Predictive Medicine by Cytomics and the Challenges of a Human Cytome Project

a report by

Professor Günter K Valet

Head, Cell Biochemistry Group, Max-Planck-Institut für Biochemie

Professor Günter K Valet has been head of the Cell Biochemistry Group at the Max-Planck-Institut für Biochemie in Martinsried, Germany, since 1989. Prior to this, he headed the Mildred-Scheel-Laboratory for Cancer Cell Research from 1981 to 1989, at the same institution, as a joint venture between the Mildred-Scheel-Foundation and the Max-Planck-Society for the Advancement of Sciences. Professor Valet was elected president of the European Society for Analytical Cellular Pathology (ESACP) between 1994 and 1999 and of the Deutsche Gesellschaft für Zytometrie (DGfZ) between 1990 and 1994. He served as councillor for the International Society for Analytical Cytology (ISAC) between 1991 and 1995 and also between 1981 and 1985, has authored more than 170 scientific publications, was a member of several editorial boards of international scientific journals and has organised and co-organised international scientific congresses and numerous courses and workshops in flow cytometry. He was appointed Professor of Experimental Medicine at the Ludwig-Maximilian University of Munich in 1981, following habilitation in 1974 and postdoctoral work from 1972 to 1973 at the Scripps Clinic and Research Foundation in La Jolla, California. Professor Valet graduated from the medical faculty of the Ludwig-Maximilian University of Munich with an MD degree in 1968, after studying in Munich, Freiburg and Montpellier.

The increased knowledge of the biomolecular capacity of organisms as derived from the genome sequence has raised hopes for the widespread identification of new molecular targets for drug development. The sequence information does, however, not currently explain the observed structural and functional multi-level biocomplexity of cells and cellular systems (cytomes) (see *Figure 1a*). Multitudes of hypotheses are required to experimentally cover the many possibilities of highly redundant molecular pathway systems, without being *a priori* certain of having focused on relevant molecular hotspots, new drug targets or disease-associated metabolic pathways.

Considering the difficulties – still present after more than 30 years of intensive research – of attempting to accurately predict three-dimensional (3-D) protein structures[1] from known sequences containing the 20 most common amino acids, systematically understanding the combinatorial complexity of the metabolic pathways of 20,000–30,000 gene products by classical deductive hypothesis formulation and experimental verification within reasonable time intervals seems an unattainable goal.

Concerning the economic side, the increased investments in the drug development sector during the last 10 years have resulted in less clinically applicable new drugs than during the preceding 10-year period.[2] It is likely that the underestimated factor of biocomplexity has significantly contributed to the observed development.

Research Strategy

This raises the issue of more efficient research strategies. The prevailing research strategy concerning functional genomics aims at understanding the details of more and more molecular pathways and mechanisms as a prerequisite for the definition of promising new drug targets. Due to the high system complexity, this is an open-ended effort with an uncertain outcome. As an alternative, one can take advantage of the genomic sequence information as an inventory of the biomolecular capacity of organisms (see *Figure 1b*) and use single-cell analysis to collect a maximum of multi-parametric molecular information to reflect the full heterogeneity of cellular systems (organs).

The cell-by-cell analysis overcomes the problem of averaged results from cell and tissue homogenates where molecular changes may be wrongly interpreted, because they can be either uniform or confined to specific cell populations, while changes in low-frequency cell populations may be lost by dilution.

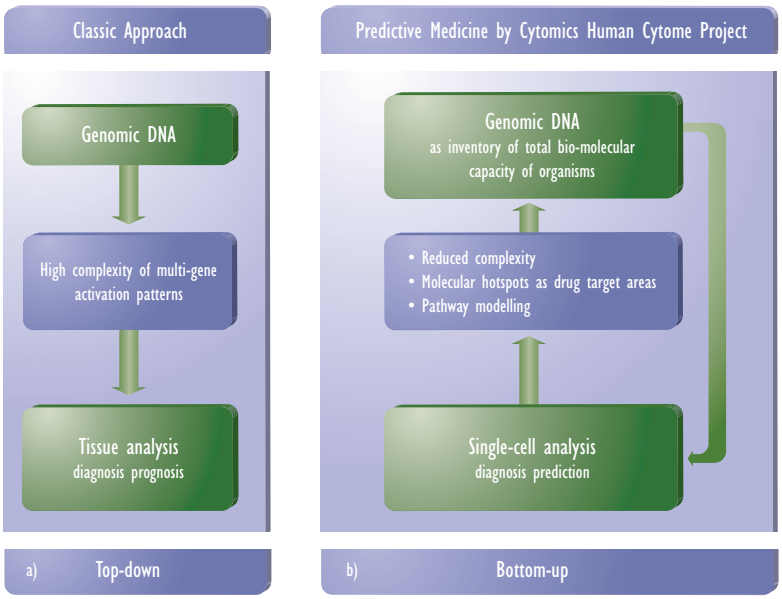
The collected information can be analysed exhaustively against the future disease course of individual patients, or against disease diagnosis. Besides their immediate use for predictive and diagnostic purposes in medicine, the resulting data patterns describe relevant molecular hotspots that can be further analysed in a second step. A framework of disease-related molecular alterations can thus be established at different levels of the genomic information, for example at the messenger RNA (mRNA), proteome or molecular cell phenotype levels, as they result from genotype and exposure to internal or external influences.

