Optimal Optimisation in Chemometrics

Een wetenschappelijke proeve op het gebied van de Natuurwetenschappen, Wiskunde en Informatica

Proefschrift

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Nijmegen
op gezag van de Rector Magnificus,
prof. dr. C.W.P.M. Blom,
volgens besluit van het College van Decanen
in het openbaar te verdedigen
op donderdag 24 juni 2004
des namiddags om 1.30 uur precies

door

Joseph Albert Hageman

geboren op 10 april 1974
te Haarlem

Promotor:
Prof. Dr. L.M.C. Buydens

Co-promotor:
Dr. R. Wehrens

Manuscriptcommissie:
Prof. Dr. E. Pretsch , Swiss Federal Institute of Technology Zurich
Prof. Dr. A.K. Smilde , Universiteit van Amsterdam
Dr. R. de Gelder

The work presented in this thesis was supported financially by the Dutch Organisation for Scientific Research (NWO).

ISBN 90-9017942-9

Printed by Ponsen & Looijen, The Netherlands.

Contents

Chapter 1

Introduction

1.1 Optimisation

Chemometrics is a discipline which originates from analytical chemistry. Chemometrics is involved in the analysis of chemical data. By applying tools from mathematics, statistics and computer science, the goal of chemometrics is to increase the information yield of chemical data [1]. With the availability of more sophisticated measuring devices, the nature of chemical analysis changed from measuring direct observable properties (e.g. colour and weight) to indirect observable properties (e.g. spectral intensities). At a certain moment, analysing this new data became as difficult and important as the measurements itself and this is where chemometrics justifies its existence.

One of the first research areas of chemometrics is the optimisation of experimental conditions, also known as experimental design. For a (chemical) process to function properly, usually a few variables have to be set. Finding optimal settings is not that difficult when only a few variables are involved, but when a larger number of variables is involved and these variables have interaction effects, finding correct settings is not straightforward anymore. Experimental design gives the possibility, by performing a minimal number of experiments, to assess the influence and to find correct settings for a given process variable.

The introduction of computers in chemistry gave rise to new possibilities. A new type of experiment became possible, the *in silico* experiment. Instead of performing a minimal number of (slow) experiments, it now became possible to perform many experiments in a relatively short period of time inside the computer. The use of computers also lead to different optimisation

strategies. Several new optimisation algorithms were invented: Simulated Annealing (SA), Genetic Algorithms (GA's) and Tabu Search (TS), all using the potential that computers had to offer.

Together with neural networks, SA and GA's form the natural computation methods. SA, GA's and TS are global optimisation methods. All three methods are able to optimise a cost function $F(x)$ by searching the parameters space of x for the optimal combination of x.

SA is an optimisation technique which stems from statistical physics. The principles were first described in 1953 by Metropolis [2]. In 1983 Kirkpatrick et al. [3] recognised the importance for general optimisation. The original technique was used to make predictions about the behaviour of a multi-particle system when the temperature was lowered. The physical system will reach a stable global ground state with minimal energy. For general optimisation purposes, predefined given cost function can be regarded as the energy of a system. By carefully lowering the temperature, the system can be trapped in a low energy conformation.

Although evolution can be considered much older, the principle for using it as an optimisation technique were first described by Holland in 1975 [4]. GA's mimic the biological process of evolution. Unlike SA or TS, GA's operate on a group of solutions (called a population) at a time. The best solutions of each population are modified by applying evolutionary inspired operators, such as crossover and mutation. By discarding not-so-good solutions and focussing on the better performing members of the population, the population as a whole moves towards the global optimum.

TS originates from the 1980's with its foundations being laid in the late 1970's by Glover et al [5,6]. TS is based on concepts from artificial intelligence. It uses basic, problem specific operators to explore a search space. Parts of the search space which have been examined are stored in memory (called the tabu list) and are not visited again. By guiding the optimisation to unknown parts of the search space, TS hopes to reach the global optimum. TS is a relatively new technique, especially within chemometrics. TS has traditionally been applied on combinatorial problems. TS can be used on continuous functions by choosing some sort of discretisation step. Applications in literature involve integer programming, routing and scheduling.

1.2 Optimisation categories

Optimisation problems can appear very diverse, but nevertheless they can all be divided into three basic categories [7]:

1. Parameter optimisation
   In this type of optimisation problem, parameter values, used as input for modelling, are optimised. Usually, the goal is to find settings for a particular model so the model is able to match experimental output. Analysis of the model parameters, lead to knowledge of the experimental system. Usually, only one set of values for the variables gives output which matches the experimental data.

2. Variable selection (also called subset selection)
   In this class of problems, the goal is to select a small group of variables from a large group. This subset usually has identical properties compared to the original, larger group of variables or performs even better. It can happen that many subsets exist which have the desired performance.

3. Sequence optimisation
   When dealing with sequence optimisation, no parameter values have to be set or variables to be selected. In this case, the optimal sequence of objects, data, processes or states that has the optimal value with respect to a certain property, has to be found. In this type of problem, there is usually only one correct solution.

1.2.1 Parameter optimisation

In parameter optimisation problems, the goal is to determine the optimal values for model parameters. When these parameters are used, the output of the model fits best to measured data. Analysis of the model parameters usually leads to a better understanding of the experimental system.

A well known example of parameter optimisation within analytical chemistry is curve fitting. The goal of curve fitting is to resolve overlapping peaks in a chromatogram or spectrum. Two steps are required to resolve a spectrum in its individual peaks. First an estimate has to be made of the number of peaks and their corresponding peak width, peak height and peak position. The next step is the parameter optimisation [8].

The fitting of strongly overlapping X-ray equator diffractometer scans of poly(ethylene naphtalate) indicated that local optimisers can perform reasonably, but only under strict constraints. Global optimisers, such as GA's do not require these estimates and are a logical tool [9]. A different example is the characterisation of thin layered materials by glancing incidence x-ray reflectometry with GA's [10].

When the underlying physics of a chromatogram or spectrum is known and can be calculated by using a model, curve fitting can be replaced by a somewhat different analysis. In this case, model parameters, other than peak width, peak height and peak position, are adjusted in such a way that the output of the model matches the experimental data. Examples are the analysis with the use of GA's of multilayer films with the use X-ray fluorescence [11] and the indexing of powder diffraction patterns with GA's by Paszkowicz [12] and later by Kariuki et al. [13]. After powder pattern indexing, GA's have also proven useful for the elucidation of a crystal structure [14, 15]. Martin et al. used the GA's for designing optical multilayers [16]. Westhead et al. successfully applied GA's and TS to a molecular docking problem [17]. GA's are also used by Forshed et al. [18] for peak alignment of NMR signals. Kasat et al. have described an application of GA's in polymer science and engineering for optimising reaction time and minimisation of reaction side products [19].

1.2.2 Variable selection

The goal of this type of optimisation problem, is to select a small group of variables from a larger group. The subset contains identical properties compared to the original group of variables or performs even slightly better for a given task. Selecting a subset of variables from a large collection is a task that cannot be easily done exhaustively. When selecting n variables from a collection of m variables, the number of possible subsets is given by:

$$\frac{m!}{n!(m-n)!} \tag{1.1}$$

The number of possibilities increase drastically even for small values of n and m.

A very well known example of variable selection problems in chemistry is wavelength selection prior to (PLS/PCR/MLR) modelling. In this particular case, m in Eq. 1.1 are spectra (consisting of hundreds of wavelengths) and n is usually a small percentage of m. The goal of wavelength selection is

to create better performing and easier understandable models by eliminating non-contributing wavelengths. Exhaustive searches for wavelength selection are not feasible with a reasonable number of wavelengths. Traditional methods, such as forward selection and backward elimination, use simple heuristics for locating an optimal solution but these solutions are likely not the best possible solutions. The application of global optimisation methods has greatly improved this research area. Comparative papers have been published, testing both the performance of SA and GA's on wavelength selection [20], giving the preference to GA's. While others demonstrate an equal performance between SA and GA's [21, 22]. Wavelength selection prior to (PLS) modelling is a much used practice, which can easily increase the predictive power of models. It is not surprising many chemometrical papers have been published applying this technique. Most of the recent examples use GA's for wavelength selection and this appears to be successful. In a number of papers, Leardi et al. describes properties of wavelength selection [23–25] and the application for prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data [26]. To mention only a few, researchers have reported beneficial effects of wavelength selection in Raman spectroscopy [27], several polarographic and stripping voltammetric data sets [28] and pyrolysis mass spectrometry [29]. Only very recently, TS has been used for variable selection in a study to assess the influence of different validation techniques [30, 31].

Feature selection for determining quantitative structure activity relationships (QSAR) is not so widespread compared to wavelength selection but it is rather similar in use. Hemmateenejad et al. [32] used GA-based variable selection to select relevant theoretical descriptors to optimally model the calcium channel activity of nifedipine analogues. Others apply feature selection to reduce the number of data points originating from one descriptor, for instance to reduce the number of points located on the electron density distribution of the molecules [33]. A selection of local and global descriptors is also possible [34]. Zupan et al. use feature selection for reducing a 120 parameter set to 15, using the correlation between the experiment output and a neural network model output as the evaluation function [35]. Xu et al. [36] compared different heuristic feature selection techniques with GA's using a QSAR data set.

A somewhat different goal for feature selection is data reduction. In this particular case, a representative subset is created from the complete data set, usually for building a model or clustering where the use of the complete data set would be troublesome.

Guo et al. proposed a feature selection method to select a subset of vari-

ables in sequential projection pursuit (SPP) analysis. The inhomogeneity of the complete data was explored by SPP, and a subset, retaining inhomogeneity information, was obtained by applying a GA [37]. A similar application was used in principal component analysis (PCA) [38] and in parallel factor analysis (PARAFAC) [39]. Other examples of feature selection for obtaining representative subsets with GA's are [40] and [41].

1.2.3 Sequence optimisation

Sequence optimisation problems are rarely encountered in chemistry, although not uncommon in other research areas. The perhaps well known travelling salesperson problem is an example of a sequence optimisation problem. Due to its rareness in chemistry, it is not possible to distinguish different types of problems. Only a few examples are found in literature and most of these problems do not involve any of the natural computation methods. Some examples of chemical related sequence problems are given below.

Massart et al. discuss the chromatographic separation of a three component mixture using a separation based ion-exchange chromatography [42]. In production level chromatographic separations, one or more rapid and clear-cut separation steps are used to resolve a mixture. Separation is better if a mixture of three components is separated in three steps where one component is eluted in each step, rather than one step where there is an overlap between components. The sequence of separation steps can be varied, giving rise to several possible sequences.

A second chemical sequence problem is the *shortest synthesis route* problem. In this case, the sequence consists of a number of sequential reactions which lead to a desired product. Alternatively, the problem can be reformulated as finding the synthesis route for a target compound in the fewest possible steps. To reduce the number of possible synthesis routes, an important methodology is the retrosynthesis [43]. Computer based systems for designing retrosynthesis are LHASA (Logic and Heuristic Applied to Synthetic Analysis) [44].

The Chemical Batch Scheduling (CBS) problem is a specific version of the Job Shop Scheduling (JSP) problem. In the chemical setting, this problem is formulated as follows: a set of orders for chemicals has to be processed on a set of facilities. For each order a given amount of a chemical must be synthesised before a certain deadline. The production consists of a sequence of reaction processes where each process has to be performed by a reactor suitable for that

process. The problem is to assign the processes to the facilities, splitting them into batches, and to schedule these batches in order to produce the demands within the given deadlines. Articles discussing the CBS problem use global optimisation methods as TS [45] and SA in a hybrid system [46]. Chang et al. uses a hybrid system of GA's and TS for the JSP [47]. A review by Cheng, Gen and Tsujimura summarises the representation and search strategies in GA for the standard JSP [48]. A comprehensive literature review on the CBS problem is written by Rippui [49] and Reklaitis [50].

Elucidation of the 3D structure of proteins can be accomplished with two- or multidimensional nuclear magnetic resonance (NMR). Wehrens et al. presented an automatic system for the interpretation of two-dimensional NMR spectra of proteins, HIPS [51]. A GA is used to tackle a search problem of huge dimensions in which spin patterns found in the NMR spectra should be mapped to amino acids in the sequence. Other global optimisation methods which have been used to analyse the spin pattern sequence are SA [52, 53] and neural networks [54].

Another sequence optimisation problem is the analysis of DNA sequence. To elucidate the sequence of the nucleotides in the DNA, the DNA is treated with small primers (consisting of several oligonucleotides) which randomly hybridise with the DNA and generate random fragments. The small fragments can be analysed. The optimisation problem reduces to identifying the overlap between the small DNA fragments. When combining the overlap information, long DNA segments can be reconstructed [55].

1.3 Problems in global optimisation

Several parts can be distinguished in SA, GA's and TS that are identical for all three algorithms. These parts are shown in figure 1.1. Since there are many tutorials describing the inner workings of SA [3], GA's [56] and TS [5,6], this section only summarises the inner workings of the three optimisation algorithms. Besides a summary, this section also stresses current problems which are associated at each stage of the optimisation algorithms.

- Choice of optimisation algorithm

  Although this is not part of an actual algorithm, this part is still very important. Before starting any optimisation, it has to be decided which optimisation algorithm is going to be used. This choice will depend on the type of optimisation problem and a possible representation of

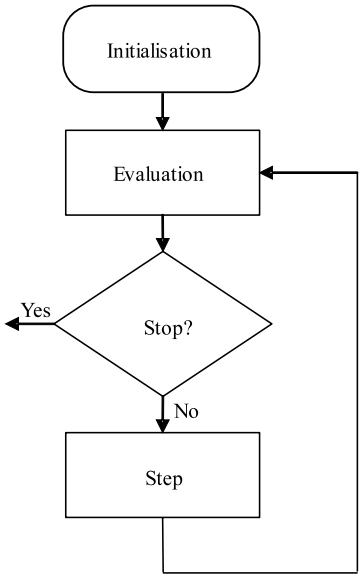

Figure 1.1: General flowchart of an optimisation algorithm.

the optimisable parameters. Some guidelines exist, for instance it has been concluded that GA's are very well suited for optimisation problems where the parameter space contains forbidden areas [56]. But for other problems, this choice may not be so clear.

- Initialisation

  The first step is the initialisation. Before starting the algorithm, values for the meta-parameters have to be chosen. This part of the initialisation is very important and determines for a large part the success of the optimisation. Examples of meta-parameters are the temperature in SA, the population size in GA's and the tabu list length in TS. Choosing suboptimal values for the meta-parameters will lead to an ill performing optimisation. Usually some sort of meta-optimisation is performed to choose correct values.

  In the case of SA and TS, a starting solution has to be chosen. In the case of GA's, a group of starting solutions has to be chosen. The starting solution(s) can be completely random or can be based on prior knowledge and carefully constructed.

- Evaluation

  The evaluation function determines the fitness (or suitability) of a trial solutions. The fitness of a trial solution is expressed in one single number. Depending on the fitness of solutions, the course of the algorithm will be decided.

  The evaluation function determines the shape of the search landscape. In an ideal case, a search landscape contains one broad optimum. In practice however, many local optima usually exist which make it hard to detect the global optimum. When a high number of local optima is present, the chance increases of converging to a (undesired) local optimum instead of the correct global optimum. Carefully designing an evaluation function will certainly yield a more smooth and clearer search space.

  Each optimisation problem requires a specific evaluation function. A specific type of evaluation function is the evaluation function that compares two spectra (or any other vector-like data). In this comparison, the similarity between a calculated spectrum (originating from the optimisation procedure) and a predefined spectrum is assessed. This predefined spectrum corresponds with the desired solution, which is usually an experimental spectrum. This type of comparison can be needed for optimisation problems in the parameter optimisation class. A conventional method for assessing the similarity between two spectra is to calculate a difference spectrum. By summing (and sometimes weighting) this difference spectrum, the similarity can be expressed in one value [10, 57]. A drawback with this type of evaluation function is that it is based on a pointwise comparison. When peaks in a spectrum shift only slightly in position, peaks do no longer overlap with the corresponding peak in the reference spectrum. In this case, a pointwise comparison will no longer recognise any similarity between two spectra. An evaluation function based on a pointwise comparison will fail to recognise similar spectra without substantial overlap between corresponding peaks.

  The evaluation function encountered in the variable selection class, is usually simpler compared to the parameter optimisation class. In the case of wavelength selection or variable selection, the quality of the subsets of variables is assessed. A much used technique to estimate the predictive ability of models is Leave-One-Out Cross-Validation (LOO-

CV). In LOO-CV, each object is removed once from the training set and used as a validation set. However, different studies pointed out that LOO-CV leads to overfitting [30,31,58]. A better technique would be to use Leave-Multiple-Out Cross Validation (LMO-CV) [30]. In this case, multiple objects are left out of the training set and treated as the validation set. When looking at computational time, LMO-CV is also faster, since it requires the use of less (but somewhat larger) validation sets. If PCR or PLS is used as a regression technique, the number of latent variables must also be chosen. The optimal number of latent variables is usually taken as the number where the RMSECV no longer significantly decreases [26,59].

- Stop ?

  The optimisation continues, usually until a predefined number of iterations has been used or when a certain result has been obtained. It is important not to stop to soon, when better results still can be obtained, but on the other hand, it is also important not to continue too long when the best solution is not going to change anymore.

- Step

  To investigate the possible existence of better performing solutions compared to the ones already found, new solutions are created and examined. The step function determines how these new solutions are created and thus how the search space is examined.

  In the case of SA, the next solution is chosen randomly from all the possible solutions that can be created by changing the current solution with a single step. Not all the possible solutions are generated and examined, but one is chosen randomly. This solution is accepted as the new solution if it performs better than the previously encountered solution. If the solution performs worse, it can still be accepted with a certain chance. The chance of acceptance is higher when the SA-temperature is higher, when the temperature is lowered, the chance of acceptance decreases.

  In the case of TS, all possible steps that can be taken from the current position, are generated and examined. The next solution will be the

solution that performs best of all these neighbouring solutions. However, this solution cannot be chosen, if it is present in the memory of TS (which is called the tabu list). This presence indicates that the solution has already been visited. This feature ensures that TS can overcome local minima by forcing it to accept solutions worse than the current solution. Newly visited solutions are added to the memory of TS.

In the above cases, it is up to the user to create a step function. There are usually many step functions possible, but not all are likely to yield good results. By taking too large steps, there is a risk that there is no cohesion between successive steps. When this happens, the algorithm cannot sense better performing areas in search space and the global optimum can be overlooked. By taking too small steps, the explorative power of the optimisation algorithm is limited to a small area of the search space which could lead to missing the global optimum.

In GA's, the user does not have to come up with a step function, the algorithm takes care of that. New solutions are creating by using biologically inspired operators such as crossover (random combination of solutions) and mutation (random adjustment of variables). The search space is examined by replacing not so good performing solutions by modifications of more promising solutions. The use of a predefined step function in GA's requires a correct representation of the problem parameters. The choice of the representation determines which solutions are close to each other (which means that these solutions can be reached in one steps) or far apart in the search space (which means they are separated by several steps). In GA's, the predefined step functions may be advantageous, but this has only shifted the problem to the representation of problem parameters. This representation will greatly affect the search characteristics.

1.4 Objectives

Unfortunately, despite the many SA, GA's and TS examples, the use of global optimisation methods is still not straightforward in all cases. The previous section listed many difficulties which are still encountered when using global optimisation algorithms. This thesis aims at investigating and improving the use of global optimisation algorithms. Not all the problems from the previous section, such as the choice of optimisation method, initialisation and

GA representation, will be treated as exhaustively as the main focus of this thesis. This thesis will focus at three specific problems associated with global optimisation:

- Evaluation function

  This thesis aims at finding a similarity criterion which deals with the problem of correctly comparing spectra when many shifted peaks are present. An evaluation function of this type is encountered in the parameter optimisation class. It compares an experimental state with a theoretical state (e.g. an experimental spectrum and a theoretical one) and expresses the (dis)similarity in one single number. This comparison can be a problem when small changes in peak positions are present. Current comparison methods use a point wise comparison and will fail to find any similarity when no overlap between peaks is present. A correct similarity assessment is necessary for a correct optimisation of model parameters.

- Tabu Search

  SA, GA's and TS are global optimisation methods and all three methods should, in theory, be able to locate an optimum of a given optimisation problem. However, implementational aspects of the optimisation algorithms and problem specific characteristics could make some optimisation algorithm more effective than others.

  TS is an optimisation technique with different characteristics compared to SA and GA's. The existence and careful examination of a neighbourhood, requires the definition of a step function. This step function can be similar to the one used in SA. The use of memory in TS is not found in SA or GA's and it requires special bookkeeping. By implementing TS to solve several chemical optimisation problems, this thesis investigates the properties and the possibilities of TS.

- Meta-optimisation

  SA and GA's have been applied in the past to a variety of chemical optimisation problems. TS is a relatively new technique, with examples outside chemistry. Unfortunately, for all three methods, there exists no standard recipe on how or when to use SA, GA's or TS. For each optimisation problem, the meta-parameters have to be set correctly, a representation has to be chosen and an evaluation function has to be

devised if any useful results are to be obtained. By studying and solving several chemical optimisation problems, another goal is to detect guidelines on how to use them and when to use them.

The previous objectives are studied by solving four chemical optimisation problems, which are divided over the three different optimisation categories. The optimisation problems solved in this thesis, are not exhaustively solved with all three optimisation methods, as this is not a comparative study.

1.5 Outline of thesis

Chapter 2 deals with the problem of finding a similarity criterion which is able to deal with shifting peaks when comparing spectra or any other vector like data. In this chapter, the weighted cross correlation (WCC) is introduced. This criterion is able to recognise similar spectra even while shifts in peak positions are present. The WCC criterion is applied to the classification of crystal structures of cephalosporin complexes, based on the comparison of their calculated powder diffraction patterns. The WCC criterion will be used in chapter 3 and chapter 4 as basis for the evaluation function.

Based on the different optimisation problem categories, chapters 3 to 6 can be divided into the next three categories:

- Part A: Parameter optimisation

  Chapter 3 and 4 deal with problems which can be classified as parameter optimisation problems. A common aspect of both problems is the evaluation function. In both cases, it compares many theoretical and experimental states with the possibility of many shifted peaks. The weighted cross correlation, as introduced in chapter 2, provides the necessary approach to correctly deal with shifted peak positions.

  Chapter 3 deals with the elucidation of rotational constants from laser induced fluorescence spectra with the aid of GA's and the weighted cross correlation. Rotational constants are an important tool in determining the spatial geometry of molecules.

  Chapter 4 deals with the important first step of structure elucidation from (X-ray) powder diffraction patterns: powder pattern indexing. The aim of indexing is to retrieve the unit cell parameters of a sample. In

chapter 4, GA's in combination with the weighted cross correlation are used for indexing.

- Part B: Variable selection

  Chapter 5 investigates properties and possibilities of TS by applying TS to the well known problem of wavelength selection, a problem from the variable selection class. Several implementational aspects of wavelength selection with TS will be discussed. Eventually, two implementations of wavelength selection with TS are tested, one which searches for a solution with a constant number of wavelengths and one with a variable number of wavelengths. Optimising the meta parameters of TS will be done by using an experimental design. The performance and characteristics of both implementations are compared with results which are obtained by wavelength selection methods which are based on SA and GA's.

- Part C: Sequence optimisation

  Chapter 6 can partly be classified as a sequence optimisation problem. In this chapter, a new method for optimising and refining multilayer optical coatings (MOCs) will be presented. MOCs are coatings which consists of a stack of thin layers of materials with differences in refractive indices. Depending on the total number of layers, the composition and thickness of each layer, a MOC is able to reflect certain wavelengths while other wavelengths are transmitted unhindered. By carefully choosing an appropriate representation, the GA is able to determine the total number and the sequence of the materials in a MOC. Together with the sequence, the GA will provide an estimate of the optimal thickness of each layer. After the optimisation with GA's, the thicknesses will be refined by a method based on TS. Optimising the thicknesses can be classified as a parameter optimisation problem. In this approach, the starting solution of TS will be the best solution obtained with GA's.

Table 1.1 summarises the outline of the thesis.

Chapter	Subject	Method
1	Introduction	
2	Weighted cross correlation	
3, 4	Parameter optimisation	GA's
5	Variable selection	SA, GA's, TS
6	Sequence and parameter optimisation	GA's, TS
7	Conclusions and future prospects	

Table 1.1: Outline of this thesis, together with division in optimisation problem categories and optimisation methods.

References

[1] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, and J. Smeyers-Verbeke. *Handbook of Chemometrics and Qualimetrics: Part A*. Elsevier, Amsterdam, 1998.

[2] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. *J. Chem. Phys.*, 21:1087, 1953.

[3] S. Kirkpatrick, C.D. Gelatt Jr., and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[4] J. Holland. *Adaptation in natural and artificial systems.* The MIT Press, cambride, 1975.

[5] F. Glover. Tabu search – part I. *ORSA Journal on Computing*, 1(3):190, 1989.

[6] F. Glover. Tabu search – part II. *ORSA Journal on Computing*, 2(1):4, 1990.

[7] Carlos B. Lucasius. *Towards genetic algorithm methodology in chemometrics.* PhD thesis, Proefschrift Katholieke Universiteit Nijmegen, 1993.

[8] A.P. de Weijer, L. Buydens, G. Kateman, and H.M. Heuvel. Spectral curve fitting of infrared spectra obtained from semi-crystalline polyester yarns. *Chemometrics and Intelligent Laboratory Systems*, 28:149–164, 1995.

[9] A.P. de Weijer, C.B. Lucasius, L. Buydens, and G. Kateman. Curve fitting using natural computation. *Anal. Chem.*, 66:23–31, 1994.

[10] A.D. Dane, A. Veldhuis, D.K.G. de Boer, and L.M.C. Buydens. Application of genetic algorithms for characterisation of thin layered materials by glancing incidence X-ray reflectometry. *Physica B*, 253:254–268, 1998.

[11] A.D. Dane, P.A.M. Timmermans, H.A. van Sprang, and L.M.C. Buydens. A genetic algorithm for model-free X-ray fluorescence analysis of thin films. *Analytical Chemistry*, 68(14):2419, 1996.

[12] W. Paszkowicz. Application of the smooth genetic algorithm for indexing powder patterns - tests for the orthorombic system. *Materials Science Forum*, 228-231:19–24, 1996.

[13] B.M. Kariuki, S.A. Belmonte, M. I. McMahon, R.L. Johnston, K.D.M. Harris, and R.J. Nelmes. A new approach for indezing powder diffraction data based on whole-profile fitting and global optimization using a genetic algorithm. *J. Synchrotron Rad.*, 6:87–92, 1999.

[14] K. Shankland, W.I.F. David, and T. Csoka. Crystal structure determination from powder diffraction data by the application of a genetic algorithm. *Z. Kristall.*, 212:550–552, 1997.

[15] B.M. Kariuki, H. Serrano-Gonzalez, and R.L. Johnston. The application of a genetic algorithm for solving crystal structures from powder diffraction data. *Chem. Phys. Lett.*, 280:189–195, 1997.

[16] S. Martin, A. Brunet-Bruneau, and J. Rivory. Darwinian evolution of homogeneous multilayer systems: a new method for optical coatings design. *SPIE*, 2253:168–174, 1994.

[17] D.R. Westhead, D.E. Clark, and C.W. Murray. A comparison of heuristic search algorithms for molecular docking. *J. Comput.-Aided Mol. Des.*, 11:209–228, 1997.

[18] Jenny Forshed, Ina Schuppe-Koistinen, and Sven P. Jacobsson. Peak alignment of NMR signals by means of a genetic algorithm. *Analytica Chimica Acta*, 487:189–199, 2003.

[19] R. Kasat, A. Ray, and S. Gupta. Applications of genetic algorithm in polymer science and engineering. *Materials and Manufacturing Processes*, 18(3):523–532, 2003.

[20] C.B. Lucasius, M.L.M. Beckers, and G. Kateman. Genetic algorithms in wavelength selection: a comparative study. *Anal. Chim. Acta*, 286:135–153, 1994.

[21] U. Horchner and J.H. Kalivas. Further investigation on a comparative study of simulated annealing and genetic algorithm for wavelength selection. *Anal. Chim. Acta*, 311:1–13, 1995.

[22] U. Horchner and J.H. Kalivas. Simulated-annealing-based optimization algorithms: fundamentals and wavelength selection applications. *Journal of Chemometrics*, 9:283–308, 1995.

[23] R. Leardi, R. Boggua, and M. Terrile. Genetic algorithms as strategy for feature selection. *Journal of Chemometrics*, 6:267–281, 1992.

[24] R. Leardi and A.L. Gonzalez. Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemometrics and intelligent laboratory systems*, 41:195–207, 1998.

[25] R. Leardi. Application of genetic algorithm-PLS for feature selection in spectral data sets. *Journal of Chemometrics*, 14:643–655, 2000.

[26] Riccardo Leardi, Mary Beth Seasholtz, and Randy J. Pell. Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from fourier transform-infrared spectral data. *Analytica Chimica Acta*, 461(2):189–200, 2002.

[27] F. Estienne, D. L. Massart, N. Zanier-Szydlowski, and Ph. Marteau. Multivariate calibration with Raman spectroscopic data: a case study. *Analytica Chimica Acta*, 424(2):185–201, 2000.

[28] Ana Herrero and M. Cruz Ortiz. Qualitative and quantitative aspects of the application of genetic algorithm-based variable selection in polarography and stripping voltammetry. *Analytica Chimica Acta*, 378(1-3):245–259, 1999.

[29] David Broadhurst, Royston Goodacre, Alun Jones, Jem J. Rowland, and Douglas B. Kell. Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. *Analytica Chimica Acta*, 348(1-3):71–86, 1997.

[30] K. Baumann, H. Albert, and M. von Korff. A systematic evalution of the benefits and hazards of variable selection in latent variable regression. part I. search algorithm, theory and simulations. *Journal of chemometrics*, 16:339–350, 2002.

[31] K. Baumann, H. Albert, and M. von Korff. A systematic evalution of the benefits and hazards of variable selection in latent variable regression. part II. practical applications. *Journal of chemometrics*, 16:351–360, 2002.

[32] Bahram Hemmateenejad, Ramin Miri, Morteza Akhond, and Mojtaba Shamsipur. QSAR study of the calcium channel antagonist activity of some recently synthesized dihydropyridine derivatives. an application of genetic algorithm for variable selection in MLR and PLS methods. *Chemometrics and Intelligent Laboratory Systems*, 64(1):91–99, 2002.

[33] Bjørn K. Alsberg, Nathalie Marchand-Geneste, and Ross D. King. A new 3D molecular structure representation using quantum topology with application to structure-property relationships. *Chemometrics and Intelligent Laboratory Systems*, 54(2):75–91, 2000.

[34] U. Cosentino, G. Moro, D. Bonalumi, L. Bonati, M. Lasagni, R. Todeschini, and D. Pitea. A combined use of global and local approaches in 3D-QSAR. *Chemometrics and Intelligent Laboratory Systems*, 52(2):183–194, 2000.

[35] Jure Zupan and Marjana Novic. Optimisation of structure representation for qsar studies. *Analytica Chimica Acta*, 388(3):243–250, 1999.

[36] Lu Xu and Wen-Jun Zhang. Comparison of different methods for variable selection. *Analytica Chimica Acta*, 446(1-2):475–481, 2001.

[37] Q. Guo, W. Wu, D. L. Massart, C. Boucon, and S. de Jong. Feature selection in sequential projection pursuit. *Analytica Chimica Acta*, 446(1-2):85–96, 2001.

[38] Q. Guo, W. Wu, D. L. Massart, C. Boucon, and S. de Jong. Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems*, 61(1-2):123–132, 2002.

[39] W. Wu, Q. Guo, D. L. Massart, C. Boucon, and S. de Jong. Structure preserving feature selection in parafac using a genetic algorithm and

procrustes analysis. *Chemometrics and Intelligent Laboratory Systems*, 65(1):83–95, 2003.

[40] Yukio Tominaga. Representative subset selection using genetic algorithms. *Chemometrics and Intelligent Laboratory Systems*, 43(1-2):157–163, 1998.

[41] C. Pizarro Milln, M. Forina, C. Casolino, and R. Leardi. Extraction of representative subsets by potential functions method and genetic algorithms. *Chemometrics and Intelligent Laboratory Systems*, 40(1):33–52, 1998.

[42] D.L. Massart, C. Janssens, L. Kaufman, and R. Smits. Application of the theory of graphs to the optimisation of chromatographic separation schemes for multicomponent samples. *Anal.Chem.*, 44:2390–2399, 1972.

[43] E.J. Corey and X.-M. Cheng. *The Logic of Chemical Synthesis*. John Wiley and Sons, New York, 1989.

[44] E.J. Corey, A.K. Long, and S.D. Rubenstein. Computer-assisted analysis in organic synthesis. *Science*, 228:408–418, 1985.

[45] P. Brucker and J. Hurink. Solving a Chemical Batch Scheduling Problem by Local Search. *Ann.Op.Res.*, 96:17–38, 2000.

[46] P. Van Bael and M. Rijckaert. A practical decomposition approach for a chemical substance scheduling problem. *Int. J. Production Economics*, 68:161–167, 2000.

[47] P-T. Chang and Y-T. Lo. Modelling of job-shop scheduling with multiple quantative and qualitative objectives and a GA/TS mixture approach. *Int.J.Comp.Integr.Manufac.*, 14:367–384, 2001.

[48] R.W. Cheng, M. Gen, and Y. Tsujimura. A tutorial survey of job-shop scheduling problems using genetic algorithms, Part II: Hybrid genetic search strategies. *Comp.Ind.Eng*, 36:343–3664, 1999.

[49] D.W.T. Rippui. Batch Process Systems Engineering: a retrospective and prospective review. *Comp.Ind.*, 1993.

[50] G.V. Reklaitis. Overview of Scheduling and planning in Batch Process Operations. In G.V. Reklaitis, editor, *Batch Processin Systems Engineering*, pages 660–705. Springer Verlag, Berlin, 1996.

[51] R. Wehrens, C. Lucasius, L. Buydens, and G. Kateman. HIPS, a hybrid self-adapting expert system for NMR spectrum interpretation using genetic algorithms. *Anal.Chim.Acta*, 277:313–324, 1993.

[52] R. Bernstein, C. Cieslar, A. Ross, H. Oschkinat, J. Freund, and T.A. Holak. Computer-assisted assignment of multidimensional NMR spectra of proteins: application to 3D NOESY-HMQC and TOCSY-HMQC spectra. *J.Biomol.NMR*, 1993.

[53] N.E.G. Buchler, E.R.P. Zuiderweg, H. Wang, and R.A. Goldstein. Protein heteronuclear NMR assingments using mean-field simulated annealing. *J.Mag.Res.*, 125:34–42, 1997.

[54] B.J. Hare and J.H. Prestegard. Application of neural networks to automated assignment of NMR spectra of proteins. *J.Biomol.NMR*, 1994.

[55] D.E. Zimmerman, C.A. Kulikowski, Y. Huang, W. Feng, M. Tashiro, S. Shimotakahara, C-Y. Chien, R. Powers, and G.T. Montelione. Automated Analysis of Protein NMR Assignments Using Methods from Artificial Intelligence. *J.Mol.Bio.*, 269:592–610, 1997.

[56] R. Wehrens and L.M.C. Buydens. Evolutionary optimisation: a tutorial. *Trends in Analytical Chemistry*, 17(4):193–203, 1998.

[57] J. Dods, D. Gruner, and P. Brumer. A genetic algorithm approach to fitting polyatomic spectra via geometry shifts. *Chem. Phys. Lett.*, 261:612–619, 1996.

[58] H.A. Martens and P. Ardenne. *Chemometrics and Intelligent Laboratory Systems*, 44:91–121, 1998.

[59] H. Kubinyi. Evolutionary variable selection in regression and PLS analyses. *Journal of Chemometrics*, 10:119–133, 1996.

Chapter 2

A Generalised Expression for the Similarity of Spectra: Application to Powder Diffraction Pattern Classification

A generalised expression is given for the similarity of spectra, based on the normalised integral of a weighted crosscorrelation function. It is shown that various similarity and dissimilarity criteria previously described in literature can be written as special cases of this general expression. A new similarity criterion, based on this generalised expression, is introduced. The benefits of this criterion are that it properly recognises shifted but otherwise similar details in spectra and that the resulting similarity measure is normalised. Moreover, the criterion can easily be adapted to specific properties of spectra resulting from various analytical methods. The new criterion is applied to the classification of a series of crystal structures of cephalosporin complexes, based on comparison of their calculated powder diffraction patterns. The results are compared with those obtained using previously described criteria.

This chapter is published as: R. de Gelder, R. Wehrens, J.A. Hageman, J. Comp. Chem., 22(3), 273-289, 2003.

2.1 Introduction

Many chemical and physical methods for the analysis of compounds in solution and/or the solid state yield one-dimensional spectra or diagrams that consist of isolated and/or (partly) overlapping sharp peaks. Examples of such spectra are 1D-NMR spectra and powder diffraction patterns. In the case of powder diffraction patterns, the positions of the peaks are very sensitive to small deviations in unit cell parameters. This means that in the case of crystal structure prediction, or *ab initio* structure determination, strongly related structures may give (calculated) powder diffraction patterns that look similar from an overall point of view but differ significantly on a more local scale. The same situation may occur for isomorphous compounds that differ only slightly in unit cell volume or unit cell shape. These compounds may give experimental or calculated powder patterns that by visual inspection are definitely related and recognisable as isostructural compounds. Unfortunately, in all those cases the calculation of a reliable and objective measure of similarity or dissimilarity, even when the whole pattern is considered, is nontrivial.

A conventional method for the comparison of powder diffraction patterns is calculating the difference pattern and quantifying the dissimilarity between the patterns as the sum of the differences or the sum of the squared (and weighted) differences. In the field of Rietveld refinement such criteria of fit are known as R_p (R-pattern) and R_{wp} (R-weighted pattern) [1]. Because these measures are based on a pointwise comparison of patterns, small shifts in peak positions may result in a large (undesired) increase of the dissimilarity measures.

In the present article, it is shown that there is a simple relationship between the conventional criterion based on squared differences, the Pearson product-moment correlation coefficient [2] and the overlap integral described by Lawton and Bartell [3] who define a measure for the similarity of powder diffraction patterns on an absolute scalewhen these criteria are written in terms of the correlation function. With respect to the correlation function, the drawback of these criteria is that they only consider one point (the value at the origin) from the auto- and crosscorrelation functions and neglect the information that is present in the remainder of the auto- and crosscorrelation functions.

