A fuzzy approach to image analysis in HLA typing using oligonucleotide microarrays

G.B. Ferrara[a, b], L. Delfino[b], F. Masulli[a, c, *], S. Rovetta[a, d], R. Sensi[a]

[a]*INFM-Istituto Nazionale per la Fisica della Materia, Via Dodecaneso 35, 16146 Genova, Italy*
[b]*IST-Istituto Nazionale per la Ricerca sul Cancro, Largo R. Benzi 10, 16132 Genova, Italy*
[c]*Dipartimento di Informatica, Università di Pisa, Via F. Buonarroti 2, 56125 Pisa, Italy*
[d]*Dipartimento di Informatica e Scienze dell'Informazione, Università di Genova, Via Dodecaneso 35, 16146 Genova, Italy*

Abstract

The human leukocyte antigen (HLA) region is a part of genome which spans over 4 Mbases of DNA. The HLA system is strongly connected to immunological response and its compatibility between tissues is critical in transplantation. We have developed an application of oligonucleotide microarrays to HLA typing. In this paper, we present a method based on a fuzzy system which interactively supports the user in analyzing the hybridization results, speeding-up the decision process moving from raw array data obtained from the scanner to their interpretation (genotyping). The two-level procedure starts with evaluation of spot activity, then it estimates probe hybridization levels from activity levels. The method is designed for being readily usable by the biologist, by adopting fuzzy linguistic variables which are familiar to the user and by featuring a standard and complete graphical interface.
© 2004 Elsevier B.V. All rights reserved.

* Corresponding author. Istituto Nazionale per la Fisica della Materia, Università di Pisa, Via Dodecaneso 35, Genova 16146, Italy. Tel.: +39 010 535 6604; fax: +39 010 353 6699.

 E-mail address: masulli@di.unipi.it (F. Masulli).

1. Introduction

The major histocompatibility complex, known as human leukocyte antigen (HLA) region, is a part of human genome which spans over 4 Mbases of DNA on the short arm of chromosome 6 and consists of a large number of immunologically relevant genes. HLA class I and class II regions encode cell surface glycoproteins involved in the recognition by T cells [11]. One of the hallmarks of HLA complex is the extensive polymorphism of its loci. The number of HLA alleles reported in the last decade has risen at a rapid rate. More than 1700 HLA allelic variants have been described to date. [1]

The HLA system is strongly connected to immunological response. In transplantation, the match between donor's and receiver's HLA is critical for *histocompatibility* (compatibility between tissues). Therefore, characterization and identification (*typing*) of HLA is crucial for transplantation, as well as for antigen presentation, autoimmune disease and many others areas of clinical interest [11]. Among the molecular methodologies, DNA microarray technology can provide a feasible and reliable approach for HLA typing.

Oligonucleotide microarrays [3] make it possible to perform a large quantity (even thousands) of simultaneous experiments. Each experiment corresponds to a given oligonucleotide probe, a DNA strand of 20–30 bases which selectively combines with a complementary sequence in the target RNA sample (this process is termed *hybridization*). The probes are affixed to specific positions (spots) on the surface of a glass substrate, organized as an array. The target is fluorescently labelled, therefore a fluorescence measurement by laser scanning gives information about the amount of RNA hybridized at each *spot*, or specific location on the chip.

The use of microarray technology in HLA typing is promising [6,18] but to date not yet widespread. However, the high throughput provided by this method allows the collection and analysis of thousand of single nucleotide polymorphism in parallel [3,6]. This spot classification task on the basis of the microarray images is complex and can be very time consuming.

Soft computing and artificial intelligence techniques are routinely proposed for the analysis of microarray data (reviews can be found for instance in [9,13,17,20]). The use of a fuzzy logic system is well matched to the inherent uncertainty in data obtained by measurement of physical phenomena. Nevertheless, while the most typical application of soft computing to microarray data aims at measuring *gene expression levels*, which are inherently continuous-valued variables, the present work addresses a less typical microarray problem. In our case, the aim is to model a situation which (in the ideal setting) should feature binary variables, indicating hybridization. Fuzzy modeling accounts for uncertainty in the representation of these quantities. The system interactively supports the user in analyzing the hybridization results, speeding-up the decision process moving from raw array data obtained from the scanner to their interpretation, i.e., genotyping.

