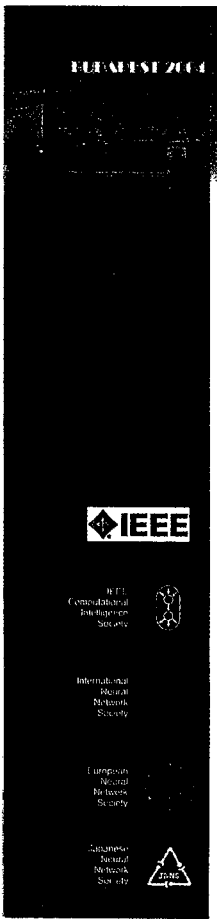


Proceedings

2004 IEEE

International Joint Conference on Neural Networks

International Conference on Fuzzy Systems

IEEE Catalog Number: 04CH37541C
ISBN: 0-7803-8360-5

IEEE Catalog Number: 04CH37542C
ISBN: 0-7803-8354-0

Budapest, Hungary
25-29 July 2004

IJCNN2004 CD-ROM Conference Proceedings
IEEE Catalog Number: 04CH37541C
ISBN: 0-7803-8360-5

FUZZY IEEE 2004 CD-ROM Conference Proceedings
IEEE Catalog Number: 04CH37542C
ISBN: 0-7803-8354-0

Support

If you have questions regarding the installation of this CD's contents, please e-mail to Dániel Bíró, danman@prim.hu.

Thursday, July 29, 2:00PM - 3:40PM

Robotics and Learning
Chair: Jeen-Shing Wang, Room: A

Neurodynamics
Chair: Desire Bolle, Room: H

Bioinformatics and Biomedical Comp. (ENNS Special Session)
Chair: Francesco Masulli and RobertoTagliaferri, Room: B

HLA Typing Using a Fuzzy Approach

Giovanni Battista Ferrara
Immunogenetics Laboratory
National Cancer Research Institute and
Dept of Oncology, Biology and Genetics
University of Genova, Largo R. Benzi 10
I- 16132 Genova, Italy
giovannibattista.ferrara@istge.it

Francesco Masulli
INFM Genova and
Dept of Computer Science
University of Pisa
Via F. Buonarroti 2
I-56125 Pisa, Italy
masulli@di.unipi.it

Stefano Rovetta
INFM Genova and Department of
Information and Computer Science
University of Genova
Via Dodecaneso 35
I-16146 Genova, Italy
rovetta@disi.unige.it

***Abstract*— The Human Leukocyte Antigens (HLA) system consists of three regions in the human genome. In transplantation, the match between donor's and receiver's HLA is critical for histocompatibility. HLA typing problem consists in the donor's and receiver's HLA systems matching. We describe the image analysis module of a Decision Supporting System supporting the application of the oligonucleotide microarray technology to the HLA typing. The Decision Supporting System is based on a fuzzy modeling approach that allows the biologist to describe and classify the probe activations using its language and concepts in a natural way, and, at the same time, supports a robust interactive image filtering thanks to the usage of a Fuzzy Basis Functions network.**

I. Introduction

The Human Leukocyte Antigens (HLA) system consists of three regions in the human genome. In transplantation, the match between donor's and receiver's HLA is critical for histocompatibility. HLA typing problem consists in matching the donor's and receiver's HLA systems.

In the last few years, the availability of the oligonucleotide microarrays technology [2] stimulated the development of new methodologies for HLA typing [13][3].

Oligonucleotide microarrays [2] make it possible to perform a large quantity (100-10000) of simultaneous experiments. Each experiment corresponds to a given oligonucleotide probe (a DNA strand of 15-20 bases) hybridizing with a target RNA sample. The probes are affixed to specific positions of a chip's surface. The target is fluorescently labelled. Therefore a fluorescence measurement by laser scanning gives information about the amount of RNA hybridized at each specific location on the chip (spot).

There are many techniques available for tackling HLA typing using oligonucleotide microarrays, but none of them seems able to resolve all ambiguities [4].

