in vivo *20*: 49-54 (2006)

Simulating Soft Data to Make Soft Data Applicable to Simulation*

MATHIAS WAGNER[1], MALGORZATA ZAMELCZYK-PAJEWSKA[2], CONSTANTIN LANDES[3], HOLGER SUDHOFF[4], JOANNA KOSMIDER[2], TEREZA RICHARDS[5], ULRIKE-MARIE KRAUSE[6], ROBIN STARK[6], ANDREAS GROH[7], FRANK WEICHERT[8] and ROLAND LINDER[9]

[1]*Department of Pathology, Saarland University, Homburg-Saar, Germany;*
[2]*Air Odour Quality Laboratory, Institute of Chemical Engineering and of Process of Environmental Protection, Technical University of Szczecin, Szczecin, Poland;*
[3]*Maxillofacial and Plastic Facial Surgery, Johann Wolfgang Goethe-University Frankfurt, Frankfurt am Main;*
[4]*Department of Otorhinolaryngology, Head and Neck Surgery, St Elisabeth Hospital, Ruhr-University of Bochum Medical School, Bochum, Germany;*
[5]*Scientific Library, University of the West Indies, Mona, Kingston, Jamaica, West Indies*
[6]*Institute of University Education and* [7]*Institute for Applied Mathematics, Saarland University, Saarbruecken;*
[8]*Department for Computer Sciences VII, University of Dortmund, Dortmund;*
[9]*Institute for Medical Informatics, University of Lübeck, Lübeck, Germany*

Abstract. *Background: Biomedical processes are often influenced by measures considered "non-crisp", "soft" or "subjective". Despite the growing awareness of the importance of such measures, they are rarely considered in biomedical simulation. This study introduces an input generator for soft data (input generator SD) that makes soft data applicable to simulation. Materials and Methods: Machine learning approaches and standard regression techniques were applied to simulate odour intensity ratings. Results: The performance of all the applied methods was satisfactory and the results can be used to modify systems biological mathematical models. Conclusion: Soft data should no longer be discounted in systems biological simulations. Exemplarily, it can be demonstrated that the input generator SD produces results that are similar to those that the simulated system can generate. Machine learning and/or appropriate conventional mathematical approaches may be applied to simulate non-crisp processes that can be used to modify mathematical models of any granularity.*

Simulations in systems biology are primarily based on mathematical modelling. Over the past decades, mathematical modelling has become an increasingly important tool to assist in analysing and understanding biomedical processes. Numerous systems biological models and simulations are driven by experimental data. These data primarily include laboratory findings, classified as hard data. Such data are often quantifiable and, hence, can be statistically described with some specificity. Mathematical models in systems biology often process such entries. For instance, hard data on nucleotide polymorphism and other nucleotide substitutions have been the subjects of mathematical modelling (*e.g.*, 1). However, soft data such as human perception and emotional reactions may be difficult to assess or are unattainable. Soft data are of notable importance especially with regards to, for example, psychoneuroimmunological and psychosomatic approaches.

The present work describes the concept of an input generator for soft data ("input generator SD"), which simulates soft data for further processing in subsequent systems biological simulations. As an example of this concept, the human olfactory perception was simulated based on hard laboratory findings. Following this, the perception rating can be incorporated as a psychoneuroimmunological aspect into artificial immune systems. For a better understanding, the implications of psychoneuroimmunology in (human) immune systems are briefly explained.

*This article is dedicated to Professor Dr. med. Gerhard R. F. Krueger, an excellent immunopathologist, scientist and academic teacher.

Correspondence to: Dr. med. Mathias Wagner, Institut für Allgemeine und Spezielle Pathologie, Universitätsklinikum des Saarlandes, 66421 Homburg-Saar, Germany. Tel: ++49 (0)6841 16 23864, Fax: ++49 (0)6841 16 23880, e-mail: trouth@gmx.net

Key Words: Crisp data, non-crisp data, mathematical modelling, systems biology, simulating soft data.