Since the identified hotspots represent only the disease-related fraction of the entire biocomplexity, the requirements for the subsequent detailed analysis of molecular pathways for drug target identification are significantly lower than those for facing the entire complexity of the genomic information.

This approach does not seem to be open-ended, since parameter numbers in the order of up to 105 or more can be currently processed for the enrichment

1. *P Aloy, A Stark, C Hadley and R R Russell, "Predictions Without Templates: New Folds, Secondary Structure and Contacts in CASP5", Proteins, 53 (2003), pp. 436–456.*
2. *F Kermani, "The Future of Biopharmaceutical Research and Development",* Business Briefing: Future Drug Discoveries, *2002, pp. 16–18.*

Figure 1: Classical, molecular pathway and mechanism oriented functional genome analysis (A) as contrasted to the data pattern-oriented molecular reverse engineering concept for predictive medicine by cytomics and the proposed human cytome project (B)



of discriminatory data patterns, each containing typically between 10 and 30 parameters .[3,4]

The essential characteristics of this research concept[5–7] are the use of molecular data patterns instead of metabolic pathways, the cell-to-patient approach and the standardised description of cell and disease states in a relational molecular framework. Following this line of thought, one may ask how difficult it is to perform multi-parameter single-cell analysis on patient cells and extract the necessary knowledge on molecular hotspots.

Single-cell Analysis

Flow and image cytometry systems represent routine instrumentation in hospitals and research institutions. Flow cytometers are preferentially used for the characterisation of single-cell suspensions from peripheral blood, effusions, urine or spinal fluid, while image analysis is preferable for the characterisation of cytologic, bioptic or histologic tissue section specimens. Several cell properties are typically stained by specific fluorescence-labelled reporter molecules. Fluorescence and biophysical signals like forward and sideways light-scattering or electrical cell-sizing signals are simultaneously collected at typical rates between 100 to more than 10,000 cells per second.

Each cell is only measured once in a flow cytometer, with the options to image cells in flow[8] or to preparatively separate them according to a predetermined parameter profile by a cell-sorting module. Image analysis systems, in contrast, can relocate cells, which is essential for several subsequent cell stainings. Cells may be initially stained for cell functions such as intracellular pH, transmembrane potentials or Ca2+ levels. They are then fixed to remove the functional stains and restained for specific extra- or intracellular constituents such as antigens, lipids or carbohydrates. After destaining, specific nucleic acids may be stained. Multi-spectral imaging,[8,9] as well as serial optical or histological sections, permits 3-D reconstructions of the molecular morphology of the cell membrane, nucleus, organelle and cytoplasmic compartments, including the parametrisation of 3-D shapes. Such data is useful for the standardised analysis of the proximity and interaction patterns of intra-cellular structures such as nuclei and organelles, as well as for different cell types within the tissue architecture (for more details, see Valet, Leary and Tárnok (2004)[7]).

Information Collection and Knowledge Extraction

Traditional visual and quantitative evaluations of 2-D or 3-D cytometric histograms such as those in flow cytometry only collect a very limited amount of the available information and one is never certain whether the most relevant information has been captured. Furthermore, experience has shown that quality-controlled consensus strategies for multi-parameter data evaluation are not easy to develop and

