Available online at www.sciencedirect.com

ScienceDirect

Journal of Chromatography A, 1158 (2007) 226–233

JOURNAL OF CHROMATOGRAPHY A

www.elsevier.com/locate/chroma

Review

Genetic algorithms in chemistry

Riccardo Leardi [*]

Department of Chemistry and Food and Pharmaceutical Technologies, University of Genoa, Via Brigata Salerno (ponte), I-16147 Genoa, Italy

Available online 19 April 2007

Abstract

Genetic algorithms (GAs) are a quite recent technique of optimization, whose basic concept is mimicking the evolution of a species, according to the Darwinian theory of the "survival of the fittest." The application of genetic algorithms to complex problems usually produces much better results than those obtained by the standard techniques. This paper explains in detail the different steps of the algorithm and the most relevant problems to be solved in order to obtain an efficient optimization tool.

Keywords: Genetic algorithms; Optimization methods; Chemometrics

Contents

1. Introduction

When the complexity of the system under study is not extremely high, then "standard" techniques (such as those of experimental design) give quite good results, with a limited number of experiments.

However, when a problem is extremely complex these techniques cannot detect the global maximum (or minimum), because they are very sensitive to global maxima or because the response surface cannot be modelled by simple empirical models.

What is a "very complex system"? Basically, the complexity can be ascribed to three basic causes:

(a) High number of independent variables. For a linear model without interactions, the Plackett–Burman matrices allow

[*] Tel.: +39 010 3532636; fax: +39 010 3532684.
E-mail address: riclea@dictfa.unige.it.

0021-9673/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2007.04.025

the screening of many variables, with a number of experiments equal to the first multiple of 4 higher than the number of variables. If one is interested also in studying the interactions among variables and the non-linear terms, then the mathematical model would be too complex and the number of experiments would become extremely high.

(b) Very complex or irregular response surface. This can happen when the mathematical function describing the response contains higher degree terms or trigonometrical functions, or when there is no function that can adequately describe the response. This happens typically when several local maxima are present; in those cases it is not possible to use local search techniques, such as simplex, that would stop at the first local maximum they find.

(c) Presence of discontinuities in the experimental domain. It can happen that it is not possible to run experiments in some regions of the experimental domain: if the limitations define a continuous zone, then the "classical" techniques can be applied without problems, while if the limitations define more than one region, separated by discontinuities, then neither experimental design nor the local search techniques can reach the global maximum (e.g., a simplex cannot get out the region in which it started).

(d) Response to be optimised function of several "subresponses." Sometimes it can happen that what is defined as "response" to be optimised is instead something much more complex, whose value depends on the value of different responses, each of them describing a specific aspect of the problem. Just to give a simple example, in the case of a chemical reaction the global "response" can depend on yield, time, percentage of impurities, and what one is looking at is a solution being a good compromise. The desirability functions can be used, but it has to be taken into account the fact that the error of the final "response" is the sum of the errors of the different subresponses. If there are many of them, it is clear that, though each one of them can be predicted with a sufficiently small error, the global function will have a rather low predictive ability.

In all these cases, the only way to be sure of finding the global maximum would be a "grid search." Under this strategy, each variable is divided into intervals, and the response is measured in every possible combination. It is easy to realise that the number of measurements increases very fast as the number of variables increases: with four intervals (and therefore five levels) for each variable, with n variables 5^n experiments must be performed. This means 25 experiments with 2 variables, but with 10 variables it is almost 10 million experiments! Furthermore, if the surface is quite irregular, with thin peaks, the number of intervals must be quite high, not to have peaks located between the points of the grid.

It is therefore evident that with very complex problems a systematic search is out of question, because the accomplishment of the required task would be impossible (let us suppose to evaluate the response of 100 experiments per second: 10 million experiments would require one hundred thousand seconds, i.e. more than one day!).

A second strategy could be performing experiments at random points, retaining the points giving a good response and trying to improve the response by somehow using the obtained information. This could be done, for instance, by performing a local search around the randomly selected point, or by trying to exploit what the best points have in common.

