INSTITUTO DE ESPAÑA
REAL ACADEMIA NACIONAL DE FARMACIA

MONOGRAFÍA XVII

LAS ÓMICAS:

GENÓMICA, PROTEÓMICA, CITÓMICA Y METABOLÓMICA. MODERNAS TECNOLOGÍAS PARA EL DESARROLLO DE FÁRMACOS

Editores:
María Cascales
María José Gómez Lechón
José Enrique O'Connor

Madrid, 2005

8. Human Cytome Project: A New Potential for Drug Discovery

GÜNTER K VALET

RESUMEN (Ed: Real Acad.Nacional de Farmacia, Madrid 2005, p207-228)

El descubrimiento de nuevas dianas de fármacos se enfoca cada vez más a las secuencias genómicas, a los transcritos RNA o a los análisis de proteínas, pero las expectativas para identificar un mayor número de nuevas estructuras para la investigación farmacéutica no han sido cumplidas durante la pasada década. En esta situación, parece importante adquirir información más detallada acerca de las condiciones moleculares en enfermos durante infecciones, enfermedades malignas, alergias, enfermedades reumáticas o diabetes, en lugar de guiarse preferentemente por el comportamiento de sistemas de cultivos celulares o modelos animales. Tales sistemas pueden no poseer el mismo *comportamiento regulador* que el organismo humano.

La información molecular puede ser obtenida eficientemente a partir de células humanas por citometría de flujo de imagen o química, llevando a cabo análisis simultáneos multiparamétricos de sistemas celulares *(citomas)* de donde emergen las enfermedades como consecuencia de procesos celulares moleculares aberrantes. La citómica, como análisis molecular multiparamétrico de células aisladas, en combinación con conocimientos exhaustivos bioinformáticos, proporciona la terapia relacionada con predicciones individualizadas en el curso de la enfermedad en pacientes, sobre la base de pautas o modelos bioparamétricos discriminatorios. Estos modelos son importantes para estudiar i) la ingeniería molecular de las vías metabólicas de la enfermedad, ii) el descubrimiento de nuevas dianas de fármacos, como también iii) la medicina clínica de cada día.

La propuesta de un proyecto de citoma humano presenta la posibilidad de avanzar sistemáticamente en este tema que presenta tan amplias aplicaciones.

SUMMARY

The discovery of new drug targets is increasingly conceptualized from genome sequences, RNA transcripts or proteome analysis but the expectations to identify an increased numbers of new target structures for pharmaceutical research have not fulfilled during the last decade. It seems in this situation important to acquire more detailed information on the molecular conditions in diseased *humans* like during infections, malignancies, allergies, rheumatic diseases or diabetes instead of being preferentially guided by the behavior of cell cultures or animal model systems. Such systems may ultimately not possess the same *regulatory behavior* as the human organism.

Molecular information can be efficiently collected from human single cells by flow, image or chemical cytometry performing simultaneous multiparametric analysis of cell systems (*cytomes*) from where diseases emerge as consequence of aberrant molecular cell processes. Cytomics, as multiparametric molecular single cell analysis in combination with exhaustive bioinformatic knowledge extraction provide therapy related *individualized disease course predictions* for patients on the basis of characteristically discriminatory bioparameter patterns. These patterns are of importance i) for the molecular reverse engineering of disease pathways, ii) for the discovery of new drug targets as well as iii) for everyday clinical medicine.

The proposal of a *human cytome project* has the potential to systematically advance this widely applicable approach.

1. INTRODUCTION

Drug development efforts have shifted from physiology and natural product driven strategies to target-based drug discovery (1) in recent years. Combinatorial chemistry, high throughput and high content screening (HTS/HCS), mRNA expression arrays, proteomics or systems biological approaches to name only some of the directions of research have, however, not fulfilled the expectations for the identification of substantial numbers of new targets and lead structures (2,3). It is presently suggested to remember earlier strategies as useful and necessary complements to target-based drug discovery (2-6).

The cause for the limited success is in all likelihood not due to the incapacity of the new technologies but amongst various reasons, like the multitude of

possibilities for combining the various approaches (3), primarily to the relative lack of knowledge on the relevant molecular processes in complex human diseases. Such knowledge represents a necessary precondition to mimic the in-vivo human disease situation as perfect as possible by disease-relevant models like for example molecular systems, single or mixed type cell cultures, ex-vivo tissue samples or research animal models.

It is the purpose of this contribution to show how the knowledge on molecular processes in complex human diseases can be substantially increased by *single cell* oriented multiparametric analysis of *aberrant molecular cell phenotypes* in disease associated cell systems of *individual patients*. This exploration strategy represents an essential feature in the context of a suggested *human cytome project*.

2. GROUP OR INDIVIDUALIZED ORIENTATION?

Individuals are heterogeneous for their genetic background as well as for their cumulated exposure to the many permanently existing external influences. Further heterogeneity exists in the specialized organs or tissues, each consisting of multiple cell types of significant internal heterogeneity, for example according to cell cycle, functional status, size and molecule content.

The observed multiparametric cellular and molecular heterogeneity prompts for a simplified evaluation of patients as groups to obtain average molecular information. Group oriented Kaplan-Meier statistics (7,8) are for example frequently used during therapy development to stratify patient groups according to their behavior according to various combinations of molecular or clinical parameters. Hierarchical clustering (9) as another example characterizes similarly expressed groups of gene products at the mRNA level. Although informative as a trend, there is the inherent problem that only a certain proportion of patients in this averaged heterogeneity will respond to a given therapy or show the average gene expression profile.

It is for example not possible to predict the therapy response of individual patients on the basis of group oriented evaluations. There is, however, significant clinical interest in the future disease course of individuals in stratified patient groups to optimize therapy and to minimize irreversible tissue damage by the disease process or by adverse drug reactions. This means for example to predict the therapeutic benefit on a person to person basis prior to the start of therapy and to individually select the most appropriate therapy amongst several possible therapeutic options in a given situation. It seems furthermore important to explore the molecular causes for the lack of response to therapy in an individualized/personalized fashion.

This may seem at first glance impossible considering the many existing heterogeneities. But individualized predictions can be obtained when investigating the molecular heterogeneity at the single cell level using a *differential bioparameter pattern* approach to characterize the disease induced aberrant molecular cell phenotypes (10). Molecular cell phenotype analysis can be performed by high-throughput image (11,12) or flow cytometric (13) analysis.

