# Genetic algorithms in near-infrared spectroscopy and chemometrics: past and future

Janne Koljonen, Torbjörn E.M. Nordling, and Jarmo T. Alander

University of Vaasa, email: *Firstname.Lastname*@uwasa.fi
URL: ftp://ftp.uwasa.fi/cs/report05-4/GANIR.pdf

## Abstract

Global optimization and search problems are abundant in science and engineering, including spectroscopy and its applications. Therefore it is hardly surprising that general optimization and search methods such as genetic algorithms have found applications also in the area of near-infrared spectroscopy. We here give a brief introduction to genetic algorithms, their objectives and applications in near-infrared spectroscopy, as well as chemometrics.

The most popular applications of genetic algorithms in near-infrared spectroscopy are wavelength, or more generally speaking variable, selection. Genetic algorithms are both frequent and convenient in multicriteria optimization, e.g. selection of pre-processing methods, wavelength inclusion, and selection of latent variables can be optimized simultaneously. Wavelet transform has recently been applied to pre-processing of near-infrared data. Especially, hybrid methods of wavelets and genetic algorithms have in a number of research papers been applied to pre-processing, wavelength selection and regression with good success.

In all calibration, and in particular when optimizing, it is essential to validate the model and to avoid overfitting. Genetic algorithms have a large potential when addressing these two major problems and we believe that many future applications will be seen. To conclude, optimization give good opportunities to simultaneously develop an accurate calibration model, regulate model complexity and prediction ability, within a considered validation framework.

## Introduction

Genetic algorithms (GA) have been successfully applied to many difficult search and optimization problems in a diversity of research domains, including chemometrics and near-infrared spectroscopy (NIRS). Table 1 shows, how the interest to use GAs in NIRS aroused in the mid 90s, and the journals that have published the most GA-NIRS articles.

**Table 1:** Annual frequencies of published GA-NIRS papers (above) and the most popular journals for GA-NIRS articles (below).

| Year | Items | Year | Items |
|---|---|---|---|
| 1987 | 2 | 1998 | 14 |
| 1988 | 0 | 1999 | 21 |
| 1989 | 0 | 2000 | 13 |
| 1990 | 0 | 2001 | 8 |
| 1991 | 0 | 2002 | 13 |
| 1992 | 4 | 2003 | 11 |
| 1993 | 3 | 2004 | 6 |
| 1994 | 4 | 2005 | 9 |
| 1995 | 8 | 2006 | 6 |
| 1996 | 13 | 2007 | 5 |
| 1997 | 22 | | |

| Journal | Items |
|---|---|
| Analytica Chimica Acta | 15 |
| Analytical Chemistry | 5 |
| Applied Spectroscopy | 5 |
| Journal of Chemometrics | 5 |
| Journal of Biomolecular NMR | 4 |
| Chemometrics and Intelligent Laboratory Systems | 4 |
| Journal of Chemical Information and Computer Science | 4 |
| Journal of Near Infrared Spectroscopy | 3 |

## Example of a genetic algorithm in NIRS

Genetic algorithms are an optimization method based on stochastic global search, which mimics biological evolution in the "survival of the fittest" sense. Experiments (trials, individuals) are encoded in *chromosomes* using an alphabet that guarantees a unique mapping between *genotype* and *phenotype*.

In the chemometrics example in Figure 1, 18 wavelength variables of a near-infrared (NIR) spectra are encoded using a binary alphabet, such that '1' denotes for the inclusion of a variable and '0' the exclusion of it. The last *gene* consists of two bits (**c**(19:20)) that map to {1, 2, 3, 4} denoting the number of components to be used in the partial least squares (PLS) model, which is used to predict the concentration of organic compounds ($CH_2$ methylene). The example hence illustrates that chromosomes easily can contain different types of variables, which promote the use of GAs in *multicriteria* optimization.

The *fitness function* transforms the genotype **c** of an individual into a single number the *fitness* value, which characterizes the performance of the individual. In our example, the objective is to minimize the root-mean-square errors of prediction (RMSEP) of organic compound concentration, while using a PLS model with as few components and included wavelengths as possible:

$$\text{Fitness} = -\left[ w_1 \underbrace{\sqrt{\frac{1}{I}\sum_{i=1}^{I}(\hat{y}_i - y_i)^2}}_{RMSEP} + w_2 N + w_3 C \right].$$

The relative importance of the three objectives is determined by the weights ($w_1$,$w_2$,$w_3$), while $N$ denotes the number of wavelengths included in the model and $C$ the number of components. The predicted organic compound concentration in sample $i$ is denoted by $\hat{y}_i$ and the true concentration by $y_i$ (measured by some other method).

