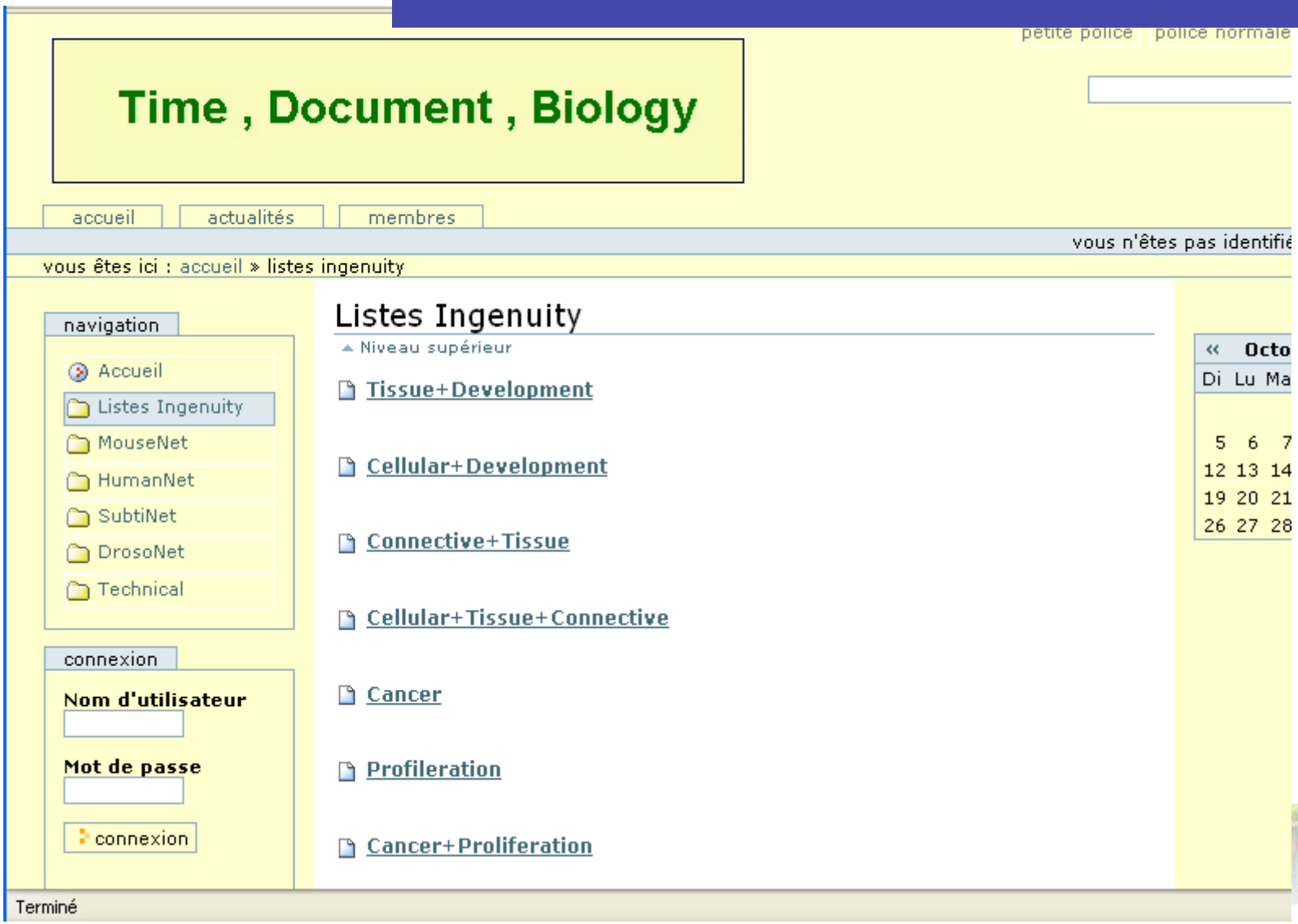
∩ CorpusM	A	B	C	A ∩ B ∩ C	D
abner + genia + lingpipe + nlprot	342	293	293	38	90

∩ CorpusH	A	B	C	A ∩ B ∩ C	D
abner + genia + lingpipe + nlprot	333	289	268	40	79

Project
Literature Issue
Database Issue
Microarray Issue
Microarray Alignment

Crossing Information Sources

http://migale.jouy.inra.fr/time/



Mathématique Informatique & Génome

What about knowledge from microarrays

- Knowledge are related to large sets of genes at a same time
  - High-throuhgput data management and analysis
- We can identify groups
  - acting in a same way ,
  - or associations between a gene and others in a same context (biological hypothesis)