2. The evolution theory

The evolution of the species can be considered by itself a form of optimisation, in which the response to be optimised is the fitness to the environment. It is anyway easy to understand that the term "fitness to the environment" is something that cannot be defined in an exact form: each individual has several characteristics, and the importance of each of them can be extremely different.

For instance, if we consider the human species, and take into account the success in life, we can see that there are some fundamental characteristics, without which an individual cannot survive (e.g., the absence of lethal pathologies), important characteristics (intelligence, strength, beauty), and non-relevant characteristics (the shoe number or the colour of the eyes). It can also be noticed how the same value of the response "success" can be obtained by individuals having totally different characteristics: a scientist winning the Nobel prize, an athlete winning the Olympic Games and a girl winning the title of Miss World all have a very high response.

The basic idea of the evolution theory is that the individuals with a greater "fitness to the environment" have a greater probability of surviving and a greater probability of winning the fights for mating. In such a way the genetic content of the best individuals will be more and more present in the following generations, since it will be transmitted by the offspring.

It is logical to think that, if the parents have a good "response," also their offspring will have a good one, sometimes even better than that of their parents.

Since centuries men applied this idea, in order to get, for instance, cows producing more milk, or horses running faster, or fruits or flowers with specific characteristics. As can be seen, in these cases the response being considered is not a general "fitness to the environment," but a specific characteristics that has to be maximised. As a consequence, we can see that a "guided evolution" can get results completely different from those of the "natural evolution."

3. How to transform the evolution theory into an optimization technique?

Genetic algorithms have been proposed by Holland in the 1960s, but it was possible to apply them with reasonable computing times only since the 1990s, when computers became much faster.

General information on genetic algorithms, relevant to the whole of this review, can be found in references [1–7]. A wealth of information can also be found on the websites of various organizations [8–15].

The basic idea is to perform a computer simulation of what occurs in nature, and the first problem to be solved is how to code the information in such a way that the computer can treat it.

It can therefore be said that the fitness to the environment is function of the genetic material, in the same way as the result of an experiment is function of the experimental conditions. Therefore, a correspondence genetic material-experimental conditions can be established.

At a lower level, we can say that the genetic material is defined by the genes, in the same way as an experimental condition is defined by the values of the variables involved in the experiment. Therefore, a correspondence genes-variables can be established.

At an even lower level, we can see that the information contained in each gene is defined by a sequence of nitrogenated bases: since there are four bases, each gene can be considered as a word of variable length, written in a four-letter alphabet. In the same way, we can use the binary code to transform the value of a variable in a word of variable length, written in bits (two-letter alphabet, 0 and 1).

So, we have the following correspondences:

(i) genetic material: experimental conditions (point in the experimental domain)
(ii) gene: variable
(iii) nitrogenated base: bit

Finally, we can see how the experimental conditions can be coded by a sequence of 0's and 1's.

4. The problem of coding

Let us suppose we are dealing with a chemical reaction, of which we want to maximise the yield, and that we want to code the following experimental condition (it has to be clear that this is just a didactical example, since it would be foolish to apply such a complex method as GAs to optimise such a simple problem):

(i) reaction temperature: 30 °C
(ii) reaction time: 20 min
(iii) stirring: yes
(iv) catalyst: type A (A and B possible catalysts)

Four genes will therefore form the corresponding "genetic material" (for sake of simplicity, GAs always work with one single chromosome), each gene containing the corresponding information in bits:

011110 10100 1 0

(blanks have been added only to make genes evident).

Two facts can be noticed:

(a) variables of different types can be dealt with at the same time: quantitative variables (time, temperature), qualitative variables (type of catalyst) and variables of type yes/no (stirring);
(b) the number of bits for each gene can be very different.

It is extremely important that the coding of the quantitative variables takes into account the interval in which the same variable can vary and the fact that the difference between two levels is significant. It is intuitive that if a large number of bits is used to describe a variable, very small variations of the same variable could be studied; these small variations must anyway have a real meaning.

In the previous example, six bits describe the temperature. This means that:

(i) the range is between 0 and 63 °C;
(ii) the difference between the two levels is 1 °C.

