Title

Matěj Holec, Filip Železný, Jiří Kléma

Czech Technical University, Prague
e-mail: {holecm1, zelezny, klema}@fel.cvut.cz

Abstract. ble ble

1 Introduction

In recent years was developed methods and equipment for measuring high-throughput genomic data. In spite of fascinating by these data principally able to solve non-trivial biological valuable questions attempts fail due a lack of samples and an abundance of attributes. In this paper is described our experiment with merging public available datasets coming from different but enough similar organisms and reducing number of attributes by using generalization for genes. We use fully-coupled fluxes (FCF) as abstract attributes for our transformation.

Genomic measurements we are using were acquired by different Affymetrix DNA microarrays (MA). As the ground for abstract attributes we are using orthologous genes which are involved in a same place and similar function in the pathway (for all used species). In particular in metabolic/signalling pathway [1] which describe chemical reactions of gene products and substrate molecules within a cell. This property offers merging data according to enzymes which exhibits same behavior in different organisms. To improve robustness we are using as long as possible linear pathway subgraphs - FCF - simply structures with predictable behavior as gene substitution for our new attributes. Unfortunately above mentioned data consist a few types of difficulties. There is not simple relationship between value gotten by MA probe corresponding to activity of some gene and activity of FCF. Affymetrix MA are using different gene identifiers (Affymetrix Id) then genes used in pathways (Entrez Gene Id) generally with relation *n:m*. Different MAs use different probe set. Different enzymes in species can be composed from different count of genes. There are missing values due non-existency of incompleteness of pathways and non-existency of translations between some Affymetrix Id and Entrez Gene Id. Thus all these data are in principle multirelational containing uncertainty and incompleteness.

2 Relational Representation of Domain

In this experiment we use datasets from three different species. From human, mouse and rat. Samples are from two different tissues. First one belongs to

[1] downloaded from KEGG genome database

hematopoetic and second one to stromal type of cells. These cells are involved in blood cells producing in bone marrow.

MA measures level of expression of genes corresponding to gene probes. Gene expression is the process of transformation genetic information to gene product as protein or RNA. The pathways are graph (see fig. 3), where nodes are protein coding genes (or gene complexes) and edges are different types of interactions[2]. Alternatively vertexes can be some compound molecules or links to another pathway, but we are only interested in reactions in FCF between enzymes where one reaction initiates another reaction and vice verse in experiments in this paper.

Nature of the data

Public available data are stored in different databases and formates. For example MA samples are available as plain text files where row contains MA probe identifier and expression value and kegg pathways as xml files which have strictly defined structure[3]. We transformed all the data to prolog facts make us possible to classify this relational data in many ways in future. All transformations were intended to preserve as much as possible features of used data. In fact MA samples are translated to prolog facts without loss and the only one difference between pathways in xml and in a prolog database is that a graphical representation of elements is cut away. Other facts except translation between gene identifiers are artificially created and can be incomplete.

entry(Organism,PatwayID,KeggNodeID,ListOfEntrezID,NodeType).

```
entry('hsa',04520,1,[hsa:4089],'gene').
entry('hsa',04520,2,[hsa:999],'gene').
entry('hsa',04520,3,[hsa:1495,hsa:1496,hsa:29119],'gene').
```

relation(Organism,PathwayID,BeginNodeID,EndNodeID, TypeOfRelation,SubtypeName,SubTypeValue).

```
relation('hsa',04520,71,77,'pprel',['activation'],['-->']).
relation('hsa',04520,14,16,'pprel',['activation','phosphorylation'],['-->','+p']).
relation('hsa',04520,4,6,'pprel',['bindingassociation'],['---']).
```

Pathway graph - transcription of kegg xml pathways into prolog facts. See Kegg pathways and Kegg markup language for more info.

e(SampleID,AffyID, ExpressionValue).

```
e('GSM101111', 'AFFX-BioB-M_at', 1016.3).
e('GSM101112', 'AFFX-BioB-3_at', 529.4).
e('GSM101113', 'AFFX-BioC-5_at', 1219.1).
```

