Chapter 2.5

DRASTIC (Diffuse Reluctance Absorbance Spectroscopy Taking in Chemometrics): A Rapid, Remote-Sensing Method for Assessing Metabolite Overproduction in Titer-Improvement Programs

Douglas B. Kell,* Mike Winson, Roy Goodacre, Bjørn Alsberg, Alun Jones,
Andy Woodward & Jem Rowland†
Institute of Biological Sciences and †Dept. of Computer Science,
University of Wales

I was really asked to talk about remote sensing, and indeed I will introduce this talk through the technology and some of the activities of the remote sensing community, in order to convey to you what obviously I believe: that by hijacking for our purposes the sort of stuff they do with their rather extensive budgets, we can actually buy ourselves some very useful technology and insights into how we can analyze complex biological systems.

First, my credit slide at the beginning (Figure 1). A colleague, Jem Rowland, in our Computer Science Department and I collaborate in this general area (Figure 2), and the aim of much of what we try and do is to acquire as many different data types as possible from the same object and crunch the numbers to turn those data into information (Figure 3). So today I am really talking about infrared spectroscopy and the folk who have done that mainly are Mike Winson and Roy Goodacre. The numbers have been crunched by Andy Woodward and Bjørn Alsberg and Alun Jones. This is in fact a collaboration with Glaxo Wellcome and Bruker Spectrospin, and the Glaxo people are from Tony Buss's group: Martin Todd, Brian Rudd and Mike Dawson.

To start then with the essential take-home messages: there are many techniques you can read about in the pages of analytical chemical journals,

* To whom all correspondence should be addressed.

including some such as CT scanners and magnetic resonance imaging. Geoffrey Hounsfield got his Nobel prize (in 1979; http://www.nobel.se/laureates/medicine-1979-press.html) not in fact particularly for that but for the radon transform that allowed them to take the data and turn it into the pretty pictures you see in CT scans. There is thus a history of using these sorts of heavy signal processing methods on spectroscopic data in biology. But I think there has been less of a trend to use these complex or modern data processing methods in many other kinds of spectroscopies perfectly well known to our medicinal chemists. My argument is, and I hope to illustrate it with examples, that we can hijack these methods too, and apply them in much more complex biological systems in which we are interested and extract important and useful information and get answers not just data.

In particular, although chemometrics (Figure 4) is not a subject to discuss in public at all, let alone after lunch on the last day of a conference, I want to make one or two points about the general sorts of methods that tend to work better, which means the supervised methods, and that is a relatively new game in town. People recognized that these modern methods can extract information from the data in a way where traditional data processing methods in fact fell over rather horribly and didn't do very well at all. This sort of area is usually called chemometrics. You could call it signal processing, of course; nowadays it is often referred to bioinformatics, although chemometrics was coined about 25 years ago so it does have some historical precedence. The major journal in the field *Chemometrics and Intelligent Laboratory Systems* started in 1986, so it is perfectly reasonable to call these methods chemometrics.

The usual definition is the application of mathematical and statistical methods to chemical data, though of course chemical data can come from biological organisms. Although there are many in the Artificial Intelligence (AI) community who would not accept this statement, I will also say that this therefore includes any type of AI-based approaches to extracting information from complex vectors and matrices.

The analogy I like to use about chemometrics (Figure 5) is that it is a tool box. There is a huge zoo of methods. We try not to let the animals escape too often. You can see partial least squares, discriminant function analysis, hierarchical cluster analysis, genetic algorithms, genetic programming, principal components analysis, principal components regression, denoising, classification and regression trees, fuzzy rules, multivariate induction and many more. The point is that the tool box is very rich, it contains a lot of useful tools, and what we have to learn to do is to apply the right tools to extract the information from our data. In particular, these tools coming into their own when we deal with highly multivariate data.

The general area of multivariate data is covered in Figure 6. In the jargon, the samples are called objects and the characters are called variables. In many cases, traditional methods cannot be applied to circumstances in which the number of variables is greatly in excess of the number of objects. That circumstance is often (probably generally) true for spectroscopic data, where you often have relatively few samples. Of course, in the HTS community it rings less true, but the point is that you have huge numbers of characters, particularly spectroscopic characteristics, from each object.

