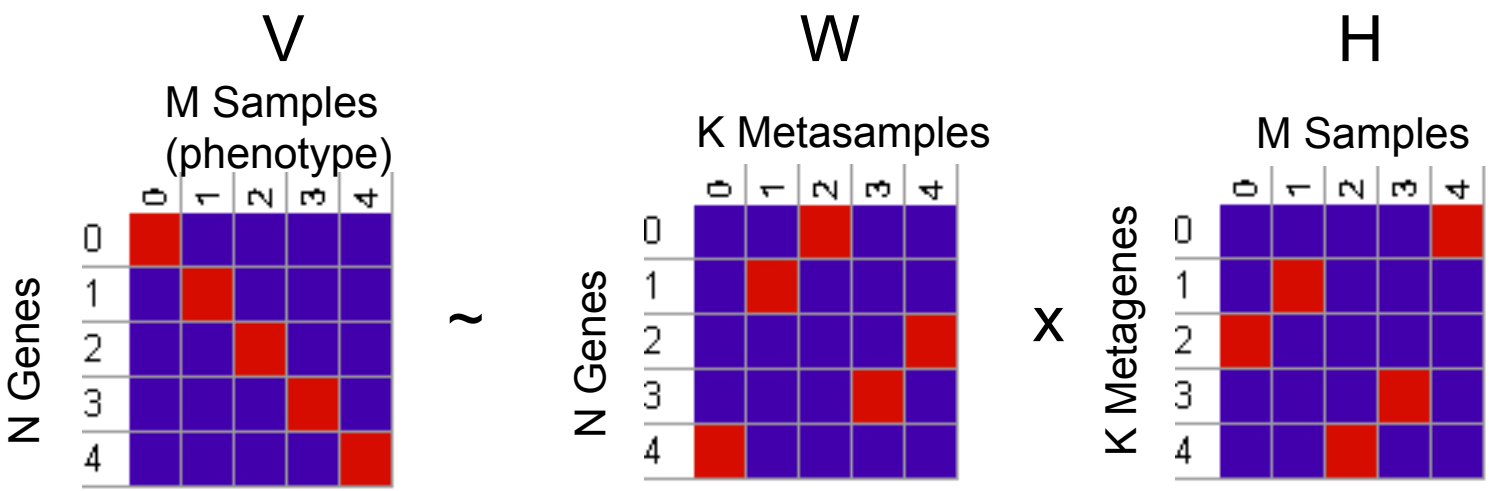
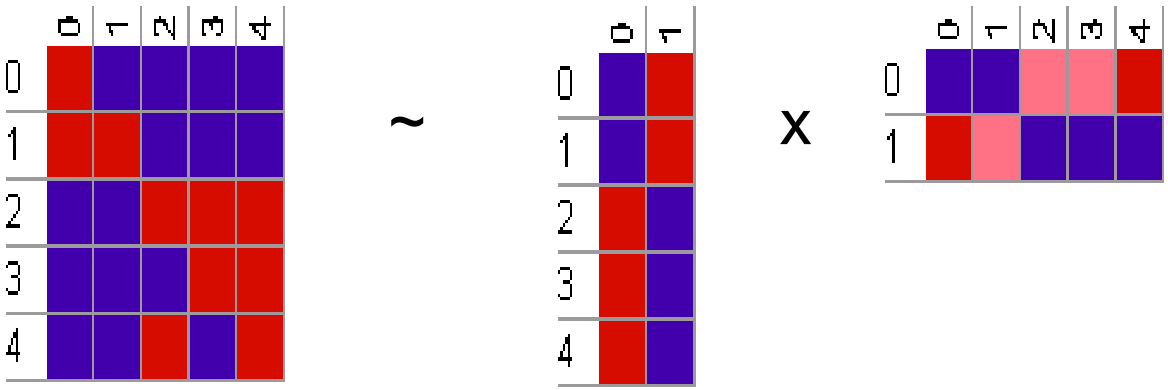
Non-Negative Factorization Examples and Analysis of Encoding and Clustering Tendencies (as a way to interpret the meaning of “metasamples” and “metagenes” in the context of “discovered” patterns, clusters and elements of a biological “by parts” representation).

