FEBS Letters 507 (2001) 114–118 FEBS 25367

The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data

Marco Crescenzi, Alessandro Giuliani*

Laboratory of Comparative Toxicology and Ecotoxicology, Istituto Superiore di Sanita', Viale Regina Elena 299, 00161 Rome, Italy

Received 13 July 2001; revised 20 September 2001; accepted 25 September 2001

First published online 5 October 2001

Edited by Veli-Pekka Lehto

Abstract By using principal components analysis (PCA) we demonstrate here that the information relevant to tumor line classification linked to the activity of 1375 genes expressed in 60 tumor cell lines can be reproduced by only five independent components. These components can be interpreted as cell motility and migration, cellular trafficking and endo/exocytosis, and epithelial character. PCA, at odds with cluster analysis methods routinely used in microarray analysis, allows for the participation of individual genes to multiple biochemical pathways, while assigning to each cell line a quantitative score reflecting fundamental biological functions. © 2001 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.

Key words: Gene expression data; Multidimensional data analysis; Biological model; Cancer; Gene regulation

1. Introduction

DNA microarray technology has made it possible to monitor gene expression levels on a genomic scale involving thousands of different gene products. In order to treat very large amounts of data and derive useful information, it is of paramount importance to use data analysis approaches exquisitely suited for multidimensional problems. Principal components analysis (PCA) is a powerful method routinely used to extract meaning from multidimensional data [1–4].

Other authors have demonstrated the value of PCA in analyzing microarray data [1,5–7]. We exploited PCA to describe the high-dimensional space of gene expression by means of a low number of derived variables constituting the unifying concepts of the studied data set. The classification of 60 tumor cell lines based on a cluster analysis performed on 1375 genes [8] was reproduced by only five independent principal components, thus providing evidence of the capture of all the information relevant for tumor discrimination. Subsequently, the components were assigned a functional meaning by inspecting the variables (genes) more correlated (highest component loadings) with them. This allowed the quantitative characterization of tumor cell lines according to their most prominent biological features at the gene expression level.

*Corresponding author. Fax: (39)-6-49902255.
E-mail address: alessandro.giuliani@iss.it (A. Giuliani).

Abbreviations: PCA, principal components analysis

2. Materials and methods

2.1. Biological data

The data sets used in this work have been generated by Ross et al. [8] and are publicly available at http://discover.nci.nih.gov/nature2000/. mRNA was extracted from 60 cancer cell lines and hybridized to cDNA microarrays including 9703 human cDNA clones representing approximately 8000 individual genes. The polished data set includes 1375 genes chosen as those showing the strongest variations among the cell lines [8].

We based our analysis on 1416 variables (named T-matrix at the web site) corresponding to the highest variance 1375 genes, with some repetitions. The data are expressed as the log ratio between the gene expression level of each cell line and a reference made by a mixture of 12 of the 60 cell lines [8].

2.2. Modeling methods

The original data set is a matrix having the cellular lines (statistical units) as rows and different gene expression levels (statistical variables) as columns. PCA projects this multidimensional space into a derived space spanned by new variables called principal components ordered in decreasing amount of explained variability. Thus, the first components will retain the maximal amount of correlated information (i.e. coordinated activity of genes) confining the uncorrelated portion of information to higher order components. This allows a strong compression of the original information by simply retaining the first extracted components and discarding the others [3,4,9].

2.3. Strategy of analysis

The data matrix with 60 statistical units and 1416 variables (corresponding to 1375 genes, some being represented more than once) was subjected to PCA. A direct visual screen test [10] on the eigenvalue distribution (distribution of explained variability) allowed us to select the bona fide signal components. The principal component scores were computed and the cell lines projected into the component space. A non-hierarchical clustering procedure (k-means) [3,11] was applied to the component space. The classification was optimized by comparing the amount of variance explained by each choice of k = number of classes with the corresponding variance explained by the clustering of a multivariate normal distribution and then choosing the value of k maximizing the difference between the explained variance relative to the actual classification and the one expected under multivariate normal assumption. The clustering so obtained was compared by means of a chi-square analysis with the clustering obtained by Ross et al. in order to assess the consistency between the original data and their low dimension representation. Computations were performed with the statistical software SAS v. 8.0 for personal computers.