In this paper, after a description of our approach to HLA typing with microarray (Section 2), we will describe in Section 3 the system supporting the measurement of probe hybridization we have developed. In Section 4 we discuss a case study for validation of the proposed approach. Discussion and Conclusions are in Section 5. The Appendix presents the fuzzy basis functions network that we used in order to speed-up the spot activity labelling.

---

[1] HLA sequences are available from the IMGT/HLA Sequence Database at http://www.ebi.ac.uk/imgt/hla/, which is continuously updated.

2. Procedure for HLA typing with microarrays

Our oligonucleotide array approach for HLA typing involves a fluorescently labelled locus specific amplification of genomic DNA followed by hybridization with a panel of probes selected to detect a specific pattern of sequence motifs corresponding to an HLA allele.

The procedure starts with the design of a set of oligonucleotides, of about 15–20 bases, able to discriminate the alleles in high resolution. Each oligonucleotide probe will only anneal to sequences that match it perfectly, a single mismatch being sufficient to prevent hybridization under appropriate conditions [21]. In such as way, to each gene allele to be discriminated we associate a code made up by a list of positive or negative expected hybridization of the ordered set of probes.

Then, oligonucleotide probes are synthesized and spotted on the microarray (chip) using a spot printing robot, and then the microarray is hybridized using the target DNA to be analyzed. After hybridization and stringent washing the slide is scanned using a slide laser scanning system obtaining in such a way the fluorescence image of the microarray.

In Fig. 1 there are the images of two microarrays produced by a Packard-Bell Bioscience Division ScanArray 4000X. In the former image probes are affixed in the central area of the array and their pattern is repeated twice. We can distinguish spots with positive (lighter ones), intermediate, or negative (darker ones) activities and outliers (noise) spread mostly in the border areas. The latter image shows an enlarged detail of another microarray image, with higher presence of outliers.

HLA typing is then obtained by comparing the pattern of hybridization of the ordered set of probes and the codes associated to gene alleles in the probe design step.

The evaluation of probe hybridization is a complex task due to the presence of spots with intermediate activity that must be ascribed either to the choice of probes with too different *melting temperatures* [2] in the probe design step, or to other experimental problems 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. As a consequence, the binary linguistic variable *Probe Hybridization* (with {*Positive*, *Negative*} term set), must be obtained from the linguistic variable *Spot Activation* that can range in a term set containing *Positive* value, *Negative* value and one or more *Intermediate* values (the last ones usually corresponding to so-called *False Negative/False Positive* spots).

Using the available knowledge about the specific probe and the experimental conditions, *Intermediate* values of *Spot Activity* can be mapped into *Positive/Negative* values of the *Probe Hybridization*. Moreover, one can exploit the redundancies of the microarray (e.g., the repetitions of spots of the same probe, such as in the case in Fig. 1(a)) in order to obtain a more reliable estimation of *Probe Hybridization*.

3. Support to probe hybridization labelling

A typical HLA typing problem can need hundred of probes to be affixed to the microarray that will contain at least a double number of spots. Because of the large throughput typical of microarrays, evaluating probe hybridization by the approach described in Section 2 will be very time consuming and

