Fuzzy Modeling for HLA Typing

G.B. Ferrara[1,2] F. Masulli[3,4] C. Pera[2] S. Rovetta[5,4] R. Sensi[5,4]

[1] Dept. of Biology, Oncology and Genetics, Univ. Genova, Largo R. Benzi 10, 16132 Genova Italy
[2] Advanced Biotechnologies Center, Largo R. Benzi 10, 16132 Genova Italy
[3] Dept. of Computer Science, Univ. Pisa, Corso Italia 40, 56125 Pisa Italy
[4] National Institute for the Physics of Matter, Via Dodecaneso 33, 16146 Genova Italy
[5] Dept. of Computer and Information Science, Univ. Genova, Via Dodecaneso 35, 16146 Genova Italy

Abstract. Oligonucleotide microarrays are able to perform a large quantity of simultaneous experiments, resulting in a digital image to be analyzed. We describe the image analysis step of a system designed for HLA typing based on microarray technology and employing an adaptive fuzzy system which can learn from user hints. The fuzzy modeling approach allows using the life scientist's language and concepts to describe and classify the probe activations in a natural way, and allows robust interactive image filtering thanks to the adaptive behaviour of the fuzzy system.

1 Introduction

Oligonucleotide microarrays [2] are able to perform a large quantity (100-10000) of simultaneous experiments. Each experiment corresponds to a given oligonucleotide probe (a DNA strand of 20-30 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).

We are developing a Decision Supporting System for HLA typing [9] using the oligonucleotide microarray technology. 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 is the problem of matching the HLA system of donor and receiver.

The main constituents of the system we are developing 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 genes of the HLA (Human Leukocyte Antigens) system which data base is available at http://www.ebi.ac.uk/imgt/hla/ and is in continuous up-dating. The task is the selection of

strings of ologonucleotides, of about 20 bases, able to discriminate groups of alleles (in high or low resolution). The ordered list of probes corresponds to the codes associated to groups of alleles to be discriminated.

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 redundance.

The DNA Microarray hybridation measurement subsystem is devoted to classify 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.

In this paper, after a sketch of the hardware and software environments (Sect.2), we will describe the DNA microarray hybridation measurement subsystem. As reported in the literature [5, 4] 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.

2 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 spot 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, redundance plays a relevant role. In facts, 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.

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

3 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.

3.1 Spot activity evaluation

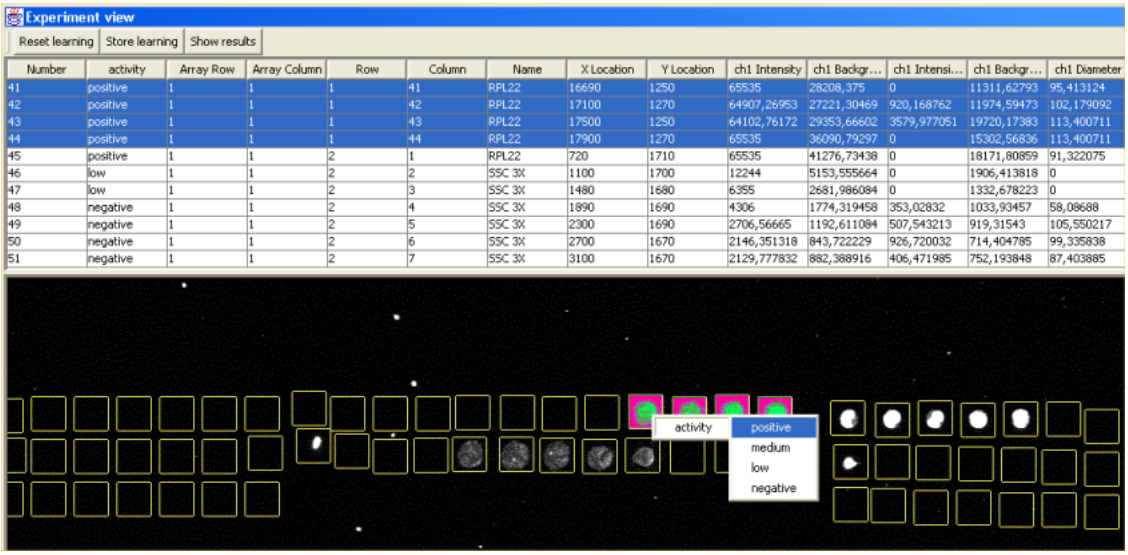
Concerning the first sub-problem, on the basis of the features measured by the scanner's driver, we define the following classes of spot activaction: *Positive*, *Intermediate*, *Negative* and *Outlier*. Each class is modelled as a fuzzy set using the adaptive fuzzy system described in the Appendix A. Intermediate fuzzy set can possibly be split into more fuzzy sets, in order to model different intermediate activaction levels of probe hybridization.

In the assessment of the overall DNA microarray genotyping system, the learning capability of the fuzzy system adopted is exploited to automatically model the shapes of the membership functions on the basis of a set of examples proposed by the user through the interactive graphical interface.

Fig. 2 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. Spots with low membership to each assigned class will be rejected (i.e., assigned to a *Reject* class). Outlier and Reject spots will not be considered in the following *fusion* step. The user can accept the classification, or else he/she can either prepare a new training set, or explicity change the membership class of each spot.

3.2 Classification fusion

The *Fusion Module* allows the user to integrate the activities 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.



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

Figure 2: 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 colour of their contours is related to the class to which each spot has the highest membership.

4 Discussion and conclusions

We have described parts of a Decision Supporting System which is being developed for the task of HLA typing. The subsystem presented classifies the spots on a DNA 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 proposed approach to HLA typing based on fuzzy modeling provides several advantages. In particular, it allows using the life scientist'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 adaptive behaviour of the fuzzy system.

A The adaptive fuzzy system

Fuzzy Logic Systems with *singleton* fuzzification, *max-product* composition, *product inference* and *height* defuzzification can be represented as [8]

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

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.)

It is worth noting that the FLS with universal function property studied by Mendel and Wang [11], 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.

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:

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

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

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

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

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

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 [6] 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 [7], 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

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

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:

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

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 [3, 10]:

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

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

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

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] 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.

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

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

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

[7] F. Masulli. Bayesian classification 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 Scientific.

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

[9] C. Pera, L. Delfino, A. Morabito, A. Longo, L. Johnston-Dow, C.B. White, M. Colona, G.B. Ferrara. HLA-A typing, *Tissue Antigens*, 50:372–379, 1997.

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

[11] 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.