3. SINGLE CELL ANALYSIS

3.1. Why single cell analysis?

Considering the high number of possibilities for investigating gene activation or proteomics array patterns, metabolic pathways, metabolite panels, cell organelles, cells or tissues it may be asked why molecular single cell analysis is the preferred aim.

The reason for this is that cells represent the *elementary building units* of cell systems, organs and organisms and diseases are caused by molecular changes in cells and cell systems. Single cell analysis by molecular content as in flow cytometry (14) or by molecular morphology as in image cytometry (12) provides a maximum of compartmentalized molecular heterogeneity. This assures a maximum of resolution to discriminate for example the molecular changes in diseased or disease associated cells from those in non affected bystander cells.

Technical progress permits detailed single cell analysis by chemical cytometry using microfluidic chips (15,16) or capillary electrophoresis (17,18). Cell microgenomics expression profiles (19) as well as single cell proteomics (17) and metabolomics (15,16,18) become accessible in this way.

Single cell techniques overcome the problem of averaged cellular information from cell homogenates or extracts where it cannot be decided whether observed changes derive from all cells or only from a particular cell subpopulation.

The analysis of humoral body compartments like blood plasma or serum, urine or cerebrospinal fluid as an alternative provides secondary information on cell derived molecules. Metabolites from cellular disease processes may have been altered in the meantime or they may not become apparent in humoral compartments for lack of secretion or fast renal or biliary excretion. As a consequence, important information may not be measurable at the secondary compartment level although significant advances in metabolite profiling have been reached (20).

3.2. Representative single cell sampling

Although multiparametric molecular single cell analysis is desirable for conceptual reasons, one may object that the single cell approach will frequently not be feasible since not all cells of a given sample can be analyzed like for example in smears, biopsies or histological sections. This is, however, not necessary as long as a representative number of disease associated cells with aberrant molecular cell phenotype can be analyzed together with some non affected reference cells as shown by the subsequent examples.

Mechanical disaggregation of tissues at 0-4ºC for cell function analysis by flow cytometry destroys between 90-95% or more of all cells. Mechanical preparation is preferable since enzymatic cell preparation at 37ºC as an alternative generates a similar degree of destruction but, in addition, may alter the metabolic condition of relevant cells during the incubation time as well as enzymatically modify cell membrane parameters. Both procedures lead to an enrichment of epithelial and inflammatory cells since fibroblasts or smooth muscle cells are mostly destroyed during cell preparation. Despite this, more than 90% of lung and colorectal cancer patients are correctly identified from flow cytometrically determined molecular cell properties (21,22). This indicates that representative fractions of surviving cancer cells and normal epithelial or inflammatory cells as reference are sufficient for the identification of cancer patients although the cellular composition of the samples has changed and tissue architecture was lost during cell preparation.

The result is not surprising since diseases represent molecular changes in cells and cell systems. The analysis of diseased cells or of disease associated inflammatory and immune cells contains therefore by itself enough molecular information about the actual state (diagnosis) and the future development (prediction) of a disease. The correct sample classification is in these conditions to a sufficient degree independent of the original location of the analyzed cells in a tissue and equally of the representative presence of all initial types of organ cells. This applies in all likelihood equally to image analysis where it may for example not be necessary to contour all nucleated single cells in a given sample to obtain discriminatory information, especially in cases where cell boundaries are difficult to detect.

A further reservation concerning single cell analysis is that cell properties may be altered during preparation for analysis (23). Cell preparation between 0-4ºC for functional studies, deep-freezing of tissues or immediate cell fixation minimizes, however, such risks.

Valuable information is for example obtained from the functional analysis of oxidative status or oxidative burst in viable ex-vivo immune cells like lympho-,

mono- and granulocytes despite the high susceptibility of these cells to external influences. Disease associated cells can be measured in tissues but advantageously also frequently in the peripheral blood where high speed multiparameter flow cytometric single cell analysis is possible and provides individualized predictions or risk assessments for example in intensive care medicine (24,25).

As a conclusion, molecular alterations by cell preparation or staining steps cannot be generally entirely excluded. Keeping the cells, however, close to the in-vivo conditions during removal and short term storage or until fixation, the determination of clinically relevant molecular cell parameters for patient care and research purposes is definitely not impaired.

4. DIAGNOSIS AND PREDICTIVE MEDICINE

Diseases are typically diagnosed by clinicians from clinical symptoms, altered clinical chemistry parameters or by the pathologists evaluating the microscopic morphology in tissue sections or in cytological specimens.

Single cell image or flow-cytometric analysis extends the diagnostic knowledge level by the determination of differential bioparameter patterns (10). Alterations may be recognized at a stage where no morphological correlate is yet detectable. The multiparametric measurements of molecular cell phenotypes address the prediction of the therapy-related future disease course of individual patients as a clinically promising new feature of the *predictive medicine by cytomics* concept (10,25).

5. BIOINFORMATIC DATA EVALUATION

5.1. Cytomics

Multiparametric single cell analysis by flow or image cytometry can provide significant amounts of data that may seem difficult to analyze (23). *Differential data pattern classification* (26) provides the means to analyze multitudes of multiparametric data like ~0.5-1x10^5 data columns or more at a time. Data of various types like from flow and image analysis, chip arrays, clinical chemistry and clinical provenience can be merged and simultaneously processed like in the context of predictive medicine by cytomics (10). *Cytomics* is defined as multimolecular cytometric analysis of cell and cell system heterogeneity in combination with exhaustive bioinformatic knowledge extraction from all measured cells (27). Highly discriminatory bioparameter patterns can be obtained by algorithmic data analysis comprising typically between 10 and 30

parameters. The bioparameter patterns can be used for further scientific analysis. Single cell and single individuum oriented analysis provide in this way a maximum of discrimination since no averaging over heterogeneous entities occurs during data acquisition and bioinformatic evaluation. Once the individualized information has been obtained, group directed analysis can be performed at a maximum level of resolution.