Karfunkel, Rohde, Leusen, Gdanitz, and Rihs [4] describe a method, that is based on the work of Stephensen and Binsch [5], in which they not only compare powder patterns pointwise but compare a point of one diagram with the neighbourhood of the corresponding point on the other diagram and vice

versa. As a matter of fact, Stephensen and Binsch transformed the conventional squared difference criterion into a criterion based on correlation functions, by introducing this neighbourhood concept in the comparison of patterns. This can easily be shown by rewriting their so-called fold in terms of auto and crosscorrelation integrals. A drawback of their criterion is that the resulting values are not on an absolute scale so that limits for acceptance can not easily be defined.

Although in principle all the information on the similarity of two patterns is contained in the crosscorrelation function, an additional function must be used to extract this similarity information. Such a function defines the effective neighbourhood and related weights in the target pattern that should be compared with the corresponding point on the reference pattern. The fold used by Karfunkel et al. contains such a function in the form of matrix **F**; however, it will be shown that a more convenient function can be used for this purpose. This alternative function is easier to tune with respect to the particular properties of the spectra of interest, because it contains only one adjustable parameter.

Because powder diffraction patterns, and also 1D-spectra from many other analytical methods, are on an arbitrary scale, an obvious question is how to scale the patterns before applying a (dis)similarity criterion. Karfunkel et al. scale the patterns by equalising the total number of counts, which is the same as normalising the area under the patterns. This choice was not based on specific arguments, although it is clear that for closely related structures the patterns should have a similar number of counts. In this article it is shown that on the basis of crosscorrelation and autocorrelation integrals you should scale the patterns according to the self-similarities of the patterns. In practice, this leads to almost the same scaling as proposed by Karfunkel et al. However, in principle, scaling according to self-similarities may lead to a different sum of counts for each individual pattern.

The similarity and dissimilarity criteria described above can be written as a special form of a generalised expression for similarity, based on normalised weighted auto and crosscorrelation functions. Using this generalised expression it can easily be seen that the various criteria only use a different weighting function and/or different normalisation factors. One of the major advantages of this generalised form is that it shows how to obtain a similarity measure on an absolute scale, given a suitable weighting function. Another advantage is that prescaling of the patterns becomes unnecessary.

Although the power of a newly proposed similarity criterion that is based

on this generalised expression is demonstrated for powder diffraction patterns corresponding to a series of crystal structures of cephalosporin complexes, the similarity concept presented in this article provides a general method for quantifying the match between spectra of various physical and chemical techniques for the analysis of matter. The applicability of the generalised expression for similarity lies in the field of pattern classification, data base searching and optimisation problems. Its recent successful use in the direct determination of molecular constants from rovibronic spectra with genetic algorithms is described by Hageman, Wehrens, De Gelder, Meerts, and Buydens [6].

2.2 Auto- and Crosscorrelation Functions

A function that describes the similarity (or overlap) of two patterns, which are here expressed as two continuous functions $f(x)$ and $g(x)$, as a function of the relative shift r between the patterns, is the correlation function. The maximum and minimum value of the relative shift r is determined by the interval for which the patterns $f(x)$ and $g(x)$ are measured or calculated. The autocorrelation function $c_{ff}(r)$ for a given reference pattern $f(x)$ is given by:

$$c_{ff}(r) = \int f(x)f(x+r)dx \tag{2.1}$$

The integral of $c_{ff}(r)$, the autocorrelation integral, is given by (see Appendix 2.8.1):

$$\int c_{ff}(r)dr = (\int f(x)dx)^2 \tag{2.2}$$

This shows that the area under the autocorrelation function $c_{ff}(r)$ is equal to the square of the area under the reference pattern $f(x)$.

Similar expressions can be defined for the autocorrelation function $c_{gg}(r)$ of target pattern $g(x)$:

$$c_{gg}(r) = \int g(x)g(x+r)dx \tag{2.3}$$

$$\int c_{gg}(r)dr = (\int g(x)dx)^2 \tag{2.4}$$

It can easily be seen from expressions (2.2) and (2.4) that setting the total number of counts for $f(x)$ and $g(x)$ to the same value also results in

setting the autocorrelation integrals for $f(x)$ and $g(x)$,which express the self-similarities of the patterns, to the same value. In this way it is possible to put the autocorrelation functions for $f(x)$ and $g(x)$ on an absolute scale (in principle, arbitrarily chosen) and compare the values of the crosscorrelation function $c_{fg}(r)$ with the values of the autocorrelation functions $c_{ff}(r)$ and $c_{gg}(r)$.

The crosscorrelation function $c_{fg}(r)$ for patterns $f(x)$ and $g(x)$ is defined in a similar way by:

$$c_{fg}(r) = \int f(x)g(x+r)dx \tag{2.5}$$

$$\int c_{fg}(r)dr = \int f(x)dx \int g(x)dx \tag{2.6}$$

From (2.6) it can be seen that the area under the crosscorrelation function $c_{fg}(r)$, the crosscorrelation integral, is always equal to the product of the areas under the patterns $f(x)$ and $g(x)$. The crosscorrelation function $c_{fg}(r)$ can be normalised by dividing it by the root of the product of the autocorrelation integrals [the product of the areas under the patterns $f(x)$ and $g(x)$], which makes prescaling of the patterns $f(x)$ and $g(x)$ unnecessary:

$$c'_{fg}(r) = \frac{c_{fg}(r)}{(\int f(x)dx \int g(x)dx)} \tag{2.7}$$

The integral of $c'_{fg}(r)$ will always be equal to 1. This means, however, that the crosscorrelation integral itself is *not* a measure for the similarity between $f(x)$ and $g(x)$. It is the shape of the correlation function $c_{fg}(r)$ [or $c'_{fg}(r)$] that contains the information on the similarity between patterns $f(x)$ and $g(x)$. In Figure 2.1, two different powder diffraction patterns are shown (corresponding to entries 2 and 20 of Table 2.2, which will be explained later). Their corresponding (normalised) auto- and crosscorrelation functions are shown in Figure 2.2. The areas under the correlation functions are the same; however, the different shapes of the curves clearly reflect the differences and dissimilarity of the patterns.

In the next sections it will be shown that various similarity and dissimilarity criteria that are described in the literature can be written in terms of auto- and crosscorrelation functions.

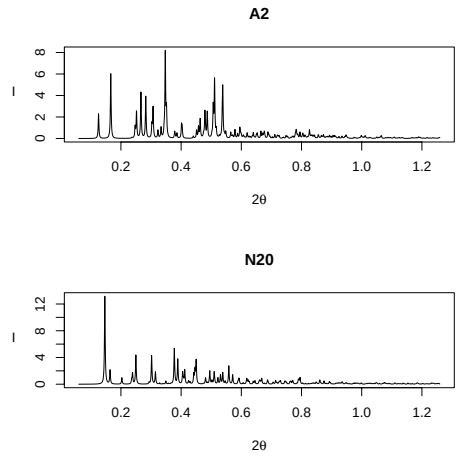

Figure 2.1: Example of two different powder diffraction patterns corresponding to the Cefradine/beta-nafthol complex (A2) and the Cefradine/methyl 3-hydroxybenzoate complex (N20) (see Table 2.2).

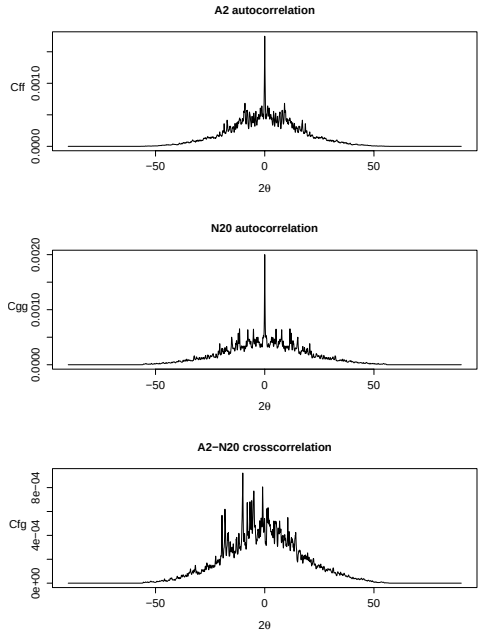

Figure 2.2: Auto- and crosscorrelation functions corresponding to the powder patterns shown in Figure 2.1.

2.3 Pointwise Similarity and Dissimilarity Criteria

A criterion often used for expressing the dissimilarity between two spectra or diagrams is the conventional pointwise criterion that includes the sum of the squared differences (see, e.g., Harris, Johnston, and Kariuki [7] or Dods, Gruner, and Brumer [8]). The difference criterion d_{fg} is given by:

$$d_{fg} = \int (f(x) - g(x))^2 dx \tag{2.8}$$

The difference criterion d_{fg} can be rewritten into an expression that only includes terms that are identical to the autocorrelation functions for $f(x)$ and $g(x)$ and the crosscorrelation function for $f(x)$ and $g(x)$ at $r = 0$ (see Appendix 2.8.2):

$$d_{fg} = c_{ff}(0) + c_{gg}(0) - 2c_{fg}(0) \tag{2.9}$$

Therefore, d_{fg} is only based on the values of the auto and crosscorrelation functions at $r = 0$ (no relative shift between the patterns is taken into account). Another criterion often used for expressing the similarity between two patterns or two vectors is the Pearson product-moment correlation coefficient r_{fg}: [2]

$$r_{fg} = \frac{\int (f(x) - < f(x) >)(g(x) - < g(x) >) dx}{(\int (f(x) - < f(x) >)^2 dx \int (g(x) - < g(x) >)^2 dx)^{\frac{1}{2}}} \tag{2.10}$$

where $< f(x) >$ and $< g(x) >$ are the mean values of patterns $f(x)$ and $g(x)$, i.e., $< f(x) >= \sum \frac{f(x_i)}{n}$ and $< g(x) >= \sum \frac{g(x_i)}{n}$ (n being the number of points x_i at which values for $f(x_i)$ and $g(x_i)$ are measured or calculated).

Defining the new patterns $f'(x) = f(x) - < f(x) >$ and $g'(x) = g(x) - < g(x) >$, this expression can also be written in terms of auto- and crosscorrelation functions (see Appendix 2.8.3):

$$r_{fg} = \frac{C_{f'g'}(0)}{(C_{f'f'}(0) C_{g'g'}(0))^{\frac{1}{2}}} \tag{2.11}$$

The value of r_{fg} is independent of the scale of the patterns $f'(x)$ and $g'(x)$, and thus independent of the scale of the patterns $f(x)$ and $g(x)$. r_{fg} can directly be used to express the similarity between two "unscaled" patterns $f(x)$ and $g(x)$. The most important conclusion is that the Pearson product-moment correlation coefficient is also based on the values of the auto- and

crosscorrelation functions at $r = 0$ only. However, the terms $< f(x) >$ and $< g(x) >$ in expression (2.10) introduce a shift of the original patterns $f(x)$ and $g(x)$ in the y-direction. Moreover, a specific scaling is implicitly applied to $f(x)$ and $g(x)$, while using criterion (2.10). When the values of $\int f'(x)^2 dx$ and $\int g'(x)^2 dx$ are normalised to 1, by scaling $f(x)$ and $g(x)$, the values of r_{fg} and $c_{f'g'}(0)$ will become the same. This is not the same as scaling by the total number of counts [setting the areas under $f(x)$ and $g(x)$ to the same value]. Instead, for a simple pointwise comparison of patterns based on the correlation function, $c_{fg}(0)$ divided by the root of the product of the autocorrelation integrals for $f(x)$ and $g(x)$ [see expression (2.7)] could be used:

$$c'_{fg}(0) = \frac{c_{fg}(0)}{(\int f(x)dx \int g(x)dx)} \tag{2.12}$$

This criterion can be used for "unscaled" patterns and measures the similarity on the basis of the crosscorrelation function in point 0, as if the patterns were prescaled according to the total number of counts. The value of $c'_{fg}(0)$ ranges from 0 to 1.

The Pearson product-moment correlation coefficient is closely related to the overlap integral $S_{\alpha\beta}$ that is described by Lawton and Bartell [3]. In principle, the method they propose is based on peak positions (lines) deduced from powder diagrams. By representing the diffraction peaks by Gaussian functions they simulate a profile that can be used to calculate an overlap integral. This overlap integral is a direct index of how well two patterns match each other.

It can easily be seen that the overlap integral is a normalised crosscorrelation function at $r = 0$ and is similar to expression 2.11, the Pearson product-moment correlation coefficient, when the simulated profile, based on lines, is replaced by the whole pattern, calculated or measured [see Appendix 2.8.4]. The only difference is a base-line shift [via the terms $< f(x) >$ and $< g(x) >$], which is not incorporated in the overlap integral. In other words, if the terms $< f(x) >$ and $< g(x) >$ are removed from expression (5a) one obtains the overlap integral of Lawton and Bartell:

$$S_{\alpha\beta} = \frac{c_{fg}(0)}{(c_{ff}(0)c_{gg}(0))^{\frac{1}{2}}} \tag{2.13}$$

To summarise for these pointwise criteria: the difference criterion can be transformed into a criterion, which includes terms that are identical to the values of the autocorrelation functions for $f(x)$ and $g(x)$ and the crosscorrelation function for $f(x)$ and $g(x)$ at $r = 0$. The difference criterion is, in principle,

the dissimilarity counterpart of the overlap integral of Lawton and Bartell if the patterns are scaled according to $\int f^2(x)dx$ and $\int g^2(x)dx$. If a baseline shift is applied to $f(x)$ and $g(x)$, according to $< f(x) >$ and $< g(x) >$, the overlap integral of Lawton and Bartell is transformed into the Pearson product-moment correlation coefficient.

2.4 Similarity and Dissimilarity Criteria Including Neighbourhoods

The pointwise difference criterion d_{fg} can be extended to a neighbourhood criterion by defining an expression $d_{fg}(r)$ in analogy with $c_{fg}(r)$ [see Appendix 2.8.5]:

$$\begin{aligned} d_{fg}(r) &= \int (f(x) - g(x+r))^2 dx \\ &= c_{ff}(0) + c_{gg}(0) - 2c_{fg}(r) \\ \int d_{fg}(r)dr &= c_{ff}(0) + c_{gg}(0) - 2\int c_{fg}(r)dr \end{aligned} \tag{2.14}$$

Although the dissimilarity criterion (2.14) contains two constants $c_{ff}(0)$ and $c_{gg}(0)$ and the same term as (2.6), this criterion is not just the dissimilarity counterpart of (2.6). When $f(x)$ and $g(x)$ are scaled on the basis of $\int f(x)dx$ and $\int g(x)dx$, the sum of the terms $c_{ff}(0)$ and $c_{gg}(0)$ may have a different value for each different pair of patterns, because these terms are related to the sum of the squared pattern values.

It will now be shown that criterion (2.14) shows close resemblance to the fold used by Karfunkel et al [4]. Their fold criterion is defined as: $\mathbf{d}^T\mathbf{F}\mathbf{d}$ where an element of the vector $\mathbf{d}$, $d(x) = f(x) - g(x)$.

The elements of the matrix $\mathbf{F}$ are defined as: $F_{ij} = 1/(1 + \alpha|i-j|^\beta)$. To compare this criterion with the crosscorrelation function we initially assume that all elements of $\mathbf{F}$ are equal to 1. This would mean that we use equal weights in the comparison of a point of the reference pattern with the neighbourhood of the corresponding point on the target pattern. It can easily be shown that in that case the fold criterion is a criterion based on correlation integrals only [see Appendix 2.8.6]:

$$\mathbf{d}^T\mathbf{F}\mathbf{d} = \int c_{ff}(r)dr + \int c_{gg}(r)dr - 2\int c_{fg}(r)dr \tag{2.15}$$

where $r = i - j$. Note the analogy of expression (2.15) with expressions (2.9) and (2.14).

If patterns $f(x)$ and $g(x)$ are scaled on the basis of the total sum of the counts, the fold criterion leads to the same but opposite results as criterion (2.6),which would mean a fold-value of 0 for any combination of $f(x)$ and $g(x)$, when all matrix elements of $\mathbf{F}$ are set to 1. Where (2.6) is a similarity criterion, (2.15) is the corresponding dissimilarity criterion. By introducing the concept of comparing one point on the reference pattern with the neighbourhood of the corresponding point on the target pattern, Karfunkel et al. transformed the conventional criterion, based on squared differences, to a correlation integral criterion. It may now be clear that, before calculating the fold, the scaling of the patterns *must* be the procedure proposed by Karfunkel et al. for reasons related to the normalisation of the autocorrelations integrals, the first and second term of (2.15). However, once the original matrix $\mathbf{F}$ of Karfunkel et al. is used, renormalisation of the crosscorrelation integral may be needed, as will be shown now.

Introducing the original matrix $\mathbf{F}$ into the expression for (2.15) would lead to [see Appendix 2.8.6]:

$$\mathbf{d}^T\mathbf{F}\mathbf{d} = \int w(r)c_{ff}(r)dr + \int w(r)c_{gg}(r)dr - 2\int w(r)c_{fg}(r)dr \qquad (2.16)$$

where $w(r) = 1/(1 + \alpha|r|^{\beta})(r = i - j)$.

The similarity counterpart of (2.16), expressing the similarity between the patterns using the same function $w(r)$, would be a weighted crosscorrelation integral:

$$\int c^w_{fg}(r)dr = \int w(r)\int f(x)g(x + r)dxdr \qquad (2.17)$$

and the associated autocorrelation integrals would include the same function $w(r)$:

$$\int c^w_{ff}(r)dr = \int w(r)\int f(x)f(x + r)dxdr \qquad (2.18)$$

$$\int c^w_{gg}(r)dr = \int w(r)\int g(x)g(x + r)dxdr \qquad (2.19)$$

An important conclusion is that the fold criterion of Karfunkel et al. can be seen as a dissimilarity counterpart of a weighted correlation integral $\int c_{fg}(r)dr$, as given in (2.6) [weighted with the function $w(r)$].

If the patterns $f(x)$ and $g(x)$ are scaled on the basis of the total number of counts, it may be clear from expressions (2.18) and (2.19) that the corresponding autocorrelation integrals do not necessarily result in the same value. On the other hand, it may now be clear that the matrix $\mathbf{F}$, or a different function $w(r)$, is needed to extract the similarity information from the crosscorrelation function. To ensure that the autocorrelation integrals will give identical values, the weighted crosscorrelation integral $\int c_{fg}^{w}(r)dr$ must be normalised to obtain a similarity measure C_{fg} on an absolute scale:

$$C_{fg} = \frac{\int c_{fg}^{w}(r)dr}{(\int c_{ff}^{w}(r)dr \int c_{gg}^{w}(r)dr)^{1/2}} \tag{2.20}$$

This similarity criterion will yield a value of 1 when patterns $f(x)$ and $g(x)$ are identical and a value between 0 and 1 for other cases. The corresponding dissimilarity criterion, which can be obtained from (2.15) by substituting C_{fg}, C_{ff} , and C_{gg} for $\int c_{fg}(r)dr$, $\int c_{ff}(r)dr$ and $\int c_{gg}(r)dr$, respectively, will yield values between 0 and 2. In fact, this dissimilarity criterion is a renormalised "fold."

2.5 A Generalised Expression for Similarity and Dissimilarity

All criteria described before can be summarised by the following expressions for similarity and dissimilarity. The generalised expression for the similarity S_{fg} between patterns $f(x)$ and $g(x)$ is given by:

$$S_{fg} = \frac{\int w_{fg}(r)c_{fg}(r)dr}{(\int w_{ff}(r)c_{ff}(r)dr \int w_{gg}(r)c_{gg}(r)dr)^{1/2}} \tag{2.21}$$

The corresponding generalised expression for the dissimilarity D_{fg} is given by:

$$D_{fg} = S_{ff} + S_{gg} - 2S_{fg} \tag{2.22}$$

The function $w_{fg}(r)$ determines the way in which the similarity information is extracted from the crosscorrelation function and the functions $w_{ff}(r)$ and $w_{gg}(r)$ determine the normalisation of the weighted crosscorrelation function via the autocorrelation functions. For obtaining a similarity or dissimilarity measure on an absolute scale the following condition must hold:

$$w_{ff}(r) = w_{gg}(r) = w_{fg}(r)$$

To include the neighbourhood in the comparison of points the weighting functions should be defined for $r \neq 0$. Both aspects are important and can easily be combined.

The differences between the various criteria described in literature can simply be explained by a different definition of the weighting functions $w_{ff}(r)$, $w_{gg}(r)$, and $w_{fg}(r)$. In Table 2.1 an overview is given of the various criteria and their corresponding use of the three weighting functions. From Table 2.1 it can be seen that none of the criteria described in literature include both the concept of neighbourhood and a correct normalisation to obtain a measure on an absolute scale. To define a similarity measure on an absolute scale in analogy with the fold of Karfunkel et al. the generalised similarity measure S_{fg} could be used with weighting functions:

$$\begin{aligned} w_{fg}(r) &= 1/(1 + \alpha|r|^{\beta}) \\ w_{ff}(r) &= w_{gg}(r) = w_{fg}(r) \end{aligned}$$

In principle, we now have defined a new criterion, a normalised fold, with different characteristics than the original fold. In the next section it is shown that the weighting function $w_{fg}(r)$, as defined for the fold, can be replaced by a simple triangle function.

2.6 Application of the Generalised Expression for Similarity to Structure Classification from Powder Diffraction Patterns

It has been shown that the various criteria for similarity and dissimilarity described in the literature can be deduced from the generalised expressions (2.21) or (2.22). To obtain a measure on an absolute scale, a suitable normalisation should be used, which is easily done by taking identical expressions for the weighting functions $w_{ff}(r)$, $w_{gg}(r)$, and $w_{fg}(r)$ in expressions (2.21 or (2.22). The question of which similarity (or dissimilarity) criterion can best be used for a given application now focusses on the choice of the weighting function $w_{fg}(r)$. This weighting function defines the neighbourhood and

Table 2.1: (Dis)similarity Criteria and Their Corresponding Weighting Functions.

Criterion	Equation	Type	$w_{fg}(r)$	$w_{ff}(r)$	$w_{gg}(r)$
Difference criterion d_{fg}	2.8 and 2.9	D_{fg}	1 if $r = 0$ 0 if $r \neq 0$	1	1
Pearson product moment correlation coefficient r_{fg}[a]	2.10 and 2.11	S_{fg}	1 if $r = 0$ 0 if $r \neq 0$	$= w_{fg}(r)$	$= w_{fg}(r)$
Overlap integral Lawton and Bartell	2.13	S_{fg}	1 if $r = 0$ 0 if $r \neq 0$	$= w_{fg}(r)$	$= w_{fg}(r)$
Fold criterion Karfunkel et al.	2.16	D_{fg}	$1/(1 + \alpha\|r\|^{\beta})$	1	1
Normalised fold criterion	2.20	D_{fg}	$1/(1 + \alpha\|r\|^{\beta})$	$= w_{fg}(r)$	$= w_{fg}(r)$
Newly proposed similarity criterion	2.21	S_{fg}	$1 - \|r\|/l$ if $\|r\| < l$ 0 if $\|r\| >= l$	$= w_{fg}(r)$	$= w_{fg}(r)$

[a] A mean centering of f(x) and g(x) should be applied first.

associated weights in the comparison of two corresponding points on the reference and target patterns. The conventional difference criterion, the Pearson product-moment correlation coefficient, and the overlap integral of Lawton and Bartell use a delta function for $w_{fg}(r)$ and do not incorporate any contributions from the neighbourhood of a point. The fold criterion of Karfunkel et al. includes a weighting function that can be tuned by two parameters α and β. These two parameters define the shape and width of the weighting function. The differences in discriminating power between the conventional difference criterion, the overlap integral of Lawton and Bartell, the fold criterion of Karfunkel et al., and a newly proposed similarity criterion, which is obtained by using a simple triangle weighting function in the generalised expression for similarity, was investigated in the classification of 20 crystal structures of complexes of cephalosporin antibiotics on the basis of their calculated powder diffraction patterns. From single-crystal X-ray diffraction analyses of these compounds (Kemperman, De Gelder, Dommerholt, Raemakers-Franken, Klunder, and Zwanenburg [9,10]) it was found that among these 20 complexes six different crystal forms are found. Ten compounds crystallise in form A, four compounds in type B, two compounds in form N and the remaining four compounds in types C, D, E, and F. In Table 2.2 the crystal data are shown for these 20 complexes. From this table it can be seen that small differences are present in the cell parameters of the compounds belonging to the same crystal form. These small differences give rise to large peak shifts in the corresponding powder diffraction patterns. In Figure 2.3, the simulated powder diffractions

patterns for the different crystal forms are shown. For forms C, D, E, and F, only one member can be shown. For forms A, B, and N, two representative members are shown, clearly illustrating the large peak shifts resulting from the small differences in unit cell parameters. The question is now whether these 20 complexes can be classified on the basis of their powder patterns using a dedicated similarity criterion. The different crystal forms should be recognised as dissimilar, however, the complexes belonging to the same crystal form should ideally be recognised as similar compounds, and should somehow be clustered together.

Similarities or dissimilarities [which can be interconverted, see (2.21) and (2.22)] were calculated with four criteria (the difference criterion, the overlap integral of Lawton and Bartell, the fold criterion, and the newly proposed similarity criterion) for each pair of patterns. For the parameters α and β corresponding to the fold criterion, the optimised values of the authors [4] were used (10^{-7} and 4, respectively). For the newly proposed similarity criterion the following simple triangle weighting function was used:

$$w_{fg}(r) = 1 - \frac{\mid r \mid}{l} \qquad \text{if } |r| < l$$
$$w_{fg}(r) = 0 \qquad \text{if } |r| \geqq l$$

The parameter l defines the width (degrees 2θ) of the neighbourhood taken into account. This function extracts information from the central part of the crosscorrelation function with a weight that decreases proportionally to the distance from the origin ($r = 0$). After a number of experiments it was found that values of l between 0.6 and 3.0 lead to stable and comparable results for the powder diffractions patterns of the 20 complexes. A value of 0.6 was chosen for further calculations. In Figure 2.4, the weighting function of Karfunkel et al. and triangle weighting function for $l = 0.6$ and $l = 3.0$ are shown for comparison. The effect of the triangle weighting function on the crosscorrelation function, that was also shown in Figure 2.2, is illustrated in Figure 2.5.

The similarity calculations lead to four (dis)similarity matrices [similarities are eventually converted to dissimilarities using expression (2.21)] that are shown in Tables 2.3, 2.4, 2.5 and 2.6. Note that only the similarity matrices obtained with the overlap integral of Lawton and Bartell and the newly proposed similarity criterion, using the triangle weighting function, contain values on an absolute scale (values between 0 and 1).

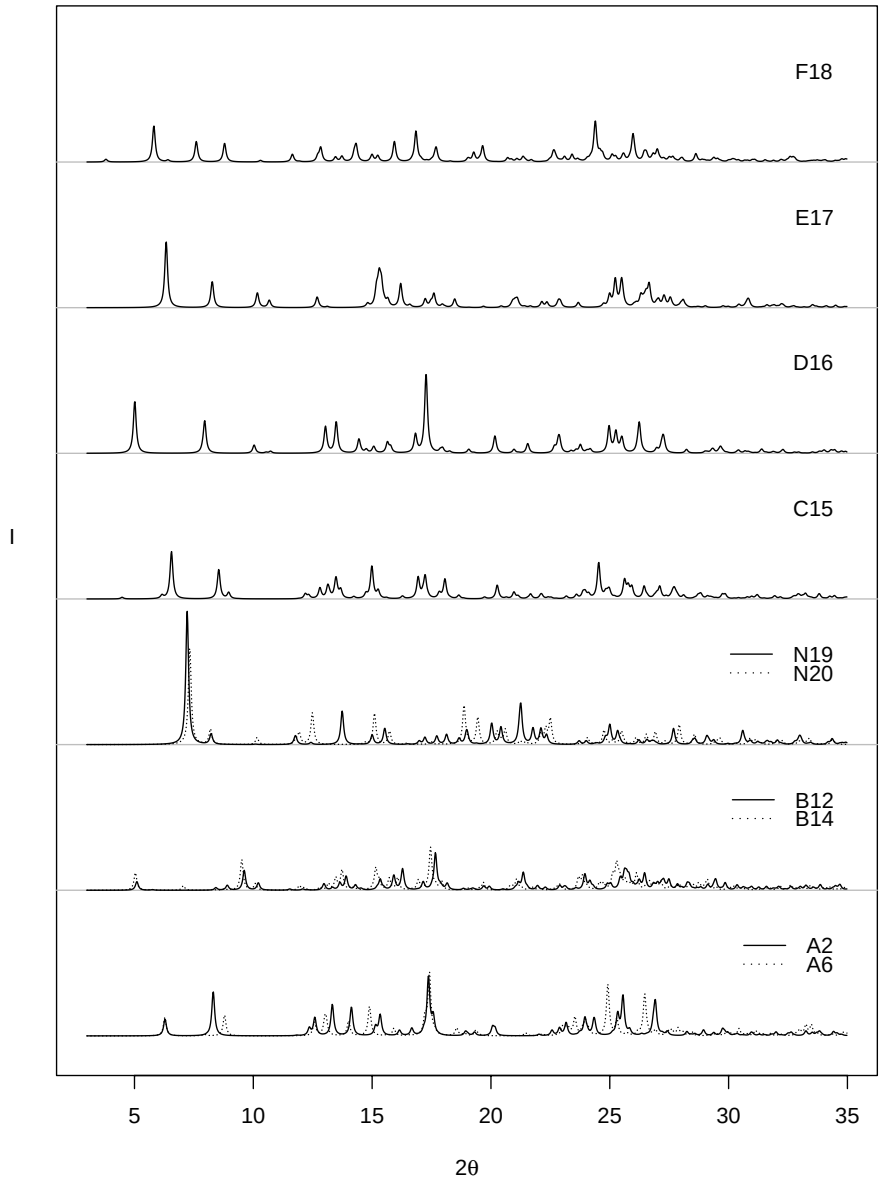

Figure 2.3: Simulated powder diffraction patterns of the various crystal forms found for complexes of cephalosporin antibiotics.

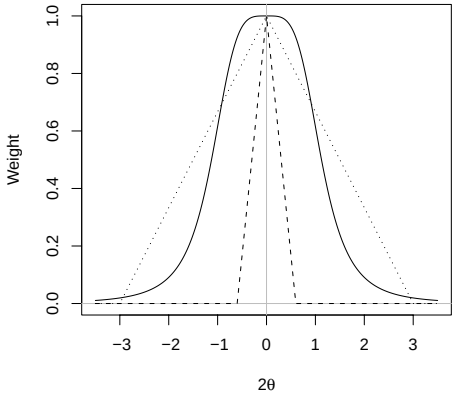

Figure 2.4: The weighting function of Karfunkel et al. (solid line), the triangle weighting function with $l = 0.6$ (dashed line) and $l = 3.0$ (dotted line).

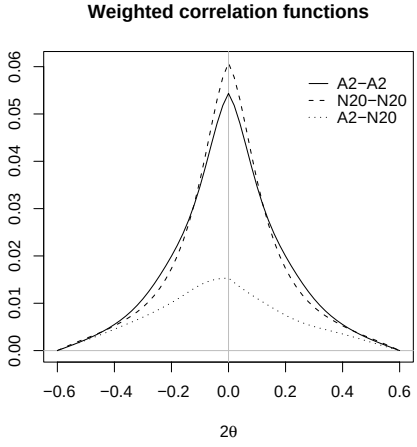

Figure 2.5: Weighted auto- and crosscorrelation functions (triangle weighting function with $l = 0.6$) for A2 and N20.

Table 2.2: Crystal Data of the Various Cephalosporin Complexes.

Compound name	a	b	c	α	β	γ	Space Group	Type
Cefradine/alpha-naphthol	23.47079	7.12154	14.93036	90.0	108.26834	90.0	C2	A1
Cefradme/beta-naphthol	23.42119	6.97147	15.00473	90.0	110.40521	90.0	C2	A2
Cefradine/naphthalene	23.45837	7.11788	14.89216	90.0	108.57098	90.0	C2	A3
Cefradine/quinoline	23.41265	7.10910	14.80598	90.0	108.14590	90.0	C2	A4
Cefradine/2-hydroxy acetonaphthone	23.38549	7.19649	14.75882	90.0	108.57974	90.0	C2	A5
Cefradine/bipyridine	23.02269	7.14670	14.55443	90.0	104.64385	90.0	C2	A6
Cephalexine/beta-naphthol	23.39759	7.06229	14.91757	90.0	109.79842	90.0	C2	A7
Cephalexine/alpha-naphthol	23.43432	7.10804	14.87538	90.0	108.19072	90.0	C2	A8
Cefaclor/alpha-naphthol	23.48840	7.07855	14.84551	90.0	108.94678	90.0	C2	A9
Cefaclor/beta-naphthol	23.44692	7.02619	14.84134	90.0	110.55009	90.0	C2	A10
Cephadroxil/beta-naphthol	7.11174	21.71704	30.95857	90.0	90.0	90.0	$P2_12_12_1$	B11
Cephadroxil/4-hydroxy-benzoezuur	6.99921	20.99127	30.69011	90.0	90.0	90.0	$P2_12_12_1$	B12
Cephadroxil/2,6-dihydroxy-naphthalene	7.10786	21.86340	32.30589	90.0	90.0	90.0	$P2_12_12_1$	B13
Cephadroxil/2,7-dihydroxy-naphthalene	7.09018	21.27323	31.00436	90.0	90.0	90.0	$P2_12_12_1$	B14
Ceftadine/4-hydroxy-benzoezuur	14.91661	7.38199	20.50296	90.0	105.77318	90.0	$P2_1$	C15
Cefradine/2-phenylphenol	23.56421	7.13203	18.68928	90.0	109.37986	90.0	C2	D16
Cefradine/hydroquinone	7.07185	10.70306	14.23422	87.15449	78.99942	89.74252	P1	E17
Cefradine/4-methylacetophenone	15.40382	7.29832	23.57355	90.0	99.35406	90.0	$P2_1$	F18
Cefradine/DMF	10.87473	9.51140	12.39035	90.0	98.70461	90.0	$P2_1$	N19
Cefradine/methyl 3-hydroxybenzoate	10.90731	9.40654	12.19924	90.0	98.53256	90.0	$P2_1$	N20

Table 2.3: Pattern Dissimilarity Values Calculated with the Difference Criterion.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B11	B12	B13	B14	C15	D16	E17	F18	N19	N20
A1	0.00	0.21	0.02	0.06	0.11	0.14	0.13	0.02	0.11	0.22	0.17	0.20	0.17	0.20	0.17	0.16	0.23	0.22	0.39	0.31
A2		0.00	0.18	0.16	0.21	0.15	0.11	0.19	0.14	0.07	0.12	0.14	0.17	0.12	0.17	0.20	0.16	0.17	0.37	0.27
A3			0.00	0.03	0.11	0.12	0.10	0.02	0.09	0.20	0.15	0.18	0.15	0.17	0.16	0.13	0.20	0.20	0.37	0.30
A4				0.00	0.13	0.11	0.09	0.03	0.08	0.19	0.14	0.18	0.17	0.17	0.17	0.15	0.17	0.20	0.38	0.30
A5					0.00	0.19	0.18	0.12	0.17	0.24	0.19	0.19	0.17	0.19	0.16	0.19	0.22	0.20	0.39	0.31
A6						0.00	0.15	0.13	0.15	0.16	0.09	0.14	0.16	0.13	0.16	0.18	0.19	0.18	0.36	0.29
A7							0.00	0.09	0.04	0.08	0.14	0.19	0.17	0.14	0.19	0.17	0.18	0.21	0.40	0.31
A8								0.00	0.07	0.20	0.15	0.20	0.17	0.18	0.17	0.15	0.21	0.22	0.39	0.31
A9									0.00	0.11	0.14	0.19	0.17	0.15	0.19	0.18	0.20	0.21	0.39	0.31
A10										0.00	0.15	0.18	0.19	0.14	0.20	0.23	0.21	0.21	0.40	0.32
B11											0.00	0.10	0.09	0.04	0.13	0.14	0.16	0.12	0.31	0.24
B12												0.00	0.10	0.08	0.12	0.18	0.14	0.10	0.28	0.23
B13													0.00	0.09	0.11	0.16	0.16	0.12	0.30	0.24
B14														0.00	0.13	0.15	0.14	0.12	0.29	0.23
C15															0.00	0.18	0.18	0.12	0.31	0.25
D16																0.00	0.24	0.21	0.38	0.33
E17																	0.00	0.18	0.35	0.27
F18																		0.00	0.32	0.23
N19																			0.00	0.22
N20																				0.00

Table 2.4: Pattern Similarity Values Calculated with the Overlap Integral of Lawton and Bartell.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B11	B12	B13	B14	C15	D16	E17	F18	N19	N20
A1	1.00	0.47	0.96	0.85	0.73	0.64	0.69	0.94	0.73	0.47	0.51	0.40	0.53	0.41	0.53	0.60	0.41	0.34	0.20	0.25
A2		1.00	0.51	0.56	0.45	0.56	0.72	0.52	0.63	0.82	0.59	0.50	0.44	0.59	0.44	0.47	0.53	0.40	0.18	0.27
A3			1.00	0.92	0.73	0.68	0.75	0.96	0.78	0.50	0.55	0.41	0.53	0.46	0.52	0.66	0.45	0.36	0.21	0.25
A4				1.00	0.68	0.70	0.77	0.93	0.80	0.53	0.57	0.40	0.49	0.48	0.50	0.61	0.53	0.35	0.19	0.23
A5					1.00	0.50	0.56	0.70	0.58	0.42	0.44	0.38	0.50	0.42	0.54	0.54	0.40	0.40	0.18	0.23
A6						1.00	0.62	0.65	0.61	0.57	0.69	0.47	0.44	0.55	0.47	0.52	0.43	0.38	0.20	0.23
A7							1.00	0.78	0.90	0.79	0.59	0.38	0.49	0.59	0.44	0.57	0.52	0.35	0.18	0.24
A8								1.00	0.83	0.52	0.56	0.39	0.52	0.46	0.52	0.62	0.44	0.34	0.19	0.25
A9									1.00	0.73	0.59	0.39	0.49	0.55	0.43	0.56	0.44	0.35	0.18	0.23
A10										1.00	0.58	0.44	0.43	0.60	0.43	0.43	0.44	0.36	0.17	0.21
B11											1.00	0.57	0.62	0.84	0.47	0.58	0.45	0.48	0.25	0.28
B12												1.00	0.55	0.63	0.48	0.41	0.49	0.52	0.27	0.25
B13													1.00	0.60	0.54	0.51	0.43	0.51	0.25	0.26
B14														1.00	0.49	0.53	0.52	0.47	0.28	0.28
C15															1.00	0.45	0.40	0.50	0.24	0.25
D16																1.00	0.35	0.33	0.19	0.18
E17																	1.00	0.35	0.21	0.27
F18																		1.00	0.20	0.28
N19																			1.00	0.54
N20																				1.00

