Cytometry Part A 59A:167–171 (2004)

Introduction

Cytomics—New Technologies: Towards a Human Cytome Project

G. Valet,[1*] J.F. Leary,[2] and A. Tárnok[3]

[1]Max-Planck-Institut für Biochemie, Martinsried, Germany
[2]University of Texas Medical Branch, Galveston, Texas
[3]Pediatric Cardiology, Heart Center Leipzig GmbH, University Hospital Leipzig, Leipzig, Germany

Received 12 January 2004; Revision Received 19 February 2004; Accepted 19 February 2004

Background: Molecular cell systems research (cytomics) aims at the understanding of the molecular architecture and functionality of cell systems (cytomes) by single-cell analysis in combination with exhaustive bioinformatic knowledge extraction. In this way, loss of information as a consequence of molecular averaging by cell or tissue homogenisation is avoided.

Progress: The cytomics concept has been significantly advanced by a multitude of current developments. Amongst them are confocal and laser scanning microscopy, multiphoton fluorescence excitation, spectral imaging, fluorescence resonance energy transfer (FRET), fast imaging in flow, optical stretching in flow, and miniaturised flow and image cytometry within laboratories on a chip or laser microdissection, as well as the use of bead arrays. In addition, biomolecular analysis techniques like tyramide signal amplification, single-cell polymerase chain reaction (PCR), and the labelling of biomolecules by quantum dots, magnetic nanobeads, or aptamers open new horizons of sensitivity and molecular specificity at the single-cell level. Data sieving or data mining of the vast amounts of collected multiparameter data for exhaustive multilevel bioinformatic knowledge extraction avoids the inadvertent loss of information from unknown molecular relations being inaccessible to an a priori hypothesis.

Challenge: It seems important to address the challenge of a human cytome project using hypothesis-driven molecular information collection from disease associated cell systems, supplemented by systematic and exhaustive knowledge extraction. This will allow the description of the molecular setup of normal and abnormal cell systems within a relational knowledge system, permitting the standardised discrimination of abnormal cell states in disease. As one of the consequences, individualised predictions of further disease course in patients (predictive medicine by cytomics) by characteristic discriminatory data patterns will permit individualised therapies, identification of new pharmaceutical targets, and establishment of a standardised framework of relevant molecular alterations in disease. This special issue of *Cytometry*, on new technologies in cytomics, focuses on prominent examples of this presently fast-moving scientific field, and represents one of the preconditions for the formulation of a human cytome project. © 2004 Wiley-Liss, Inc.

Key terms: human cytome project; cytomics; bioinformatics; multilevel biocomplexity profiling; relational cell classification system

A very significant increase in knowledge on the biomolecular capacity of organisms has resulted from the genome sequencing work. Nevertheless, only a very limited part of the observed structural and functional multilevel biocomplexity of cells and cell systems (cytomes) can yet be explained by all of this information.

The biocomplexity of the genome is further evidenced by the fact that many proteins have different functions depending on their location within individual cells. Functional proteomics will require high-resolution 3D mapping of proteins within single cells. The prediction of 3D protein structures from their amino acid sequence is a typical example of the problems already encountered at the biomolecular level, and this is still far away from the struc-

*Correspondence to: Günter K. Valet, Max-Planck-Institut für Biochemie, Arbeitsgruppe Zellbiochemie, Am Klopferspitz 18a, D-82152 Martinsried, Germany.
E-mail: valet@biochem.mpg.de
Published online 17 May 2004 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/cyto.a.20047

tural and functional complexity of entire cells. Exact predictions of protein structure from sequences containing the 20 most common amino acids are still difficult after more than 30 years of intensive research (1), and despite the explosive development of computing and software capacities in the meantime.

Quantitative analysis of cells at the single-cell level in combination with bioinformatic analysis has led to a new field called cytomics. In clinical application, it opens the way to predictive medicine for individual patients.

CHALLENGE

Considering the far more difficult predictions of the association and functionality of biomolecules in viable cells from the 20,000–30,000 coding gene sequences, and the perception of an increasing number of sense-antisense transcription units and noncoding coregulatory RNAs (2), it seems out of reach to shortly understand the enormous biocomplexity of cells or cytomes by traditional deductive top-down hypothesis development followed by experimental verification. The high redundancy of molecular pathways in cell signalling, cell proliferation, or during apoptosis requires a very substantial number of investigations for the collection of a multitude of details, and there is no certainty that these will have focused on ultimately relevant disease-associated metabolic pathways, molecular hotspots, or new pharmaceutical targets. Alternatively, the bottom-up strategy of single-cell molecular cell phenotype analysis of entire cell systems represents a kind of multilevel molecular reverse-engineering strategy as a complement to deductive approaches. A standardised framework of directly-associated related molecular interrelations can then be established and complemented by the required details.