One way of looking at these huge number of dimensions of data we have is to think that for each variable — which might be an absorbance at a certain wavelength — a sample has a certain value, and it therefore can be said to have a position in *n*-dimensional hyperspace for each of those *n* values. In pyrolysis mass spectrometry (which I am not going to discuss here) we have 150 dimensions in the way we typically use them. The infrared data I will talk about today have 882 variables. The key point of this is that **by definition**, such very high dimensional data and methods must have a high resolving power, and a much higher resolving power than the univariate or oligovariate methods more traditionally employed.

There are two major ways in which we analyze these data. First, the so-called unsupervised methods (Figure 7), which are the more traditional ones in which we might have a series of spectra, we extract some features we may be interested in, and we look to see how alike the spectra area. When an unknown spectrum comes along, we say, "which one does it look like most?." After deciding we therefore say it probably means that the sample was like it. The big disadvantage of those types of methods is they are totally undiscriminating. Every variable is effectively given equal weight, whether or not it contributes to the discrimination of the characters in which you are interested.

What the big new game in town (effectively—not so new in some areas) is the concept of supervised learning methods (Figure 8) in which we have our spectra and we have some knowledge of what the characters of interest are: the biological characters, the name of an organism, how much stuff is it making, which target in the cell have I hit, whatever it is that we are interested in which we have measured by some rather tedious method which I will call "the hard way." We want to relate the spectra to the thing we are interested in, using a mathematical model known as a multivariate calibration model. But we do it iteratively so that we muck about with the model mathematically until we can take the spectra, feed them through the model, and it gives us the answer we are interested in.

Then, of course, we come along with new spectra the thing hasn't seen, and we can tell whether it has done a good job or not. Under these circumstances, the methods essentially ignore the irrelevant variance in the data we have and concentrate on the variance that matter for the purposes we are interested in. To cut a long story short, this is a new game in town and it is an incredibly much more powerful way of extracting the information from the data we have got.

Let's go back now to chemometrics and why we are obviously going to need these mathematical and computer methods to extract the data from these high dimensional spaces. We are going to look at the remote sensing side of things, and I might start here with a quotation from Sir Alistair Hardy (Figure 9), who was a Professor of Zoology when I was at Oxford, and who latterly retired to run a Religious Experience Research Unit. He was on a trans-Atlantic flight, and noticed that the colors of the sea varied in different places. He recognized that it was probably because the numbers of chlorophyll-containing micro-organisms varied in different places. Since fish can eat these things and get nutrients from them, this might well be a good idea, if you could only understand what that meant, you could then tell your fishermen where to go and fish.

So the idea of essentially remote sensing of chemical and biologically relevant information begins to happen and, with rather expensive satellites of course, turns into reality. The object in Figure 10 is known as a Coastal Zone Color Scanner. I don't exactly know the price of these things, but that is probably the right decade to be in. The Coastal Zone Color Scanner has been flying around for some time, and it comes up with pictures like the one in Figure 11, chemical images (in which in the key there is a high concentration at the bottom for some reason). It is essentially relying on the spectral properties of chlorophyll, and you can see that the up-wellings in the shallow regions are indeed connected with high amounts of microbial productivity and out in the deeper part of the ocean, the productivity is relatively low.

Remote sensing methods can be and have been used to affect chemical imaging in a way we know quite well (Figure 12). Another nice example well known to you all I am sure is the observation of the ozone hole (Figure 13). This is done in the ultraviolet where ozone has a particular absorbance, and we all know about the ozone hole in 1992.

These sorts of traditional methods, though, Landsat and Spot for instance, the satellites you have certainly heard of, use five and seven wave bands respectively because they weren't very good at collecting photons. The human visual apparatus, of course, uses three different types of cone, like television sets: red, green, and blue. Most of these things thus use very small numbers of wave bands to produce the chemical information one is interested in.

What happens if instead of using these very small numbers of wave bands, five in this case, we use hundreds of wavelengths simultaneously (Figure 14)? Do we buy ourselves anything apart from a lot of grief? The effective answer is yes, because one way to consider this is that there are many things with the same spectrum as those where there are only five variables, but there is only one thing that has the same spectrum as when there are, say, 224 variables. Again, we come back to the point that the high dimensional data are by definition intrinsically more highly resolving than the low dimensional data, and this isn't just hypothetical.

There is a thing that has been built called AVIRIS, a predictable acronym (Figure 15), which flies around collecting data in, in fact, 224 bands between 400 nanometers and 2.5 microns, i.e., the visible and the near infrared, collecting spectra from each location. This is known as imaging spectroscopy, so you are collecting a spectrum from every place in space. The game then of course is to take these spectra and ask what was it that was on the ground? Obviously it is very easy to discriminate water and trees and so forth (Figure 16), but one would hope to be much more discriminating. How can we use these hyperspectral data (as the jargon has it; Figure 14) to go from the data we are not particularly interested in at all to the biology that we are interested in?

