GENETIC ALGORITHMS

Riccardo Leardi

Department of Pharmaceutical and Food Chemistry and Technology
University of Genoa - ITALY

People at Dow Chemical were reading the literature ...

ELSEVIER

Chemometrics and Intelligent Laboratory Systems 41 (1998) 195–207

Chemometrics and intelligent laboratory systems

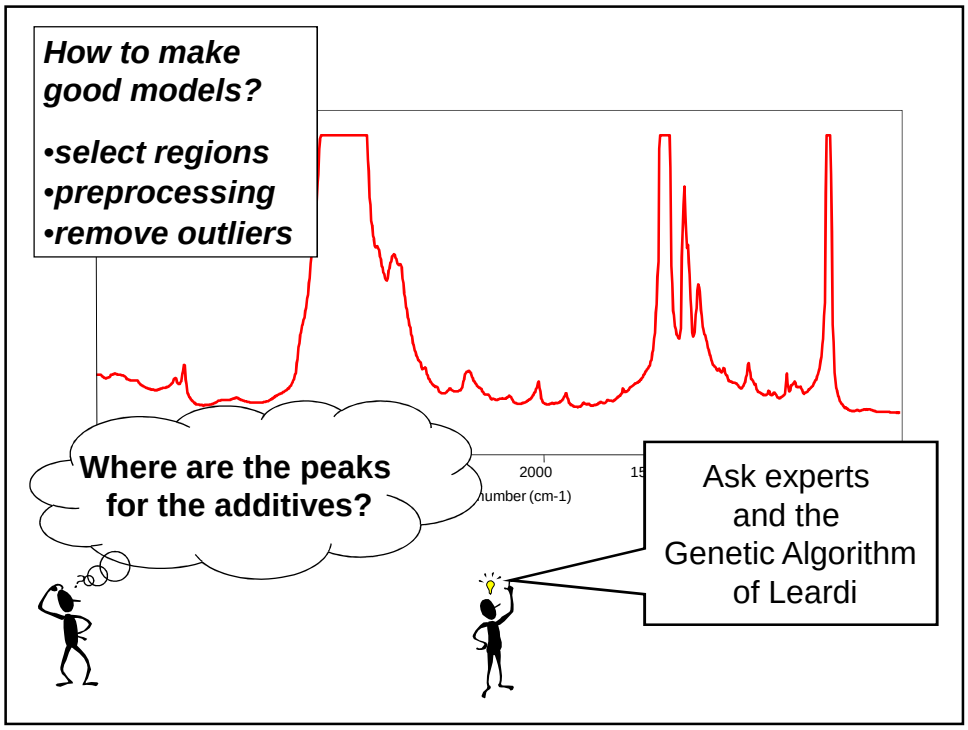
Genetic algorithms applied to feature selection in PLS regression: how and when to use them

Riccardo Leardi [a,*], Amparo Lupiáñez González [b]

[a] *Dipartimento di Chimica e Tecnologie Farmaceutiche e Alimentari, via Brigata Salerno (ponte), University of Genoa, 16147 Genova, Italy*

[b] *Departamento de Química Analítica, Facultad de Ciencias, University of Granada, Granada, Spain*

Received 7 November 1997; accepted 10 April 1998



WHAT I GOT

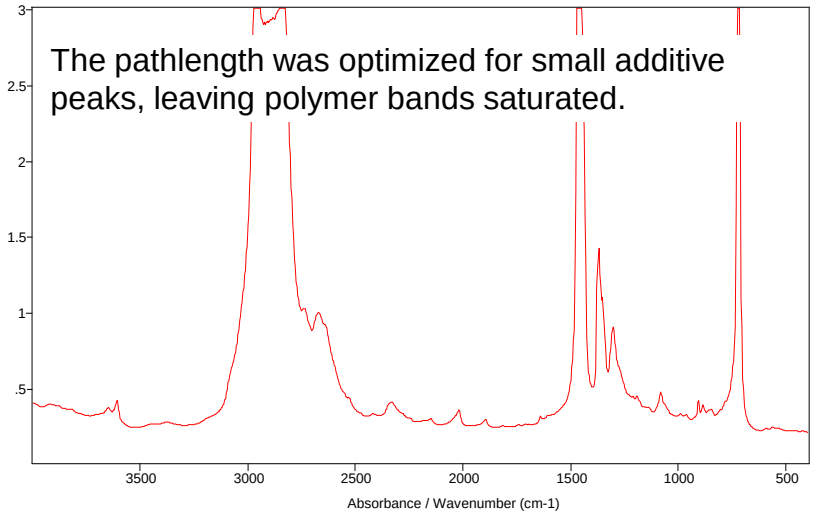
- **FTIR data of polymer films (1873 wavelengths)**
- **Concentrations of 2 additives (no names)**
  - **Additive B (42 + 28 samples)**
  - **Additive C (109 + 65 samples)**
- **NO information about suggested regions**