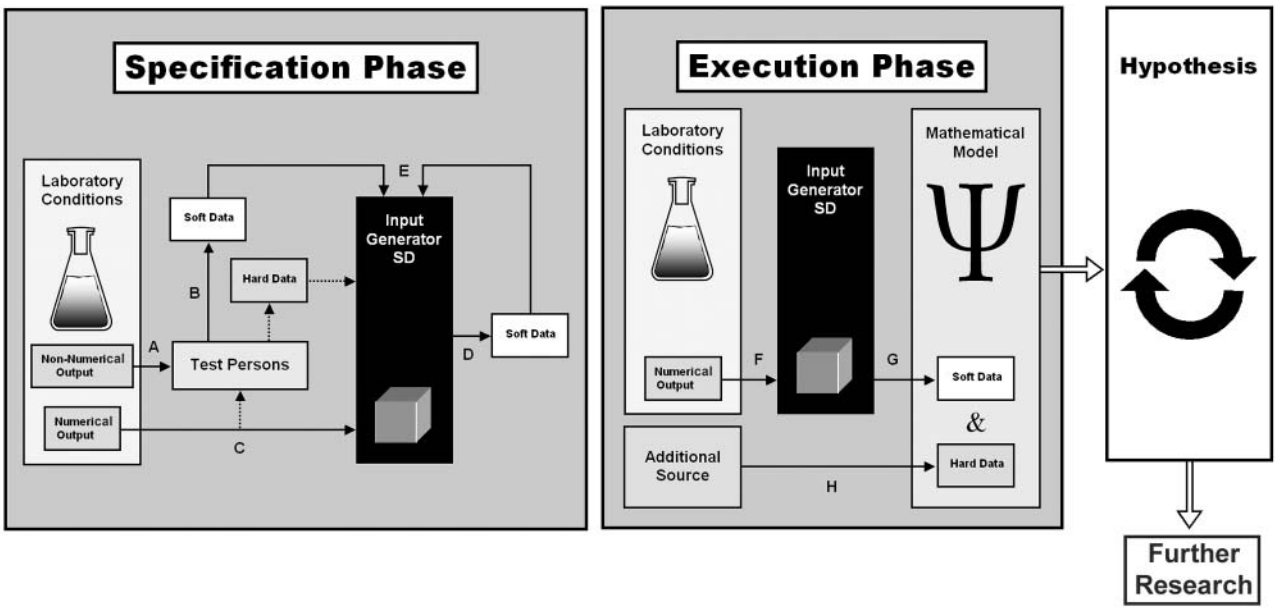

Figure 1. *A diagrammatic representation of the general set-up of the present concept. Multiple simultaneous soft data entries (visual, tactile, gustatory, etc.) may be represented by multiple parallelized "input generators SD" of which only one is displayed here. In the specification phase, test persons are made to perceive non-numerical variables, e.g. odour intensities (A). The test persons are instructed to produce rankings (B). The input generator SD is given putatively related numerical values extracted from biochemical or biophysical characteristics of the matter of interest, e.g. odorant concentrations (C). The input generator SD is set to generate rankings analogous to those of the test persons (D). Optimization strategies iteratively change the values of these artificial soft data to approximate the desired or target values provided by the test persons (E). Hard data might be collected from a test person while ranking. Such data can be made available to the input generator SD to supplement its afore-mentioned input. The specification phase is terminated once an appropriate predictive power has been reached. Test persons are omitted in the subsequent execution phase. After the specification phase, only the numerical output (C, F) is required to allow the input generator SD to generate soft data (D, G) similar to those generated by the test persons. Hard data applied (H) to run mathematical models in systems biology would be supplemented by machine-generated soft data. This might open new doors to the introduction of novel hypotheses.*

The cytotoxic T-lymphocyte (CTL) response can be raised by applying C57BL/6 or DBA/2 spleen cell alloantigens. Natural killer (NK) cell activity can also be elevated by adequate stimulus, for example, injections of an inducer of interferon alpha and beta, polyinosinic:poly-cytidylic acid (poly I:C). Interestingly, odorous conditioning stimuli can produce comparable effects (2-7). Conditioning means that the odorous stimulus was presented to the test persons in association with the alloantigens or the interferon several times previously. Current mathematical models on CTL and/or NK cells (*e.g.*, 8-13). do not include the soft data aspect of psychoneuroimmunology. The authors of the present study, therefore, attempted to modify pre-existing models by entering data on aesthetic rating, which may influence the current discussion on the possible impact of human olfaction on CTL and NK cell activity.

Materials and Methods

General set-up / preliminaries. A few general statements may help define a standard for future studies using the input generator SD approach. The idea is to generate ("simulate") results of trials that produce soft data in order to make soft data applicable to further simulation of higher granularity. The sequence of data analysis begins with a "specification phase" in which the objective is being formulated and the subsequent data collection is carried out (Figure 1). An appropriate mathematical model has to be established which maps the input features into an output. Subsequently, the relationships between input and output have to be analyzed with regards to validation sufficiency. This concept accepts the addition of hard data. Entering hard data at this stage may influence the processing of the soft data. Once declared sufficient, the results can be used in an "execution phase" to modify another mathematical model.