Data

ID_REF	NAME	GSM23324	GSM23325	GSM26511	GSM23326	GSM23327	GSM23328	GSM23330	C
3069	3069	-0.12095261	-0.159064695	-0.112117298	-0.442279081	0.044055627	-0.138586163	-0.030866648	1
2173	2173	-0.134408201	-0.160850872	-0.043401834	-0.381694889	-0.124970576	-0.249941744	0.046745013	1
1105	1105	-1.550597412	-0.675447603	-0.146603474	-2.525728644	-0.566395475	-1.945910149	-0.211309094	1
4449	4449	-0.064720191	0.066624028	-0.152385454	-0.234877715	-0.041641026	-0.162003333	0.064983488	1
1520	1520	-0.063476064	0.041528459	0.030614636	-0.186829974	-0.155733209	-0.066511481	-0.038183787	1
560	560	-0.379489622	-0.341170757	-0.538660423	-3.496507561	-0.149345289	-0.972986076	-0.035755649	1
1706	1706	-0.027779564	-0.024667232	-0.110130824	-0.304353607	-0.037582711	-0.234010656	-0.12351371	1
3334	3334	-0.236664298	-0.030277259	0.086709399	-0.394753453	-0.115896291	-0.139846692	0.056384719	1

Measure

Log (base 2)
of the ratio of the mean of Channel 2 (635 nm)
to Channel 1 (532 nm)

Value : between -10 (very inhibited) and +10 (very activated)

Datasets of interest

- GSE 1414 only kinetics about **bovine** and dealing with same biological problem : elongation and implantation in bovine embryo (2,000 unique genes)
  (Ushizawa et al, Reprod Biol Endocrinol, 2004)
  on-going INRA-home made microarray

- GSE 9046 time-course experiment with embryoid bodies of CGR8 **mouse** embryonic stem cells (12,000 unique genes)
  (Mitiku and Baker, Dev Cell. 2007)
  *INRA-home made microarray about a kinetics of development in **mouse**, based totipotent embryo stem cell (degrelle et al, dev biol, 2005)*

- GSE 3553 interesting for **human** cell differentiation in trophoblast in human under effect of BMP4 (25,000 unique genes)
  (Xu et al, Nat Biotechnol. 2002)

What about knowledge from microarrays

Issue

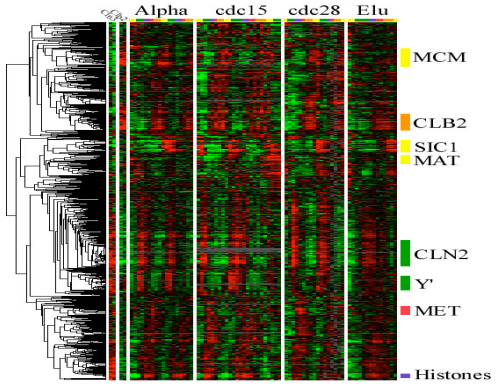
- Time-series microarrays with several time-points (3 to 10)
- Two different species (for instance bovine / human or bovine / mouse)

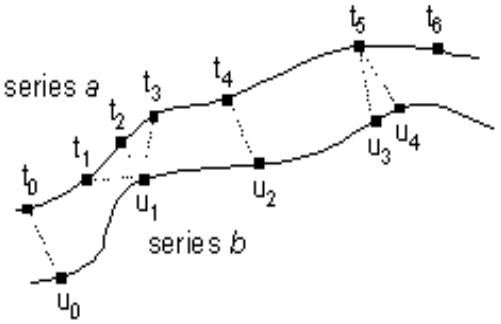
Challenge

- *state of the art : clustering is largely used but only work for same conditions , in our case , microarrays are different-conditions made*
- *state of the art : time warping is used for time-comparison scales (curve alignment) but in our case time scales are different from one species to another and a same ortholog gene can occur at different time-point because of genome evolution over time.*

[Husmeier, 2001]

[Aach, 2001]