What comes out, then, of these types of approaches, are datacubes like that in Figure 17; it bears looking at this picture for a couple of seconds to get one's mind around it. We are always going to have a 2-D image in the XY plane, and in the Z plane effectively an entire spectrum, here encoded by intensities as colors in which at the top we are talking about 400 nanometers and at the bottom 2.5 microns. So we have a data hypercube in which the XY plane is indeed the XY plane, and the Z plane is the spectral plane with hundreds of absorbances at once.

When you use these sorts of things, you can come out with very pretty pictures (Figure 18). All of these have been compared with what the remote sensing people call the ground truth which is, "if I get in there with a bucket and spade, what do I find?" It gets it all right, all these different kinds of minerals can be remotely sensed (Figure 19). There is an example now with vegetation (Figure 20), with a nice tale about how the European Union used this to advantage to discover that a particular Mediterranean country had not in fact grown 4.2 million tons of durum wheat as it claimed, but only 1.7 million tons. There was quite a big scandal about that, apparently.

The point is that one can indeed use the spectral data from reflectance measurements. The sun, of course, in this case is the source of the photons, the photons are going to hit the ground, rattle around in the target, and bounce back to where the satellite is going to see them. It is diffuse reflectance, and of those that come back, some of them have been absorbed, and therefore that is why it is diffuse reflectance absorbance spectroscopy and the DRASTIC acronym. We are going to use this, if possible, to give ourselves chemical information about complex biological systems for pharmaceutical systems in which we are interested. I tend to be interested in natural products and microbiological systems.

There are two problems with the remote sensing approach, however. First is, when it goes wrong, it goes wrong spectacularly (as in the Ariane rocket exploding). The second problem is you have to use the visible and mid-infrared really, because although you might like to use the mid-infrared, it is almost entirely absorbed by the water and the CO_2 in the atmosphere (Figure 21). You can use this imaging spectroscopy in the lab to get around the first of those, but again they are using the visible and very near infrared spectrum. What we want to do is recognize that in the mid-infrared there are huge amounts of chemical specificity, as I am sure most of you know. Every molecule has a rich infrared spectrum that is due to normally vibrational modes of the individual chemical bonds in the molecules of interest.

In the laboratory, we don't have this problem of water absorbance and CO_2 absorbance that the remote-sensing people have. We then can probably hijack this sort of approach in the laboratory to do diffuse reflectance absorbance spectroscopy in the infrared for, as it says, rapid noninvasive chemical analysis. When you combine infrared spectrometry in the lab, where we happen to collaborate with Bruker (Figure 22), obviously the price is a bit more favorable relative to the coastal zone color scanner.

An important point to make, although I am not going to discuss it in detail here, is that miniaturization is very straightforward because all the manufacturers have microscopes as well (Figure 23). Because of the wavelengths we are talking about, you tend to be able to go down to about only 10 microns direct spatial resolution, and so that is a very straightforward thing.

I am going to talk about something rather more mesoscopic today. This is based on an attachment that Bruker markets for looking at thin layer chromatography plates (Figure 24), in which they have implemented diffuse reflectance. So there is a source of infrared photons, various optics, and a detector and it bounces off the target on the TLC plate. The thing is mounted on an XY stage so that you can effectively fly over the thing doing what we want to do. But instead of having used a TLC plate and done a bit of chromatography first (which we are not very interested in), we simply use it as a sample presentation device so that we can whack all our different samples into little wells on this plate (Figure 25), as obviously we have made various 10 × 10, 100 × 100 and so on (nowadays standard 384-well plate coordinates in fact). You are talking then about typical 20 μl down to 5 μl sample volumes.

Now, I describe a couple of studies we have actually done using this diffuse reflectance approach—one a relatively easy one, as it were, and one a relatively hard one. The easy one (Figure 26) is a simple model in which we are just going to mix up what I will call drug and bug, and see if the thing can estimate the drug in a background of the bug, because if it can't do it with mixtures of known drug plus known bug, then it is clearly going to fail when you get to something much more difficult. So the model bug is *E. coli* and the model drug is ampicillin, and we can collect a lot of spectra.