Definitions:

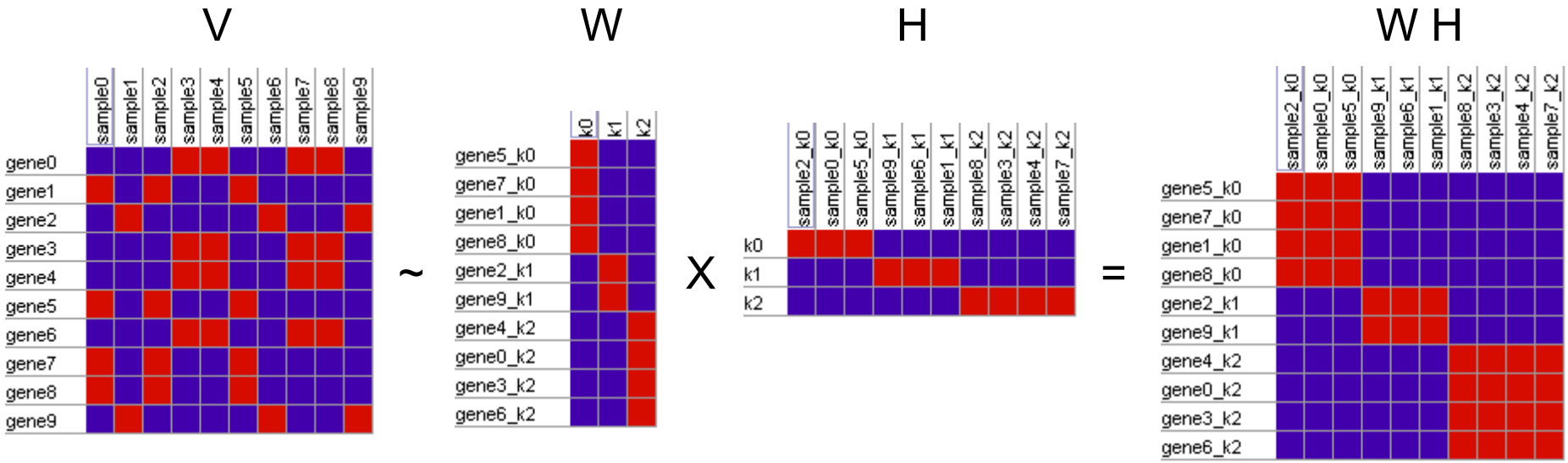


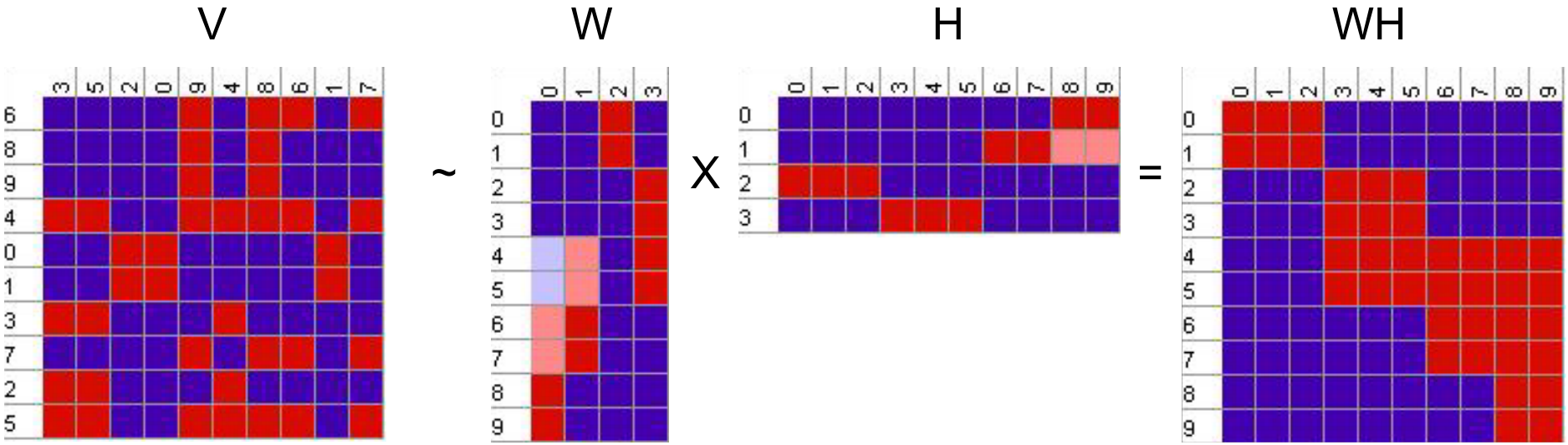
In a real microarray example the “cells” may represent a group of several genes or samples with similar expression.

