Journal of Biological Regulators and Homeostatic Agents

Predictive medicine by cytomics: potential and challenges

G. VALET

Max-Planck-Institut für Biochemie, Martinsried, Germany

ABSTRACT: Predictive medicine by cytomics represents a new concept which provides disease course predictions for individual patients. The predictive information is derived from the molecular cell phenotypes as they are determined by patient's genotype and exposure to external or internal influences. The predictions are dynamic because they are therapy dependent. They may provide a therapeutic lead time for preventive therapy or for the diminution of disease associated irreversible tissue damage.

Multiparametric data from cytometry, multiple clinical chemistry assays, chip or bead arrays serve as input for an algorithmic data sieving procedure (http://www.biochem.mpg.de/valet/classif1.html). Data sieving enriches the discriminatory parameters in form of standardized data masks for predictive or diagnostic disease classification in the individual patient (http://www.biochem.mpg.de/valet/cellclas.html). Besides predictive and diagnostic utility, the data patterns can be used in a top-down approach for the development of scientific hypotheses on disease inducing mechanisms in complex inflammatory, infectious, allergic, malignant or degenerative diseases. (J Biol Regul Homeost Agents 2002; 16: 164-7)

KEY WORDS: *Predictive medicine, Clinical cytomics, Cytome, Medical bioinformatics, Data pattern classification, Data sieving, Data mining*

Received: April 15, 2002
Revised: June 15, 2002
Accepted: June 15, 2002

Scope

Extrapolation into the future development of diseases is usually expressed as prognosis. Prognosis is, however, a statistical operator and of little value for the individual patient.

Predictive medicine, in contrast, tries to either individually predict disease occurrence or to individually predict further disease course in already diseased patients. Prenatal or preimplantation diagnostics (1) for comparatively rare genetic disorders concern the former while individualized disease course prediction of widespread postnatal inflammatory, infectious, allergic, malignant or degenerative diseases address the latter. Individualized disease course predictions may facilitate therapy e.g. in decision making for surgery (2) or in future personalized medicine by the pharmacogenomics concept (3). Individualized disease course predictions would significantly facilitate individualized therapy of disease in everyday medicine.

Concept

Predictive medicine by cytomics represents a new concept for individualized disease course predictions in patients and has the potential to overcome present limitations (http://www.biochem.mpg.de/valet/cellclas.html).

Considering that diseases are caused by molecular changes in heterogeneous cellular systems or organs (cytomes), information on disease course prediction and disease diagnosis should be collectable at the cellular level. Predictive medicine by cytomics consists therefore in the cytometric analysis of disease associated molecular alterations in cytomes. Cell analysis is intimately linked with medical bioinformatics to obtain predictive parameter patterns from multiparametric data spaces. Cytomics access a maximum of information on the apparent molecular cell phenotype which is the result of a patient's genotype and cumulated exposure to external and internal influences. Expusure is a significant factor because genotipically susceptible individuals without exposure may remain disease free (e.g. allergies, asthma rheumatoid arthritis).

The multiparameter data space from molecular cell phenotype analysis can be processed by cluster analysis, self organizing neural networks, or expert systems (4-10). The various approaches have so far not led to individualized disease course predictions while high speed algorithmic “data sieving” (11, http://www.biochem.mpg.de/valet/classif1.html) seems

0393-974X/164-04 $02.00

more promising.

Algorithmic data sieving for the most discriminatory data columms briefly works as follows: Data column values for reference patients as well as for diseased patients of the learning set are transformed into triple matrix characters: 0 (unchanged), + (increased) and - (decreased). The transformation depends on the location of the values between the lower and upper, above the upper or below the lower percentile threshold of the data values in the reference patient group. The percentile thresholds are automatically optimized. The transformation step is performed for all data columns and a patient classification mask for each patient is obtained in this way. Subsequently, the most frequent triple matrix character of each data column is determined from the patient classification masks of each group of diseased or reference patients. The most frequent triple matrix character is then entered into the disease classification mask of each disease group of patients. An unknown patient is classified according to the highest positional coincidence of the individual patient classification mask with any one of the disease classification masks.

