Cytometry Part A 59A:232–236 (2004)

# Data Pattern Analysis for the Individualised Pretherapeutic Identification of High-Risk Diffuse Large B-Cell Lymphoma (DLBCL) Patients by Cytomics

**Günter K. Valet[1]* and Heinz Gert Hoeffkes[2]**

[1]Max-Planck-Institut für Biochemie, Cell Biochemistry Group, Martinsried, Germany
[2]Interdisciplinary Cancer Center, Klinikum Fulda, Fulda, Germany

Received 12 January 2004; Revision Received 10 March 2004; Accepted 12 March 2004

**Background:** Clinical outcome predictions in phase III studies are mostly derived for patient groups, but not for individual patients, although individualised predictions are an ultimate goal to permit a personalised fine tuning of therapy. This may permit earlier application of target therapies, minimise general damage to the organism, and result in improved complete remission rates in malignant diseases.
**Methods:** In this study, Lymphochip cDNA microarray gene expression results of DLBCL patients, from a published prospective meta-analysis study on the prediction of group prognosis, were analysed for individualised predictions using a nonstatistical data pattern classification approach. The training set was comprised of the same 160 DLBCL patients as in the prognosis study, with the validation set of 80 patients remaining unknown to the learning process. This permits the assessment of prospective classifier performance towards unknown patients.
**Results:** Pretherapeutic predictions for the training and validation set patients were correct in 98.1% and 78.3% of the cases for nonsurvival and in 67.3% and 45.3% for survival. The discriminatory data pattern consisted of 14 known and 10 unknown gene products.
**Conclusions:** The better than 95% correct pretherapeutic prediction for about one-half of the ultimately nonsurviving high-risk patients of the training set is promising for clinical considerations about individualised therapy in such cases. Reliable individualised survival predictions are not possible with the information content of the present dataset. It seems necessary to investigate additional gene products, since survival may significantly depend on non–lymphocyte-associated genes that escape to the lymphocyte-oriented Lymphochip gene activation analysis. © 2004 Wiley-Liss, Inc.

**Key terms:** predictive medicine; cytomics; diffuse large B-cell lymphoma; DLBCL; data sieving; data pattern classification

The individualised response prediction of cancer patients to therapy is of significant clinical interest. Early targeted therapies may diminish disease-associated irreversible damage, favour increased quality of life, and minimise serious adverse events. Several predictive approaches in clinical medicine can be distinguished.

Predictive medicine by genomics (1) addresses the prenatal recognition of rare genetic disorders, but also the prediction of future disease occurrence in postnatal life. Pharmacogenetics (2) may be therapeutically important in certain metabolic situations. Disease induction depends, however, frequently more on exposure than on the genotypic background of a patient. As a consequence, sustained exposure of genotypically unsusceptible individuals may be disease-inducing, such as in allergies, while individuals with genetic disposition and no exposure may remain disease-free.

Prognosis prediction in large patient groups is useful for patient stratification in multicenter trials or for the development of new pharmaceuticals. Kaplan-Meier statistics (3) or hierarchical clustering (4), for instance of cDNA expression arrays (5,6), are often used in this context.

*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.20057

Table 1
*Principles of Data Pattern Classification*

| | Disease course prediction | |
|---|---|---|
| a) Disease classification masks (schematic for 10 parameters): | | b) High accuracy at low risk of random coincidence between patient classification masks and disease classification masks: |
| 0000000000 | Stationary | $1/3^{10} \rightarrow 0.0017\%$ probability for random coincidence with 10 parameter masks and $1/3^{7} \rightarrow 0.046\%$ for random coincidence with 7 parameter masks |
| ++++++++++ | Improvement | |
| −−−−−−−−−− | Deterioration | |
| c) Patient classification masks (some examples): | | d) High multiplicity of patient classification masks at correct predictions in case of partial coincidence like 7 out of 10 parameters: |
| 0−000+00+0 | Stationary | |
| 00++000+00 | Stationary | $\frac{10! \times 2^3}{7! \times 3!} = 960$ possible patient classification masks as potential result of genotype and exposure influences (7) |
| −0000−00+0 | Stationary | |
| 000+0+000− | Stationary | |
| 0−++000000 | Stationary | |
| ... | ... | |