The concept of predictive medicine by cytomics (3,4) was developed as a consequence of these considerations. It has recently generated thoughts about the challenges of a human cytome project (5).

INFORMATION COLLECTION

Cell systems are composed of various kinds of single cells, constituting the elementary building units of organs and organisms. The individualised analysis of single cells overcomes the problem of averaged results from cell and tissue homogenates in which molecular changes in low frequency cell populations may be wrongly interpreted. Single cells can be either uniform or confined to particular cell subpopulations, while changes in low frequency cell subpopulations may be lost by dilution (6). This problem is overcome by single-cell analysis of the molecular cell phenotypes resulting from genotype and exposure.

Essential progress in single-cell molecular analysis has been achieved by the continuous development of new image and flow cytometric instrumentation. Multicolor measurements (7), spectral imaging (8), confocal (9) and laser scanning cytometry (10–12), fluorescence resonance energy transfer (FRET) (13), fluorescence lifetime imaging (FLIM) (14), and second harmonic imaging (15) mark the progress in optical bioimaging. Also, a family of concepts has been developed that allows image acquisition far beyond the resolution limit, down to the nanometer range, including multiphoton excitation (16), stimulated emission depletion (STED) microscopy (17), spectral distance microscopy (18), and image restoration techniques (19). Fast fluorescence imaging in flow (20), optical stretching in flow (21), and miniaturised flow cytometry within laboratories on a chip (22,23) constitute essential progress in flow cytometry and flow imaging. Biomolecular analysis techniques like bead arrays (24), laser microdissection (25), layered expression imaging (26), single-cell polymerase chain reaction (PCR) (25), tyramide signal amplification (27), or biomolecule labeling by quantum dots (28), magnetic nanobeads (29), and aptamers (30) open new horizons of sensitivity, molecular specificity, and multiplexed analysis at the single-cell level.

The dimensionality of measured molecular cell data can be substantial, especially when six or eight repeated colour staining protocols on many different cell populations are performed (8,31) and when their spatial interrelationship within a tissue is taken into account (32,33). One of the most important outcomes of the Human Genome Project is the realisation that there is considerably more biocomplexity in the proteome than previously appreciated. Not only are there many splice variants of each gene system, but some proteins can function in entirely different ways, not only in different cells, but also in different locations of the same cell, lending additional importance to the single-cell analysis of laser scanning cytometry and confocal microscopy. These differences would be lost in the mass spectroscopy of heterogeneous cell populations. Hence, cytomics approaches may be critical to the understanding of proteomics. In extending this from cells to tissue architecture, machine vision protocols are the key to conducting hyperquantitative profiling of tissue heterogeneity (34). Viable cells may be initially stained for cell functions like intracellular pH, transmembrane potentials, or Ca2+ levels, followed by fixation to remove the functional stains, and staining for specific extra- or intracellular constituents such as antigens, lipids, or carbohydrates. After destaining, specific nucleic acids may be stained. Microscopic image capture and analysis systems, using their spatial relocation capacities, will increasingly permit such staining sequences. Serial optical or histological sections will permit 3D reconstruction of the molecular morphology of cell membrane, nucleus, organelles, and cytoplasm, including the parametrisation of 3D shapes. This will serve as basis for the standardised analysis of proximity and interaction patterns for intracellular structures such as nuclei and organelles, as well as for different cell types within the tissue architecture. Traditional visual and quantitative evaluations of gated 2D or 3D cytometric histograms as in flow cytometry collect only a very limited amount of the available information, and one is never certain whether the really relevant information has been extracted. Experience has also shown that it is not easy to provide quality controlled consensus strategies for multiparameter data evaluation. There is also little preexisting

interpretative knowledge on very complex multiparameter data spaces. Essential information may therefore be lost, simply due to lack of awareness. As a consequence, more sophisticated multidimensional data mining techniques, rather than human pattern recognition and reduction of dimensionality approaches, will be required.

DATA ANALYSIS STRATEGIES

Considering the efforts being made toward sample collection, staining, measurement, and data analysis, it seems mandatory to routinely use automated, self-gating evaluation strategies to extract the entire information content of all measured cells for subsequent knowledge extraction. This means, in practice, for example in flow cytometry, that the percent frequency, means, or medians of light-scatter and fluorescence signals, and light-scatter and fluorescence ratios, as well as the coefficients of variation for all parameters in all evaluation gates should be calculated and databased. An effort should be made to collect this information for more than 95% of all measured cells, to be reasonably certain that no relevant information escapes the analysis (3).