THE CHALLENGE

To verify if Genetic Algorithms could find a model characterized by:

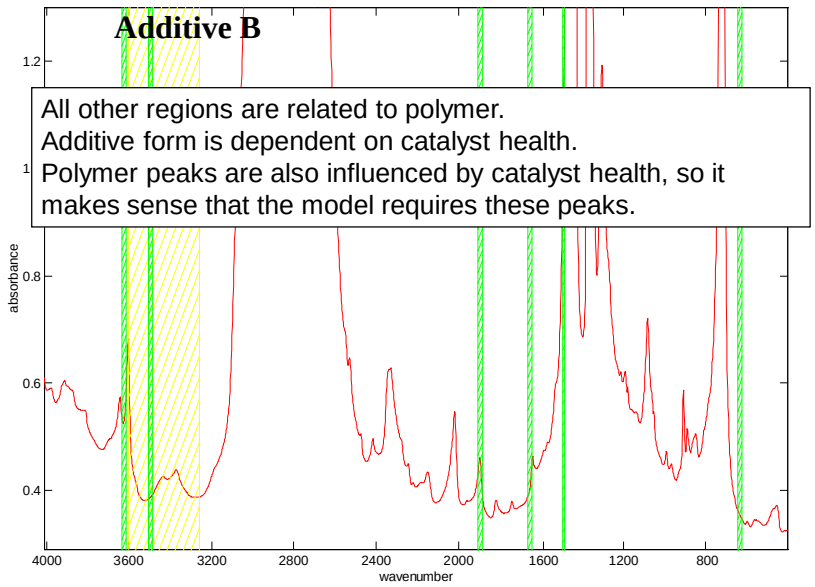
- **good predictive ability**
- **“logical” regions**

