A COMBINATORICS-BASED DATA-MINING APPROACH TO TIME-SERIES MICROARRAY ALIGNMENT

N. Turenne [1], I. Hue [2]

[1] INRA, Unité Mathématique Informatique et Génome UR1077, F-78350 Jouy-en-Josas, France, e-mail: turenne@jouy.inra.fr; [2] INRA, UMR 1198 Biologie du Développement et Reproduction, F-78350 Jouy-en-Josas, France, e-mail: isabelle.hue@jouy.inra.fr

One of the biological issues aiming at understanding bovine embryo development implies the analysis of proliferation and differentiation processes. Using published data from model species (mouse, human) we used a double-step classical clustering approach. First step runs a k-mean clustering for each chip individually. Second step runs a fuzzy consensus clustering to merge a few clusters (i.e. megaclusters) between microarrays. Hence we make temporal gene profiles using the symbolic time property of simultaneity and precedence according expression in ensemble of clusters. Finally with the help of a Jaccard coefficient between temporal gene profiles across species, we extract a list of genes revealing a similarity with a target gene of interest. Depending on the species and on the target gene, this list of genes differed in size and content, thus highlighting the interest of such cross-species comparisons to gain insights from different literature contexts.

Key words: microarray alignment, cross-species comparison, clustering, consensus, time series.

Motivation and Aim

Large-scale biological experiments such as microarrays are now available and require data analysis to process huge amounts of results (Eisen *et al.*, 1998). One of the purposes when studying not well-know species is to refer to model-species. In our case we aimed at understanding bovine embryo development (Hue *et al*., 2007) and analysed at first genes related to proliferation processes. Since embryo development and proliferation processes have been well studied in mouse and human species, we based our study on the comparison of published data sets in these three species. However, selecting temporal series for each species, it appeared that each dataset referred to a different time scale (detailed in datasets). Hence, comparing different sets of sequential time-series data could be done through alignment as largely done with DNA sequences since a few decades (Smith *et al.*, 1981; Altschul *et al.*, 1990) and even improved now (Kucherov *et al.*, 2004). For gene expression data the multidimensional property has not been handled so far since the classical way to align microarray datasets deals with curves of univariate time-series (Aach, Church, 2001; Ernst *et al.*, 2005). Nevertheless, this cannot be adapted to our purpose since curves can be deformed if they are compressed in a short-time duration. Lots of combinations are available, among which the Dobinski formula $B_n = \frac{1}{e}\sum_{k=0}^{\infty}\frac{k^n}{k!}$ that gives the number of all partitions of a set of n objects. For $n = 26$, this formula gives $1{,}6 \cdot 10^{21}$ combinations. With a 2Ghz clock, and approximating 1 cycle for a partition, time processing is about 800 years. To solve this, our approach consists in aligning only some parts of the microarrays, thus restricting the space of the alignments. We first used a classical clustering approach on each dataset and merged a few clusters by consensus to study gene interactions around genes of interest while using symbolic time property of simultaneity and precedence as described in Turenne and Schwer, 2008. Our project through microarray alignment was also to cross this kind of relational (temporal) information with known relational information from the literature or database softwares (IPA, PubMed, Gene Ontology for example). Two

corpora (document databases, PubMed) about mouse and human species have been designed and lists of gene names have been extracted from commercial software (IPA) and both corpora.

Datasets

We compared pairwise Bovine microarray data with Human and Mouse microarray data. Bovine embryos on days 7, 14 and 19/20/21, extra-embryonic membranes on days 27/28 and fetuses on days 27/28 were collected to represent early embryo, elongating embryo, pre-implantation embryo, post-implantation extra-embryonic membrane and fetus, respectively (hence it covers 7 time-points). In total, the processed microarray contains 4,607 cDNA covering about 2,000 unique genes (GSE 1414) and call hereafter this chip: BiopB (Ushizawa, 2004). Dataset of murine embryonic developmental time course consists of morphologically staged samples from E6.25 to E9.0 (at approximately 0,25 day intervals, hence it covers 11 time points). In total, the processed microarray contains 43,000 cDNA covering 12,000 unique genes (GSE 9046) hereafter we call this chip: BiopM (Mitiku, Baker, 2007). Human embryonic stem cells were treated in pairs with or without BMP4. This was followed by RNA extraction and amplification and microarray analysis on DNA chips containing 43,000 cDNA clones, which represented about 25,000 unique genes. Samples have been extracted at time 3hrs, 6hrs, 12hrs, 24hrs, 48hrs, 3days, 7days (GSE 3553), hence it covers 7 time points. Hereafter we called this chip: BiopH (Xu *et al.*, 2002).