3. Results

3.1. Data structure

The eigenvalue distribution on the components is reported in Fig. 1. The sudden drop of the eigenvalues with increasing component number suggests the possibility to select a small number of components modeling the gene expression difference among tumor cell lines [4]. We chose a five-component solution as bona fide signal, explaining 40% of the total var-

PII: S0014-5793(01)02973-8

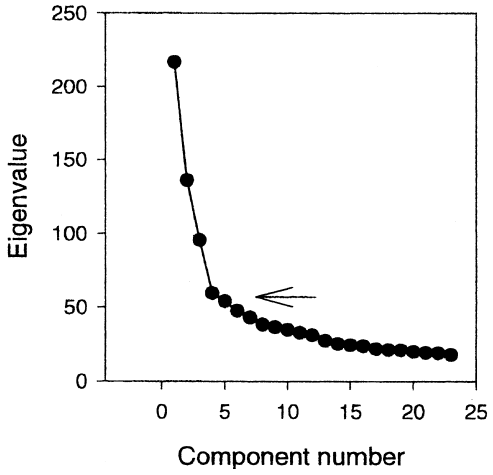

Fig. 1. Component number vs. eigenvalue. The arrow points at the five-component solution at the presumptive beginning of the 'noise floor'.

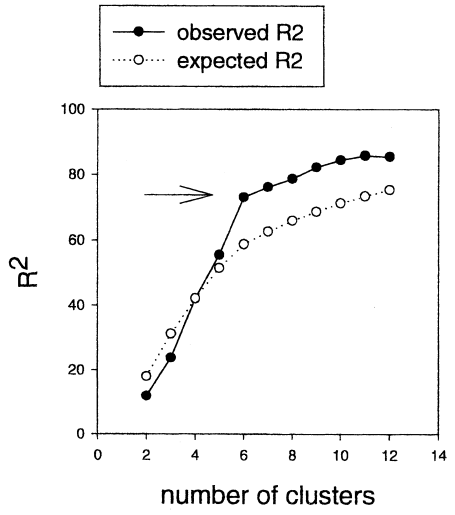

Fig. 2. Variation of R^2 (proportion of variance explained by the clustering) with increasing number of clusters for the data set analyzed and for a correspondent multivariate normal distribution. The six-cluster solution is the most parsimonious choice for which the two distributions diverge.

iability carried by the 1416 gene expression data (Table 1). Each cell line is now defined in a space spanned by five dimensions (component scores), independent of each other, representing distinct aspects of the differences between the cell lines.

3.2. Tumor classification

The *k*-means clustering procedure [11] applied to the distribution of the cell lines in the five-component space highlighted a six-cluster partition as the best one (largest difference with the expected R^2 under multivariate normal distribution), explaining 73% of the total variance of the component space (Fig. 2). The percentage of explained variance (R^2) is expressed as the ratio: between-clusters variance/total variance. The cluster composition is reported in Fig. 3.

Despite the deep diversity of the clustering algorithm used by Ross et al., a six-class structure is evident in their dendrogram representation too, as a transition in the aggregation distance. We assigned six class indicators to the dendrogram, according to the relative distances of aggregation, so that class 1 is closer to class 2 than class 5 because they are merged together by the hierarchical clustering at an early stage of aggregation (Fig. 3).

A comparison between the *k*-means clustering based on the five components and the hierarchical clustering based on the full-rank matrix showed a remarkably high concordance, with a chi-square value of 148.9 ($P < 0.0001$) and a contingency coefficient of 0.84. This implies that the five dimensional space spanned by the principal components retains all the essential variance discriminating among cell lines. It is worth noting that the concordance between the two classifications was verified a posteriori: the component-based classification was selected only for its structural properties (maximal explained variance) without any reference to the clustering obtained by Ross et al. Thus, the concordance between the two groupings is a strong proof of the accuracy of our representation. Moreover, the value of class indicator attached to the original dendrogram (ranging from 1 to 6 according to the relative similarities between classes) is strongly correlated ($r = 0.82$, $P < 0.0001$) with the first component scores. This implies that the first component alone allows a general, if approximate, metric scaling of the relative position of the tumor classes.