Note, for instance, the trace with the lowest concentration of ampicillin in this case, none, does not have the lowest reflectance (Figure 26). That is because even at the National Bureau of Standards you can't make a plate that is perfectly homogeneously reflecting, and even if you could you are only going to cover it with gloop anyway; you needn't have bothered in the first place. Chemometrics isn't going to mind that. It is worth noticing that dealing with baselines is in principle going to be a problem. Note the peak band, which is in fact at 1767 cm^{-1}. It is known that this band at 1767 cm^{-1} is actually the carbonyl stretch (Figure 27), but the carbonyl stretch is slightly different from where it would normally be because it is constrained in the beta-lactam ring of the ampicillin moiety.

You can show that in two ways, the main one of which is by zooming in and looking at it as an amount of stuff (Figure 28). Clearly, just sitting at one univariate wavelength and measuring the absorbance would fall over horribly, (i) because of the changes in the baseline I have told you about, and (ii), because the absorbance itself is evidently by inspection by eye not wonderfully linear with the amount of stuff. You could also show the peak is due to what I say by adding lactamase and it disappears.

If you then say, okay, let's just integrate the band (Figure 29). We know this band is essentially selective for the ampicillin in this case, and you ask the question, if I integrate this band and look at the distribution of absorbance over space in my 100-well plate you see something like the map in Figure 30. We clearly loaded this so that there is none down at the bottom and plenty up top, and clearly it gets it all right. But we would like to be quantitative, and the way we like to do that is by using whole spectral methods rather than the rather unimaginative band integration methods you tend to be able to get from people with commercial software and spectrophotometers.

After this, we are going to use neural networks and we are also going to use partial least squares. As you know, what you have in a typical back propagation neural network is an input, a hidden layer and an output. The spectral data (not, in this case, PyMS but the infrared absorbances) are going to be on the input. What these lines—known as weights—do is multiply the inputs by numbers and the circles—known as nodes; add them up and multiply them by a sigmoidal function and spit out a number. You muck about with the weights using a more or less intelligent algorithm until the spectra you add at the front of the neural network gives you the output you want. When you have trained the neural network appropriately, you interrogate it both with the data you have trained it with—the training set—and totally unseen data in the test set; obviously it has become quite good at giving you a reasonably accurate prediction of the amount of ampicillin in your mixture.

So the information is there. We can use it (Figure 31). Note that because of the sigmoidicity here amongst other things, the predictions at the end are really not very good. Here we are using 882 inputs, that is the absorbances at all of those 882 bands, which are the wavenumbers between about 4,000 wavenumbers and 600 wavenumbers with four-wavenumber resolution.

We can be a little more intelligent than that. I am not going to go into too many chemometrics methods, but just give you a flavor of the sort of things one can do to improve things. If I extract what are called the principal components, which is an unsupervised method, and ask for each of my samples, what happens if I input the value of the first several principal components to the net, rather than 882 variables? A) of course, it trains one heck of a lot quicker, and B) we can ask the question how many of these principal components should I use? The answer appears to be nine when we look at the test set (Figure 32) and when we look at the output then, we have cleaned that up very nicely and we have got the straight line anyone would be proud of (Figure 33).

That is ampicillin in *E. coli*. Of course, the question is, what happens when you make it a bit more demanding and throw in something else, in this case *Staphylococcus aureus*? It has learned very well to discriminate the amount of ampicillin in *S. aureus*, and clearly you see a nice straight line. What more can one say?

That was in effect an easy one, where we were adding known mixtures of drug plus bug, and we could more or less follow fairly exactly what we thought was going to happen and make it happen and get a nice straight line at the end. Now we are going to do a hard one (Figure 34), which is looking at a real *Streptomyces* fermentation in what any of you would know what *Streptomyces* fermentations are like—a pretty gloopy broth with all sorts of complex media constituents inside. We are going to be making (or, the bug is going to be making) two carbocyclic nucleosides called aristeromycin and neplanocin A. As I mentioned, this project is a collaboration with GlaxoWellcome (Figure 35).

We are going obviously to see if we can quantify them and discriminate them, and that is quite hard, because chemically they are very similar. The only difference is the chirality here, and there is a double bond there, and there isn't one there (Figure 36). The strategy essentially is as before. First we can just look at the diffuse reflectance-absorbance spectra of the compounds themselves (Figure 37). You don't need me to tell you can gawp at those for a long time and it is not going to do very much for you; only the computer will solve this.