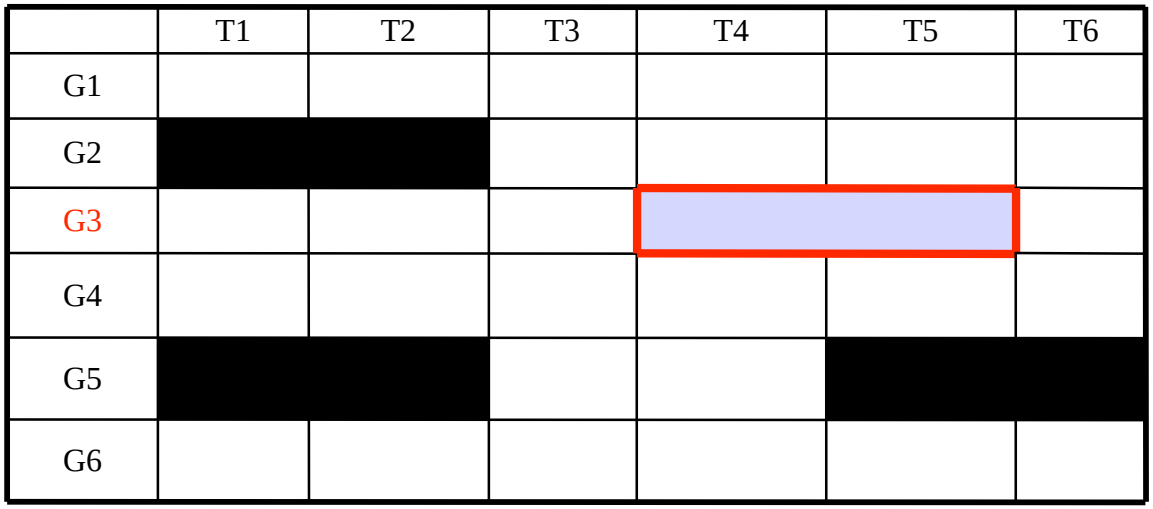




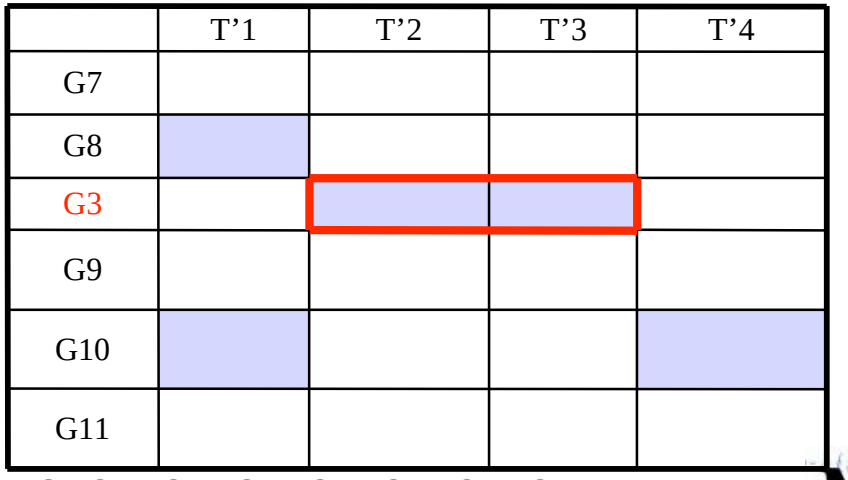
What about knowledge from microarrays

Goal

- Patterns Identification
- Data format is matrix-like
- 2 tables



	T1	T2	T3	T4	T5	T6
G1						
G2						
G3						
G4						
G5						
G6						



	T'1	T'2	T'3	T'4
G7				
G8				
G3				
G9				
G10				
G11				

A combinatorics issue

T1	T2	T3	T4	T5	T6
G2	G2				
			G3	G3	
G5	G5			G5	G5

T'1	T'2	T'3	T'4
G8			
	G3	G3	
G10			G10

The issue of Alignment

- How to place G8 before G2 or during G2 ?
- We can not fit T1 and T'1, T2 and T'2 …
- Even infer that T4 = T'2 is not jusiified by the fact it is the same gene G3

A combinatorics issue

Dobinski formula

$$B_n = \frac{1}{e}\sum_{k=0}^{\infty}\frac{k^n}{k!}$$

Number of partitions of size n

T1	T2	T3	T4	T5	T6
G2	G2				
			G3	G3	
G5	G5			G5	G5

T'1	T'2	T'3	T'4
G8			
	G3	G3	
G10			G10

Very small set of constraints about strict order (<), such as
G2 before G3
G3 before and after G10
G8 before G3
….etc

G2 G5	G3	G3 G5	G8 G10	G3	G10	
G2 G5 G8 G10	G3	G3 G5	G5 G10			
G2	G2G5	G8	G10	G3		G10
G5 G2	G5 G2	G8 G10	G3	G10	G3	G3 G5
G2 G5 G8 G10	G3	G5 G3 G10				
And many many many others …						