The faithful reconstruction of the original classification allows us to consider the extracted components as the 'collective' determinants of tumor line discrimination, thus justifying our effort to give components an interpretation on the basis of the genes most correlated with them. The fact that a component space explaining as little as 40% of total variability reproduces the classification based on the full-rank matrix points to the existence of relevant scale differences between gene aggregations. Very few major components collect the greater part of correlated gene activity that, by the very fact of being correlated, shapes the cell line classification. At the same time, a large part of the total variance of gene activity is dispersed along a plethora of minor components, presumably corresponding to activation networks comprising fewer genes. This interpretation is strengthened by the analysis of minor components' eigenvalue spectrum. When such a spectrum is

Table 1
Eigenvalue distribution

Component number	Eigenvalue	Variance (%)	Cumulative
1	216.70	15.30	15.3
2	135.85	9.59	24.9
3	95.44	6.74	31.6
4	59.42	4.20	35.8
5	53.80	3.80	39.6
6	47.35	3.34	43.0
7	42.82	3.02	46.0
8	38.04	2.69	48.7
9	36.40	2.57	51.3
10	34.49	2.44	53.7
11	32.40	2.29	56.0
12	30.90	2.18	58.2
13	27.18	1.92	60.1
14	25.12	1.77	61.9
15	24.20	1.71	63.6

The table reports the eigenvalue corresponding to each component (eigenvector) together with the relative percentage of variance explained and the cumulative variance of each subsequent global solution.

depurated by the major components that 'flatten' minor ones to the noise floor, it appears to contain signal-like information at least up to component 20 (70% of cumulative total variance explained, data not shown). Higher order components, on the contrary, appear to correspond to genuine noise of undefined origin.

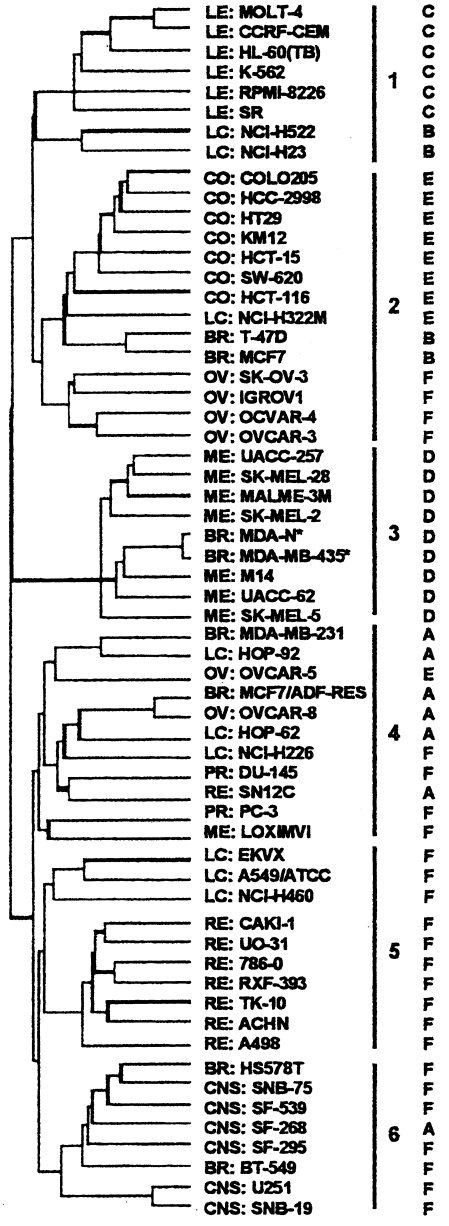

Fig. 3. The hierarchical clustering obtained on the 1416 gene matrix (modified from [7]) together with a class label indicating the relative distance of aggregation of the clusters (1 through 6). The aggregation distance corresponds to the value of mutual distance at which two clusters (or objects) are joined together to form a wider class. The class indicator, being ordered along the increasing aggregation distance at which the clusters progressively merge, gives a semiquantitative description of the relative position of the clusters and was demonstrated to scale with component 1 ($r = 0.82$, $P < 0.001$, see text). The column to the right reports the cluster membership according to the k-means procedure on the component space. Each cluster is indicated by a capital letter. BR: breast carcinoma, CNS: central nervous system tumor, CO: colon carcinoma, LC: non-small cell lung cancer, LE: leukemia, ME: melanoma, OV: ovarian carcinoma, PR: prostate carcinoma, RE: renal carcinoma.