Table 2.5: Pattern Dissimilarity Values Calculated with Fold Criterion.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B11	B12	B13	B14	C15	D16	E17	F18	N19	N20
A1	0.00	-0.12	0.07	0.37	-0.01	0.51	-0.03	0.00	0.35	0.25	0.93	1.73	0.97	1.00	0.51	0.79	2.99	0.69	6.46	4.85
A2		0.00	-0.06	0.17	0.03	0.22	0.05	-0.09	0.36	0.42	0.65	1.34	0.78	0.80	0.33	0.63	2.55	0.59	5.91	4.43
A3			0.00	0.19	0.04	0.33	0.10	0.08	0.49	0.42	0.75	1.47	0.86	0.86	0.43	0.71	2.83	0.53	6.18	4.72
A4				0.00	0.25	0.39	0.41	0.36	0.84	0.84	0.87	1.36	1.00	0.92	0.28	0.80	2.21	0.33	5.51	4.35
A5					0.00	0.49	-0.03	-0.02	0.16	0.17	1.07	1.97	1.23	1.12	0.62	0.95	3.10	0.80	6.37	4.92
A6						0.00	0.61	0.48	0.72	0.73	0.63	1.06	1.05	0.70	0.63	1.18	2.84	0.66	7.08	5.40
A7							0.00	0.00	0.43	0.34	1.06	1.98	1.09	1.20	0.58	0.79	3.35	0.93	6.49	5.03
A8								0.00	0.34	0.26	0.92	1.67	0.96	0.97	0.48	0.80	3.01	0.67	6.38	4.77
A9									0.00	0.01	1.43	1.93	1.58	1.31	0.98	1.48	3.72	1.05	7.07	5.27
A10										0.00	1.51	2.19	1.71	1.48	1.06	1.37	4.00	1.37	7.25	5.68
B11											0.00	0.39	0.30	0.11	0.76	0.69	2.22	0.58	6.68	5.11
B12												0.00	0.56	0.33	1.29	1.56	1.88	0.98	6.26	4.70
B13													0.00	0.28	1.02	0.86	1.85	0.80	6.26	5.05
B14														0.00	0.73	1.02	1.89	0.58	6.56	4.99
C15															0.00	1.16	1.59	0.35	4.11	3.33
D16																0.00	2.90	1.00	6.23	5.22
E17																	0.00	1.29	5.25	4.07
F18																		0.00	4.60	3.19
N19																			0.00	1.18
N20																				0.00

Table 2.6: Pattern Similarity Values Calculated with Generalised Similarity Criterion.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B11	B12	B13	B14	C15	D16	E17	F18	N19	N20
A1	1.00	0.80	0.99	0.96	0.93	0.88	0.93	0.99	0.92	0.82	0.78	0.60	0.78	0.72	0.75	0.80	0.63	0.66	0.34	0.41
A2		1.00	0.82	0.82	0.78	0.78	0.93	0.83	0.87	0.95	0.80	0.75	0.75	0.80	0.76	0.77	0.68	0.69	0.34	0.41
A3			1.00	0.98	0.92	0.90	0.94	0.99	0.93	0.83	0.81	0.63	0.79	0.75	0.76	0.82	0.66	0.67	0.35	0.41
A4				1.00	0.90	0.90	0.93	0.97	0.91	0.81	0.79	0.63	0.76	0.74	0.76	0.78	0.71	0.67	0.34	0.39
A5					1.00	0.82	0.86	0.91	0.85	0.79	0.72	0.61	0.75	0.69	0.81	0.75	0.58	0.73	0.31	0.40
A6						1.00	0.85	0.89	0.85	0.79	0.85	0.69	0.75	0.77	0.74	0.75	0.63	0.68	0.33	0.38
A7							1.00	0.95	0.96	0.94	0.81	0.66	0.77	0.78	0.75	0.81	0.66	0.66	0.35	0.40
A8								1.00	0.95	0.85	0.80	0.61	0.77	0.74	0.75	0.80	0.65	0.65	0.34	0.41
A9									1.00	0.93	0.79	0.66	0.75	0.77	0.71	0.78	0.63	0.64	0.34	0.39
A10										1.00	0.78	0.71	0.73	0.78	0.73	0.77	0.60	0.66	0.32	0.37
B11											1.00	0.84	0.87	0.96	0.73	0.83	0.66	0.69	0.41	0.42
B12												1.00	0.78	0.89	0.70	0.67	0.67	0.74	0.42	0.41
B13													1.00	0.87	0.77	0.83	0.66	0.75	0.42	0.44
B14														1.00	0.73	0.78	0.70	0.70	0.44	0.43
C15															1.00	0.67	0.70	0.75	0.43	0.44
D16																1.00	0.55	0.63	0.36	0.35
E17																	1.00	0.60	0.37	0.45
F18																		1.00	0.42	0.53
N19																			1.00	0.80
N20																				1.00

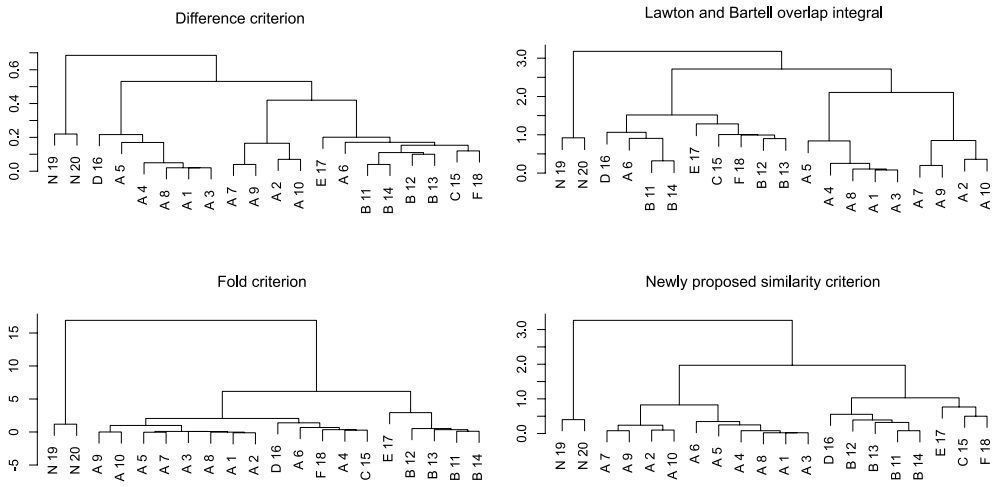

Figure 2.6: (a) Clustering dendrogram found for the difference criterion. (b) Clustering dendrogram found for the Lawton and Bartell overlap integral. (c) Clustering dendrogram found for the Fold criterion. (d) Clustering dendrogram found for the newly proposed similarity criterion.

On the dissimilarity matrices, a clustering algorithm was applied to group the patterns that are considered to be similar on the basis of the data in the matrices. A hierarchical agglomerative clustering method was used for this purpose. Initially, each object is viewed as a separate cluster; in each subsequent step, similar objects are joined according to a distance criterion, and the distances of the newly formed cluster to the other clusters or objects are recalculated. This process continues until all objects are joined into one cluster. The criterion that is used is known as "Ward's method," [11] where elements or clusters are joined in such a way that the sum of heterogeneities of all clusters (defined as the summed squared distance of each member of a cluster to the centroid of that cluster) increases as little as possible. The method performs best in cases where the clusters are approximately spherical in shape and of equal size, and is widely applied. This clustering procedure yields the dendrograms as depicted in Figure 2.6. Clearly, the results are quite different for the four (dis)similarity criteria. The newly proposed similarity criterion, using the triangle weighting function, leads to the most homogeneous classification and is able to separate the crystal structures into four groups consisting of 1: only N, 2: only A, 3: four B's and one D, and 4: E, C, and F.

To illustrate the stability of the generalised similarity criterion using the

simple triangle weighting function, similarity matrices were calculated for increasing values of the parameter l. The same clustering procedure was applied on the corresponding similarity matrices, and the resulting dendrograms are shown in Figure 2.7. For small values of l (smaller than 0.6), very inhomogeneous classification dendrograms are found. For very large values of l (larger than 3.0) also inhomogeneous classification dendrograms are found. However, in the range 0.6 to 3.0, similar dendrograms of comparable homogeniety are found.

2.7 Discussion and Conclusion

The generalised expression for the similarity of powder diffraction patterns shows that the criteria described in literature all refer to the correlation function. It also shows that the differences between the criteria can be explained by different choices of weighting functions for the auto- and crosscorrelation terms. The nature of the weighting functions used determines whether a pointwise or neighbourhood approach is applied, and whether the resulting measure for similarity or dissimilarity is on an absolute scale or not.

The importance of a neighbourhood approach has been demonstrated for the classification of crystal structures on the basis of their calculated powder diffraction patterns. Including the neighbourhoods leads to a significant improvement of discriminating power compared to pointwise approaches, and makes it possible to recognise closely related structures. The way in which the neighbourhood is taken into account depends on the shape of the weighting function used. A simple triangle function leads to very useful results and performs even better than the more complex weighting function used by Karfunkel et al. It should, however, be emphasised that we did not try to adapt the parameters defining the fold weighting function to optimise the classification of our set of structures. We just took the function and parameters the authors optimised for their specific problem.

In our tests we used calculated powder diffraction patterns as the source of structural information. In practice, one should like to apply the classification procedure on experimentally obtained powder patterns. In that case, there might be factors like zero-point shifts, preferred orientation, peak broadening, nonzero background and experimental noise that may influence the outcome of the classification procedure. However, the classification of experimentally obtained patterns and its associated problems will be the subject of our further

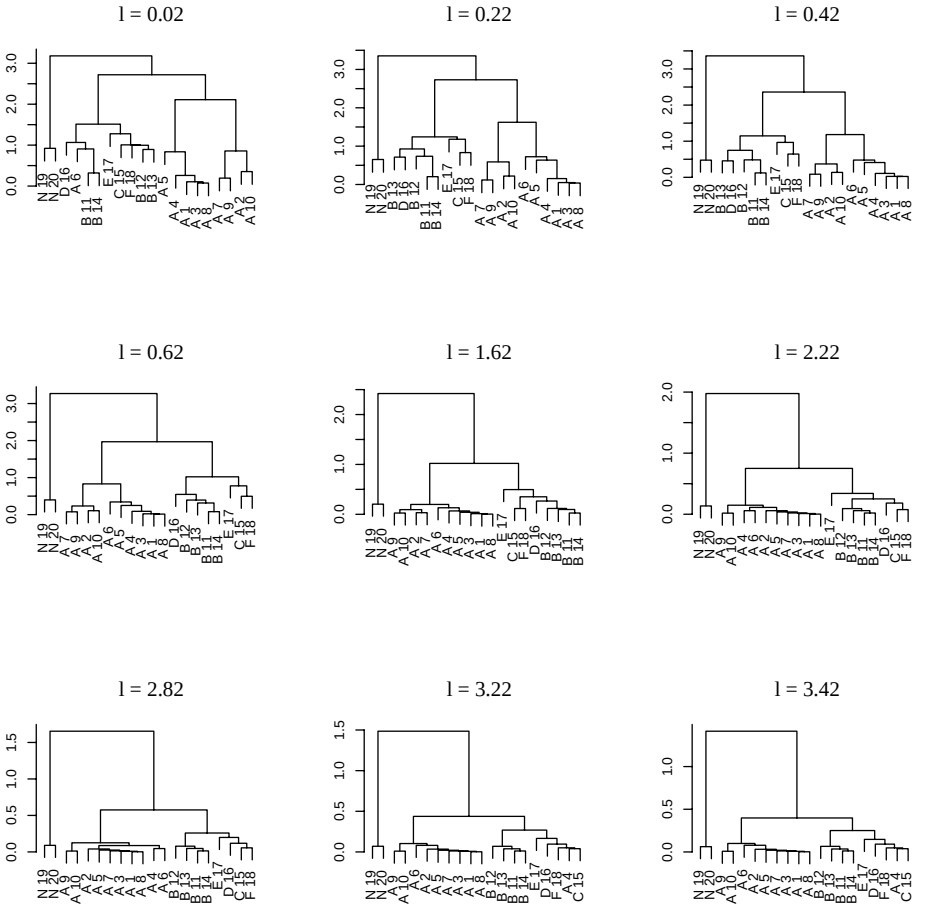

Figure 2.7: Clustering dendrograms found for increasing values of l (only unique dendrograms in the range $l = 0.02$ to 3.42 are shown).

research, and will not be discussed any further here.

In the introduction it was mentioned that the generalised expression might also be used for database searching and optimisation problems. For database searching it is of crucial importance that a measure for similarity or dissimilarity is defined on an absolute scale. It is necessary to know the numerical range of the measure to define limits for acceptance in the selection of subsets of patterns from a database. The newly proposed criterion for similarity can be used for the purpose of database searching, because its value always ranges from 0 to 1. The width of the triangle function can be adapted to change the search from a strict to a more generous one. It is up to the user to optimise the settings for his particular case.

For optimisation problems the aspect of normalisation, to obtain a measure on an absolute scale, is less important. The width of the neighbourhood, however, may determine the overall success of an optimisation procedure (see Hageman et al. [6] in which parameters defining a theoretical pattern are optimised with respect to an experimental spectrum. The effect of the inclusion of the neighbourhood is a more realistic assessment of the error, and this may guide the process to the global optimum. A possible application of the newly proposed similarity criterion could be the determination of cell and/or positional parameters of crystal structures from powder diffraction data only. In principle, the criterion allows for a gradual improvement of both peak positions and peak intensities as was shown by Hageman et al.

Although this article has focussed on powder diffraction patterns, which are one-dimensional patterns, it should be emphasised that the generalised expression for similarity is applicable to other types of spectra and is not limited to one-dimensional patterns only. If the variables x and r are replaced by vectors the expression can directly be used in multidimensional space. Then it can also be applied, for example, to express the similarity between 2D, 3D, or 4D-NMR spectra.

2.8 Appendix

2.8.1 The integral of the autocorrelation function

$$c_{ff}(r) = \int f(x)f(x+r)dx$$

$$\begin{aligned}\int c_{ff}(r)dr &= \int\int f(x)f(x+r)dxdr \\ &= \int\int f(x)f(r')dxdr' \qquad (r = r' - x) \\ &= \int f(x)dx \int f(r')dr' = (\int f(x)dx)^2\end{aligned}$$

2.8.2 d_{fg} Written as auto- and crosscorrelation function

$$\begin{aligned}d_{fg} &= \int (f(x) - g(x))^2 dx \\ &= \int f^2(x)dx + \int g^2(x)dx - 2\int f(x)g(x)dx\end{aligned}$$

A pointwise comparison of patterns $f(x)$ and $g(x)$ (neglecting the neighbourhoods of points in the patterns) would, speaking in terms of the crosscorrelation function $c_{fg}(r)$, be identical to calculating the crosscorrelation function for patterns $f(x)$ and $g(x)$ at $r = 0$:

$$c_{fg}(0) = \int f(x)g(x)dx$$

Therefore, d_{fg} can be written as:

$$d_{fg} = c_{ff}(0) + c_{gg}(0) - 2c_{fg}(0)$$

2.8.3 r_{fg} Written as auto- and crosscorrelation function

$$r_{fg} = \frac{\int (f(x) - < f(x) >)(g(x) - < g(x) >)dx}{(\int (f(x) - < f(x) >)^2 dx \int (g(x) - < g(x) >)^2 dx)^{\frac{1}{2}}}$$

Defining the new patterns $f'(x) = f(x) - < f(x) >$ and $g'(x) = g(x) - < g(x) >$ and considering these new patterns $f'(x)$ and $g'(x)$ as vectors **f'** and **g'** in n-dimensional space (n being the number of points x_i at which values for $f'(x_i)$ and $g'(x_i)$ are measured or calculated) this expression reduces to the cosine of the angle $\alpha_{f'g'}$ between vectors **f'** and **g'**:

$$
\begin{aligned}
r_{fg} &= \cos(\alpha_{f'g'}) \\
&= \int f'(x)g'(x)dx/(\int f'^2(x)dx \int g'^2(x)dx)^{1/2} \\
&= \mathbf{f'} \cdot \mathbf{g'}/|\mathbf{f'}||\mathbf{g'}| \\
c_{f'f'}(0) &= \int f'^2(x)dx \\
c_{g'g'}(0) &= \int g'^2(x)dx \\
c_{f'g'}(0) &= \int f'(x)g'(x)dx \\
r_{fg} &= c_{f'g'}(0)/(c_{f'f'}(0)c_{g'g'}(0))^{1/2}
\end{aligned}
\tag{2.23}
$$

2.8.4 The overlap integral of Lawton and Bartell

The overlap integral for peaks i from pattern α and peaks j from pattern β, defined by Lawton and Bartell, is expressed as follows:

$$
\begin{aligned}
S_{\alpha\beta} &= \sum\sum(I_\alpha(i)/(\sigma_\alpha(i)a_\alpha^{1/2}))(I_\beta(j)/(\sigma_\beta(j)a_\beta^{1/2})) \\
&\times \quad exp[-(d_\alpha(i) - d_\beta(j))^2/4(w_\alpha(i) \cdot w_\beta(j))]
\end{aligned}
$$

where

$$
\begin{aligned}
a_\alpha &= \sum\sum(I_\alpha(i)/(\sigma_\alpha(i))(I_\alpha(j)/\sigma_\alpha(j)) \\
&\times \quad exp[-(d_\alpha(i) - d_\alpha(j))^2/4(w_\alpha(i) \cdot w_\alpha(j))] \\
a_\beta &= \sum\sum(I_\beta(i)/(\sigma_\beta(i))(I_\beta(j)/\sigma_\beta(j)) \\
&\times \quad exp[-(d_\beta(i) - d_\beta(j))^2/4(w_\beta(i) \cdot w_\beta(j))]
\end{aligned}
\tag{2.24}
$$

I_α and I_β correspond to the diffracted relative intensities in patterns α and β, σ_α, and α_β are parameters representing the characteristic variations in these intensities, d_α and d_β correspond to the interplanar spacings in patterns α and β and w_α and w_β are window parameters describing the windows of acceptance.

2.8.5 $d_{fg}(r)$ Written as auto- and crosscorrelation function

$$\begin{aligned} d_{fg}(r) &= \int (f(x) - g(x+r))^2 dx \\ &= \int f^2(x)dx + \int g^2(x+r)dx - 2\int f(x)g(x+r)dx \\ &= c_{ff}(0) + c_{gg}(0) - 2c_{fg}(r) \\ \int d_{fg}(r)dr &= c_{ff}(0) + c_{gg}(0) - 2\int c_{fg}(r)dr \end{aligned}$$

2.8.6 The fold written as auto- and crosscorrelation integrals

$$fold = \mathbf{d}^T\mathbf{F}\mathbf{d}$$

where an element of the vector $\mathbf{d}$, $d(x) = f(x) - g(x)$. The elements of the matrix $\mathbf{F}$ are initially set to 1: $F_{ij} = 1$.

$$\begin{aligned} \mathbf{d}^T\mathbf{F}\mathbf{d} &= \int\int (f(x) - g(x))(f(x+r) - g(x+r))dxdr \\ &= \int\int f(x)f(x+r)dxdr + \int\int g(x)g(x+r)dxdr \\ &\quad - \int\int f(x)g(x+r)dxdr - \int\int f(x+r)g(x)dxdr \\ &= \int c_{ff}(r)dr + \int c_{gg}(r)dr - 2\int c_{fg}(r)dr \end{aligned}$$

where $r = i - j$.

Introducing the original matrix $\mathbf{F}$ into the expression for the fold leads to:

$$\begin{aligned} F_{ij} &= 1/(1 + \alpha|i-j|^\beta) \\ \mathbf{d}^T\mathbf{F}\mathbf{d} &= \int\int (f(x) - g(x))w(r) \times \quad (f(x+r) - g(x+r))dxdr \\ &= \int\int w(r)f(x)f(x+r)dxdr + \int\int w(r)g(x)g(x+r)dxdr \\ &\quad - \int\int w(r)f(x)g(x+r)dxdr - \int\int w(r)f(x+r)g(x)dxdr \\ &= \int w(r)c_{ff}(r)dr + \int w(r)c_{gg}(r)dr - 2\int w(r)c_{fg}(r)dr \end{aligned}$$

where $w(r) = 1/(1 + \alpha|r|^{\beta})(r = i - j)$.

References

[1] R.A Young. *The Rietveld Method; International Union of Crystallography*. Oxford University Press, Oxford, 1993.

[2] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, and J. Smeyers-Verbeke. *Handbook of Chemometrics and Qualimetrics: Part B*. Elsevier, Amsterdam, 1998.

[3] S.L. Lawton and L.S. Bartell. Application of the overlap integral in X-ray diffraction powder pattern recognition. *Powder Diffraction*, 9(2):124 – 135, 1994.

[4] H.R. Karfunkel, B. Rohde, F.J.J. Leusen, R.J. Gdanitz, and G. Rihs. Continuous similarity measure between nonoverlapping X-ray powder diagrams of different crystal modifications. *J. Comp. Chem.*, 14(10):1125–1135, 1993.

[5] D.S. Stephenson and G. Binsch. Automated analysis of high-resolution NMR spectra. II illustrative applications of the computer program DAVINS. *J. Magn. Reson.*, 37:409–430, 1980.

[6] J.A. Hageman, R. Wehrens, R. de Gelder, W.L. Meerts, and L.M.C. Buydens. Direct determination of rotational constants from rovibronic spectra with genetic algorithms. *J. Phys. Chem.*, 113(18):7955–7962, 2000.

[7] K.D.M. Harris, R.L. Johnston, and B.M. Kariuki. The genetic algorithm: Foundations and applications in structure solution from powder diffraction data. *Acta Cryst.*, A54:632–645, 1998.

[8] J. Dods, D. Gruner, and P. Brumer. A genetic algorithm approach to fitting polyatomic spectra via geometry shifts. *Chem. Phys. Lett.*, 261:612–619, 1996.

[9] G.J. Kemperman, R. de Gelder, F.J. Dommerholt, P.C. Raemakers-Franken, A.J.H. Klunder, and B. Zwanenburg. *Chemistry: a European Journal*, 5:2163, 1999.

[10] G.J. Kemperman, R. de Gelder, F.J. Dommerholt, P.C. Raemakers-Franken, A.J.H. Klunder, and B. Zwanenburg. *Perkin Trans II*, 7:1425, 2000.

[11] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data, An Introduction to Cluster Analysis*. Wiley, New York, 1989.

Chapter 3

Direct Determination of Molecular Constants from Rovibronic Spectra with Genetic Algorithms

It is shown that a new procedure, based on Genetic Algorithms (GA's), can be used for direct determination of molecular constants, in particular rotational constants, from rovibronic spectra. This new approach only requires an estimate of the acceptable range of the parameters. The power of the method is demonstrated on the rotationally resolved fluorescence spectra of indole, indazole, benzimidazole and 4-aminobenzonitril. A rigid asymmetric rotor Hamiltonian is used to calculate the theoretical spectra. The GA matches the generated spectra with an experimental spectrum with the use of a new method for spectra comparison. This spectra comparison function is able to deal with frequency shifts which are caused by (small) changes in the rotational constants and it yields better results in comparison with traditional spectra comparison methods, like RMS. In addition, the robustness of the method is tested.

This chapter is publised as: J.A. Hageman, R. Wehrens, R. de Gelder, W. Leo Meerts and L.M.C. Buydens, Direct Determination of Molecular Constants from Rovibronic Spectra with Genetic Algorithms, J. Chem. Phys., 113(18), 7955-7962, 2000

3.1 Introduction

Rotational constants are an important tool in determining the spatial geometry of molecules. These constants give access to intra- and intermolecular bond lengths and their changes upon excitation. Rotational constants can be obtained from a large variety of methods, among others micro wave spectroscopy, IR Fourier transform, diode laser spectroscopy and high resolution laser induced fluorescence (LIF) excitation spectra. Especially the last 3 methods deal with a two state problem, either two vibrational or two electronic states. The resolution of such spectra is such that individual rotational transitions can be observed and the spectra contain large number of lines. Usually, to determine molecular rotational constants a spectrum is simulated using a model (for instance an asymmetric rotor Hamiltonian) which uses rotational constants estimated from other experiments or from ab initio calculations and the appropriate selection rules. The spectrum is calculated and compared with the experimental one. In what we will call the classical method, an initial assignment in terms of theoretical quantum numbers of transitions is made. In a least squares fit procedure the molecular parameters are adjusted to reproduce the assigned lines. The assignments are refined and the process is repeated until all lines in the spectrum are successfully reproduced [1]. The quality of the fit is, amongst other factors, dependent on the sophistication of the used model. The determination can be facilitated and speeded up by using reasonably accurate estimates of the molecular constants although this is not crucial.

Recently, attempts to automate the interpretation of rovibronic spectra have been undertaken. Automation becomes increasingly important when spectra become more difficult to interpret and/or prior-knowledge about the molecule is little or lacking. The group of Neusser et al. [2] used a procedure which directly fits the experimental data, without any preceding assignment of lines, with the help of the so called correlation automated rotational fitting algorithm. This algorithm still relies on accurate initial estimates of the rotational constants obtained from other experiments. Their experimental data were also analysed by assigning lines and using a least squares fit procedure. They concluded, by visual inspection, that the correlation method yielded better results [2]. Unfortunately, the method still has limited applicability.

The approach of fine-tuning the parameters of the Hamiltonian model so that the theoretical spectrum is in close agreement with the experimental one, can be seen as an optimisation problem. The process of determining

molecular constants can be automated with global optimisation methods like Simulated Annealing (SA) [3], Tabu Search (TS) [4] or Genetic Algorithms (GA's). In this paper it is shown that a GA with a specially developed fitness function is very successful in directly determining the molecular constants from LIF spectra. This is done without using any initial estimates of these constants, except their global limits. This new approach is demonstrated for 4 rotationally resolved (LIF) spectra from indole, indazole, imidazole and 4-aminobenzonitril (4-ABN). The spectra were measured by Berden et al. [5,6]. The essence of the analysis of Berden et al. was an assignment of quantum numbers of the initial and final states of the transitions in the spectrum. In a second step an overall fit of the intensities was carried out in which only the intensity parameters were determined. By carefully adapting the parameters Berden et al. succeeded in minimizing the difference between the experimental and simulated spectrum and obtained the complete set of molecular constants.

In the next section, a description is given of the parameters that appear in the Hamiltonian model, the use of GA's and the new method for comparing spectra. It will be shown in section 3.3 that a GA is very capable of determining the molecular parameters that reproduce the experimental spectra. In addition, the robustness of this GA-based method will be assessed by artificially deteriorating the quality of the data. It is shown that the method is quite robust and, therefore, widely applicable.

3.2 Theory

3.2.1 Model Representation

Given a set of molecular parameters, a theoretical rovibronic spectrum can be calculated using a rigid asymmetric rotor Hamiltonian. All experimental spectra analysed in this paper are fitted to this type of calculated spectra. It is assumed that, if a theoretical spectrum matches the experimental one, the parameters used to calculate the spectrum are very close to the true values. Since a discussion of the theory of the rigid asymmetric rotor Hamiltonian is beyond the scope of this paper, we will suffice to say that all rotational levels of the molecules under study can be calculated with this model [5,6]. The important details of the model are described briefly below. It contains 13 parameters, which are optimised by the GA. They can be divided in 5 groups.

1. Six rotational constants. Three parameters (A", B", C") describing the ground state and three parameters (ΔA, ΔB and ΔC) describing the

difference between the ground and excited state values, ΔA=(A' - A") etc. Here the double and single primes label the ground and excited states, respectively. These parameters are responsible for the positions of the transition frequencies.

2. A frequency shift parameter (ν). This parameter shifts the whole spectrum relative to an arbitrary zero point.

3. Three parameters that describe the relative intensities of the transitions (T_1, T_2, W). A three-parameter two-temperature model has been used [5]. By definition, T_2 must be higher that T_1. W is a weighting factor.

4. The direction of the electronic transition moment of the electronic excitation (θ) and a parameter (θ_T), which is the angle between the principal axes systems in the ground and in the excited states. θ_T is not optimised in this approach as it influences only a very small number of lines (< 10) and can only be determined by visual inspection of the appropriate lines. See for an example figure 5 in Ref. [5].

5. The line width ($\Delta\nu$) of lines in the spectrum. In Ref. [5] it is shown that the transitions in all 4 spectra are best described by a Lorentzian profile. However, this is not an essential limitation for the present discussion.

3.2.2 Genetic Algorithms

GA's are a special class of global optimisers, based on the theory of evolution. A GA is able to minimise (or maximise) a function G(x), where x represents a parameter vector, by searching the parameter space of x for the optimal solution. Several general steps can be distinguished that are identical for all GA's.

1. Initialisation: GA's do not operate on an individual solution for searching the parameter space but rather on a group of solutions (called population) at a time. A solution, called string or chromosome, consists of several parts, called genes. Each part is a parameter which needs optimisation. All initial solutions are set to random values. In the present examples each chromosome contains 12 genes which are the 12 parameters of the rigid Hamiltonian model.

2. Evaluation: All strings in the population are evaluated by an objective function. This results in a measure of quality of the string, expressed

in a single number. The evaluation function is usually tailor made for the specific GA application. It should be able to discriminate between 'good' and 'bad' strings in a given generation, to allow selection in a later phase.

3. Stop: A stop criterion is checked.

4. Selection: A percentage of the best strings in a population is selected and placed in the next generation.

5. Recombination: To form the new population, new solutions are created by combining two existing solutions (parents) to yield two different ones (children). This is called crossover.

6. Mutation: Genes on a string in the new population are selected randomly and modified. An example of a mutation operator is addition of a (small) random number. To prevent the search from random behaviour the probability of mutation is usually chosen to be quite low.

Several parameters, for instance the rate of crossover and mutation, regulate the performance of the GA and each specific problem has its own specific set of parameters for which the GA performs at its optimum. This so-called meta-optimisation of the GA parameters can be tedious and can be considered a disadvantage of GA in general [7]. In this paper it is shown that one set of GA parameters can successfully be used for estimating molecular constants of indole, indazole, benzimidazole and 4-ABN, so it is not necessary to repeat this meta-optimising for each new compound. The most important advantage of the GA approach is that it does not need prior knowledge of the molecular constants. All that is required, is an estimate of the accessible range for each parameter. The narrower this range is chosen the faster the optimisation will be. These ranges can be chosen, for instance, on physical grounds or be adapted from similar molecules known from literature.

Some literature is available about GA's. For an introduction to the subject Ref. [8] or for a more sophisticated level Ref. [9] are very well suited.

3.2.3 Evaluation or objective function

The parameters on each string are used in the rigid asymmetric rotor Hamiltonian model to calculate a theoretical spectrum. The similarity between the calculated spectrum and experimental spectrum has to be expressed in a

single number. Several methods are known to compare spectra. The more traditional methods perform a pointwise comparison between two spectra and express this as the sum of the squared differences [9, 10]. More sophisticated comparison methods include a comparison of the neighbourhood to deal with peak shifts [11].

Our initial attempts clearly demonstrated the inability of evaluation functions based on a sum of squared differences to differentiate reliable between spectra originating from nearly identical sets of parameters. Other approaches, based on peak picking and minimising the distance to neighbouring peaks in both spectra, failed as well. Moreover, since the relative position of peaks can change dramatically, one is never sure if the correct peak pairs are compared. With these types of evaluation functions, similar spectra with shifts in peak positions will not properly be recognised as similar. An improvement over the RMS-based evaluation function is the correlation coefficient Cr:

$$Cr = \frac{\sum_{x=0}^{x=k}(f(x) - < f >)(g(x) - < g >)}{\sqrt{\sum_{x=0}^{x=k}(f(x) - < f >)^2 \sum_{x=0}^{x=k}(g(x) - < g >)^2}} \tag{3.1}$$

Here $< f >$ and $< g >$ are the average intensity value of spectrum f and g respectively, $f(x)$ and $g(x)$ are the spectra f and g with length k. Further improvement is obtained if the average value is no longer subtracted from each point and (a possible) elevated baseline is removed. This ensures that all signals present in the spectrum are originating from peaks. This yields an evaluation function cos γ:

$$cos\gamma = \frac{f \cdot g}{\parallel f \parallel \parallel g \parallel} \tag{3.2}$$

Here $f \cdot g$ is the dot product of the experimental (f) and calculated (g) spectrum, $\parallel f \parallel$ and $\parallel g \parallel$ represent the length of spectrum f and g respectively. Cos γ ranges from $< -1 \mid 1 >$. For equal spectra cos γ=1.

Eq. 3.2 does not take into account small frequency shifts in the peak position. The evaluation function of Eq. 3.2 can be improved to deal with shifts if a crosscorrelation function is used:

$$C_{fg}(r) = \frac{\sum_{x=0}^{x=k} f(x) \cdot g(x+r)}{\| f \| \cdot \| g \|} \tag{3.3}$$

The crosscorrelation function compares two spectra shifted by r. In order to deal with 'end points' the sum should run from $-\infty$ to $+\infty$. Formally this can be realised by adding to the spectra points of zero intensity. In this way the normalisation by $\| f \|$ and $\| g \|$ is properly defined. Figure 3.1 shows several C_{fg} with r ranging from [-100,100]. The solid line is the autocorrelogram where both f and g are the original calculated spectrum of indole from Ref. [5]. The dashed and the dashdotted line are crosscorrelograms of the calculated spectra of indole with two calculated spectra in which A" and ΔA, respectively, are slightly changed by 1.0 MHz. The dotted line is the crosscorrelogram of the calculated spectrum of indole with the calculated spectrum of benzimidazole from Ref. [5]. It can be seen from figure 3.1 that no shift whatsoever of the benzimidazole spectrum will significantly increase C_{fg}, indicating no correlation at all between the spectra of indole and benzimidazole. Although the solid, dashdotted and dashed lines originate from spectra calculated with nearly identical parameters, the change in $C_{fg}(0)$ (which equals Eq. 3.2) is quite large. This implies that almost identical spectra may have quite different values for $C_{fg}(0)$. However, the area under the curve can be used as a convenient measure if a suitable weight function is used.

To penalise larger shifts, Eq. 3.3 is modified by introducing a weight function $w(r)$:

$$w(r) = 1 - \frac{\mid r \mid}{l} \tag{3.4}$$

The parameter l defines the width of the neighbourhood that is taken into account, typically in the order of 100 data points in the current work. Several weight functions were tested, including the sigmoidal function from Ref. [11]. Eventually the simple triangle function (Eq. 3.4) is used, because it depends on only one parameter. The sigmoidal function showed no improvement over Eq. 3.4.

The final overlap function is obtained by integrating Eq. 3.3 multiplied by the weight function and normalising between 0 and 1:

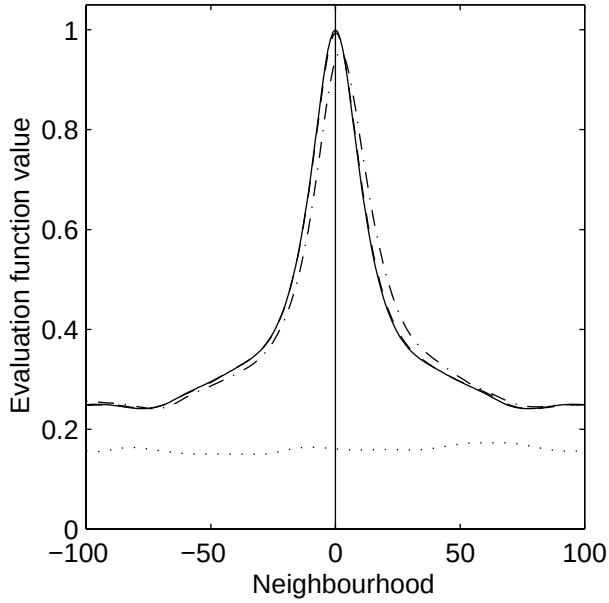

Figure 3.1: Correlogram of the calculated spectrum of Indole. Autocorrelogram: solid line, Crosscorrelogram: dashed line (A" increased by 1.0 MHz), dash dotted line (ΔA increased by 1.0 MHz) and dotted line (calculated spectrum of Benzimidazole).

$$C_{fg}^{ws} = \frac{\sum_{r=-l}^{r=l} C_{fg}(r)w(r)}{\sqrt{\sum_{r=-l}^{r=l} C_{ff}(r) * w(r) *} \sqrt{\sum_{r=-l}^{r=l} C_{gg}(r) * w(r)}} \tag{3.5}$$

For two identical spectra C_{fg}^{ws} is 1 and for two distinctly different spectra C_{fg}^{ws} is close to zero. The final evaluation function used in the GA calculations is defined as

$$F = 100 * (1 - C_{fg}^{ws}) \tag{3.6}$$

and its value is minimised.

Error landscapes of an RMS-based evaluation function and F are plotted in figure 3.2. In both plots A" and B" are varied over a grid covering the complete range, while the remaining parameters are held fixed. The effect

of Eq. 3.6 clearly shows a more smooth error landscape, which reduces the number of local minima.

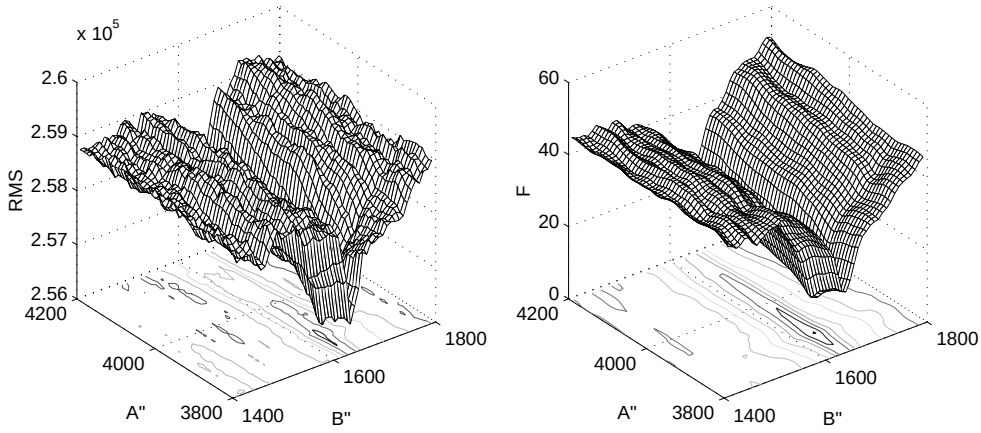

Figure 3.2: Difference in error landscape between a RMS-based evaluation function (left) and one based on Eq. 6 (right).