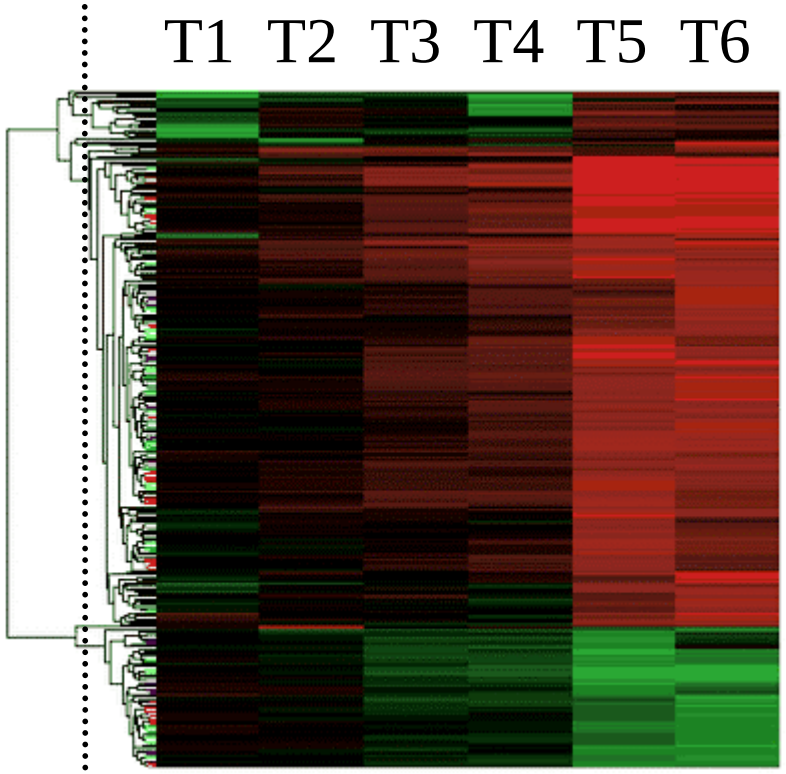
A solution in a two-step clustering

- Step 1: make clusters of similar genes into a unique time-series
  - relative expression profile

- Step 2 : make a clustering between 2-sets of clusters through common points
  - consensus clustering over two sets of clusters

Step 1

- make clusters of similar genes expression profile
- using a classical euclidian-distance metrics and dendrogram computation
- See TreeView (1998) http://rana.lbl.gov/EisenSoftware.htm



Cut-off

Step 2

- make consensus clustering between two sets of clusters
- Works if some objects belongs to both sets of clusters
- Result is a set of MegaClusters overlapping microarrays (idea of alignment)

Dictionary of Genes [G1-G6] from microarray Bio1, [G1;G7-G12] from microarray Bio2

(G1 , G2 , G3 , G4 , G5 , G6 , G7 , G8 , G9 , G10, G11, G12)

partition Bio1 (**C1 , C1 , C1 , C2 , C2 , C2** , C3 , C4 , C5 , C6 , C7 , C8)

partition Bio2 (C16, C10, C11, C12, C13, C14, **C15, C15, C15**, C16, C16, C16)

result (**C1 , C1 , C1** , C2 , C2 , C2 , C3 , C3 , C3 , **C1 , C1 , C1**)

Because G1 belongs to C1 and C16, C1 and C16 are merged

Consensus clustering approach

Definition

Merging of several clustering into a unique clustering

Three kinds of clusterings:

- axiomatic (we suppose we can formalize property of the resulting partition
- constructive (some rules are given to achieve the merging)
- optimization (a criteria to minimize is defined)

Consensus clustering approach

membership matrix (clusters/partitions)

	P_1	P_2
C_1	1	0
C_2	0	1

we deduce from the matrix

$$\begin{cases} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} \text{simultaneous clusters} \\ C_1 C_2 \text{ cluster 1 precedes} \\ C_2 C_1 \text{ cluster 2 precedes} \end{cases}$$

membership matrix (megaclusters/partitic

	P_1	P_2
M_1	1	0
M_2	1	1

we deduce from the matrix

$$\begin{cases} \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} (t, P_1) \, if \, S \leqslant q_1(t) \wedge S \leqslant q_2(t) \\ M_1 M_2 (t, P_1) \, if \, S \leqslant q_1(t) \leqslant q_2(t) \\ M_2 M_1 (t, P_1) \, if \, q_1(t) \geqslant q_2(t) \geqslant S \end{cases}$$

optimization approach for consensus

$\{C_1, \ldots, C_B\}$ → $\sum_{b=1}^{B} w_b d(C, C_b)^p \Rightarrow \min_{C \in \mathcal{C}}$

d is a dissimilarity measure

Let $M_1, \ldots, M_B$ and M denote the membership matrices of the elements of the ensemble and their sought least squares consensus partition.

$\sum_b w_b \min_{\Pi_b} \|M - M_b \Pi_b\|^2 \Rightarrow \min_M$ permutation matrices $\Pi_1, \ldots, \Pi_B$ Euclidean partition dissimilarity

fix the Π_b let $\bar{M} = s^{-1} \sum_b w_b M_b \Pi_b$ where $s = \sum_b w_b$

$$\sum_b w_b \|M - M_b \Pi_b\|^2 = s(\|M - \bar{M}\|^2) + \sum_b w_b \|M_b\|^2 - s\|\bar{M}\|^2$$

$$\text{maximizing } s^2 \|\bar{M}\|^2 = \sum_{\beta, b} w_\beta w_b \operatorname{tr}(\Pi'_\beta M'_\beta M_b \Pi_b)$$