The aims of the Decision Supporting System (DSS) presented in this work is to support the design of oligonucleotide microarrays for HLA typing, and to simplify their analysis. Actually, as noted in [4], combinations of techniques may be necessary for reliable typing, and, moreover, serological analysis may still be necessary. Therefore it is important that the usage of this molecular characterization tool is as simple as possible.

The main constituents of the DSS for HLA typing are:

1) Support to the oligonucleotide probe design;
2) Spotter system programming;
3) DNA microarray hybridation measurement;
4) Genotyping.

The first subsystem of our Decision Supporting System is based on the analysis of the alleles of the HLA system[1]. The task of this subsystem is to support the design of a set of oligonucleotides, of about 15-20 bases, able to discriminate the alleles (in high or low resolution). The ordered list of probes corresponds to the codes associated to the alleles to be discriminated. Note that to each probe corresponds an optimal temperature for the hybridization (*melting temperature*) depending on its constituents. This subsystem tries to suggest to the user oligonucleotides with similar melting temperatures; in facts, as the hybridization will be at the same temperature for all the probes on the microarray, if the probes' melting temperatures are too different there will be the risk of *false positive* and *false negative* spots.

The second subsystem interacts with the user in order to program a spotter to print the selected probes on the target microarrays, with an assigned redundancy level.

The DNA Microarray hybridation measurement subsystem is devoted to classifing the probe activations on the basis of the information coming from the microarray's scanner.

The last subsystem computes the probe activation codes and compares them with the codes associated to groups of alleles to be discriminated supplied by the first subsystem supporting the oligonucleotide probe design.

As reported in the literature [9], [8] the analysis of the information embedded into a DNA microarray is a complex task. In this work, we address in particular the presence of outliers and the possibility that a probe can produce spots with intermediate activation that, in principle, could be ascribed either to the positive or negative activation classes.

In the next section, we will give a short description of the hardware and software environments. Then we will present the DNA microarray hybridation measurement subsystem (Sect. III). Discussion and Conclusions are in Sect IV. The Appendix

[1]The IMGT/HLA Sequence Database is available on-line at the URL http://www.ebi.ac.uk/imgt/hla/, and is continuously updated.

Fig. 1. A sample image produced by the scanner. It is possible to distinguish spots with positive or intermediate activation and outliers, as for instance the bright spots in the bottom area of the image.

contains some details on the classification system based on the Fuzzy Basis Functions network that are used by the DSS.

II. Hardware and software environments

The Decision Supporting System has been developed on a 500MHz PC Pentium in *Sun Java 2*, and is based on an interactive graphical user interface, making extensive use of pure *Sun Java Swing* graphical components such as *tables*, *trees*, *menus* and *image panels*.

The instrumentation setup considered is as follows:

- a Packard-Bell Bioscience Division SpotArray 24 printing system that prints, on one or more slides (DNA microarrays), the probes to be used in the hybridization process;
- a Packard-Bell Bioscience Division ScanArray Express slide laser scanning system.

The available drivers for those instruments are designed for the *Microsoft Windows NT* operating system.

In the spotting task, redundancy plays a relevant role. The spotter robot cannot print single spots but only groups of 5 adjacent spots, in order to prevent printing errors and to consume all the probe "ink" loaded by pins. Moreover, it is important to program the robot to spot the same probe in several different zones of the slide, in order to prevent the effects of local problems due to low quality zones in hybridization process.

The ScanArray Express reads the DNA microarray by laser scanning and produces an high resolution image with spots corresponding to the hybridization activity results of oligonucleotide probes (see, e.g., Fig. 1). Moreover, the scanner driver provides a data base associating each spot to a vector of features, to be used for classification, including:

- the evaluation of intensity level, background level, diameter, area, footprint, circularity, spot uniformity, background uniformity and signal-to-noise ratio and
- the position of spot centers and other geometrical information coming from the spot printing system.

III. DNA microarray hybridization measurement subsystem

We have to consider two sub-problems:

1) The classification of the activity measured by each spot on the basis of the scanner outputs.
2) The integration (or *fusion*) of the activities of spots corresponding to the same probe, in order to obtain a robust evaluation to the results of the hybridization process.