5.2. Data pattern classification

Data pattern classification represents an *algorithmic data sieving procedure* for the individualized analysis of multiparametric data. The goal of the analysis concerns i) the determination of the most discriminatory bioparameter patterns between several patient categories being characterized in a multiparametric way and furthermore ii) the therapy related disease course predictions in individual patients with an accuracy >95% as belonging to one of the categories. The 95% level represents a first approximation. It seems possible to reach 99% or even higher levels when more multiparametric molecular single cell knowledge in disease will become available with time.

The classification process is automated and does not require personal interference once initiated. The principle of data pattern classification consists in the transformation of numerical data column values into the *triple matrix* characters (—)= decreased, (0)= unchanged and (+)= increased, followed by the iterative elimination of non-discriminatory data columns. Individual patients are classified according to the highest positional coincidence between their patient classification mask and any one of the disease classification masks (**table 1**). The classification principle assures high classification accuracy at high multiplicity of patient classification masks to account for the many combinatorial possibilities between genotypic and exposure influences on the selected discriminatory parameters. The patient classification masks can be further analyzed for similarities and dissimilarities in an effort to systematically investigate genotypic and exposure influences on patients with complex diseases such as infections, diabetes, rheumatic diseases, allergies, asthma, atherosclerosis, myocardial infarction, malignancies and others.

5.3. Classification of multiparametric patient data

The numerical values for the subsequent classification are localized in databases (**table 2**) (28,29). The databases classified here derive from an earlier

TABLE 1. *Classification Principle of the CLASSIF1 Algorithm*

A.) disease classification masks	
(schematic for 10 parameters)	**disease course prediction**
0000000000	stationary
++++++++++	improvement
————————	deterioration
B) patient classification masks	
(some examples)	
0+000-00+0	stationary
00+-00+-00	stationary
++000-0+0-	stationary
0-0—+000-	stationary
+-0000-0+0	stationary
++0-0++0+0	improvement
-0-0-+-0+-	deterioration
...	...

patient classification: highest positional coincidence between patient classification mask and any one of the disease classification masks

characteristic features of data pattern classification:

high accuracy at low risk of random coincidence between patient and disease classification mask. Probability is 0.0017% ($1/3^{10}$) for random coincidence with 10 parameter masks and 0.046% ($1/3^{7}$) for 7 parameter masks.

high multiplicity of patient classification masks at correct classification in case of partial positional coincidence like 7 out of 10 parameters:

$\frac{10! \times 2^3}{7! \times 3!}$ = 960 possible patient classification masks as potential result of genotype and exposure influences on the molecular cell phenotype or other para-meters.

study on the identification of angiographically defined myocardial infarction risk patients from surface antigen patterns of circulating and partially activated peripheral blood thrombocytes (28). The data column values in the databases can be either directly measured or calculated from multiparametric single cell flow or image cytometric measurements. Merged databases containing numerical results from various types of measurements can be jointly classified.

Prior to the classification, the available data are divided into a learning set of patients and into an unknown test (validation) set remaining inaccessible to the learning process. Typically, patients 1,5,10,15,20 etc of each classification category are a-priori assigned to the unknown test set. In this way it is assured that the data of this embedded test set are objectively selected and representative for the learning set since they have been collected under similar conditions.

An upper and lower percentile is now calculated for the reference category patients or samples of each available data column (**figure 1**). The reference cate-

TABLE 2. *CD62 Database parameters*

nr	parameters
1	%CD62 positive thrombocytes
2	thrombocyte mean FSC
3	thrombocyte mean SSC
4	thrombocyte mean ratio SSC/FSC
5	mean CD62 on all thrombocytes
6	mean CD62 on CD62 positive thrombocytes
7	mean CD62 on CD62 negative thrombocytes
8	mean CD62 surface density on all thrombocytes
9	mean CD62 surface density on CD62 positive thrombocytes
10	mean CD62 surface density on CD62 negative thrombocytes
11	% CD62 negative thrombocytes

Four databases containing a total of 4x11 parameters for CD62, CD63 and thrombospondin expression as well as for immunoglobulin gamma (IgG) binding to thrombocytes of an earlier study (28) were used to illustrate the principle of data pattern classification.

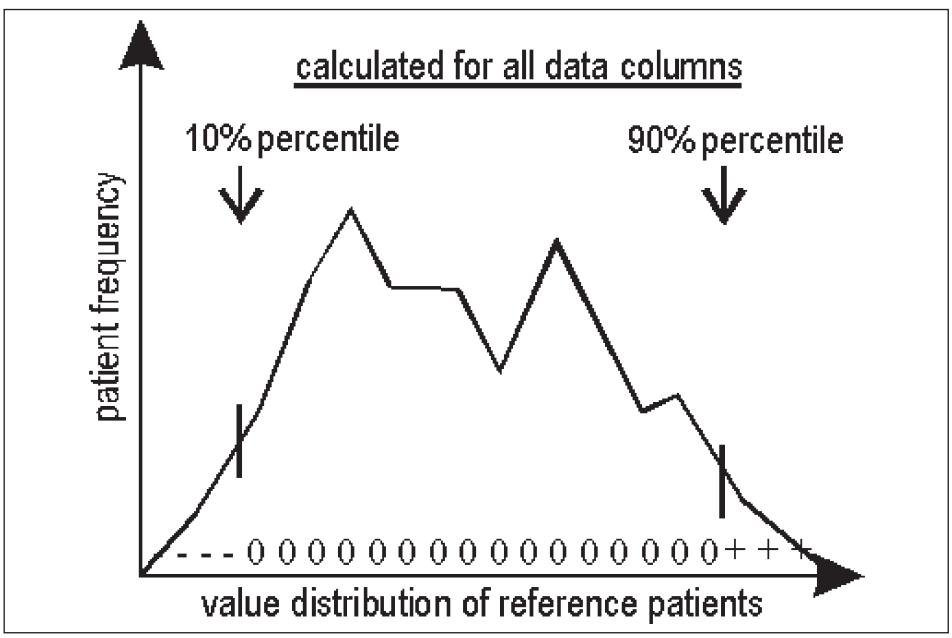

FIGURE 1. *Calculation of lower and upper percentile thresholds for the value distribution of each data column of the reference group of patients. In reality the percentile thresholds 1, 2, 5, 10, 15, 20, 25, 30, 70, 75, 80, 85, 90, 95, 98 and 99% are calculated. This permits to find the optimally discriminating percentile thresholds for the transformation of the numerical data column values into triple matrix characters during the subsequent iterative optimisation procedure.*

patient	numeric database columns 1	2	3	triple matrix database columns 1	2	3
1	6.64	2.38	40.74	-	-	+
2	6.68	3.16	41.68	-	0	+
3	2.24	2.54	39.54	-	-	0
4	9.10	2.27	40.96	0	-	+
5	9.72	2.10	42.12	0	-	+
6	4.49	2.76	38.85	-	0	0
7	6.64.	2.71	39.72	-	-	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.