Consensus clustering approach

- CLUE library
  - R-project
  - function *cl_consensus(method="DWH")*
  - Fuzzy clustering

• E. Dimitriadou, A. Weingessel and K. Hornik (2002). A combination scheme for fuzzy clustering. *International Journal of Pattern Recognition and Artificial Intelligence*, **16**, 901–912

Consensus clustering approach

- CLUE library
  - heuristic-based
  - locally single-pass through the ensemble of clusterings
  - starting with

$$\tilde{M}_1 = M_1$$

$\tilde{M}_b$ is obtained from $\tilde{M}_{b-1}$ by optimally matching $M_b\Pi_b$ to this taking a weighted average of $\tilde{M}_{b-1}$ and $M_b\Pi_b$ in a way that $\tilde{M}_b$ is the weighted average of the first b $M_\beta\Pi_\beta$

Result
is a fuzzy membership
but it is possible to get
a hard clustering

C1(1, 1, 2, 2) C2(3,3,3,4)

Memberships:

	[,1]	[,2]
[1,]	0.0	1.0
[2,]	0.0	1.0
[3,]	0.5	0.5
[4,]	1.0	0.0

Hard clustering (1 1 2 2)

Temporal profile

Time Correlation Matrix

• Use notion of precedence and simultaneity, using the symbol **B** for *before*, **A** for *after* and **D** for during

- about expression
- for a given gene
- comparison between time neigbourghood

$$\begin{cases} A_g(t,P) & if\ S \leqslant q_g(t+1,P) \\ B_g(t,P) & if\ S \leqslant q_g(t-1,P) \\ D_g(t,P) & if\ \dfrac{q_g(t,P)}{q_G(t,P)} \geqslant S \end{cases}$$

Temporal profile

Cluster	*Target*	*T1(Bio1)*	*T2(Bio1)*	*T3(bio1)*	*T4(Bio1)*	*T1(Bio2)*	*T2(Bio2)*	*T3(Bio2)*
1	*4*	*AD*	*ABD*	*ABD*	*BD*			
2	*4*					*B*	*A*	*D*

Cluster	*Target*	*T1(Bio1)*	*T2(Bio1)*	*T3(bio1)*	*T4(Bio1)*	*T1(Bio2)*	*T2(Bio2)*	*T3(Bio2)*
p	*4*	*AD*	*ABD*	*ABD*	*BD*	*B*	*A*	*D*

For a given Gene, for instance G4,
We take its MegaCluster (c1, c2) obtained from consensus clustering
For each timepoint and for each cluster, for instance T3 (microarray 1) and cluster 1 we test if expression is high during (D), before (T2)or after (at T4). It is ok for before and during so the value for T3-C1 is BD.

Comparison of temporal profile

- Jaccard index similarity $J(A,B) = |A \cap B| / |A \cup B|.$

- A given a gene G and its Time matrix correlation TMC(G)
- We look for all genes have similar their TMC to G one.

- for each gene in both microarray (dictionary of gene)
  - Compute J(TMC(G), TMC(g))
  - Export all genes if J > 0.99

Algorithm – AlibR (R Script)

1. Read 2 Datasets (D) and input a Given Gene (G)
2. Compute mean expression values for clusters
3. Create Gene Dictionary
4. Create Partition of Gene Dictionary with Clusters for D
5. Apply consensus
6. Create a Mapping MegaCluster <-> clusters (MGC)
7. Generate the Temporal Matrix (TM) for all clusters
8. Compute a submatrix of TM for G (TMG) using MGC
9. For each gene g
    1. compute submatrix (TMg) using MGC and
    2. compute Jaccard value J
10. Export Temporally Similar Gene List with J < 0.99

Complexity

- Tests has been done on 30% of microarrays (~9000 genes)

- Time-computation

20-lines microarray	0.42	s	0.5 Mb
600-lines microarray	18.25	s	100 Mb
2000-lines microarray	60.50	s	900 Mb
15000-lines microarray	18000	s	7000 Mb

- DHW consensus method complexity
  - O(n x k) in memory
  - O(n x k^3) in time
  - Optimisation solver O(n^2) in memory (Hungarian algorithm)

Similar genes…

target genes	similarity threshold	Bovine (B) & Human (H) arrays				Bovine (B) & Murine (M) arrays			
		megacluster (#cluster)	B & H genes	B genes	H genes	megacluster (#cluster)	B & M genes	B genes	M genes
alg5	Tb=0.7 ;T=0.9	16	14	18	0	12	25	43	37
	Tb=0.7 ;T=0.1	11	14	18	0	15	12	20	0
eif2s3	Tb=0.7 ;T=0.9	16	12	10	0	15	208	298	2265
	Tb=0.7 ;T=0.1	10	76	81	574	5	6	16	0

