A Hybrid Evolutionary Strategy to Optimise Early-Stage Cancer Screening

Grazziela P. Figueredo
The Advanced Data Analysis Centre
The School of Computer Science
The University of Nottingham, UK
Grazziela.Figueredo@nottingham.ac.uk

Peng Shi
ASAP Research Group
School of Computer Science
The University of Nottingham, UK
psxps4@exmail.nottingham.ac.uk

Andrew J. Parkes
ASAP Research Group
School of Computer Science
The University of Nottingham, UK
Andrew.Parkes@nottingham.ac.uk

Keith Evans
The Digital Research Service
The University of Nottingham, UK
Keith.Evans@nottingham.ac.uk

Jonathan M. Garibaldi
IMA Research Group
School of Computer Science
The University of Nottingham, UK
Jon.Garibaldi@nottingham.ac.uk

Ola Negm
School of Medicine
The University of Nottingham, UK
Faculty of Medicine
Mansoura University, Egypt
Ola.Negm@nottingham.ac.uk

Patrick J Tighe
Immunology School of Life Sciences
The University of Nottingham, UK
Paddy.Tighe@nottingham.ac.uk

Herbert F. Sewell
Immunology School of Life Sciences
The University of Nottingham, UK
mbahs3@exmail.nottingham.ac.uk

John Robertson
School of Medicine
The University of Nottingham, UK
John.Robertson@nottingham.ac.uk

***Abstract*—Current methods to identify cut-off values for tumour-associated molecules (antigens) discrimination are based on statistics and brute force. These methods applied to cancer screening problems are very inefficient, especially with large data sets with many antigens being investigated. There is a long wait to produce outcomes for clinicians, high performance computing is required, the best solution is not likely to be achieved and scalability is an issue. Cancer research is therefore limited in the number of antigens the methods can efficiently handle, and good solutions are potentially missed. We present an alternative evolutionary method based on Genetic Algorithms and Harmony Search to accelerate clinical research and to enable the consideration of a larger number of candidate antigens during the designing of the screening. We show that compared to the traditional methodology employed by clinicians, our approach is able to produce better results in a timely manner.**

***Index Terms*—genetic algorithms, multiple objectives, composite chromosome, Monte Carlo, harmony search, cancer-screening, colorectal cancer, breast cancer, lung cancer**

I. Introduction

The immune system, which protects us from microbes, also mounts a response to molecules overproduced and released by cancer cells within a tumour. This response includes the generation of auto-antibodies to these tumour-associated molecules. Clinicians have evidence that detecting these auto-antibodies in the blood of patients can provide a route to improved methods for early detection of tumours [1]–[8]. The approach only requires a blood sample, avoiding invasive techniques. Auto-antibodies are tested to multiple tumour molecules (antigens) simultaneously in a microarray assay. The signal intensity of the auto-antibodies response is subsequently stored in a numerical dataset - anonymised as needed. For each antigen, a cut-off is established computationally to distinguish those patients likely to be cancer positive and those who are cancer negative (namely, controls in clinical research exercises).

In our research, it is expected that cancer patients will have upregulated responses (higher numerical values) to auto-antibodies compared to non-cancer patients. Comparison to cut-offs based upon examination of known cancer patients and normal controls therefore allows the identification of samples positive for tumour-molecule auto-antibodies, indicating that further investigation is required. For cancer types such as lung, colorectal and breast cancer, there is an urgent need to identify relevant tumour markers from the blood showing high sensitivity (high rate of detection of true cancer positive patients) and specificity (low rate of false positives on true cancer negative patients) for its early detection. Such detection will enable the creation of tests that would improve clinical outcomes for patients. That means less aggressive treatments for early stage disease, opportunity for targeted chemoprevention, lower incidence of advanced stage and improved survival.

Current detection of markers is performed by medical researchers via brute-force, combinatorial methods coupled with random number generators, known by clinicians as Monte Carlo approaches. These approaches are inspired by traditional

Monte Carlo Methods [9], [10]. Auto-antibodies responses to antigens for cancer and control patients are processed from the microarray raw data and cut-offs for discrimination are established by the Monte Carlo approach. The cut-offs determine the sensitivity (SENS) and specificity (SPEC) of the panel of antigens investigated and therefore their suitability for early-stage cancer screening. In the screening process, a patient is labelled as cancer positive if their microarray values are above the cut-off for at least one antigen within the panel. Conversely, to be classified as negative, all values need to be below the cut-off thresholds for all antigens. This problem therefore, does not fall into classical machine learning classification (although might be regarded as a very specific kind of decision tree).

There are several issues with the existing methodology employed to identify the best set of antigens for the screening test: (a) The optimal values for cut-offs are not guaranteed or are too computationally expensive to achieve. This means that clinicians might be missing on more suitable antigens (or combinations of antigens); (b) the analytical approach is laborious and resource intensive, requiring the use of the high performance computing facilities and the assistance of experts in computer science; (c) the search mechanism is ineffective, as millions of sub-optimal solutions are generated, evaluated and discarded during the analysis; (d) there is a long wait from researchers until results are produced; (e) the method is mostly limited to investigations of 40 to 50 antigens, as it is not scalable within feasible timelines.