Patient stratification does not, however, provide individualized predictions concerning therapeutic susceptibility or final disease outcome. A significant number of patients are for this reason over- or undertreated. The prediction of prognosis in DLBCL patients by hierarchical data clustering of results from Lymphochip cDNA gene expression microarrays discriminate, for example, between the subgroups of activated B-cell-like, type3, and germinal-center B-cell-like patients with average 10-year survival rates of 25, 35, and 47% (5). The analysis does not, however, provide predictions on individualised disease outcome, which is of potential importance for an early decision on stem cell transplantation. Individualized disease course predictions can be obtained by nonstatistical data pattern classification with a data sieving algorithm (7) for the simultaneous multilevel profiling of cellular, clinical chemistry, and clinical multiparameter data. The goal of the present study is to provide individualised disease course predictions from the same cDNA array results that had been previously used for prognosis prediction (5) of patient groups.

## MATERIALS AND METHODS

### Patients

Tumor biopsy specimens of 240 untreated diffuse large-B-cell lymphoma patients were processed for Lymphochip cDNA array analysis to screen for preferentially-expressed genes in lymphoid cells, or for genes being involved, according to present knowledge, in the development of cancer or in various immune functions (5). Resulting data were downloaded from the Internet (http://llmpp.nih.gov/DLBCL [the web supplement to (5)]) (8). The original, randomised assignment of patients to a training set of 160 patients and an a priori defined unknown validation set of 80 patients for prospective verification of the learned classifiers was strictly respected.

### Data Pattern Classification

Numeric multiparameter data sets of more than 10,000 data columns are accessible to data pattern classification by the CLASSIF1 algorithm (7). The parameter values of the 7,399 available spots of the Lymphochip arrays were first converted into following the triple-matrix characters, (–) for diminished, (0) for unchanged, and (+) for increased, depending on their position toward a lower and an upper percentile threshold of the respective parameter value distributions for reference patients. Reference patients were the survivor patients. An essential feature of the data classification algorithm consists, during the learning phase, of the maximization of the diagonal sum of a confusion matrix between clinical and predicted patient future. The improvement or deterioration of the classification result is iteratively recorded following sequential temporary removal and reinsertion of single parameters or combinations of two parameters in all permutations. At the end of this iterative data sieving process, parameter columns that have improved the classification are retained, while those that have deteriorated the classification are definitively excluded. Disease classification masks for each predicted disease course are established by insertion of the most frequent triple-matrix character for each selected parameter column into the respective masks (Table 1a). Disease classification masks typically contain between five and 30 parameters.

The patient classification masks show a high accuracy, i.e., low chances for random occurrence. As soon as the disease classification masks are constituted by five or more parameters, the statistical coincidence probabilities are less than 1%. They are, for example, 0.046% and 0.0017% for seven-parameter and 10-parameter masks (Table 1b), respectively, and are therefore negligible in the context of typical clinical studies of a few hundred patients.

At the same time, the frequently-observed partial coincidence between the patient-classification masks and the positionally-most-similar disease-classification mask (Table 1c) nevertheless results in stable classifications for a high multiplicity of different patient classification masks. The multiplicity arises from influences of genotype and exposure on the expression levels of certain of the predictive

Table 2
*Classification of Patient Training set and Unknown Validation set*

| | CLASSIF1 prediction | | | | | |
|---|---|---|---|---|---|---|
| | a. Training set | | | b. Unknown validation set | | |
| Clinical outcome | Patients (n) | Survival (%) | Non survival (%) | Patients (n) | Survival (%) | Nonsurvival (%) |
| Survival | 71 | **98.6** | 1.4 | 29 | **82.8** | 17.2 |
| Nonsurvival | 86 | 39.5 | **60.5** | 47 | 61.7 | **38.3** |
| Negative/positive predictive values | | **67.3** | **98.1** | | **45.3** | **78.3** |