Similar genes… case of ALG5

Microarray bovine/human : similarity threshold0.1/0.7

Microarray Bovine

gene: **bp107457**
gene: **bpl11933**
gene: **bpl10819**
gene: **af069434**
gene: **y16359**
gene: **bp111692**
gene: **bp110718**
gene: **loc536818**
gene: **cfdp2**
gene: **bp110964**
gene: **loc509824**
gene: **bp112639**
gene: **u01924**
gene: **bp109437**
gene: **loc531522**
gene: **sepx1**
gene: **aa112300**
gene: **v00125**

Microarray Bovine & Human

gene: **vsig4**
gene: **cask**
gene: **hdac1**
gene: **mmp14**
gene: **vegfa**
gene: **syt1**
gene: **actr2**
gene: **akap9**
gene: **furin**
gene: **alg5**
gene: **mmp1**
gene: **foxred1**
gene: **npepps**
gene: **sdf4**

Similar genes… case of ALG5

Crossing with IPA (ingenuity)

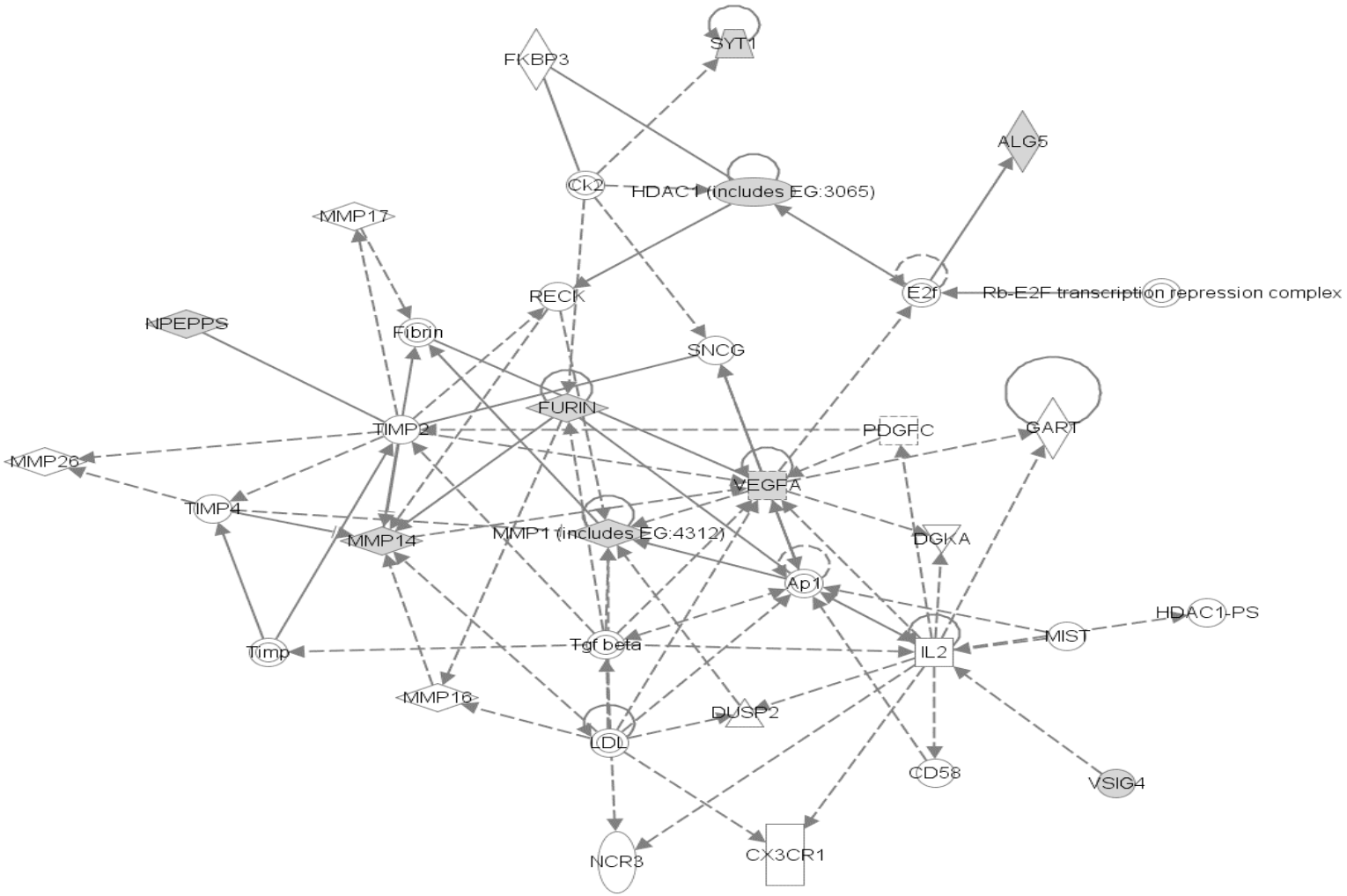
genes	networks	score
Alg5 bov hum	Connective tissue disorders, genetic disorders, cancer	22
Alg5 bov mus	Cancer, cell to cell signalling and interaction, cellular assembly and organisation	14