A. Spot activity evaluation

In the oligonucleotide microarray technology, the spot's *hybridization* is, in principle, a binary variable $\{0,1\}$, associated with the outputs of the scanner. In practical experimental cases, it is necessary to take into account some intermediate levels of probe hybridization due to environment factors like the discrepancy of the melting temperature of probes with respect the experimental temperature of the microarray (*false positive* and *false negative* problems), and to others like, e.g., the (partial) probe curling due to the presence of auto-complementary sequences or to the bad anchorage of the probe to the glass.

Given the large throughput of the oligonucleotide microarray technology, an automatic approach to spot hybridization evaluation, on the basis of the features measured by the scanner driver, is necessary.

A direct approach to the design of such sub-systems is not easy, as an expert biologist can discriminate the hybridization level of a spot on the basis of the image produced by the scanner, while he cannot obtain a reliable classification of spots using only the features measured by the scanner driver.

As no useful rules, based on the scanner's features, can be obtained a-priori from the domain expert, we defined the linguistic variable *hybridization* on the *virtual* universe of discourse [0,1]. The term "virtual" means that no direct measurements of the hybridation are possible, while information related to it is embedded in the scanner's features, even if a clear relationship is unknown.

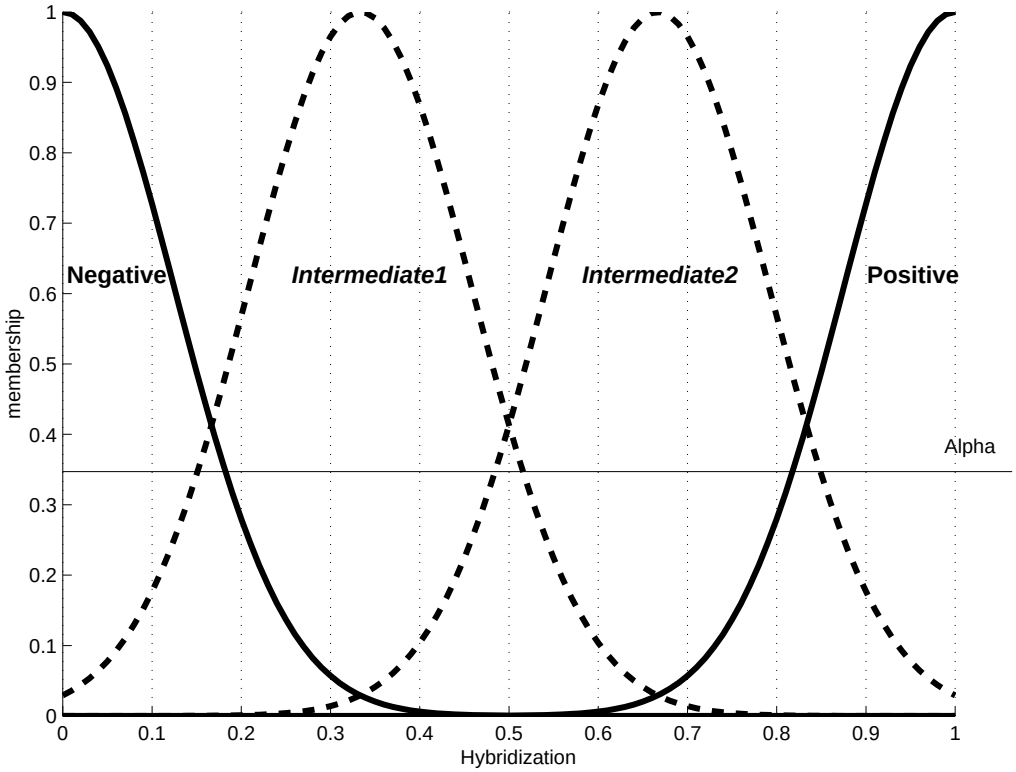


Fig. 2. Linguistic variable *hybridization*. In this figure the *intermediate* fuzzy set is split into two different fuzzy sets.