Classification for 25–75% percentile thresholds, classifiable patients: 157/160 (98.1%) and 76/80 (95.0%), database: S1R12P25.Bl4. Unclassifiable patients (1.9%, 5.0%) had equal numbers of positional coincidences for the survival and nonsurvival data pattern. These transitional classification patients were not considered for the above tables. Specificity and sensitivity values on the diagonal as well as negative and positive predictive values are printed in **bold** for easier comparison.

parameters. The combinatorial analysis of a 10-parameter patient-classification mask coinciding at seven positions with a given disease-classification mask correctly results in the disease course prediction "stationary" (Table 1c). The noncoincident three positions, each with the two possible states (+) = increased or (-) = decreased, generate a multiplicity of 960 possible patient classifications masks (Table 1d).

Characteristic features of the triple-matrix classification algorithm are, therefore, high classification accuracy (Table 1b), at high multiplicity of patient classification masks (Table 1d), as a consequence of genotypic and exposure influences on individual patients. The analysis of the deviations (Table 1c) provides a systematic approach to the investigation of disease induction or progress at variable genotypic and exposure backgrounds.

## RESULTS

The application of data pattern classification to the 7,399 spot Lymphochip cDNA expression arrays of the 160 training set patients provided a predictive value of 98.1% (Table 2a) for the individualised identification of high-risk patients prior to therapy at a predictive value of 67.3% for the survival prediction.

The prospective classification of the unknown validation set patients (Table 2b) indicates that unknown patients are classified by the nonstatistical data pattern classification, similar to the known patients of the training set. Most patients (98.1%, 95.0%) of the dataset are classifiable; unclassifiable patients show transitional classifications with equally-frequent positional coincidence of the patient classification masks with the disease classification masks for survival and nonsurvival.

The disease classification masks (Table 3) contain 24 gene products, 14 of which are known, while 10 are unknown. The selected mask parameters are mostly correlated with coefficients of correlation below 0.500. Exceptions are correlation coefficients of 0.855 for survivors and 0.895 for nonsurvivors between mask parameters nr. 11 and nr. 12, corresponding twice to HLA DP-alpha1 |Hs.914|*H60848| and |Hs.914|*H62848| gene products as two different clones of the same molecule. Further correlation coefficients were 0.591 between nuclear receptor subfamily 3, group C, member 1 |Hs.75772| and HLA DP-alpha1 |Hs.914|*H62848| for survivors, and 0.537 between hypothetical protein FLJ10116 |Hs.79741| and IFN-gamma-inducible protein 30 |Hs.14623| for nonsurvivors. The expression differences of the selected mask parameters were mostly statistically highly significant ($P < 0.05$–$0.0005$, Student's *t*-test) with the exception of $P < 0.10$ for mask parameters nr. 10, 16, 22, and 23 (|Hs.170195|, |Hs.574|, ||||LC_20218, |Hs.88411|).

## DISCUSSION

Data pattern classification provides the individualised pretherapeutic identification of approximately half of the ultimately nonsurviving high-risk patients prior to chemotherapy at a >95% level in the training set. The individualisation of predictions is of significant clinical interest. It may favour the individualised reconsideration of therapy schedules for high-risk patients, leading to therapy individualisation within clinically-stratified patient groups. The individualisation of outcome predictions by data pattern classification represents an advantage over prognosis prediction (3,5), because it addresses the individual patient.

The comparison between the parameters of the disease classification masks for individualised predictions with the 17-parameter classification mask for the prediction of prognosis (5) indicates that they are similar for HLA DP-alpha1, but different for all other parameters. The difference between predictive and prognostic data patterns is not surprising, since the characteristic average behaviour of entire patient groups (3,5) is not paralleled by similar characteristics of individual group members. This has been equally observed in clinical multiparameter data pattern classification of immunophenotype CD-antigen expression in combination with various cytogenetic abnormalities for patients of an multicenter study of AML (9). Like for DLBCL, the individualized pretherapeutic detection of high-risk AML patients is of significant therapy related clinical interest.