This is like saying that:

(i) we are interested in studying the reaction from 0 to 63 °C;
(ii) the difference of one degree is significant (the reaction at 25 °C can be different from the reaction at 26 °C) and the temperature can be set with the precision of 1 °C (for an experiment to be performed at 25 °C, I can actually set it between 24.5 °C and 25.5 °C).

It is clear that in this case the number of bits used is too large for the characteristics of the variable. If the total number of bits of a chromosome is n, the total number of combinations is 2^n (in our example, with 13 bits, the total number of possible combinations is 8192); a number of bits larger than the required has the only effect of increasing the complexity of the system, and therefore decreasing the efficiency of the search.

Always taking into account our previous example, let us suppose that the range of temperatures we are interested in is from 25 to 60 °C, with an interval of 5 °C: as a consequence, eight levels describe completely our variable, and three bits are enough:

000 = level 0 = 25 °C
001 = level 1 = 30 °C
010 = level 2 = 35 °C
011 = level 3 = 40 °C
100 = level 4 = 45 °C
101 = level 5 = 50 °C
110 = level 6 = 55 °C
111 = level 7 = 60 °C

The information that can be obtained with this coding is comparable to that of the original coding, but just three bits instead of six have been used.

In the same way, if for the variable time we are interested in the range 10–40 min, with an interval between levels of 2 min, 16 levels, and therefore four bits, will be required:

0000 = level 1 = 10 min
0001 = level 2 = 12 min
. . . .
1110 = level 15 = 38 min
1111 = level 16 = 40 min

The final coding would be 001 0101 1 0, with nine bits and 512 possible combinations. It is easy to realise that a more clever coding can reduce by 16 times the search complexity, without losing almost any information.

After having coded the experimental conditions and having generated the corresponding chromosome, the response (in this case the yield of the reaction) is measured: this response will always be associated to the chromosome.

5. Steps of the genetic algorithms

According to the evolution theory, the improvement of a species occurs because, through a very high number of generations, the genetic material of its individuals is constantly improving. The reason of this is because the "bad" individuals do not survive and the best ones have a greater probability of spreading their genetic material to the following generation. Beyond this "logical" development, mutations allow the exploration of new "experimental conditions"; usually mutations produce bad results (e.g., severe pathologies), but it can happen that these random changes of a nitrogenated basis end up in a better genome.

Several GAs have been developed; beyond the common basic idea (mimicking the evolution of a species), they can have relevant differences. All of them have three fundamental steps that can anyway be performed in different ways. These three steps are:

(i) creation of the original population;
(ii) reproduction;
(iii) mutations.

Let us now have a short description of each one of them.

5.1. Creation of the original population

The population size stays constant throughout the elaboration. The number of individuals can be quite different, and usually is in the range 20–500 (later on, we will describe the influence of this parameter on the performance of the GAs).

After having decided the population size (p), the genetic material of the p individuals is randomly determined. This means that every single bit of each chromosome is randomly set to 0 or 1.

If this chromosome corresponds to a possible experimental condition (i.e., inside the experimental domain), its response is evaluated.

5.2. Reproduction

After having created the original population (or first generation), the individuals start "mating" and "producing offspring." This is the step in which the different GAs have the greatest variations, though all of them follow the same idea: the probability of the best chromosomes (the ones giving the best responses) of producing offspring is higher than that of the worst chromosomes, and the offspring originated by their "mating" are a recombination of the parents.

Basically, the first step is creating the population of the second generation simply by randomly copying p times a chromosome of the first generation. If the drawing would be totally random, then each chromosome would have the same probability of going to the next generation and therefore the average response of the generation $n+1$ would be statistically the same as that of generation n.

In nature, each individual always has a possibility of reproducing itself, but the best ones have a greater probability of winning the fights for mating. In the same way, the drawing performed to select the chromosomes that will be copied must take into account the response of the individuals, giving the best ones a higher probability. To do this, a biased drawing is performed, in which the probability of each individual of being selected is function of its response. To visualise this process in a simple way, it is as performing the selection with a roulette wheel in which the slots corresponding to the best individuals are larger than those corresponding to the worst ones.