FIGURE 2. *Transformation of numerical data column values into triple matrix characters (—), (0), (+). Two percentile thresholds are required for the transformation. The numerical values of each data column that is the values of the reference as well as of the other classification categories are transformed into the triple matrix characters (—) = diminished, for values below the lower threshold, (0)= unchanged, for values between the lower and upper threshold and (+) = increased, for values above the upper threshold.*

gory consists for example of healthy normal individuals but stationary or survivor patients may also serve as reference depending on the purpose of the classification. Percentiles like the 10% or 90% percentiles represent the threshold values at which 10% or 90% of the values of a reference category data column are reached. The calculation is performed in the same way for all data columns.

Percentile value calculations do not require assumptions regarding the mathematical distribution of data column values for reference or other category patients. This is advantageous since value distributions from flow or image analysis are frequently not distributed according to standard statistical functions.

The numerical values for the reference as well as for the non reference patients are subsequently transformed for each data column into the triple matrix characters (+) for values above the respective upper percentile, (—) for values below the respective lower percentile and (0) for values between both percentile thresholds. This yields a triple matrix replica (**figure 2**) of numerical databases as starting condition for the iterative selection of the discriminatory data

columns. Percentile based data classifications perform fast since the iterative calculations require only value comparisons that represent an elementary functionality of computer processing units.

Following the triple matrix transformation, the molecular status of each patient is represented by the *patient classification mask* consisting of a sequence of triple matrix characters (+), (—) and (0). The most frequent triple matrix character of each data column is placed for each classification category into a *disease classification mask*. The disease classification mask for reference patients contains typically (0) values because for example in the 10-90% percentile condition, (0) is the most frequently encountered triple matrix character. Its occurrence is 80% in each data column with additional 10% (+) and 10% (—) characters.

Data columns with (0) characters in the various classification categories are immediately removed from the classification process since they do not contain discriminatory information. This leaves 23 of the 44 initial data columns (**figure 3**).

In a first classification round, patients are classified according to the highest positional coincidence of their patient classification mask with any one of the disease classification masks. The classification results are entered into a classification (confusion) matrix with the present (diagnosis) or future (prediction) clinical truth typically on the ordinate while the classification results from the multiparametric data set is displayed on the abscissa (**table 3A**). Ideally, that is when the classification from the measured parameters reflects exactly the clinical situation, there will be 100% values on the diagonal of the classification matrix while all other boxes contain 0% values. This is not the case at the initial classification step. It is now important to determine the discriminatory information content of each data column to ultimately enrich a maximum of discriminatory capacity within the disease classification masks.

The discriminatory capacity of each data column is determined by its temporary removal from the classification process, followed by reclassification of the remaining data columns. Deterioration of the previous classification result indicates a positive discriminatory capacity of the removed data column while improvement reveals it as non-discriminatory. The removed data column is subsequently reinserted and the next data column is checked in the same way. Discriminatory data columns are internally labeled with a virtual green stamp while non-discriminatory columns obtain a virtual red stamp. Only data columns with green stamps are kept in the final disease classification masks. The «tracer» like temporary removal of individual database columns assures that the discriminatory capacity of the individual data columns is determined against all re-

Nr.	classification category	category abbreviation	mask coinc. factor	disease classification masks IgG CD62 CD63 thrombosp.
1	normal	N	1.00	0 0 0 **0** 0 0 0 0 **0** 0 **0** 0 0 0 0 0 0 0 0 **0** 0 0 **0** 0 0 0 0
				- - +- +- - - - - +- +- ++- +- - - +-
2	infarction risk	R	1.00	++- **+** - ++ ++ +- +- +- - **+** - + **+** - +

Nr.	clinical classification (T6LERN)		CLASSIF1 classification	mask coinc. factor	patient classification masks
1	#102	N	N	0.91	+0 0 0 0 0 0 0 **0** 0 **0** 0 0 +0 +0 - - 0 0 0
2	#103	N	N	0.65	0 0 + + + 0 - + + + - 0 - 0 0 0 0 0 + 0 0 0 0
3	#104	N	N	0.61	+0 0 0 0 0 0 0 + + - 0 - 0 - - 0 - + 0 0 0 0
4	#106	N	N	0.83	0 + 0 0 0 + 0 0 0 **0** + 0 + + 0 + + + - 0 0 0 0
5	#107	N	N	0.87	0 0 0 0 0 0 0 0 0 **0** 0 0 + 0 + 0 0 0 0 0 0 + 0 0
6	#108	N	N	1.00	0 0 0 0 0 0 0 0 0 **0** 0 0 + 0 0 0 0 0 0 0 - + 0
7	#109	N	N	0.96	0 - 0 0 0 - 0 - **0** - 0 - 0 - 0 - 0 0 0 0 - 0 - -
8	#111	N	N	0.91	0 0 + **0** + 0 - 0 0 0 0 0 0 0 + **0** + 0 + 0 0 +
9	#112	N	N	0.91	0 0 0 0 0 0 0 0 0 **0** 0 0 0 0 0 0 0 0 0 0 0 + 0 +
10	#113	N	N	1.00	0 0 + 0 0 0 0 0 0 **0** 0 0 0 0 0 0 0 0 0 0 0 0 0 0
14	#137	R	R	0.87	++- +- ++ +++ 0 +- +- 0 + 0 + ++- +
15	#138	R	R	0.74	+0 - +- 0 + 0 **+** + - 0 - +0 - 0 - + 0 **+** - +
16	#139	R	R	0.91	0 + - **+** - ++ + **+** + - 0 - + - - **+** - + + **+** - +
17	#141	R	R	0.78	++- +- ++ +++ 0 + 0 + 0 + + ++ + + 0 +
18	#142	R	R	0.87	++- +- ++ +++ - +- + 0 + + ++ + + 0 +
19	#143	R	R	0.78	0 0 - **+** - ++ 0 **+** + - 0 - + - - **+** - + 0 **+** - +
20	#144	R	R	0.78	0 + - **+** - ++ + **+** + - 0 - + - - 0 - + 0 + 0 +
21	#146	R	R	0.74	- - - + 0 0 + 0 **+** + - - - + - - 0 - + - + - +
22	#147	R	R	0.65	+0 - +- ++ 0 + + 0 - - + - - 0 - + 0 + - +
23	#148	R	R	0.65	0 0 - **+** - ++ 0 **+** + 0 0 - + - 0 **+** - + 0 + 0 +