Experimentation. Information on odour intensity rating was obtained as follows. In a ventilated laboratory environment, tedlar foil bags were filled with atmospheric air and 73 different mixtures of cyclohexanol (CAS number: 108930), cyclohexanone (CAS: 108941) and/or cyclohexane (CAS: 110827) in different concentrations. These samples were presented to an average of 60 healthy Caucasian human individuals aged between 21 and 24 years (both genders). Persons with a history of olfactory malfunction were excluded from the further course of the present study. The individual olfactory threshold was determined for each of the remaining individuals. Subsequently, they were asked to rate the unknown intensity of a given odorous compound of cyclohexanol, cyclohexanone and/or cyclohexane in relation to those of serially numbered and standardized ascending dilutions of n-butanol (CAS:

71363). For each compound solution, the individuals were instructed to sort the n-butanol standards according to their intensities in comparison with the compound odorants. This led to a list of numbers of standard dilutions rated with regard to their relative intensities of smell. The difference between the number of the particular standard dilution that was regarded as equal and the individual's odour detection threshold determined the value of the odour intensity rating. The median of these aesthetic ratings was assessed to reduce the impact of outliers.

Specifying the input generator SD. The input generator SD can be run using a wide range of conventional statistics or machine learning approaches. The aesthetic ratings were entered to specify Linear Regression (LR), Non-Linear Regression (NLR) as well as Artificial Neural Network (ANN) approaches and Support Vector Machines (SVM). Training was conducted utilizing error minimization. To determine LR, the statistical software package SPSS V12.0.1 (SPSS Inc., Chicago, USA) was applied, while NRL was computed using MATLAB 7.0.1. R14. The ANN was trained by prototypical software named ACMD (14). The SVM was also specified by a prototypical software based on libSVM (15). All four approaches were evaluated using the leave-one-out method. The precision of the prediction of the aesthetic ratings was assessed using Pearson's Correlation Coefficient as well as the statistic R^2 indicating the significance of the slope of the regression line; the higher the R^2 value, the more efficient the model.

Results

A total of 73 gas mixtures were generated, and the sum of the assessments by the test persons was 4,403. The full data set was used to compute the input generator SD approach. When specifying the ANN and LR approach and validating them by the leave-one-out method, the ratings correlated significantly with the aesthetic ratings provided by the test persons (Table I, Figure 2).

As a consequence, these soft data can be used as an input for subsequent mathematical modelling. Implications will be outlined that refer to the implementation of two aspects of psychoneuroimmunology in artificial immune systems.

Soft data for modifying models of the immune system. Appropriate preparation can result in an increased activity of CTL and NK cells when the individual is exposed to an odorous conditioning stimulus. Antigen-mediated stimulation induces the differentiation of naïve and memory cells into CTL. In experimental settings without distinction on the ground of conditioning, stimulation takes the form of a saturating function (16, 17). Should conditioned CTL activity be proportional to the estimated odour intensity ($aCTL \propto I_{odour}$), soft data (I_{odour}) might have the following impact on CTL stimulation:

$$Stimulation = \frac{I_{odour} \cdot \sum \frac{e_i A_i}{K_i}}{1 + I_{odour} \cdot \sum \frac{e_i A_i}{K_i}} \in [0.1) \qquad \text{(Eq. 3)}$$

Table I. *Results produced by LR, NLR, ANN, and SVM. Investigators planning to apply the input generator SD need to find the best approach to obtain results most suited to for respective models. To date, this can only be achieved by trial and error.*

	LR	NLR	ANN	SVM
Pearson's r	0.770	0.916	0.883	0.881
R^2	0.667	0.771	0.757	0.749

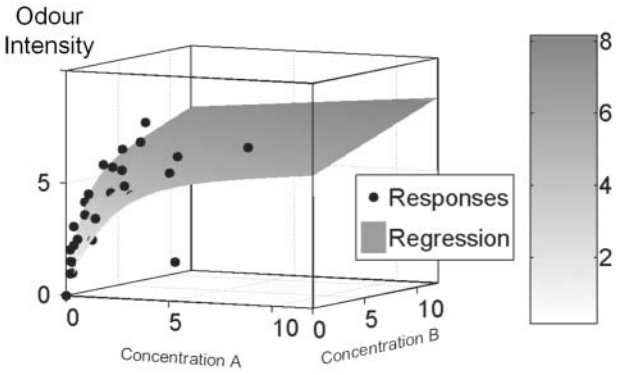

Figure 2. *Non-linear regression (NLR). Measured responses (i.e., laboratory data, represented by black dots) and calculated responses (grey field) in relation to data-points with the concentration C = 0. The concentrations A and B were both scaled (factor: 0.01). A view parallel to the second bisecting line suggests a regression function that contains exponential and linear terms. Thus, in this computation the attempt was:*