The use of an alternative, more effective approach is therefore necessary and timely. In this paper, we investigate the use of Genetic Algorithms (GA) [11] combined with Harmony Search (HS) [14] (GAHS) to achieve optimal screening solutions. We have the following main research objectives: (1) To improve cut-off optimisation using our approach (2) To compare our results to the existing Monte Carlo techniques. (3) To replace effectively and reliably the current methodology. To achieve these, we investigate five different data sets. We show that our approach obtains better solutions much faster. We also show that for a problem where the data set is larger (58 antigens), the Monte Carlo approach would take more than one hundred days to produce the results, which is not a feasible timeline for clinical research.

II. PROBLEM DESCRIPTION

Auto-antibodies from a set of patient's blood samples are tested to various tumour-associated antigens simultaneously in a microarray assay. The signal intensity of the auto-antibodies response is processed, the microarray signal intensity is converted into a number and stored in a dataset, with a format shown in Table I. From the larger set of antigens investigated with microarray technology, it is of interest to select a subset of antigens with optimal cut-offs separating cancer and control patients to form a panel to be used for early-cancer detection.

Cancer researchers are interested in four aspects of the data: (1) to determine a subset (panel) of antigens able to distinguish with high SENS and SPEC cancer patients from controls in an initial training data; (2) to determine the optimal cut-off for each antigen within the panel for the separation between the two categories of patients; (3) to maximise the pair (SENS, SPEC) for the patient cohort; and (4) when possible, to identify multiple panels that achieve maximum SENS and SPEC to evaluate which panel will allow for better reproducibility of results in a validation set.

TABLE I
TAB:MICROARRAY DATA EXAMPLE

Patient	**Antigen 1**	**Antigen 2**	**...**	**Antigen n**	Class
1	0	0.34	0.8	...	cancer
2	0	0.55	0.22	...	control
3	0	0.84	0.22	...	control
...	...	...	...	...	...
n	0	0.11	0	...	cancer

In this research, we are assisting with the training stage of the research to produce suitable panels and cut-offs for the validation stage. From the training data, it is expected that cancer patients will have higher numerical values for auto-antibodies responses compared to non-cancer patients. Those antibodies that do not produce different responses between at least a sub-group of patients should be discarded and not make part of a screening panel. A panel of antigens is necessary, as it is clinically unlikely that a single antibody will produce similar responses for all patients.

A. Screening Process

To screen and subsequently to classify patients as cancer positive or negative (controls), cut-off values for each antigen combination are determined and their SENS and SPEC are calculated. The objectives are to maximise (SENS, SPEC) for a panel and to obtain a subset of antigens with the desirable classification performance. Figure 1 shows an example of an antigen panel with n antigens and their corresponding cut-offs. In the figure, 9 patients are considered; four patients are controls (orange circles) and five patients have cancer (blue circles). The cut-offs are represented by the red lines. For each patient screening, the following rules are applied, given a set of cut-offs defined for each antigen:

- If the patient response to the antigen is higher than the cut-off for **at least one antigen**, then the patient is labelled as *cancer positive*.
- If the patient response to **all** antigens is lower than the cut-off for all antigens, then the patient is labelled as *cancer negative*.

The SENS and SPEC are therefore calculated as follows:

- If a cancer patient is *True Positive* (above threshold) for at least one antigen then the number of *True Positive* patients on the screening is incremented by one. Otherwise, *False Negative* is incremented.
- If the control patient is *True Negative* (below the threshold) for all antigens then the number of *True Negative* patients on the screening is incremented by one. Else, *False Positive* gets incremented.