The discriminatory potential is optimized by an iterative temporary removal of single data columns, followed by reclassification of the learning set patients. A confusion matrix is used as indicator of the classification result. The confusion matrix for predictive metaanalysis typically indicates the known future disease course on the ordinate and the predicted disease course on the abscissa. Ideally the diagonal values of the confusion matrix as well as the predictive values are 100% and non diagonal values are 0%. This is typically not the case when all available data columns are considered at the beginning (Fig. 1A). A decrease of the diagonal sum upon sequential temporary removal of single data columns indicates informative parameters while non informative parameters are indicated by an increase of the diagonal sum. The classification result for each data column is retained, the data column is reinserted and the next data column is temporarily removed. At the end of the iteration process, the most discriminatory data columns are obtained. Only data columns improving the classification result either alone or paired with a second data column in all possible combinations are retained in the optimized disease classification masks (Fig. 2). This improves the classification result significantly (Fig. 1B). The robustness of the optimized classifier is verified by the prospective classification of unknown test set patients (Fig. 1C).

Risk Assessment for Myocardial Infarction

A. learning set before

clinical outcome	pat. (n)	CLASSIF1 classification (%) norm.	inf. risk
normal	13	100.0	0.0
inf.risk	77	64.9	42.9
neg/pos predval	–	20.6	100.0

→ 44 data columns IgG,CD62,CD63,TRSP

B. after iteration

clinical outcome	pat. (n)	CLASSIF1 classification (%) norm.	inf. risk
normal	13	100.0	0.0
inf.risk	77	0.0	100.0
neg/pos predval	–	100.0	100.0

→ 5 data columns (Fig.2)

C. unknown test set patients

clinical outcome	pat. (n)	CLASSIF1 classification (%) norm.	inf. risk
normal	4	100.0	0.0
inf.risk	20	0.0	100.0
neg/pos predval	–	100.0	100.0

10-15% percentiles, THSLEARN database

Fig. 1 - Classification of thrombocyte activation antigens CD62, CD63 and thrombospondin as well of surface IgG binding in angiographically defined myocardial infarction risk patients. The thrombocyte database established in (11) was used to demonstrate the classification characteristics of the CLASSIF1 algorithm (11, 16, 19, 20). The iterative optimization reduces the number of data columns from 44 columns as extracted from four forward/sideward light scatter/single colour fluorescence measurements to 5 by the elimination of 39 non informative data columns.

Optimized Disease Classification Mask for Myocardial Infarction Risk Assessment

#	classification parameters	assay	N	R
1	IgG on IgG positive thrombocytes	FSC/SSC/IgG	0	+
2	CD62 on CD62positive thrombocytes	FSC/SSC/CD62	0	+
2	CD63 surf.dens. on CD63 positive thrombocytes	FSC/SSC/CD63	0	+
4	thrombosp. on thrombosp.pos. thrombocytes	FSC/SSC/TRSP	0	+
5	thrombosp. surf.dens.on thrombosp.pos. thrombocytes	FSC/SSC/TRSP	0	+

Fig. 2 - Optimized disease classification mask for the identification of myocardial infarction risk patients from thrombocyte surface antigens. It is of interest that the increased antigen expression (+) of all four antigen assays carries the discriminatory information while the routinely used % antigen positive fraction of thrombocytes is less informative and eliminated during the optimization process. Abbreviations: FSC/SSC = forward/sideward light scatter, TRSP = thrombospondin, N = normal, R = myocardial infarction risk patient.

Consequences

Data sieving represents an inductive approach for the exhaustive information extraction from large multiparametric data spaces in view of predictive or diagnostic goals. Hypothesis driven data collection (deductive) is followed by data sieving (inductive) and hypothesis driven interpretation (deductive) of the resulting predictive data patterns (Tab. I). The data patterns may serve as input for repetitive rounds of deductive, inductive and deductive refinement steps. This should further improve the individualized disease course predictions in medical or clinical cytomics. The approach differs from hypothesis driven data mining which may inadvertantly leave relevant information unconsidered.

Besides immediate clinical utility, the predictive data patterns may be useful for the detection of unknown disease inducing mechanisms. Such mechanisms could be hidden to direct deductive hypothesis because no knowledge on their existence within the high molecular complexity of cytomes may be available a priori. It seems promising to use the readily accessible and clinically relevant predictive data patterns for a top-down molecular reverse engineering process. Since even a high degree of knowledge on the totatility of cytome based biomolecules by bottom-up strategies does not include knowledge on their mutual spatial arrangement and functionality in intact cells.