---

[2] The melting temperature of a probe is the optimal temperature for its hybridization and depends on its basis. The quantity of RNA hybridized can increase/decrease if the hybridization experiment has been done at a temperature higher/lower that the probe's melting temperature.
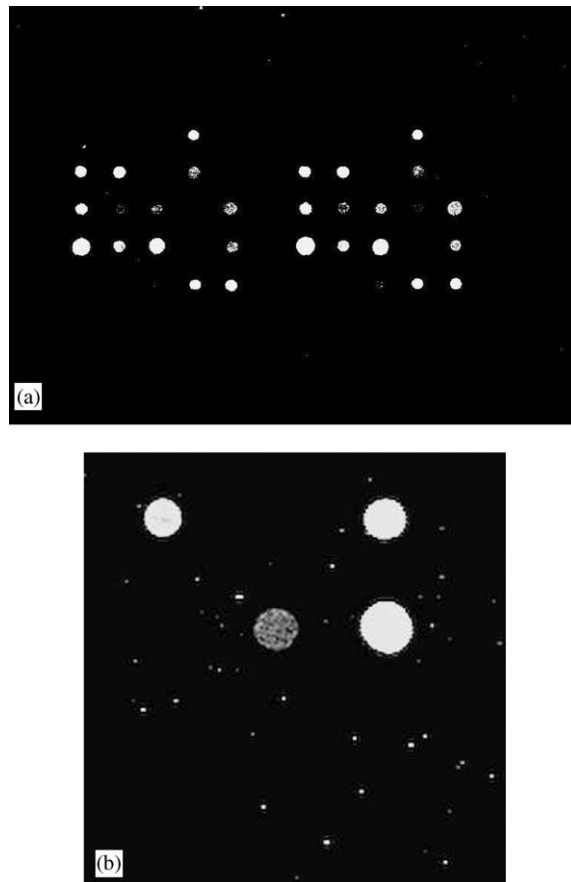
Fig. 1. Scanned images of two microarrays for HLA typing. (a) The probes are affixed in the central area of the array. One can distinguish spots with positive (lighter ones), intermediate, or negative (darker ones) activities and outliers (noise) spread mostly in the border areas. The experiment is described in Section 4. (b) There is an enlarged detail of another microarray image, with higher presence of outliers.

complex, computer-assisted analysis is of value in order to provide large-scale allele typing, improve data management, and streamline overall quality control processes.

A direct approach to computer-assisted labelling of *Probe Hybridization* can consist in the definition of a bank of (fuzzy) rules evaluating the probe hybridization on the basis of the image features obtained from the spots. But this approach is not easy, as an expert biologist can discriminate the hybridization level of a probe on the basis of the image produced by the scanner and of the nature of the probe itself, while s/he cannot obtain a reliable classification of spots using only spots' features. As a consequence, a machine learning approach trying to correlate spots' features (inputs) and the *Probe Hybridization* classification made by the biologist (labels) can be more fruitful in supporting the user's labelling task.

In principle, if the microarray shows a sufficient spot redundancy (e.g., if each probe has been spotted many times), a learning machine can help modeling the *Probe Hybridization* of each probe by evaluating its memberships to fuzzy sets (terms) *Positive* and *Negative* on the basis of the spot's sub-image features

(inputs) and the expert biologist's classification (labels). But, since usually a probe is spotted few times on the microarray, the spot redundancy is not sufficient to guarantee a reasonable generalization.

The approach we followed to design a system for the support to probe hybridization labelling is based on two sequential interactive steps:

(1) A *Spot Labelling Step* modeling the *Spot Activity* of a probe by evaluating the memberships of spots to terms (fuzzy sets) *Positive*, *Negative* and (one or more) *Intermediate*. Moreover, an additional term *Outlier* is considered grouping spots contaminated by noise. To this aim we use a learning machine that takes as input patterns the values of the spot's sub-image features and as labels the expert biologist's classification based on visual inspection of the spot sub-image.

(2) A *Probe Labelling Step* supporting the biologist in the association of *Spot Activity* values to those of *Probe Hybridization*. After the previous step, *Positive* and *Negative* values of *Spot Activity* are univocally associated to the same terms as *Probe Hybridization*. In the current step the biologist will associate to each probe the *Intermediate* values of *Spot Activity* to either *Positive* or *Negative* values of the *Probe Hybridization* on the basis of the available knowledge about the specific probe and the experimental conditions and exploiting the redundant spots on the microarray.

We have developed the system on a 500 MHz Pentium PC in *Sun Java 2*, providing it with an interactive graphical user interface making use of pure *Sun Java Swing* graphical components such as *tables*, *trees*, *menus* and *image panels* (see Fig. 2). In this way, the user has access to a familiar look-and-feel which makes user training much easier.