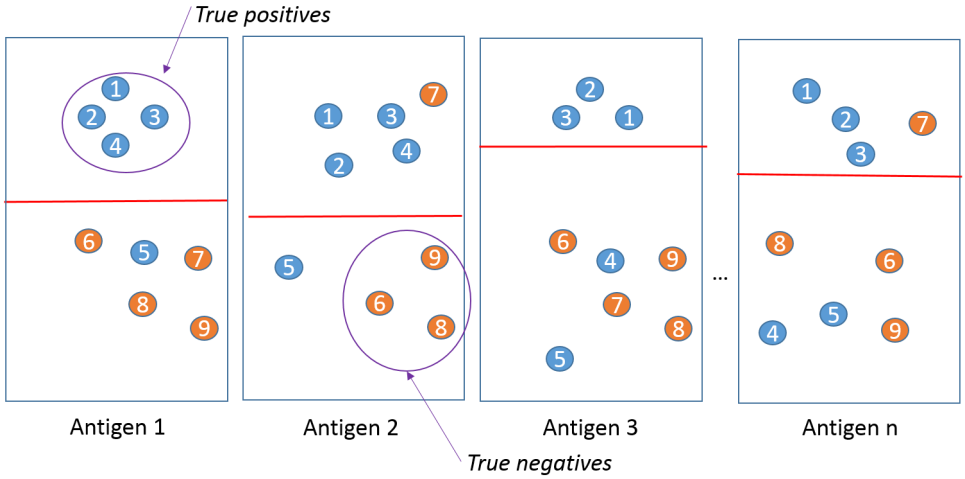

Fig. 1. Cancer screening process. In the figure, each rectangle contains the antigens responses for a set of patients. The blue circles represent the cancer patients; the orange circles are the control patients. For each antigen a cut-off (red horizontal line) needs to be established. Those values for each patient above the red line are classified as cancer positive; values below are negatives. The true positive patients are the blue circles above the red cut-off for at least one antigen. True negative patients are the orange circles (control patient values) below the red cut-off for all antigens

It is worth noting that the process described above is used for patient screening, which is different from patient classification. Classical machine learning methods for classification are therefore not suitable for this problem.

III. Multi-Objective Optimisation (MOO) Formulation

As usual in supervised machine learning, the input data is a set of tuples of class and feature values; Specifically, each person p, in the data set, has a tuple (y_p, x_{pa}) with meaning:

- Label: $y_p = 1$ if p has cancer, and 0 otherwise
- Data: x_{pa} is the strength of the response of person p on antigen a.

A candidate solution for a screening system, consists of

- Selection vector: $S_a = 1$ if antigen a should be included in the screening panel, and 0 otherwise
- Cut-offs: C_a is the cutoff or threshold for each antigen a. Only the cut-offs for the selected antigens, with $S_a = 1$, are relevant.

To compute the fitness functions we just need formulas for the sensitivity and specificity. For convenience, let us define z_{pa} to be 1 if and only if person p exceeds the cut-off for antigen a:

$$\begin{aligned} z_{pa} &= 1 \text{ if } x_{pa} \geq C_a, \\ &= 0 \text{ otherwise} \end{aligned} \tag{1}$$

and also define z_p to be 1 if and only if person p is screened as potential cancer, that is, exceeds the cut-off for at least one antigen a:

$$\begin{aligned} z_p &= 1 \text{ if } \left(\sum_a z_{pa} \right) \geq 1, \\ &= 0 \text{ otherwise} \end{aligned} \tag{2}$$

The sets of True/False Positives/Negatives are then simply

$$TP = \{ p \mid y_p = 1 \ \& \ z_p = 1 \} \tag{3}$$
$$FP = \{ p \mid y_p = 0 \ \& \ z_p = 1 \} \tag{4}$$
$$FN = \{ p \mid y_p = 1 \ \& \ z_p = 0 \} \tag{5}$$
$$TN = \{ p \mid y_p = 0 \ \& \ z_p = 0 \} \tag{6}$$

and then

$$\text{Sens} = |TP| \ / \ (\ |TP| + |FN| \ ) \tag{7}$$
$$\text{Spec} = |TN| \ / \ (\ |TN| + |FP| \ ) \tag{8}$$

Note that the denominator of Sens, $|TP| + |FN|$ is just the number of positives in the data set, and so a constant factor (for a given data set). Similarly, for the denominator in Spec. (This is helpful, as it means the problem will be a linear optimisation problem.) Furthermore, note that if changes to the cut-off C_a do not change its comparison to any measured response x_{pa} then the values of z_p also do not change and therefore they do not affect the objectives. Hence, if desired, the continuous spectrum of values for C_a could be replaced with a discrete set of representatives from each of the ranges of values defined from the data x_{pa}. Jointly maximising Spec and Sens can therefore be regarded as a (discrete linear) bi-optimisation problem.

Generally, we could also want to optimise (reduce) the number of antigens that are used. Hence, for example, there could be a third objective, to be minimised:

$$\text{Cost} = \sum_a w_a S_a$$

where w_a are some costs associated with performing the test for antigen a. Overall, this now gives a 3-objective (discrete linear) optimisation problem. In this paper, however, we do not address the cost problem; instead we just limit the number of antigens used and focus on optimising a given combination

of SPEC and SENS rather than doing the full multi-objective version.

A. Single-Objective Case

In this paper, for simplicity, we do not address the full multi-objective version but instead focus on an important special case with a single objective, but that will illustrate the potential of the methodology.