$y = f(x,b) = b^1 + b^2 \exp(b^{3x1} + b^{4x2} + b^{5x3}) + b^{6x1} + b^{7x2} + b^{8x3}$ (Eq. 1)

and b (rounded to three decimal places) is calculated as:

$b \approx (4.918, -4.816, -0.526, -0.702, . -0.379, 0.080, 0.196, -0.182)$ (Eq. 2)

where e_i represents the epitope density on cells with antigen i, A_i is the number of affected cells that express antigen i and K_i is the amount of antigen i necessary to generate half-maximal stimulation for the T cell.

In case conditioned NK cell activity is proportional to the estimated odour intensity ($a_{NK} \propto I_{odour}$), then either the sequential model or the two-step selection model could apply (10, 11, 18). The term:

$$S_{min} \leq E_s \leq S_{max} \qquad \text{(Eq. 4)}$$

is valid for all selected NK cells in the two-step selection model, where E_S represents the number of expressed self-major histocompatibility complex (self-MHC) receptor genes, while S_{min} represents the minimum number of

expressed self-specific genes required (S_{max} represents the maximum number). Once the expression of inhibitory genes is reduced as follows:

$$a_{NK} \propto \frac{1}{E_s} \quad \text{(Eq. 5)}$$

a high aesthetic rating (soft data I_{odour}) might reduce the upper threshold:

$$S_{max} \propto \frac{1}{I_{odour}} \quad \text{(Eq. 6)}$$

and thereby modify gene expression in NK cells as follows:

$$E_s \propto S_{max} \quad \text{(Eq. 7)}$$

It may also be argued that the selection process remains unaltered. A higher odour intensity rating may then result in a reduced number of activated genes which could be expressed as:

$$p \propto \frac{1}{I_{odour}} \quad \text{(Eq. 8)}$$

where p represents the probability that a given gene is activated in a cell of the total pre-selection inventory. This means that the average number of inhibitory receptors decreases along with the gene activation probability:

$$E_s \propto p \quad \text{(Eq. 9)}$$

which raises NK cell activity.

Discussion

The current body of biomedical literature has only occasional reports focussing on problems associated with the computer-assisted processing of hard and soft data (19-22), none of which directly applies to systems biological mathematical modelling.

A synoptic overview of the body of literature was undertaken to assess research activity and output on soft data that was dealt with in mathematical modelling in the area of systems biology (as of March 10th, 2005). This was done by serially interrogating the following databases: PUBMED, BIOMED CENTRAL, PROQUEST, EBSCO, FIRSTSEARCH and INSPEC. This resulted in no relevant outcome, which suggests that the input generator SD represents an innovative approach in systems biology. The input generator SD described in the present work therefore appears to be the first to be designed to generate a systems biological input that is based on soft data.

Machine learning approaches have previously been used to predict the human observers' notion of similarity in digital mammography examination (23). A similar approach has been applied to the automated ranking of "facial attractiveness" in a project conducted by Gideon Dror from the School of Computer Science at the University of Tel-Aviv-Yaffo, Israel (unpublished data). Appropriate training set sizes may help achieve human-like performance as machines can be trained to generate soft data. None of the results of the aforementioned trials, however, have yet been used to modify a subsequent systems biological mathematical model.

The present paper does not aim to give a detailed and fair comparison of the methods used. For the odour intensity rating the performance of all applied methods was sufficient, the results being similar to those of experimentally assessed human aesthetic ratings. For other approaches, the suitability of the methods will be far-reaching problem-dependent. It is noteworthy that the input generator SD itself does not constitute a novel method, but adds a new dimension to machine learning when combined with systems biological simulation.