So you grow the cultures, take the broth (this is an organic, wholemeal experiment—nothing added, nothing taken away), and you look at some spectra. Here are four different classes of producer (Figure 38). One makes lots of aristeromycin; one makes lots of neplanocin A; one makes aristeromycin only; and one that makes both. These are typical spectra. The one on the right side of Figure 38 is a blow up of the region between 1700 and 1000 wave numbers. Clearly, there are differences but equally clearly, in contrast with the ampicillin case, there are no unique bands; no univariate analysis could possibly give you a vague chance of extracting the concentration of either the aristeromycin or the neplanocin A. Only the whole spectral, hyperspectral analysis can do that.

What we have to do in all of these things is set up a training set and a test set, but the one thing not widely recognized is what you should choose to be in the training set and what you should choose to be in the test set (Figure 39). In principle, the training sets should encompass the test set, so that it doesn't extrapolate, but if there are 882 dimensions, it is kind of hard to tell that. Usually people just put them in at random. This can cause these methods to fail horribly, and people don't recognize why. To cut a long story short, we (or more precisely Alun Jones in the group) have got around it by writing some software that can recognize these problems and deal with them and make the training and test sets correctly.

So you make the training set, and you look in this case first with the partial least squares method, which is essentially a regression method for when the number of variables is grossly in excess of the number of objects. Happily enough, if you are trying to predict metabolite concentrations, this is aristeromycin *plus* neplanocin A (Figure 40, left); obviously it does jolly well. For aristeromycin only, it also does jolly well (Figure 40, right). It doesn't do as well as we would like in terms of being quantitative, but it does pick all the high-producing strains which, after all, is all it is supposed to be doing.

We do the same with neural nets, the same sort of story, just showing the test set this time as one really should, and there is a similarly nice prediction of the combined, and a nice prediction of the aristeromycin on its own (Figure 41). When we look at the neplanocin A tests—again with PLS (Figure 42) or with neural nets (Figure 43)—it has again gotten all the high yielding strains, although we would be happier if it also got the right answer (Figure 44).

How are we going to deal with that? There are lots of potential ways. The way I will describe to you is that the model is probably getting confused because there is a lot of spectral overlap between the aristeromycin and the neplanocin A (Figure 36). There are not enough objects in the training set anyway in reality in this experiment, and it just hasn't the ability to discriminate them. But if you could tell it which class of producer it was first, you would have a different model then to the aristeromycin only or the neplanocin A only. Then you would have solved the problem because you wouldn't be trying to look at organisms in which you are making both of them. You would only be looking at organisms making one or the other, so you want to make that decision first. This is known as the 'divide-and-conquer' approach.

You encode high aristeromycin only, both, neither, or high neplanocin, and when you do that, you can then just produce the model secondarily. The way we do this is with a discriminant function analysis (Figure 45). The four different classes are represented. If you take away class 1, you don't initially have the ability to discriminate 2 and 3 very easily, but once you have identified 1, you take them out, and you can easily affect the discrimination of the others. Then you have essentially solved the problem because now you only need to make the model for the ones you are interested in.

The conclusion then (Figure 46) is that these whole broth, whole spectral, hyperspectral analyses do have sufficient chemical information to affect the kind of discrimination in which we are interested. In this case, I have concentrated on high yielding strains and titer improvement programs. I could make the same arguments about many other kinds of pharmacological assay of interest.

The nice thing, of course, is it is rapid and non-invasive. I didn't say how rapid. The spectrometer can do twenty a second. Obviously it is noninvasive. There has been talk about reagent costs during this meeting; here, there are none, a slight advantage. There has also been talk to the effect that, okay, we can do 10^5 assays, but no one can provide me with 10^5 molecules from the combichem. But in the high-throughput screening, titer-improvement programs, neither of those statements is true. It really is the more the merrier. If you have 10^5 organisms in a milliliter and observe them in terms of optical density you won't even see them. We are not short of possible numbers of organisms so in titer-improvement programs, it really is a matter of the more the merrier because it is easy to generate the mutants we would want to put through a titer improvement program.

You may well want to ask me about sensitivity and limit of detection. That is somewhat harder to speak about than I would wish because the thing is limited not so much by the intrinsic sensitivity of the infrared machine, which is easily in the nanomolar, but by the variance in the background that could, on a bad day, swamp it. Equally, it can work to your advantage because you are interested in looking at things that correlate with what you are interested in, and it doesn't necessarily mean you have to measure them directly in these cases. So small chemical changes can be amplified, and indeed will be amplified by the cells in a way that you can use those as the spectral features of interest rather than looking directly. So if you were to ask me about limits of detection, I have to say there isn't a simple answer to that.