[2] http://www.genome.jp/kegg/document/help_pathway.html

[3] http://www.genome.jp/kegg/docs/xml/

Expressions of probes as prolog facts.

array(CellType, GDSnumber, SampleID, TissueState, Organism, ?’LSK’, ?’RNA’, MArray_ID, Comments).

```
array(hematopoetic,’GDS2718’, ’GSM169465’, ’normal’, ’Mus musculus’,
’LSK’, ’RNA’, ’GPL1261’, ’GCOS 1.4 software (Affymetrix)’).
array(stromal,’GDS2231’, ’GSM75448’, ’treated’, ’Rattus norvegicus’,
’msc’, ’RNA’, ’GPL341’, ’Affymetrix Microarray Suite version 5.0’).
array(stromal,’GDS1288’, ’GSM38627’, ’normal’, ’Homo sapiens’, ’msc’,
’RNA’, ’GPL96’, ’Affymetrix Microarray Suite version 5.0’).
```

Dataset description holding information about a sample.

affy2entrez(ChipID,AffyID,EntrezID).

```
affy2entrez(’GPL1261’,’1452692_a_at’,72900).
affy2entrez(’GPL1261’,’1422823_at’,13860).
affy2entrez(’GPL1261’,’1443502_at’,329581).
```

Facts containing info necessary for translating between affymetrix ID and Entrez ID. Extracted from BioConductor.

flux(FCF id,PathwayID,List of Kegg nodes).

```
flux(flux8,61,[102,107]).
flux(flux9,62,[6,7,8,9,10,11,12,13,14,15,16,17,18,19]).
flux(flux10,62,[22,23,24]).
```

FCF generated from all pathways from kegg.

kChipID(AffyID_1,AffyID_2,Correlation).

```
kGPL1261(’1422433_s_at’,’1422433_s_at’,1).
kGPL1261(’1422433_s_at’,’1419821_s_at’,0.795838356519116).
kGPL1261(’1422433_s_at’,’1444265_at’,0.613436992054866).
```

Generated fact how much are different probes correlated for all samples in some MA.

candidate(ChipID,FluxID,Set of candidate AffyIDs,Correlation between AffyIDs)

```
candidate(’GPL1261’,flux0,[’1451002_at’,’1450048_a_at’],2.86).
candidate(’GPL1261’,flux1,[’1415918_a_at’],1.0).
candidate(’GPL1261’,flux2,[’1448894_at’,’1419456_at’],2.91).
```

Generated candidates for best repersentants of FCF.

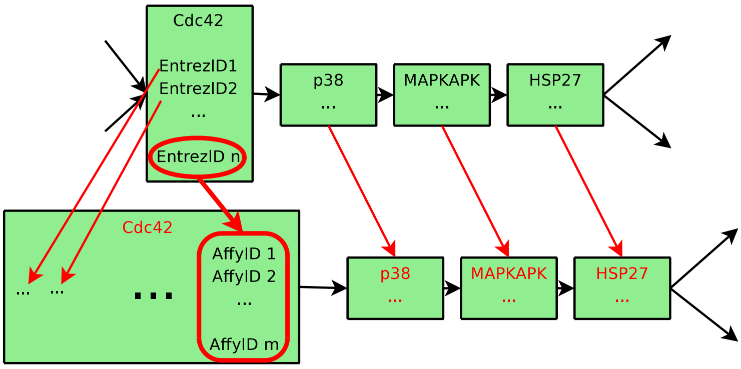

Fig. 1. Relation between Affymetrix and Entrez identifiers in fluxes.

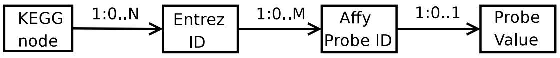

Fig. 2. Types of relationship between identifiers and values in fluxes.

Translation between expression in a sample and expression of a some FCF

????