The terms (fuzzy sets) of the linguistic variable *hybridization* are, e.g., *Negative* (N), *Intermediate* (I), *Positive* (P), and *Outlier* (O) [2]. The membership functions of those terms must be indirectly evaluated on the basis of the scanner's features (*semantic map*).

In our DSS, the semantic map is obtained using a learning system able to estimate the membership functions of terms on the basis of a small set of labeled samples supplied by the biologist through a graphical interface. The main constituent of this module is a learning machine based on a network of Fuzzy Basis Functions (FBF) [15], [12] (see Appendix).

A FBF network is a Mamdami fuzzy logic system [10] with *singleton* fuzzification, *max-product* composition, *product inference* and *height* defuzzification, equivalent to the ANFIS model by Jang [5], that can learn its parameters from a labelled data set using a gradient descent procedure.

Each FBF network's task is the discrimination of class against the remaining others. We use the mean square error (MSE) as the cost function (empirical risk). In this way the FBF network estimates the posterior class conditional probability of any unknown data sample [11], [1], that we can consider as the fuzzy membership to the class. A representation of the linguistic variable *hybridation* using two different *Intermediate* terms is shown in Fig. 2.

Fig. 3 shows an example of the graphical interface. The user can select a small set of samples for each class and in few seconds the adaptive fuzzy system generalizes the classification to all spot in the image. After learning, the spot activity (hybridization) will be classified using a Winner Take All (WTA) rule that associates the data points related to the spot to the highest membership class. Points not belonging to the alpha-cuts shown in Fig. 2 will be rejected (i.e., assigned to a *Outlier* class). Rejected spots will not be considered in the following *fusion* step. The user can accept the classification, or else can either prepare a new training set, or explicitly change the membership class of each spot.