For the (preliminary) studies in this paper, the fitness (to be maximised) of the individual is simply given as:

$$fitness \quad = \quad \text{Min}(\ \text{SENS}\ ,\ \text{SPEC}\ ) \tag{9}$$

This fitness is also equivalent to minimisation of $\text{Max}((1 - \text{SENS})), (1 - \text{SPEC}))$ and so is essentially a form of the Chebyshev scalarisation method [12], [13]. This is an approach to multi-objective optimisation using just a single reference (target) point of $\text{SENS} = \text{SPEC} = 1$, and simply uses the maximum, over the different axes, of the distances from target. It hence drives the solutions towards increasing SENS and SPEC with equal weight in a more focussed fashion than simply using their direct sum. In practical medical situations SENS and SPEC might not be of equal importance. The fitness could therefore easily be adjusted to take account of this by using standard Chebyshev with a different reference point. E.g. by maximising $\text{Min}(\text{SENS} - a, \text{SPEC})$, for various values of a, or similar; this is likely to give a more focussed search than by using the standard weighted sum method.

This fitness is employed in both our Monte-Carlo studies in the next section, and the hybrid evolutionary approach in Section V.

IV. THE MONTE CARLO APPROACH

For the Monte Carlo (with brute force) approach currently employed by clinicians in [5], [8], fixed sizes of panels of interest are established. Normally, panels with 6, 8, 10, 12 and 15 antigens are investigated. All possible combinations of antigens from the microarray data are produced and the panels' effectiveness in distinguishing cancer and control patients are assessed. The number of possible combinations, considering N antigens is calculated by the standard "N choose n", or binomial coefficient $\binom{N}{n} = N!/(n! \cdot (N-n)!)$, where n is the size of the panel (in our case, 6, 8, 10, 12 or 15 antigens).

For each antigen a_i $(i : 1..n)$ the cut-off calculation is given by the Monte Carlo Approach. Fifty thousand random cut-off values are generated for each antigen in the panel. The random combinations for cutoff_i range between the minimum and the maximum value for antigen a_i. The panels and cut-offs produced are subsequently evaluated (according to their SENS and SPEC) and the best results are saved. The best result in our case is the panel that achieves maximum non-dominated solution for the pair (SENS,SPEC). The maximum non-dominated pair is chosen after processing all combinations, by looking at the Pareto front graph of solutions output by the Monte Carlo, as shown in Figure 2. The criteria for the best solution is defined for the purpose of output comparison.

In cases where N is a high number (larger than 40), the number of combinations reaches the order of 10^{10}, which requires very large processing power and is a very ineffective, laborious way of tackling the problem. The resource-intensive search strategy generates millions of sub-optimal solutions that have to be evaluated and subsequently discarded during the analysis. In addition, as there is no mechanism to improve the search of the cut-offs, their optimal values are not guaranteed. This means that clinicians are limited to partial searches, which might be missing out on more suitable panels.

The limitations of the current methodology, has led us to work with clinicians to improve their tool set for the results computation. We have therefore created a solution by

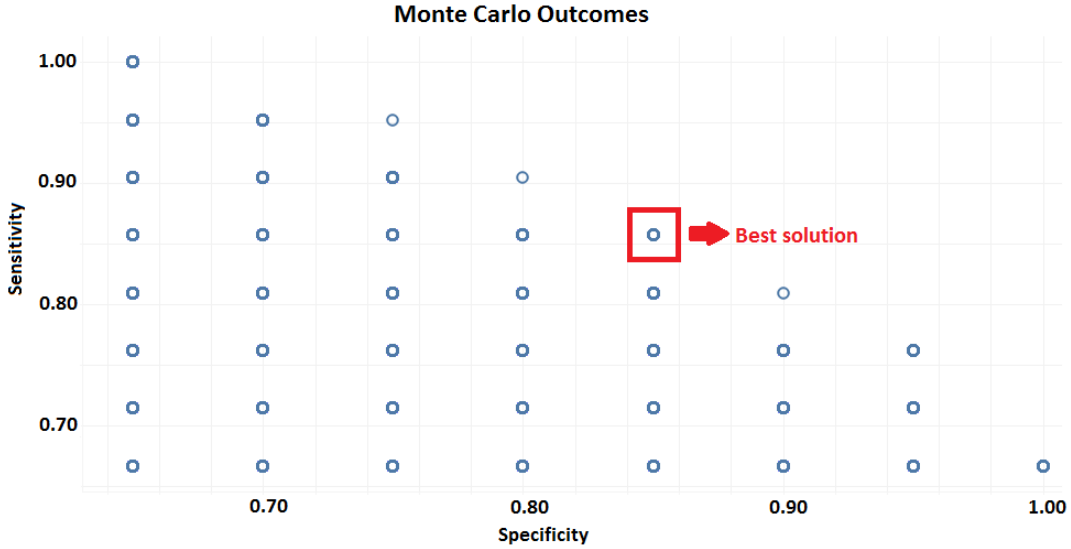

Fig. 2. Example of how the best solution is selected for the Monte Carlo Approach. The best pair (SENS,SPEC) is obtained by extracting the maximum non-dominated solution, according to the fitness function.

employing Genetic Algorithms (GA) coupled with Harmony Search (HS) to achieve better cut-offs and determine better possible panels, as further discussed next.