Acknowledgments

We thank the BBSRC, GlaxoWellcome, and Bruker Spectrospin for partial support of the work described.

QUESTIONS AND ANSWERS

Question: Who uses this sort of technology to sort out which one is which, what have you made, and that sort of thing?

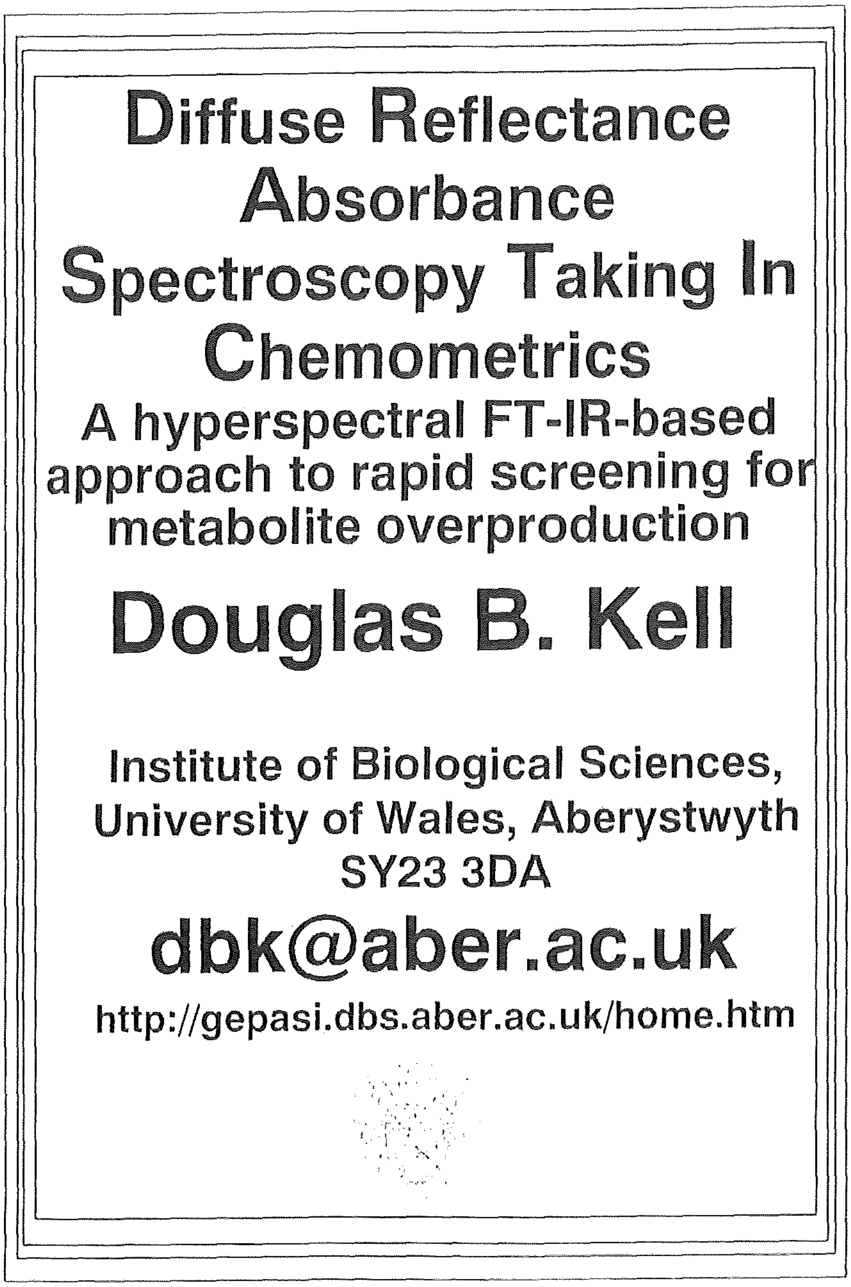
Prof. Kell: There is a group at Merck* that has published several papers on measuring solid-phase combichem beads, essentially saying, yes, we can take microspectroscopic data from individual beads and see that it has an infrared spectrum consistent with what I thought I had made. But I haven't seen any other group yet publishing it.

* Yan, B., Fell, J.B., and Kumaravel, G. Progression of organic-reactions on resin supports monitored by single bead FTIR microspectroscopy. *Journal of Organic Chemistry* 61:7467-7472, 1996.

Yan, B., and Kumaravel, G. Probing solid-phase reactions by monitoring the IR bands of compounds on a single flattened resin bead. *Tetrahedron* 52:843-848, 1996.