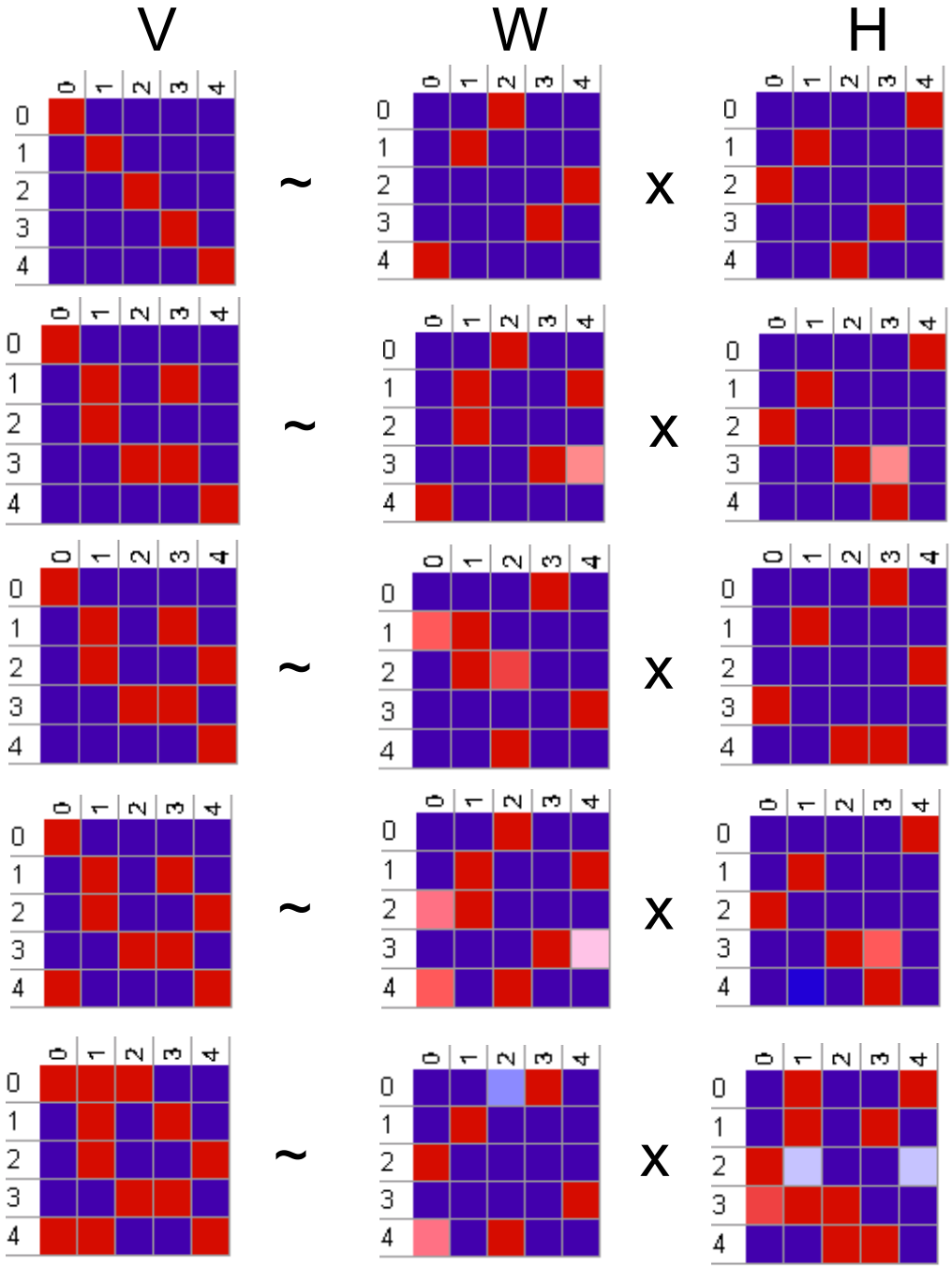
V W H



The phenotypes have mainly expression in genes 0-1 and 2-4. It recognizes that and separates them into to groups of metasamples. The same clustering takes place with the metagenes.


V
W
H
W H
sample0
sample1
sample2
sample3
sample4
sample5
sample6
sample7
sample8
sample9
gene0
gene1
gene2
gene3
gene4
gene5
gene6
gene7
gene8
gene9
~
k0
k1
k2
gene5_k0
gene7_k0
gene1_k0
gene8_k0
gene2_k1
gene9_k1
gene4_k2
gene0_k2
gene3_k2
gene6_k2
X
sample2_k0
sample0_k0
sample5_k0
sample9_k1
sample6_k1
sample1_k1
sample8_k2
sample3_k2
sample4_k2
sample7_k2
=


V
W
H
WH
~
X
=
3 5 2 0 9 4 8 6 1 7
6 8 9 4 0 1 3 7 2 5
0 1 2 3
0 1 2 3 4 5 6 7 8 9



Encoding Tendency

Same number of factors as genes and samples

Each phenotype and gene is unique and orthogonal and this is preserved by the encoding. As the data is already sparse the encoding doesn't produce new patterns.