3. *G Valet and H G Hoeffkes, "Data Pattern Analysis for the Individualised Pre-therapeutic Identification of High-risk Diffuse Large B-cell Lymphoma (DLBCL) Patients by Cytomics",* Cytometry, *xx, 2004; in press.*
4. *G Valet, R Repp, H Link, G Ehninger, M Gramatzki and the SHG-AML Study Group, "Pre-therapeutic Identification of High-risk Acute Myeloid Leukemia (AML) Patients from Immunophenotypic, Cytogenetic and Clinical Parameters",* Cytometry, *53B (2003), pp. 4–10.*
5. *G Valet, "Predictive Medicine by Cytomics: Potential and Challenges",* JBRH, *16 (2002), pp. 164–167.*
6. *G K Valet and A Tárnok, "Cytomics in Predictive Medicine",* Cytometry, *53B (2003), pp. 1–3.*
7. *G K Valet, J Leary and A Tárnok, "Cytomics – New Technologies: Towards a Human Cytome Project",* Cytometry, *xx (2004), in press.*
8. *T C George, B E Hall, C A Zimmermann, D A Basiji, W E Ortyn, M J Seo, D H Lynch and P J Morissey, "Distinguishing Modes of Cell Death Using the ImageStream Multi-spectral Imaging Flow Cytometer",* Cytometry, *xx (2004), in press.*
9. *R C Ecker, R de Martin, G E Steiner and J A Schmid, "Application of Spectral Imaging Microscopy in Cytomics and Fluorescence Energy Transfer (FRET) Analysis",* Cytometry, *xx (2004), in press.*

there is little pre-existing interpretation knowledge on very complex multi-parameter data spaces. Essential information may be lost due to a lack of awareness.

The use of automated, self-adjusting evaluation strategies for the systematic collection of the entire information content of all measured cells for subsequent knowledge extraction is therefore important. Biophysical parameters such as light scattering or cell volume signals, as well as the presence of DNA or certain antigens such as CD45 on leukocytes, can be used as gating parameters to assure information collection from more than 95% of the measured cells. Besides relative and absolute cell frequencies, it is also essential to automatically evaluate fluorescence intensities of cell populations, as well as the relative packing densities of the specifically labelled biomolecules in the form of relative surface densities for cell membrane components and relative concentrations for molecules in the cell interior. The calculation of parameter ratios and coefficients of variations provides essential complements for determining the parametric heterogeneity of cells.

The knowledge extraction from this information sometimes stretches to several thousand database columns and can be performed in various ways. Mathematical and statistical knowledge extraction methods may require assumptions on the mathematical distribution of parameter values or, occasionally, develop problems with missing values. Algorithmic data sieving[5,10] as a mathematical assumption-free alternative provides highly discriminatory, predictive or diagnostic data patterns from thousands of data columns. The analysis is suitable for parallel computing and the processing of image analysis parameters.

Relational Data Classification

Multi-parametric flow cytometers or microscopes represent a complex form of instrumentation. As a consequence, two instruments built with the same parts will not provide identical results from a given sample. This is due to the existing tolerances in the multitude of electronical and optical components of such instruments. Fluorescence and light-scattering signals are measured on relative scales and the cell population-oriented histogram-gating procedures remain, to some extent, arbitrary. These accuracy errors cancel themselves out in the majority of cases on the relational expression of parameter values as a fraction of the results from a reference group. The

10. http://www.biochem.mpg.de/valet/classif1.html

relational expression conserves the relative positions of individual parameters, including the coefficients of variation, as indicators of parametric dispersion. Reference groups of the same type, when established in different laboratories, will be indistinguishable by classification from each other, provided they are composed of representative reference individuals and measured with long-term precision using specific reagents. The definition of reference groups can be obtained by consensus. In this way, the standardised and laboratory-independent classification of relational data is possible. Relational databases from different laboratories can be merged into larger, standard databases. Where the classification reveals differences between reference groups from various laboratories, this indicates methodological or ethnic differences. A relational system for the objective molecular description of diseases and elementary cellular states, such as differentiation, maturation, divisions and malignancy, at the cellular level can be established as a disease-, cell type- or cell differentiation-oriented molecular framework. Different cell types will be in a standardised relation to each other in a kind of periodic system of cells.

Predictive Medicine by Cytomics

Diseases represent alterations of molecular pathways in disease-associated cellular systems. These alterations contain predictive information on the future disease course in individual patients, as well as diagnostic information.[5] Cytomics as a form of molecular cell systems research combines the multi-parametric, cytometric analysis of the heterogeneity of cellular systems with the exhaustive bioinformatic analysis of the apparent molecular phenotypes of all analysed single cells. The correlation of the collected multi-parameter data with the future disease development reveals predictive aspects for the individual patient, while diagnostic conclusions are obtained by correlating the data with the patient's present status.