These spectra are not pretty



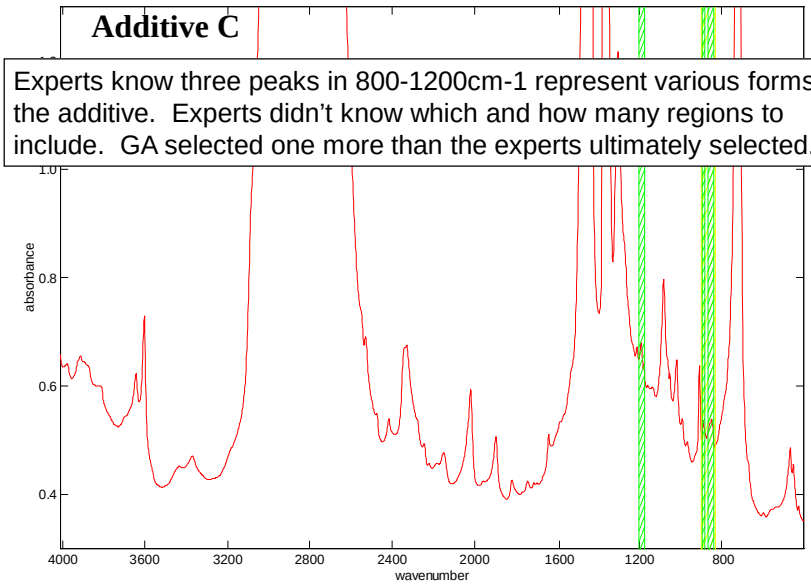
THE RESULTS

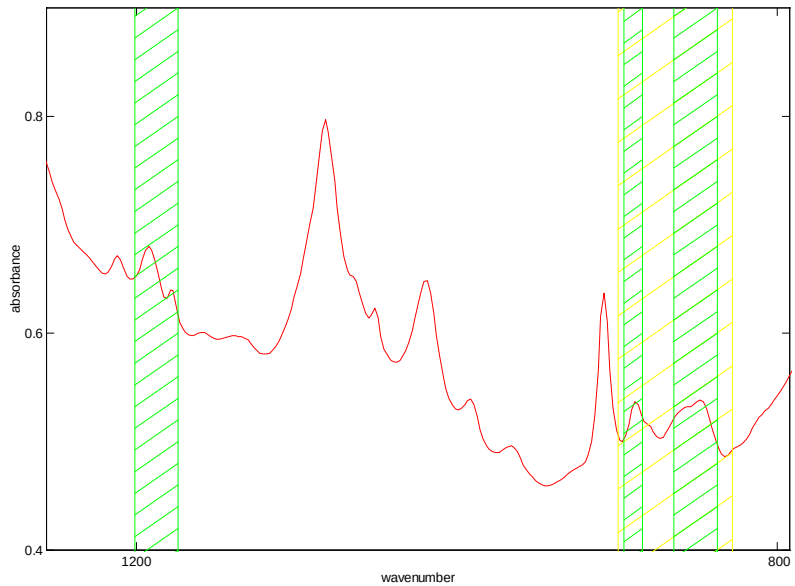
	Additive B	Additive C
RMSEP (GA)	48	47
RMSEP (expert)	54	48
regions (GA), cm-1	*3634-3616* *3506-3485* 1906-1884 1662-1645 1493-1487 644-623	1200-1175 *895-885* *864-839*
regions (expert)	3600-3260	899-829



Additive C

Experts know three peaks in 800-1200cm-1 represent various forms of the additive. Experts didn't know which and how many regions to include. GA selected one more than the experts ultimately selected.





- **Variable Selection is a very important step for developing a good multivariate model.**
- **This tool provides an automated approach when expertise is not available or the variables are not known (ex. Octane number).**

OPTIMIZATION STRATEGIES

The “standard” approaches (experimental design, steepest ascent, simplex...) work well with relatively simple problemes but fail when the complexity of the problem is too high (e.g., trapped on local maxima)

When does a problem become “very complex”?

- High number of independent variables

- Very complex or irregular response surface

- Presence of discontinuities in the experimental domain

- Response to be optimised function of several “subresponses”

WHAT TO DO WITH A "VERY COMPLEX PROBLEM"?

The only way to be "sure" of finding the global maximum would be a "grid search". Problem: the number of measurements increases very fast as the number of variables increases

An alternative strategy:
- performing experiments at random points
- retaining the points giving a good response
- trying to improve the response by somehow using the obtained information (local search around the randomly selected point, or exploiting what the best points have in common).

GENETIC ALGORITHMS

Genetic Algorithms (GA) mimick the evolution of a species according to the Darwinian theory ("**survival of the fittest**").

The fitness to the environment is function of the genetic material, (the result of an experiment is function of the experimental conditions).
Genetic material →experimental conditions

The genetic material is defined by the genes (an experimental condition is defined by the values of the variables involved in the experiment).
Genes ⟶ variables

The information contained in each gene is defined by a sequence of nitrogenated bases; 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).
Nitrogenated bases ⟶ bits

Each experimental condition, coded by a **sequence of 0's and 1's**, is treated as the genome of an individual, whose "performance" is considered as its "fitness"

CODING

How to code the following experimental condition?
- reaction temperature: 30°C
- reaction time: 20 minutes
- stirring: yes
- catalyst: type A (A and B possible catalysts)

011110 10100 1 0
(blanks have been added only to make genes evident)

- 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)

- the number of bits for each gene can be very different

CODING

011110 10100 1 0

011110 = 30°C

This means that:
- the range is between 0 and 63°C
- the difference between two levels is 1°C