For each spot we considered its position onto the microarray and the following features computed on the spot's sub-image: average intensity, average background intensity, intensity standard deviation, diameter, circularity, and uniformity. All those data are a sub-set of those produced by the ScanArray Express software equipping the ScanArray 4000X. If necessary, when using a different scanner type, these features can be evaluated directly from the spot sub-image.

The learning machine used in the Spot Labelling step is a network of fuzzy basis functions (FBF) [16,22,23] that is a Mamdani fuzzy logic system [14] with *singleton* fuzzification, *max-product* composition, *product* inference and *height* defuzzification, equivalent to the ANFIS model [7]. A FBF network can learn its parameters from a labelled data set using a gradient descent procedure. A description of the FBF network and of its learning rules is presented in the Appendix and a Java implementation is available at http://mlsc.disi.unige.it/person/HLA/FBF/.

For each class to be modeled we use a FBF network whose task is the discrimination of that class against the remaining others on the basis of the considered spot sub-images features. We use the mean square error (MSE) as a cost function (empirical risk) to be minimized by the gradient descent procedure. In this way the FBF network estimates the posterior class conditional probability of any spot [2,15], that we can consider as the fuzzy membership to the class.

Fig. 2 shows an example of interaction in the *Spot Labelling Step*. The user starts by selecting a small set of spots for each *Spot Activity* class: *Positive*, *Medium*, *Negative*, and *Outlier*. In few seconds the FBF networks generalize the classification to all spots in the image. Labels are assigned to spots by using a Winner Take All (WTA) rule that associates the spot to the highest membership class. If the user accepts the classification, the step terminates. Otherwise, the user can either explicitly change the membership class of some spots and terminate the step, or prepare a new sample and retrain. Moreover, every time the user can mark a spot as *Reject*. Rejected spots will not be considered in the *Spot Labelling* and *Probe Labelling Steps*.