The situation is, by analogy, similar to the practical impossibility of assembling a modern car from its disassembled parts in the absence of blue prints by deductive hypothesis alone. A chance for success exists, however, by the progressive disassembly of a fully assembled car with the aim of generating blue prints for reassembly.

Data classifications are presently considered predictive for individual patients at predictive values >95% for each classified disease category of the learning set. They are prognostic at values <95% (http://www.biochem.mpg.de/valet/cytomics.html). The effort will be to elevate this level to >99% through the search for more efficiently discriminating molecular data patterns.

Potential

Individualized disease course predictions by cytomics are dynamic predictions due to their therapy dependance. Patients with prediction for “disease aggravation” may convert under therapy within some time into “no complication” patients such as in intensive care medicine. The early prediction of disease aggravation or amelioration provides in principle a lead time for therapy onset and offset.

The clinical potential of the approach is provided by the possibility of increased overall therapeutic efficiency by individualized therapy. This may help to preventively reduce irreversible tissue damage by preventive therapy as well as to avoid unwanted therapeutic side effects. The cytomics approach as evidence based medicine at the cellular level, may also diminish the number or length of clinical therapy trials by utilizing the predictive data pattern changes in patient cytomes as indicators of therapeutic effects instead of the entire patient. It seems especially important in case of fast progressing or life threatening diseases or of significant therapeutic side effects.

The clinical potential of predictive medicine by cytomics has so far been demonstrated for the prediction of sepsis and shock in intensive care patients (12), postoperative edema and effusion (POEE) in children’s cardiac surgery (13, 14), the overtraining syndrome risk in competition cyclists (11), complications in bone marrow stem cell transplantation (15) and of life threatening conditions in sepsis (16) or colorectal cancer (17). Further applications of the concept concern the risk assessment for myocardial infarction (11), diagnostic classification of leukemias, lymphomas (18, 19) and juvenile asthma (20) as well as disease staging in human immunodeficiency virus (HIV) infection (20). Multiplex bead assays (21, 22) are potentially a very valuable complement to cell biochemical parameters. They provide multiparametric information on the molecular environment of cellular systems. This seems of especially high importance for the understanding of complex regulatory processes in diseases of the immune or hematopoietic systems.

TABLE I

Predictive Medicine by Cytomics

Analysis Concept

http://www.biochem.mpg.de/valet/cellclas.html

1. **deductive:**	hypothesis driven data collection
2. **inductive:**	algorithmic data sieving (data mining) for discrimination
3. **deductive:**	hypothesis driven interpretation of discriminatory data patterns

The scientific potential of predictive medicine by cytomics consists in the above mentioned possibility for the successful retrograde analysis of molecular disease processes from disease associated cytomes. It may in this way be possible to gain indirectly access to causative mechanism of complex disease processes.

Challenges

The evident challenges in advancing to the patient level is up to the concerted effort of scientists, clinicians and industry.

ACKNOWLEDGEMENTS

The present concept was significantly advanced by collaborations within the European Working Group for Clinical Cell Analysis (EWGCCA, http:// www.ewgcca.org) under EU contract BMH4-CT97-2611 as well as by earlier funding from the Deutsche Krebshilfe (Mildred-Scheel-Stiftung), Germany.

Reprint requests to:
Prof. Dr. Günter K. Valet
Max-Planck-Institut für Biochemie
Arbeitsgruppe Zellbiochemie
Am Klopferspitz 18a
D-82152 Martinsried
Germany
valet@biochem.mpg.de