This coding is reasonable if:
- we are interested in studying the reaction from 0 to 63°C
- the difference of one degree is significant (the reaction at 25°C can be different from the reaction at 26°C)
- 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)

If 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

If for the variable time we are interested in the range 10-40 minutes, with an interval between levels of two minutes, 16 levels, and therefore four bits, will be required

Final coding: 001 0101 1 0 (nine bits and 512 possible combinations), instead of the original 13 bits and 8192 combinations (search complexity reduced by a factor of 16)

Operators of a classical GA:

Creation of the original population

Select-copy: simulates the fights for mating, in which the best individuals have the highest probability of success, and therefore of spreading their genome

Cross-over: simulates the mating between two individuals, producing two offsprings, whose genetic material is derived from that of the two parents

Mutation: as in nature, rarely occurring random phenomena, producing random changes in the genetic material

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).

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.

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

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

Reproduction

After having created the original population (or first generation), the individuals start "mating" and "producing offspring."

Two basic concepts common to all the GAs:

- The probability of the best chromosomes (the ones giving the best responses) of producing offspring is higher than that of the worst chromosomes

- The offspring originated by their "mating" are a recombination of the parents.

First step: creating the population of the generation $x+1$ simply by randomly copying p times a chromosome of the generation x, taking into account the response of the individuals, giving the best ones a higher probability (e.g., $p_i = \text{resp}_i / \Sigma \text{resp}_i$).

Sort the population and compute the selection probability:

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

Draw 10 random numbers between 0 and 1:
0.353 0.038 0.367 0.324 0.414 0.903 0.150 0.353 0.428 0.915

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 ten 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

For each gene a random number is drawn, determining to which offspring the genes of the parents will be assigned. 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

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.

Mutations

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.

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

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.

VARIABLE SELECTION METHODS:

"UNIVARIATE": select those variables that have the greatest correlation with the response

"SEQUENTIAL": select the best variable and then the best pair formed by the first and second and so on in a forward or backward progression. A more sophisticated approach applies a look back from the progression to reassess previous selections

"MULTIVARIATE (PLS-ORIENTED)": Interactive Variable Selection, Uninformative Variable Elimination, Iterative Predictor Weighting PLS, Interval PLS, ...

GENETIC ALGORITHMS

AN EXAMPLE OF GA APPLIED TO FEATURE SELECTION

(for sake of simplicity, assume 10 variables)

chromosome 1: 0010011001 (model made by variables 3, 6, 7, 10)

chromosome 2: 1000110011 (model made by variables 1, 5, 6, 9, 10)

Cross-over: genes 1, 4, 6, 8 are swopped

offspring 1: 1010011001
offspring 2: 0000110011

Mutation: gene 2 of offspring 2 is mutated

offspring 1: 1010011001 (variables 1, 3, 6, 7, 10)
offspring 2: 0100110011 (variables 2, 5, 6, 9, 10)

The main problems of "Classical GA"

- **overfitting**

- **lack of reproducibility**

When applied to spectral data sets (as any other selection method)

- **non "spectroscopically logical" selections ("dispersed" wavelengths rather than regions)**

Modifications have been made to the standard GA in order to:

- **make it more suitable to the feature selection problem**
- **reduce the risk of overfitting**

Further modifications have been made to make it especially suitable for spectral data sets

Detailed description of the algorithm goes well beyond the scope and the time of this talk

Data set APPLE JUICES:

- **367 German apple juices from three different years (1999, 2000, 2001)**
- **Training set: 229 samples (1999, 2000)**
- **Validation set: 138 samples (2001)**
- **7 responses**
- **FT-IR spectra (1054 wavelengths) by Wine Scan FT120 (Foss Electric A/S) (only wavelengths 1-550 are taken into account)**

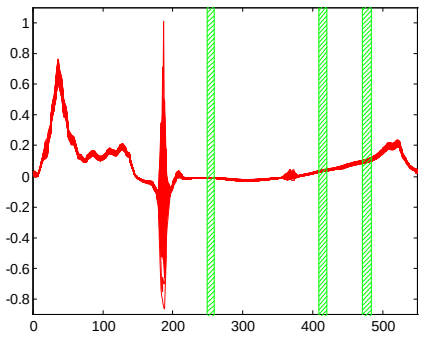
Research Institute of Geisenheim (Germany), Department of Wine Analysis and Beverage Research