A more detailed discussion and comparisons with other methods for the assessment of similarity between 1-dimensional spectra can be found in the work of De Gelder et al [12].

3.3 Experimental

The spectra of indole, indazole, benzimidazole and 4-ABN are shown in figure 3.3. The spectra of indole and benzimidazole contain 65536 equidistant data points, the spectrum of indazole 61821 data points and the spectrum of 4-ABN contains 40972 data points. All 12 parameters were coded as 10-bit gray binary numbers. T_2 is coded on the string as α, with T_2=α*T_1 and $\alpha > 1$. The calculated spectra always contain the same number of data points as the corresponding experimental ones. The optimal settings of the GA were determined by trial and error and based on previous experience using the experimental spectrum of benzimidazole and are shown in Table 3.1.

The optimal size of the neighbourhood in Eq. 3.5 has been established from several experiments. The optimal value for l was 100 data points. A larger range also results in a correct solution but leads to longer run times. For a significantly smaller range no correct solution is obtained indicating that the inclusion of neighbourhood information is crucial. After establishing the

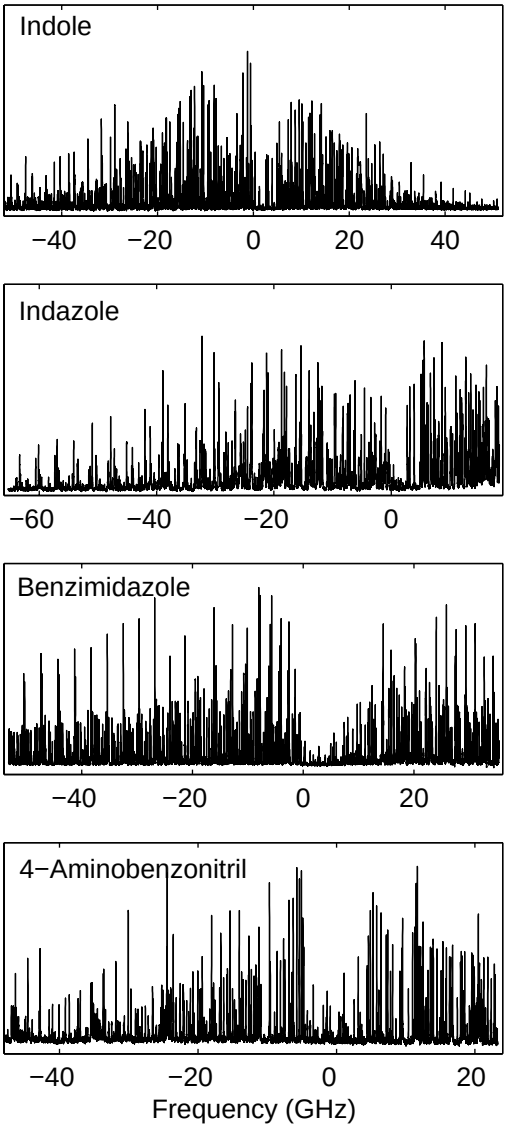

Figure 3.3: High resolution LIF spectra of Indole, Indazole, Benzimidazole and 4-ABN. In all cases the absolute frequency is set to 0.0 according to Ref. 5 and Ref. 6. The intensity on the vertical scale is in arbitrary units.

optimal settings, the experimental spectra of indole, indazole, benzimidazole and 4-ABN were fitted using boundary constraints as given in Table 3.2. The duration of a run has been set to 500 generations, long enough to converge to

Table 3.1: GA settings.

Setting	Value
maximum number of generations	500
population size	300
elitism	150
crossover type	two-point-crossover
crossover probability	0.85
mutation type	new random value within boundaries
mutation probability	0.05
selection type	probabilistic
fitness type	raw[a]

[a]Fitness value increases inversely proportional with evaluation value of a string.

a minimum. All runs were repeated 5 times with different random generator seeds to exclude lucky and/or unlucky runs.

The robustness of the GA method was investigated in a number of runs. We investigated the influence of (high) noise levels, increased line widths and the total number of points in a spectrum. Synthesised spectra of indole and benzimidazole were modified with different levels of normally distributed (white) noise, increased line widths and a combination of these two factors. Spectra with a reduced number of data points were also investigated. Figure 3.4 shows parts of the spectrum of indole with (a) a signal-to-noise level (S/N) of 10 (for the peak with the largest intensity), (b) a line width of 90 MHz and (c) a S/N of 10 combined with a line width of 90 MHz. The spectrum with a combination of large line widths and low S/N can be considered as very extreme.

All GA calculations were performed with the GA library PGAPack version 1.0 [13], which can run on parallel processors. PGAPack and the evaluation function are written in ANSI-C, the rigid asymmetric rotor Hamiltonian function was written in Fortran. All calculations were performed on a Sun-Ultra-Enterprise-10000 with 24 processors each running at 333 MHz. With 16 processors, the average runtime was about half an hour for 500 generations and 65536 data points. In practice this runtime can be reduced drastically, because often runs converge to their final solution long before the maximum number of generations is reached. If the boundaries are taken narrower runtimes can be further reduced because runs will converge even earlier. This will also lead to increased reproducibility and this decreases the need for more

Table 3.2: Boundary constraints for all 12 parameters used for Indole, Indazole and Benzimidazole and 4-ABN[a].

Parameter	Boundary constraints	
	Indole and Benzimidazole	4-ABN
A"	3800 - 4200	5000 - 6000
B"	1400 - 1800	800 - 1200
C"	800 - 1400	600 - 1000
T_1	1 - 6[b]	1 - 6
T_2^d	1.5 - 5	1.5 - 5
W	0 - 1	0 - 1
θ	0° - 90°	90°, fixed[e]
ν	-300 - 300[c]	-5000 - 5000
ΔA	-200 - 0	-400 - 400
ΔB	-50 - 0	-100 - 100
ΔC	-50 - 0	-100 - 100
$\Delta\nu$	10 - 40	10 - 90

[a]Rotational constants in the ground state are indicated by A", B" and C". Rotational constants in the excited state are given by their deviations from the ground state (ΔA, ΔB and ΔC). $\Delta\nu$ is line width of the Lorentzian peaks. Rotational constants, ν and $\Delta\nu$ are in MHz, T_1 and T_2 in K.

[b]Range is 2 - 8 for the spectrum taken from Benzimidazole.

[c]The frequency of the origin (ν) is set to zero. The area of deviation is taken to be $\pm$ 10% of the reported value from Refs. [5] [6]

[d]$T_2=\alpha*T_1$ where α has been optimized with the constrained $\alpha > 1$

[e]Determined by the geometry of the molecule.

replicated runs. On a single processor (500 MHz) one complete analysis can be executed in about 12 hours.

3.4 Results and Discussion

Table 3.3 shows the 12 parameters for all four experimental spectra as they resulted from the GA, together with the results of a re-evaluation of the spectra reported in Ref. [5] (indole, indazole and benzimidazole) and Ref. [6] (4-ABN) using the classical methods. The molecular constants from Ref. [5] are averages from multiple spectra and were determined using very accurate ground rotational constants from microwave experiments. Results reported in

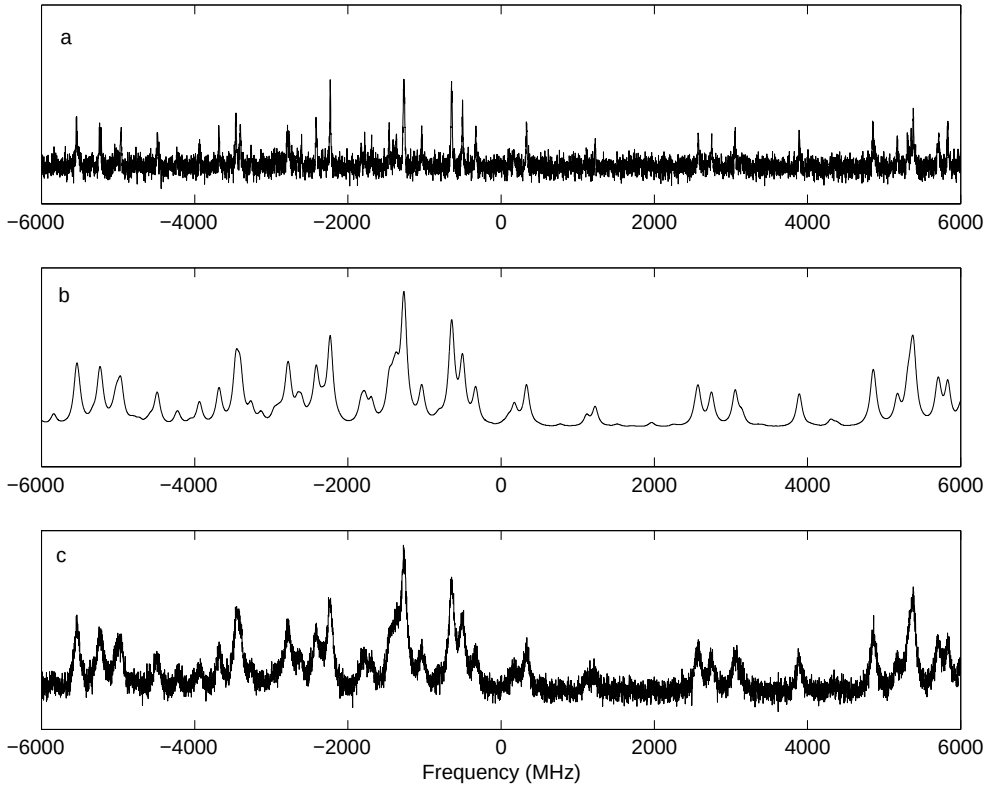

Figure 3.4: Synthesized spectra of indole with (a) S/N=15 (for strongest line), (b)$\Delta\nu$= 90 MHz and (c) S/N =15 (for strongest line) together with a $\Delta\nu$= 90 MHz. The intenstity on vertical scale is in arbitrary units.

Table 3.3 are based on a spectral analysis of the same spectrum as used for the GA-method and the ground rotational constants were also determined from that spectrum. The values obtained in the present GA approach are in close agreement with those from the classical method. For most of the parameters the results are within the experimental error. If the results are outside the error, the deviations are very small. These deviations are probably caused by the lack of precision of a GA. It is know that GA's can locate the global minimum but that they are not as precise as, for instance, local optimisers.

The GA method gives no information about the accuracy of the best fit parameters. However it should be possible to assign quantum numbers to the experimental transitions after a GA-fit. The experimental errors can then be estimated by performing a classical calculation like in Ref. [5], where it is no

longer necessary to go through the sometimes tedious assignment process.

All GA runs were repeated 5 times with different seeds for the random number generator and the solution with the lowest evaluation values are shown in Table 3.3. Results from Ref. [5] can be expected to be more accurate because the ground rotational constants were determined by microwave experiments which are more precise.

The parameters that describe the relative intensity of a transition (T_1, T_2, W) have different values in comparison with the reported values from Ref. [5]. (Ref. [6] used a one-temperature model so this cannot be compared with the present results).The deviation is due to the fact that for these parameters several sets can be used with equal spectral intensities as a result.

The GA was able to find the correct solution for the indole, indazole and benzimidazole spectra in all 5 replicated runs. For the 4-ABN data, the correct solution was only found in 2 of the 5 cases, as shown in figure 3.5. The cause of the reduced reproducibility of the 4-ABN run is probably due to larger boundary constraints, which makes it more difficult for the GA to locate the correct solution.

The absolute evaluation function values do not reach the same level for the 4 compounds. This is due to noise level, line width and total number of data points in a particular spectrum. High noise levels intrinsically give rise to large evaluation function values. However, the minimum reached in each case is the global minimum irrespective of the absolute evaluation value. The similarity between all four experimental and the corresponding calculated spectra is very high. As an example this is shown in figure 3.6 for a representative area of the spectrum of indole.

3.5 Applicability of the GA-method to partly resolved spectra

Figure 3.7 shows results for synthesised spectra of indole and benzimidazole with increased noise levels, line widths and spectral resolutions. Again, the GA-runs were repeated 5 times with different seeds for the random number generator. The best set of parameters found in these runs was used to generate spectra which are free of noise, have normal line widths and the same number of data points as the source spectra. The source spectra and the generated spectra are then compared with the evaluation function F. The evaluation values calculated in this way can directly be used to compare the quality of

Table 3.3: Results from GA runs for Indole, Indazole, Benzimidazole and 4-ABN.[a]

	Indole		Indazole	
	GA	Ref [5][b]	GA	Ref [5][b]
A"	3879.8	3880.7 (1.0)	3979.9	3979.2 (0.8)
B"	1637.0	1637.5 (0.4)	1633.8	1633.9 (0.3)
C"	1151.3	1152.1 (0.4)	1158.4	1158.6 (0.3)
T_1	2.22	1.50	2.60	2.60
T_2	7.93	5.03	7.35	8.18
W	0.1	0.22	0.23	0.19
θ	37.4°	±38.3°	62.3°	62.2°
ν^c	0.78	0.0 (1.6)	-1.7	0.0 (1.7)
ΔA	-134.70	-134.66 (0.09)	-102.44	-102.30 (0.09)
ΔB	-18.08	-17.96 (0.18)	-29.23	-29.20 (0.13)
ΔC	-20.72	-20.77 (0.32)	-23.31	-23.20 (0.28)
$\Delta\nu$	16.158	20.05	26.452	32.75
	Evaluation Values			
best	4.18		0.68	
mean	4.24		0.74	
std. dev.	0.08		0.06	
	Benzimidazole		4-ABN	
	GA	Ref [5][b]	GA	Ref [6]
A"	3929.0	3930.5 (1.0)	5579.7	5579.3 (0.5)
B"	1679.2	1679.5 (0.2)	990.23	990.26 (0.09)
C"	1177.1	1176.7 (0.2)	841.45	841.39 (0.08)
T_1	5.63	4.88	2.63	3
T_2	21.52	20.0	4.56	-
W	0.42	0.42	0.84	-
θ	22.1°	±22.0°	0°	0°
ν^c	1.04	0.0 (1.64)	-1.61	0.0
ΔA	-155.62	-155.70 (0.03)	-315.54	-316.61 (0.06)
ΔB	-15.30	-15.37 (0.08)	10.66	10.849 (0.003)
ΔC	-21.41	-21.31 (0.13)	0.29	0.095 (0.001)
$\Delta\nu$	19.33	19.45	16.16	26
	Evaluation Values			
best	0.65		1.2	
mean	0.71		14.7	
std. dev.	0.06		13.9	

[a]Values from Ref. [5] and Ref. [6] are listed in the respective columns. Rotational constants, ν and $\Delta\nu$ are in MHz, T_1 and T_2 in K.

[b]Results in this column differ partly from those reported in Ref. [5]. See text for details.

[c]The absolute frequency of the origin is given as the deviation from the reported value from Ref. [5] and Ref. [6].

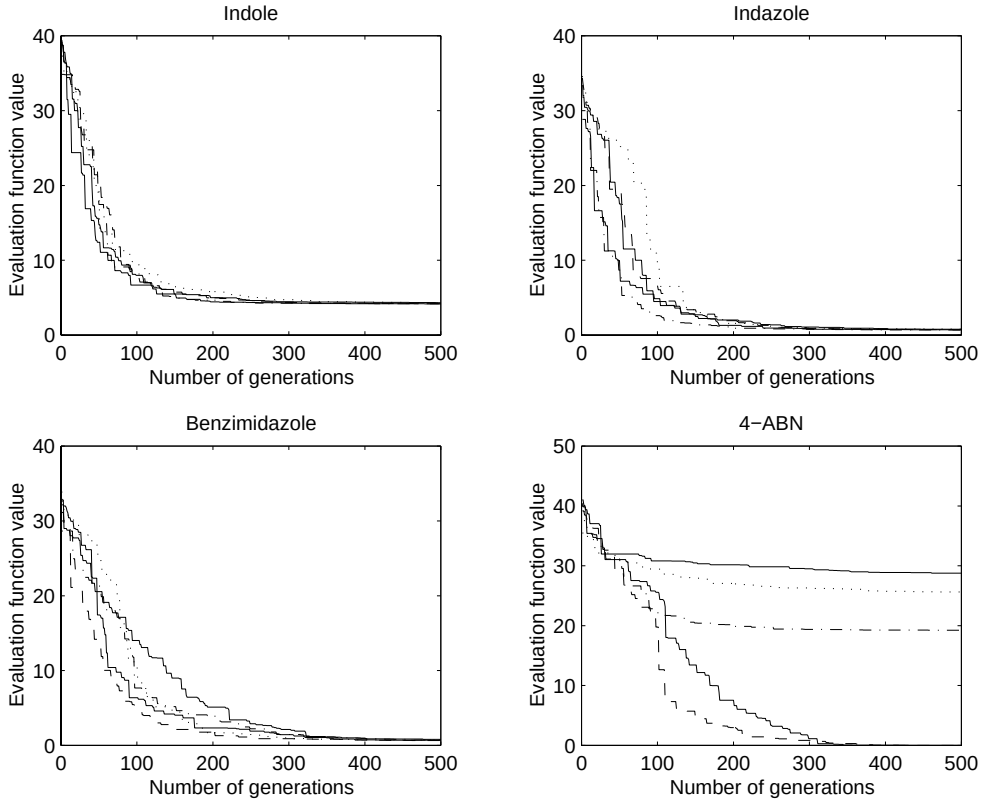

Figure 3.5: Progression of the best solution during a run for indole, indazole, benzimidazole and 4-ABN.

the different fits since the effects of added noise and line widths is removed from the evaluation function. Figure 3.7 demonstrates the results for indole and benzimidazole.

In both cases, all modifications to the calculated spectra lead to an increase in evaluation value and thus in a deterioration in quality of the solution. However, the effect of the increased line widths is somewhat less and more constant than the effects of other contributions. If the noise levels increase, the quality of the solutions decreases. The combination of both increased line widths and high noise levels does not lead to further deterioration of the best solution. A decrease of the number of data points (where the frequency is kept constant) only shows an effect on the benzimidazole spectrum. For a smaller number of data points, the solutions become worse. This is due to the fact

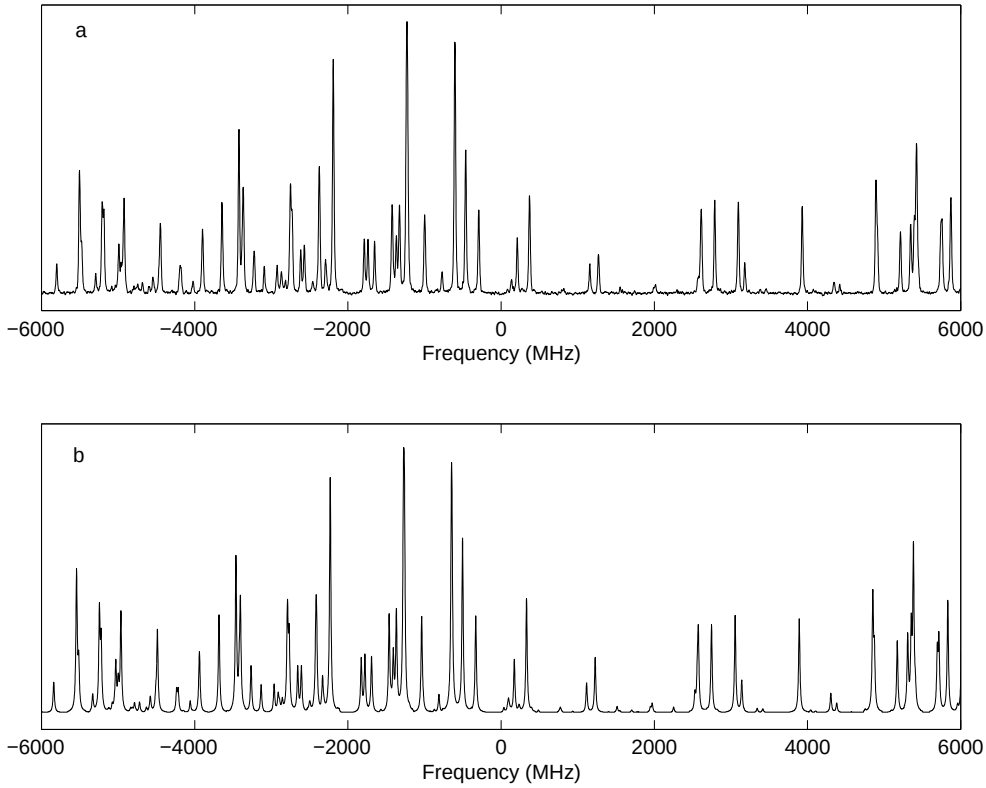

Figure 3.6: Representative area of the experimental (a) and calculated (b) spectrum of indole. The intensity on the vertical scale is in arbitrary units.

that spectral information gets lost if the distance between two successive data points becomes too large. Although the quality of the obtained parameters deteriorated, the rotational constants are hardly influenced by the elevated noise levels. The deviations are mostly found in T_1 and T_2 and in θ and θ_T. Because one is mostly interested in the rotational constants the method can be considered quite robust for the determination of these parameters.

3.6 Conclusion

The automated interpretation of high resolution spectra becomes of great importance if the interpretation by other methods is no longer feasible, too time-consuming or more a routine matter. In the approach presented in this

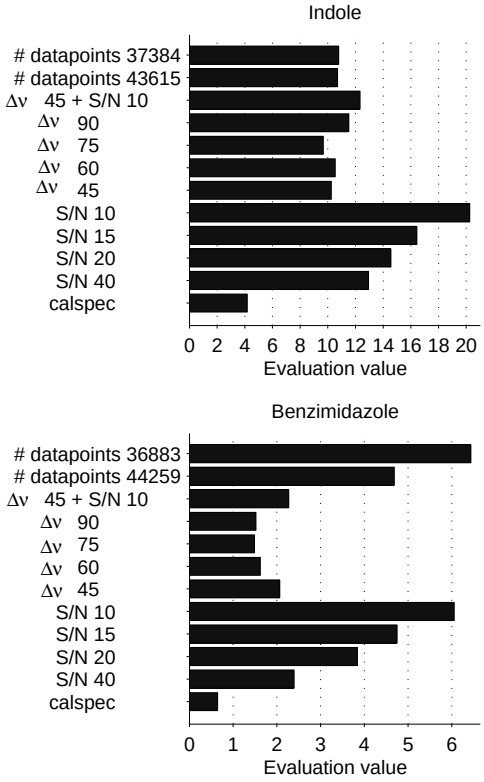

Figure 3.7: Influence of noise (S/N), line width ($\Delta\nu$) and the total number of data points in a spectrum on the best solution found of indole (top) and benzimidazole (bottom). Calspec indicates the spectrum which fits with the experimental one best.

paper, only knowledge of the range of the parameters is needed for the deduction of molecular constants. In general, feasible ranges can be given and may even be quite large. The meta optimising can be tedious for GA's. However, in the present case it is demonstrated that one set of GA settings suffices to retrieve the molecular constants from different rotationally resolved spectra. The success of the GA method crucially depends on the newly developed evaluation function. Other, more standard, evaluation functions lead to no results.

The problem of spectrum comparison in this particular application is related to peak shifts which are caused by small changes in the rotational constants. This makes it necessary to include a comparison of the neighbourhood of a given point in the spectrum. All attempts based on the sum of squared

differences without considering the neighbourhood of points failed, precisely because these criteria do not properly deal with peak shifts. This demonstrates that a special tailor-made evaluation function is crucial to obtain any results. It shows that, apart from an optimisation of the settings of the GA, GA's in combination with a standard evaluation function can not be used as a black box to solve any optimisation problem.

The GA method is quite robust. It is insensitive to large line widths in the spectrum, and only at very high noise levels do the results deteriorate. It is shown that the GA is able to use all information present in the spectrum and therefore its performance increases with the number of data points. The method of matching experimental data (represented as a vector) with simulated model data by optimising model parameters with a GA can be successfully used in other fields, especially with the newly developed evaluation function.

References

[1] G. Berden, W.L. Meerts, D.F. Plusquellic, I. Fujita, and D.W. Pratt. High resolution electronic spectroscopy of 1-aminonaphthalene: S_0 and S_1 geometries and $S_1 \leftarrow S_0$ transition moment orientations. *J. Chem. Phys.*, 104:3935, 1996.

[2] R.M. Helm, H.-P. Vogel, and H.J. Neusser. *Chem. Phys. Lett.*, 270:185, 1997.

[3] J.H. Kalivas, editor. *Adaption of simulated annealing to chemical optimization problems*, volume 15 of *Data handling in science and technology*. Elsevier, Amsterdam, 1995.

[4] F. Glover and M. Laguna. *Tabu Search*. Kluwer Academic Publishers, Dordrecht, 1998.

[5] G. Berden, W.L. Meerts, and E. Jalviste. Rotationally resolved ultraviolet spectroscopy of indole, indazole and benzimidazole: Inertial axis reorientation in the $S_1(^1L_b) \leftarrow S_0$ transitions. *J. Chem. Phys.*, 103(22):9596–9606, 1995.

[6] G. Berden, J. Van Rooy, W.L. Meerts, and Z.A. Zachariasse. *Chem. Phys. Lett.*, 278:373, 1997.

[7] R. Wehrens, E. Pretsch, and L.M.C. Buydens. The quality of optimisation by genetic algorithms. *Anal. Chim. Acta*, 388:265–271, 1999.

[8] R. Wehrens and L.M.C. Buydens. Evolutionary optimisation: a tutorial. *Trends in Analytical Chemistry*, 17(4):193–203, 1998.

[9] K.D.M. Harris, R.L. Johnston, and B.M. Kariuki. The genetic algorithm: Foundations and applications in structure solution from powder diffraction data. *Acta Cryst.*, A54:632–645, 1998.

[10] J. Dods, D. Gruner, and P. Brumer. A genetic algorithm approach to fitting polyatomic spectra via geometry shifts. *Chem. Phys. Lett.*, 261:612–619, 1996.

[11] H.R. Karfunkel, B. Rohde, F.J.J. Leusen, R.J. Gdanitz, and G. Rihs. Continuous similarity measure between nonoverlapping X-ray powder diagrams of different crystal modifications. *J. Comp. Chem.*, 14(10):1125–1135, 1993.

[12] R. de Gelder, R. Wehrens, and J.A. Hageman. A generalized expression for the similarity of spectra: application to powder diffraction pattern classification. *J. Comput. Chem.*, 22(3):273–289, 2001.

[13] D. Levine. PGAPack V1.0. PGAPack can be obtained from anonymous ftp from: ftp://ftp.mcs.anl.gov/pub/pgapack/pgapack.tar.Z.

Chapter 4

Powder pattern indexing using the Weighted Cross Correlation and Genetic Algorithms

X-ray diffraction is a powerful technique for investigating the structure of crystals and crystalline powders. Unfortunately, for powders the first step in the structure elucidation process, retrieving the unit cell parameters (indexing), is still very critical. In the present paper, an improved approach to powder pattern indexing is presented. The proposed method matches peak positions from experimental X-ray powder patterns with peak positions from trial cells using a recently published method for pattern comparison (weighted cross correlation). Trial cells are optimised with Genetic Algorithms. Patterns are not pretreated to remove any existing zero point shift as this is determined during optimisation. Another improvement is the peak assignment procedure. This assignment is needed for determining the similarity between lines from trial cells and experiment. It no longer allows calculated peaks to be assigned twice to different experimental peaks, which is beneficial for the indexing process. The procedure proves to be robust with respect to false peaks and accidental or systematic absences of reflections and is successfully applied to powder patterns originating from orthorhombic, monoclinic and triclinic compounds measured with

This chapter is published as: J.A. Hageman, R. Wehrens, R. de Gelder and L.M.C. Buydens, J. Comp. Chem., 24(9), 1043-1051, 2003.

synchrotron as well as with conventional laboratory X-ray diffractometers.

4.1 Introduction

X-ray diffraction is perhaps the most important method for investigating the structure of crystalline materials. Structure determination from single crystals is nowadays a standard technique for the identification of new compounds, the analysis of stereochemistry and the elucidation of intra- and intermolecular interactions in various materials [1]. As it may be impossible to grow suitable crystals, not all compounds can be examined by single-crystal diffraction. Although powder diffraction data contain less information, it is often possible to use powder data for a full structure determination. For this reason, structure determination from powder diffraction data is becoming increasingly important and consequently it receives more and more attention in literature.

Elucidation of a crystal structure from powder diffraction data consists of four stages [1]:

1. Identification of the repeating unit of the sample (the unit cell)
2. Space group determination
3. Structure solution (building of an initial model)
4. Structure refinement (obtaining a final and accurate model)

Today, steps two, three and four can be performed reasonably successful, although there always remain specific problems. Step one, identifying the unit cell (this process is also referred to as indexing), has become the bottle neck in the complete process of structure elucidation and this is often the reason why a structure elucidation of a compound from powder data is unsuccessful.

There are a number of reasons why indexing is difficult and often fails, especially for low symmetry cases. Several reasons are of an experimental nature; e.g. peak broadening, which leads to a loss in resolution, and peak shifts. Other reasons are more fundamental in nature, for instance the increasing peak density at higher angles (with low intensities), which makes correct peak assignment difficult, systematic and accidental absences of reflections, dominant zones and/or the possible presence of impurity phases, leading to spurious peaks.

Methods for indexing powder patterns have adopted several strategies. They might use either peak positions or the whole profile. Whole-profile fitting methods have to deal with peak profile description: asymmetry, peak shape variation, etcetera. Using the whole profile requires incorporation of reflection intensities, when calculating trial patterns. This knowledge is in principle not present at the indexing stage, so peak intensities have to be assigned (for example using a Pawley refinement [2]). When a continuous pattern is converted into a list of peaks, some of these disadvantages are overcome. Information on peak shapes and peak width is simply not used, so this cannot interfere with the fitting process. Crucial will be the quality of the peak picking process. Peak picking has two problems. The first is picking spurious peaks (false positives), the second one is not picking existing peaks (false negatives). In the first case there will be an extra peak in the experimental peak list, which might interfere with the fitting of the experimental pattern. This extra peak can also be seen as a peak originating from an impurity phase. In the second case the peak list will be one peak short. As usually not every theoretical reflection is present in a powder pattern, due to accidental and systematic absences, an extra absent peak will very likely not cause a problem when enough peaks are present.

All indexing strategies generate trial powder patterns and calculate the agreement between experiment and theory. When the agreement is sufficiently high, the results are presented as possible solutions to the user. A number of criteria exist for the quantification of the agreement between experimental and trial powder patterns [3–5]. Most criteria sum, in some way, the discrepancies between experimental and calculated lines. An often used figure of merit (called M_{20} or M_N) was proposed by De Wolff [6]. This criterion sums all the discrepancies in peak positions, after assigning the nearest calculated peaks to experimental peaks. This figure of merit is very sensitive to peak accuracy, as was demonstrated by De Wolff [6], and also to false peaks, as will be shown in this paper. Moreover, as part of De Wolff's strategy, when two experimental peaks are assigned to one calculated peak, one of the experimental peaks is marked as unindexed and will not contribute to M_N but to the auxiliary X_{20}, a situation which should not occur for the true cell parameters.

Trial solutions are generated by either varying the h, k, l indices or by direct modification of the unit cell parameters. This can be done exhaustive or in a grid like manner. Several methods have applied specific crystallographic knowledge for aiding exhaustive searches. This knowledge diminishes the search space and increases the chances for obtaining a good solution. A

property of all current methods is that they do not correct for zero point shifts (originating e.g. from sample misalignments) on the fly, which leads to erroneous results if a significant zero point shifts is present in the experimental data. An overview of existing indexing methods can be found in Ref. [1].

A more recent approach for generating trial cells is the use of Genetic Algorithm's (GA's) [4, 7], a global optimisation method that has been shown to be successful in a variety of applications. In this paper, an improved method for indexing powder patterns is presented. It is based on GA's and the use of the weighted cross correlation (WCC). As seen in previous work [8–10], the use of the WCC figure of merit ensures a certain insensitivity to peak broadening and peak shifts. This figure of merit is applied here to peak positions alone and uses no information on intensities. The zero point shift correction is treated as an extra variable during optimisation, so patterns do not need to be pretreated. The peak assignment procedure is different from the one suggested by De Wolff [6]. When a calculated peak should be assigned twice to different experimental peaks, as a results of distance considerations, a different assignment is made here, allowing only the assignment of one calculated peak to one experimental peak. This also has beneficial effects on the indexing process, as will be shown. The procedure uses a minimum of prior knowledge of the molecule under study and proves not to be very sensitive to impurity phases. The method is successfully applied to powder patterns originating from orthorhombic, monoclinic and triclinic compounds, measured with synchrotron as well as conventional laboratory X-ray diffractometers.

4.2 Theory

4.2.1 Calculating powder lines

A unit cell can be described in terms of three cell edge lengths (a, b, c) and the angles between the edges (α, β, γ) as is shown in figure 4.1, where α is the angle between b and c, β is the angle between a and c and γ is the angle between a and b. All unit cells can be grouped into seven crystal systems, which range from simple (cubic, where $a = b = c$ and $\alpha = \beta = \gamma = 90°$) to more complex (triclinic, where $a \neq b \neq c$ and $\alpha \neq \beta \neq \gamma$).

It is possible to calculate all theoretical peak positions in a powder pattern for a given compound provided that its unit cell parameters are known. The formula for determining the positions of peaks in a theoretical powder pattern is shown in Eq. 4.1. A detailed description of this formula is beyond the scope

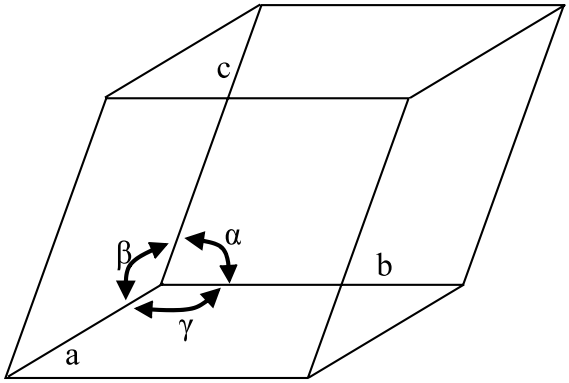

Figure 4.1: Definition of the unit cell, where α is the angle between cell edge length a and b, β is the angle between a and c and γ is the angle between a and b.

of this paper, but details can be found in textbooks like ref. [11]. In short, it calculates, assuming the most general crystal system, all lattice plane spacings d_{hkl}.

$$d_{hkl} = \sqrt{(h^2A + k^2B + l^2C + 2klD + 2hlE + 2hkF)} \tag{4.1}$$

Here $A = a^* \cdot a^*, B = b^* \cdot b^*, C = c^* \cdot c^*, D = b^* \cdot c^*, E = c^* \cdot a^*, F = a^* \cdot b^*$. The asterixes indicate reciprocal lattice unit cell vectors. The d_{hkl} values are converted by Bragg's law (Eq. 4.2) into 2θ values.

$$2\theta = 2sin^{-1}\frac{\lambda}{2d_{hkl}} \tag{4.2}$$

These 2θ values can be compared with the location of the peaks in the experimental powder pattern. When we have no information about the structure of our compound under investigation, or any information on electron densities, we are not able to calculate the intensity of the peaks. So a peak from a calculated powder pattern is only represented by its 2θ value, and the complete pattern is given by a vector of peak positions.

Theoretical peak positions are determined by 11 parameters:

- three cell edge lengths and three cell angles. When there is reason to assume that some of these lengths are equal or that some angles are 90° or 120° this knowledge can be used.

- one lambda parameter which gives the radiation wavelength which was

used during the experiment. This parameter is usually known and it is kept fixed during optimisation.

- a zero point shift parameter. This parameter counteracts a shift which arises from the physical thickness and/or misplacement of the sample or any misalignments in the apparatus.

- one parameter giving the starting angle, one parameter giving the maximum angle (both in degrees) and one parameter giving the sampling rate. These three parameters are in principle defined by the experimental pattern but the maximum angle must be chosen carefully to select an appropriate number of lines for the indexing process.

Since four parameters are experimental or optimisation settings, seven parameters are determined by the indexing procedure.

4.2.2 Fitting an experimental pattern

When the collection of peak positions in a theoretical powder pattern is very similar to the one determined from the experimental pattern it is very likely that the unit cell parameters, used for the calculation of the theoretical pattern, are the true values. Determining the set of parameters which result in calculated powder lines equal to the experimental ones can be seen as a parameter optimisation problem in seven dimensions with many local optima. The global optimum should represent the true parameters in this case. This type of problem can be solved with global optimisation methods such as Genetic Algorithms (GA), Simulated Annealing (SA) or Tabu Search (TS), using a suitable evaluation function. In the approach described here, GA's are used to optimise the model parameters. GA's are a special class of global optimisers based on the theory of evolution. A GA is able to minimise (or maximise) a function G(x), where x represents a parameter vector, by searching the parameter space of x for the optimal solution [12, 13]. In general, GA's operate not on a single trial solution but on a group of solutions, called a population. Using evolutionary inspired operators such as fitness, crossover and mutation, the best solutions are modified and passed on to the next generation. In this way the population as a whole moves towards better solutions and hopefully to the global optimum. A critical part in powder indexing is the assessment of the similarity between an experimental powder pattern and a trial powder pattern (or assessing the fitness of a trial solution). This assessment is required

to give the GA the ability to discriminate between similar and dissimilar solutions so it can select accordingly. For a tutorial on Genetic Algorithms, the reader is referred to Ref. [14].

4.2.3 Evaluation function

One way to determine the dissimilarity between two (continuous) patterns is to sum the (weighted) differences between the intensities of the patterns pointwise, for instance via a root-mean-square-error (RMS) approach like in Eq. (4.3):

$$\mathrm{RMS} = \sqrt{\frac{\sum_{x=1}^{x=k} \left(f(x) - g(x)\right)^2}{k}} \tag{4.3}$$

Here *f(x)* and *g(x)* are vectors f and g with equal size k. The weighted R-profile in the work of Harris et al. [7, 15] is an example of this. However, in the indexing stage only information on the positions of peaks in the calculated patterns and no information on intensities is available, so this approach is only applicable in combination with Pawley refinement. In most methods for indexing, the experimental powder pattern is converted by a peak picking method to a list of peak positions. Figure 4.2 shows the conversion, for one of the test compounds, C6F, from a continuous pattern to peak positions (indicated at the bottom of the figure with PP).