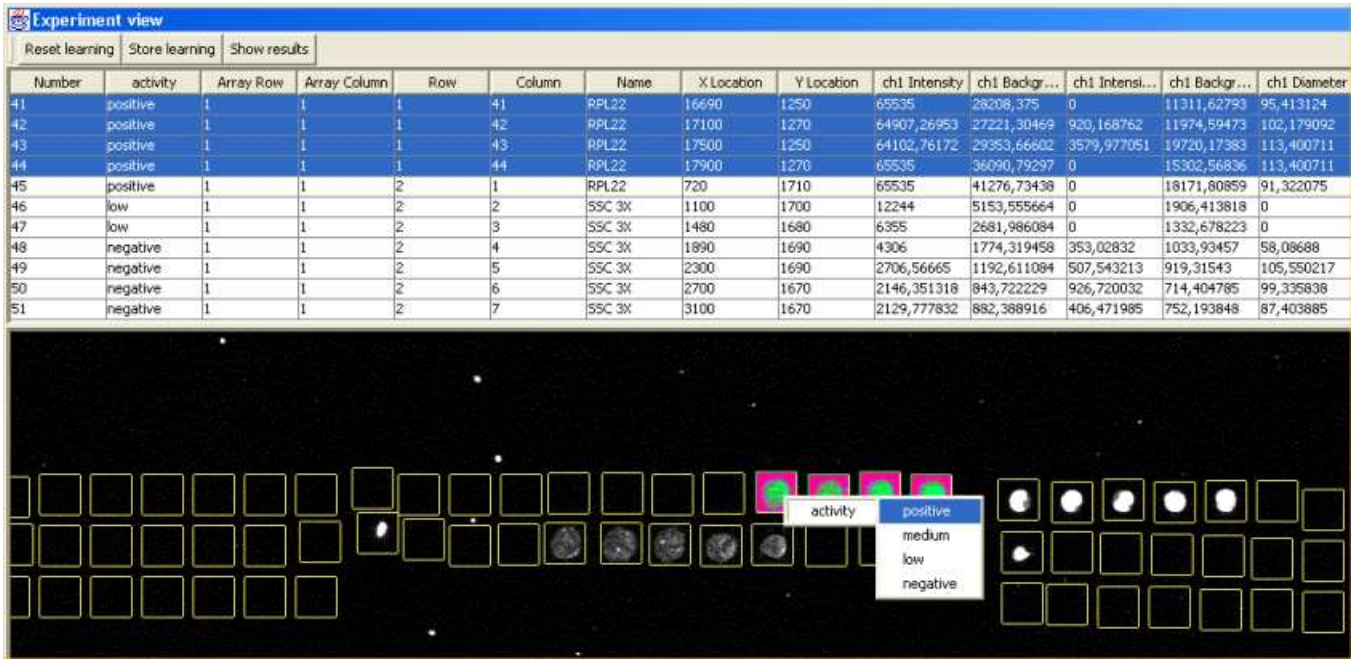
B. Classification fusion

The *Fusion Module* allows the user to integrate the activity levels of the redundant spots corresponding to the same probe and to obtain in such a way a more robust evaluation to the results of hybridization process. For each probe, the fusion can be obtained by a choice of operators such as maximum, minimum, averaging, voting, etc., possibly referred to each sequential group of spots and between groups of spots.

IV. Discussion and conclusions

We have described the image analysis module of a Decision Supporting System designed for supporting HLA typing. This

[2]Note that, in order to model different intermediate activation levels of probe hybridization, such as false negative hybridization and false positive hybridization, *intermediate* fuzzy set can possibly be split into more fuzzy sets.



Number	activity	Array Row	Array Column	Row	Column	Name	X Location	Y Location	ch1 Intensity	ch1 Backgr...	ch1 Intensi...	ch1 Backgr...	ch1 Diameter
41	positive	1	1	1	41	RPL22	16690	1250	65535	28208,375	0	11311,62793	95,413124
42	positive	1	1	1	42	RPL22	17100	1270	64907,26953	27221,30469	920,168762	11974,59473	102,179092
43	positive	1	1	1	43	RPL22	17500	1250	64102,76172	29353,66602	3579,977051	19720,17383	113,400711
44	positive	1	1	1	44	RPL22	17900	1270	65535	36090,79297	0	15302,56836	113,400711
45	positive	1	1	2	1	RPL22	720	1710	65535	41276,73438	0	18171,80859	91,322075
46	low	1	1	2	2	SSC 3X	1100	1700	12244	5153,555664	0	1906,413818	0
47	low	1	1	2	3	SSC 3X	1480	1680	6355	2681,986084	0	1332,678223	0
48	negative	1	1	2	4	SSC 3X	1890	1690	4306	1774,319458	353,02832	1033,93457	58,08688
49	negative	1	1	2	5	SSC 3X	2300	1690	2706,56665	1192,611084	507,543213	919,31543	105,550217
50	negative	1	1	2	6	SSC 3X	2700	1670	2146,351318	843,722229	926,720032	714,404785	99,335838
51	negative	1	1	2	7	SSC 3X	3100	1670	2129,777832	882,388916	406,471985	752,193848	87,403885

Fig. 3. The user interface used in the DNA microarray Hybridization Measurement Subsystem. Each row of the table corresponds to a spot and contains the feature values computed by the scanner driver, plus other information, such as the class of membership (*activity*). On the bottom, on the scanned image, the squares represent the position of spots, and the color of their contours is related to the class to which each spot has the highest membership.

module classifies the spots on a oligonucleotide microarray image on the basis of user hints. The subsequent processing steps transform the list of class memberships of probe hybridization activation into codes. Then the computed codes are compared with whose designed using the Oligonucleotide Probe Design Subsystem.

In the coding process, probes belonging to Positive and Negative fuzzy sets will be coded, respectively, as 1 and 0. The probes belonging to the Intermediate fuzzy set(s) could be assigned either code (1 or 0) depending on domain knowledge obtained by an interaction with the user. This piece of knowledge will be recorded in the probe data base.

As a general comment, the DDS we are developing and testing adds novel tools to those already available for the HLA typing problem [3], [4]. The approach to HLA typing shown in this paper is based on fuzzy modeling, and can be fruitfully integrated with others already published, as each approach presents its own strengths and weakness.