Several functions have been studied to perform this step: the simplest one is that the probability of each chromosome of being selected is equal to its response divided by the sum of the responses:

$$p_i = \frac{\mathrm{resp}_i}{\Sigma\,\mathrm{resp}}$$

As a result of such a drawing, the best individuals will be copied more than once in the population $n+1$, while the worst ones will disappear. It can be easily understood that the average response of generation $n+1$ will be higher than the average response of generation n.

Continuing with the previous example, let us simulate this step, supposing that the population size is 10 individuals.

Original population

Chromosome	Experimental conditions	Yield
001 1001 0 1	30 °C, 28 min, no, B	54.9
010 0100 1 1	35 °C, 18 min, yes, B	67.2
000 1010 0 0	25 °C, 30 min, no, A	66.0
100 0101 1 1	45 °C, 20 min, yes, B	70.3
110 0001 1 0	55 °C, 12 min, yes, A	79.1
010 1111 0 1	35 °C, 40 min, no, B	62.1
101 0111 1 1	50 °C, 24 min, yes, B	71.3
001 0010 1 0	30 °C, 14 min, yes, A	83.4
100 1001 1 0	45 °C, 28 min, yes, A	89.6
001 0011 1 1	30 °C, 16 min, yes, B	59.7

After sorting the population and computing the selection probability, we have:

Chromosome	Experimental conditions	Yield	Probability
100 1001 1 0	45 °C, 28 min, yes, A	89.6	0.127
001 0010 1 0	30 °C, 14 min, yes, A	83.4	0.119
110 0001 1 0	55 °C, 12 min, yes, A	79.1	0.112
101 0111 1 1	50 °C, 24 min, yes, B	71.3	0.101
100 0101 1 1	45 °C, 20 min, yes, B	70.3	0.100
010 0100 1 1	35 °C, 18 min, yes, B	67.2	0.096
000 1010 0 0	25 °C, 30 min, no, A	66.0	0.094
010 1111 0 1	35 °C, 40 min, no, B	62.1	0.088
001 0011 1 1	30 °C, 16 min, yes, B	59.7	0.085
001 1001 0 1	30 °C, 28 min, no, B	54.9	0.078

The average response is 70.36. Let us now suppose to draw 10 random numbers between 0 and 1: these numbers will decide which chromosomes will be copied: Let the 10 numbers be:

0.353	0.038	0.367	0.324	0.414	0.903	0.150	0.353	0.428	0.915

The value 0.353 correspond to the selection of chromosome 3, whose space in the roulette wheel corresponds to the values between 0.247 (0.127 + 0.119 + 0.001) and 0.358 (0.127 + 0.119 + 0.112); in the same way, the other values correspond to chromosomes 1, 4, 3, 4, 9, 2, 3, 4, 9; in the second generation chromosome 1 will be copied once, chromosome 2 once, chromosome 3 three times, chromosome 4 three times and chromosome 9 twice. The second generation will be the following:

Chromosome	Experimental conditions	Yield
100 1001 1 0	45 °C, 28 min, yes, A	89.6
001 0010 1 0	30 °C, 14 min, yes, A	83.4
110 0001 1 0	55 °C, 12 min, yes, A	79.1
110 0001 1 0	55 °C, 12 min, yes, A	79.1
110 0001 1 0	55 °C, 12 min, yes, A	79.1
101 0111 1 1	50 °C, 24 min, yes, B	71.3
101 0111 1 1	50 °C, 24 min, yes, B	71.3
101 0111 1 1	50 °C, 24 min, yes, B	71.3
001 0011 1 1	30 °C, 16 min, yes, B	59.7
001 0011 1 1	30 °C, 16 min, yes, B	59.7

The average response of this population is 74.36.

With this process chromosomes existing in the previous population have simply been copied and placed in the following population, without exploring any new experimental condition.