A possibility to assess the dissimilarity between two lists of peaks is to sum the differences in positions of corresponding peaks, like in the approach of De Wolff [6]. An example is given in figure 4.3. The differences are summed like in Eq. (4.4):

$$D_{fg} = \sqrt{\frac{\sum_{i=1}^{i=N} min(L_{f,i} - L_g)^2}{N}} \tag{4.4}$$

Here the list of peaks of experimental pattern f and calculated pattern g are indicated by L_f and L_g, N is the number of experimental peaks and $min(L_{f,i} - L_g)$ indicates the shortest distance between peak i in list f and any peak in list g. As we will demonstrate later in this paper, the D_{fg} dissimilarity measure is not robust towards false peaks.

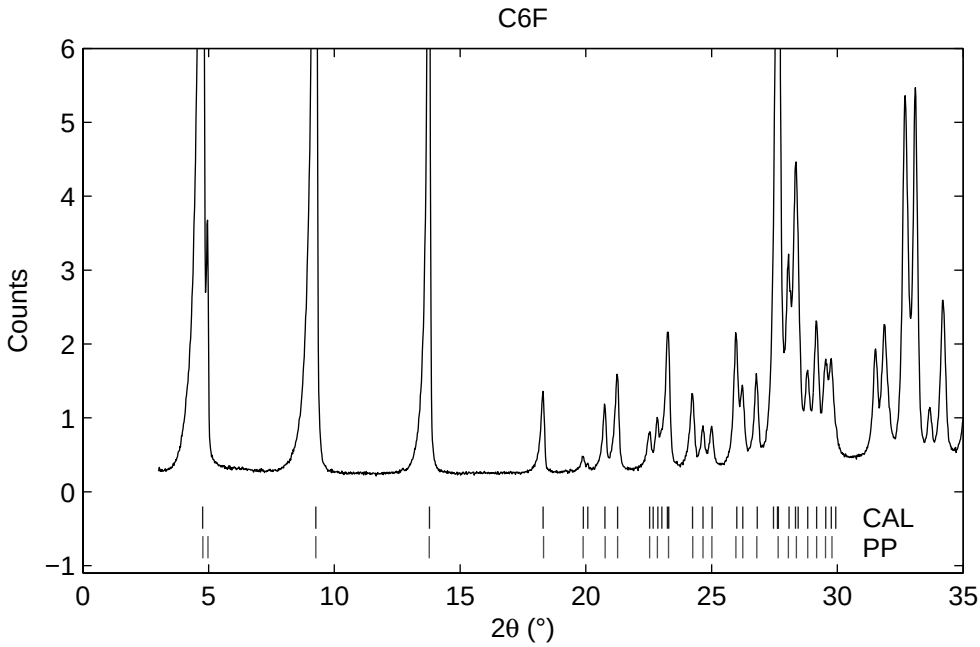

Figure 4.2: Part of the powder pattern of C6F. The vertical lines in the bottom of the figure indicate which peaks are selected with a peak picking method (indicated with PP), the lines indicated with CAL give the positions of the peaks resulting from the best fit. Only the range of 0-30° is considered.

A recently developed function which is robust against variations in peak positions, is the Weighted Cross Correlation (WCC) measure. This function has been applied for clustering powder diffraction data [10] and for the determination of molecular constants from laser induced fluorescence (LIF) spectra [8,9]. In short, the WCC function is based on the normal cross correlation function as given in Eq. (4.5):

$$C_{fg}(r) = \sum_{x=1}^{x=k} f(x) \cdot g(x+r) \tag{4.5}$$

Again, *f(x)* and *g(x)* are vectors f and g with equal size k. Eq. (4.5) calculates the cross correlation coefficient between two vectors when one vector has been shifted relative to the other by an amount r. Coefficients calculated with larger shifts (r-values) can be penalised by a simple triangular weighting function *w(r)*, where t is the maximum displacement that is taken into account:

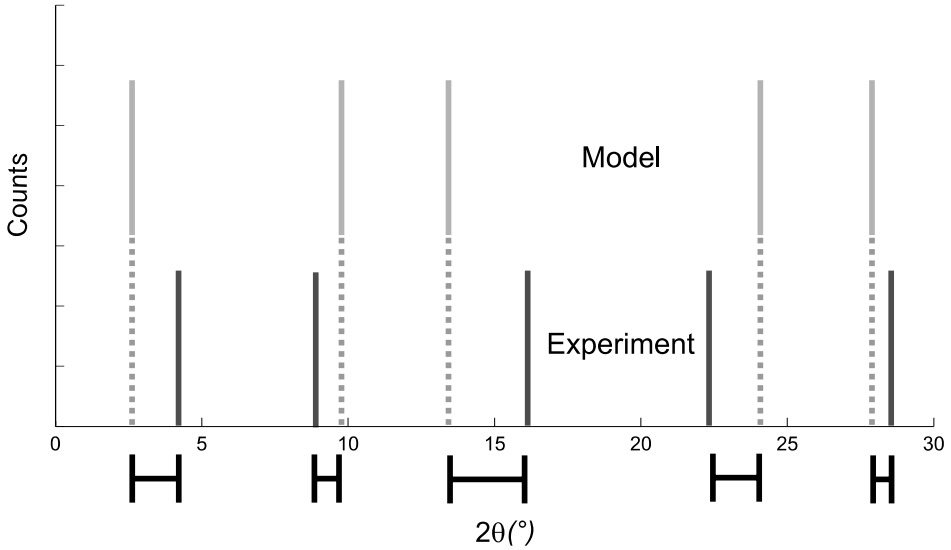

Figure 4.3: Example of a similarity measure which sums the distances between peak positions in an experimental and calculated powder pattern. Vertical lines indicate the position of each peak and horizontal lines below x-axis indicate the shortest distance between two pairs of peaks.

$$w(r) = 1 - \frac{\mid r \mid}{t} \qquad \text{if } r \leqq t$$
$$w(r) = 0 \qquad \text{if } r > t \tag{4.6}$$

Combining Eqs. (4.5) and (4.6) and applying normalisation yields Eq. (4.7), the weighted crosscorrelation function, which describes the area under the weighted cross-correlogram:

$$C_{fg}^{ws} = \frac{\sum_{r=-t}^{r=t} C_{fg}(r)w(r)}{\sqrt{\sum_{r=-t}^{r=t} C_{ff}(r) * w(r)} * \sqrt{\sum_{r=-t}^{r=t} C_{gg}(r) * w(r)}} \tag{4.7}$$

For two identical patterns C_{fg}^{ws} is one and for distinctly different vectors C_{fg}^{ws} is close to zero. An example of the behaviour of $C_{fg}^{ws}(r)$ is given in Figure 4.4. The top row shows the unweighted crosscorrelogram, the weights and the weighted

crosscorrelogram, respectively, for the comparison of the experimental C6F pattern with the theoretically correct pattern. The area under the curve (C_{fg}^{ws}) is equal to 1.00. The bottom row shows the same for an arbitrary theoretical pattern, compared with the same experimental C6F pattern. Shifts from -5° to 5° are considered.

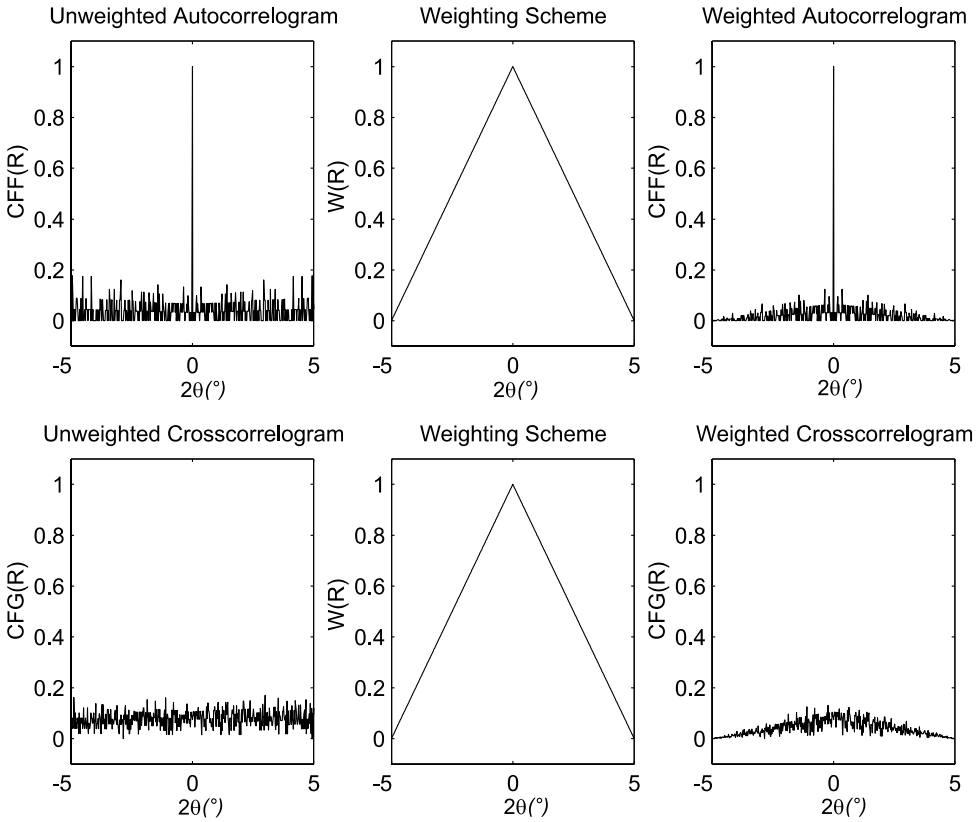

Figure 4.4: The top row shows the unweighted crosscorrelogram, weights and weighted crosscorrelogram corresponding to the comparison of the experimental C6F pattern with the correct theoretical pattern. The bottom row shows the same for the comparison of the experimental C6F pattern with an arbitrary theoretical pattern.

In our case, patterns f and g are not continuous patterns but consist of discrete peaks with equal intensities. It is very likely that a calculated pattern will contain more peaks than the experimental pattern, due to accidental and/or systematic absences. This can, for example, be seen in figure 4.2 where there are six extra peaks in the correct pattern (indicated by CAL) compared

to the number of peaks resulting from the peak picking procedure (indicated by PP). If a calculated pattern is correct, it should match every experimental peak (neglecting the possibilities of impurity phases for the moment). If there are more peaks present in the calculated list, the extra peaks should be left out before comparison. To get equal numbers of peaks, a subset of peak positions must be selected from the calculated pattern. The subset is chosen in such a way that the peak closest to a given peak in the experimental pattern is selected. It is also made sure that every peak in the experimental pattern is matched with one unique peak in the calculated pattern, unlike in the original De Wolff approach [6] which uses an auxiliary criterion to deal with this.

The value returned by the WCC function is transformed into the final evaluation function F by using information about the estimated volume of the unit cell. An estimation of the molecular volume can be made by using the 18Å^3-rule, which multiplies all non-hydrogen atoms by 18Å^3. As the correct number of molecules in a unit cell is in principle unknown, several runs are executed in which the number of molecules present in a unit cell is varied. During these runs the estimated molecular cell volume is simply multiplied by the number of molecules in the unit cell to obtain an estimation for the unit cell volume.

If the calculated cell volume deviates more than 20% from the estimated volume, the evaluation function value F is given by:

$$F = C_{fg}^{ws} * \frac{V_{maxdiff}}{\mid (V_{calc} - V_{est}) \mid} \tag{4.8}$$

If not, F equals C_{fg}^{ws}. In Eq. (4.8) $V_{maxdiff}$ is the maximum allowed deviation from the estimated volume, V_{calc} is the cell volume corresponding to the trial pattern and V_{est} is the estimated experimental cell volume. The evaluation function F is maximised by the GA.

4.3 Experimental

4.3.1 Indexing powder patterns

The procedure described in the previous sections has been applied to a set of 16 powder patterns. The data set consists of five triclinic, eight monoclinic and three orthorhombic cases. Eleven cases have been reported in literature before and were solved earlier by other methods. Five cases were measured in-house (crystal data were available from single-crystal diffraction experiments).

Part of this data set has also been used to demonstrate the usefulness of several additions to TREOR [5], an existing indexing program.

All continuous powder patterns were converted into a list of peak positions by the Bruker AXS Diffrac^{plus} EVA V4.0 software. This method performs a Savitsky-Golay smoothing combined with a second derivative method. Peak width and threshold parameters were set to their default values. The results of the peak picking procedure were checked visually and for eight patterns this resulted in the removal of a few peaks per pattern. Removed peaks were either very low in intensity, using a signal-to-noise ratio of 3 as upper limit for the intensity, had strongly deviating peak widths compared to the majority of peaks or could clearly be recognized as artifacts of the peak picking procedure. The removed peaks in the eight patterns are shown in figure 4.5.

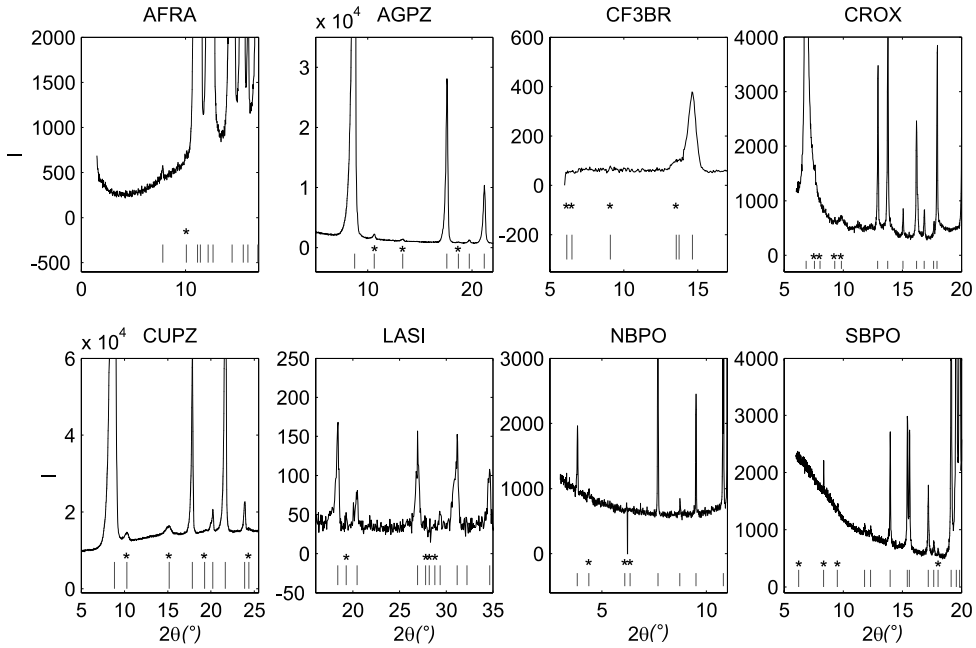

Figure 4.5: Peaks which were removed after the peak picking procedure. Vertical lines indicate peaks found by the peak picking procedure, asterisks indicate manually removed peaks. Removed peaks were either very low in intensity, had strongly deviating peak widths or could clearly be recognised as artifacts of the peak picking procedure.

Each pattern was fitted for all seven crystal systems. For each fit, a corresponding number of parameters was optimised with the GA (for instance one

for the cubic and six for the triclinic crystal system). An extra parameter that was included in the fit was the zero point shift. All parameters were coded as 20 bit gray binary numbers [14]. All trial solutions were transformed into reduced cell parameters and the reduced cell parameters were put back on the string. This speeded up convergence considerably. However, calculating reduced cell parameters for the monoclinic and tetragonal crystal systems can lead to a non-standard setting. To overcome this, it was decided not to use reduced cell parameters for these crystal systems. Our evaluation function F, based on the WCC criterion, was compared with the D_{fg} function. As will be shown below, results for the WCC criterion are much better.

The patterns were fitted with the boundary constraints given in Table 4.1. For the AFRA, AGPZ, CIME, CUPZ, NBPO, SUCROS and YONO samples, the maximum axis length was set to 30Å. The other patterns were fitted with a maximum axis length of 20Å. Setting boundary constraints to large values ensures that no prior knowledge is used on cell parameters. However, when some prior knowledge is present, this can be used to narrow the lattice parameter space and this will speed up the complete optimisation process.

The first 20-30 experimental peaks were usually used as input. This means that roughly the area up to 25-30° was used.

Table 4.1: Typical boundary constraints for unit cell parameters and zero point shift[a].

Parameter	Boundary constraints
a	3 - 20 or 30Å
b	3 - 20 or 30Å
c	3 - 20 or 30Å
α	60 - 120 °
β	60 - 120 °
γ	60 - 120 °
zero point shift	-0.5 - 0.5 θ

[a]If required by the crystal system, some parameters have been kept at a constant value.

The optimal size of the triangular weight function (l in Eq. 4.6) was determined to be 5.0°. Obtaining good solutions with values smaller then 5.0° proved difficult for some patterns, while larger values only slowed down the calculations. The settings of the GA are given in Table 4.2. All runs were repeated five times to exclude any effects of the random starting population.

Table 4.2: GA settings.

Setting	Value
maximum number of generations	300
population size	500
starting population	randomly initialised
elitism	250
crossover type	two-point-crossover
crossover probability	0.85
mutation type	flipping random bit
mutation probability	0.25
selection type	proportional
fitness type	raw[a]

[a]Fitness value increases inversely proportional with evaluation value of a string.

All GA calculations were performed with the GA library PGAPack version 1.0 [16]. PGAPack, the model and the evaluation function are written in ANSI-C, the function which converts unit cell parameters to their reduced counterparts was written in Fortran. All calculations were performed on a Sun-Ultra-10 running at 440 MHz. The average runtime was about five minutes per fit, which comes to roughly three hours for a complete fit for every crystal system in five fold. In practice, run times can be reduced considerably because it will probably not be necessary to try all crystal systems and very likely not in five fold.

4.3.2 Robustness

The robustness towards false peaks is tested for both D_{fg} and WCC based evaluation functions (based on Eq. 4.4 and 4.7, respectively) by using the powder pattern of C6F, as shown in figure 4.2. The continuous powder pattern of C6F is converted into a list of peak positions. During this conversion the first peak (around 5°) was mistakenly represented by two peaks in the peak list. The effect of this false peak on one parameter, cell edge length b, will be shown. To this end, the experimental pattern will be compared with several patterns which are calculated with all parameters set to their correct values (which are known from single crystal experiments and are shown in table 4.3) except for cell edge length b which is varied over the range 3.5 - 5.5 Å. Together with the influence of false peaks, two methods for selecting the subset of peaks from

the calculated pattern are tested. In the first scheme, the nearest calculated peak is assigned to an experimental peak. It is, however, possible that one calculated peak gets assigned to two (or more) experimental peaks, a situation which cannot be right. The second scheme prohibits these double assignments by stating that a calculated peak can only be assigned once.

4.4 Results and Discussion

4.4.1 Indexing powder patterns

Table 4.3 shows the unit cell parameters and the zero point shift for the 16 powder patterns as they resulted from the GA fit, together with results from literature. As can be seen, the results from the GA fit are in close agreement with earlier reported values. In some cases the precision of the results can be improved by repeating the run with narrower boundaries. An example of this is also given in Table 4.3 for the fit of the powder pattern of YONO. The values in brackets indicate the results with the smaller boundaries. It is a well known fact that GA's are very well able to locate a global minimum (the correct set of unit cell parameters in this case), but they are not as precise as to pinpoint the exact location, as for instance local optimisers can [13]. Narrowing the search space will increase the ability to locate the exact solution. Only the results with the best evaluation value are shown in table 4.3. Usually, the global optimum was located several times out of five runs, with solutions only differing in precision. As an example, the progression of all five runs of C6F is shown in figure 4.6. For this particular pattern all runs lead to the same solution, with one being slightly better. The powder pattern C6F could not be solved by the CRYSFIRE package [17], a collection of available indexing techniques which contains methods like ITO, DICVOL and TREOR. This is likely caused by the false peak around 5° and the large zero point shift. Our GA method is not sensitive to this peak.

4.4.2 Robustness

When fitting the powder pattern of C6F with an evaluation function based on the D_{fg} function, the results do not describe the true cell parameters very well. Both cell edge lengths as well as angles show sometimes (large) deviations from the cell parameters obtained from single crystal data. Results for C6F are shown in table 4.4. The fit for C6F with the D_{fg} evaluation function is also

Table 4.3: Compounds in the data set and unit cell parameters.

Compound[a]		a	b	c	α	β	γ	Z.P.[e]
AFRA(X)	[b]	8.09	8.16	11.83	96.37	102.53	97.56	0.038
		8.10	8.26	11.84	96.24	102.84	97.43	
AGPZ(X)	[18]	6.46	6.52	20.03				0.101
		6.47	6.53	20.06				0.101
C6F(X)	[b]	4.02	4.60	19.76	85.80	87.80	80.80	0.294
		4.02	4.59	19.71	85.04	87.54	80.40	0.247
CF3BR(N)	[19]	8.11	5.84	7.93		111.65		0.333
		8.14	5.85	7.96		111.72		0.243
CIME(S)	[20]	6.82	10.38	18.82			106.45	0.007
		6.83	10.39	18.83			106.44	
CROX(X)	[21]	5.43	6.54	12.02	75.00	89.34	78.12	0.028
		5.42	6.56	12.05	74.87	89.25	77.95	
CUPZ(X)	[18]	6.09	6.67	19.85				0.002
		6.08	6.67	19.83				
LASI(N)	[22]	5.41	8.78	13.15		90.33		0.079
		5.41	8.80	13.18		90.50		
NBPO(S)	[23]	8.72	8.80	29.89	91.72			0.000
		8.72	8.79	29.87	91.77			
NIZR(S)	[24]	8.83	8.94	12.41		90.56		
		8.84	8.93	12.39		90.55		
PBS(S)	[25]	6.92	7.18	16.11				0.066
		6.92	7.18	16.11				
SBPO(S)	[26]	8.36	8.84	11.94		91.289		0.088
		8.32	8.73	11.94		91.12		
SUCROS(S)	[b]	7.76	8.71	10.86		102.90		0.10
		7.72	8.68	10.82		102.98		
TITUL6(X)	[b]	6.45	9.50	12.21	96.90	98.37	94.64	0.041
		6.41	9.50	12.16	97.10	98.24	94.55	
VFXBS1(X)[c]	[b]	10.12	10.23	13.83	72.17	70.00	63.89	0.033
		10.11	10.21	13.81	72.22	69.97	63.92	
YONO(S)	[d] [27]	3.80	9.19	16.33			91.83	0.045
		3.63	9.38	16.38			101.06	0.001
		3.63	9.38	16.39			101.070	

[a]All unit cell parameters are given in reduced cell parameters, second line per compound are literature values. [b]Compounds are measured inhouse and values on the second line are results from single crystal analysis. [c] peak at 7.6° has been removed after identification as an impurity phase. [d]Values on second line indicate results when smaller boundaries are used during optimisation (a=[3,10], b=[5,15], c=[10,20], β=[95,110]). Third line are literature results. [e]Zero point shift: not all zero point shifts of all compounds from literature are known. X: data collected on conventional laboratory diffractometer, N: neutron data, S: synchrotron data.

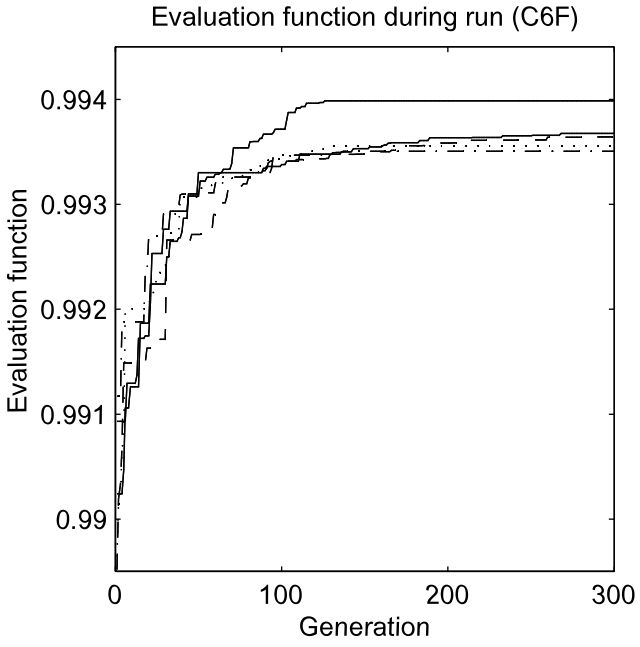

Figure 4.6: Progression of the best solution during all five replicate runs for the fit of compound C6F.

repeated with the false peak removed. Although the results are better, they do not completely agree with the single-crystal data. Clearly, the WCC-based evaluation finds the correct solution.

Figure 4.7 shows the results for the tests of the influence of the false peak on both types of evaluation functions. In the top figure the comparisons are shown for the D_{fg} function. The solid line shows the values obtained with the correct peak position list. The minimum of this line corresponds to the correct value for b (which is 4.61 Å). When we use the peak list with the false peak present, the solid line changes to the dashed line. The minimum has shifted to the position marked with MIN. The solid and dashed line in the top figure both exhibit significant roughness. This has its origins in the way the subset is chosen. In the top figure the subset is chosen in such a way that every experimental peak gets matched with the closest calculated peak from the model. When applying the extra restraint on the subset selection so that every calculated peak can only be assigned once to an experimental peak, we get the middle figure. The dashed and solid line are now much smoother, which is usually beneficial for optimisation methods in general because it makes it

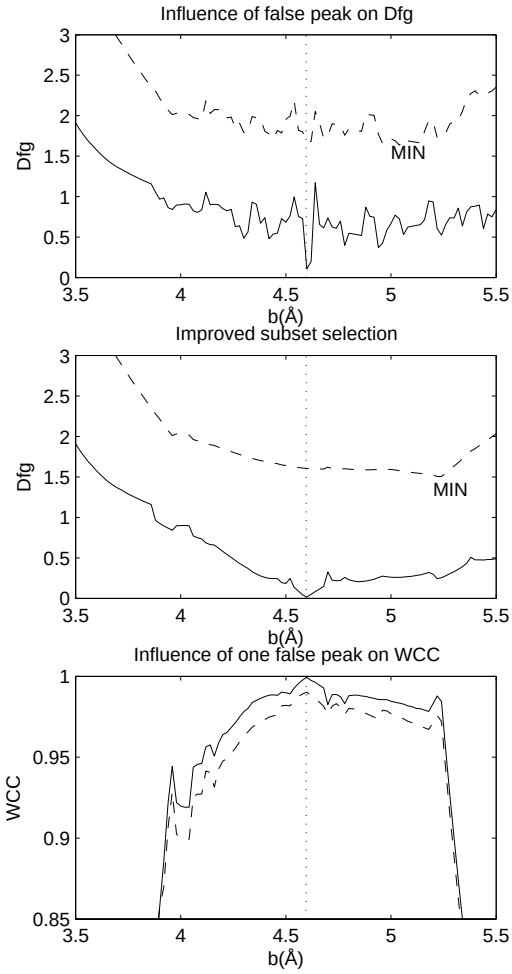

Figure 4.7: Effect of a false peak on D_{fg} with different subset selection approaches (top and middle figure) and the effect of a false peak on the WCC (bottom) based evaluation function. Solid lines indicate no false peak present, dashes lines indicate one false peak present. Vertical dotted lines indicate location of correct minimum. MIN indicates shifted minimum of D_{fg} function with one false peak.

easier to locate the minimum. Nevertheless, we can see that the addition of one false peak still shifts the minimum to the right. A fit with this type of evaluation function will lead to incorrect results if peaks are shifted or wrong peaks are present.

If we use an evaluation function based on the WCC and allow for calculated

Table 4.4: Results of the fit for C6F

C6F[a]	a	b	c	α	β	γ	zero point shift
Single Crystal	4.02	4.59	19.71	85.04	87.54	80.40	0.247
D_{fg} +f.p.	3.39	4.35	24.73	94.45	93.04	99.79	0.495
D_{fg} -f.p.	3.80	4.53	19.56	86.22	86.23	84.91	0.162
WCC +f.p.	4.02	4.60	19.76	85.80	87.80	80.80	0.294

[a]All unit cell parameters are given in reduced cell parameters, +f.p., -f.p. indicate with or without the false peak respectively.

peaks to be assigned only once, recalculate our previous example of C6F, we get the bottom figure. Here we see that the only difference between the comparison with or without a false peak is an offset difference between the two curves. The shape and the location of the minimum remain the same. This is a much better behaviour compared to the D_{fg} because the minimum remains at the same position. However, there is a limit to the increased robustness. Some false peaks will remain troublesome, even with the WCC figure of merit. An example of this is VFXBS1, where the presence of a false peak at 7.6° disturbs the indexing process leading to a failure.

The shape and location of the minimum are thus influenced by the presence of false peaks. However, this influence is dependent on the location of the false peak, being larger at lower angles.

4.5 Conclusions

Powder pattern indexing is an area of the complete process of structure elucidation which deserves attention as it usually is a slow and cumbersome process. The method presented in this paper, demonstrates several improvements. The use of the WCC function as a new figure of merit and the application of it to peak positions, yields a certain robustness. The improved way to select a subset of peaks yields a more smoothed error landscape which is beneficial for the optimisation process. Zero point shifts do not need to be pretreated, as the GA is capable of correcting this, by treating it as an extra optimisation variable.

An important advantage is the increased robustness against false peaks, as these peaks are usually hard to detect and mostly heavenly disturb the

indexing process. The increased robustness may result in the indexing of patterns which could not be indexed before.

The method is very general and uses no crystallographic knowledge on the sample. However, when prior knowledge is available, this can be used to speed up the optimisation process, by shrinking the search space.

Results are in close agreement with known literature values, so no fine tuning with local optimisers is necessary. No tedious fine tuning of the meta-parameters of the GA is needed, since one standard set of GA parameters is able to correctly index a wide variety of examples. The method is demonstrated with orthorhombic, monoclinic and triclinic examples, measured on synchrotron and conventional laboratory X-ray diffractometers.

References

[1] K.D.M. Harris. Crystal structure determination from powder diffraction data. *Chem. Mater.*, 8:2554–2570, 1996.

[2] G.S. Pawley. Unit-cell refinement from powder diffraction scans. *J. Appl. Crystallogr.*, 14:357–361, 1981.

[3] D. Taupin. Information in powder pattern indexing. *J. Appl. Cryst.*, 21:485–489, 1988.

[4] W. Paszkowicz. Application of the smooth genetic algorithm for indexing powder patterns - tests for the orthorombic system. *Materials Science Forum*, 228-231:19–24, 1996.

[5] A. Altomare, C. Giacovazzo, A. Guagliardi, A.G.G. Moliterni, R. Rizzi, and P.E. Werner. New techniques for indexing: N-TREOR in EXPO. *J. Appl. Cryst.*, 33:1180–1186, 2000.

[6] P.M. de Wolff. A simplified criterion for the reliability of a powder pattern indexing. *J. Appl. Cryst.*, 1:108–113, 1968.

[7] B.M. Kariuki, S.A. Belmonte, M. I. McMahon, R.L. Johnston, K.D.M. Harris, and R.J. Nelmes. A new approach for indezing powder diffraction data based on whole-profile fitting and global optimization using a genetic algorithm. *J. Synchrotron Rad.*, 6:87–92, 1999.

[8] J.A. Hageman, R. Wehrens, R. de Gelder, W.L. Meerts, and L.M.C. Buydens. Fitting fluorescence spectra with genetic algorithms. In *Parallel*

Problem Solving from Nature VI, pages 702–711. Springer Verlag, 2000. Lecture Notes in Computer Science.

[9] J.A. Hageman, R. Wehrens, R. de Gelder, W.L. Meerts, and L.M.C. Buydens. Direct determination of rotational constants from rovibronic spectra with genetic algorithms. *J. Phys. Chem.*, 113(18):7955–7962, 2000.

[10] R. de Gelder, R. Wehrens, and J.A. Hageman. A generalized expression for the similarity of spectra: application to powder diffraction pattern classification. *J. Comput. Chem.*, 22(3):273–289, 2001.

[11] C. Hammond. *The basics of Crystallography and Diffraction*. Oxford University Press, 1997.

[12] J.H. Holland. *Adaptation in naturial and artificial systems*. MIT Press, 1992.

[13] D.E. Goldberg. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley, 1989.

[14] R. Wehrens and L.M.C. Buydens. Evolutionary optimisation: a tutorial. *Trends in Analytical Chemistry*, 17(4):193–203, 1998.

[15] K.D.M. Harris, R.J. Johnston, and B.M. Kariuki. *Evolutionary Algorithms in Molecular Design*, volume 8. Wiley-VHC, 2000.

[16] D. Levine. PGAPack V1.0. PGAPack can be obtained from anonymous ftp from: ftp://ftp.mcs.anl.gov/pub/pgapack/pgapack.tar.Z.

[17] R. Shirley. The CRYSFIRE system for automatic powder indexing: user's manual. The Lattice Press, 41 Guildford Park Avenue, Guildford, Surrey GU2 5NL, England.

[18] N. Masciocchi, M. Moret, P. Cairati, A. Sironi, G.A. Ardizzoia, and G. La Monica. *J. Am. Chem. Soc.*, 116:7668–7676, 1994.

[19] A. Jouanneaux, A.N. Fitch, and J.K. Cockroft. *Mol. Phys.*, 71:45–50, 1992.

[20] R.J. Cernik, A.K. Cheetham, C.K. Prout, D.J. Watkin, A.P. Wilkinson, and B.T.M. Willis. *J. Appl. Cryst.*, 24:222–226, 1991.

[21] P. Norby, A. N. Christensen, H. Fjellvåg, M.S. Lehmann, and M. Nielsen. *J. Solid State Chem.*, 94:281–293, 1991.

[22] A.N. Christensen. *Z. Kristallogr.*, 209:7–13, 1994.

[23] J.J. Zah-Letho, A. Jouanneaux, A.N. Fitch, A. Verbaere, and M. Tournoux. *Eur. J. Solid State Inorg. Chem.*, 29:1309–1320, 1992.

[24] A. Jouanneaux, A. Verbaere, Y. Piffard, A.N. Fitch, and M. Kinoshita. *Eur. J. Solid State Inorg. Chem.*, 28:683–699, 1991.

[25] A.N. Christensen, R.G. Hazell, A.W. Hewat, and K.P.J. O'Reilly. *Acta Chem. Scand.*, 45:469–473, 1991.

[26] A. Jouanneaux, A. Verbaere, D. Guyomard, Y. Piffard, S. Oyetola, and A.N. Fitch. *Eur. J. Solid State Inorg. Chem.*, 28:755–765, 1991.

[27] A.N. Christensen, M. Nielsen, K.P.J. O'Reilly, and T. Wroblewski. *Acta Chem. Scand.*, 46:224–230, 1992.

Chapter 5

Wavelength Selection with Tabu Search

This paper introduces Tabu Search in analytical Chemistry by applying it to wavelength selection. Tabu Search is a deterministic, global optimisation technique loosely based on concepts from artificial intelligence. Wavelength selection is a method which can be used for improving the quality of calibration models. Tabu Search uses basic, problem-specific operators to explore a search space, and memory to keep track of parts already visited. Several implementational aspects of wavelength selection with Tabu Search will be discussed. Two ways of memorising the search space are investigated: storing actual solutions or storing steps necessary to create them. Parameters associated with Tabu Search are configured with a Plackett-Burman design. In addition, two extension schemes for Tabu Search, intensification and diversification have been implemented and are applied with good results. Eventually, two implementations of wavelength selection with Tabu Search are tested, one which searches for a solution with a constant number of wavelengths and one with a variable number of wavelengths. Both implementations are compared with results obtained by wavelength selection methods based on Simulated Annealing (SA) and Genetic Algorithms (GA's). It is demonstrated with three real world data sets, that Tabu Search performs equally well and can be a valuable alternative to SA and GA's. The improvements in predictive abilities increased with a factor of 20 for data set 1 and with a factor of 2 for data set 2 and 3. In addi-

This chapter is publised as: J.A. Hageman, M. Streppel, R. Wehrens and L.M.C. Buydens, Wavelength Selection with Tabu Search, J. Chemometrics., 17:1-11, 2003

tion, when the number of wavelengths in a solution is variable, measurements on the coverage of the search space show that the coverage is usually higher for Tabu Search compared to SA and GA's.

5.1 Introduction

In analytical chemistry, Tabu search [1–4] is a relatively new technique with only a few published examples, dealing with graph theory [5] and molecular docking [6]. Tabu Search is an iterative deterministic global optimisation method. It examines the search space in an highly ordered fashion using memory to keep track of parts already visited. Given a starting solution, it will always come up with the same end solution. In this paper, Tabu Search will be introduced for wavelength selection. Wavelength selection is a much used procedure for improving the quality of calibration models for example. After wavelength selection, predictive abilities are usually higher and the models are simpler and more robust [7–10]. The easiest way for finding the optimal combination of wavelengths would be an exhaustive search. However, an exhaustive search for wavelength selection would require the examination of an astronomical number of combinations. As this is usually not feasible due to large computation times, other wavelength selection methods have been designed. Originally these methods used simple heuristics for locating a solution, but given the characteristics of the methods, these were likely not the best obtainable solution. With the recognition that wavelength selection is an optimisation problem and the increasing availability of faster computers, more sophisticated optimisation techniques such as Simulated Annealing (SA) [11–13] and Genetic Algorithms (GA's) [7, 10–12] have frequently been used. SA and GA's are both iterative probabilistic global optimisation methods. As a consequence, both methods do not always end up with the same end solution, given identical starting solutions.

Several implementational aspects of wavelength selection with Tabu Search will be discussed. As memorising the search space is an important characteristic of Tabu Search, two possible ways for memorising are investigated. In addition, it will be shown that configuring the parameters associated with Tabu Search can be done with an experimental design. To further improve results, two extension schemes for Tabu Search for applying wavelength selection have been implemented: intensification and diversification. It will be demonstrated that both are valuable assets. Two implementations of Tabu Search

have been made, one that searches for solutions with a constant number of wavelengths and one with a variable number of wavelengths. Results of Tabu Search are applied to three real world data sets and are compared with results obtained by SA- and GA-based methods. Results indicate that Tabu Search works equally well and is a valuable alternative to SA and GA's. In addition, when the number of wavelengths in a solution is variable, measurements on the coverage of the search space show that the coverage is usually higher for Tabu Search compared to SA and GA's.