The input generator SD can be used to expand systems biological simulations, for instance in the field of psychoneuroimmunology which has emerged in the neurosciences over recent decades. This conceptual frame describes links between the nervous and immune systems (24-26). Psychoneuroimmunological issues, however, are not commonly addressed in mathematical models of the immune system (27). The currently presented approach may, therefore, help increase awareness to modelling such possible interactions between psychic and physical processes: where there is a virtual immune system in the current body of literature it can now be enriched by soft data integration.

The formulae depicted above lack *in vitro* confirmation and respective details of CTL and NK cell conditioning are still unknown. Odour intensity ratings could, therefore, also correlate positively with the average CTL cell cycle time or be reciprocal to the delay before a stimulated naïve cell becomes a CTL or to the naïve-derived active CTL death rate (9). The result of the virtual odour intensity rating I_{odour} then has to be applied to the respective formula. This introduces a major advantage of the input generator SD concept: the simulation of soft data generation can be retrieved and/or new soft data can be generated by interpolation at any time.

The input generator SD concept does not simulate internal processes of the system that is being simulated. It rather generates results that are similar to those that the simulated system may generate. It is anticipated that controversial discussion may ensue concerning whether the term "simulation" aptly describes the simulation of soft data by the input generator SD. Its applicability in the context of the input generator SD may be interpreted as a matter of

granularity, since the incorporation of additional relevant hard data (*e.g.*, blood test results) is possible. Problems based on over-determined input spaces have to be considered here (28-32). To the knowledge of the authors, there appears to be no generally accepted rule to define whether a given simulation is of appropriate granularity. It may, therefore, be possible to interpret the input generator SD concept as an extremely coarse-grained simulation, reminiscent of the adage: "All models are wrong, but some are useful" (33). Future systems biological mathematical models may, therefore, benefit from the option to enter imitations of ways in which individuals perceive, rank and weigh the variety of elements that make up an attribute, without having to pay attention to the details of the underlying decision-making process.

Acknowledgements

The present study was supported by grants from the Polish State Committee for Scientific Research (KBN), science project no. 1544/T09/2001/21.

References

1 Nei M and Li WH: Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA *76*: 5269-5273, 1979.

2 Demissie S, Ghanta VK, Hiramoto NS and Hiramoto RN: NK cell and CTL activities can be raised *via* conditioning of the CNS with unrelated unconditioned stimuli. Int J Neurosci *103*: 79-89, 2000.

3 Ghanta VK, Hiramoto NS, Solvason HB, Tyring SK, Spector NH and Hiramoto RN: Conditioned enhancement of natural killer cell activity, but not interferon, with camphor or saccharin-LiCl conditioned stimulus. J Neurosci Res *18*: 10-15, 1987.

4 Ghanta VK, Hiramoto NS, Soong SJ and Hiramoto RN: Conditioning of the secondary cytotoxic T-lymphocyte response to YC8 tumor. Pharmacol Biochem Behav *50*: 399-403, 1995.

5 Hiramoto RN, Hsueh CM, Rogers CF, Demissie S, Hiramoto NS, Soong SJ and Ghanta VK: Conditioning of the allogeneic cytotoxic lymphocyte response. Pharmacol Biochem Behav *44*: 275-280, 1993.

6 Hiramoto RN, Rogers CF, Demissie S, Hsueh CM, Hiramoto NS, Lorden JF and Ghanta VK: Psychoneuroendocrine immunology: site of recognition, learning and memory in the immune system and the brain. Int J Neurosci *92*: 259-285, 1997.

7 Solvason HB, Ghanta VK, Lorden JF, Soong SJ and Hiramoto RN: A behavioral augmentation of natural immunity: odor specificity supports a Pavlovian conditioning model. Int J Neurosci *61*: 277-288, 1991.

8 Bocharov G, Klenerman P and Ehl S: Predicting the dynamics of antiviral cytotoxic T-cell memory in response to different stimuli: cell population structure and protective function. Immunol Cell Biol *79*: 74-86, 2001.

9 Chao DL, Davenport MP, Forrest S and Perelson AS: A stochastic model of cytotoxic T cell responses. J Theor Biol *228*: 227-240, 2004.

10 Salmon-Divon M, Höglund P and Mehr R: Generation of the natural killer cell repertoire: the sequential *versus* the two-step selection model. Bull Math Biol *65*: 199-218, 2003

11 Salmon-Divon M, Höglund P and Mehr R: Models for natural killer cell repertoire formation. Clin Dev Immunol *10*: 183-192, 2003.