To do this, a reproduction is simulated: the 10 individuals are randomly paired in five pairs, and from each pair (the “parents”) two new individuals (the “offspring”) will be obtained after a “crossover,” by which the genes of the parents will be shuffled.

Let us suppose the pairs are: 1–10, 2–9, 5–8, 4–6 and 3–7. Let us take into account the first one:

100 1001 1 0	45 °C, 28 min, yes, A
001 0011 1 1	30 °C, 16 min, yes, B

There are several ways of shuffling the genes of the parents. The two most frequently applied are the “single crossover” and the “uniform crossover.”

In the former, a “breaking point” in the chromosome is randomly selected: the first offspring will be formed by the genes of parent 1 at the left of the breaking point and by the genes of parent 2 at the right of the breaking point; the other way round for offspring 2.

With four genes a number between 1 and 3 is drawn: let us suppose 2.

100 1001 |1 0	45 °C, 28 min, yes, A
001 0011 |1 1	30 °C, 16 min, yes, B

The two offspring will be:

100 1001 |1 1	45 °C, 28 min, yes, B
001 0011 |1 0	30 °C, 16 min, yes, A

It is easy to understand that, with this method, the order of the variables in the chromosome is very important: with g genes, two variables coded as contiguous genes will have a probability of just $1/(g-1)$ of being transmitted each one to a different offspring, while the first and the last variables will always be transmitted to the first and the second offspring, respectively.

In the latter method, for each gene a random number is drawn, determining to which offspring the genes of the parents will be assigned: if the value is <0.5, then the gene of parent 1 will be given to offspring 1 (and the gene of parent 2 will be given to offspring 2), if it is >0.5, then the gene of parent 1 will be given to offspring 2 (and the gene of parent 2 will be given to offspring 1). Of course, this drawing will not take place for the genes being the same in both parents.

Let us suppose that the values are 0.334 for the first gene, 0.719 for the second one and 0.265 for the fourth one (the third one is the same in both parents).

The two offspring will be:

100 0011 1 0	45 °C, 16 min, yes, A
001 1001 1 1	30 °C, 28 min, yes, B

Doing the same for all the pairs, the following population is obtained:

100 0011 1 0	45 °C, 16 min, yes, A
001 1001 1 1	30 °C, 28 min, yes, B
001 0011 1 0	30 °C, 16 min, yes, A
001 0010 1 1	30 °C, 14 min, yes, B
110 0111 1 0	55 °C, 24 min, yes, A
101 0001 1 1	50 °C, 12 min, yes, B
101 0001 1 0	50 °C, 12 min, yes, A
110 0111 1 1	55 °C, 24 min, yes, B
101 0111 1 0	50 °C, 24 min, yes, A
110 0001 1 1	55 °C, 12 min, yes, B

The 10 individuals that have been obtained after this step are different from the 10 individuals of the first generation, and also different from each other (in some algorithms this is a necessary condition, while in some other ones the “twins” are accepted). Though different individuals have been obtained, by continuing in this way only already tested values of the variables would be used; furthermore, in this case the third gene (stirring) has value 1 in all the population: therefore, an experimental condition without stirring could never more occur.

5.3. Mutations

To overcome these problems an operator simulating mutations is implemented: in nature, mutations take place with an extremely low probability and have as effect the variation of a “letter” of the word coding the gene; a nitrogenated basis in the DNA, a bit in our chromosome.

The main difference between crossover and mutation is that, while the crossover is applied at gene level (it involves all the bits coding the variable), the mutation affects single bits.

The usual probability is 1–2%. In our case, having a population of 10 chromosomes, each of them described by nine bits, a mutation probability of 2% would lead to an average of 1.8 mutations per generation.