A typical GA contains fours steps, Figure 2: (i) creating an initial population, (ii) evaluating the fitness of each individual in the population, (iii) checking the stop condition, (iv) generating a new generation, which includes selection of parents, crossover and mutation.

## Conclusions

Most applications of genetic algorithms in near-infrared spectroscopy are in wavelength selection. **Wavelength selection improves** model **accuracy** even in conjunction with latent variable extraction, **robustness** of the model against disturbances, and makes the **inference** of the model, as well as its **interpretation** easier. In comparative studies, GA usually outperforms other methods in many aspects. The **drawback** is usually longer **computation time** in comparison to deterministic methods.

GA seems to suffer **less** from **overlearning**, probably due to its ability to get stuck in local optima of a noise landscape, despite that GAs are regarded as a global search methods. The cause of overlearning is that measurement noise specific to the data set, used to evaluate the fitness function, end up being used as a predictor variable. Hence the regression model (or classifier) is overfit and its ability to generalize is inferior. Both the setup of the search strategy and the selection of a proper validation method are of great importance for tackling the problem. One should also remember that the final optimizer can be overfit to the calibration or/and validation sets.

GAs are well suited for **multicriteria optimization**. In chemometrics, it is possible to simultaneously optimize e.g. wavelength inclusion, other **pre-processing** steps, the **number of latent variables** in the model, and the regression model (or classifier) itself. Results in the literature indicate that optimized pre-processing can greatly enhance the spectral model. A promising recent trend in chemometrics is to the use of **wavelet transformation** together with GA optimization **to pre-process** spectra.

In conclusion, genetic algorithms have found many applications in chemometrics and NIRS, particularly wavelength selection but more general approaches to optimization schemes have also been introduced. The **generality** and **diversity of GAs** is, among others, perhaps the most important reason for their popularity and applicability in chemometrics. There are, however, still a lot of **potential new applications** for GAs in NIRS, hints about these can be found in other research domains. For example method testing and development should be considered.

## Acknowledgements

The EU Interreg project NIRCe is gratefully acknowledged for funding this research.

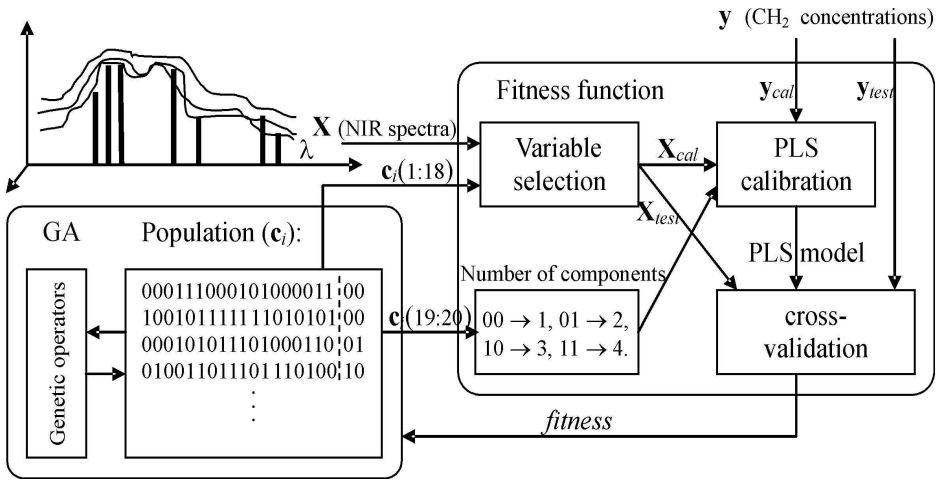


**Figure 1**: A typical simple genetic algorithm for wavelength selection.

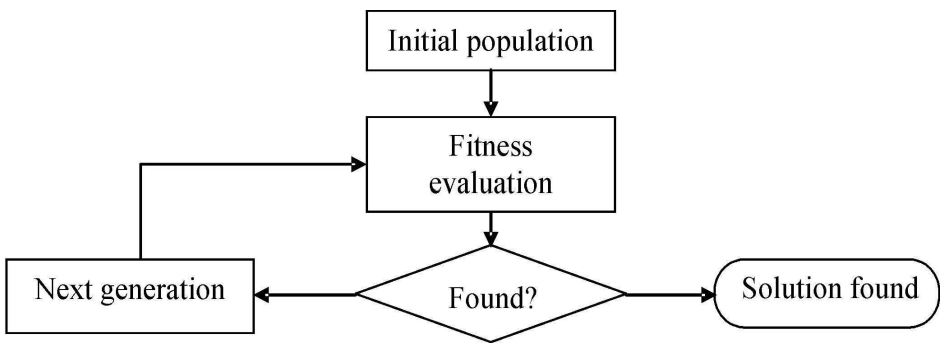


**Figure 2:** The four steps of a typical genetic algorithm (GA).