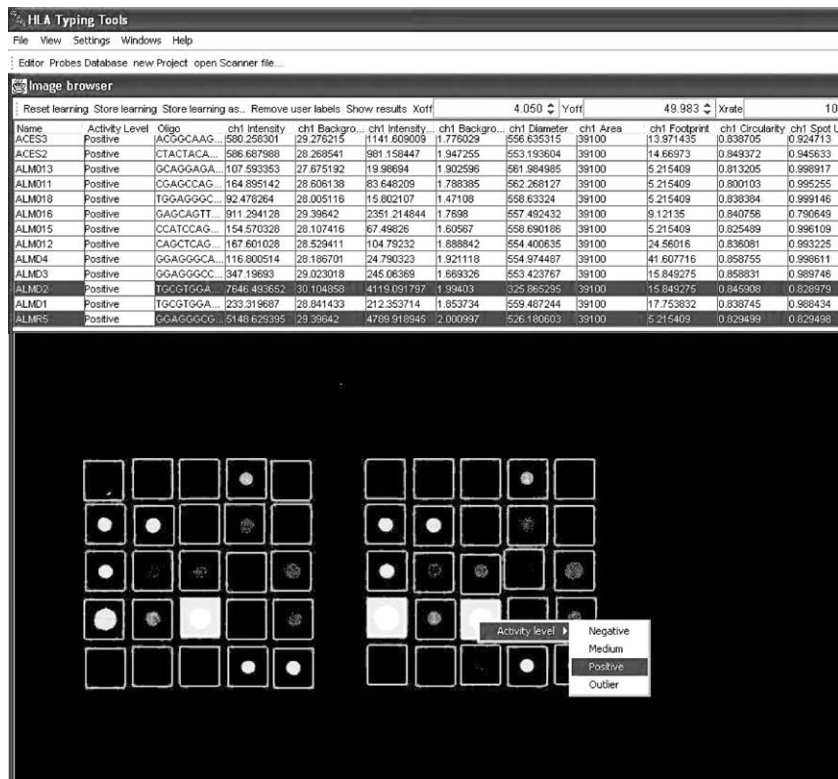| Name	Activity Level	Oligo	ch1 Intensity	ch1 Backgro...	ch1 Intensity...	ch1 Backgro...	ch1 Diameter	ch1 Area	ch1 Footprint	ch1 Circularity	ch1 Spot U
ACES3	Positive	ACGGCAAG...	580.258301	29.276215	1141.609009	1.776029	556.635315	39100	13.971435	0.838705	0.924713
ACES2	Positive	CTACTACA...	586.687988	28.268541	981.158447	1.947255	553.193604	39100	14.66973	0.849372	0.945633
ALM013	Positive	GCAGGAGA...	107.593353	27.675192	19.98694	1.902596	561.984985	39100	5.215409	0.813205	0.998917
ALM011	Positive	CGAGCCAG...	164.895142	28.606138	83.648209	1.788385	562.268127	39100	5.215409	0.800103	0.995255
ALM018	Positive	TGGAGGGC...	92.478264	28.005116	15.802107	1.47108	558.63324	39100	5.215409	0.838384	0.999146
ALM016	Positive	GAGCAGTT...	911.294128	29.39642	2351.214844	1.7698	557.492432	39100	9.12135	0.840756	0.790649
ALM015	Positive	CCATCCAG...	154.570328	28.107416	67.49826	1.60567	558.690186	39100	5.215409	0.825489	0.996109
ALM012	Positive	CAGCTCAG...	167.601028	28.529411	104.79232	1.888842	554.400635	39100	24.56016	0.836081	0.993225
ALMD4	Positive	GGAGGGCA...	116.800514	28.186701	24.790323	1.921118	554.974487	39100	41.607716	0.858755	0.998611
ALMD3	Positive	GGAGGGCC...	347.19693	29.023018	245.06369	1.669326	553.423767	39100	15.849275	0.858831	0.989746
ALMD2	Positive	TGCGTGGA...	7646.493652	30.104858	4119.091797	1.99403	325.865295	39100	15.849275	0.845908	0.828979
ALMD1	Positive	TGCGTGGA...	233.319687	28.841433	212.353714	1.853734	559.487244	39100	17.753832	0.838745	0.988434
ALMR5	Positive	GGAGGGCG...	5148.629395	29.39642	4789.918945	2.000997	526.180603	39100	5.215409	0.829499	0.829498

Fig. 2. The interactive user interface. Each row of the table corresponds to a spot and contains the values of its features, and other information, including the class of membership (*Positive*, *Medium*, *Negative*, *Outlier*, and *Reject*). On the bottom, on the scanned image, the squares represent the positions of spots, and the color of their contours indicate the class of membership. The experiment is described in Section 4.

Note that many outlier points in the microarray image are implicitly filtered out as they are outside the spots' sub-images. Concerning the outliers belonging to spots' areas, the hosting spot areas are grouped in the class *Outlier* during the previously described interactive learning procedure and, moreover, we add to this class also the spots with low membership to the other classes. The spots assigned to class *Outlier* will not be considered in the subsequent *Probe Labelling Step*.

As our main goal is the development of a system for computer-assisted analysis of scanned images of microarrays for HLA typing in order to speedup the usual manual labelling task, there are not any strict generalization or speed requirements for the learning machine to be used in *Spot Labelling Step*, but, obviously, the higher those performance indexes, the faster the whole HLA typing.

To this aim we have performed a model selection on FBF networks using a K-fold validation method [1,19] that is particularly suitable when only a small data set is available, as in the present application. We considered a data set of 50 vectors of spot features randomly extracted from a scanned image of a microarray for HLA typing. The data vectors are labelled with 3 classes *Spot Activity*: *Positive*, *Medium*, and *Negative*. The best generalization results, using 5 folders of 10 patterns each, have been obtained with FBF networks with 8 hidden units (i.e., 8 rules), and this is the architecture we selected for implementation in the system.

Concerning spot redundancy, there are two typical cases that arise from microarray design. The first one (*local redundancy*) is due to a constraint of some spot printing robots (e.g. the Packard-Bell Bioscience Division SpotArray 24) that 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. The second source of redundancy (*global redundancy*) is a microarray designer's shrewdness consisting in spotting the same probe in several regions of the slide, in order to prevent the effects of local experimental problems due e.g. to low-quality zones in hybridization process.