V. The Evolutionary Approach to the Solution

A GA hybrid with HS (GAHS) is designed to tackle the proposed problem. Given the data set of antigens and a specific panel size, the steady-state GA is implemented to produce the best antigen panel. In the GA, each chromosome represents a possible panel. The chromosome fitness is based on the SENS and SPEC calculation for a set of cut-offs, which are determined by the HS.

In this problem, the GA solution is encoded as a binary chromosome representation of length N, where N is number of input antigens from the microarray data. The uniform crossover operator is adopted. Single point and two points crossover operators are also implemented, however in our experiments the uniform crossover produces better results. The flip mutation operator is adopted. The whole procedure of the genetic algorithm is summarised as follows:

1) Initialise the population. A list of binary chromosomes $[g_1, g_2, \cdots, g_N]$ is randomly constructed, where $g_i = 1$ if this antigen is selected in the panel for further evaluation and 0, otherwise. Further validation of the individual is required to ensure that the sum of ones in the chromosome does not exceed the panel size (ps). A valid chromosome needs to satisfy the following constraint:
$$\left(\sum_{i=1}^{N} g_i\right) = ps \tag{10}$$
2) For each individual generated, calculate the fitness value of each cut-off chromosome obtained from using HS.
3) Construct a new population. Randomly split the whole population into pairs of parents. With the selected parents, two new individuals are produced through crossover and mutation operators. Each offspring chromosome is evaluated and corrected (as in step 1) to satisfy the required panel size constraint.
4) Update the population. Replace parents if better fitness values are achieved in the offspring. Replacement of individuals occurs in the following manner:

if *Min(Offspring1.fitness, Offspring2.fitness)* $\geq$ *Max(Parent1.fitness, Parent2.fitness)* **then**
  Keep offspring for the next generation;
end
else if *Max(Offspring1.fitness, Offspring2.fitness)* $\leq$ *Min(Parent1.fitness, Parent2.fitness)* **then**
  keep parents for the next generation;
end
else
  keep the better offspring and the better parent for the next generation;
end

5) Repeat the procedure from step 2 to step 4 until satisfying the termination criterion. Output the panel by decoding the best chromosome from the last generation.

The HS is also a population-based solving procedure, similar to GAs. Figure 3 shows an example of how HS determines the cut-off for 5 antigens. An individual in HS is a harmony and the whole population is defined as harmony memory (HM). Each position of harmony is a pitch. The HS procedure implemented in this paper is described next:

1) Initialise the harmony memory. Given a specific chromosome $[g_1, g_2, \cdots, g_n]$ from the GA, the HM $[x_1^m, x_2^m, \cdots, x_n^m]$ is randomly constructed. m is the harmony index and n is the antigen index. x_n^m is a random value from the pitch range $[anti_{min}^n, anti_{max}^n]$ if $g_n = 1$. This is the cut-off for further evaluation. Otherwise x_n^m is assigned as 0. As the example showed in Figure 3, the number of input antigens is 5 and the selected panel is a combination of antigens 2 and 3. In this example, the range of antigen 2 is [0, 1] and [1, 2] for antigen 3. The related SENS and SPEC of each harmony is calculated and stored in the memory.
2) Construct a new harmony $[x_1^{'}, x_2^{'}, \cdots, x_n^{'}]$. The harmony value is determined by two parameters, the harmony memory considering rate (HMCR) and pitch adjusting rate (PAR), which is similar to crossover and mutation rates in the GA, respectively. If $x_i^{'}$ is chosen from the HM, then any values of pitch i in the memory can be selected. This value could be further adjusted to a neighbour among the pitch range if $g_i = 1$. If $x_i^{'}$ is not chosen from the memory, it is assigned by a random value as step 1. As shown in the example in Figure 3, the value of pitch 2 in the new harmony is chosen from the memory. Data of the same pitch in the memory is $[0.28, 0.47, 0.83]$ and it is preliminary assigned as 0.47. This value could be further adjusted to a neighbour value, for example 0.4, or kept into the harmony. The value of pitch 3, conversely, is randomly assigned from [1, 2]. The remaining pitches are kept unchanged.
3) Update the harmony memory. Calculate SENS and SPEC of the harmony from step 2. Replace the worst harmony in the recent memory if both values are higher. Otherwise, eliminate the constructed harmony.
4) Repeat the procedure of step 2 and step 3 until satisfying the termination criterion. Return the harmony with the best SENS and SPEC values.

The GA and HS are inter-linked with each other to produce the desired search. On one hand, the output of the HS is dependent on the chromosomes constructed by the GA. On the other hand, the population of the GA could be efficiently updated if a better fitness value is obtained by the HS.