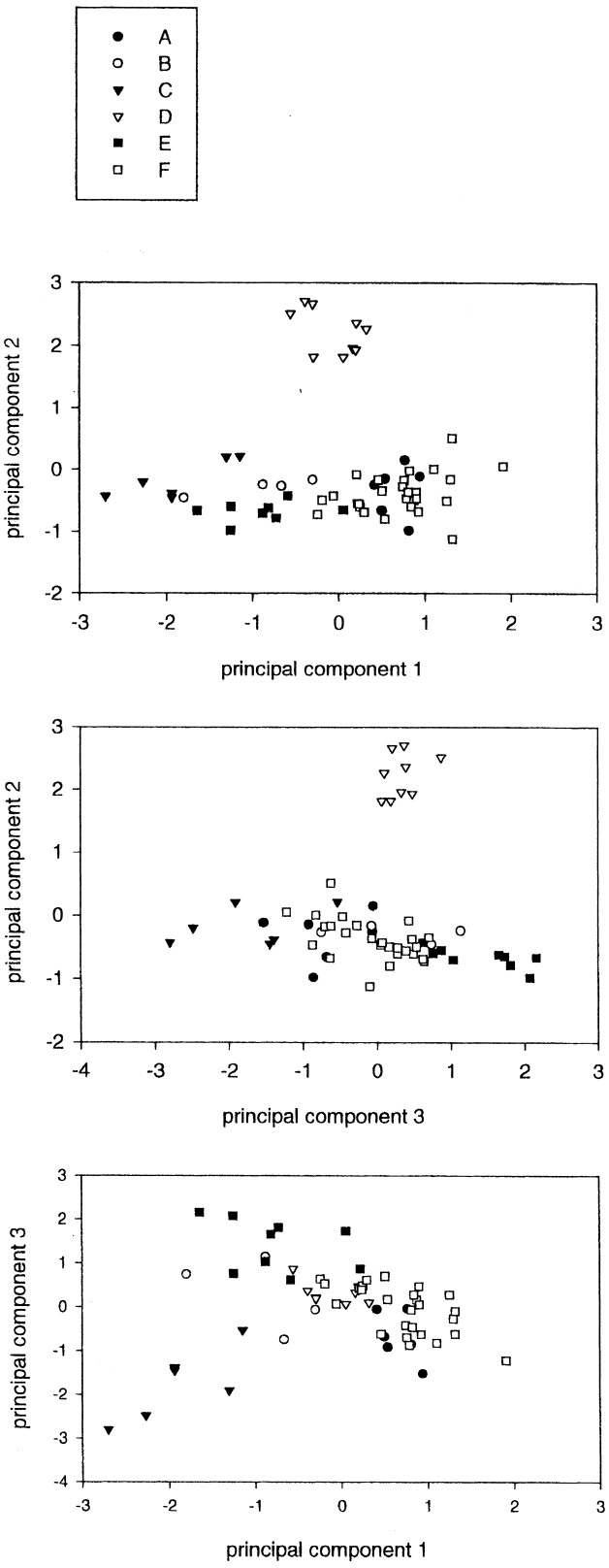

Fig. 4. The three possible planes having as axes the three main components. The coordinates of each element correspond to its component scores. The individual tumor cell lines are indicated in the plots by their class identifiers with reference to Fig. 3.