Project
Literature Issue
Database Issue
Microarray Issue
Microarray Alignment

Similar genes… case of ALG5

Crossing with IPA (ingenuity) Microarray bovine/human

Network 1

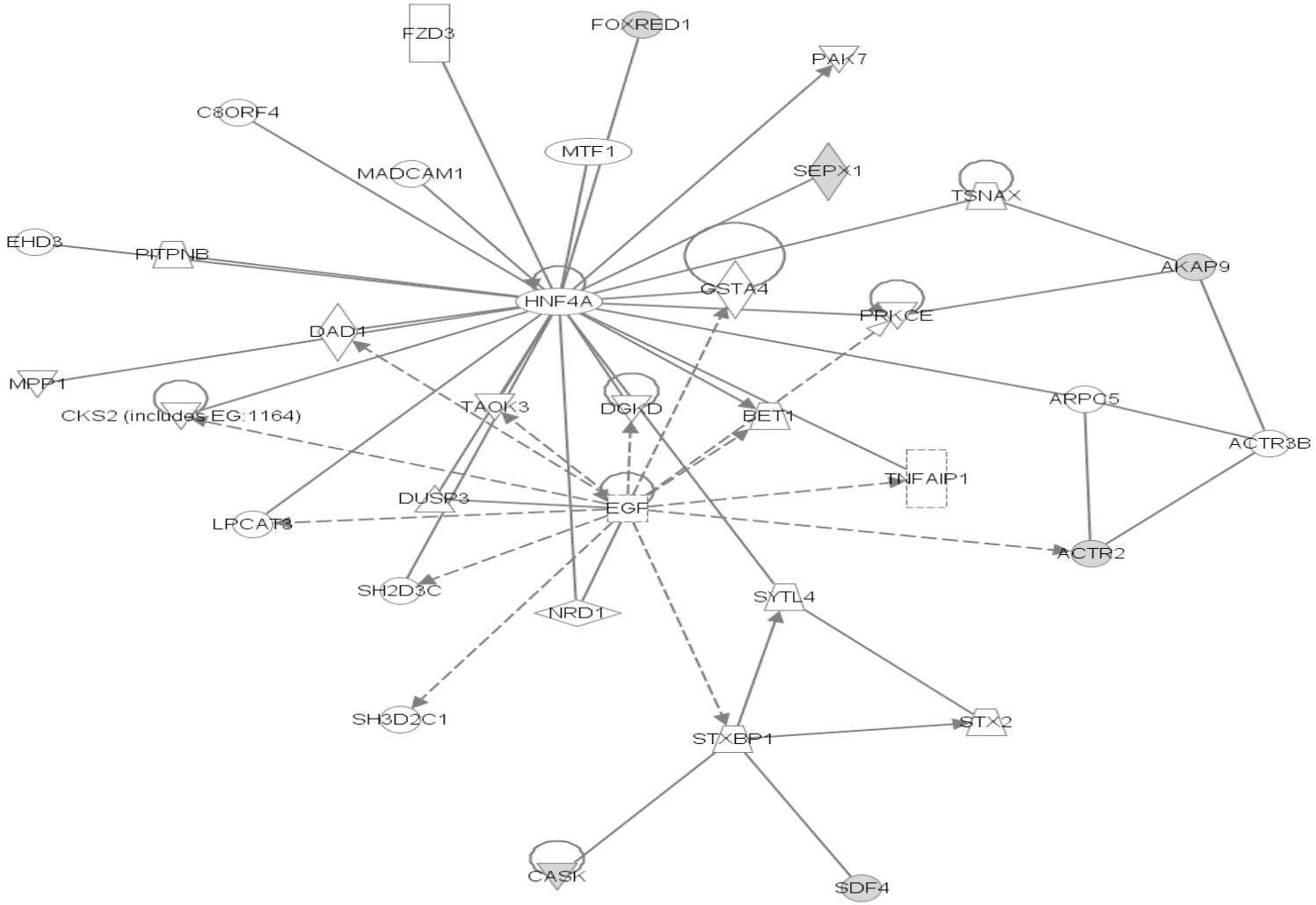


Project
Literature Issue
Database Issue
Microarray Issue
Microarray Alignment

Similar genes… case of ALG5

Crossing with IPA (ingenuity) Microarray bovine/human

Network 2

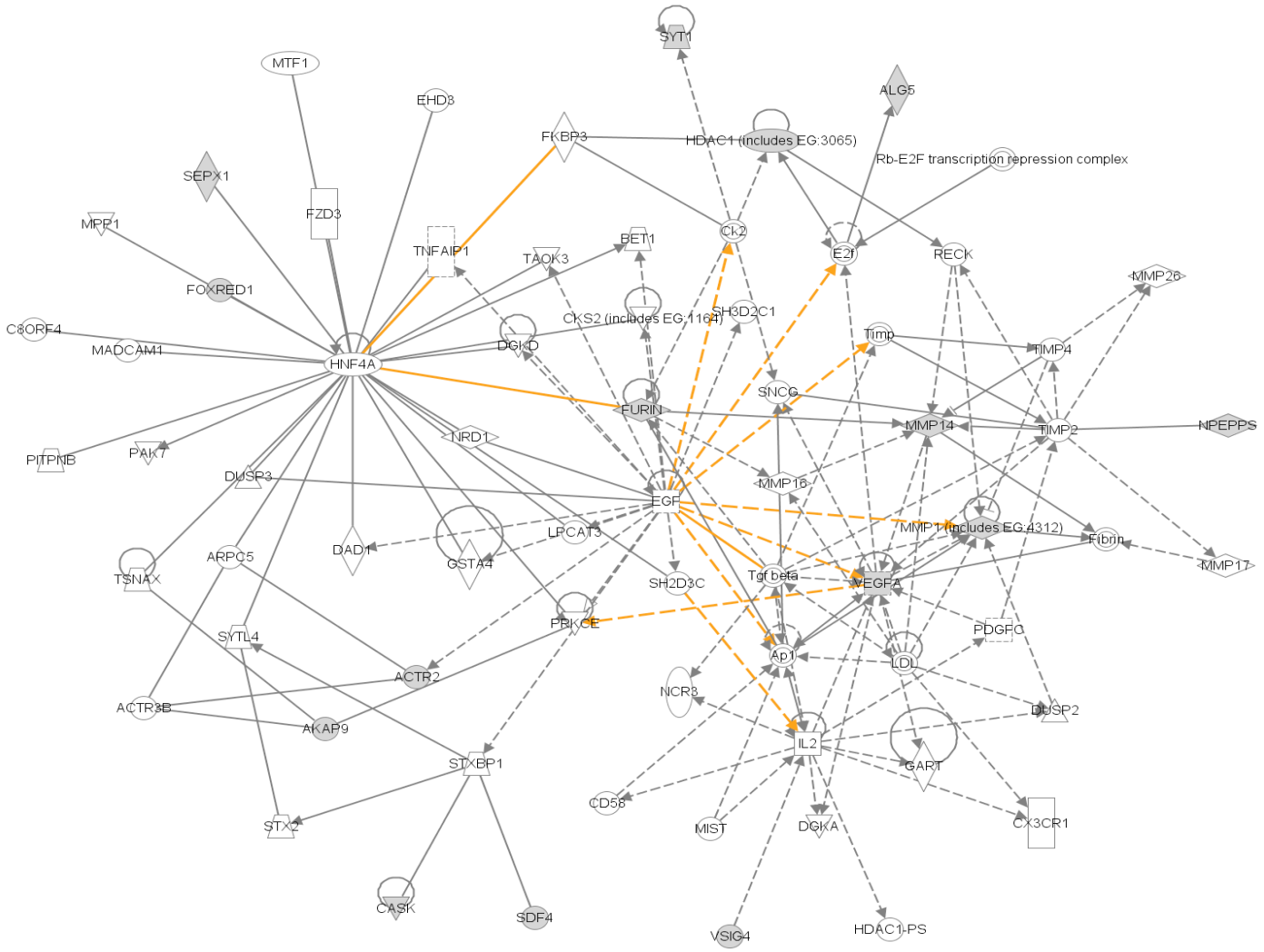


Project
Literature Issue
Database Issue
Microarray Issue
Microarray Alignment

Similar genes… case of ALG5

Crossing with IPA (ingenuity) Microarray bovine/human

Network 1 & 2



INRA
Institut National de la Recherche Agronomique

Genes with similar time matrix correlation

- role of relationships (interaction)
  - not only based on genomic data
  - transcriptomics approach
- role of expression over time
  - not only facts about inhibition / activation
  - comparison of relative expression
  - comparative transcriptomics

conclusion

- Approach with a double-step clustering using time-dependent molecular high-throughput expression data

- Make a temporal profile over two datasets by consensus clustering even if a gene does not belong to one of them

- Fast and easy to understand

- Need to make deeper benchmark with Ingenuity Usage for validation
- Need re-programming for time/memory optimization (R + C-language)

Co-operations…

Dr Isabelle Hue (INRA, BDR Unit)
(Reproductive and Developmental Biology)

INRA has recently signed a cooperation agreement with the **Russian Foundation for Basic Research (RFBR/RFFI)**

call for project proposals on 1st septembre 2008

MERCI
СПАСИБО