FIGURE 3. *Triple matrix database following removal of non informative data columns showing (0) triple matrix characters for normal individuals and infarction risk patients. Twenty-three of the original 44 database columns remain after this step. They come from each of the four initial databases (IgG, CD62, CD63, thrombospondin). The finally selected 5 most discriminatory data columns are indicated in boldface characters.*

maining data columns. The iterative procedure identifies usually most of the data columns as being non-discriminatory for the classification.

Only 5 of the initial 44 data columns improve (**table 3B**) the initial classification result (**table 3A**) significantly. The proof that this classification result is not due to the classification of random statistical aberrations in multiparametric

TABLE 3. *Risk Assessment from Peripheral Blood Thrombocyte Surface Antigens*

A. Before Optimization				**B. After Optimization**			**C. Unknown Test Set**		
		CLASSIF1 risk status (%)							
clinical risk status	pat (n)	normal	infarction risk	pat (n)	normal	infarction risk	pat (n)	normal	infarction risk
normal	13	**100.0**	0.0	13	**100.0**	0.0	4	**100.0**	0.0
inf.risk	74*	41.9	**58.1**	77	0.0	**100.0**	20	0.0	**100.0**
neg/pos pred.val	—	**29.5**	**100.0**	—	**100.0**	**100.0**	—	**100.0**	**100.0**

4x11=44 data columns CD62,CD63,thrombospondin and bound IgG classified with 10+15+20% optimized percentile thresholds against the angiographically defined risk status
*) 3 patients show the transitional classification normal/infarction risk. They are not included into the classification result before optimization.
Bold numbers on the classification diagonals indicate specificity and sensitivity with negative and positive predictive values in the bottom line.

TABLE 4. *Selected Classification Parameters*

nr	*classification parameters*	*assay*	*N*	*IR*
1	IgG on IgG positive thrombocytes	FSC/SSC/IgG	0-	+
2	CD62 on CD62 positive thrombocytes	FSC/SSC/CD62	0-	+
3	CD63 mean surface density on CD63 positive thrombocytes	FSC/SSC/CD63	0-	+
4	thrombospondin (TRSP) on TRSP positive thrombocytes	FSC/SSC/TRSP	0-	+
5	TRSP mean surface density on TRSAP positive thrombocytes	FSC/SSC/TRSP	0-	+

FSC= forward light scatter, SSC = sideward light scatter, N=normal, IR= infarction risk

data sets, is provided by the unknown embedded data set (**table 3C**). It is i) correctly classified by the selected 5-parameter data pattern (**table 4**) and ii) there are statistically significant parameter differences (**table 5**) between normal individuals and the angiographically defined risk patients for myocardial infarction.

5.4. Standardized classifiers

Data pattern classifiers are inherently standardized on the data column means of the reference groups. Reference groups can be constituted by normal individuals but also by patients with stationary disease or by survivor patients.

TABLE 5. *Numerical Characteristics of Selected Thrombocyte Parameters*

parameter	*normal means±SEM (AU*, n=13)*	*infarction risk means±SEM (AU*, n=77)*	*statistical significance (Student P %)*	*selected lower & upper percentiles (AU*)*	*standardized percentiles (perc/mean)*
IgG on IgGpos	13.6±0.1	16.7±0.2	< 0.001	12.9/13.9(10/90%)	0.95/1.02
CD62 on CD62pos	14.3±0.4	19.0±0.2	< 0.001	13.4/14.7(15/85%)	0.94/1.03
CD62 SURFD on CD63pos	1.49±0.03	1.77±0.03	< 0.001	1.34/1.60(10-90%)	0.90/1.07
TRSP on TRSPpos	14.1±0.2	20.3±0.2	< 0.001	13.4/14.8(10-90%)	0.95/1.05
TRSP SURFD on TRSPpos	1.55±0.94	2.46±0.06	< 0.001	1.33/1.71(10-90%)	0.86/1.10

* AU = arbitrary units, IgG=immunoglobulin gamma, SURFD=mean surface density

The identity of reference groups between institutions is verified by classifying the respective reference data sets against each other. In case, reference groups from different institutions are undistinguishable by data pattern classification, they can be considered identical for classification purposes. This provides the potential to merge data between institutions with the aim to enlarge data sets and to establish public databases.

When differences during the comparison of reference groups are detected, they may be due to methodological, compositional like for sex or age, ethnic or exposure differences of the included individuals.

5.5. No artifactual misclassification of random number data sets

It is important to investigate the susceptibility of data pattern classification to the artifactual classification of random statistical aberrations in multiparametric data sets.

A random number data set of 133 data columns was classified for this purpose. The data set consisted of 40 data records. Every second record was a-priori assigned to either category#1 or category#2 yielding 20 data records of either type. The coefficients of variation (CV=100*standard deviation/mean) of the data columns varied between 0.97-25.7%. Data records 1,5,10,15 and 20 of each category were a-priori assigned to the unknown embedded test set, leaving a learning set of 15 category#1 and 15 category#2 records. The classification of all data columns (**table 6A**) indicated no difference between category#1 and category#2 records. The automated iterative removal of 130 non informative data columns resulted in 3 data columns providing a certain degree of discrimination between category#1 and category#2 records (**table 6B**). The classification is, however, not informative because (i) the classification of the unknown test