As already stated, the *Probe Labelling Step* supports the biologist in mapping the *Spot Activity* values already evaluated into the two values of the *Probe Hybridization*. To this aim the biologist exploits the spot redundancy using a choice of operators including maximum, minimum, averaging, and voting, in order to fuse the *Spot Activity* values corresponding to all instances of a given probe, and then he will exploit his knowledge about the probes and the experimental conditions in order to map *Intermediate* values of *Spot Activity* to either *Positive* or *Negative* values of the *Probe Hybridization*.

After the *Probe Labelling Step*, we obtain the typing of the target HLA allele by comparing the ordered list of *Probe Hybridization* levels obtained with the list of alleles' codes produced during probe design.

4. Case study: validation of the system

In this section, we describe an experimental validation of our procedure for HLA typing with microarrays and of our system assisting probe hybridization labelling. We used a small number of probes organized in two identical squares of 5×5 spots (see Fig. 1(a)).

A panel of 20-mer oligonucleotide probes was designed for identifying polymorphic positions located in exon 2 and exon 3 of HLA-A and B loci and in exon 2 of HLA-DRB1 locus (see Table 1). Each probe contained a $5'$ aminolink for immobilization chemistry and a 12-mer spacer, followed by the 20-mer hybridization sequence. The polymorphic sequence was situated near the center of each hybridization sequence.

Oligonucleotide probes were synthesized and spotted on an array by MWG Biotech Srl. The microarray was hybridized with single strand PCR product amplified from human genomic DNA.

The target DNA was previously HLA typed as A*0216/0301 HLA with an independent approach (high-resolution sequencing with a capillary sequencer, for different HLA loci) and was prepared as follows: double stranded PCR product first generated using locus specific primers and then purified to remove the excess primers. Single stranded DNA molecules were then generated from asymmetric PCR using one $5'$ Cy5-labelled primer as described in [21].

After hybridization and stringent washing the slide was scanned using a ScanArray 4000XL and the fluorescence image was analysed using the interactive system described in the previous section.

We expected to get positive hybridization with the following probes: ACES2, ACES3, ALMR5, ALMD2, ALM016, ALM017 and negative with the others. The analysis with the support of the interactive system leads to the expected results with few user interactions. Fig. 2 shows a screen-shot of the interactive interface. [3]

———————

[3] On the web site http://mlsc.disi.unige.it/person/HLA/DSS/ there is a detailed description of the interaction with the system.

Table 1
List of probes spotted in the HLA typing microarray

PROBE	Sequence	Array deploy
ACES2	5′ CTA CTA CAA CCA GAG CGA GG 3′	A1; A19
ACES3	5′ ACG GCA AGG ATT ACA TCG CC 3′	A2; A20
ALMR1	5′ G TAT TTC TAC ACC TCC GTG TC 3′	A3
ALM4R	5′ GGG ACC GGA ACA CAC GGA A 3′	A4
ALMR6	5′ TC ACA TCC ATG TCC CGG CC 3′	A5
ALMR2	5′ CAC TCA CAG ATT GAC CGA GTG 3′	A6
ALMR3	5′ GAC GGG CGC CTC CTC CGC 3′	A7
ALMR5	5′ GG AGG GCG AGT GCG TGG A 3′	A8
ALMD1	5′ TG CGT GGA CGG GCT CCG C 3′	A9
ALMD2	5′ TG CGT GGA GTG GCT CCG C 3′	A10
ALMD3	5′ GG AGG GCC GGT GCG TGG A 3′	A11
ALMD4	5′ GG AGG GCA CGT GCG TGG A 3′	A12
ALM012	5′ CAG CTC AGA TTA CCA AGC GC 3′	A13
ALM015	5′ C CAT CCA GAT GAT GTA TGG CT 3′	A14
ALM016	5′ GA GCA GTT GAG AGC CTA CCT 3′	A15
ALM018	5′ TGG AGG GCT GGT GCG TGG 3′	A16
ALM011	5′ CG AGC CAG AAG ATG GAG CC 3′	A17
ALM013	5′ GCA GGA GAG GCC TGA GTA TT 3′	A18
ALM014	5′ AGG TAT TTC TCC ACA TCC GTG 3′	A21
ALM017	5′ GAC ACG GAA TGT GAA GGC CC 3′	A22

The table shows the name assigned to probes, theirs bases sequence with DNA orientation (5′ and 3′) and their deployment position in the microarray.