It is empirically advisable to use self-adapting and contiguous gates for the automated evaluation of flow cytometrically well-known cell population entities like lymphocytes, monocytes, or granulocytes, as defined by forward (FSC) or sideward (SSC) light scatter or by typical antigenic properties like the expression of CD45 antigen. The subsequent fluorescence gating can be equally automated by using standard quadrant evaluation at fixed threshold levels in gated two-parameter histograms. The evaluation should always include fluorescence-negative as well as single and double fluorescence-positive cells. It is not of primary importance at this stage that cell population boundaries be respected, since relevant information will be picked up anywhere by subsequent data sieving analysis, provided the information for more than 95% of all cells has been accessed during the information collection phase (3).

The situation is reminiscent of genome analysis by shotgun sequencing, in combination with subsequent computer realignment of sequenced DNA strands, as opposed to the sequencing of a priori overlapping DNA strands. This translates by analogy into generalised versus cell population–oriented information collection in flow and image cytometry. Generalised information collection constitutes a prerequisite for the exhaustive knowledge extraction in cytomics, while cell population–oriented evaluations are of interest for cell differentiation or cell function studies. Multidimensional clustering of histograms in cell population–oriented studies relies on cluster definitions being to some extent arbitrary, because they depend on the cluster model used. They are also computationally intensive, and may require human supervision. Restriction of the analysis to major clusters may miss essential information, while the definition of too many clusters may overstress the information content of the measurement, and produce partially-defined mathematical clusters without biological significance. These problems are reliably avoided by the strategy of automated and generalised information collection, in combination with automated knowledge extraction by data sieving (3).

BIOINFORMATIC KNOWLEDGE EXTRACTION

Knowledge extraction after generalised information collection represents a very essential task. Collected information may easily represent several thousand data columns per set of measurements. The classification of such numbers of data columns by statistical, principal component, fuzzy logic, or neuronal network analysis is frequently beyond the capacity of typical software packages, and may require distributed computing (35). Classification results by these analysis strategies may furthermore depend, to some degree, on the assumption of certain mathematical distributions of parameter values or on predefined levels of correlation coefficients (in the case of cluster analysis). Missing experimental values may have to be reconstituted or data records may have to be discarded, which may influence the final classification result. A further important complication is due to the mixed data-type format, particularly from proteomic databases.

Data sieving (36) as an alternative nonstatistical knowledge extraction strategy does not require mathematical assumptions, missing values do not have to be reconstituted, and the analysis is suitable for parallel computation and inherently fast, because only data thresholding is required for classification.

RELATIONAL CELL CLASSIFICATION SYSTEM

Multiparametric flow cytometers or microscopes represent complex instrumentation, and no two instruments will provide identical results on a given sample, despite the use of the same parts. This is caused by tolerances existing in the multitude of electronic and optical components contained in such instruments. Fluorescence and light-scatter signals are measured on relative scales and the gating of cell populations in histograms remains to some extent arbitrary. These errors of accuracy mostly cancel out when all parameter values are relationally expressed as a fraction of the means of results from a reference group. The relational expression conserves the relative individual parameter means and their coefficients of variation.

Reference groups of the same type, when established in different laboratories, will be indistinguishable by classification against each other, provided they are composed of representative reference individuals and are measured with long-term precision and specific reagents. Reference groups can be defined by consensus. In this way, the standardised and laboratory-independent classification of relational data is possible, and relational databases from different laboratories can be merged into larger standard databases. If the classification reveals differences between reference groups from various laboratories, this is an indication of methodological or location-specific 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 in this way. Different cell types will be in a standardised relation to each other in a specific kind of periodic system of cells.

Human Cytome Project

With these considerations in mind, three major levels of a human cytome project can be distinguished at present.

The first level addresses the behaviour of cells in their life cycle, including cell-cycle control, biomolecule synthesis, import and export of molecules, energy and oxidoreductive balance, and organelle functions, to name only some of the important phenomena. It also seems important to address the very significant dispersion of cell parameters ranging from narrow coefficients of variation (CV), such as DNA G0 peaks between 1% and 3% as opposed to CVs greater than 100% for some widespread immunophenotype distributions. The multiparametric molecular heterogeneity of cells in their combinatorial multiplicity may be of high importance for the reliable adaptation of cells to new conditions, and for the susceptibility or resistance to disease or therapy.