Method and Algorithm

Our method is currently developed under R tool and we used Clue library (method DWH) for clustering consensus (Hornik, 2005). The first step of the methodology relies on gene clustering for each microarray. This part could be done by *k*-means or descendant agglomerative hierarchy using an Euclidian distance for similarity. We obtained clusters of resembling genes through resembling expression profiles. At the second stage we merged clusters from each microarray in a way explained by Mirkin and Cherny (1970) or Meila (2005). A megacluster is a set of clusters (resulting from consensus clustering) assigned to a specific gene. A consensus distance was used for merging two partitions. The size of a partition is a vector of unique objects to classify (i.e. genes). For instance 12 items are clustered leading to a partition P_1 and a partition P_2. In Partition P_1 the 6 first items belong to the first microarray, the 6 following ones to the second microarray as illustrated in Table 1. Let us suppose that the first item belongs to cluster 1 (in P_1) and cluster 16 (in P_2). Because of this item the clusters 1 and 16 merged in the resulting partition.

Table 1

Example of consensus result

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	Gene8	Gene9	Gene10	Gene11	Gene12
P_1	1	1	1	2	2	2	3	4	5	6	7	8
P_2	16	10	11	12	13	14	15	15	15	16	16	16
Consensus	1	1	1	2	2	2	3	3	3	1	1	1

Notes. Values represent indices of clusters.

From this example we see that gene 3 belongs to cluster 1 but as cluster 1 is merged with cluster 16, a megacluster for gene 3 will be (cluster 1, cluster 16). Hence for a target gene we can identify to which megacluster it belongs and assess a time correlation matrix across both microarrays. Indeed, as we see below, if a cluster C_1 belongs to P_1 and C_2 belongs to P_2, without intersection, three possibilities are possible and all of them are:

$$\begin{array}{ccc} & P_1 & P_2 \\ C_1 & 1 & 0 \\ C_2 & 0 & 1 \end{array} \quad \text{we deduce from the matrix} \quad \begin{cases} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} \\ C_1 C_2 \ \text{ cluster 1 precedes} \\ C_2 C_1 \ \text{ cluster 2 precedes} \end{cases} \qquad (1)$$

In our alignment method, through megaclusters, as shown below with M_1, M_2, a cluster M_2 can be compared to M_1 for a given time point t (of P_1 in this case). Let $q_1(t)$ (resp. q_2) be the measurement of a megacluster at time t for partition P_1 (resp. P_2) and S and a threshold of level expression.

$$\begin{array}{c|cc} & P_1 & P_2 \\ M_1 & 1 & 0 \\ M_2 & 0 & 1 \end{array} \quad \text{we deduce from the matrix} \quad \begin{cases} \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} (t, P_1) & \text{if } S \le q_1(t) \wedge S \le q_2(t) \\ M_1 M_2 (t, P_1) & \text{if } S \le q_1(t) \le q_2(t) \\ M_2 M_1 (t, P_1) & \text{if } S \ge q_2(t) \ge S \end{cases} \tag{2}$$

Through megaclusters we can more easily combine using now measurement of a partition (i.e. P_1 as above for instance). Hence we can compute a time correlation matrix according to the preceding and use the symbol B for *before*, A for *after* and D if a current megacluster has a measurement greater than a target megacluster. We can compute the time occurrence of a given Gene at a given time point compared to next or previous time point knowing the level expression (measurement) of either activation (state p), either inhibition (state m) of clusters in which it is supposed to belong, as:

Let Mod_t be a temporal mod e $\in$ {A, D, B}, t a time point, g a given gene, G a t arg et gene and P a biochip:

$$\begin{cases} A_g(t, P) & \text{if } S \le q_g(t+1, P) \\ B_g(t, P) & \text{if } S \le q_g(t-1, P) \\ D_g(t, P) & \text{if } \dfrac{q_g(t, P)}{q_G(t, P)} \ge S \end{cases} \tag{3}$$

Table 2

Time correlation matrix for a target gene

(a)	Cluster	Target gene	$T1(P_1)$	$T2(P_1)$	$T3(P_1)$	$T4(P_1)$	$T1(P_2)$	$T2(P_2)$	$T3(P_2)$
	1	3	AD	ABD	BD	0	0	0	0
	16	3	0	A	D	B	A	D	0
(b)	State	Target gene	$T1(P_1)$	$T2(P_1)$	$T3(P_1)$	$T4(P_1)$	$T1(P_2)$	$T2(P_2)$	$T3(P_2)$
	P	3	AD	ABD	BD	B	A	D	0

Above, Table 2(a) shows an example of time correlation matrix for a target gene and its megacluster (1,16) and activation measurement; Table 2(b) gives the final resulting matrix for activation state (p), and should be completed by inhibition state (m). We use a Jaccard similarity index to compare the sub-matrix corresponding to a gene and the sub-matrix corresponding to the target gene and decide whether two genes are close to each other in their temporal profiles. These are the main steps of the algorithm in two stages. First stage is a preprocessing of data. Second stage applies consensus (Table 3).

Results

Complexity of consensus approach from CLUE library is O (n x k) in memory and O (n x k^3) in time but the DHW use an optimization solver is found on Hungarian algorithm and takes O (n^2) in space. We used a multiprocessor cluster of 162 nodes running under Sun Grid Engine, each node having 4 processors exploiting between 4 Gb and 8 Gb each one (processor Xeon EMT 64 3,2 Ghz / 4 Go; processor WoodCrest 2,33 Ghz / 8 Go). A job is assigned to the most available node. On the cluster a selection of 30 % of BiopM or BiopH microarray

Table 3

Stage 1 (left), Stage 2 of Microarray Alignment algorithm

Input: Two microarray datasets (D) (matrix with in column time-points), a measurement threshold (M) 1 – implement a classical k-means clustering on 2 Datasets (D) adding a column of unique cluster index 2 – delete measurement values < M Output: D cleaned with a column pointing to a cluster index for each gene.	Input: Datasets (D), a Given Gene (G), a threshold of expression level (S), a threshold of temporal similarity (Js) 1 – Compute mean expression values for clusters 2 – Create Gene Dictionary D 3 – Create 2 Partitions of Gene Dictionary with Clusters for D 4 – Apply consensus to obtain a unique partition P 5 – Create a Mapping MegaCluster ↔ clusters (MGC) using P 6 – Generate the Temporal Matrix (TM) for all clusters 7 – Compute a submatrix of TM for G (TMG) using MGC (as in Table 2. (b)) 8 – For each gene g of D – compute submatrix (TMg) using MGC according to expressions in (3) – compute Jaccard value J between TMG and TMg Output: List of Genes Temporally Similar to G having J > Js

size and overall of BiopB lead to memory limit of computation: 8 Gb of computation to manage a unique gene partition out of 9000 genes and 5 hours (40 minutes for consensus). We have tested the new method presented in the chapter below, running it on data described in chapter Datasets. We thus observed that (i) each target gene has a different context in each array dataset (ii) this context varies depending on the similarity threshold or megacluster size and (iii) the use of temporal gene profiles based on a symbolic time property of simultaneity and precedence identifies target gene contexts which might be of high interest (Table 4).

By studying these contexts with the IPA software, we found similar gene networks around the alg5 target gene with the bovine-human and bovine-mouse megaclusters whereas those surrounding the eif2s3 target gene were rather different (Table 5). At first glance, these networks make sense with those identified as important in embryos from cows or from other ungulates such as sheep or pig (recently reviewed in Blomberg, 2008). We thus feel confident that this approach is interesting to pursue. Whether these gene contexts make sense at certain stages more than others along the GSE time-series analysed here, in the understanding of the proliferation and differentiation processes involved in bovine embryo development, clearly awaits further studies.