5.2 Theory

Where SA is based on the physical process of cooling down a heated liquid and GA's are inspired by the process of evolution, Tabu Search is based on concepts from artificial intelligence [4]. It uses basic, problem-specific operators to explore a search space and memory (which is called the tabu list) to keep track of parts already visited. By guiding the optimisation to areas not present in memory, Tabu Search hopes to find the global optimum. The foundations for Tabu Search were laid out in the late 1970's by Glover and the principles were described in general terms in 1989 and 1990 also by Glover [1–3]. In recent years, tutorials documenting successes accomplished with Tabu Search have been published [3, 4, 14].

5.2.1 Tabu Search

Tabu Search is an optimisation technique which tries to optimise a function $G(x)$ where x represents a parameter vector, by iteratively searching the parameter space of x for the optimal solution. The framework of Tabu Search consists of several steps which are described below and depicted in Figure 5.1.

1. Initialisation: a starting solution s is generated by choosing random values for x. This solution is evaluated by the evaluation function and solution s is stored in the algorithm's memory. This memory is called the tabu list.

2. Neighbourhood exploration: all possible neighbours of solution s are generated and evaluated. Neighbouring solutions are solutions which can be reached from the current solution by a simple, basic transformation of the current solution. Solutions which are present in the tabu list are considered unreachable neighbours.

```
choose an initial solution s in x
s* = s;      % best solution so far
k = 0;       % iteration number
kbest = 0;   % last improving iteration
kmax = ...; % maximum non-improving iterations
T = [];      % Tabu list
stop = false
while not stop
    generate V* ⊆ N(s) - T
    if ((k - kbest > kmax) or (V* == []))
        stop = true;
    else
        k = k + 1;
        choose best s' in V*
        s = s';
        if (f(s) < f(s*))
            s* = s;
            kbest = k;
        end
        update tabu list
    end
end
```

Table 5.1: Pseudo-code of the Tabu Search algorithm.

3. New current solution: a new current solution is chosen from the explored neighbourhood. This solution cannot be in the tabu list and it has to have the best evaluation value from all reachable neighbours. The evaluation value can be worse compared to the current solution. In this way the algorithm is able to overcome local minima. The new current solution is added to the tabu list.

4. Stop: If no more neighbours are present (all are tabu) or a certain evaluation value or a predetermined number of iterations is reached, the algorithm stops, otherwise the algorithm continues with step 2.

In Table 5.1 pseudo-code of the Tabu Search Algorithm is given.

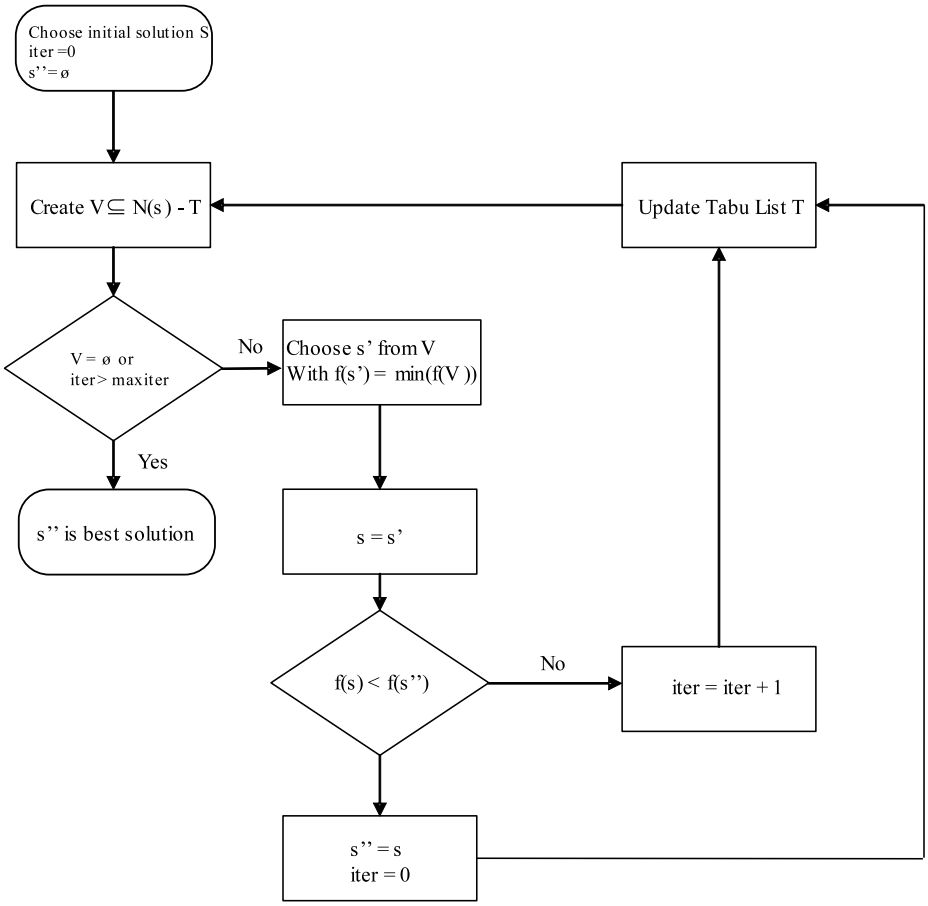

Figure 5.1: General flowchart of the Tabu Search algorithm. Iter keeps track of the number of iterations without an improvement, s is the current solution, s' is the solution with the lowest evaluation value of all neighbours of s, s" is the best obtained solution, V are all neighbours of solution s, T is the tabu list, maxiter is the allowable maximum number of iterations without an improvement.

5.2.2 Neighbourhood exploration of wavelength selection

The neighbourhood of a solution is defined by those solutions which can be reached in one step. A solution reachable in one step is called a neighbour. These steps are specific for each optimisation problem. Usually, two neighbouring solutions resemble each other closely. In the case of wavelength selection, a solution is a combination of a number of wavelengths and changing this combination can be accomplished by 3 different types of steps:

1. Selecting or adding a number of wavelengths
2. Deselecting or removing a number of wavelengths
3. Moving a number of wavelengths

Selecting and deselecting wavelengths are operators which are used in classical methods for wavelength selection. The move operator can be seen as a combination of both, it first deselects a wavelength and subsequently selects a different one. The number of possible neighbours of a solution increases drastically when more than one operator is allowed per step. Equations 5.1-5.3 give the maximum number of unique neighbouring solutions by using the operators select, deselect and move respectively.

$$\texttt{\#select} = \sum_{i=1}^{m_s} \binom{t-n}{i} \tag{5.1}$$

$$\texttt{\#deselect} = \sum_{i=1}^{m_d} \binom{n}{i} \tag{5.2}$$

$$\texttt{\#move} = \sum_{i=1}^{m_m} \binom{t-n}{i}\binom{n}{i} \tag{5.3}$$

where m_s, m_d and m_m are the maximum number of wavelengths considered for selection, deselection and move, respectively, n is the number of selected wavelengths in s and t is the total number of wavelengths in the spectrum. In a data set with 150 wavelengths of which 50 wavelengths are selected the total number of neighbouring solutions is 6,075,075 when two select, deselect or move steps are allowed. When only one select, deselect or move step is allowed, there are only 5150 neighbouring solutions. Since a NIR-spectrum can contain several hundred wavelengths, it is not feasible to allow more than

one select, deselect or move step per iteration ($m_s = 1$, $m_d = 1$, $m_m = 1$). The total number of neighbours per iteration is then given by

$$\begin{aligned}\binom{t-n}{1} + \binom{n}{1} + \binom{t-n}{1}\binom{n}{1} \\ = t - n + n + (t-n) \cdot n \\ = t + t \cdot n - n^2\end{aligned} \tag{5.4}$$

The number of neighbours which have to be evaluated is at a maximum when $n = t/2$. This maximum is $t + t^2/4$. To further reduce the number of neighbouring solutions a restriction is placed on the maximum distance of a move. When a wavelength is moved, it is only allowed to shift a distance d to left or right. This causes Tabu Search to search the solution space more structured, because the difference between two solutions is smaller when a wavelength is moved over a short distance since the intensities at wavelengths close to each other tend to be correlated. The total number of neighbours that should be evaluated is now given by $t + 2dn^2$, where d is half the distance allowed for the move.

Another possibility is to only allow the move operator (Eq. 5.3) and not the select or deselect operator (Eq. 5.1 and 5.2). A consequence is that the number of selected wavelengths can be predetermined and will be kept constant during the optimisation. When the distance of a move is restricted, the number of possible neighbours which needs to be evaluated is $2dn^2$.

5.2.3 Tabu list

In Tabu Search, the tabu list plays an important role. It keeps track of previous explored solutions and prohibits Tabu Search from revisiting them again. In this way Tabu Search can overcome local minima by forcing it to accept solutions worse than the current solution. The tabu list has a finite length l. After l iterations the first tabu restriction is removed and this first solution becomes available again for selection. Keeping the tabu list too short may result in visiting the same sequence of solutions over and over again. The algorithm then ends up in a cycle and will not be able to locate better solutions. A list too long may cause unnecessary long run times but also may prevent the algorithm from reaching an optimal solution. In general, a list with length l will prevent cycles with length l. In the list, the actual solution can be stored (called explicit memory). In this case, the tabu list

Table 5.2: Overview of operators on wavelength p and q and the subsequent tabu restrictions in recency based Tabu Search.

Operator	**Tabu list**
select p	deselect p
deselect p	select p
move q to p	select q
	deselect p

contains the actual combination of wavelength indices that are selected. It is also possible to store the steps necessary to generate the new solution (called recency-based memory). In the case when a wavelength is deselected, the selection of this wavelength becomes tabu, to prevent the algorithm to retract to the original solution. Table 5.2 shows the tabu restrictions that take place after the execution of all three operators. If steps are stored, all operators have their own tabu list. The tabu list is referred to as short-term memory when it contains previously visited solutions. It deals with the most recent history of the search trajectory.

5.2.4 Intensification and diversification

Two extension schemes are common for Tabu Search: intensification and diversification. Both schemes are referred to as a form of long term memory as they use information not available in short term memory. Intensification focuses on the part of the solution space which seems promising and has often been visited with good results. As an intensification approach, the best solutions obtained after several runs with different starting solutions are stored. Wavelengths which contribute to a good model will likely be selected more often in best found solutions. So for a next run, a new starting solution is generated consisting of the wavelengths which where selected in 30 and 60% of the best results of these previous runs. This provides Tabu Search with a promising starting solution.

Diversification is the opposite of intensification. It guides Tabu Search towards unexplored parts of the search space. In this way, the solution space will be covered more thoroughly and the chance of missing the optimal solution will be reduced. To be able to guide Tabu Search to an unexplored part of

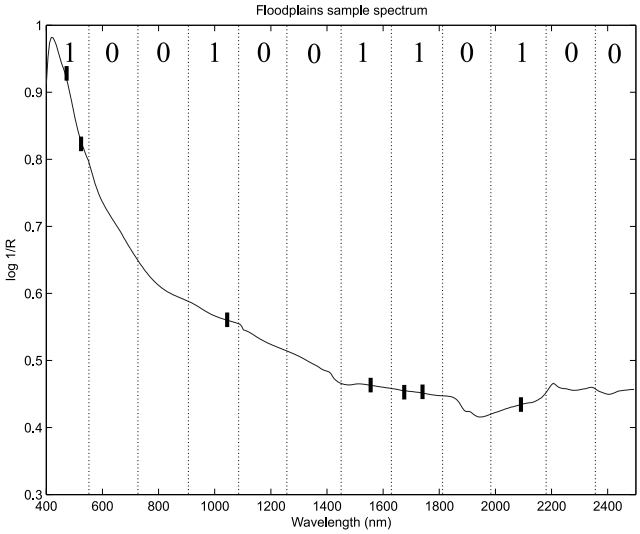

Figure 5.2: Example how a combination of selected wavelengths are transformed into a bitstring. The bitstring indicates which region of the solution space has been examined by that solution.

the solution space it is necessary to keep track of the parts of the solution space which Tabu Search has explored during its search. To accomplish this, each spectrum is divided into several bins, 12 in our case. When one or more wavelengths in a bin are selected, this bin is represented by a one. If no wavelengths in a bin are selected, it will be represented by a zero. All 12 ones and zeros put together form a bitstring. Each bitstring can be seen as a point in the (simplified) solution space. By keeping track of which bitstrings have been visited, the regions which have not been explored can be identified. Figure 5.2 shows how a combination of selected wavelengths is transformed into a bitstring.

After the end of a run it is checked which regions have not been visited. A solution in the region which is the furthest from the visited regions is taken as the initial solution for a new run. By comparing the number of visited regions, it is also possible to measure the coverage of the search space. The higher this number, the better the algorithm has been able to scan different regions of the search space and thus should have been able to locate a good solution.

The distance between regions is calculated using the Hamming distance

[15]. The Hamming distance between two regions j and k is given by

$$d_{j,k} = \sum_{i=1}^{N} \delta_{j,k}(i) \tag{5.5}$$

where $\delta_{j,k}(i) = 1$ when the ith bin of regions j and k do not contain the same value and $\delta_{j,k}(i) = 0$ when they contain the same value, N is the number of bins.

5.2.5 Evaluation function

The goal of wavelength selection is to find a set of wavelengths for the creation of an optimal predictive model. The prediction model used is Partial Least Squares regression (PLS) [16] and in particular SIMPLS [17]. A problem with PLS is that the number of latent variables should be specified. As the information in each set of wavelengths is not the same, this number cannot be kept constant but has to be determined again for each solution. For determining the optimal number of latent variables and to prevent overfitting, data sets are divided into two parts, a training set and a test set.

The correct number of latent variables for each subset is determined by leave-p-out cross validation on the training set, with p being a number which divides each data set in roughly 15-20 groups. Outliers should be removed from any used data set as these will negatively influence the predictive abilities of the obtained models. Crossvalidation results in an array of predictive abilities for each number of latent variables. The correct number of latent variables is obtained by comparing two consecutive values. When the next value still increases the predictive ability with more than 10%, the number of latent variables is increased by one. The complete trainingset in combination with the correct number of latent variables is used to calculate the PLS coefficients. The test set in combination with the PLS coefficients is used to calculate the RMSEP (Eq. 5.6) value:

$$\texttt{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n} (y_{pred,i} - y_i)^2}{n}} \tag{5.6}$$

where $y_{pred,i}$ and y_i are respectively the predicted and the measured response values and n is the number of samples. This value is used as an evaluation value in Tabu Search. The complete procedure for the evaluation function is shown in Figure 5.3.

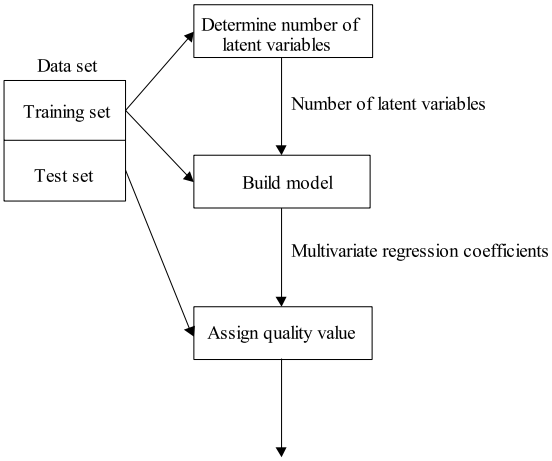

Figure 5.3: Flow-chart of the evaluation function.

5.3 Experimental

5.3.1 Data sets

Three near infrared spectral data sets are used to test and compare the performance of Tabu Search for wavelength selection. All three data sets are assumed to be free of outliers.

1. Gasoline data set [18]: NIR spectra of gasoline samples with measured octane numbers. Samples are measured from 900 to 1700 nm with a sampling rate of 2 nm. The first 200 nm have been omitted as they show no signal. This results in a total of 300 wavelengths. The training set consists of 40 spectra and the test set of 20 spectra.

2. Wheat data set [18]: NIR spectra of wheat samples. Two responses have been measured which are the moist and protein content. The spectra have been recorded from 1100 to 2500 nm with a 2 nm interval. Every two wavelengths have been averaged to reduce the number of wavelengths which resulted in 350 wavelengths. The training set consists of 67 spectra and the test set of 33 spectra.

3. Floodplains data set [19]: 67 NIR spectra of floodplains with four measured response values which are the Cd, Zn, clay and organic matter content. The spectra have been measured from 400 to 2500 nm with a

sampling rate of 2 nm. To reduce the number of wavelengths every three wavelengths have been averaged, which resulted in 350 wavelengths. The training set consists of 54 spectra and the test set of 13 spectra.

5.3.2 Tabu Search configuration

The Tabu Search algorithm for wavelength selection has been implemented as explained in the previous section. To be able to decide whether to store the actual solutions or the steps necessary to create them in the tabu list, both methods are implemented. For an optimal performance, the parameters associated with Tabu Search need to be optimised. For both cases these parameters are the locality of the move operator (parameter d), the maximum number of iterations and the number of selected wavelengths in the initial solution. When storing solutions, the length of the tabu list is kept long enough to obtain all encountered solutions. When storing steps, four extra parameters need to be optimised: the length of the tabu list for the *select*, *deselect*, *move from* and *move to* operators. For selecting optimal values for these parameters, a Placket-Burman experimental design [20] was used in combination with the gasoline data set and leave one out crossvalidation. The high and low values used in the experimental design are shown in Table 5.3. All experiments in the Plackett-Burman design were executed in five fold, each time with a different starting solution, to cancel out random effects. The response variable in the experimental design was the RMSEP. After configuring Tabu Search while storing steps or storing solutions, it was decided which storing method would be used by running both algorithms with five different starting seed and comparing the results.

Eventually, two implementations of Tabu Search have been made: one using a variable number of wavelengths and one using a constant number of wavelengths, only allowing the move operator. The configuration of the parameters associated with Tabu Search in combination with a constant number of wavelengths was based on the results of the experimental design, but were modified slightly after additional experiments. Both Tabu Search implementations have been used to select optimal combinations of wavelengths of the three different data sets. When the number of wavelengths was constant, solutions consisting of 15, 30 and 45 wavelengths were optimised. All runs were repeated five times, to exclude any random effect of a starting solution.

Table 5.3: The values used for the high and low levels for the variables in the Plackett-Burman design.

Variable	High Level	Low Level
Locality	45	10
Maximum iteration	65	30
Number initially selected	20	5
Length select	50	10
Length deselect	50	10
Length move from	50	10
Length move to	50	10

5.3.3 Comparison with other wavelength selection methods

The performance of Tabu Search for wavelength selection is compared with the results of 4 other methods. Two of these are simple heuristic methods: forward selection and backward elimination [21]. The other two methods are SA and GA's. As it is possible for Tabu Search to search with a variable number of wavelengths as well as a constant one, both ways have also been incorporated into SA and GA's. The implementation for the GA approach with a variable number of wavelengths is based on Refs. [7, 11, 12] with one exception, which is the mutation operator. The mutation operator used in this paper has a 90% chance for selecting a zero and a 10% chance for selecting a one. This ratio ensures that not too many wavelengths will be selected, as this is disadvantageous for a good predictive model. When the number of wavelengths must be kept constant, a different approach is used. Instead of a bitstring representation, an integer array representation is used. This array contains the indices of selected wavelengths. It is made sure that a wavelength index can be selected only once. The implementation of wavelength selection with SA with a constant number of wavelengths is based on Refs. [11, 12]. When allowing a variable number of wavelengths, some additions have been made. Besides moving selected wavelengths in the step generating function, it is also possible to add or remove wavelengths. The step generating function of SA can apply the same operators as in Tabu Search. The optimal settings for the SA and GA's based methods were determined by trial and error and are shown in Table 5.4 and in Table 5.5.

Table 5.4: The GA settings used for all data sets for implementations with a variable and constant number of wavelengths.

GA-method	Variable number of wavelengths	Constant number of wavelengths
Number of generations	800	800
Population size	300	300
Elitism	150	150
Crossover type	Uniform crossover	1-point crossover
Crossover probability	0.7	0.7
Mutation type	90/10% Bit flip	Uniform random replacement
Mutation probability	0.05	0.05
Selection type	Proportional selection	Proportional selection
Fitness type	Linear normalisation fitness	Linear normalisation fitness

Table 5.5: Settings for the SA method used for all data sets for implementations with a variable and constant number of wavelengths.

SA-method	Variable number of wavelengths	Constant number of wavelengths
Starting temperature	0.01	0.01
Cooling constant	0.999	0.999
Maximum Markov Chain Length	300	300
Minimum Markov Chain	150	150
Minimum temperature	0.001	0.001
Minimum number of constant function values	300	300
Chance being moved[a]	0.1	0.1
Locality of move[b]	10	10
Chance being added[a]	0.01	-
Chance being deselected[a]	0.1	-

[a]Only one step of all possible three will be executed for each wavelength, but with different chances.

[b]Locality indicates the maximum distance a wavelength can be moved.

5.3.4 Software

All software was programmed in ANSI-C. The Tabu Search and SA methods were programmed from scratch. For the GA's approach PGAPack [22] was used as a basis. The SVD routine used in the SIMPLS algorithm was adopted from Ref. [23]. Calculations were performed on a Sun-Ultra 10 running at 440 MHz. Runtimes were dependant on the size of the data set and the number of properties that needed to be predicted. For the gasoline data set, runtimes were in the order of two hours for the SA, GA and Tabu Search based methods for one run.

5.4 Results and Discussion

5.4.1 Tabu Search Configuration

The main effects of the Plackett-Burman experimental design are shown in Figure 5.4 for storing solutions (part A) and storing steps (part B) in the tabu list. In both cases a larger number of non improving iterations is beneficial for obtaining a good solution, which is logical because this prolongs the searching time. The number was set to 65. A high number of initial wavelengths is also beneficial. A high number increases the chance of selecting good contributing wavelengths from the start. However, experiments which were performed to choose the high and low limits for the experimental design, demonstrated that when this number is too high, Tabu Search spends a lot of time getting rid of non-contributing wavelengths, which in the end leads to a decreased predictive power. This number was set to 20. The effect of the locality between both approaches differs. When storing solutions, increasing the locality means that a solution can be refined faster. When steps are being stored, a moved wavelength cannot be removed for a certain number of iterations. In that time, other wavelengths are changed and Tabu Search crosses that (local) optimum without ever locating the exact minimum. This is also reflected in the length of the tabu lists of the different steps. When tabu lists are too long, certain wavelengths are not available anymore and this forces Tabu Search to different areas in the search space. One exception is the *move from* length, this list prohibits wavelengths from being reselected after being moved. When this list is too short, it will lead to cycles. Lengths for the *select*, *deselect*, *move from* and *move to* were set to 10, 10, 50 and 10 respectively.

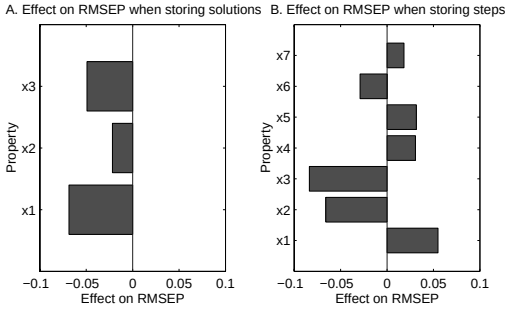

Figure 5.4: Results of the Plackett-Burman experimental design for determining an optimal configuration of Tabu Search. Part A shows results when storing complete solutions. Part B shows results when only storing steps. (x1: locality of the move operator, x2: maximum number of iterations, x3: number of wavelengths in initial solution, x4: length *select* tabu list, x5: length *deselect* tabu list, x6: length *move from* tabu list, x7: length *move to* tabu list)

Based on the results of the Plackett-Burman design, optimal configurations were determined for both approaches, as shown in Table 5.6. Analysis of the five replicate runs with different starting solutions with both approaches showed that there was no significant difference between the means of all runs. However, storing steps yielded the best solution with the lowest RMSEP value. A comparison of selected wavelengths in the best solution between all replicate runs also showed that the reproducibility is higher. When steps are being stored it appears that the algorithm is much more able to select the same wavelengths during different runs. When the selection, deselection or moving of wavelengths is made tabu, the number of neighbours decreases, which is beneficial for running times. Therefore, storing steps in the tabu list has more advantages and will be used for the remainder of the work in this paper. The best RMSEP value for each of the five runs is given in Table 5.7.

Applying intensification resulted in an improvement of the RMSEP values. When the new starting solution consisted of wavelengths which were present in 30% of the best found solutions, the new RMSEP value was 6.99e-2, when wavelengths were chosen which were present in 60% of the final solutions, the improvement was even greater: 6.59e-2. It is very likely that wavelengths which are selected more often in best found solutions contribute more to a good predictive model. A starting solution based on these wavelengths enables Tabu Search to come up with an improved predictive model.

Table 5.6: Configuration of Tabu Search when storing actual solutions or steps leading to solutions. WL constant, indicates the settings when using a constant number of wavelengths

Tabu Search Approach	Solutions	Steps	WL constant
Locality	45	10	30
Maximum iterations	65	65	65
Number initially selected	20	20	15
Length select TL		10	
Length deselect TL		10	
Length move from TL		50	5
Length move to TL		10	20

Table 5.7: Results of Tabu Search while storing steps in the tabu list and testing different additions. Results are obtained with the gasoline data set and leave one out crossvalidation.

	Run 1	Run 2	Run 3	Run 4	Run 5
Basic Tabu Search	7.88e-2	1.05e-1	1.25e-1	7.00e-2	7.09e-2
Intensification[a]	30%: 6.99e-2		60%: 6.59e-2		
Diversification[b]	7.69e-2	4.56e-2	4.56e-2	9.28e-2	6.68e-2
Both combined	1.30e-1	8.01e-2	1.02e-1	1.27e-1	1.14e-1

[a]Intensification is performed on the combined results of all previous five replicate runs. 30% and 60% indicate the occurrences of selected wavelengths in the best found solutions.

[b]After each initial run five diversification runs have been performed, the best of these results is shown.

Diversification is also able to improve RMSEP values. Results are shown in Table 5.7. After each replicate run, five diversification rounds were used. In four out of five, diversification yielded RMSEP values which were lower. The best RMSEP value after diversification is 4.56e-2.

Intensification and diversification have also been combined. Since wavelengths which are important for a good solution are likely to be in every best solution, diversification is performed before intensification. By first applying diversification, a large part of the solution space will be covered. After the initial best solutions have been found, intensification is used to zoom in on interesting wavelengths and perhaps locate a better solution. Results of the combination are given in Table 5.7. Diversification appears to be highly effective and renders intensification superfluous, because intensification does not lead to an improvement in all five cases. Therefore, only diversification is used.

5.4.2 Comparison with other methods

Table 5.8 show the results for all runs with all three data sets, including the runs performed with the SA and GA based methods. Table 5.9 shows the number of regions each method has examined. It can be seen that all tested wavelength selection methods are able to increase the predictive abilities of PLS models. It can also be seen that the simple heuristic methods backward elimination and forward selection perform poorly. The resulting RMSEP values are higher and the coverage of the search space is also less.

Implementations of GA, SA and Tabu Search with a variable number of wavelengths show comparable results with respect to RMSEP values, especially for the wheat and floodplains data set. The results for the gasoline data set show some variations. Comparing results between all three methods is somewhat complicated. Not all methods come up with the same number of selected wavelengths and latent variables. Tabu Search selects the most wavelengths, but uses a smaller number of latent variables. This might indicate a suboptimal solution. When a smaller number of latent variables is preferable, Tabu Search has a slight advantage, if absolute predictive values are important SA scores better.

When applying Tabu Search, the coverage of the search space is high for the gasoline and wheat data set but the coverage of the floodplains data set is lower than for SA and GA's based methods. This is probably due to the large number of selected wavelengths, which makes it difficult to move from one region to another when only changing one wavelength. Coverage for backward

Table 5.8: The results of the different wavelength selection methods. WL indicates the number of wavelengths present in the solution. LV indicate the number of latent variables.

Method	Gasoline			Wheat			Floodplains		
	RMSEP	WL	LV	RMSEP	WL	LV	RMSEP	WL	LV
All wavelengths	6.88e-1[a]	301	4	7.84e-1	350	7	2.18	350	5
Stochastic methods									
Backward elimination	2.21e-1	181	8	6.14e-1	161	7	1.60	77	3
Forward selection	1.54e-1	29	7	4.72e-1	13	10	1.59	42	3
Implementations with a variable number of wavelengths									
GA	4.54e-2	30	6	3.37e-1	35	11	1.54	8	3
SA	3.33e-2	33	7	3.15e-1	21	12	1.55	6	3
Tabu Search	5.01e-2	38	5	3.19e-1	49	11	1.56	39	3
Implementations with a constant number of wavelengths									
GA	7.15e-2	15	4	3.43e-1	15	9	1.48	15	4
SA	5.82e-2	15	5	3.33e-1	15	10	1.56	15	3
Tabu Search	5.70e-2	15	6	3.36e-1	15	12	1.55	15	3
GA	6.00e-2	30	5	3.33e-1	30	9	1.49	30	4
SA	4.06e-2	30	7	3.34e-1	30	10	1.63	30	3
Tabu Search	6.59e-2	30	7	3.41e-1	30	10	1.57	30	3
GA	7.59e-2	45	8	3.50e-1	45	9	1.49	45	4
SA	1.11e-1	45	7	3.47e-1	45	10	1.63	45	3
Tabu Search	6.03e-2	45	7	3.42e-1	45	11	1.58	45	3

[a]The number of latent variables has been chosen visually as automated selection resulted in the selection of one latent variable.

elimination is very low because with all wavelengths selected in the initial solution, it is difficult to deselect all wavelengths in a bin, which is necessary for changing the bitstring.

The implementations with a constant number of wavelengths show a somewhat different trend. Tabu Search is able to build models which usually have slightly higher predictive abilities, especially for the gasoline and wheat data sets. Again, the largest deviations can be found with the gasoline data set, whereas the results obtained with the wheat and the floodplains data set are more homogeneous. For all three methods, results deteriorate somewhat when solutions are forced to contain more wavelengths. The coverage of the search space shows some trends when keeping the number of wavelengths constant in a solution. GA's usually have the higher coverage, followed by SA and Tabu Search. This reflects the degree of randomness is each method. Where GA's are able to take large steps in the search space, this ability is somewhat reduced in SA and in Tabu Search this is highly structured. Nevertheless, so-

Table 5.9: Coverage of the solution space by the different methods.

Method	Gasoline Regions	Wheat Regions	Floodplains Regions
Stochastic methods			
Backward elimination	1	1	1
Forward selection	78	63	68
Implementations with a variable number of wavelengths			
GA	202	161	873
SA	226	219	656
Tabu Search	334	362	413
Implementations with a constant number of wavelengths[a]			
GA(15)	702	725	743
SA(15)	301	440	876
Tabu Search(15)	394	279	376
GA(30)	141	207	356
SA(30)	106	125	196
Tabu Search(30)	48	64	142
GA(45)	26	57	248
SA(45)	40	59	71
Tabu Search(45)	12	22	72

[a]Number between brackets indicate which number of wavelengths in a solution is kept constant.

lutions obtained with Tabu Search perform equally well and sometimes better as obtained with GA's and SA.

Figure 5.5 shows the selected wavelengths in the best solutions obtained with the SA, GA's and Tabu Search based methods for each data set. When looking at wavelengths which are selected by the different methods in the best solutions, there is great overlap. Wavelengths selected in the gasoline data set can be found more or less throughout the spectrum, but for the wheat and even more for the floodplains data set, specific regions of selected wavelengths can be identified. In these cases, wavelengths from specific regions contain the most valuable information. For the wheat and floodplains data set, differences in the positions of selected wavelengths in the best solutions between the replicate runs were small. For the gasoline data set these differences were slightly larger. As wavelengths near each other are often highly correlated, small differences have only small effects on the predictive ability of models. Of all three optimisation methods, replicate runs with different starting solutions, performed with GA's and Tabu Search have a higher reproducibility. An

example of this is shown in Figure 5.6. In contrast to GA's and Tabu Search, it would appear that SA is easier caught in a local optimum.

The gasoline and wheat data set have also originally been used for demonstrating the possibilities of wavelength selection with GA's [10]. When comparing the results from [10] and our findings, selected wavelengths and RMSEP values differ at some points. These differences are likely to be caused by the differences in evaluation functions and the number of latent variables which may have been used. In this paper, for each solution, a different number of latent variables is calculated with leave-p-out crossvalidation, while Ref. [10] used a constant number.

In this paper, data sets have not been preprocessed before wavelength selection and PLS modelling. It is very likely that spectral preprocessing would increase the predictive abilities of the models. It is also very likely that the beneficial effects of preprocessing will be the same for the different types of wavelength selection techniques, so this has been left out in this work. To obtain the best possible predictive models, spectral preprocessing can be recommended.

5.5 Conclusion

This paper shows the potential and an implementation for wavelength selection with Tabu Search. Tabu Search is a relatively new method in analytical chemistry, and in contrast to SA and GA's it is not probabilistic but deterministic. As a consequence it will always come up with the same best solution, if it is provided with the same starting solution.

Wavelength selection is a much used procedure for easily increasing the predictive ability of models. Even simple heuristic methods are able to improve models. However, better improvements are obtained by using more sophisticated methods like SA, GA's and Tabu Search. It is demonstrated that the implementation described here, yields results as good as those obtained by other well established methods like SA and GA's. Configuring the parameters of Tabu Search, or meta-optimising, is no difficult task. It can be done with an experimental design, or if some experience is present with trial and error. The intensification and diversification approaches applied in this paper are valuable extensions of Tabu Search.

It is possible to use Tabu Search for locating solutions with a variable and constant number of wavelengths. In general, results are better when the

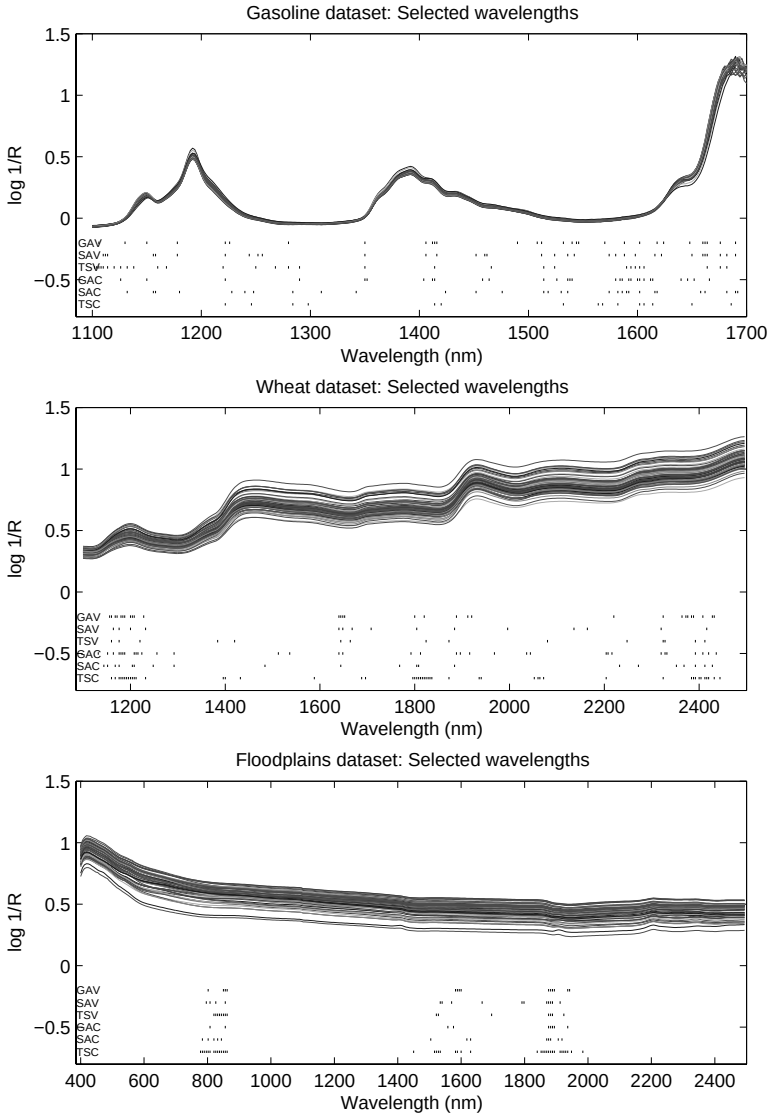

Figure 5.5: Selected wavelengths in the best solutions obtained with wavelength selection methods based on SA, GA's and Tabu Search. The labels before the solutions indicate which method was used: GA's, SA or TS (Tabu Search), V or C indicate variable or constant number of wavelengths. From the solutions with a different constant number of wavelengths (15,30 or 45) only the one with the best RMSEP values is shown.

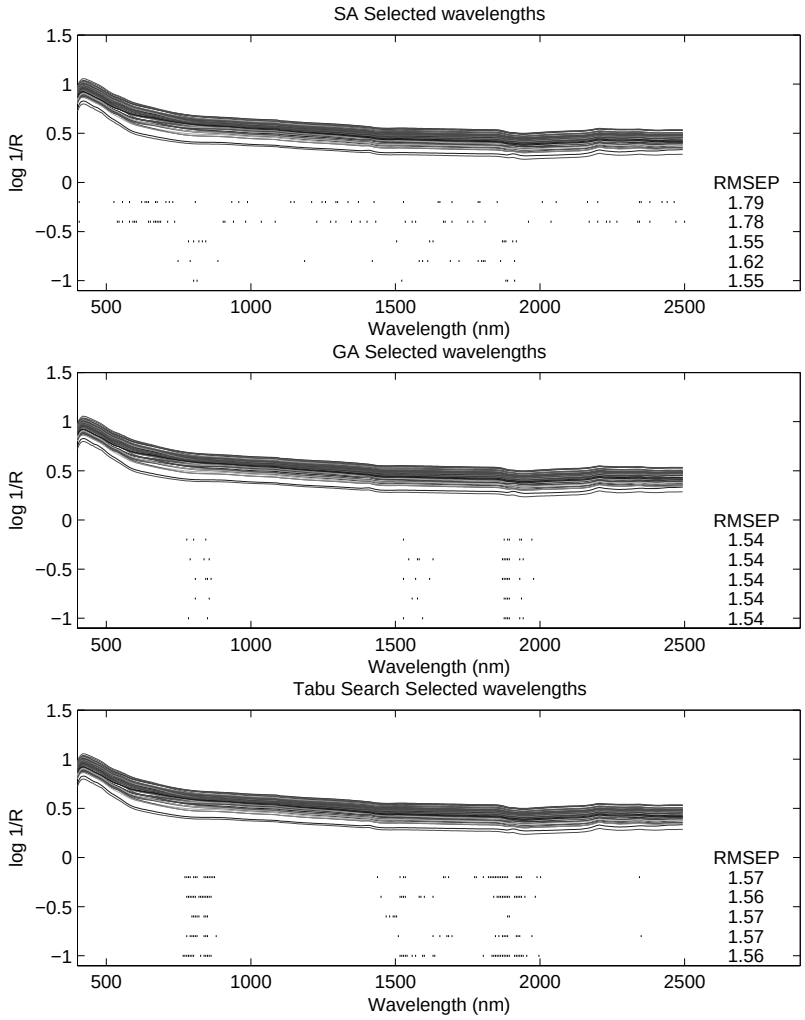

Figure 5.6: Results of five replicate runs, with different starting solution. Wavelength selection is performed on the floodplains data set with SA, GA's and Tabu Search with a variable number of wavelengths.

number of wavelengths is variable. Using Tabu Search with a variable number of wavelengths, the coverage of the search space is usually better compared with GA's and SA. This can become important when more local minima exist and it becomes harder to avoid getting trapped in them.