References

1. Miny P, Tercanli S, Holzgreve W. Developments in laboratory techniques for prenatal diagnosis. Curr Opin Obstet Gynecol 2002; 14: 161-8.
2. Taylor ChA, Draney MT, Ku JP, et al. Predictive medicine: Computational techniques in therapeutic decision-making. Comp Aided Surgery 1999; 4: 231-47.
3. Manincelli L, Cronin M, Sadée. Pharmacogenomics: The promise of personalized medicine. AAPS PharmSci 2002; 2, (1) article 4 (http://www.aapspharmsci.org/).
4. Weinstein JN, Scherf U, Lee JK, et al. The bioinformatics of microarray gene expression profiling. Cytometry 2002; 47: 46-9.
5. Boddy L, Wilkins MF, Morris CW. Pattern recognition in flow cytometry. Cytometry 2001; 44: 195-209.
6. Wilkins MF, Hardy SA, Boddy L, Morris CW. Comparison of five clustering algorithms to classify phytoplankton from flow cytometry data. Cytometry 2001; 44: 210-7.
7. Cualing HD. Automated analysis in flow cytometry. Cytometry 2000; 42: 110-3.
8. Decaestecker Ch, Remmelink M, Salmon I, et al. Methodological aspects of using decision trees to characterize Leiomyomatous tumors. Cytometry 1996; 24: 83-92.
9. Frankel DS, Frankel SL, Binder BJ, Vogt RF. Application of neural networks to flow cytometry data analysis and real-time cell classification. Cytometry 1996; 23: 290-302.
10. Diamond LW. Nguyen DT, Andreeff M, Maiese RL, Braylan RC. A knowledge-based system for the interpretation of flow cytometric data in leukemia and lymphomas. Cytometry 1994; 17: 266-73.
11. Valet G, Valet M, Tschöpe D, et al. White cell and thrombocyte disorders: Standardized, self-learning flow cytometric list mode data classification with the CLASSIF1 program system. Ann NY Acad Sci 1993; 677: 233-51.
12. Rothe G, Kellermann W, Valet G. Flow cytometric parameters of neutrophil function as early indicators of sepsis- or trauma-related pulmonary or cardiovascular organ failure. J Lab Clin Invest 1990; 115: 52-61.
13. Tarnok A, Bocsi J, Pipek M, et al. Preoperative prediction of postoperative edema and effusion in pediatric cardiac surgery by altered antigen expression patterns on granulocytes and monocytes. Cytometry 2001; 46: 247-53.
14. Tarnok A, Pipek M, Valet G, Richter J, Hambsch J, Schneider P. Children with post-surgical capillary leak syndrome can be distinguished by antigen expression on neutrophils and monocytes. In: Cohn GE, Owickivc, eds. Systems and technologies for clinical diagnostics and drug discovery II, SPIE Progress in Biomedical Optics 1999, Bellingham, WA, USA, Vol. 3603, 61-71.
15. Valet G, Cornelissen J, Lamers C, Gratama JW. Predictive medicine by cytomics: Outcome prediction in bone marrow stem cell transplantation (SCT) (Abstract). Cytometry 2002; (Suppl 11): S54.
16. Valet GK, Roth G, Kellermann W. Risk assessment for intensive care patients by automated classification of flow cytometric data. In: Robinson JP, Babcock GF, eds. Phagocyte function. New York: Liss Inc 1998; 289-306.
17. Van Driel BEM, Valet GK, Lyon H, Hansen U, Song JY, Van Noorden CJF. Prognostic estimation of survival of colorectal cancer patients with the quantitative histochemical assay of G6PDH activity and the multiparameter classification program CLASSIF1. Cytometry 1999; 38: 176-83.
18. Bartsch R, Arland M, Lange St, Kahl Ch, Valet G, Höffkes HG. Lymphoma discrimination by computerized triple matrix analysis of list mode data from three-color flow cytometric immunophenotypes of bone marrow aspirates Cytometry 2000; 41: 9-18.
19. Valet G, Höffkes HG. Automated classification of patients with chronic lymphatic leukemia and immunocytoma from flow cytometric three colour immunophenotypes. Cytometry 1997; 30: 275-88.
20. Valet G, Kahle H, Otto F, Bräutigam E, Kestens L. Prediction and precise diagnosis of diseases by data pattern analysis in multiparameter flow cytometry: Melanoma, juvenile asthma and human immuno-deficiency virus infection. Methods in Cell Biology 2001; 64: 487-508.
21. Chen R, Lowe L, Wilson JD, et al. Simultaneous quantification of six human cytokines in a single sample using microparticle based flow cytometric technology. Clin Chem 1999; 45: 1693-4.
22. Lund-Johansen F, Davies K, Bishop J, de Waal Malefyt R. Flow cytometric analysis of immunopreciptates: High-throughput analysis of protein phosphorylation and protein-protein interactions. Cytometry 2000; 39: 250-9.