In future versions of our tool, the search for different configurations of the pair (SENS,SPEC) will be changed via input parameters, for instance to (a) maximise, only SENS; (b) maximise only SPEC; or to find a specific value for the

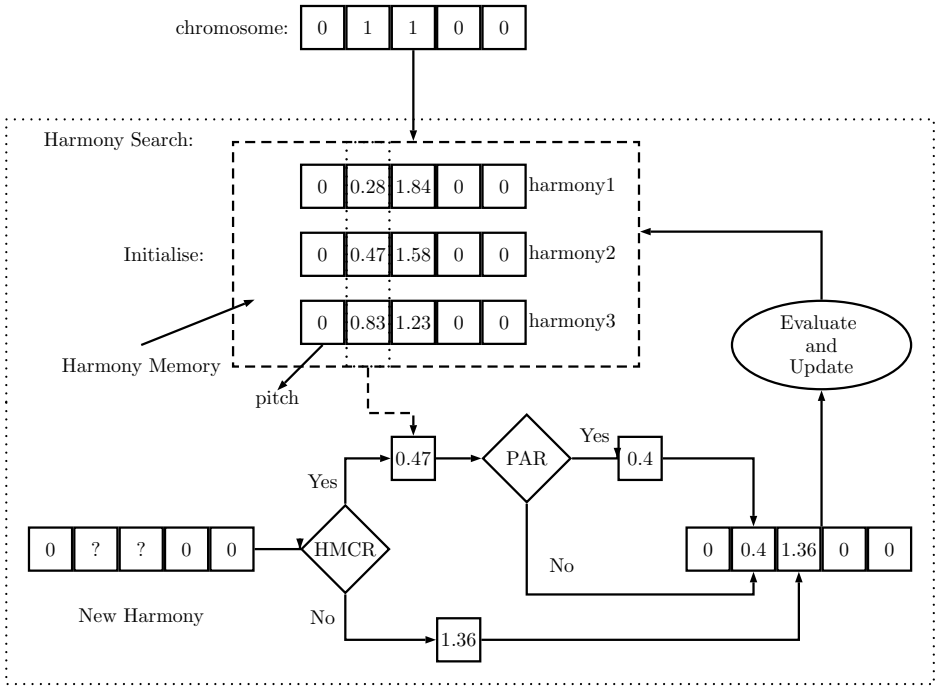

Fig. 3. Example of Harmony Search for 5 antigens

pair (e.g. SENS = 0.65 and SPEC = 0.85) depending on the clinicians research questions. In such cases, the comparison with the Monte Carlo is also performed by searching for the equivalent solution in the Pareto front.

VI. EXPERIMENTAL DESIGN

Experiments are run in five data sets, described in Table II. They vary in number of patients and in number of antigens considered. The data is obtained from clinicians from the School of Medicine at the University of Nottingham. It has been anonymised regarding the type of cancer, the origin of the patient samples and the name of antigens considered to protect patient privacy and the clinical research intellectual property.

TABLE II
DATA SETS INVESTIGATED

Dataset	Number of Patients	Total Antigens
Data 1	20	32
Data 2	41	32
Data 3	202	32
Data 4	263	32
Data 5	181	58

For each data set, panels of 6, 8, 10, 12 and 15 antigens are considered (6 and 10 for Data 1, 2, 3 and 4; and 6, 8, 10, 12 and 15 for Data 5). The objective is to find the panel with the highest SENS and SPEC.

The Monte Carlo approach was run on the University of Nottingham's High Performance Computer (HPC) platform, known and hereafter referred to as "Minerva" which has since been decommissioned. Minerva consisted of approximately 180 compute nodes, each of which in turn consisted of two 8 core x86 Intel processors and 32Gb of RAM; Minerva boasted a total compute capacity of around 40 TFLOPS with each user limited to 3 TFLOPS (or 192 cores) per compute job. In terms of software, a bespoke package was authored in Fortran to the 95 standard and parallelised using OpenMPI [1]. For the Monte Carlo, all combinations of antigens have been generated and fifty thousand random cut-offs were considered.

The GAHS [2] described in section V is implemented in Java (JDK 1.7) and all computations were performed on an Intel (R) Core (TM) i7 CPU with 3.2 GHz and 6 GB of RAM. The fitness value of of the candidate solutions is calculated as Function 9. The parameter settings of GAHS are given in Table III.

TABLE III
PARAMETERS OF GENETIC ALGORITHM AND HARMONY SEARCH

Genetic Algorithm		Harmony Search	
Population	50	Harmony Memory	100
Generations	100	Iterations	1000000
Crossover	Uniform Crossover	HMCR	0.7
Crossover Rate	0.5	PAR	0.2
Mutation Rate	0.03	Neighbour Gap	0.15

The parameter values from Table III are set based on preliminary experiments using Data 1 from Table II. In the future, we intend to investigate potential improvements in the results exhibited in Section VII by employing parameter tuning procedures.

[1] https://bitbucket.org/ADAC_UoN/montecarlo/src
[2] https://github.com/PengSimonShi/ProjectGAHS.git

VII. Results

In order to analyse the performance of GAHS compared with the Monte Carlo method, five data sets were selected for testing with the parameter settings described previously. The results obtained by Monte Carlo and GAHS are presented in Table IV. For GAHS, the table shows the best run achieved among 10 independent experiments. The highest values for sensitivity, specificity and times are highlighted in bold.