Some patients were considered unclassifiable because of equal numbers of positional coincidences of their patient classification mask with the survival or nonsurvival disease classification masks (see bottom of Table 2). These transitional classification patients may be borderline patients. More information will become accessible through

Table 3
*Disease Classification Masks*

| Nr. | Parameter | ID-code | S | NS |
|---|---|---|---|---|
| 5 | Glutathione synthetase | HS.82327 | −,0 | + |
| 10 | Bone morphogenetic protein 7 | HS.170195 | −,0 | + |
| 17 | Caspase 6 cysteine proteinase | Hs.3280 | −,0 | + |
| 20 | Intercellular adhesion molecule 2 | Hs.347326 | −,0 | + |
| 21 | Chemokine (C-X3-C) receptor | Hs.78913 | −,0 | + |
| 23 | CD117 (stem cell growth factor receptor) | Hs.88411 | −,0 | + |
| 7 | Nuclear receptor subfamily 3, group C, member 1 | Hs.75772 | 0,+ | - |
| 11 | HLA DP-alpha1 | Hs.914 * H60848 | 0,+ | - |
| 12 | HLA DP-alpha1 | Hs.914 * H62848 | 0,+ | - |
| 13 | Solute carrier family 2 (facilitated glucose transport), member 3 | Hs.7594 | 0,+ | - |
| 15 | IFN-gamma-inducible protein 30 | Hs.14623 | 0,+ | - |
| 16 | Fructose-1,6,-biphosphatase 1 | Hs.574 | 0,+ | - |
| 18 | CD9 antigen (p24) | Hs.1244 | 0,+ | - |
| 19 | Adenosine kinase | Hs.94382 | 0,+ | - |
| 2 | ‖‖LC_28024 | not available | −,0 | + |
| 3 | DKFZP434F2021 protein | Hs.78277 | −,0 | + |
| 4 | ESTs | Hs.22635 | −,0 | + |
| 6 | Hypothetical protein MGC4189 | Hs.334808 | −,0 | + |
| 8 | MAD homolog 4 (Drosophila) | Hs.75862 | −,0 | + |
| 9 | H.sapiens mRNA; cDNA DKFZp586L141 | Hs.140945 | −,0 | + |
| 22 | ‖‖LC_20218 | not available | −,0 | + |
| 1 | MAD homolog 5 (Drosophila) | Hs.37501 | 0,+ | - |
| 14 | Hypothetical protein FLJ10116 | Hs.79741 | 0,+ | - |
| 24 | ESTs, hypothetical protein FLJ2024 | Hs.159556 | 0,+ | - |

The positional sequence (nr.) of the selected parameters in this table is rearranged to better visualise systematic increases (+) and decreases (–) of known and unknown gene products for nonsurvivor (NS) patients. Nonsurvivors (NS) in data column 5 differ from survivors (S) as reference in data column 4 by increases (+) or decreases (–) of the discriminatory parameters. The occurrence of two characters (–,0) or (0,+) indicates that reference patients are classified S for decreased (–) and unchanged (0) parameters and NS for increased parameters while the classification NS at decreased (–) parameters leads to the classification S at unchanged (0) and increased (+) parameters. Genes are identified according to the UniGene nomenclature (Hs.number). No UniGene numbers are available for the Lymphochip clones LC_20218 and LC_28024.

the systematic classification of the triple-matrix patterns of survivors and nonsurvivors for the occurrence of particular pattern types. This has not yet been done in a systematic way, although it seems possible to discriminate between various genotypic or exposure-induced patterns in this way.