W and H are similar to V (with different column and row order) with one notable difference: gene 1 is decoupled and produced by combining metagenes 1 and 4. Metagene 4 is a new pattern that helps to build the profiles for genes 1 and 3. Notice the small increase in sparseness in matrix H (6 red cells) compared with V (7 red cells). This is an example of the encoding "by parts." As the V matrix is symmetric the same encoding trick could have been applied to the samples (e.g. making sample 3 and 1 with a metasample that has only gene 3 expressed. This break of symmetry is probably due to the fact that W is updated before H.

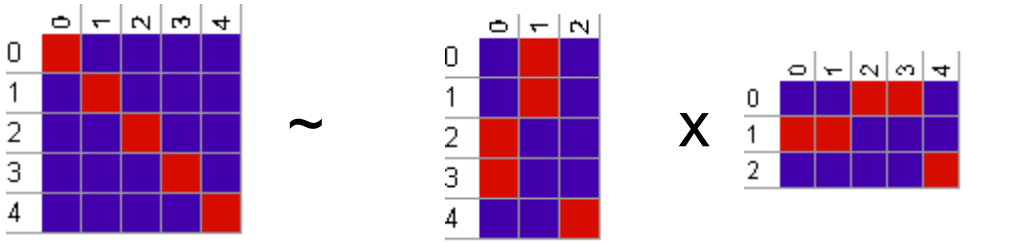
Similar metagenes as previous example. Previous metasamples 1, 2 and 3 are now 1, 3 and 4. The additional expression of gene 2 in sample 4 changes the encoding of previous metasamples 0 and 4. The new metasamples 0 and 2 allow to reproduce a V with higher density (8 red cells) with the same density in W (7 red cells).