TABLE 6. *Classification of Random Number Dataset*

A. Before Optimization				**B. After Optimization**			**C. Unknown Test Set**		
		CLASSIF1 category (%)							
assigned category	rec (n)	cat#1	cat#2	rec (n)	cat#1	cat#2	rec (n)	cat#1	cat#2
cat#1	15	**100.0**	0.0	15	**100.0**	0.0	5	**33.7**	66.7
cat#2	15	100.0	**0.0**	15	60.0	**40.0**	5	80.0	**20.0**
neg/pos pred.val	—	**50.0**	**0.0**	—	**62.5**	**100.0**	—	**33.5**	**25.0**

Category#1 records are indistinguishable from category#2 records when all data columns are classified together (**A**), while a certain degree of discrimination is obtained after removal of 130 non discriminatory data columns (**B**). The classification quality for the unknown test set (**C**) is insufficient. It is based on differences of the 3 selected data columns. The differences are statistically not significant between the two classification categories. This indicates the robustness of the CLASSIF1 algorithm against the artifactual interpretation of random statistical aberrations as real differences. The data records were classified using the optimal percentile thresholds 15% and 85%. Bold numbers are as in **tab. 3**.

set is insufficient (**table 6C**) and (ii) the selected data columns are not statistically different between the two categories (col29: 73.2±6.6/72.8±11.1, col77: 53.3±8.0/58.9±11.3, col133: 24.9±3.7/24.9±5.8, means±standard deviation).

The classification result indicates robustness of the triple matrix data pattern classifier against the artifactual interpretation of random statistical aberrations as real differences.

6. HUMAN CYTOME PROJECT AND DRUG DISCOVERY

6.1. Bioparameter patterns

The determination of individually predictive bioparameter patterns by data pattern classification opens a way to the i) reverse engineering of molecular causes inducing favorable or unfavorable bioparameter patterns and ii) to the discovery of new drug targets.

The initial selection of the investigated multiparameter parameters in a given situation is *hypothesis driven* and varies from investigator to investigator. The cytomics data analysis, in contrast, is *hypothesis free* because it effectuates the exhaustive differential knowledge extraction from all analyzed cells or other data. This opens the way for the detection of so far unknown molecular me-

Nr.	classification category	category abbreviation	mask coinc. factor	disease classification masks
1	normal	N	1.00	0 0 0 0 0
				- - - - -
2	infarction risk	R	1.00	+ + + + +

Nr.	clinical classification (TH5LEARN)		CLASSIF1 classification	mask coinc. factor	patient classification masks (.=no value)
1	#102	N	N	1.00	0 0 0 0 0 ←
2	#103	N	N	0.60	+ + 0 0 0
3	#104	N	N	0.80	0 + 0 0 0
4	#106	N	N	1.00	0 0 0 0 0
5	#107	N	N	0.80	0 0 0 + 0
6	#108	N	N	1.00	0 0 0 - 0 ←
7	#109	N	N	1.00	0 0 - 0 - ←
8	#111	N	N	0.60	0 0 + 0 +
9	#112	N	N	0.60	0 0 0 + +
10	#113	N	N	1.00	0 0 0 0 0
14	#137	R	R	1.00	+ + + + +
15	#138	R	R	0.80	+ + 0 + +
16	#139	R	R	1.00	+ + + + +
17	#141	R	R	1.00	+ + + + +
18	#142	R	R	1.00	+ + + + +
19	#143	R	R	0.80	+ + 0 + +
20	#144	R	R	0.80	+ + 0 + +
21	#146	R	R	0.80	+ + - + +
22	#147	R	R	0.80	+ + 0 + +
23	#148	R	R	0.80	+ + 0 + +

FIGURE 4. *Reclassification of the learning set. The first ten patients of either classification category are displayed. All patients are correctly classified according to the highest positional coincidence of the patient classification mask with any one of the disease classification masks. The mask coincidence factor indicates the degree of positional coincidence between the patient classification mask and the selected disease classification mask. In some instances the mask coincidence factor is 1.00 (arrows) despite the fact that not all characters of the disease classification mask are (0) for reference patients. Since infarction risk show increases (+) for the five selected data columns, unchanged (0) as well as decreased (-) values indicate normal individuals. Patients #101, 105, 110, 136, 140, 145 are missing in the table because they had been a-priori assigned to the unknown test (validation) set.*

chanisms or of the action of particular regulatory networks in heterogeneous cell systems remaining typically hidden to the global analysis of tissues or biopsies and as a consequence also to hypothesis driven approaches.

In the course of cytomics studies, new drug targets may be recognized by multiparametric single cell oriented tissue (11) or high throughput (30) cell analysis. The analysis may also provide the relevant information for a particular in-vitro composition of disease-relevant complex human cell systems (31) mimicking the molecular in-vivo behavior of defined patient groups in disease.

6.2. Systems biology

The suggested molecular reverse engineering of the observed discriminatory bioparameter patterns prompts for the subsequent mathematic modeling of molecular disease pathways as one of the preconditions for the rational understanding of complex disease processes in man.

Systems biology aims at the understanding of the integral functionality of single cells, organs or organisms by molecular analysis and mathematical modeling using controlled perturbations for the differential screening of molecular changes induced at different levels (23,32-35). This task is significantly more complex for organisms than for single cell systems like for example bacteria or yeast. The prediction of the reactivity of biological systems under predefined conditions, represents an important further goal of systems biology. Predictions should be indeed feasible, once a majority of possible reactivities in a biological system have been explored by differential perturbations.

6.3. High complexity by bottom-up

Numerous perturbations may, however, be required to describe major molecular pathways by hypothesis driven *bottom-up* analysis from the genome level via the proteome, the metabolome (27), the organelle compartments up to the level of cells, cell systems, organs and organisms (36). This is especially true when considering the high number of potential interrelations amongst 30.000 to 40.000 human genes as well as the cellular complexity of tissues and organs.

It is unknown, how much molecular knowledge is necessary to predict for example the disease susceptibility or future disease course in individual patients. It may also be impossible to obtain sufficient information by systematic bottom-

up forward engineering of mammalian organisms within reasonable time intervals, seen the complexity of the entire system.

As an alternative, cell cultures, or diseased as well as genetically modified animals, may be used as model systems for human disease. Significant concerns as to the ultimate validity of the conclusions from such model systems for the human in-vivo situation, especially in disease or in the search for new drug targets, have, however, been raised (37). Cell or animal model systems may frequently not permit to generate valid mathematical models of the human organism by systems biology especially if there is insufficient information on the heterogeneous single cell complexity of diseased human tissues.

