Independent component analysis of gene expression data

Wolfram Liebermeister, Theoretische Biophysik Humboldt-Universität zu Berlin, Germany wolfram.liebermeister@rz.hu-berlin.de

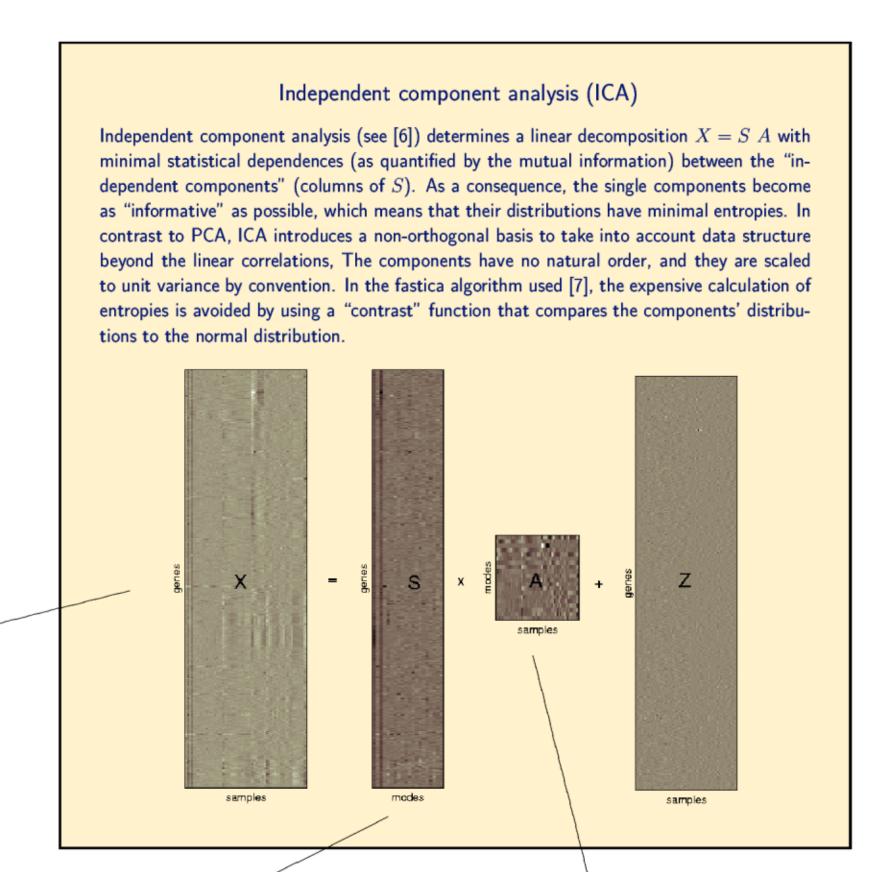
We applied independent component analysis (ICA) to gene expression data, inferring hidden variables which we term "expression modes". According to the ICA model, the modes exert linear influences on the genes with minimal statistical dependences between them. The dominant modes obtained from a set of yeast data could be related to separate biological functions. A projection to these modes helps to determine sets of coregulated genes, to visualize the data and to compress them in a biologically meaningful way.

The problem

Cells react to external stimuli and to their internal needs by the induction or repression of genes. Genomic scale patterns of gene expression can be observed using high-throughput methods like the microarray technique. One may hope that correlations in these large data sets reveal causal relationships between the genes. This is the main idea behind various kinds of genetic network models, like linear, nonlinear, or discrete dynamical systems, or Bayesian networks. However, it seems that until now the microarray data are too noisy and that the number of experiments is not yet sufficient to reconstruct detailed large-scale genetic networks. Thus, methods are needed to reduce large amounts of data to their most relevant aspects, in particular the coregulation of genes and characteristal patterns of cell sample types. One has to keep in mind that with all multivariate methods the results my depend heavily on data metric, and thus on the normalization scheme used.