The metasamples are very similar to the samples. The metagenes decompose the gene patterns in an efficient way using almost orthogonal rows. This allows to reproduce the V matrix with 9 red cells with only six red cells in H.

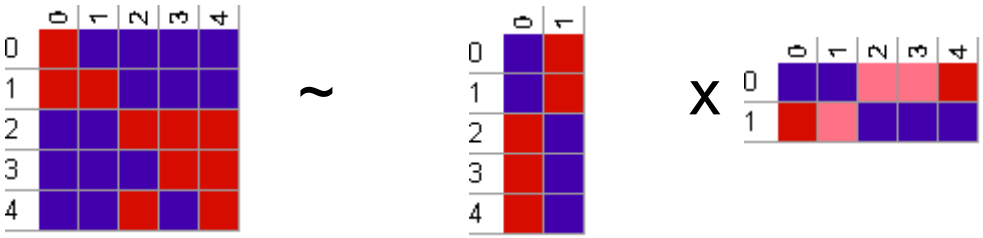
Here the "by parts" decomposition works better for W than H. The gene patterns are harder to decouple and the metagenes look closer to the original genes (no increase sparseness. In contrast W is more sparse with 6-7 red cells instead of 11 for V). There is more sample "overlap" in this case and then for example samples 1 is decomposed in metasamples 3, 2, 1 and 0.

Clustering Tendency

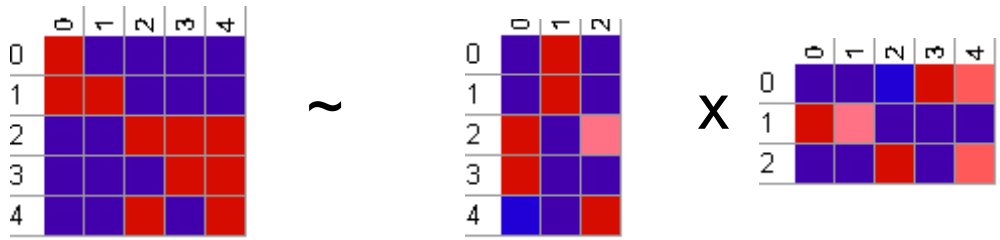
Same number of factors as genes and samples



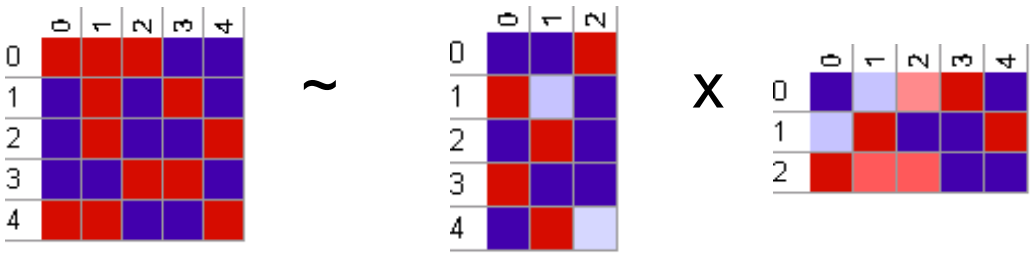
Five orthogonal phenotypes and only three metasamples. It groups them in two groups of two and then one apart.



The phenotypes have mainly expression in genes 0-1 and 2-4. It recognizes that and separates them into to groups of metasamples. The same clustering takes place with the metagenes.