If the bits affected by a mutation are bit number 4 of chromosome 2 and bit number 3 of chromosome 7, the “final” population for the second generation will be:

100 0011 1 0	45 °C, 16 min, yes, A
001 0001 1 1	30 °C, 12 min, yes, B
001 0011 1 0	30 °C, 16 min, yes, A
001 0010 1 1	30 °C, 14 min, yes, B
110 0111 1 0	55 °C, 24 min, yes, A
101 0001 1 1	50 °C, 12 min, yes, B
100 0001 1 0	45 °C, 12 min, yes, A
110 0111 1 1	55 °C, 24 min, yes, B
101 0111 1 0	50 °C, 24 min, yes, A
110 0001 1 1	55 °C, 12 min, yes, B

After having evaluated the response and having sorted the chromosomes, we have:

101 0111 1 0	50 °C, 24 min, yes, A 89.2
100 0011 1 0	45 °C, 16 min, yes, A 86.5
110 0111 1 0	55 °C, 24 min, yes, A 85.8
100 0001 1 0	45 °C, 12 min, yes, A 84.0
001 0011 1 0	30 °C, 16 min, yes, A 83.8
110 0111 1 1	55 °C, 24 min, yes, B 68.9
101 0001 1 1	50 °C, 12 min, yes, B 68.4
110 0001 1 1	55 °C, 12 min, yes, B 67.6
001 0010 1 1	30 °C, 14 min, yes, B 65.0
001 0001 1 1	30 °C, 12 min, yes, B 64.6

The average response of the second generation is higher than that of the first one; it can be noticed that all the best chromosomes have stirring and catalyst A. The whole process (select-copy, cross-over, mutations) is repeated on these chromosomes and the third generation is obtained. With this kind of selection, each generation will usually have an average response higher than the previous generation.

New generations will be created until a stop criterion is satisfied, the most common of which are: predefined number of generations, predefined time of elaboration, obtention of a target response value.

6. Comments about the parameters of the genetic algorithms

When describing a GA, the details about the different parameters must be given: they can have very different values and can have a very strong effect on the final result. It has to be well understood that an "optimal" form of the GA does not exist, and that for each problem the best results can be obtained by a specifically designed GA.

Basically, the strength of the GAs is in the joint application of two strategies: exploration and exploitation. The former is typical of the random searches, in which different points of the experimental domain are randomly tested: this allows testing points in different regions of the space, without being worried by what happens around them. The latter is typical of the local searches, such as simplex, that try to reach the local maximum closer to the starting point, without caring of what happens in different regions of the experimental domain.

The basic problem, and the secret to get a good algorithm, is getting a good balance between exploration and exploitation: in the next subsections the influence of each parameter will be described.

6.1. Population size

A population formed by many individuals allows keeping a great difference among the chromosomes, and therefore exploring at the same time several different regions; with a small population it can happen that all the individuals are extremely similar.

On the other side, in the same computing time a greater population will produce a smaller number of generations than a larger population. This means that a very good chromosome found in generation n will need much more time in producing its effects, by generating offspring: this will happen only in generation $n+1$.

In literature populations ranging between 20 and 500 individuals can be found. To decide the population size the time required to evaluate the response is also important: if it is quite short, then a larger population can be used, since the time interval between the generations will be short; on the other side, if it is quite long, then it would be better to work with a reduced genetic variability, keeping anyway acceptable the time interval between generations.

Generally speaking, it can be said that by increasing the population size an increase in the exploration and a decrease in the exploitation is obtained.

6.2. Reproduction

In the previous example we saw that, usually, generation $n+1$ has an average value of the response higher than generation n. In the same example, it occurred that the best chromosome of generation 2 is worse than the best chromosome of generation 1. This can happen because all the parents "die," and it is not sure that in the following generation chromosomes better than the best of the parents are present.

To avoid this problem it is possible to set the "elitism": the k best individuals of generation n go directly to generation $n+1$. Therefore, if the global maximum is found, the corresponding chromosome will never die: it is like saying that the concept of immortality has been added. With a population of c chromosomes, the select-copy operation, performed on all the c chromosomes, will produce $c-k$ parents, and therefore $(c - k)/2$ pairs, and finally $c-k$ offspring, to which the k best chromosomes of generation n will be added, in order to reconstitute a population size of c chromosomes. It has to be noticed that the k elitist chromosomes can also be selected as parents and therefore they can continue producing offspring. As an extreme case, it is possible to copy just two chromosomes, whose offspring will enter the population only if they are better than the worse chromosomes of the previous population: by doing that, the concept of "generation" is completely lost.