However, our approach provides several advantages and is therefore expected to obtain notable results by itself. In particular, it allows using the standard biologist's language and concepts to describe and classify the probe activations in a natural way. Furthermore, it allows robust interactive image filtering thanks to the usage of a learning machine based on a fuzzy system.

ACKNOWLEDGEMENTS

Work funded by the INFM and by the Italian Ministry of Education, University and Research ("Prin2002"). The Istituto Scientifico Oncologico (ISO) made available the scientific instrumentation. We thank Cinzia Pera and Riccardo Sensi for their support.

APPENDIX

Fuzzy Logic Systems with *singleton* fuzzification, *max-product* composition, *product inference* and *height* defuzzification can be represented as [12]

$$y = f(\mathbf{x}) = \sum_{l=1}^{M} \overline{y}^{\,l} \phi_l(\mathbf{x}) \tag{1}$$

where $\overline{y}^{\,l}$ denote the center of gravity of the output fuzzy set, and $\phi_l(\mathbf{x})$ are called *fuzzy basis functions* and are given by

$$\phi_l(\mathbf{x}) = \frac{\prod_{i=1}^{p} \mu_{F_i^l}(x_i)}{\sum_{l=1}^{M} \prod_{i=1}^{p} \mu_{F_i^l}(x_i)} \tag{2}$$

where $l = 1, 2, ..., M$. We can refer to those FLS as *fuzzy basis expansions* or *networks of fuzzy basis functions* (FBF network) [3].

It is worth noting that the FLS with universal function property studied by Mendel and Wang [15], which is a singleton FLS using product inference, product implication, Gaussian membership and height defuzzification, can be rewritten as a FBF network expansion. The universal function approximation property gives a strong mathematical ground when applying FLSs in critical applications, ranging from control, to time series prediction, to pattern recognition.

[3]The relationships between fuzzy basis expansions and other basis functions have been extensively studied in [7].

Let us consider a fuzzy logic system based on a multi-input-multi-output version of this FBF network. Specifically, if there are K units in the input layer, J fuzzy inference rules and I outputs, the rule activations can be expressed as $r_j = \prod_k \mu_{jk}(x_k)$, where the quantity $\mu_{jk}(x_k)$ represents the value of the membership function of the component x_k of the input vector for the jth rule and is defined as:

$$\mu_{jk}(x_k) = \exp\left(-\frac{(x_k - m_{jk})^2}{2\sigma_{jk}^2}\right), \quad (3)$$

and m_{jk} and σ_{jk}^2 are the means and variances of the Gaussian membership functions. The values of the output units are:

$$y_i = \frac{\sum_j r_j \overline{y}_{ij}}{\sum_j r_j} = \sum_j \overline{y}_{ij}\phi_j(\mathbf{x}) \ , \quad (4)$$

where $\overline{y}_{ij}$ is the center of gravity of the output fuzzy membership function of the jth rule associated with the output y_i, and

$$\phi_j = \frac{\prod_k \mu_{jk}(x_k)}{\sum_j \prod_k \mu_{jk}(x_k)} \quad (5)$$

is the fuzzy basis function associated to rule *j*, and represents its normalized activation. (Without loss of generality, we could assume that the fuzzy membership functions are singletons: $\overline{y}_{ij} \equiv s_{ij}$.)

The FBF network can be regarded as a feedforward connectionist system with one hidden layer whose units correspond to the fuzzy rules. It can be identified [10] both by exploiting the linguistic knowledge available (*structure identification problem*) or by using the information contained in a data set (*parameter estimation problem*), which is the approach followed in the present context.