Yan, B., Kumaravel, G., Anjaria, H., Wu, A., Petter, R.C., Jewell Jr., C.F., and Wareing, J.R. Infrared spectrum of a single resin bead for real-time monitoring of solid-phase reactions. *J. Org. Chem.* 60:5736-5738, 1995.

Yan, B., Sun, Q., Wareing, J.R., and Jewell, C.F. Real-time monitoring of the catalytic oxidation of alcohols to aldehydes and ketones on resin support by single-bead Fourier transform infrared microspectroscopy. *J. Org. Chem.* 61:8765-8770, 1996.



Diffuse Reflectance Absorbance Spectroscopy Taking In Chemometrics

A hyperspectral FT-IR-based approach to rapid screening for metabolite overproduction

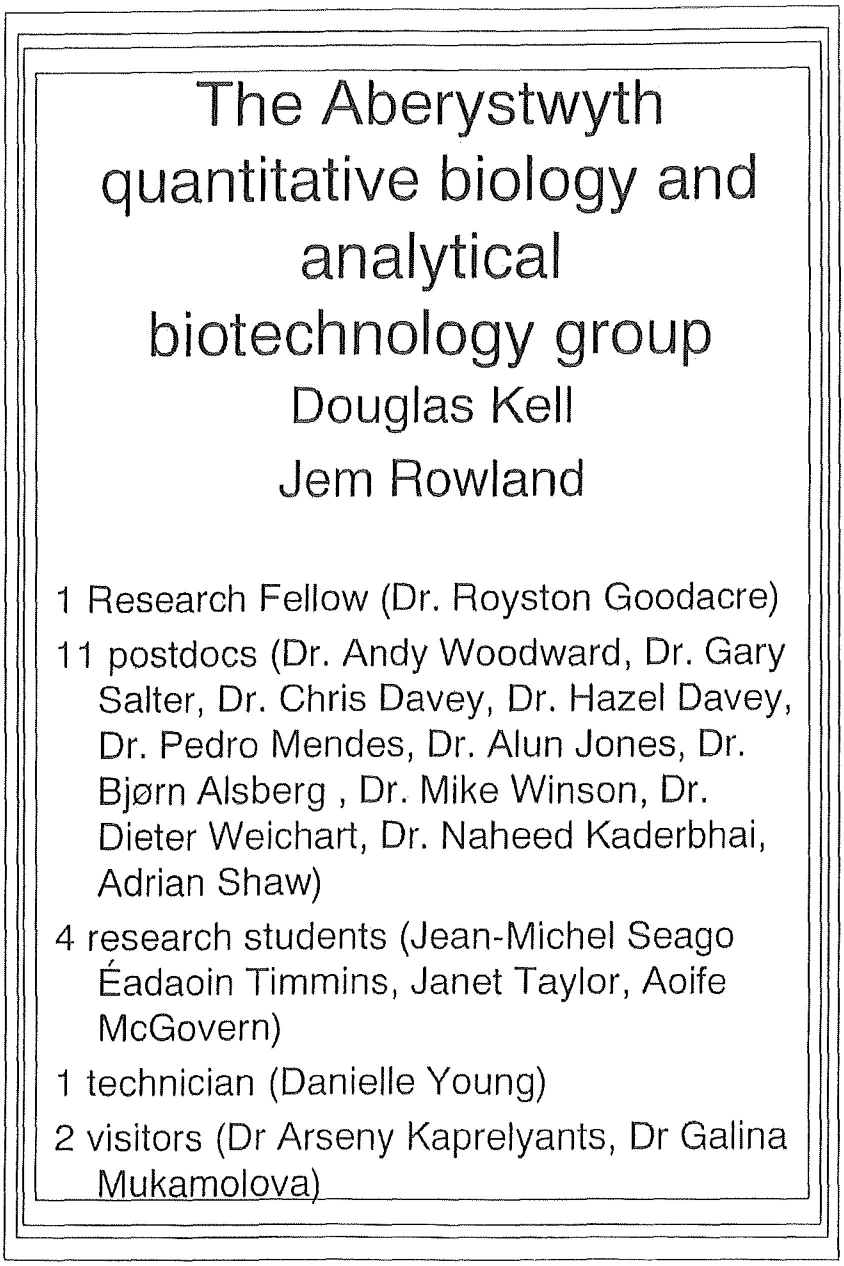
Douglas B. Kell

Institute of Biological Sciences,
University of Wales, Aberystwyth
SY23 3DA

dbk@aber.ac.uk

http://gepasi.dbs.aber.ac.uk/home.htm

Figure 1.