The Monte Carlo method is inherently paralleliseable due to the independence of each combination and as such, compute cost scaled approximately linearly with the number of possible combinations. The largest analysis feasible on Minerva was for Dataset 5 with 58 antigens and a panel size of 8, resulting in approximately 2 billion combinations; this job took 4 days on 192 cores an average of 500 million combinations per day. For comparison, a panel size of 10 for dataset 5 would have 52 billion combinations and take an estimated 104 days to complete. Theoretically, with full utilisation of Minerva's resources at 100% parallel efficiency, an analysis of dataset 5 with a panel size of 10 would still not have been feasible due to a wall time constraint of 7 days. We have therefore limited our processing time to 144 hours and collected the best results produced. However, for 15 antigens, satisfactory results were still not achievable within this time frame.

In general, SENS and SPEC obtained by GAHS is better than the Monte Carlo (MC) approach. Even though specificity for Data 5 obtained by GAHS is worse for 8, 10 and 12 antigens, the gap between these two approaches is of 8% at most. Importantly, the computational time of generating these results by GAHS is much shorter than MC. Efficiency is therefore the advantage of the proposed approach comparing with Monte Carlo. This means that clinicians can get insights and make decisions in a much faster manner. This will also allow for a larger volume of experiments to be run considering large numbers of antigens, which will hopefully improve clinical research.

With our method, we were also able to identify multiple panels with optimal SENS and SPEC. For instance, for Data 1 we identified 5 different panels with SENS = 1 and SPEC = 0.91. We expect that in the future, when we tackle the multi-objective problem, we can also perform a systematic comparison regarding the number and diversity of the panels found compared to Monte Carlo.

VIII. Conclusions and Future Work

In cancer research there is an urgent need to identify relevant tumour markers from the blood for its early detection. Methods for fast detection are important as they enable the creation of tests that can potentially improve clinical outcomes for patients. Early disease diagnosis means less aggressive treatments, opportunity for targeted treatments, lower chances of reaching advanced cancer stages and reduced mortality. Microarray research for early stage screening identifies tumour auto-antibody responses to antigens and relies on computational methods to determine thresholds that classify patients between cancer positives and negatives within a panel. Current screening algorithms to calculate those cut-offs and to determine the best panels are based on Monte Carlo methods and brute force. These methods perform poorly, as they do not guarantee that the best solution will be achieved; they are also very resource-intensive and take long periods of time to produce the desired outcomes. This means that early stage cancer research is being delayed by a less suitable computational tool kit.

In this paper we investigated a more effective alternative to address the problem based on evolutionary algorithms. We implemented a tool based on GAs and HS to search for the best panel of antigens with the highest sensitivity and specificity for patient classification. We wanted to address the research questions: (1) Can we improve cut-off optimisation using our approach? (2) How our results compare to the existing Monte Carlo techniques? (3) Can this approach replace effectively and reliably the current methodology? To answer these research questions, we investigated five different real-world data sets, collected by clinicians. For each data set, we ran both

TABLE IV
COMPARATIVE RESULTS BETWEEN THE ORIGINAL IMPLEMENTATION OF THE MONTE CARLO APPROACH AND THE EVOLUTIONARY APPROACH. NOTE THAT TIME IS EXPRESSED AS HH:MM::SS

Data set	Patients	Total Antigens	Panel Size	Methods' performance					
				Monte Carlo			GA + HS		
				Sensitivity	Specificity	Time	Sensitivity	Specificity	Time
Data 1	21	32	6	0.9	0.9	43:45:20	**1**	**0.909**	**1:51:23**
			10	0.9	0.8	72:18:21	**1**	**0.909**	**3:08:13**
Data 2	41	32	6	0.85	0.85	43:47:56	**0.864**	**0.895**	**3:09:35**
			10	0.85	0.8	72:20:47	**0.905**	**0.9**	**5:34:27**
Data 3	202	32	6	0.7	0.77	42:36:10	**0.825**	**0.828**	**11:48:10**
			10	0.78	0.8	72:31:18	**0.825**	**0.828**	**14:33:10**
Data 4	263	32	6	0.7	**0.8**	43:49:50	**0.791**	0.791	**22:25:24**
			10	0.73	0.75	72:20:10	**0.793**	**0.797**	**24:55:35**
Data 5	181	58	6	0.58	0.6	48:00:00 (aprox.)	**0.714**	**0.719**	**22:51:26**
			8	0.6	**0.8**	96:00:00 (aprox.)	**0.728**	0.739	**27:11:55**
			10	0.6	**0.8**	144:00:00	**0.745**	0.745	**24:43:49**
			12	0.6	**0.8**	144:00:00	**0.767**	0.767	**25:57:42**
			15	—	—	—	**0.767**	**0.767**	**27:6:36**

Monte Carlo and our approach, and collected the results for the best pair of sensitivity and specificity coupled with the time taken to obtain the desired outcome.