12 Van Baalen CA, Guillon C, van Baalen M, Verschuren EJ, Boers PH, Osterhaus AD and Gruters RA: Impact of antigen expression kinetics on the effectiveness of HIV-specific cytotoxic T lymphocytes. Eur J Immunol *32*: 2644-2652, 2002.

13 Wodarz D and Jansen VA: A dynamical perspective of CTL cross-priming and regulation: implications for cancer immunology. Immunol Lett *86*: 213-227, 2003.

14 Linder R and Pöppl SJ: ACMD: a practical tool for automatic neural net based learning. Lect Notes Comp Sci *2199*: 168-173, 2001.

15 Chang CC and Lin CJ: Training nu-support vector regression: theory and algorithms. Neural Comput *14*: 1959-1977, 2002.

16 Davenport MP, Fazou C, McMichael AJ and Callan MF: Clonal selection, clonal senescence, and clonal succession: the evolution of the T cell response to infection with a persistent virus. J Immunol *168*: 3309-3317, 2002.

17 De Boer RJ, Oprea M, Antia R, Murali-Krishna K, Ahmed R and Perelson AS: Recruitment times, proliferation, and apoptosis rates during the CD8+ T-cell response to lymphocytic choriomeningitis virus. J Virol 75: 10663-10669, 2001.

18 Salmon-Divon M, Höglund P, Johansson MH, Johansson S and Mehr R: Computational modelling of human natural killer cell development suggests a selection process regulating coexpression of KIR with CD94/NKG2A. Mol Immunol *42*: 397-403, 2005.

19 Adler RH: Die Anamneseerhebung - ein "unmögliches" Unterfangen. [Interviewing the patient - an "impossible" task]. Ther Umsch *61*: 728-731, 2004.

20 Morse JM: The hardening of soft data. Qual Health Res *14*: 591-592, 2004.

21 Sosa MMC, Pablos HJL and Santos AD: Guía para elaborar el protocolo de investigación. V. Material y métodos. Variables y su clasificación. [Guide for clinical research methodology. V. Material and methods. Variables and their classification]. Acta Pediatr Mex *17*: 8-12, 1996.

22 Zeller C and Scherrer M: Harte und weiche Daten in der Medizin; ihre Verarbeitung durch den Arzt und durch den Computer. [Hard and soft data in medicine; their processing by the physician and by the computer]. Schweiz Med Wochenschr *109*: 773-780, 1979.

23 El-Naqa I, Yang Y, Galatsanos NP, Nishikawa RM and Wernick MN: A similarity learning approach to content-based image retrieval: application to digital mammography. IEEE Trans Med Imaging *23*: 1233-1244, 2004.

24 Hausotter W: Begutachtung des Chronic-Fatigue-Syndroms. [Expert assessment of chronic fatigue syndrome]. Versicherungsmedizin *48*: 57-59, 1996.

25 Linder R, Dinser R, Wagner M, Krueger GR and Hoffmann A: Generation of classification criteria for chronic fatigue syndrome using an artificial neural network and traditional criteria set. In Vivo *16*: 37-43, 2002.

26 Wagner M, Krueger GRF, Ablashi DV, Whitman JE and Rojo J: Síndrome de fatiga crónica (SFC): Revisión de los datos clínicos de 107 casos. [Chronic fatigue syndrome (CFS): Review of clinical data of 107 cases]. Rev Med Hosp Gen Mex *61*: 195-210, 1998.

27 Yates A, Chan CC, Callard RE, George AJ and Stark J: An approach to modelling in immunology. Brief Bioinform *2*: 245-257, 2001.

28 Gelsema E: Pattern recognition and artificial intelligence in medical research and clinical practice. Methods Inf Med *28*: 63-65, 1989.

29 Heisele B, Serre T, Prentice S and Poggio T: Hierarchical classification and feature reduction for fast face detection with support vector machines. Pattern Recognit *36*: 2007-2017, 2003.

30 Jain A and Waller W: On the optimal number of features in the classification of multivariate Gaussian data. Pattern Recognit *10*: 365-374, 1978.

31 Kanal L and Chandrasekaran B: On dimensionality and sample size in statistical pattern recognition. Pattern Recognit *3*: 225-234, 1971.

32 Liu H, Li J and Wong L: A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Informatics *13*: 51-60, 2002.

33 Box GEP: Robustness is the strategy of scientific model building. *In*: Launer RL, Wilkinson GN (eds.). Robustness in Statistics, Academic Press, New York, pp. 201-236, 1979.

Received August 23, 2005
Accepted September 21, 2005