Tabu Search is a valuable alternative to SA and GA's, especially in cases where there is a clear definition possible of a neighbourhood of a solution.

References

[1] F. Glover. Tabu search – part I. *ORSA Journal on Computing*, 1(3):190, 1989.

[2] F. Glover. Tabu search – part II. *ORSA Journal on Computing*, 2(1):4, 1990.

[3] F. Glover, E. Taillard, and D. de Werra. A user's guide to tabu search. *Annals of Operations Research*, 41:3–28, 1993.

[4] F. Glover and M. Laguna. *Tabu Search*. Kluwer Academic Publishers, Dordrecht, 1998.

[5] V. Kvasnicka and J. Pospichal. Fast evaluation of chemical distance by tabu search algorithm. *J. Chem. Inf. Comput. Sci.*, 34:1109–1112, 1994.

[6] D.R. Westhead, D.E. Clark, and C.W. Murray. A comparison of heuristic search algorithms for molecular docking. *J. Comput.-Aided Mol. Des.*, 11:209–228, 1997.

[7] R. Leardi, R. Boggua, and M. Terrile. Genetic algorithms as strategy for feature selection. *Journal of Chemometrics*, 6:267–281, 1992.

[8] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B. Vandenginste, and C. Sterna. Elimination of uninformative variables for multivariate calibration. *Analytical Chemistry*, 68:3851–3858, 1996.

[9] D. Jouan-Rimbaud, B. Walczak, R.J. Poppi, O.E. de Noord, and D.L. Massart. Application of wavelet transform to extract the relevant component from spectral data for multivariate calibration. *Analytical Chemistry*, 69:4317–4323, 1997.

[10] R. Leardi. Application of genetic algorithm-PLS for feature selection in spectral data sets. *Journal of Chemometrics*, 14:643–655, 2000.

[11] C.B. Lucasius, M.L.M. Beckers, and G. Kateman. Genetic algorithms in wavelength selection: a comparative study. *Anal. Chim. Acta*, 286:135–153, 1994.

[12] U. Horchner and J.H. Kalivas. Further investigation on a comparative study of simulated annealing and genetic algorithm for wavelength selection. *Anal. Chim. Acta*, 311:1–13, 1995.

[13] U. Horchner and J.H. Kalivas. Simulated-annealing-based optimization algorithms: fundamentals and wavelength selection applications. *Journal of Chemometrics*, 9:283–308, 1995.

[14] D. de Werra and A. Hertz. Tabu search techniques, a tutorial and an application to neural networks. *OR Spektrum*, 11:131–141, 1989.

[15] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, and J. Smeyers-Verbeke. *Handbook of Chemometrics and Qualimetrics: Part B*. Elsevier, Amsterdam, 1998.

[16] P. Geladi and B.R. Kowalski. Partial least-squares regression: a tutorial. *Analytica Chemica Acta*, 185:1–17, 1986.

[17] S. de Jong. SIMPLS: an alternative approach to partial least squares regression. *Chemometr. Intell. Lab. Syst.*, 18:251–263, 1993.

[18] J.H. Kalivas. Two datasets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 37:255–259, 1997.

[19] L. Kooistra, R. Wehrens, R.S.E.W. Leuven, and L.M.C. Buydens. Possibilities of VNIR spectroscopy for the assessment of soil contamination in river floodplains. *Analytica Chimica Acta*, 446:97–105, 2001.

[20] R.L. Plackett and J.P. Burman. The design of optimum multifactorial experiments. *Biometrika*, 19:305–325, 1946.

[21] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, and J. Smeyers-Verbeke. *Handbook of Chemometrics and Qualimetrics: Part A*. Elsevier, Amsterdam, 1998.

[22] D. Levine. PGAPack V1.0. PGAPack can be obtained from anonymous ftp from: ftp://ftp.mcs.anl.gov/pub/pgapack/pgapack.tar.Z.

[23] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical recipes in C.* Cambridge University Press, Cambridge, 1988.

Chapter 6

A hybrid Genetic Algorithm - Tabu Search Approach for Optimising Multilayer Optical Coatings

Constructing multilayer optical coatings (MOCs) is a difficult large-scale optimisation problem due to the enormous size of the search space. In the present paper, a new approach for designing MOCs is presented using genetic algorithms (GA's) and Tabu Search (TS). In this approach, it is not necessary to specify how many layers will be present in a design, only a maximum needs to be defined. As it is generally recognised that the existence of specific repeating blocks is beneficial for a design, a specific GA representation of a design is used which promotes the occurrence of repeating blocks. Solutions found by GA's are improved by a new refinement method, based on TS, a global optimisation method which is loosely based on artificial intelligence. The improvements are demonstrated by creating a visible transmitting / infrared reflecting filter with a wide variety of materials.

This chapter is published as: J.A. Hageman, R. Wehrens, H.A. van Sprang and L.M.C. Buydens, A hybrid Genetic Algorithm - Tabu Search Approach for Optimising Multilayer Optical Coatings, Analytica Chimica Acta, 490, 211-222, 2003

6.1 Introduction

Multilayer optical coatings (MOCs) are coatings which consists of a stack of thin layers of materials with differences in refractive indices [1]. Depending on the total number of layers, the composition and thickness of each layer, a MOC is able to reflect certain wavelengths while other wavelengths are transmitted unhindered [1]. This property allows the design of filters with specific spectral characteristics. An important use of MOCs is the use as a visible transmitting / infrared reflecting filter applied on halogen lamps. These types of filters increase the efficiency of halogen lamps by reflecting the infrared radiation, emitted by the filament, back to the filament for re-absorption and possible re-emission in the visible wavelength range. Energy losses due to energy being radiated in the infrared region are of the order of 80% for a halogenlamp operating at 2800K but with the use of this type of filters, these losses can be reduced [2]. The synthesis of a visible transmitting / infrared reflecting filter has been the subject of a contest in 1996 [3].

Designing MOCs, or determining the optimal number of layers as well as the composition and thicknesses of each layer, is an elaborate optimisation problem, especially when considering that using state of the art deposition techniques, coatings can be made up of 75 or even more layers and a number of different materials can be used.

Several methods are available for designing MOCs. They can roughly be divided into two categories: refinement methods and synthesis methods [4]. Refinement methods need a starting design which should be close to the optimal design, otherwise no good results are obtained. These methods usually modify the thicknesses of the layers but do not influence the total number or the sequence of the layers. Synthesis methods are more general. They create a promising design without a starting design. This promising design can be refined afterwards.

A recent development for designing MOCs is the introduction of Genetic Algorithms (GAs) [5]. The GA-based method can be classified as a synthesis method, as GA's do not require a starting design. By using GAs, a 90% rejection filter [5,6], a nonpolarising edge filter [7], an antireflection filter [6,7] and a beam splitter [6] were designed by Martin et al [5–7]. These filters were designed for the wavelength range of 200 until 600 nm.

This paper introduces several improvements for designing MOCs with GAs, allowing for the design of more complex filters for a larger wavelength range. In this method, only the maximum number of layers is specified. The algorithm

will decide how many layers are optimal. As it is generally recognised that the existence of specific repeating blocks can be beneficial for a design, a special GA representation of a design is used, which promotes the occurrence of repeating blocks. Solutions found by GA's are improved by a new refinement method, based on Tabu Search (TS), a global optimisation method which is loosely based on artificial intelligence. The improvements are demonstrated by creating a visible transmitting / infrared reflecting filter with various sets of materials. However, as this is a general system, it is possible to design a multitude of different MOCs.

6.2 Theory

6.2.1 Genetic Algorithms

GA's are a special class of global optimisers, based on the theory of evolution. A GA is able to minimise (or maximise) a function G(x), where x represents a parameter vector, by searching the parameter space of x for the optimal solution [8, 9]. GA's do not operate on a single trial solution, but on a group of solutions, called a population. A solution, which is called a string, is a vector of all parameters which are to be optimised. Using evolutionary inspired operators such as fitness, crossover and mutation, the best solutions are modified and passed on to the next generation. In this way, the population as a whole moves towards better solutions, ideally to the global optimum. For a better understanding of GA's the reader is referred to [9, 10].

Representation

A trial solution, containing values of all parameters that are to be optimised, can be represented by a vector of bits, real values or integers. A design consisting of p layers is completely described by a vector, one half containing the thicknesses of each layer $\mathbf{t} = (t_1, ..., t_p)$ and one half containing the type of material per layer $\mathbf{m} = (m_1, ..., m_p)$. Therefore, the search space for designing MOCs consists of real values for the thicknesses and integer values indicating what type of material is used for each layer. Figure 6.1 shows a schematic of the representation used in the GA. This is also the representation used in the work of Martin et al [5–7]. In addition, a few improvements are introduced. If a solution is to be found with three types of material, these types are indicated by 1,2 and 3. The first major improvement is the introduction of a zero-type

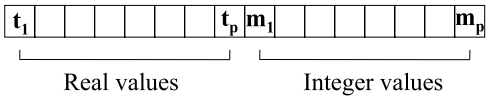

Figure 6.1: Example of the representation of a design with a maximum of eight layers in GA's. The representation is a mixture of real values for the thicknesses (t) and integer values for the material types (m).

material in the representation. When the material type is zero, a layer will have no material assigned to it and it will not be used in the design. In this way, the number of layers in a design is flexible, and the GA can decide how many layers actually will be present in a design. Only the maximum number of layers has to be specified. After applying the GA operators to a string, a cleanup and back coding of this string is performed. This is a consequence of using the zero-type material. An example of the cleanup process and the back coding is shown in figure 6.2. Two adjacent layers with identical material types are combined into one thicker layer and layers with a zero material type are removed. After this cleanup, each design is coded back to the GA string. A further essential improvement is that the representation for a design is not coded back starting at position zero in the GA string, but at a random position. In this way, the representation of a design can move across the complete string in the GA and is not only located at the first part of a string. When applying the crossover operator during a GA run, it is now possible to copy complete blocks from one string to another. After applying a number of crossover operators, it is possible that, within one string, a block is repeated. The repeated occurrence of specific blocks is considered beneficial for the performance of a design. Layers in the beginning or end in the GA representation which are not used, are given a zero material type. At the beginning of a GA run, the first generation is seeded with small designs. These design have only a few layers with a material index other than zero. Positions in a string which are not used are given material type zero.

Evaluation Function

The quality of a solution is given by the fitness value, which is calculated by the evaluation function. The fitness value is used by the GA to discriminate between good and not so good designs, so it can select accordingly. For optical filters, the fitness of a design is determined by calculating the corresponding

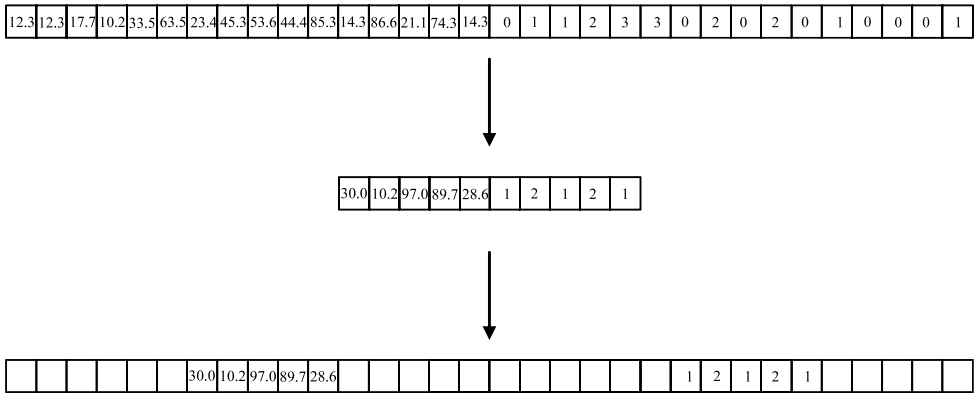

Figure 6.2: Example of the cleanup process of a string and of putting the cleaned up design back to the string. A random offset (in the example six is used) determines the position were the design is put back. Empty places in the lower string are filled with zeros, but these are omitted for clarity.

transmission spectrum using the matrix formalism [1, 11]. This spectrum is compared with a target transmission spectrum by using evaluation function F, which sums the differences between the intensities of the calculated transmission spectrum, indicated by $T(\lambda)$, and the target transmission spectrum, indicated by $\tau(\lambda)$, as shown in Eq. 6.1:

$$F = \sum_{\lambda=\lambda_{min}}^{\lambda=\lambda_{max}} \frac{\mid \tau(\lambda) - T(\lambda) \mid \cdot W(\lambda)}{N} \tag{6.1}$$

The differences in intensities are multiplied by a weighting factor $W(\lambda)$ to stress the relative importance of some areas over others. N indicates the number of wavelengths in the transmission spectrum. The resulting evaluation value F is minimised by the GA. By changing the target transmission spectrum $\tau(\lambda)$ it is possible to construct filters with specific properties. The target spectrum, in combination with a set of weights, determines what kind of filter is designed. Figure 6.3 gives a few examples of target spectra, corresponding to a beam splitter, a visible reflecting / IR transmitting and a 90% rejection filter.

6.2.2 Refinement with Tabu Search

GA's are able to locate promising regions for global optima in a search space, but sometimes have difficulty finding the exact minimum of this optima

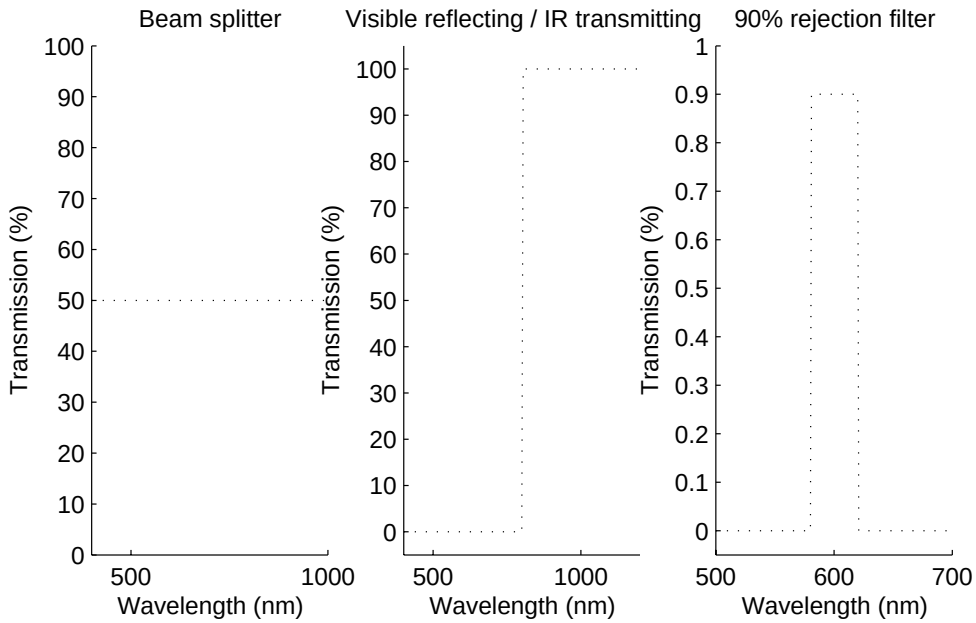

Figure 6.3: Examples of different target spectra for the construction of filters with different characteristics. Left figure shows the target spectrum for a beam splitter, the middle figure for a visible reflecting/ IR transmitting filter, the right figure shows a 90% rejection band filter.

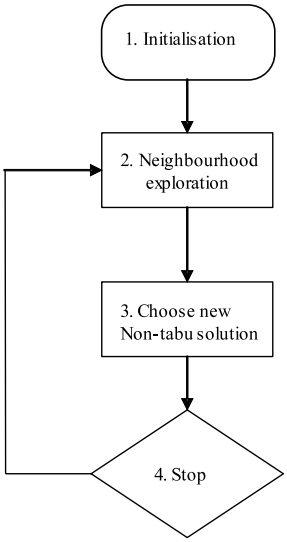

Figure 6.4: General flowchart of the Tabu Search algorithm.

[9]. Especially since the search space for constructing a visible transmitting / infrared reflecting filter is very large, it is likely that designs found by GA's can still be improved. Several examples are known from literature where a solution, obtained by GA's, are improved by a second optimisation method [12–14]. In this paper, a new procedure, based on Tabu Search (TS), is used to refine the designs. Where GA's are inspired by the process of evolution and work on a group of solutions at a time, TS is based on concepts from artificial intelligence and operates on a single solution at a time [15]. TS uses basic, problem-specific operators to explore a search space and memory (which is called the tabu list) to keep track of parts already visited. By guiding the optimisation to new areas, TS is able to overcome local minima and hopefully reach the global optimum. Refining MOCs with local optimisers could easily yield poor results, since the search space is very complex and consists of many local optima in which a local optimisers would get stuck. The foundations for TS were laid out in the late 1970's by Glover and the principles were described in general terms in 1989 and 1990 by Glover [16–18]. In recent years, tutorials documenting successes accomplished with TS have been published [15, 18, 19]. The framework of TS consists of several steps which are described below and depicted in Figure 6.4.

1. Initialisation: a starting design s is chosen. For the construction of

simple filters, this starting design can be chosen randomly. Here, the best result from the GA is used.

2. Neighbourhood exploration: all possible neighbours of design s are generated and evaluated. A neighbour is a design which can be reached from the current design by a simple, basic transformation of this current design. Two neighbouring designs resemble each other closely. For refinement of MOCs, the sequence of materials will be kept as it resulted from the GA run and only the thicknesses of the layers will be adjusted. Each thickness in turn can be adjusted by adding or subtracting a small random value. So during one iteration, $2p$ neighbours have to be evaluated for a design with p layers.

3. Choose a new non-tabu design: a new design is chosen from the explored neighbourhood. This design has the best fitness value from all neighbours and is not in the tabu list. The tabu list keeps track of previously explored designs and prohibits TS from revisiting them again. Thus, if the best neighbouring design is worse than the current design, TS will go uphill. In this way, local minima can be overcome. Instead of storing previously explored designs, it is also possible to store moves (changes to previous designs). Any reversal of these moves is then tabu, and they will remain so for a prespecified number of iterations.

4. Stop: If no more neighbours are present (all are tabu), or when during a predetermined number of iterations no improvements are found, the algorithm stops. Otherwise, the algorithm continues with step 2.

6.3 Experimental

The left part of figure 6.5 shows the target requirements for a visible transmitting/infrared reflecting filter, viz. 100% transmission in the visible wavelength range and zero transmission in the infrared region. In the evaluation function, a weight of 5.0 is given to the visible wavelength range, as it is important that the transmission is as high as possible. For the infrared range, the weights are determined differently because the spectral power distribution of a tungsten coil is not constant in the IR range [2]. To emphasise areas which have more output, the wavelengths in the IR range are weighted with the coefficients of a black body radiator at 3000K. All weights are shown in the right part of figure 6.5. Transmission spectra were calculated from 380 nm

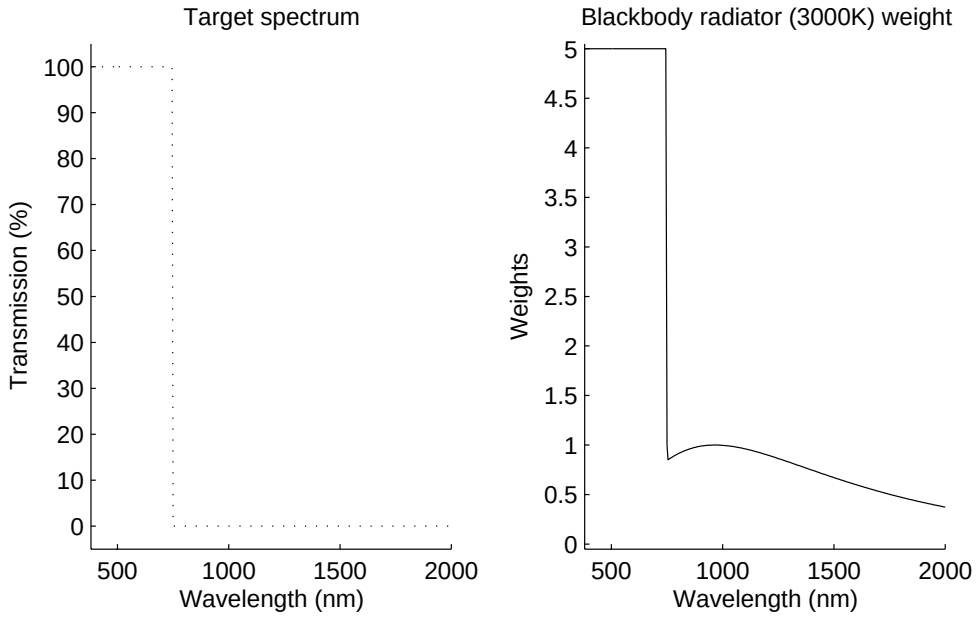

Figure 6.5: Target spectrum (left) and weights (right) for the visible transmitting/infrared reflecting filter problem.

to 2000 nm with a total of 200 datapoints. The sampling rate for calculating the transmission spectra was 5 nm in the visible wavelength range and 10 nm in the infrared range.

Several visible transmitting/ infrared reflecting filters were constructed using four different combinations of materials. A combination consists of a material with a low refractive index (SiO_2, for instance) and a high refractive index (rutile TiO_2 or SiC, for instance). The choice of materials greatly influences the quality of the filters after optimisation. To demonstrate the influence of the number and the properties of the materials and the ability of the GA-TS approach, several combinations of materials are tested. The combinations are shown in table 6.1 and described below:

1. Combination I consists of SiO_2 and Nb_2O_5. These materials are currently in use for sputtering visible transmitting/infrared reflecting filters.

2. Combination II consists of SiO_2, ZrO_2, and rutile TiO_2. Currently, filters are used which are based on two materials (as in combination I). However, in the future filters based on three different materials will be

used, as it is expected that with three different materials more efficient filters can be constructed [11]. The combination of materials is chosen in such a way that the refractive indices are ideal for a three material system. As a rule of thumb, when the refractive index of the intermediate material equals the square root of the product of the refractive indices of the other two materials, the refractive indices are considered optimal [11]. Both the dispersive and absorbing properties of the materials in combinations I and II have been taken into account, which is closest to reality.

3. The materials in combination III are identical to combination II. However, in this combination nonabsorbing properties have been assumed for all materials. Assuming nonabsorbing properties, is forcing the materials to behave more ideally. This should have a positive effect on the resulting filter and demonstrates the negative influence of the absorptive properties and thus the need for nonabsorbing materials.

4. Combination IV consists of MgF_2, ZrO_2 and SiC. Again, this is a three material system which is expected to perform very well. It represents the case were the properties of materials are ideal (nondispersive and nonabsorbing). In this case the materials also span a wide range of refractive indices. Using materials which span a wide range in refractive indices is considered beneficial for the construction of a visible transmitting/ infrared reflecting filter [2]. The refractive index of the intermediate material (ZrO_2) has, just as in combination II and III, the ideal value in combination with MgF_2 and SiC.

The wavelength dependencies of the refractive indices for the materials are shown in figure 6.6. For all combinations, the incident medium was air, normal incidence was assumed and glass was used as a substrate.

The optimal search settings of the GA and TS were determined based on previous experience and by trial and error. The settings are shown in table 6.2 and table 6.3. The maximum number of layers was set to 60. The initial population was seeded with designs which contained a maximum of 6 layers, with thicknesses ranging from 10 to 75 nm. During a run, the lower limit for the thickness of a layer was set to 10 nm, while there was no upper limit.

All software was programmed in ANSI-C. GA calculations were performed using the GA library PGAPack, version 1.0 [20]. TS was programmed from scratch. Calculations were performed on a Sun-Ultra 10 running at 440 MHz.

Table 6.1: The materials used in each combination with refractive indices at 550 nm.

Comb	n_{low}		n_{int}		n_{high}		Properties
I	SiO_2	1.46	-		Nb_2O_5	2.34	dispersive/absorbing
II	SiO_2	1.46	ZrO_2	2.06	rutile TiO_2	2.74	dispersive/absorbing
III	SiO_2	1.46	ZrO_2	2.06	rutile TiO_2	2.74	dispersive/nonabsorbing
IV	MgF_2	1.35	ZrO_2	2.00	SiC	2.60	nondisp/nonabsorbing

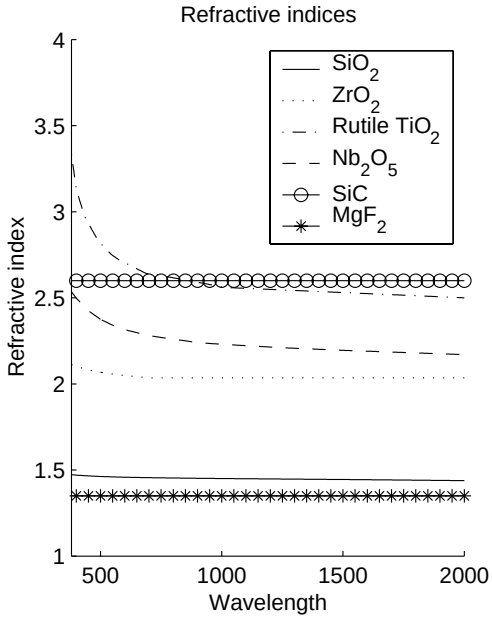

Figure 6.6: Refractive indices for the material types.

Table 6.2: Settings used for GA optimisation.

Setting	Value
Number of generations	250
Population size	600
Elitism	50
Crossover type	two-point
Crossover probability	0.7
Mutation type	Gaussian distribution with zero mean and standard deviation of 0.4
Mutation probability	0.05
Selection type	tournament selection
Fitness type	Raw

Table 6.3: Settings used in the TS-refinement of layer thicknesses in GA-generated designs.

Setting	Value
Stepsize	Random value within [0...5] nm
Length of tabu list	20 iterations
Number of iterations without improvement	50

Runtimes were in the order of 15 minutes per run. The GA runs were repeated with different random seeds to exclude any negative effects of the random starting population. At the end of a GA run, only designs which have a transmission average of over 96.5 % in the visible range are considered. An average below 96.5% is considered too low to yield an effective filter.

6.4 Results and Discussion

The left parts of figures 6.7, 6.8, 6.9 and 6.10 show the best transmission spectra after applying the GA method and refinement with TS when using the combinations of materials from table 6.1. In each of these figures, the dashed line indicates the target spectrum. Table 6.4 contains the characteristics of all four spectra. In figure 6.7, the dotted line indicates a transmission spectrum of a filter which is currently used on halogen lamps. This reference design uses the materials of combination I, consists of 47 layers and was created by refinement of a stack of blocks which influence different spectral ranges. The number of layers was kept constant during the design of this reference filter. The design located with GA's transmits an equal amount of radiation in the visible region but is able to reflect 7% more radiation in the IR range. The right parts of figures 6.7, 6.8, 6.9 and 6.10 show the composition of each design together with the thicknesses per layer in nm. All GA-runs start with completely random designs of only a few layers, but in the end, the best designs show a high degree of repetition, like one would expect when using analytical methods for filter design. Apparently, the GA-TS approach mimics the results of analytical methods.

When comparing the transmission spectra obtained with combination I and combination II, the results are better for combination II wich has three materials. The reflective properties for combination II in the IR range are good. However, mainly due to the absortance of rutile TiO_2 in the visible wavelength range, wavelenghts near 400 nm are absorbed by the filter, as shown in figure 6.8. A lamp coated with a filter based on this design will not be color neutral and will have limited applicability. Restricting the use of rutile TiO_2 improved the performance in the visible wavelength range, but deteriorated the performance of the IR wavelength range considerably. In combination III, nonabsorbtivity of all materials has been assumed. Immediately, results increased in quality, as the best transmission spectrum, shown in figure 6.9, greatly improves.

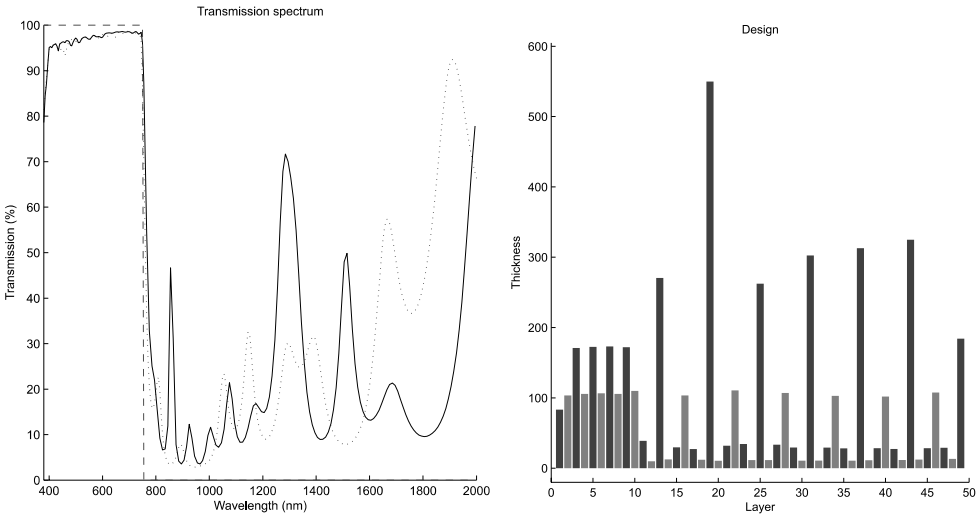

Figure 6.7: Transmission spectrum (left, solid line) and the thicknesses for each layer (right figure) for the design with combination I. In the left figure, the target spectrum is shown with the dashed line, the dotted line is the transmission spectrum of a reference filter which is actually used on halogen lamps. In the right figure: black indicates SiO_2, grey indicates Nb_2O_5.

The best results are obtained with combination IV in table 6.1, where both materials have nondispersive and nonabsorbing properties. In this example, transmission averages in both the visible and IR wavelength ranges have excellent properties. When using nondispersive and nonabsorbing materials for filter design, the problem is somewhat simplified which leads to a simpler search space and, subsequently, to the location of better solutions.

The use of more than three materials yielded no better designs. As the GA is capable of selecting the materials, it often used mainly three materials when four materials were offered. It seems more important that the three materials span a reasonable range of refractive indices.

The number of layers is not constant during a GA run. Figure 6.11 shows the number of layers in the best design of each generation for combination II from table 6.1. In the beginning of a run, the best designs contains a very small number of layers and during a run, this number steadily increases. It is very likely that the best solution will contain a large number of layers (but still below the maximum) since better designs can be created with more layers [11].

Refinement with TS is a worthwhile process as all designs optimised with

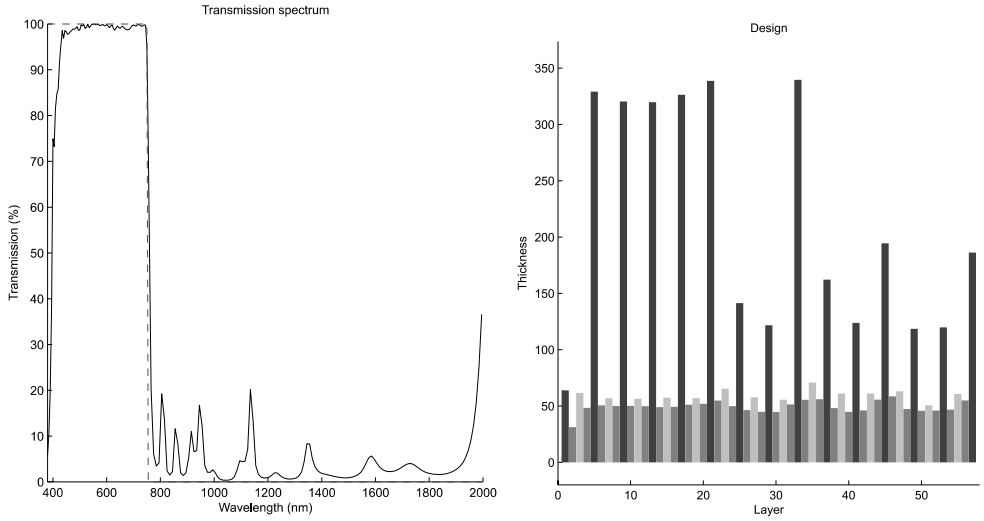

Figure 6.8: Transmission spectrum (left, solid line) and the thicknesses for each layer (right figure) for the design with combination II. In the left figure, the target spectrum is shown with the dashed line. In the right figure: black indicates SiO_2, dark grey indicates ZrO_2, light grey indicates rutile TiO_2.

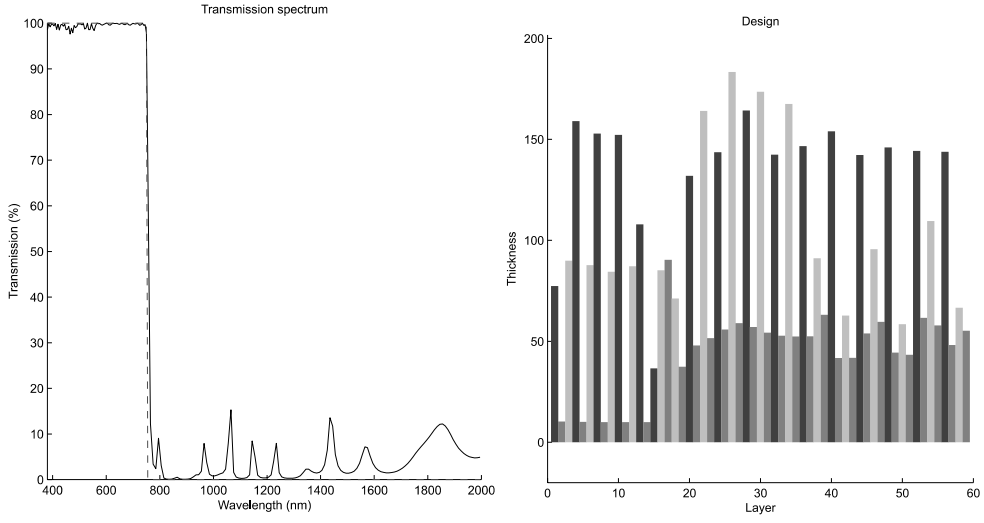

Figure 6.9: Transmission spectrum (left, solid line) and the thicknesses for each layer (right figure) for the design with combination III. In the left figure, the target spectrum is shown with the dashed line. In the right figure: black indicates SiO_2, dark grey indicates ZrO_2, light grey rutile TiO_2.

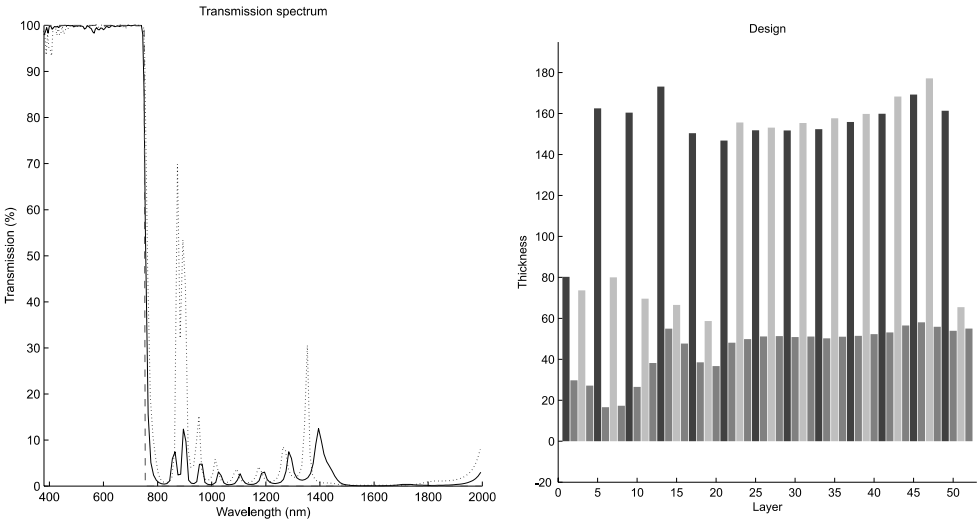

Figure 6.10: Transmission spectrum (left, solid line) and the thicknesses for each layer (right figure) for the design with combination IV. In the left figure, the target spectrum is shown with the dashed line, the dotted line indicates the transmission spectrum after the GA-run but before the TS refinement. In the right figure: black indicates MgF_2, dark gray indicates ZrO_2, light gray indicates SiC.

Table 6.4: Average transmittances for the visible and IR wavelength range for each combination of materials from table 6.1, together with the number of layers, physical thickness and measure of improvement for refinement with TS and the number of TS iterations.