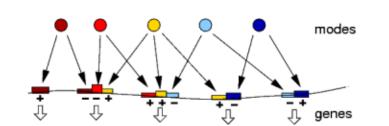
Yeast data from Eisen et al. (1998)

The gene expression matrix X contains intensity ratios related to relative mRNA levels of 2467 yeast open reading frames (ORFs). The samples represent timecourses different situations: cell replication cycle (synchronization with the mating α factor or using small G1 cells obtained by elutriation) sporulation, heat shock, response to a reducing agent, cold shock and diauxic shift from fermentation to respiration. We preprocessed the data by shifting the gene and sample means to zero, replacing the missing values by zeros and projecting the data to their first 20 principal components. The higher principal components are contained in the additive noise term Z.



Linear models of gene expression data

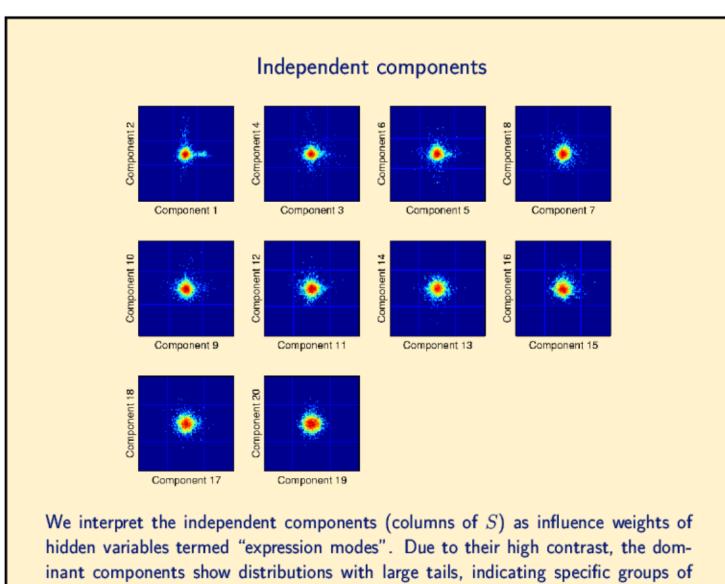
Clusterings (for instance hierarchical, k-means, self-organized maps) are a widely used method to determine sets of coregulated genes or sets of cell samples with similar gene expression. In a more detailed view, each gene's expression depends on a group of cellular regulators that may act together in some nonlinear way. Linear models implement the idea of a combinatorial control, describing the expression levels of genes as linear functions of common hidden variables.



Technically, the gene expression matrix X is split into a product X = S A, representing each gene profile (row of X) as a linear combination of "mode profiles" (the rows of A), the coefficients ("components") being contained in the columns of S. Linear models like principal component analysis (PCA) [1], the plaid model [3], REDUCE [4], or ICA rely on different criteria to determine the modes. It would be desirable that some of the modes could be related to biological causes of variation, like regulators of gene expression, cellular functions, or responses to experimental treatments. The components would then describe the modes' influences on the genes. Once a small number of effective key variables has been identified, they max be described by simple dynamic models (see for example [5]).

Principal component analysis (PCA)

Principal component analysis rotates the data to a new orthonormal basis which is formed by the eigenvectors of the data covariance matrix. In this new basis, linear correlations are removed, and a minimal number of the new variables ("principal components") explains a maximal amount of data variance. PCA is widely used to reduce the data dimensionality, while maintaining a maximal amount of variance and filtering out variation that is likely to represent noise. Usually, there is no obvious biological interpretation for the single components.



"target" genes. For each component, we determined these targets by excluding the most outlying genes until all remaining genes were inside $n_{\sigma}=4$ standard deviations from their median.

Expression mode profiles

Mode profiles (rows of A, samples shown on the abscissa) from yeast data obtained by ICA. The different gene profiles (rows of X) are represented as linear combinations of these basis patterns (up to a noise term). The colors indicate different experiments.

PCA expression modes

Mode 1	Mode 2
Mode 3	Mode 4
Mode 5	Mode 6
Mode 7	Mode 8
Mode 9	Mode to
Mode 1	Mode 52
Mode 13	48 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
WWW MAN	Mode 15
Mode 17	Mode 15
Mode 13	Mode 20
samalas	samples.