Table 2
Genes strongly correlating with components 1–5

Component no./gene name and [GenBank accession number]	Interpretation of component/gene function(s)
Component 1	cell motility and migration
Caveolin, caveolae protein, 22 kDa [AA047106]	membrane, signaling
Collagen, type IV, α1 [AA054624]	ECM
Laminin, γ1 [AA035021]	ECM
Caldesmon [AA026215]	contraction
Caveolin 2 [AA036724]	membrane
Quiescin (Q6) [W79188]	?
Transforming growth factor β [N29100]	growth factor
Integrin, β1 [AA044145]	integrin
Calponin 3, acidic [AA043227]	contraction
Component 2	membrane trafficking, endocytosis
hUNC18a alternatively spliced mRNA [AA053982]	?
Small GTP binding protein Rab7 [AA034507]	membrane trafficking, endocytosis
Highly similar to clathrin coat assembly protein AP19 [W30851]	membrane trafficking, endocytosis
β-Tubulin [AA009881]	multifunctional; involved in intracellular transport
Small GTP binding protein Rab7 [AA043679]	membrane trafficking, endocytosis
Phosphatidylinositol glycan biosynthesis, class F [AA042803]	membrane anchoring, endomembrane system, secretion
DNA-directed RNA polymerase II 14.4 kDa polypeptide [W79319]	RNA synthesis
Putative 32 kDa heart protein PHP32 [W88869]	?
Gal-β(1-3/1-4)GlcNAc α-2,3-sialyltransferase [W47425]	ECM catabolism, invasion, tissue/development-specific
Putative 32 kDa heart protein PHP32 [AA021369]	?
Component 3	epithelial cell proteins
E-MAP-115 [W90783]	microtubule-associated protein, mostly epithelial cells
E-MAP-115 [W01846]	microtubule-associated protein, mostly epithelial cells
P311 HUM (3.1) [AA047647]	unknown function, downregulated in transformation
Cytochrome *c* oxidase assembly protein COX11 [W37274]	transcription factor, hematopoiesis
Intestinal peptide-associated transporter HPT-1 [AA053188]	intestin, membrane, oligopeptide absorption
Cyclin-dependent kinase inhibitor 2A [AA055721]	p16 cell cycle inhibitor
Major GI tumor-associated protein GA733-2 precursor [AA055858]	homotypic cell adhesion molecule, most normal epithelia
Aldehyde dehydrogenase 10 (fatty ald. dehydrog.) [H63829]	metabolic enzyme
Moesin [AA043008]	links adhesion molecules to the cortical cytoskeleton
Epithelial-specific transcription factor ESE-1b [H27938]	epithelial-specific transcription factor
Component 4	?
Low affinity IgG Fc receptor II C precursor [R78402]	leukocytes, placenta, intestine, Ig transport
Antileukoproteinase E 1 precursor [AA026192]	secreted by epithelia; protective against leukocytes
Zn-15 related zinc finger protein [R15988]	transcription factor
Actin, α2, smooth muscle, aorta [AA040833]	smooth muscle
Myeloperoxidase [R05886]	leukocytes
P8 protein [AA024560]	?
Crystallin, αB [AA037471]	lens
Highly similar to POL protein (Moloney virus) [W87891]	polymerase?
CDK6 cyclin-dependent kinase 6 [H66259]	cell cycle regulator
Guanine nucleotide binding protein, α stimulating activity polypeptide 1 [3′: AA057701]	G protein stimulator
Component 5	?
ICH-2 protease precursor [AA029875]	caspase-4
Proteasome component C13 precursor [W74742]	akin to multicatalytic protease macropain
Cleavage signal-1 protein [AA045603]	protease
V-ets avian erythroblastosis virus E26 oncogene homolog 2 [R25353]	transcription factor, hematopoiesis
Transforming growth factor, β receptor II (70–80 kDa) [AA053517]	receptor
Glutaredoxin (thioltransferase) [AA033593]	REDOX
V-ets avian erythroblastosis virus E26 oncogene homolog 2 [W19687]	transcription factor, hematopoiesis
Elongation factor TU, mitochondrial precursor [AA011453]	protein synthesis
Erythropoietin receptor [H16867]	receptor
Cyclin-dependent kinase 6 [H66259]	cell cycle regulator
Inositol polyphosphate-1-phosphatase [H30231]	signaling, calcium chelation

The first 10 named genes most strongly correlating with components 1–5 are shown. Underlined functions: functions in agreement with the biological interpretation of the component. ?: unknown or poorly defined component/function.

3.3. Biological interpretation of the components

From the interpretive point of view, the major merit of PCA resides in the possibility to attach a biological meaning to the components. Interpretation is based on component loadings, that is, on the correlation coefficients between original variables (genes in this case) and the components [3]. The extracted dimensions represent the 'linearly independent systems' organizing the data [4]. The genes more correlated with the components are the ones more important to assign biological meanings. Each gene participates to all the extracted components, to different degrees, so fulfilling the notion that the same gene can participate to more than one activation network.

The first principal component is by far the major order

parameter structuring the observed variability (15% of explained variance, see Table 1). In order to assign a meaning to this component, we considered the first 100 genes possessing the highest absolute loadings, that is, the 100 genes showing the highest positive or negative correlation with component 1 and thus mostly involved in the corresponding coregulation circuit. Inspection of the component loadings pattern (web supplementary material) readily shows that most of these genes are known to be involved in the synthesis of the extracellular matrix (ECM), adhesion to ECM components, ECM-mediated signaling, cell contraction, and/or migration. Collectively, most genes highly loaded on component 1 are involved in all intra- and extracellular aspects of cell motility and migration. Table 2 shows the 10 previously characterized genes in this subset possessing the highest absolute loadings on the component. Although this result had not been anticipated, in retrospective it is not surprising that the first component isolates the leukemic from all other cancer cell lines (Fig. 4).