Combination	Visible	IR	Layers	Thickness	TS impr	TS iter
I	96.6	22.1	49	4690 nm	11%	277
Ref MOC Fig.6.7	96.5	29.4	47	3718 nm		
II	93.5	4.8	57	5421 nm	22%	409
III	99.5	4.2	59	5056 nm	52%	375
IV	99.4	2.2	52	4690 nm	56%	754

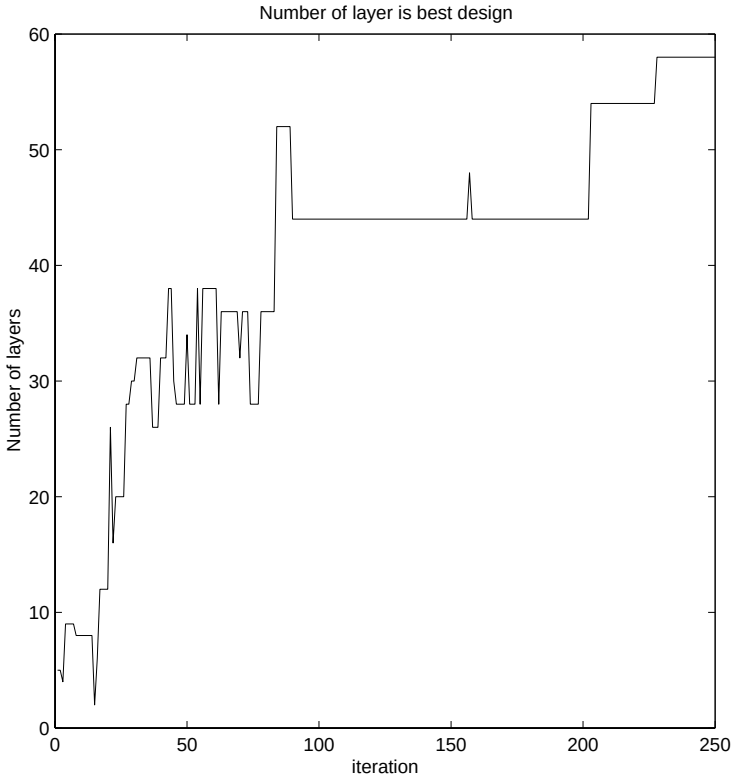

Figure 6.11: The number of layers in the best design in each iteration.

GAs could still be improved by TS. GA's are able to locate promising designs, but lack a certain precision to obtain the exact (global or local) minimum. This minimum is obtained by refinement with TS.

The improvements in evaluation value F for all four combinations are shown in table 6.4 and range from 11% to 56%. The main improvements in the transmission spectra for all four combinations are located in the IR range. The visible range usually performs well after the GA optimisation. The largest improvement (56%) can be found for combination IV. In the left part of figure 6.10, the best design obtained with the GA is shown with the dotted line. After refinement with TS (shown with the solid line), the transmission spectrum clearly shows improvement, mainly reached by the decrease of the large transmission peak around 900 nm. The number of iterations for the TS refinement varies somehow, the minimum being 277 (combination I), the maximum being 754 iterations (combination IV).

In this approach, TS is used for refining designs which were obtained with GA's. However, designing simple filters without a good starting point is also possible with TS. When using two materials, an alternating sequence of materials is used. When using more than two materials, the sequence of materials also needs to be optimised. If the sequence has to be optimised with TS, it is necessary to determine operators which modify the material type of a layer when a neighbourhood is explored. However, in the approach demonstrated here, it is shown that GA's are very well able to establish a good sequence of materials with reasonable values for the thicknesses. These thicknesses can be refined with TS with good results.

It is unknown how the filters, optimised in this paper, will perform when implemented. Since small errors in the deposition of layers could change the characteristics, some changes could be expected. It is possible to create filters which are more robust against small deviations in thicknesses. To asses the influence of these deviations, filters can be evaluated multiple times while some random deviations to the layers thicknesses are added. Filters which are sensitive to these deviations while deteriorate more and get a lower evaluation value. In the end, the obtained filter will more robust.

It is also possible to perform the TS refinement during a GA run and refine the members of the GA population at each generation. It might be possible, due to synergetic effects, that better solutions might be obtained. However, calculation times would become too long and this makes any practical application virtually impossible.

The transmission spectra obtained with the GA-TS method for construct-

ing a visible transmitting / IR reflecting filter, cannot directly be compared with results from the contest [3]. The contest used different materials, a different evaluation function which also took into account the number of layers while the wavelength ranges between 380-400 and 720-750 were left out. However, these wavelengths are crucial for some real-world applications, such as the halogen lamp filters. Furthermore, it was not necessary to penalise larger designs compared to smaller designs because in our approach only a maximum number of layers was important.

6.5 Conclusion

As a typical design for a MOC can contain up to 75 layers, each with their own thickness and material type, the search space for creating MOCs is enormous, which makes constructing MOCs a difficult optimisation process.

In this paper, several improvements are introduced for optimising MOCs with GA's. Firstly, the representation employed, including a zero-type material, makes it unnecessary to specify beforehand how many layers will be present in a design. The algorithm can decide this, which is very efficient. Secondly, the special cleanup of the GA representation of a MOC and subsequently back coding at a random position makes it possible, by applying the crossover operator, to copy complete blocks from one design to another design at any position. The repeated occurrence of blocks is considered beneficial for the performance of MOCs. Finally, GA's are able to locate promising designs, but lack a certain precision. Designs optimised with GA's, are refined by a new refinement method, based on TS. TS refinement leads to improvements in the range of 10 - 50% for the examples shown in this paper.

All the improvements are demonstrated by the creation of a visible transmitting / infrared reflecting filter using several combinations of materials. The choice and number of materials influences the quality of the best designs. As expected, the use of three materials yielded better designs compared to the use of two materials. Using dispersive and absorbing materials make it harder to obtain a good filter, for instance in combination II where the specific absorbance of rutile TiO_2 is a problem. Using nondispersive and nonabsorbing materials yield good filters by simplifying the problem. This also stresses the importance of the use of materials with the correct properties. Three materials, with no or a minimum of absorbing or dispersive properties and a large difference between the lowest and highest refractive index, while the third

has an intermediate value, seems to work the best. As this is a very general method, it is possible to design a multitude of different MOCs.

References

[1] H.A. Macleod. *Thin Film Optical Filters*. Adam Hilger Ltd, Bristol, 1986.

[2] R.S. Bergman and T.G. Parham. Applications of thin film reflecting coating technology to tungsten filament lamps. *IEE Proceedings-A*, 6:418–428, 1993.

[3] A. Thelen. Design of a hot mirror: contest results. *Applied optics*, 35:4966–4977, 1996.

[4] J.A. Dobrowolski and R.A. Kemp. Refinement of optical multilayer systems with different optimization procedures. *Applied optics*, 29:2876–2893, 1990.

[5] S. Martin, J. Rivory, and M. Schoenauer. Simulated darwinian evolution of homogeneous multilayer systems: a new method for optical coatings design. *Optics Communications*, 110:503–506, 1994.

[6] S. Martin, J. Rivory, and M. Schoenauer. Synthesis of optical multilayer systems using genetic algorithms. *Applied optics*, 34:2247–2254, 1995.

[7] S. Martin, A. Brunet-Bruneau, and J. Rivory. Darwinian evolution of homogeneous multilayer systems: a new method for optical coatings design. *SPIE*, 2253:168–174, 1994.

[8] J.H. Holland. *Adaptation in naturial and artificial systems*. MIT Press, 1992.

[9] D.E. Goldberg. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley, 1989.

[10] R. Wehrens and L.M.C. Buydens. Evolutionary optimisation: a tutorial. *Trends in Analytical Chemistry*, 17(4):193–203, 1998.

[11] A. Thelen. *Design of optical interference coatings*. McGraw-Hill, New York, 1989.

[12] D. Hibbert. A hybrid genetic algorithm for the estimation of kinetic parameters. *Chemometrics and intelligent laboratory systems*, 19:319–329, 1993.

[13] G. Vivo-Truyols, J. R. Torres-Lapasio, and M. C. Garcia-Alvarez-Coque. A hybrid genetic algorithm with local search: I. discrete variables: optimisation of complementary mobile phases. *Chemometrics and intelligent laboratory systems*, 59:89–106, 2001.

[14] S. Wong. Hybrid simulated annealing/genetic algorithm approach to short-term hydro-thermal scheduling with multiple thermal plants. *International journal of electrical power & energy systems*, 23:565–575, 2001.

[15] F. Glover and M. Laguna. *Tabu Search*. Kluwer Academic Publishers, Dordrecht, 1998.

[16] F. Glover. Tabu search – part I. *ORSA Journal on Computing*, 1(3):190, 1989.

[17] F. Glover. Tabu search – part II. *ORSA Journal on Computing*, 2(1):4, 1990.

[18] F. Glover, E. Taillard, and D. de Werra. A user's guide to tabu search. *Annals of Operations Research*, 41:3–28, 1993.

[19] D. de Werra and A. Hertz. Tabu search techniques, a tutorial and an application to neural networks. *OR Spektrum*, 11:131–141, 1989.

[20] D. Levine. PGAPack V1.0. PGAPack can be obtained from anonymous ftp from: ftp://ftp.mcs.anl.gov/pub/pgapack/pgapack.tar.Z.

Chapter 7

Conclusions and future research

7.1 Conclusions

SA, GA's and TS are global optimisation techniques. All three methods are able to optimise a cost function $F(x)$ by searching the parameters space of x for the optimal combination of x. All three methods should perform equally well, however, given their characteristics, one method could be more suitable than others for a given problem.

In this section, the conclusions of the research presented in this thesis in relation to the question posed in section 1.4, will be discussed.

7.2 Evaluation function

7.2.1 Parameter optimisation

An evaluation function is an important part of all three optimisation algorithms. It determines the shape of the search landscape and greatly influences the chance of finding the global optimum. This thesis investigated a specific type of evaluation function, a criterion which determines the (dis)similarity between two spectra (e.g. an experimental and a theoretical one). An evaluation function of this type can be encountered in the parameter optimisation class. Traditional comparison methods, which are based on pointwise comparisons between the two states, often fail in the presence of shifted peaks. This is clearly demonstrated in chapters 3 and 4. Both applications are exam-

ples where a correct comparison between experimental and theoretical states is crucial. When peaks do not completely overlap anymore, a point wise comparison reveals no information on similarity. This effect is only strengthened by the comparison of the position of peaks without line widths in chapter 4.

The newly developed weighted crosscorrelation (WCC), as introduced in chapter 2, overcomes the drawbacks of traditional comparison methods. By taking the neighbourhood of peaks into account, the presence of shifted peaks is detected and can be taken into account when calculating the (dis)similarity. The use of the WCC function proved to be critical for the analysis of rotationally resolved spectra (chapter 3) and for the indexing of powder patterns (chapter 4). In these cases, an ordinary comparison function yielded no results, because of the possibility of shifting peaks after small adjustments in the optimisation parameters.

7.2.2 The use of error landscapes

The influence of changes to the evaluation function can be spotted at forehand by visualising the error landscape. It is not straightforward to visualise an error landscape for every optimisation problem. However, visualising the effects of changes to the parameters of parameter optimisation problems is simple. By varying two parameters in a grid like manner while keeping the others constant, calculating the response values gives the error landscape. A 2D error landscape cannot show the changes as they would appear in a multidimensional error landscape, but still it will demonstrate trends. This is clearly demonstrated in figure 3.2 in chapter 3. In this way, any influence can be judged before, rather than after numerous experiments. These experiments can then be used for other purposes, e.g. an experimental design for optimising the meta-parameters.

In the case of including the neighbourhood of peaks in a comparison, this leads to a decrease in the number of local optima and a widened and smoothed global optimum. These effects are beneficial for optimisation in general. This shows that it is important to know the characteristics (e.g. the possible existence of shifting peaks) of an optimisation problem, so an evaluation function can be devised which is able to deal with it properly.

7.2.3 Variable selection

The evaluation function used in the wavelength selection application (chapter 5) is less complicated, compared to the previous examples, but nevertheless

requires attention. The use of PLS requires the calculation of the number of latent variables which need to be done correctly since the use of too many latent variables can lead to overfitting. In our work, we used leave-m-out cross validation to determine the number of latent variables, with m being a number (4 or 5) which divided the training set into roughly 15-20 groups. By leaving multiple observations out, overfitting was avoided, something what easily happens when using leaving-one-out crossvalidation [1, 2].

7.3 Tabu Search in relation with SA and GA's

TS is a relatively new technique, especially within chemometrics but with equal possibilities compared with SA and GA's. In contrast to SA and GA's, TS is deterministic, which means that given a starting solution, it will always come up with the same end solution. This makes replicate runs unnecessary. However, when TS is started with a random solution, the probabilistic nature of SA and GA's is also introduced in TS.

SA and TS are both methods that can be started with a single starting solution, unlike GA's which need a group of starting solutions. SA and TS can therefore be used to refine a single solution. When used as a refinement method, SA and TS use a suboptimal solution as a starting solution. This suboptimal solution can be the end result of another method. An example is the construction of a infrared reflecting mirror as described in chapter 6. In this case, TS uses the solution obtained by GA's as a starting solution.

7.3.1 Characteristics of neighbourhood

TS carefully examines the neighbourhood of a current solution with a step function. This step function can be identical to the one used in SA. To examine the neighbourhood, every single solution which can be reached from the current solution, is generated and tested. This is an important difference with SA, since SA just picks one single solution from the neighbourhood. When designing a step function, it is important that the step function keeps the number of neighbours limited. Since all neighbours have to be examined in every iteration, this part of the algorithm can easily make TS too slow in practice. A good step function, keeps the numbers of neighbours minimal but is still able to cover a lot of search space. The latter characteristic is also influenced by the tabu list, as is discussed later on. If an optimisation problem has many

parameters, changing only some of the parameters per iteration can help to keep the number of possible neighbours at a minimum.

When an optimisation problem consists of integer variables, as in the case of the wavelength selection problem (chapter 5), the examination of the neighbourhood can be exact. In this case, the step function is the addition or removal of a wavelength from the current solution. When the neighbourhood of a solution consists of real variables (for instance in the case of parameter optimisation problems), the examination of the neighbourhood uses some discretisation step. A grid like approach could also be possible, but was not investigated in this thesis. An example of a discretisation step can be found in the refinement of multilayer optical coatings with TS (chapter 6). In this case, the step function added small random values to the optimisable parameters to explore their neighbourhood. Using a discretisation step introduces two drawbacks. Firstly, randomness is added to the otherwise deterministic nature of TS. Secondly, it now becomes possible when reaching the global minimum to miss the exact minimum because of the random step size. These drawbacks do not exist when an exact scan of the neighbourhood is possible.

7.3.2 Characteristics of tabu list

An aspect which is not encountered in SA or GA's is the tabu list. The tabu list is the memory of TS and prevents returning to previously visited solutions. In this way, TS is able to overcome local minima. When the tabu list is too short, TS will end up in a cycle and is not able to examine new parts of the error landscape. When the tabu list is too long, checking solutions will take longer than (perhaps) necessary. It is also possible that the exclusion of parts of the search space (still present in the tabu list) will hinder TS by finding the global minimum.

It is possible to store the exact solutions that have been encountered. However, when many parameters are involved, storing and checking the memory of TS becomes a rather computer memory and computer intensive process. It is also possible to store the steps which were necessary to create previously visited solutions. The main advantage of storing steps, is that it takes less memory and less time to find restricted solutions.

7.4 Choice of optimisation algorithm

TS is very useful for optimisation problems where an exact definition of the neighbourhood is possible. This is the case when dealing with optimisation problems defined by integers, for instance with problems from the subset selection class (for example the wavelength selection problem in chapter 5). When there is no clear definition of a neighbourhood of a solution, the use of SA or GA's seems more justified. The drawbacks of using a discretisation step have no extra influence on SA, since this algorithm already has a probabilistic nature. When it is difficult to define a step function which limits the number of neighbours, TS can become a rather slow process. In that case, it might be advantageous to use SA, since SA chooses one random neighbouring solution rather than generate all first and pick one.

When forbidden areas in the search space can be expected, GA's are a more logical choice. An advantage of GA's is that they do not need a user defined step function, but this has only shifted the problem since the user has to come up with a correct representation of the problem parameters. This representation is very important and will determine the possibilities of optimisation with GA's. An example of an efficient GA representation is shown in chapter 6. In this representation, a zero material design was introduced which allows for a flexible number of layers. The special cleanup of this representation after each GA iteration demonstrates that the representation is not static but can be changed at any moment.

7.5 Meta-optimisation of SA, GA' and TS

Finding correct values for the meta-parameters of all three global optimisation methods can be a tedious work. Especially when, at the same time, a correct expression for the evaluation function has to be devised. In this thesis, finding the settings for the meta-parameters was done by using an experimental design (chapter 5) or, when more experience is present, trial-and-error can be used while using results from past experiments (chapters 3,4 and 6).

When developing an optimisation application, starting simple can help clarify things. It can be useful to test an optimisation application by using synthetic data. In this way, the influence of noise and other experimental artifacts is removed. When, in this case, the application fails, this can be ascribed to the method and not to other (experimental) effects. The introduction of different levels of noise also gives the opportunity to investigate the effects of

deteriorated data on the optimisation method, which can help to discover the limits of the method in terms of the experimental data. An example was given in chapter 3.

By optimising only a few of all optimisable parameters (and keeping the rest constant at the correct values), the influence of incorrect meta-parameters diminishes somewhat, while the evaluation function remains as important as before. In this way, it can be determined if the meta-parameters or perhaps the evaluation function is the cause of a failure to produce good results.

7.6 Future research

7.6.1 The use of error landscapes

The development of the weighted cross correlation (WCC) was of paramount importance for obtaining correct results when analysing LIF fluorescence spectra (chapter 3) and the indexing of powder patterns (chapter 4). By examining the search landscape, the influence of using the WCC criterion instead of a point wise criterion could very well be visualised. The search landscape also played an important role during the development of the WCC criterion. In many areas of sciences, multi-dimensional error landscapes (or sometimes called hypersurfaces) are encountered [3,4]. Understanding the characteristics of error landscapes can facilitate in understanding and developing applications. An example is given in chapter 3. Changes to the WCC criterion have a clear effect on the error landscape. Visualising these changes can help understand how the method works. In this example, the visualisation was done in 2D, but other possibilities exist. For instance, a plot in 1D or an aspect of an error landscape(e.g. roughness) can simply be expressed in one single number.

The visualisation of the search landscape can be an important tool for the development of an evaluation function. In this thesis, error landscapes were constructed by varying two parameters while keeping the other constant. When dealing with a multiple parameter problem, this gives a limited impression of the error landscape. Nevertheless, in our cases it sufficed to demonstrate the effects of changes to the evaluation function (see chapter 3). A different way to visualise parts of a multi-dimensional error landscape would be by taking a trajectory through the hypersurface by allowing changes to all parameters simultaneously [5]. This method can be extended to a plane. This method allows general trajectories through a multi-dimensional param-

eter space in which the values of more than one variable are changed simultaneously. An advantage of this approach is that it is possible to show the dependencies of more than two variables.

A interesting combination, which has not been made in this thesis, would be the visualisation of the search trajectory of the global optimisation method in combination with the error landscape [6]. Analysis could possibly aid in finding correct settings for the meta-parameters and/or help devise a correct step function in SA and TS. Alternatively, such a technique would allow the visualisation of the quality criteria for evaluating the performance of optimisation algorithms [7].

A future line of research in global optimisation methods should focus more on the use of search landscapes. In this thesis, a visualisation has been made only for problems from the parameter optimisation class. However, problems from the variable selection and sequence optimisation class could benefit as well from insights gained during the investigation of the search landscape. The visualisation of the search landscapes from these problem classes is more difficult and requires more research. Ultimately, an 1D or 2D impression of a multi dimension search landscape would be ideal. A difficult challenge will be the visualisation of a multi-dimensional error landscape.

7.6.2 One step structure elucidation from X-ray powder diffraction patterns

The elucidation of crystal structures from X-ray powder diffraction data can be divided into four steps [8]: (i) indexing of powder diffraction patterns (determination of unit cell parameters), (ii) space group determination, (iii) crystal structure solution (initial model) and (iv) crystal structure refinement (final model). Traditional methods for solving crystal structures follow these steps. Recent developments and improvements have mainly been focussed on step (iii) [9–12], but the bottleneck for a complete structure elucidation via this scheme is step (i). Chapter 4 describes a method based on GA's and the WCC for indexing powder patterns (step (i)).

Something which has not been investigated in this thesis, is a different approach for structure elucidation. When indexing has been unsuccessful, crystal structure prediction methods [13] may still be able to predict a structure that matches the experimental powder pattern. Crystal structure prediction relies on knowledge of the contents of the unit cell and the use of force field calculations. By varying the unit cell parameters (a, b, c, α, β and γ) and the

parameters on the molecular structure (describing the location per (group of) atom(s) in the unit cell), crystal structure prediction is able to optimises a structure. Comparison of the patterns calculated with the predicted structure and the experimental pattern determines whether the predicted structure was correct. The unit cell parameters determine the location of the peaks while the molecular structure parameters determine the intensity of the peaks. The number of molecular structure parameters can be reduced by working with molecular fragments and expressing their location in rotational and translational terms.

The principle of optimising model parameters and matching the output to experimental data has been described in chapters 3 and 4. For such an application to correctly work, a suitable optimisation algorithm and a suitable evaluation function has to be found. In chapters 3 and 4, GA's have been used as optimisation technique, while the similarity between theory and experiment was determined by the WCC. It is possible that a similar approach could work for a one-step structure prediction from powder patterns. GA's (or other global optimisation techniques) can be used for optimising the unit cell parameters and structural parameters, while the WCC determines the match between theory and experiment. Since peak positions can easily shift after small changes to the unit cell parameters, the WCC will still determine a correct match between calculated and experimental pattern.

An advantage of this approach over the traditional method would be that structure elucidation from powder patterns is now possible in one step while using all the information present in a powder pattern a the same time. Practical problems such as peak overlap and peak picking will become less important in such an approach, while other (such as impurities and zero shift) may remain a problem.

References

[1] K. Baumann, H. Albert, and M. von Korff. A systematic evalution of the benefits and hazards of variable selection in latent variable regression. part I. search algorithm, theory and simulations. *Journal of chemometrics*, 16:339–350, 2002.

[2] K. Baumann, H. Albert, and M. von Korff. A systematic evalution of the benefits and hazards of variable selection in latent variable regression. part II. practical applications. *Journal of chemometrics*, 16:351–360, 2002.

[3] J.C. Schon and M. Jansen. Determination, prediction, and understanding of structures, using the energy landscape of chemical systems - part i. *Z. Kristallogr.*, 216:307–325, 2001.

[4] J.C. Schon and M. Jansen. Determination, prediction, and understanding of structures, using the energy landscape of chemical systems - part ii. *Z. Kristallogr.*, 216:361–383, 2001.

[5] G.W. Turner, E. Tedesco, K.D.M. Harris, R.L. Johnston, and B.M. Kariuki. A method for understanding characteristics of multi-dimensional hypersurfaces, illustrated by energy and powder profile r-factor hypersurfaces for molecular crystals. *Z. Kristall.*, 216:187–189, 2001.

[6] S. Habershon, K.D.M. Harris, R.L. Johnston, G.W. Turner, and J.M. Johnston. Gaining insights into the evolutionary behaviour in genetic algorithm calculations, with applications in structure solution from powder diffraction data. *Chemical Physics Letters*, 353:185–194, 2002.

[7] R. Wehrens, E. Pretsch, and L.M.C. Buydens. Quality criteria of genetic algorithms for structure optimisation. *J. Chem. Inf. Comput. Sci.*, 38(2):151–157, 1998.

[8] K.D.M. Harris. Crystal structure determination from powder diffraction data. *Chem. Mater.*, 8:2554–2570, 1996.

[9] K. Shankland, W.I.F. David, and T. Csoka. Crystal structure determination from powder diffraction data by the application of a genetic algorithm. *Z. Kristall.*, 212:550–552, 1997.

[10] B.M. Kariuki, H. Serrano-Gonzalez, and R.L. Johnston. The application of a genetic algorithm for solving crystal structures from powder diffraction data. *Chem. Phys. Lett.*, 280:189–195, 1997.

[11] B.M. Kariuki, S.A. Belmonte, M. I. McMahon, R.L. Johnston, K.D.M. Harris, and R.J. Nelmes. A new approach for indezing powder diffraction data based on whole-profile fitting and global optimization using a genetic algorithm. *J. Synchrotron Rad.*, 6:87–92, 1999.

[12] K.D.M. Harris, R.L. Johnston, G.W. Turner, E. Tedesco, E.Y. Cheung, and B.M. Kariuki. Recent advances in the opportunities for solving molecular crystal structures directly from powder diffraction data. *Molecular Crystals And Liquid Crystals*, 389(1):123–129, 2002.

[13] P. Verwer and F.J.J Leusen. Computer simulation to predict possible crystal polymorphs. *Rev. Comp. Chem.*, 12:327–365, 1998.

Summary

The goal of chemometrics is to increase the information yield of chemical data by applying tools from mathematics, statistics and computer science. A specific group of chemometrical tools are the global optimisation algorithms such as Simulated Annealing (SA), Genetic Algorithms (GA's) and Tabu Search (TS). These tools can be used for finding optimal parameters for optimisation problems. Global optimisation algorithms are very useful since optimisation problems are present in abundance in chemistry. In general, optimisation problems can be divided into three categories, (i) parameter optimisation, (ii) subset selection and (iii) sequence optimisation.

Unfortunately, the use of global optimisation methods is not straightforward, especially for the more difficult optimisation problems. Solutions have to be found for items such as the evaluation function, representation, step function and meta-parameters, before any useful results can be obtained. This thesis aims at investigating and improving the use of global optimisation algorithms. In particular, this thesis will focus at three specific problems which are associated with global optimisation. (1) It aims at finding a similarity criterion which deals with the problem of correctly comparing spectra when many shifted peaks are present and that can be used as an evaluation function for optimisation purposes. (2) TS is a relatively new optimisation technique with different characteristics compared to SA and GA's. By implementing TS to solve several chemical optimisation problems, this thesis investigates the properties and the possibilities of TS. (3) Unfortunately, for all three methods, there exists no standard recipe on how or when to use SA, GA's or TS. By studying and solving several chemical optimisation problems, a third goal is to detect guidelines on how and when to use global optimisation algorithms.

Chapter 2 describes a newly developed similarity criterion, the weighted cross correlation (WCC). When comparing spectra which contain many shifted peaks relative to each other, traditional methods often fail to recognise any

similarity. The WCC criterion is able to deal with shifting peaks when comparing spectra or any other vector like data by including a comparison of the neighbourhood of peaks. In chapter 2, the WCC criterion is applied to the classification of crystal structures of cephalosporin complexes, based on the comparison of their calculated powder diffraction patterns. Inclusion of the neighbourhoods of peaks leads to a significant improvement of discriminating power and in this way, makes it possible to recognise closely related structures.

Chapter 3 and 4 deal with optimisation problems which fall in (i) the parameter optimisation category. Chapter 3 describes a new procedure for the analysis of rotationally resolved fluorescence spectra based on GA's. Chapter 4 describe a method for the indexing of powder diffraction patterns based on GA's and TS, respectively. In these cases, the success crucially depended on the newly developed similarity criterion, the weighted cross correlation (see chapter 2). The problem of a correct comparison of rotationally resolved fluorescence spectra or the comparison of powder patterns is related to peak shifts which are caused by small changes in the model parameters. To be able to make a meaningful comparison, it is necessary to include a comparison of the neighbourhood of a given point in the spectrum. This demonstrates that a special tailor-made evaluation function is crucial to obtain any results.

Chapter 5 investigates the properties and possibilities of TS in chemometrics by applying it to wavelength selection, a problem from (ii) the subset selection category. In this class of problems, the goal is to select a small group of variables from a large group. This subset usually has identical properties compared to the original, larger group of variables or performs even better. TS is a deterministic, global optimisation technique loosely based on concepts from artificial intelligence. TS uses basic, problem-specific operators to explore a search space. TS stores solutions in memory to keep track of parts already visited. Several implementational aspects of wavelength selection with TS are discussed. Several intensification and diversification approaches are applied and are shown to be valuable extensions of TS. It is possible to use TS for locating solutions with a variable and constant number of wavelengths. It is demonstrated with three real world data sets, that TS performs equally well and can be a valuable alternative to SA and GA's, especially in cases where there is a clear definition of a neighbourhood of a solution.

Chapter 6 deals with a problem from (iii) the sequence optimisation category. In this chapter, GA's and TS are used for constructing multilayer optical coatings (MOCs). MOCs are coatings which consists of a stack of thin layers of materials with differences in refractive indices. Depending on the

total number of layers, the composition and thickness of each layer, a MOC is able to reflect certain wavelengths while other wavelengths are transmitted unhindered. Constructing multilayer optical coatings (MOCs) is a difficult large-scale optimisation problem due to the enormous size of the search space. The approach described in chapter 6 makes it unnecessary to specify at forehand the total number of layers. A special GA representation is used which allows specific blocks to be copied from one solution to another. This approach promotes the occurrence of repeating blocks which is beneficial for the performance. Solutions which are found by the GA approach are refined by a new refinement method based on TS. The improvements are demonstrated by creating a visible transmitting / infrared reflecting filter with a wide variety of materials.

Some conclusions and ideas for future research are given in chapter 7.

Samenvatting

Het doel van chemometrie is het verhogen van de informatie opbrengst van chemische data door de toepassing van technieken uit de wiskunde, statistiek en informatica. Een specifieke groep technieken zijn de globale optimalisatie algoritmen zoals Simulated Annealing (SA), Genetische Algoritmen (GA's) en Tabu Search (TS). Door het gebruik van deze algoritmen is het mogelijk om optimale instellingen voor optimalisatie problemen te vinden. Globale optimalisatie algoritmen zijn zeer nuttig voor chemici aangezien optimalisatie problemen in overvloed binnen de scheikunde aanwezig zijn. Optimalisatie problemen kunnen in drie categorieën worden verdeeld, (i) parameter optimalisatie, (ii) subset selectie en (iii) volgorde optimalisatie.

Het gebruik van globale optimalisatie methoden is niet zonder problemen, zeker niet voor de wat moeilijkere optimalisatie vraagstukken. De evaluatiefunctie, de representatie van het probleem, de stap functie en de meta-parameters zijn onderdelen van de algoritmes die eerst gekozen moeten worden, voordat nuttige resultaten geboekt kunnen worden. Dit proefschrift onderzoekt het gebruik van globale optimalisatie algoritmen en hoopt zo het gebruik ervan te verbeteren. Dit proefschrift richt zich op drie specifieke problemen die kunnen voorkomen bij optimalisatie vraagstukken. (1) Het vinden van een similariteitscriterium dat op een correcte wijze spectra, met vele verschoven pieken daarin, kan vergelijken. Dit criterium moet geschikt zijn om gebruikt te kunnen worden als evaluatiefunctie voor optimalisatie doeleinden. (2) TS is een vrij nieuwe optimalisatie techniek. TS heeft andere kenmerken in vergelijking met SA en GA's. De eigenschappen en mogelijkheden van TS worden onderzocht door TS te implementeren om verschillende chemische optimalisatie problemen op te lossen. (3) Er bestaat helaas geen standaardrecept hoe en wanneer SA, GA's en TS, te gebruiken. Een derde doel van dit proefschrift is om richtlijnen voor het gebruik van globale optimalisatie algoritmen te ontdekken. Hiertoe worden verschillende chemische optimalisatie problemen

bestudeerd.

Hoofdstuk 2 beschrijft een recent ontwikkeld similariteitscriterium, de weighted crosscorrelation (WCC). Traditionele methodes slagen er vaak niet in om een correcte similariteit te bepalen tussen spectra, waar vele, ten opzichte van elkaar verschoven pieken, in voorkomen. Het WCC criterium is in staat om om te gaan met verschoven pieken door tijdens het vergeleken van twee spectra, of andere vector georiënteerde data, ook de omgeving van een piek mee te nemen. In hoofdstuk 2, wordt het WCC criterium gebruikt in de classificatie van kristalstructuren van cephalosporine complexen. Deze classificatie is gebaseerd op de vergelijking van berekende poederdiffractie patronen. Het meenemen van de omgeving van pieken tijdens het bepalen van de similariteit, leidt tot een significant verbeterde bepaling van de similariteit en maakt het mogelijk om sterk verwante structuren te herkennen.

Hoofdstuk 3 en 4 behandelen optimalisatie problemen uit categorie (i), de parameteroptimalisering. Hoofdstuk 3 beschrijft een nieuwe procedure voor de analyse van rotationeel opgeloste fluorescentiespectra. Deze methode is gebaseerd op GA's. Hoofdstuk 4 beschrijft een methode, eveneens gebaseerd op GA's, voor het indexeren van poeder diffractie patronen. In beide gevallen, werd het succes bepaald door het recent ontwikkelde similariteitscriterium, de weighted crosscorrelation (zie hoofdstuk 2). Een correcte vergelijking van rotationeel opgeloste fluorescentiespectra of van poederpatronen wordt bemoeilijkt door piekverschuivingen die door kleine veranderingen in de modelparameters worden veroorzaakt. Om toch een zinvolle vergelijking te kunnen maken, is het noodzakelijk om de omgeving van pieken in een spectrum mee te nemen. Dit toont aan dat een op maat gemaakte evaluatiefunctie essentieel kan zijn om goede resultaten te krijgen.

Hoofdstuk 5 onderzoekt de eigenschappen en de mogelijkheden van TS in de chemometrie door TS op het probleem van de golflengteselectie toe te passen. Golflengteselectie is een probleem uit categorie (ii), subsetselectie. In deze klasse, is het doel om een kleine groep variabelen uit een grote groep te selecteren. Deze subset heeft gewoonlijk identieke eigenschappen in vergelijking met de originele set, de grotere groep variabelen, of presteert nog beter. TS is een deterministische, globale optimalisatie techniek die op concepten van kunstmatige intelligentie is gebaseerd. TS gebruikt simpele, probleem-specifieke operatoren om een zoekruimte te onderzoeken. TS slaat oplossingen in het geheugen op, om zo reeds bezochte oplossingen bij te houden. Verscheidene implementationele aspecten van golflengteselectie met TS worden besproken. Verschillende intensivering en diversificatie benaderingen worden toegepast en

blijken waardevolle uitbreidingen van TS te zijn. Het is mogelijk om TS te gebruiken om oplossingen te vinden met een variabel of een constant aantal golflengten. Met behulp van drie data sets, wordt aangetoond, dat TS even goed presteert als GA's en SA. TS kan een waardevol alternatief zijn voor SA en GA's, vooral in gevallen waarbij een duidelijke definitie is van de omgeving van een oplossing.

Hoofdstuk 6 behandelt een probleem uit categorie (iii), de volgorde optimalisatie. In dit hoofdstuk worden GA's en TS gebruikt voor het ontwerpen van meerlaags optische coatings (MOCs). MOCs zijn coatings die bestaan uit een aantal zeer dunne lagen van materialen met verschillende brekingsindex. Afhankelijk van het totale aantal lagen, de samenstelling en de dikte van elke laag, kunnen MOCs bepaalde golflengten reflecteren terwijl andere golflengten ongehinderd kunnen passeren. Het construeren van meerlaags optische coatings (MOCs) is een moeilijk optimalisatie probleem. Dit komt door de enorme grootte van de zoekruimte. De benadering die in hoofdstuk 6 wordt beschreven zorgt ervoor dat het totale aantal lagen niet gespecificeerd hoeft te worden. Een speciale GA representatie zorgt ervoor dat specifieke blokken van de ene oplossing naar de andere gekopieerd kunnen worden. Deze benadering stimuleert herhaling van dezelfde blokken in een meerlaags structuur. Dit is gunstig voor de prestaties van het filter. De oplossingen die met de GA aanpak gevonden worden, worden verfijnd met een nieuwe methode die gebaseerd is op TS. De verbeteringen worden gedemonstreerd door het ontwerpen van een filter dat zichtbaar licht doorlaat en IR licht reflecteerd.

In hoofdstuk 7 worden enkele conclusies en ideëen voor toekomstig onderzoek gegeven.

Dankwoord

Ruim vier jaar heb ik op de Afdeling Analytische Chemie in Nijmegen gewerkt. Deze mooie tijd is omgevlogen. Het resultaat van deze periode heeft zich in dit boekje laten opschrijven. Ondanks het feit dat er slechts één auteur op de kaft van dit proefschrift staat vermeld, heeft de inzet van vele mensen tot dit resultaat geleid. Daarom wil ik een aantal mensen bedanken.

Allereerst wil ik graag mijn promotor Lutgarde Buydens en co-promotor Ron Wehrens bedanken. Ik heb veel van jullie geleerd. Misschien wel het belangrijkste daarvan is om kritisch naar mijn eigen werk te kijken.

De mensen met wie ik tijdens de verschillende projecten heb samengewerkt: Leo Meerts, Rene de Gelder en Hans van Sprang. Doordat elk van ons zijn (vak)specifieke steentje bijdroeg, konden we samen vele problemen oplossen. Bedankt hiervoor.

De vaste staf van de afdeling analytische chemie: Willem Melssen, en Geert Postma. Bedankt voor de ruimte die er altijd was voor vragen en hulp. Brigitte Lamzari, hartelijk bedankt voor de o zo belangrijke secretariële ondersteuning.

De mensen met wie ik een kamer deelde: Han Witjes, Theo Reymers, Arjan Simonetti, Ruifen Huo en Than Tran. Bedankt voor de gezellige tijd.

De andere AIO's/OIO's: Erik Swierenga, Philip de Groot, Lammert Kooistra, Uwe Thissen, Jacco Noordam, Egon Willighagen en postdoc's Fabien Zsabo en Velitchka Mihaleva. Bedankt voor de gezellige tijd. Lammert, bedankt voor de uitnoding voor je mooie bruiloft. Het was de dag dat Jacqueline en ik elkaar leerden kennen, en die dag zullen Jacqueline en ik nooit meer vergeten!

De studenten die mij geholpen hebben met een scriptie of een stage: Wouter van Bommel, Micha Streppel en Egon Willighagen (toen bij ons als student). Bedankt voor jullie hulp.

En alle andere mensen die er waren maar wiens naam ik hier vergeten ben. Jullie ook bedankt.

Bij het bedanken van alle mensen die bij mijn proefschrift betrokken zijn

geweest, kunnen mijn ouders niet ontbreken. Zij hebben de basis gelegd die mij in staat stelde dit proefschrift te voltooien. Pap, mam, bedankt hiervoor.

Lieve Jacqueline. Het was fijn om jou aan mijn zijde te hebben tijdens het werken aan dit proefschrift. Bovendien, de wetenschap dat jij in verwachting was, gaf mij de juiste hoeveelheid energie om dit proefschrift af te ronden. Samen met jou en onze zoon Joris zie ik een mooie toekomst tegemoet.

Gegroet, Jos.

Curriculum Vitae

Joseph Albert Hageman werd geboren op 10 april 1974 te Haarlem. Na het behalen van zijn VWO diploma, begon hij in 1993 aan de studie Farmacochemie aan de Vrije Universiteit te Amsterdam. Hoofdafstudeerrichting was de synthese van biologisch actieve verbindingen. Nevenafstudeerrichting was chemische informatica, waar neurale netwerken bestudeerd werden. In 1998 werd hij aangesteld als onderzoeker in opleiding (OIO) bij de vakgroep Analytische Chemie aan de Katholieke Universiteit Nijmegen. Het daar uitgevoerde onderzoek staat beschreven in dit proefschrift.