As shown in [11], in order to obtain a "fuzzy" classifier approximating the Bayes discriminant functions in the large training set size limit, we must find the values of the parameters (or *weights*) that minimize the *mean square error* (MSE) defined as

$$MSE = \frac{\sum_{k,n}(y_k^n - t_k^n)^2}{N} \ , \quad (6)$$

where N is the size of the training set, $\mathbf{y}^n = (y_k^n)$ is the network output, and $\mathbf{t}^n = (t_k^n)$ is the n-th label of the associative pair of the training set. The components of $\mathbf{t}^n$ are defined as follows:

$$t_j = \begin{cases} 1 & \text{if the pattern belongs to class } j, \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

The cost function (6) can be minimized by many different techniques. In our experiments, the FBF network parameters (i.e., m_{jk}, σ_{jk} and $\overline{y}_{ij}$) were obtained by performing a gradient descent with respect to the MSE across the training set.

The learning formulas are as follows [6], [14]:

$$\Delta\overline{y}_{ij} = \eta_s[t_i - y_i]\phi_j \quad (8)$$

$$\Delta m_{jk} = \eta_m\phi_j \sum_i [t_i - y_i][\overline{y}_{ij} - y_i][x_k - m_{jk}]/\sigma_{jk}^2 \quad (9)$$

$$\Delta\sigma_{jk} = \eta_\sigma\phi_j \sum_i [t_i - y_i][\overline{y}_{ij} - y_i][x_k - m_{jk}]^2/\sigma_{jk}^3 \quad (10)$$

where η_s, η_m, and η_σ are the learning rates of $\overline{y}_{ij}$, m_{jk}, and σ_{jk}.

REFERENCES

[1] F. Casalino, F. Masulli, and A. Sperduti. Rule specialization in networks of fuzzy basis functions. *Intelligent Automation and Soft Computing*, 4:73–82, 1998.

[2] R. Ekins and F.W. Chu. Microarrays: their origins and applications. *Trends in Biotechnology*, 1999, 17, 217-218

[3] Z. Guo, L. Hood, E.W. Petersdorf, Oligonucleotide arrays for high resolution HLA typing, *Rev Immunogenet*, 1(2):220-230, 1999.

[4] J.A. Gerlach, Human lymphocyte antigen molecular typing: how to identify the 1250+ alleles out there, *Arch Pathol Lab Med*, 126(3):281-284 2002.

[5] J.S.R. Jang. ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Trans. on Systems, Man, and Cybernetics*, 23:655–684, 1993.

[6] C.C. Jou. Comparing learning performance of neural networks and fuzzy systems. In *IEEE International Conference on Fuzzy Systems*, pages 1028–1033, San Francisco, 1993. IEEE, New York, NY.

[7] H.M. Kim and J.M. Mendel. Fuzzy basis functions: Comparisons with other basis functions.

[8] M. Katzer, F. Kummert, G. Sagerer. Robust automatic microarray image analysis In *BREW Bioinformatics Research and Education Workshop*, Hinxton, UK, 2002.

[9] A. Kuklin, A. Petrov, S. Shams. Quality control in micorarray image analysis *G.I.T. Imaging & Microscopy*, pages 2–3, no. 1, 2001.

[10] C.C. Lee. Fuzzy logic in control systems: fuzzy logic controller. I. *IEEE Transactions on Systems, Man and Cybernetics*, 20:404–418, 1990.

[11] F. Masulli. Bayesian classifi cation by feedforward connectionist systems. In F. Masulli, P. G. Morasso, and A. Schenone, editors, *Neural Networks in Biomedicine - Proceedings of the Advanced School of the Italian Biomedical Physics Association - Como (Italy) 1993*, pages 145–162, Singapore, 1994. World Scientifi c.

[12] J.M. Mendel. Fuzzy logic systems for engineering: A tutorial. *Proceedings of the IEEE*, 83:345–377, 1995.

[13] C. Pera, L. Delfi no, A. Morabito, A. Longo, L. Johnston-Dow, C.B. White, M. Colona, G.B. Ferrara. HLA-A typing, *Tissue Antigens*, 50:372–379, 1997.

[14] L. X. Wang. *Adaptive Fuzzy Systems and Control*. Prentice Hall, Englewood Cliffs, New Jersey, 1994.

[15] L. Wang and J.M. Mendel. Fuzzy basis functions, universal approximation, and orthogonal least-squares learning. *IEEE Trans. on Neural Networks*, 5:807–14, 1992.