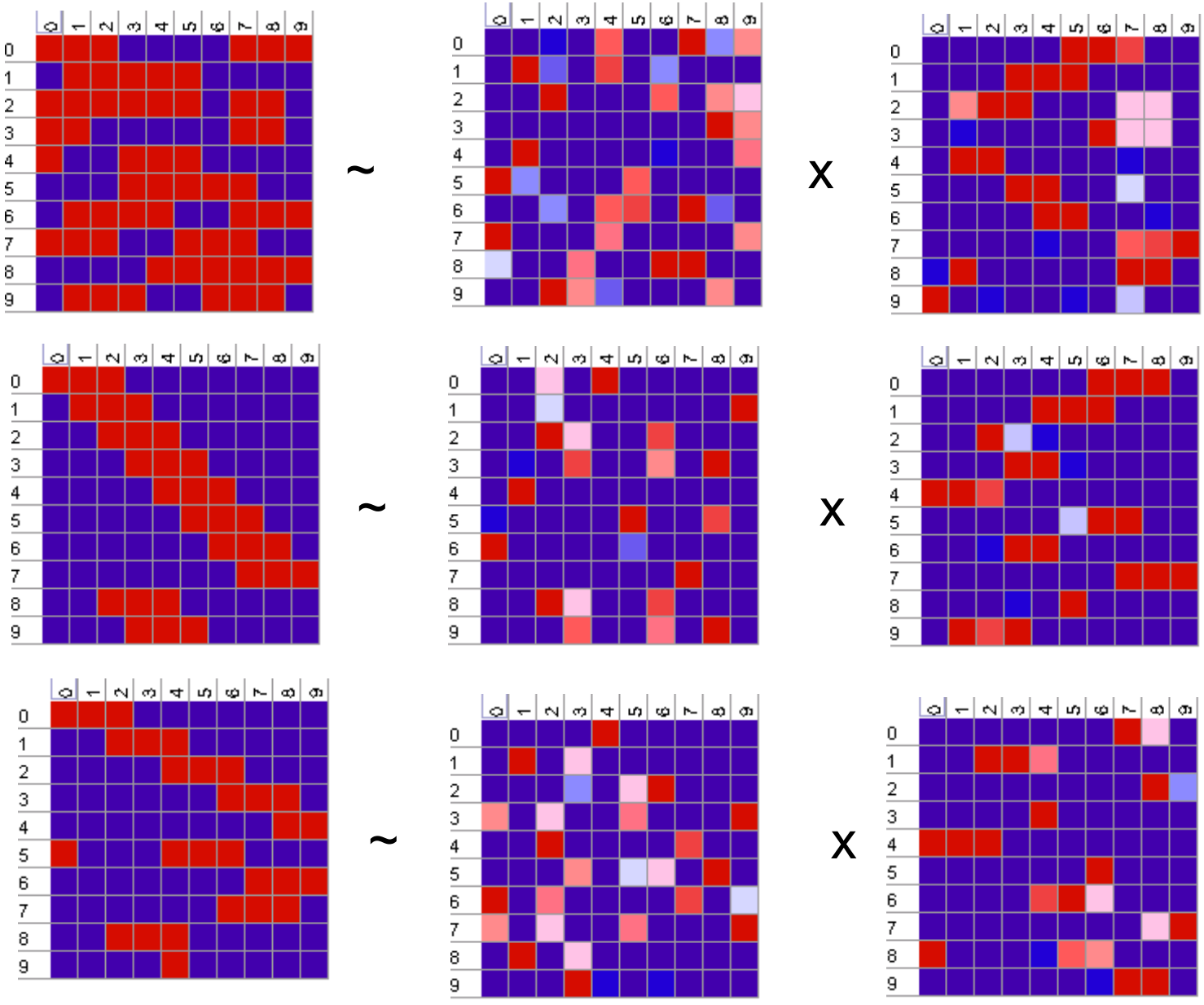


Same as before but there is one extra metasample that picks up the phenotype of sample 2. Metagene 0 from the previous example now decouples in two. One (2) now accounts for the new metasample (2).



The samples so not easily cluster in less than 5 groups. Here we see a combination of clustering and encoding. It separates the phenotypes of samples 0 and 3 in metasamples 2 and 0 but encodes samples 1, 2 and 4 into metasample 1 and the partially in metasamples 0 and 2. This is illustration of the encoding by parts: sample 1 is mainly made of metagenes 1 and 2 which combine metasamples 1 and 2.