Mode profiles (basis vectors) from yeast data obtained by PCA.

Target genes

ribosomes RPL10 RPL11A RPL11B RPL12A RPL12B RPL13A RPL14A RPL14B RPL15A RPL15B RPL16A RPL16B RPL17A RPL17B RPL18A RPL18B RPL19A RPL19B RPL1A RPL1B RPL20A RPL20B RPL21A RPL21B RPL22A RPL23A RPL23B RPL24A RPL24B RPL25 RPL26B RPL27A RPL27B RPL28 RPL28 RPL30 RPL30 RPL31A RPL33A RPL33B RPL34B RPL35A RPL35B RPL37A RPL37B RPL38 RPL40B RPL41A RPL42A RPL42B RPL43A RPL4A RPL5 RPL6B RPL7A RPL7B RPL8A RPL8B RPL9A RPL9B RPP0 RPP1B RPP2A RPP2B RPS0A RPS0B RPS10A RPS11A RPS12 RPS13 RPS14A RPS15 RPS16B RPS17B RPS18A RPS19A RPS19B RPS1A RPS1B RPS2 RPS20 RPS21A RPS21B RPS22A RPS22B RPS23A RPS23B RPS24A RPS24B RPS25A RPS25B RPS26A RPS26B RPS26A RPS28B RPS29A RPS29B RPS3 RPS4A RPS4B RPS5 RPS6A RPS6B RPS7A RPS7B RPS8A RPS8B RPS9A RPS9B SIK1 protein synthesis EFT1 EFT2 YEF3 protein degradation RPS31 stress SSB2 YHB1

cell cycle CDC2 CDC21 CDC45 CDC9 CLB5 CLB6 CLN1 CLN2 EGT2 GIC2 MSB2 RAD53 SWE1 SWI4 , down: CDC20 CDC5 dna replication CTF4 DPB2 POL1 POL12 POL2 POL30 PRI2 RFA1 RFA2 RNR1 DUN1 MSH2 MSH6 OGG1 PMS1 RAD27 RAD51 RHC18 RNR3 , down: ALK1 chromatin HIF1 SMC3 cytoskeleton BNI4 BUD9 RSR1 SPH1 SRO4, down: BUD4

meiosis BBP1 cell wall CSI2, down: CHS2 CWP1 transcription RNH35 protein synthesis and targeting SEN34 GOG5 MNN1 OCH1 SUR1 , down: NCE102 metabolism + transport PHO8 HXT2 PYC1 , down: PDR5 signal ASF1 ASF2 HCM1 SPT21 SVS1, down: ACE2 HST3

cell cycle CLB1 CLB2 PDS1 SIM1, down: CDC46 CLN3 EGT2 FAR1 GIC2 PCL2 dna replication ALK1, down: MCM3 chromatin HHF1 HHF2 HHO1 HHT1 HHT2 HTA1 HTA2 HTA3 cytoskeleton SPC98 STU2, down: CHS1 cell wall CWP1 ECM33 FKS1 WSC2, down: CTS1 transcription FIR1 protein synthesis PMT4 stress, down: CTT1 HSP150 HSP26 TIP1 YGP1 mating, down: ASH1 GPA1 STE2 sugar metabolism, down: ALD6 metabolism, down: PHO11 PHO5

mating, down: AFR1 FIG1 FIG2 FUS1 FUS2 FUS3 KAR3 KAR4 KAR5 MID2 SST2 STE2 cytoskeleton, down: CHS1 CIK1 RVS161 meiosis, down: IME4 cell cycle CDC6 CLB1 CLB2, down: FAR1 PCL2 cell wall CHS2 CWP1, down: CHS3 CTS1 GFA1 stress YGP1 protein synthesis MNN1, down: KTR2 protein targeting NCE102 metabolism ELO1 GCV1 CAR2 PHO11 PHO12 PHO5 signal SVS1, down: HAP4