5. Discussion and conclusions

We have described a system for assisting the biologist in the analysis of oligonucleotide microarray images for HLA typing [11]. Because of the large number of probes used in HLA typing computer-assisted analysis is of value in order to provide high-throughput allele typing, improve data management, streamline overall quality control processes. In Section 4 we have shown an experimental verification of the system.

The approach we followed is based on a fuzzy modeling of spot activity and the mapping of spot activity evaluations into the evaluation of the hybridization of related probes. The system interactively supports the user in analyzing the hybridization results. The spot activity evaluations made by the biologists on a sub-set of spots are generalized to the other spots by means of learning machines based on FBF networks [16,22,23] that relate the labels of spots given by the biologist with a set of features measured on the spots' sub-images. In this way we obtain a significant speeding-up of the decision process moving from raw array data obtained from the scanner to their interpretation, i.e. genotyping.

Concerning the set of features used as the input vectors of the learning machine, in the experiments reported in this paper we have used a subset of those produced by the ScanArray Express software equipping the ScanArray 4000X. Obviously, other, (if possible more specific) sets of features can be considered in its place to improve the classification results (see, e.g., [5]).

On the basis of the promising results obtained with the method described in this paper, we have started the development of a decision-support system for the full processing of HLA typing using the

oligonucleotide microarrays technology. Note that the proposed computer-assisted approach increases in relevance while the complexity of the typing task increases and hundred or thousand of spots have to be labelled.

Although (to the best of our knowledge) the complete system we are implementing is a novel realization, the design of a decision-support system is a natural choice in this application field. Computer assistance in decision making (expert systems, decision-support systems [24]) is especially widespread in medical practice.

As an example, related to applications of our work, in transplantation there are two main scenarios, related to the type of transplant. Transplantation of solid organs usually require fast access to information and donor–receiver compatibility. Decision-support systems in this scenario should enable fast decision making in ubiquitous computing environments [12]. Very close match between the immune systems is somewhat less critical in this case. In contrast, when it comes to bone marrow transplantation, even small mismatch can result in unacceptably increased risk [4]. In this case molecular technologies (based on Polymerase Chain Reaction, sequencing, or microarrays) have to be adopted. The methods described in this paper can be fruitfully exploited in this scenario.

Acknowledgements

Appendix

Fuzzy logic systems (FSL) with *singleton* fuzzification, *max-product* composition, *product inference* and *height* defuzzification can be represented as [16]

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

where $\overline{y}^l$ denote the center of gravity of the output fuzzy sets, $l = 1, 2, \ldots, M, \mathbf{x} = (x_i), i = 1, 2, \ldots, p$, are the input patterns, and $\phi_l(\mathbf{x})$ are called *fuzzy basis functions* and are given by

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

We can refer to those FLS as *fuzzy basis expansions* or *networks of fuzzy basis functions* (FBF network). [4]

It is worth noting that the FLS with universal function property studied by Mendel and Wang [23], 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

---

[4] The relationships between fuzzy basis expansions and other basis functions have been extensively studied in [10].

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. Moreover, without loss of generality, we can assume that the fuzzy membership functions are singletons, i.e., $\overline{y}_{ij} \equiv s_{ij}$.

Note that, the FBF network can be regarded as a feedforward connectionist system with one hidden layer whose units correspond to the fuzzy rules and it can be identified [14] both by exploiting the linguistic knowledge available (*structure identification problem*) or by using the information contained in a data set (*parameter estimation problem*), as we done in this paper.