On figure 1 is depicted process of transforming of expression of genes to expression of FCF. Our data samples have values of expression of genes encoded in Affymetrix IDs. Pathways have genes encoded by Entrez Gene identifiers, thus for evaluating of expression of FCF we get with method depicted on figure 2. This way we have set of values describing activity of whole FCF subgraph.

On figure 2 are depicted relations between elements in some FCF. FCF is compsed from KEGG nodes. To every node can be assigned set of expression of particular probes according to these relations.

Experiments - expression of maximal correlated candidates and average expression of FCF

????

Our goal is to get a single value describing of activity of FCF. We are using two different approaches (presne oduvodneni?). The first one consists in averaging of all set values assigned to FCF - particulary independent on fact from which FCF node some value comes from (see fig. 5) . The second one is generated by selecting genes for every FCF node (described by Affymetrix Id) which are maximally correlated with another ones from other FCF nodes. Sets of those

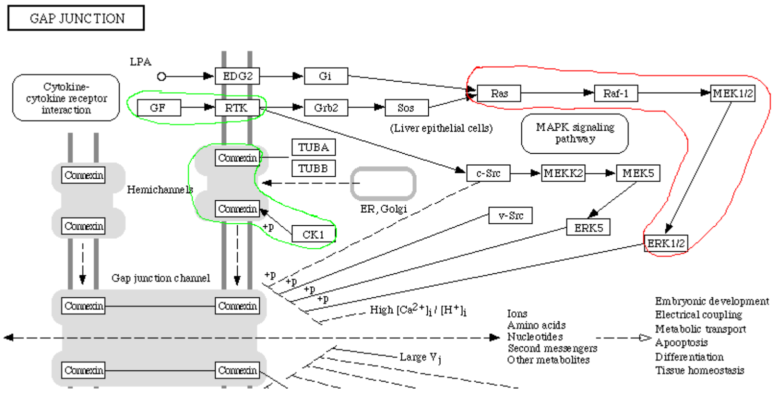

Fig. 3. Pathway with distinguished high-correlated fully-coupled fluxes with significantly higher average expression in some class.

genes (with same length as FCF) are called candidates. Value in output dataset are computed as average expression of genes from candidate for every sample. (see fig. 4)

Because measured level of gene expression by probes is relative we are using normalization on every row of output. Normalization consist in substraction of avarage expression of fluxes and dividing by its own standard deviation.

-max corr: najdou se hodnoty exprese kandidata a jejich suma se vydeli delkou kandidata.

-u avg se vytvori list vsech affyIDs, udela se suma odpovidajcih expresi, vysledek je suma vydelena delkou listu.

-pri zobrazovani do pca se vyhodi sloupce tech fluxu, kde se nekde vyskytuje NA. takze misto 476 fluxu je ve vyslednych obrazcich pouzito pouze 208fluxu

3 Baseline Results

We generated merged dataset from all three species with FCF as attributes. We selected a few FCFs according to flold change for both classes (see highlighted FCFs in fig. 3) and the variance ... . Secondly we transformed merged datasets by PCA into 2D space. Figure 4 depicts PCA of FCF which have values computed as average of maximally correlated candidates, figure 5 showes same samples but FCF values are computed as average of all expressions of genes contained by some FCF. Note the normalization which was improved by cut of columns which contains some ”not availabile” value only for PCA.