cell cycle APC4 CDC10 CDC20 CDC27 CDC3 CDC5 CLB3 CLB4 dna replication PES4 DHS1 cytoskeleton BNR1 CNM67 cell wall CWP1 transcription MIP6 protein synthesis SOL1 sugar metabolism GCR3 GIP1 HXT10 HXT14 PIG1 respiration GDS1 transport CCC1 GNP1 SUT1, down: MEP2

transcription MAK16 MRT4 NMD3 RPA49 ribosomes DBP3 HCA4 LCP5 NOP4 RLP7 ROK1 SOF1 protein synthesis PUS1 ENP1 HMT1 protein targeting BFR2 NSR1 transport PHO84 rest NOP2 URA7 cell cycle, down: EGT2 TFS1 cell wall, down: CTS1 GLC3 stress, down: HSP30 HSP78 sugar metabolism, down: GIP2 GLK1 PGM2

cell cycle, down: EGT2 PCL9 SIC1 cytoskeleton, down: BUD9 CHS1 meiosis, down: RME1 cell wall CWP1, down: CTS1 EXG1 SUN4 mating, down: ASH1 stress CTT1 HSP26 SSA3 SSA4, down: CTA1 HSP150 HSP30 protein targeting NCE103 sugar metabolism ALD6 respiration NCA3 CIT1 IDH1 PYC1 metabolism ARO9 DIP5 CAR2, down: FAA4 FAS1 transport PDR15 rest, down: CUP1A CUP1B

The first modes represent biological functions

mode	induced functions	downregulated functions
1	ribosomes	
2	S-phase, replication	M-phase, cytoskeleton
3	cell division, glucose repression	
4	histones, cell wall	stress, mating
5	protein production	cell wall, stress
6		mating
7	stress, TCA cycle	

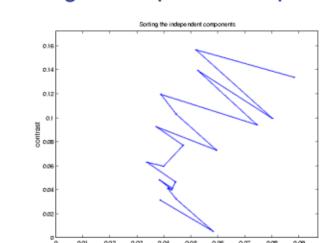
The first 7 expression modes act on sets of genes related to particular biological functions. Accordingly, the modes could be related to cell cycle phases (2 and 4), protein synthesis (1), sporulation (3), stress (5), and to the mating response (6).

ICA infers a non-orthogonal basis that can be used to visualize data, to reduce their dimension, and to define data metrics that highlight relevant aspects in the data. Although ICA uses no external knowledge, the modes obtained in our example could be related to biological functions.

Conclusion

Can ICA generally be expected to detect biological modes? In particular, is the independence postulate a plausible assumption to determine effective regulators or global modes of gene expression? One may expect biological regulators to act on specific (though overlapping) sets of genes. Accordingly, their influences on all genes should show a "supergaussian" distribution with heavy tails (related to the influenced genes) and a high peak in the center. ICA is sensitive to such modes, as their influences are far from being normally distributed.

Sorting the independent components



We expect to find among the independent components some which are of biological significance. These can be supposed to show large values of variance (abscissa) and contrast (ordinate) a measure of information content. We sorted the independent components according to a linear combination of both quanti-

References

- [1] Orly Alter et al. (2000). Singular value decomposition for genome-wide expression data processing and modeling, PNAS 97 (18), 10101-10106
- [2] Pablo Tamayo et al. (1999). Interpreting patterns of gene expression with selforganising maps: methods and application to hematopoietic differentiation, PNAS 96, 2907-2912
- [3] Laura Lazzeroni and Art Owen (2000), Plaid models for gene expression data, Technical report 211, Stanford Biostatistics Series
- [4] Harmen J. Bussemaker et al. (2001). Regulatory element detection using correlation with expression. Nature Genetics 27, 167-171
- [5] Neal S. Holter et al. (2001). Dynamic modeling of gene expression data, PNAS 98, 1693-1698
- [6] Aapo Hyvärinen, Survey on independent component analysis (1999), Neural Computing Surveys, 94-128
- [7] Aapo Hyvärinen and Erkki Oja (1997), A fast fixed-point algorithm for independent component analysis, Neural computation 9, 1483-1492
- [8] Michael B. Eisen et al. (1998). Cluster analysis and display of genome-wide expression patterns, PNAS 95, 14863-14868