Individualised disease course predictions according to the predictive medicine-by-cytomics concept[3,4] are different from the patient group-oriented predictions of prognosis,[11,12] such as Kaplan-Meier survival or therapy response curves. Such curves are of recognised importance for the development of new therapies by multi-centre trials but of little value for the individual patient and the treating physician. As a result of non-individualised therapy schemes, a substantial number of patients do not benefit from therapy and may even suffer from adverse drug reactions (ADRs). The predictive medicine-by-cytomics concept overcomes this problem by using the therapeutic lead time for preventive therapies, with the goal of, for example, minimising irreversible tissue losses. The goal may also be to suppress disease declaration, as it was in the case of potential screenings for the degree of sensitisation for asthma in risk families.

Bypassing the mechanisms of genome realisation by differential molecular cell phenotype screening represents a self-focusing mechanism for identifying disease-related molecular hotspots.

Challenges of a Human Cytome Project

In this situation, a human cytome project can simultaneously advance the biomedical sciences in two important directions. The individualised disease course predictions and pre-therapeutic identification of high-risk patients are of immediate use in everyday medicine. This is likely to improve the general efficiency of healthcare. At the same time, the initial bypassing of the complex mechanisms of genome realisation by differential molecular cell phenotype screening for medical purposes (see *Figure 1b*) represents a self-focusing mechanism for the identification of disease-related molecular hotspots. New drug targets may be detectable through the subsequently localised reverse-engineering strategy of functional genome analysis.

The focused effort of a human cytome project will profit from substantial quantities of existing clinical, clinical-chemistry and cytometric single-

11. *A Rosenwald, G Wright, W C Chan, J M Connors, E Campo, R I Fisher, R D Gascoyne, H K Muller-Hermelink, E B Smeland and L M Staudt, "The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-cell Lymphoma",* N. E. J. M., *346, 2002, pp. 1,937–1,947.*
12. *R Repp, U Schaekel, G Helm, C Thiede, S Soucek, U Pascheberg, H Wandt, W Aulitzky, H Bodenstein, R Sonnen, H Link, G Ehninger, M Gramatzki and the AML-SHG Study Group, "Immunophenotyping as an Independent Factor for Risk Stratification in AML",* Cytometry, *53B (2003), pp. 11–19.*

cell data available in many clinical institutions. New studies do not require sampling beyond established and ethically admitted limits since sufficiently informative single-cell analysis can be performed within the 1,000 to 100,000-cell range, amounting to microlitre-scale requirements for blood samples.

Some of the major challenges of a human cytome project concern the development of specially adapted cytometric instrumentation with regard to automated sample preparation, staining, measurement and information extraction. The miniaturisation of flow cell, light sources and photon capture[13] are further important goals in view of the large-scale uses of this technology, especially for flow cytometry as a point-of-care or even home-version technology for the early detection of complications in risk patients. Such instrumentation will require the development of sensitive, multi-spectral nanoparticle[14] or other labelling reagents to provide suitable reagent kits for predictive and diagnostic demands in general medicine.

There is also a significant need for bioinformatic software development concerning fast, efficient and multi-parameter data analysis and knowledge extraction. This is especially true for image analysis during the repeated acquisition of molecular information, including high-throughput systems. It will also be important to describe molecular interrelations within cells simultaneously with the architectural interactions of cells within 3-D reconstructed tissue areas and to parametrise complex 3-D contours of cell and organelle shapes. Furthermore, the development of new strategies of fast knowledge extraction from highly multi-parametric data – for example, for online knowledge extraction in high-throughput systems – has to be advanced and structures for the permanent storage of relational cell classification systems for various diseases and cell states have to be generated.

Altogether, a human cytome project constitutes a substantial challenge for biomedical and bioinformatic scientists, clinicians and those capable of producing innovative technological developments. Furthermore, it has the potential to provide essential leads towards achieving new molecular drug targets. ■

13. *Z Palková, L Váchová, M Valer and T Preckel, "Single-cell Analysis of Yeast, Mammalian Cells and Fungal Spores with a Microfluidic, Pressure-driven, Chip-based System", Cytometry, xx (2004), in press.*
14. *W J Parak, D Gerion, T Pellegrino, D Zanchet, C Micheel, S C Williams, R Boudreaur, M A Gros, C A Larabel and A P Alivisatos, "Biological Applications of Colloidal Nanocrystals",* Nanotechnology, *14 (2003), pp. 15–27.*