GOAL OF THIS STUDY

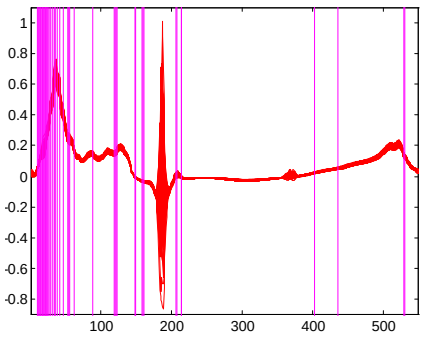
Compare GA to a commercial package for variable selection (Foss) in what concerns:

- **predictive ability**
- **interpretability of the selected wavelengths**

TOTAL ACIDITY AS TARTARIC

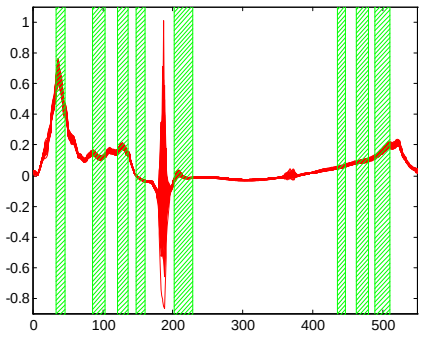


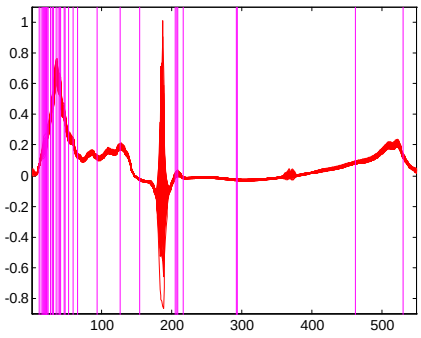
GA (RMSEP 0.3)



Wine Scan (RMSEP 0.5)

TEAC

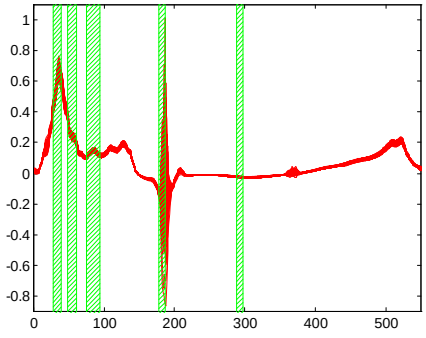


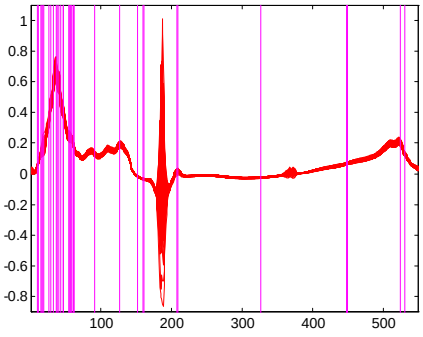


GA (RMSEP 1.1)

Wine Scan (RMSEP 1.3)

BRIX





GA (RMSEP 0.12)

Wine Scan (RMSEP 0.09)

COMPARISON OF THE PREDICTIVE ABILITY

	RMSEP			signif.		
	GA	WS	PLS	GA-WS	GA-PLS	WS-PLS
Total ac. (Malic)	0.21	0.25	0.36	* (G)	*** (G)	*** (W)
Total ac. (Tartar.)	0.3	0.5	0.4	*** (G)	* (G)	** (P)
Brix	0.12	0.09	0.11	*** (W)		* (W)
Extract	0.4	0.7	0.6	*** (G)	*** (G)	
Folin C	149	137	164			* (W)
pH	0.05	0.06	0.07		** (G)	
TEAC	1.1	1.3	0.9	* (G)		*** (P)

Data set PINE SEEDS:

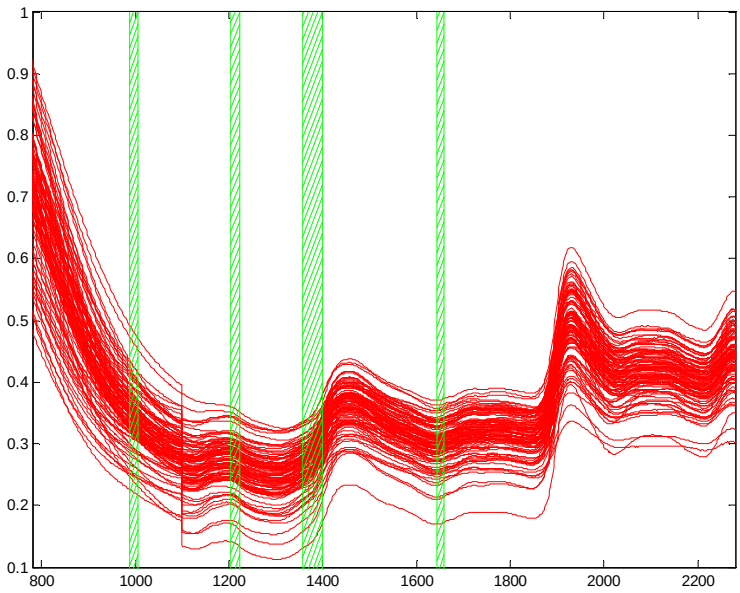
- **Moisture measured on 155 single seeds of Scots pine (*Pinus sylvestris* L.)**
- **Training set: 103 samples**
- **Validation set: 52 samples**
- **NIR spectra (751 wavelengths in the range 780-2280 nm) by NIRS 6500 (NIRSystems, Silver Spring, MD, USA)**

Torbjörn Lestander (Dept. of Silviculture, Swedish University of Agricultural Sciences, Umeå) and **Paul Geladi** (Unit of Biomass Technology and Chemistry, Swedish University of Agricultural Sciences, Umeå)

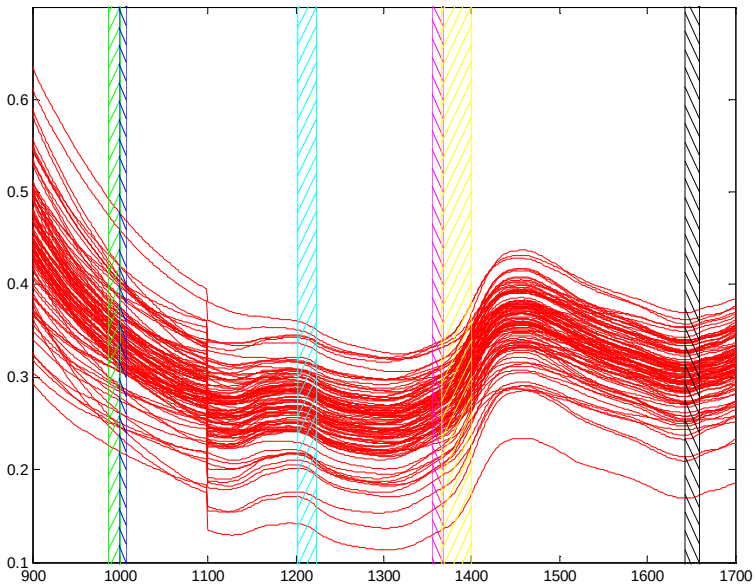
GOAL OF THIS STUDY

Select wavelengths that could be used in a few NIR filter sensors to predict moisture content in single seeds of Scots pine.

The results are of importance to the construction of an apparatus that uses parallel NIR-sensors for automatic and fast moisture determinations of conifer seeds.



RMSEP full spectrum: 1.9; RMSEP selected regions (50 wl.): 1.6



RMSEP full spectrum: 1.9; RMSEP six uniform density filters: 2.1

CONCLUSIONS

The application of GA as a technique of wavelength selection produced models that

- **were able to emulate region choices of experts**
- **gave results better than a well-known commercial software (lower RMSEP, better interpretation of selected wavelengths)**
- **allowed to detect relevant regions for the construction of filter instruments**