As shown in [15], 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 nth 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 [8,22]:

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

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

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

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

References

[1] S. Amari, N. Murata, K.-R. Müller, M. Finke, H.H. Yang, Asymptotic statistical theory of overtraining and cross-validation, IEEE Trans. Neural Networks 8 (5) (1997) 985–996.

[2] F. Casalino, F. Masulli, A. Sperduti, Rule specialization in networks of fuzzy basis functions, Intelligent Automat. Soft Comput. 4 (1998) 73–82.

[3] R. Ekins, F.W. Chu, Microarrays: their origins and applications, Trends Biotechnol. 17 (1999) 217–218.

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

[5] A. Ghosh, S.K. Pal, (Eds.), Soft Computing Approach to Pattern Recognition and Image Processing, World Scientific, Singapore, 2002.

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

[7] J.S.R. Jang, ANFIS: adaptive-network-based fuzzy inference system, IEEE Trans. Systems Man Cybern. 23 (1993) 655–684.

[8] C.C. Jou, Comparing learning performance of neural networks and fuzzy systems, in: IEEE Internat. Conf. Fuzzy Systems, San Francisco, IEEE, New York, 1993, pp. 1028–1033.

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

[10] H.M. Kim, J.M. Mendel, Fuzzy basis functions: comparisons with other basis functions, IEEE Trans. Fuzzy Systems 3 (1995) 158–168.

[11] I. Klein, A. Sato, The HLA system—second of two parts, New Engl. J. Med. 343 (11) (2000) 782–786.

[12] A. Krause, D. Hartl, F. Theis, M. Stangl, K.E. Gerauer, A.T. Mehlhorn, Mobile decision support for transplantation patient data, Internat. J. Med. Inform. 73 (5) (2004) 461–464.

[13] A. Kuklin, A. Petrov, S. Shams, Quality control in microarray image analysis, G.I.T. Imaging & Microscopy 1 (2003) 2–3.

[14] C.C. Lee, Fuzzy logic in control systems: fuzzy logic controller. I, IEEE Trans. Systems Man Cybern. 20 (1990) 404–418.

[15] F. Masulli, Bayesian classification by feedforward connectionist systems, in: F. Masulli, P.G. Morasso, A. Schenone (Eds.), Neural Networks in Biomedicine—Proc. Advanced School of the Italian Biomedical Physics Association—Como (Italy) 1993, World Scientific, Singapore, 1994, pp. 145–162.

[16] J.M. Mendel, Fuzzy logic systems for engineering: a tutorial, Proc. IEEE 83 (1995) 345–377.

[17] A. Narayanan, E.C. Keedwell, B. Olsson, Artificial intelligence techniques for bioinformatics, Appl. Bioinformatics 1 (4) (2002) 191–222.

[18] C. Pera, L. Delfino, A. Morabito, A. Longo, L. Johnston-Dow, C.B. White, M. Colonna, G.B. Ferrara, HLA-A typing, Tissue Antigens 50 (1997) 372–379.

[19] M. Stone, Cross-validatory choice and assessment of statistical predictions, J. Roy. Statis. Soc. Ser. B 36 (1974) 111–147.

[20] F. Valafar, Pattern recognition techniques in microarray data analysis: a survey, Ann. NY Acad. Sci. 980 (2002) 41–64.

[21] R.B. Wallace, J. Shaffer, R.F. Murphy, J. Bonner, T. Hirose, K. Itakura, Hybridization of synthetic oligodeoxyribonucleotides to Φ_χ 174 DNA: the effect of a single base pair mismatch, Nucleic Acid Res. 6 (1979) 3543–3557.

[22] L.X. Wang, Adaptive Fuzzy Systems and Control, Prentice-Hall, Englewood Cliffs, NJ, 1994.

[23] L. Wang, J.M. Mendel, Fuzzy basis functions, universal approximation, and orthogonal least-squares learning, IEEE Trans. Neural Networks 5 (1992) 807–814.

[24] H.-J. Zimmermann, Fuzzy Set Theory and its Applications, 3rd ed., Kluwer Academic Publishers, Dordrecht, 1996.