Our experiments showed that our approach in general obtains better solutions much faster. We also show that for a problem where the data set is larger (58 antigens), the Monte Carlo approach would take more than one hundred days to produce the results, which is not a feasible timeline for clinical research. GAHS is still limited however when compared to the Monte Carlo in some aspects. As the problem was converted to a single objective, we do not produce a Pareto front as does the Monte Carlo approach. Our current methodology therefore will be extended to address the multi-objective character of the problem. This will require adapting multi-objective evolutionary methods to handle the combined system of a GA for selecting antigens, and the HS for selecting cut-offs. Another issue that will need to be addressed is that the solution space is "multi-modal", in the sense that multiple panels can produce the same SENS and SPEC; the search should be enhanced to ensure that a diverse set of such modes are captured and provided to the clinicians.

Another aspect to be investigated by our tool is the potential of using combined antigens rather than a single one flagging up possible cancer. This would assist clinicians to understand the interplay between different antigens and potentially point to more robust screening approaches.

Future work could also study other machine learning mechanisms, but adapted to the screening problem and retaining the ease of understanding and explaining of the results to clinicians. In addition, it would be important to investigate whether the assumption of needing a simple decision method leads to a significant loss of classification quality. We will also assess the effectiveness of other alternative evolutionary and swarm methods and perform a comparative study with our current approach.

IX. Acknowledgements

Brute-force computational analyses were performed using the University of Nottingham's HPC "Minerva". Software tools were authored by members of the School of Computer Science together with the University of Nottingham Research Software Engineering (RSE) service and The Advanced Data Analysis Centre, Digital Research Service (DRS).

References

[1] Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin 2011; 61: 6990.

[2] The Independent UK Panel on Breast Cancer Screening The benefits and harms of breast cancer screening: an independent review. Lancet 2012; 380, 1723.

[3] Chapman C, Murray A, Chakrabarti J Thorpe A, Woolston C, Sahin U, Barnes A, Robertson J. Autoantibodies in breast cancer: their use as an aid to early diagnosis. Ann Oncol 2007;18:868-73.

[4] Zhong L, Coe SP, Stromberg AJ, Khattar NH, Jett JR, Hirschowitz EA. Profiling tumor-associated antibodies for early detection of non-small cell lung cancer. J Thorac Oncol. 2006;1(6):513-9.

[5] Robertson JFR, Chapman C., Cheung KL, Murray A, Pinder SE, Price MR and Graves RL. Autoantibodies in early breast cancer Journal of Clinical Oncology, 2005 ASCO Annual Meeting Proceedings. Vol 23, No 16S (June 1 Supplement), 2005: 549.

[6] Lam S, Boyle P, Healey G, Maddison P, Peek L, Murray A, Chapman CJ, Allen J, Wood WC, Sewell HF, Robertson JFR. EarlyC-Lung: an Immuno-biomarker Test as an Aid to Early Detection of Lung Cancer. Cancer Prev Res 2011; 4 (7) 1126-1134

[7] Eiermann W, Jackson L, Murray A, Chapman CJ, Peek LJ, Widschwendter P, Allen J, Healey G, Robertson JFR Serum autoantibodies to breast cancer associated antigens reflect tumor biology: an opportunity for early detection & prevention? Cancer Res 2011;71(24 Suppl):Abstract P4-08-03

[8] Negm OH, Hamed MR, Schoen RE, Whelan RL, Steele RJ, Scholefield J, Dilnot EM, Shantha Kumara HM, Robertson JF, Sewell HF. Human Blood Autoantibodies in the Detection of Colorectal Cancer PLoS One. 2016 Jul 6;11(7):e0156971.

[9] Kahn, Herman, 1956; Applications of Monte Carlo, AECU-3259 (April).

[10] Kalos, M.H., and P.A. Whitlock, 1986;Monte Carlo Methods, Volume 1: Basics, John Wiley & Sons

[11] Goldberg, D. E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, Mass.: Addison-Wesley, 1989.

[12] Chebyshev, P. L. (1854). "Thorie des mcanismes connus sous le nom de paralllogrammes". Mmoires des Savants trangers prsents lAcadmie de Saint-Ptersbourg. 7: 539586.

[13] Rivlin, Theodore J. The Chebyshev polynomials. Pure and Applied Mathematics. Wiley-Interscience [John Wiley & Sons], New York-London-Sydney,1974. Chapter 2, "Extremal Properties", pp. 56–123.

[14] Geem ZW, Kim JH, Loganathan GV. A New Heuristic Optimisation Algorithm: Harmony Search. Simulation. Sage Publications Sage CA: Thousand Oaks, CA. 2001;76(2):60–68.

[15] Shi P *et al.*: Genetic Algorithm and Harmony Search Project. `https://github.com/PengSimonShi/ProjectGAHS`.