-data pripravena pro hlubsi multirelacni analyzu

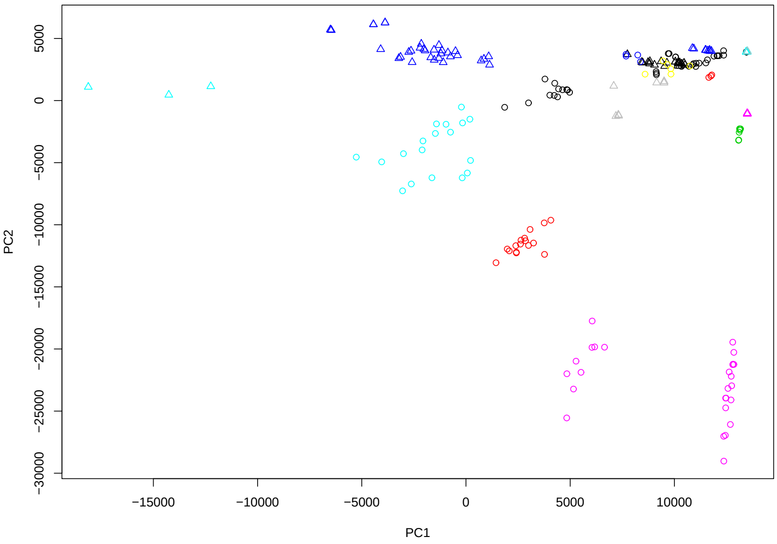

Fig. 4. PCA of fluxes.

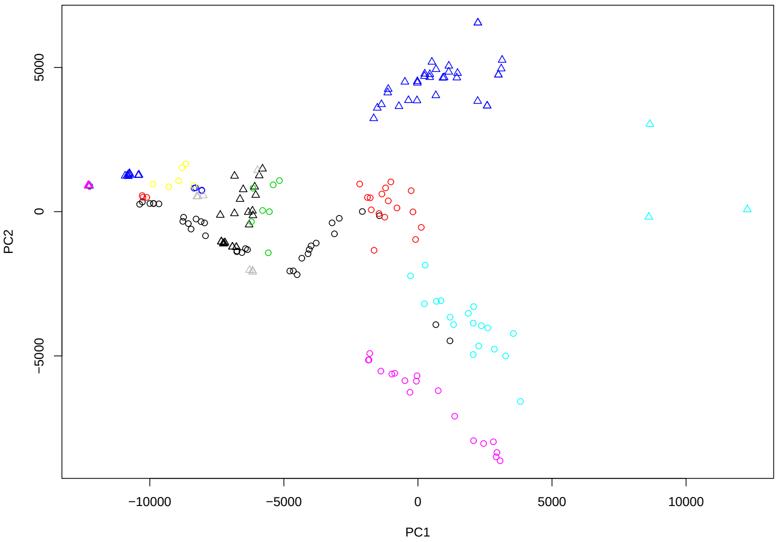

Fig. 5. PCA of fluxes.

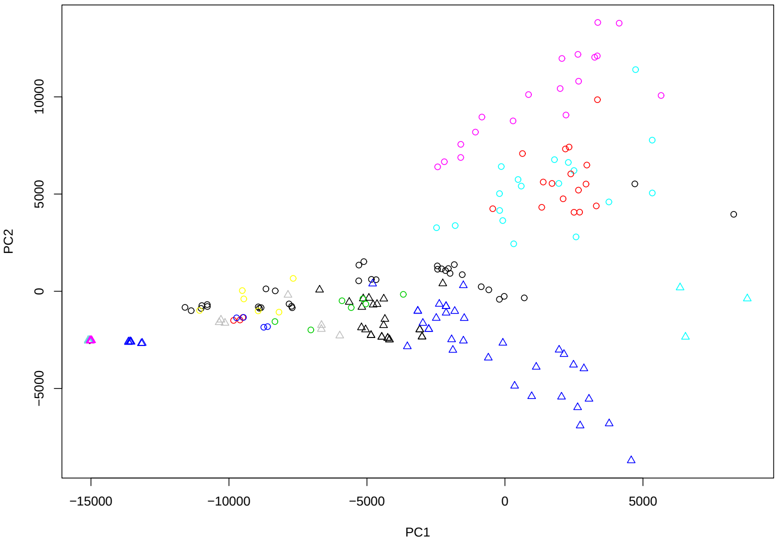

Fig. 6. PCA of pathways.
Points correspondes to fluxes and pathways, triangles and rings distinguishes the class, MAs have different colors.

References

1. Notebaart RA, Teusink B, Siezen RJ, Papp B (2008) *Co-Regulation of Metabolic Genes Is Better Explained by Flux Than Network Distance*. PLoS Computational Biology 4(1): e26 doi:10.1371/journal.pcbi.0040026