While component 1 represents the 'coarsest grain' classification structure of tumors, subsequent components focus on more refined details of the differential profile of cells. Table 2 shows the first 10 named genes with the highest absolute loadings on components 2 through 5 (see also web supplementary material). Components 2 and 3 are almost as strongly characterized as the first one. Component 2 comprises a high percentage of genes involved in membrane trafficking and endo/exocytosis, while component 3 is heavily loaded with genes selectively expressed in epithelial cells. Considering progressively higher order components, it becomes more and more difficult to distinguish a dominant key. Accordingly, we found it hard to assign a biological specificity to components 4 and 5. These difficulties are easily explained considering that each successive component accounts for a smaller amount of variability among tumor cell lines and, as a consequence, is less strongly characterized and more and more affected by both intrinsic and experimental noise. Nevertheless, the elucidation of the first three components illustrates the power of PCA in providing a biological meaning to microarray data. Fig. 4 shows the scattering of the cell lines in the space spanned by the first three components. Component 2 sharply isolates cluster D from all other cell lines. Similarly, both components 1 and 3 isolate cluster C, while component 3 essentially discriminates cluster E.

4. Discussion

In this paper, we apply PCA to an existing set of microarray data. Our aim was to use a classical strategy to make a general interpretive frame emerge from the data by a simple model-free analysis.

As already discussed, the concordance of PCA-based cell line classification with the published dendrogram distribution provides assurance of internal consistency. More important, as theoretically predicted, principal components correspond to biologically meaningful characteristics. Indeed, the dendrogram developed by Ross et al., reported in Fig. 3, is blind to the functional bases of the aggregation. On the contrary, our analysis highlights the biological determinants shaping the classification. By way of example, cluster C cell lines, comprising all the leukemias analyzed, are grouped at the negative extreme of component 1 (Fig. 4). Hematopoietic cells, among those considered, are least endowed with several functions identified by component 1, including production of ECM, adherence, and cell-autonomous motility and migration. These functions are actually exerted in a high degree by differentiated cells belonging to some hematopoietic lineages, e.g. macrophages. However, leukemia cell lines represent comparatively early stages of cell differentiation, not displaying the properties of their fully differentiated, normal counterparts. Even more strikingly, component 2, representing membrane trafficking and exo/endocytosis, distinctly segregates cluster D. This cluster includes all but one of the melanomas and two breast carcinomas. Melanocytes are almost exclusively devoted to producing pigment granules, which are then exported to neighboring cells through intense membrane trafficking and exocytosis. Two breast carcinomas associate with melanomas in both the PCA and the hierarchical clustering [8]. Ross et al. proposed that this may be due to the neuroendocrine features of some breast cancers [8]. Similarly, component 3, including many genes expressed in epithelial cells, identifies at the positive extreme cluster E (seven colon, one non-small cell lung, and one ovary cancer cell lines) and, at the negative end, again the leukemias. It should be stressed that the biological nature of the principal components molding our classification could not have been anticipated on theoretical grounds.

On the methodological ground, it is worth noting the possibility to attach a quantitative value to so far purely qualitative concepts such as 'adhesion behavior' or 'secretory character'. This value is computable for each tumor specimen on the basis of the expression of a relatively low number of the genes mostly correlated with the appropriate components. In any case it is important to stress that the picture sketched by PCA is based on empirical data. The microarray technology is still very young and further experience is needed to build confidence on any conceptualization based upon it.

Acknowledgements: We gratefully acknowledge the Weinstein group for making their data publicly available, without which this work would not have been possible. We thank R. Benigni for his continuing interest. This work has been partially supported by an AIRC grant to M.C.

References

[1] Alter, O., Brown, P.O. and Botstein, D. (2000) Proc. Natl. Acad. Sci. USA 97, 10101–10106.
[2] Beltrami, E. (1873) G. Mat. Battaglini 1, 98–106.
[3] Benigni, R. and Giuliani, A. (1994) Am. J. Physiol. 266, R1697–R1704.
[4] Broomhead, D.S. and King, G.P. (1986) Physica 20D, 217–236.
[5] Holter, N.S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J.R. and Fedoroff, N.V. (2000) Proc. Natl. Acad. Sci. USA 97, 8409–8414.
[6] Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L. and Somogyi, R. (1998) Proc. Natl. Acad. Sci. USA 95, 334–339.
[7] Raychaudhuri, S., Stuart, J.M. and Altman, R.B. (2000) Pac. Symp. Biocomput., 455–466.
[8] Ross, D.T. et al. (2000) Nat. Genet. 24, 227–235.
[9] Meloun, M., Capek, J., Miksik, P. and Brereton, R.G. (2000) Anal. Chim. Acta 423, 51–68.
[10] Cattell, R.B. (1966) Multiv. Behav. Res. 1, 245–276.
[11] Brazma, A. and Vilo, J. (2000) FEBS Lett. 480, 17–24.