At the second level, single-cell preparations, either as collected or after mechanic or enzymatic preparation, are investigated by flow or image cytometry to determine the molecular status of normal or diseased cells as descriptors for health and disease. This discrimination does not necessarily depend on the representative original in situ assembly of cells within tissues, provided conclusions can be derived from the molecular status of specific cells or combinations of cells. The gene expression profiles of specific cell subpopulations can be studied after cytomic analyses and cell purification strategies, to determine the true differences between diseased and normal cells (6). Mapping back the location of specific proteins within single cells will increasingly become important for validation of proteomic information obtained on cell extracts.

The third level concerns cells in assembled tissues. Since cell–cell signalling is important in virtually all tissues, cytomics technologies capable of mapping the functional interaction of molecules within tissues will become increasingly important to understanding all of these cell–cell interactions, in addition to just physical proximity at the tissue level. The molecular interrelation and proximity of cells within an intact cellular architecture can thus be studied under the most complex conditions and at the ultimate organisational level of cell systems in health and disease.

CONCLUSIONS

- With diseases emerging from deviations of typical molecular processes in cells or cell systems, a detailed molecular knowledge of the enormous biocomplexity of such systems in normal and disease conditions is required to understand the mechanisms of disease processes.
- The necessary knowledge is obtained by multilevel single-cell molecular analysis close to in vivo conditions, in combination with exhaustive extraction of bioinformatic knowledge.
- Results are stored in a standardised relational knowledge system or framework for scientific hypothesis development as well as for direct medicine-related purposes. These results concern predictive medicine by cytomics as the individualised prediction of therapy-dependent future disease course in individual patients, the potential for personalised therapy, and the search for new pharmaceutical targets.
- The establishment of such a system using the various single-cell oriented molecular technologies, in conjunction with specific biomolecule labelling in a specially focused human cytome project, represents a combined challenge to science, medicine, and technological innovation.

ACKNOWLEDGMENTS

We thank R.C. Ecker (in Vienna) and A. Kriete (in Philadelphia) for stimulating and constructive discussions during the preparation of this manuscript.

LITERATURE CITED

1. Aloy P, Stark A, Hadley C, Russell RR. Predictions without templates: new folds, secondary structure, and contacts in CASP5. Proteins 2003;53:436–456.
2. Herbert A. The four Rs of RNA-directed evolution. Nat Genet 2004; 36:19–25.
3. Valet G. Predictive medicine by cytomics: potential and challenges. J Biol Regul Homeost Agents 2002;16:164–167.
4. Valet GK, Tarnok A. Cytomics in predictive medicine. Cytometry 2003;53B:1–3.
5. Van Osta P. A Human Cytome Project? Bionet.Cellbiol Newsgroup 2003. http://news-reader.org/article.php?group=bionet.cellbiol&post_nr=14902.
6. Szaniszlo P, Wang N, Sinha M, Reece LM, van Hook J, Luxon BA, Leary JF. Getting the right cells to the array: gene expression microarray analysis of cell mixtures and sorted cells. Cytometry 2004;59A:000–000.
7. Ecker RC, Steiner GE. Microscopy-based multicolor tissue cytometry at the single-cell level. Cytometry 2004;59A:000–000.
8. Ecker RC, de Martin R, Steiner GE, Schmid JA.. Application of spectral imaging microscopy in cytomics and fluorescence energy transfer (FRET) analysis. Cytometry 2004;59A:000–000.
9. Pawley J, editor. Handbook of biological confocal microscopy, 2nd edition. New York: Plenum Press, 1995.
10. Megason SG, Fraser SE. Digitizing life at the level of a cell: high-performance laser-scanning microscopy and image analysis for in toto imaging of development. Mech Dev 2003;120:1407–1420.
11. Gerstner AOH, Laffers W, Lenz D, Bootz F, Steinbrecher M, Tarnok A. Near-infrared dyes for immunophenotyping by LSC. Cytometry 2002; 48:115-123.
12. Tarnok A, Gerstner A. Clinical applications of laser scanning cytometry. Cytometry 2002;50:133–143.
13. Jares-Erijman EA, Jovin TA. FRET imaging. Nat Biotechnol 2003;21: 1387–1395.
14. Murata S, Herman P, Lin HJ, Lakowicz JR. Fluorescence lifetime imaging of nuclear DNA: effect of fluorescence resonance energy transfer. Cytometry 2000;41:178–185.
15. Campagnola PJ, Loew LM. Second-harmonic imaging microscopy for visualizing biomolecular arrays in cells, tissues and organisms. Nat Biotechnol 2003;21:1356–1360.
16. Manconi F, Kable E, Cox G, Markham R, Fraser LS. Whole-mount sections displaying microvascular and glandular structures in human uterus using multiphoton excitation microscopy. Micron 2003;34: 351–358.
17. Hell SW. Towards fluorescence nanoscopy. Nat Biotechnol 2003;21: 1347–1355.
18. Esa A, Edelmann P, Kreth G, Trakhtenbrot L, Amariglio N, Rechavi G, Hausmann M, Cremer C. Three-dimensional spectral precision distance microscopy of chromatin nanostructures after triple-colour