6.4. Top-down in the proposed human cytome project

In this situation, a *top-down* research strategy simplifies the task. *Nature induced* bioparameter perturbations or differentials such as between diseased versus healthy, progressive versus stationary disease or survivor versus non-survivor patients can be directly analyzed instead of generating hypothesis driven systematic perturbations in model systems. Individualized disease course prediction for patients are possible in this way (10) without prerequisite to a-priori fully understand the entire molecular network of disease associated cell system changes. Discriminatory bioparameter patterns are obtained by multiparameter data analysis as described above. These patterns can be further investigated by a molecular reverse engineering strategy (38) to understand disease inducing molecular pathways and to find new drug targets (**figure 5**). It is advantageous for this *biomedical cell systems biology* concept (39, 40) that many in all likelihood suitable data sets are already available as starting material from current or past clinical studies where patients are routinely followed for diagnostic or therapeutic purposes. Further multiparameter data are continuously being produced in the clinical and medical environment. The concept of the human cytome project fits furthermore within the top-down approach of the physiome project (41).

Initial clinical areas suitable for a human cytome project concern the individually predictive molecular disease course characterization of leukemia/lymphomas, cancers, rheumatoid diseases, allergies and infections, while stem cell differentiation, cell cycle regulation, cell proteomics and cell organelle functionality represent initial areas of interest for basic research.

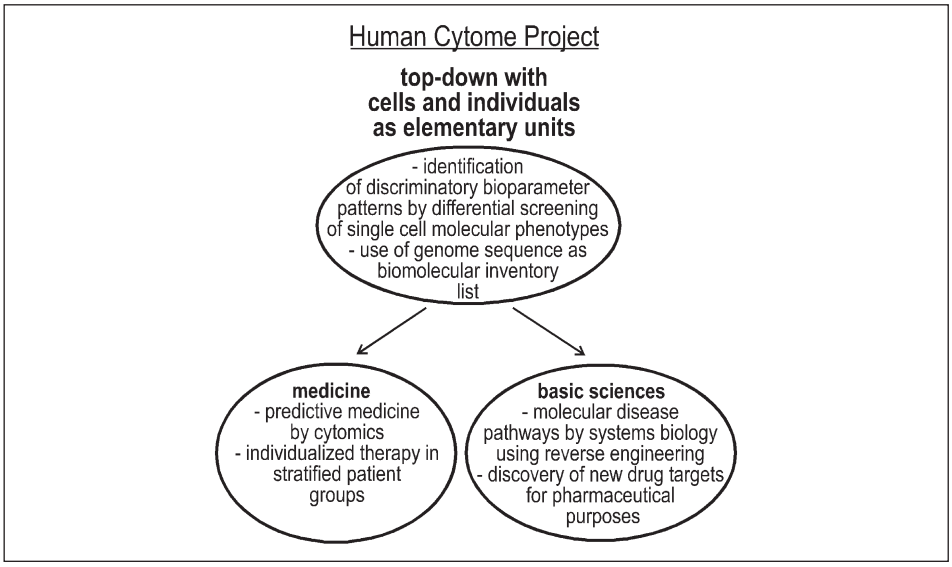

FIGURE 5. *Concept of the human cytome project*

7. CONCLUSIONS

The value of the *single cell*, *single individual* analysis concept resides in its clinical significance for the individual patient as well as in the bioparameter patterns being of interest for i) molecular reverse engineering of molecular disease pathways by systems biology and for ii) the discovery of new drug targets.

The research concept for the proposed *human cytome project* is deductive for the initially selected analytical parameters but inductive during the bioinformatic data evaluation phase. It is therefore capable of accessing new knowledge being *unreachable* by traditional hypothesis formulation. Most of the initially available multiparametric molecular information is typically eliminated as irrelevant during the algorithmic data-sieving phase, leaving typically a bioparameter pattern of between 10 to 30 discriminatory parameters for further analysis. This provides comparatively simple starting points for retrograde analysis and mathematical modeling efforts by systems biology.

There is also the possibility of consecutive *cyclic knowledge enrichment* steps. The information of the preceding sieving step is used to conceptualize and perform a new round of multiparametric molecular measurements at the single cell level to provide the information input for the next data sieving step.

ACKNOWLEDGEMENT

The generation of the random number data set by Dr.W.Meyer and PD Dr.G.Haroske (Institute of Pathology, University of Dresden, Germany) is gratefully acknowledged.

LITERATURE REFERENCES

1. Sams-Dodd, F. (2005) Target-based drug discoveries: is something wrong? *DDT* **10**, 139-147.
2. Butcher, E.C., Berg, E.L., Kunkel, E.J. (2004) Systems biology in drug discovery. *Nature Biotechnol* **22**, 1253-1259.
3. Schneider M. (2004) A rational approach to maximize success rate in target discovery. *Arch Pharm Pharm Med Chem* **337**, 625-633.
4. Lansbury, T.L. (2004) Back to the future: the 'old-fashioned' way to new medications for neurodegeneration. *Nature Rev Neuroscience* **5**, S51-S57.
5. Piggott, A.M., Karuso, P. (2004) Quality, not quantity: the role of natural products and chemical proteomics in modern drug discovery. *Comb Chem High Throughput Screening* **7**, 607-630.
6. Butler, M.S. (2004) The role of natural product chemistry in drug discovery. *J Nat Prod* **67**, 2141-2153.
7. Rosenwald, A., Wright, G., Chan, W.C., Connors, J.M., Campo, E., Fisher, R.I., Gascoyne, R.D., Müller-Hermelink, H.K., Smeland, E.B., Staudt, L.M. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *NEJM* **346**, 1937-1947.
8. Repp, R., Schaekel, U., Helm, G., Thiede, C., Soucek, S., Pascheberg, U., Wandt, H., Aulitzky, W., Bodenstein, H., Sonnen, R., Link, H., Ehninger, G., Gramatzki, M., AML-SHG study group. (2003) Immunophenotyping as an independent factor for risk stratification in AML. *Cytometry* **53B**, 11-19.
9. Eisen, M., Spellman, P., Brown, P. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *PNAS* **95**, 14863-14868.
10. Valet, G. (2002) Predictive medicine by cytomics: potential and challenges. *JBRHA* **16**, 164-167.
11. Beesley, J., Roush, C., Baker, L. (2004) High-throughput molecular pathology in human tissues as a method for driving drug discovery. *DDT* **9**, 182-189.
12. Huang, K., Murphy, R.F. (2004) From quantitative microscopy to automated image understanding, *J. Biomed. Optics* **9**, 893-912.