Conclusion

Using published microarray datasets and comparing two time series from two different species, we addressed a question for correlation between time-points and gene comparison across microarrays where time was neither equivalent nor linear between species. All combinations of occurrence

Table 4

Gene contexts identified with our combinatorial approach
(Tb is threshold for Bovine, T is threshold for the other microarray)

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

Table 5

Gene contexts identified with our combinatorial approach and analysed with the IPA software.
The highest scores indicate the more significant functions (or top functions) identified in these gene contexts

Genes	Networks	Score
Alg5 bov hum = Alg5 bov mus	Connective tissue disorders, genetic disorders, cancer	22
	Cancer, cell to cell signalling and interaction, cellular assembly and organisation	14
Eif2s3 bov hum	Molecular transport, organ morphology, reproductive system development and function	24
Eif2s3 bov mus	Reproductive system disease, cardiovascular system development and function, organismal development	46
	Cancer, cell to cell signalling and interaction, cellular function and maintenance	22
	Cell cycle, cellular assembly and organisation, DNA replication, recombination and repair	22
	Cell cycle, cell morphology, connective tissue development and function	18
	Post-translational modification, protein folding, cancer	11

of genes between time points are possible, hence all partitions. We propose a methodology based on merged clusters and use a time correlation matrix to compute a time profile over two microarrays. This kind of combinatorial-mining approach could be a first step to multidimensional sequence alignment which seems close to known 1-dimension sequence alignment but is more highly combinatorial.

References

Aach J., Church G.M. Aligning gene expression time series with time warping algorithms // Bioinformatics. 2001. V. 17. № 6. P. 495–508.

Altschul S.F., Gish W., Miller W. *et al.* Basic local alignment search tool // J. Mol. Biol. 1990. 215 (3). P. 403–410.

Blomberg L., Hashizume K., Viebahn C. Blastocyst elongation, trophoblastic differentiation, and embryonic pattern formation // Reproduction. 2008. 135. P. 181–195.

Eisen M.B., Spellman P., Brown P.O., Botstein D. Cluster Analysis and Display of Genome-Wide Expression Patterns // Proc. Natl Acad. Sci. USA. 1998. 95(25). P. 14863–14868.

Ernst J., Nau G.J., Bar-Joseph Z. Clustering short time series gene expression data // Bioinformatics. 2005. V. 21. Suppl. 1. P. 159–168.

Hornik K. A CLUE for CLUster Ensembles // J. Stat. Software. 2005. 14(12).

Hue I., Degrelle S.A., Campion E., Renard J.P. Gene expression in elongating and gastrulating embryos from ruminants // Soc. Reprod. Fertil Suppl. 2007. 64. P. 365–377. Review.

Kucherov G., Ной L., Ponty Y. Estimating seed sensitivity on homogeneous alignments // Proc. of the IEEE 4th Symp. on Bioinformatics and Bioengineering (BIBE). May 19-21, 2004, Taichung (Taiwan). P. 387–394. IEEE Computer Society Press.

Meila M. Comparing clusterings - an axiomatic view // Proc. of the 22nd Intern. Conf. on Machine Learning (ICML'2005).

Mirkin B., Cherny L. Deriving a distance between partitions on a finite set // Automation and Remote Control. 1970. № 5. P. 120–127.

Mitiku N., Baker J.C. Genomic analysis of gastrulation and organogenesis in the mouse // Dev. Cell. 2007. 13(6). P. 897–907.

Smith T.F., Waterman M.S. Identification of common molecular subsequences // J. Mol. Biol. 1981. 147. P. 195–197.

Turenne N., Schwer S.R. Temporal Representation of Gene Networks // J. of Data Mining and Bioinformatics (JDMB). 2008. V. 2(1).

Ushizawa K., Herath C.B. *et al*. cDNA microarray analysis of bovine embryo gene expression profiles during the pre-implantation period // Reprod. Biol. Endocrinol. 2004. 24. 2:77.

Xu R.H., Chen X., Li D.S. *et al.* BMP4 initiates human embryonic stem cell differentiation to trophoblast // Nat. Biotechnol. 2002. 20(12). P. 1261–1264.