DNA labelling: a study of the BCR region on chromosome 22 and the Philadelphia chromosome. J Microsc 2000;199(Pt 2):96–105.
19. Holmes TJ, Liu YH. Image restoration for 2-D and 3-D fluorescence microscopy. In: Kriete A, editor. Visualization in biomedical microscopies 3-D imaging and computer applications. Weinheim: VCH-Publisher;1992. p 283–323.
20. George TC, Basiji DA, Hall BE, Lynch DH, Ortyn WE, Perry DJ, Seo MJ, Zimmermann CA, Morissey PJ. Distinguishing modes of cell death using the ImageStream multispectral imaging flow cytometer. Cytometry 2004;59A:000–000.
21. Lincoln B, Erickson HM, Schinkinger S, Wottawah F, Mitchell D, Ulvick S, Bilby C, Guck J. Deformability-based flow cytometry. Cytometry 2004;59A:000–000.
22. Kruger J, Singh K, O'Neill A, Jackson C, Morrison A, O'Brien P. Development of a microfluidic device for fluorescence activated cell sorting. J Micromech Microeng 2002;12:486–494.
23. Palková Z, Váchová L, Valer M, Preckel T. Single-cell analysis of yeast, mammalian cells and fungal spores with a microfluidic pressure driven chip-based system. Cytometry 2004;59A:000–000.
24. Lund-Johansen F, Davis K, Bishop J, de Waal Malefyt R. Flow cytometric analysis of immunoprecipitates: high-throughput analysis of protein phosphorylation and protein-protein interaction. Cytometry 2000;39:250–259.
25. Taylor TB, Nambiar PR, Raja R, Cheung E, Rosenberg DW, Anderegg B. Microgenomics: identification of new expression profiles via small and single-cell samples. Cytometry 2004;59A:000–000.
26. Englert CR, Baibakov GV, Emmert-Buck MR. Layered expression scanning: rapid molecular profiling of tumor samples. Cancer Res 2000; 60:1526–1530.
27. Freedman LJ, Maddox MT. A comparison of anti-biotin and biotinylated anti-avidin double-bridge and biotinylated tyramide immunohistochemical amplification. J Neurosci Methods 2001;112:43–49.
28. Parak WJ, Gerion D, Pellegrino T, Zanchet D, Micheel C, Williams SC, Boudreaur R, Le Gros MA, Larabel CA, Alivisatos AP. Biological applications of colloidal nanocrystals. J Nanosci Nanotechnol 2003;14:15–27.
29. McCloskey KE, Chalmers JJ, Zborowski M. Magnetic cell separation: characterization of magnetophoretic mobility. Anal Chem 2003;75: 6868–6874.
30. Ulrich H, Martius AHB, Pesquero JB. RNA and DNA aptamers in cytomics analysis. Cytometry 2004;59A:000–000.
31. Lenz D, Gerstner A, Laffers W, Steinbrecher M, Bootz F, Tarnok A. Six and more color immunophenotyping on the slide by laserscanning cytometry (LSC). In: Nicolau DV, Enderlein J, Leif RC, Farkas DL, editors. Manipulation and analysis of biomolecules, cells and tissues. Proceedings of SPIE 2003;4962:364–374.
32. Gerstner AOH, Racz P, Osmancik P, Tenner-Racz K, Tarnok A. Quantitative histology by multicolor slide-based cytometry. Cytometry 2004;59A:000–000.
33. Smolle J, Gerger A, Weger W, Kutzner H, Tronnier M. Tissue counter analysis of histologic sections of melanoma: Influence of mask size and shape, feature selection, statistical methods and tissue preparation. Anal Cell Pathol 2002;24:59–67.
34. Kriete A, Freund J, Anderson M, Love B, Caffrey J, Young B, Sendera T, Magnuson S, Braughler M. Combined histomorphometric and gene expression profiling applied to toxicology. Genome Biol 2003;4:R32.
35. Snow CD, Nguyen H, Pand VS, Gruebele M. Absolute comparison of simulated and experimental protein-folding dynamics. Nature 2002; 420:102–106.
36. Valet G, Hoeffkes HG. Data pattern analysis for the individualized pre-therapeutic identification of high risk diffuse large B-cell lymphoma (DLBCL) patients by cytomics. Cytometry 2004;59:000–000.