Concerning the linked expression of gene products, correlation coefficients, as well as the statistical significance of selected parameters, are calculated for informative purposes. The information is, however, not used for classification since the learning process is entirely directed towards the maximisation of the diagonal sum of the confusion matrix (Table 2a). In datasets with many discriminatory parameters, only the most discriminating parameters are selected, while correlated parameters will mostly fall out during the learning process. Specific exceptions are mentioned in Results. The majority of the selected classification parameters are highly significantly different between survivor and nonsurvivor patients, except that the four mask parameters nr. 10, 16, 22, and 23 (Table 3) show only borderline statistical significance. The explanation for this is that the CLASSIF1 algorithm screens parameters according to their discriminatory capacity, but not according to statistical or correlative criteria. As a consequence, parameter distributions with only slightly different means but substantial right skew in combination with a left skewed distribution, or a biologically widespread parameter distribution in combination with a narrowly spread distribution, may contain significant classification information.

Although informative for the early identification of high-risk patients, the present Lymphochip gene array pattern does not contain enough information for a clinically usable survival prediction at the individual patient level. It seems important to extend the gene activation screen beyond the Lymphochip addressed genes, since patient survival may not only depend on lymphocyte-related genes, but rather on gene products in cells counteracting the malignant process.

The data pattern of the 14 known gene-expression products (Table 3) contains a number of mechanistically interesting parameters. These are CD9, CD117, caspase6, IFN-gamma-inducible protein 30, or chemokine receptor; however it is presently not possible to integrally understand the predictive changes in the disease-classification masks for ultimate nonsurvivor patients. The 10 unknown gene products are equally of interest. Their selection indicates a good discriminatory potential among a large majority of less informative parameters. The unknown gene products may represent activity signs of unknown structures, or metabolic pathways that have so far remained inaccessible to the development of deductive hypothesis.

## CONCLUSIONS

Complex multiparameter microarray data sets can be classified for individualised predictions of a patient's further disease course. This may permit the development of individualised therapies for high-risk patients. The most discriminatory parameter combinations of the disease-classification masks can be used as a starting point for the formulation of testable new hypotheses on possible mechanisms of disease progress or disease induction. Unknown molecular interrelations and genome activities that remained so far hidden to deductive hypothesis development can be uncovered by the observed discriminatory data patterns.

## ACKNOWLEDGEMENT

We thank H.W. Valet for helpful discussions concerning the calculations in Table 1.

## LITERATURE CITED

1. Kaplan J. Genomics and medicine: hopes and challenges. Gene Ther 2002;9:658–661.
2. Lindpaintner K. Pharmacogenetics and the future of medical practice. J Mol Med 2003;81:141–153.
3. 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. Immunophenotyping as an independent factor for risk stratification in AML. Cytometry 2003;53B: 11–19.
4. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:1486–1488.
5. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltnane JM, Hurt EM, Zhao H, Averett L, Yang L, Wilson WH, Jaffe ES, Simon R, Klausner RD, Powell J, Duffey PL, Longo DL, Greiner TC, Weisenburger DD, Sanger WG, Dave BJ, Lynch JC, Vose J, Armitage JO, Montserrat E, Lopez-Guillermo A, Grogan TM, Miller TP, LeBlanc M, Ott G, Kvaloy S, Delabie J, Holte H, Krajci P, Stokke T, Staudt LM; Lymphoma/Leukemia Molecular Profiling Project. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 2002;346:1937–1947.
6. Ntzani EE, Ioannidis JPA. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Lancet 2003;362:1439–1444.
7. Valet G. Predictive medicine by cytomics: potential and challenges. J Biol Regul Homeost Agents 2002;16:164–167.
8. Lymphoma/Leukemia Molecular Profiling Project. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. Bethesda, MD: National Institutes of Health. (http://llmpp.nih.gov/DLBCL).
9. Valet G, Repp R, Link H, Ehninger G, Gramatzki M, SHG-AML study group. Pretherapeutic identification of high-risk acute myeloid leukemia (AML) patients from immunophenotypic, cytogenetic, and clinical parameters. Cytometry 2003;53B:4–10.