Basically the elitism has the advantage that, at the end of the elaboration, the k best chromosomes of the population will be the k best chromosomes ever found. Furthermore, the time required for a generation is smaller, and therefore the same result as discussed in the subsection "population size" is obtained. Here too, with a high elitism the risk is that all the chromosomes are quite similar, around a good maximum, and that it will be impossible to get out of that region: the only possibility would be landing by chance on a higher peak, with a higher response.

Summarizing, a reproduction without elitism has a higher exploration, while the higher the elitism the higher the exploitation. Here too, the problem is finding a good balance.

6.3. Mutation probability

Also in this case, the goal is finding a good balance between exploration and exploitation. A low mutation probability does not allow changing easily the study region, and therefore favours exploitation. With a high mutation probability changing region is much easier, but the local search is very poor, the process is much more random and exploration is favoured.

7. Hybrid algorithms

The main difference between the classical techniques and GAs has already been reported: while the former focus on local search, succeeding in the identification of the local maximum, without taking into account what happens in the different regions of the experimental domain, the latter perform a very good exploration of the space, without directly trying to go to the top of the local maxima. It is obvious to think that the application of both techniques could lead to results being better of what could be obtained by each of them separately.

One of the problems of the classical techniques is that they need previous knowledge, in order to set their parameters (e.g., for a simplex, experimental domain, step and starting point). This knowledge could instead be obtained by a GA: the best chromosome would be a good starting condition for a classical technique.

Even better results are produced by alternating the two techniques: in this case an "hybrid" algorithm is obtained. The first step is a GA; when a predefined stop criterion is satisfied (computation time, number of evaluations, ...), a local search is performed starting from the best chromosome. The result of this local search will be treated as an offspring, and therefore enter the population. At the end of the local search, a new GA is started, until the next stop, and so on ...

By joining two totally different approaches the drawbacks of both are reduced, and a new technique with both a good exploration and a good exploitation is obtained.

8. Looking for the global maximum: Is it worthwhile?

Many times the validity of a GA is measured only by taking into account its capability of finding the global maximum and the time required for it.

This can be really important only if the global maximum is much better than the other local maxima: in this case, if the GA stops on a different maximum (as in the case of simplex), then a much poorer result is obtained.

A different situation happens when the response corresponding to one or more local maxima is not significantly worse than that of the global maximum. In this case, the detection of the global maximum is much more difficult; furthermore, this is not relevant from the practical point of view, provided that the value found by the GA is not very different.

It has also to be considered that, when working with chemical data, we are dealing with data affected by an experimental error; therefore, the experimental conditions corresponding to the global maximum could be quite different simply as a consequence of measurement errors. Let us suppose to have two local maxima, with similar height: it is very easy that which one of the two peaks will be the higher will depend only on the data taken into account (and on their experimental errors). What is really important is that at least one of them can be detected.

A great advantage of GAs over the classical techniques is that at the end of the elaboration the user is given not just an "optimal" solution, but also a population of extremely good solutions, usually having very similar responses. With a complex response surface, with many local maxima, it is possible that the same response can be obtained from completely different experimental conditions. In this case, it is quite common that experimental conditions leading to the same response are not equivalent in what concerns their cost and/or practical operating difficulties. The user can therefore select the "best" chromosome based also on practical aspects.

9. Applications in separation sciences

The most practical application of GAs in separation sciences is related to variable (peak) selection (the problem is pretty much the same as the wavelength selection in spectroscopic sciences [16]).

Nowadays, chromatograms with hundreds of peaks can be quite easily obtained, and the detection of the most relevant ones is often one of the key points for the success of an application; this can be either in case of quantitative analysis or in case of discrimination among classes of samples. While in relatively simple cases experts can select by themselves the peaks of interest, according to their chemical knowledge, this becomes much more challenging when analysing much more complex samples.

Computer-aided variable selection is important for several reasons. Variable selection can improve model performance, provide for more robust models and models that can transfer more readily and allow non-expert users to build reliable models with only limited expert intervention.