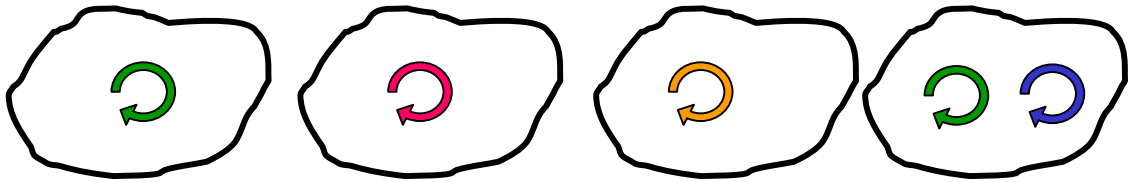
V W H



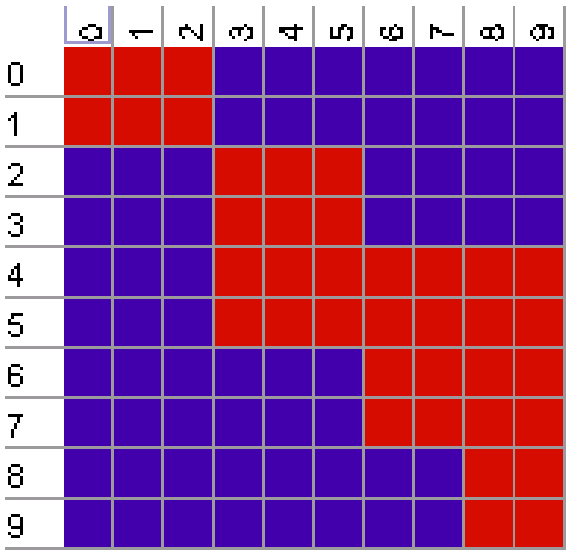
Biological Model and example

4 metagenes

4 metasamples

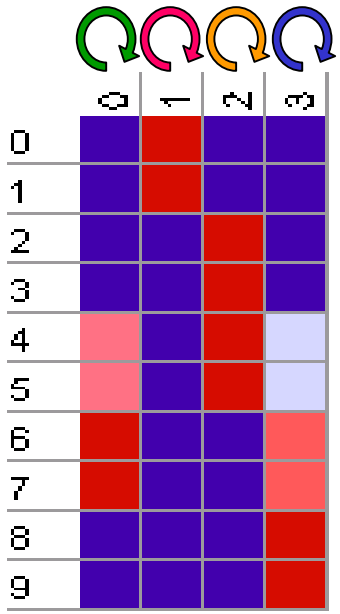


V



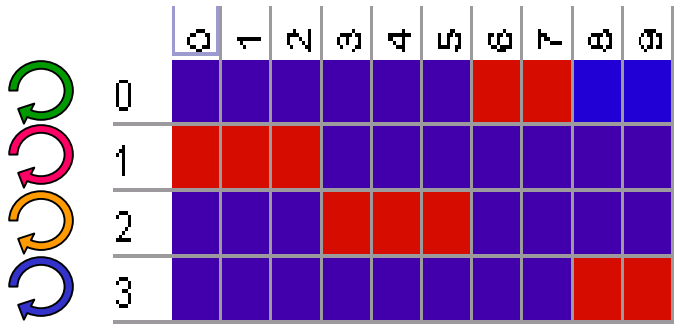
~

W

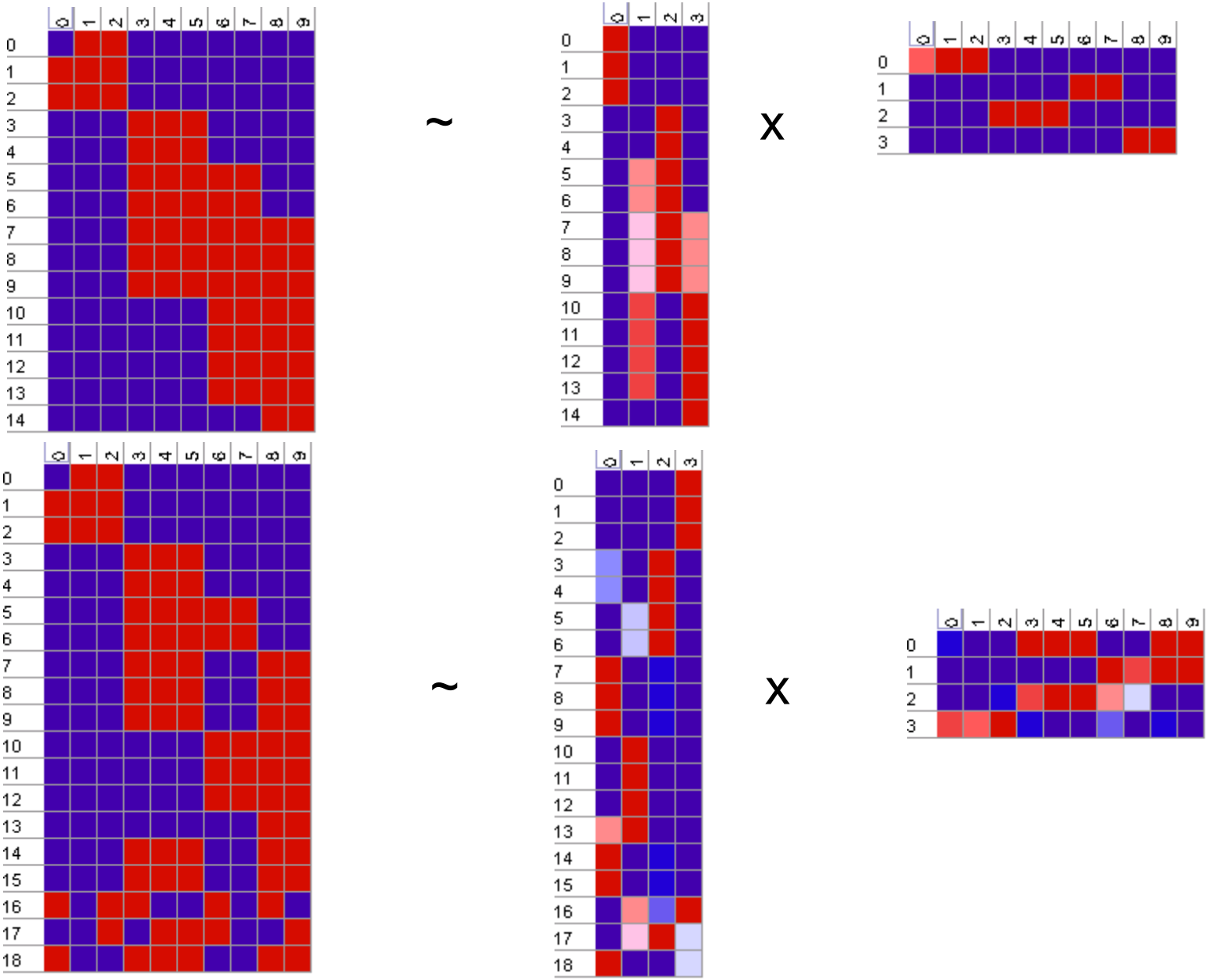


x

H



V W H



Summary of trends:

Very similar patterns of genes or samples get cluster together. This is a consequence of the number of factors being less than the number of genes or samples but also of the trend of increase sparseness in the “by parts” decomposition.

A sparse W (or H) as a result of a significant decomposition by parts may imply a less sparse H (or W) because more parts are needed to recreate the samples (or genes).

If V mainly contains orthogonal samples and gene patterns overlap significantly (i.e. V has “row patterns”) then the genes get decomposed in parts and H will be sparse.

If V mainly contains orthogonal genes and samples patterns overlap significantly (i.e. V has “column patterns”) then the samples get decomposed in parts and W will be sparse.