13. Edwards, B.S., Oprea, T., Prossnitz, E.R., Sklar, L.A. (2004) Flow cytometry for high-throughput, high content screening. *Curr Opin Chem Biol* **8**, 392-298.

14. Szaniszlo, P., Wang, N., Sinha, M., Reece, L.M., van Hook, J., Luxon, B.A., Leary, J.F. (2004) Getting the right cells to the array: Gene expression microarray analysis of cell mixtures and sorted cells. *Cytometry* **59A**, 191-202.

15. Wu, H., Wheeler, A., Zare, R.N. (2004) Chemical cytometry on a picoliter-scale integrated microfluidic chip. *PNAS* **101**, 12809-12813.

16. Palkova, Z., Vachova, L-, Valer, M-, Preckel, T. (2004) Single-cell analysis of yeast, mammalian cells , and fungal spores with a microfluidic pressure-driven chip-based system. *Cytometry* **59A**, 246-253.

17. Dovichi, J., Hu, S. (2003) Chemical cytometry. *Curr Opin Chem Biol* **7**, 603-608.

18. Arkhipov, S.N., Berezovski, M., Jitkova, J., Krylov, S.N. (2005) Chemical cytometry for monitoring metabolism of a Ras-mimicking substrate in single cells. *Cytometry* **63A**, 41-47.

19 Taylor, T.B., Nambiar, P.R., Raja, R., Cheung, E., Rosenberg, D.W., Anderegg, B. (2004) Microgenomics: identification of new expression profiles via small and single-cell sample analysis. *Cytometry* **59A**, 254-261.

20. Fernie, A.R., Trethewey, R.N., Krotzky A.J., Willmitzer L. (2004) *Nature Rev Mol Cell Biol* **5**, 1-7.

21. Valet, G., Rüssmann, L-, Wirsching, R. (1984) Automated flow-cytometric identification of colo-rectal tumor cells by simultaneous DNA, CEA-antibody and cell volume measurements. *J Clin Chem Clin Biochem* **49**, 83-90.

22. Liewald. F., Demmel, N., Wirsching, R., Kahle, H-, Valet, G. (1990) Intracellular pH, esterase activity and DNA measurements of human lung carcinomas by flow-cytometry. *Cytometry* **11**, 341-348.

23. Hood, L., Heath, J.R., Phelps, M.E., Lin, B. (2004) Systems biology and new technologies enable predictive and preventative medicine. *Science* **306,** 640-643.

24. Valet, G., Roth, G., Kellermann, W. (1998) Risk assessment for intensive care patients by automated classification of flow cytometri data. In: Phagocyte function. Eds: Robinson, J.P., Babcock, G.F., Wiley-Liss Inc, New York, p 289-306.

25. Valet, G., Kahle, H., Otto, F., Bräutigam, E., Kestens, L. (2001) Prediction and precise diagnosis of diseases by data pattern analysis.in multiparameter flow cytometry. Melanoma, juvenile asthma and human immunodeficiency virus infection. *Meth Cell Biol* **64**, 487-508.

26. Valet, G.K., Hoeffkes, H.G. (2004) Data pattern analysis for the individualised pretherapeutic identification of high-risk diffuse large B-cell lymphoma (DLBCL) patients by cytomics. *Cytometry* **59A**, 232-236.

27. Chitty M. (2005) -Omes and -omics glossary. http://www.genomicglossaries.com/content/omes.asp

28. Valet, G., Valet, M., Tschöpe, D., Gabriel, H., Rothe, G., Kellermann, W., Kahle, H. (1993) White cell and thrombocyte disorders: Standardized, self-learning flow cytometric list mode classification with the CLASSIF1 program system. *Ann NY Acad Sci* **677**, 183-191.

29. Valet, G.K., Höffkes, H.G. (1997) Automated classification of patients with chronic lymphatic leukemia and immunocytoma from flow cytometric three-color immunophenotypes. *Cytometry* **30**, 275-288.

30. Perlman, Z.E., Slack, M.D., Feng, Y., Mitchison, T.J., Wu. L.F., Altschuler, S.J. (2004) Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194-1198.

31. Butcher, E.C., Berg, E.I., Kunkel, E.J. (2004) Systems biology in drug discovery. *Nature Biotechnol* **22**, 1253-1259.

32. Kitano, H. (2002) Systems biology: a brief overview. *Science* **295**, 1662-1664.

33. Ideker, T., Galitski, T., Hood, L. (2001). A new approach to decoding life: systems biology. *Ann Rev Genomics Hum Genet* **2**, 343-372.

34. Hood, L. (2003) Systems biology: integrating technology, biology and computation. *Mechanisms of Aging and Development* **124**, 9-16.

35. Weston, A.D., Hood, L. (2004) Systems biology, proteomics, and the future of health care: towards predictive, preventative, and personalized medicine. *J Prot Res* **3**, 179-196.

36. Collins, F.S., Green, E.D., Guttmacher, A.E., Guyer, M.S. (2003) A vision for the future of genomics research. *Nature* **422**, 835-847.

37. Horrobin, D.F. (2003) Modern biomedical research: an internally self-consistent universe with little contact with medical reality? *Nature Rev Drug Discovery* **2**, 161-164.

38. Valet, G. (2005) Human Cytome Project, Cytomics and Systems Biology: The Incentive for New Horizons in Cytometry. *Cytometry* **64A**, 1-2.

39. Valet, G. (2005) Cytomics: an entry to biomedical cell systems biology. *Cytometry* **63A**, 67-68.

40. Valet, G., Murphy, R.F., Robinson, J.P., Tarnok, A., Kriete, A. (2005) Cytomics - from cell states to predictive medicine. In: Computational Systems Biology. Eds: Kriete, A., Eils, R., Elsevier, Amsterdam, in press.

41. Hunter, P.J., Borg, T.K. (2003) Integration from protein to organs: the physiome project. *Nature Rev Mol Cell Biol* **4**, 237-243.