In a data set in which each sample is described by v variables (peaks), each chromosome is composed by v genes, each gene being formed by a single bit [17]. As an example, with 10 variables, a set that only uses variables 1, 5, 8 and 9 will be coded as 1000100110.

The response to be optimised will be the cross-validated variance explained by the selected set of variables (in case of multivariate calibration) or the percentage of cross-validated correct classifications (in case of a classification problem).

While most GAs are intended to work in a continuous space, in this case the space under investigation is absolutely discontinuous: when working with v variables, it is like studying only the vertices of a v-dimensional hypercube. Several changes must therefore be applied to the standard GA to best adapt it to this specific purpose.

The procedure of variable selection, apparently so simple, is indeed very complicated and needs a careful validation to avoid the risk of overestimating the predictive ability of the selected model. In such cases, when using it on new data, one could be strongly deceived, discovering that it has no predictive ability at all.

This is mainly due to random correlations: if you try to describe 10 hypothetical objects with 100 random variables and a random response, you will surely have some of the variables perfectly modelling the response. This risk is, of course, higher when the variables/objects ratio is very high.

Therefore, this technique can be applied only if the number of samples is relatively high (at least some tens of samples).

The risk of overfitting is also higher the longer the GA runs (i.e. the more models that are tested); a good solution consists of performing a large number of independent short runs, and obtaining the final model by taking into account the results of all the runs. By doing this, a much more consistent (and less overfitted) solution can be obtained.

Of course, after this automated variable selection, the chemist will have to use his expertise and his knowledge of the chemical problem in order to interpret and give a chemical explanation to the results of this procedure.

References

[1] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Berkeley, 1989.

[2] R. Leardi, J. Chemom. 15 (2001) 559.

[3] R. Leardi (Ed.), Nature-Inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks (Data Handling in Science and Technology, Vol. 23), Elsevier, Amsterdam, 2003.

[4] C.B. Lucasius, G. Kateman, Chemom. Intell. Lab. Syst. 19 (1993) 1.

[5] C.B. Lucasius, G. Kateman, Chemom. Intell. Lab. Syst. 25 (1994) 99.

[6] R.E. Shaffer, G.W. Small, Anal. Chem. 69 (1997) 236A.

[7] R. Wehrens, L.M.C. Buydens, Trends Anal. Chem. 17 (1998) 193.

[8] LIPS (Laboratory for Intelligent Process Systems), Purdue University, http://cobweb.ecn.purdue.edu/~lips/.

[9] S. Schulze-Kremer, Genetic Algorithms and Protein Folding, Faculty of Technology, University of Bielefeld, 1996, http://www.techfak.uni-bielefeld.de/bcd/Curric/ProtEn/proten.html.

[10] Wikipedia, Genetic algorithm (2007), http://en.wikipedia.org/wiki/Genetic_algorithm.

[11] T. Wong, H. Wong, S. Drossopoulou, Genetic Algorithms, Imperial College of Science Technology and Medicine, London, 1996, http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/tcw2/report.html.

[12] J.-P. Rennard, Introduction to Genetic Algorithms (2000), http://www.rennard.org/alife/english/gavintrgb.html.

[13] Association for the Advancement of Artificial Intelligence, Genetic Algorithm and Genetic Programming, 2006, http://www.aaai.org/AITopics/html/genalg.html.

[14] A. Marczyk, The TalkOrigins Archive, Genetic Algorithms and Evolutionary Computation (2004), http://www.talkorigins.org/faqs/genalg/genalg.html.

[15] J. Holland, Genetic Algorithms, in: L. Tesfatsion homepage, Department of Economics, Iowa State University, 2007, http://www.econ.iastate.edu/tesfatsi/holland.GAIntro.htm.

[16] R. Leardi, in: R. Leardi, (Ed.), Nature-Inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks (Data Handling in Science and Technology, Vol. 23), Elsevier, Amsterdam, 2003, pp. 169-196.

[17] R. Leardi, R. Boggia, M. Terrile, J. Chemom. 6 (1992) 267.