The Aberystwyth quantitative biology and analytical biotechnology group

Douglas Kell

Jem Rowland

1 Research Fellow (Dr. Royston Goodacre)

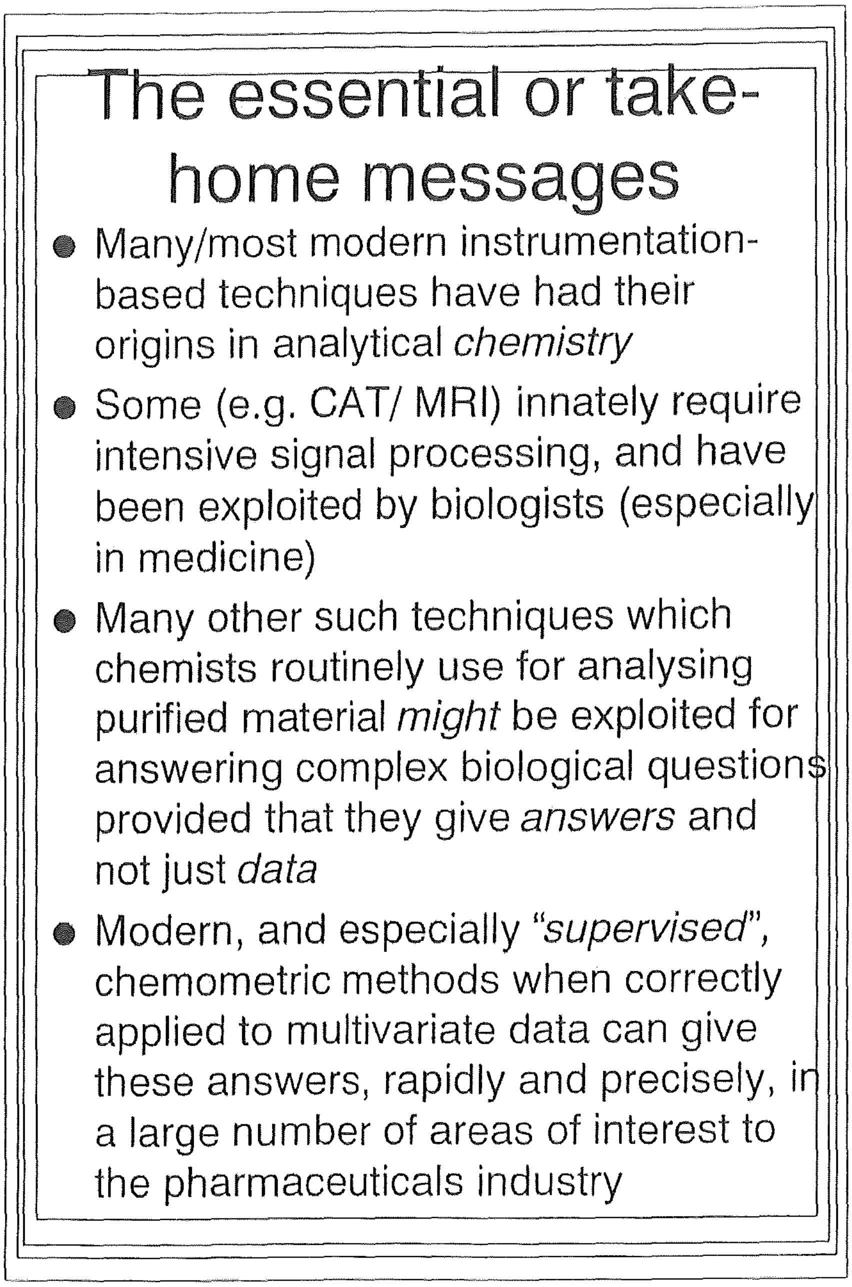
11 postdocs (Dr. Andy Woodward, Dr. Gary Salter, Dr. Chris Davey, Dr. Hazel Davey, Dr. Pedro Mendes, Dr. Alun Jones, Dr. Bjørn Alsberg , Dr. Mike Winson, Dr. Dieter Weichart, Dr. Naheed Kaderbhai, Adrian Shaw)

4 research students (Jean-Michel Seago Éadaoin Timmins, Janet Taylor, Aoife McGovern)

1 technician (Danielle Young)

2 visitors (Dr Arseny Kaprelyants, Dr Galina Mukamolova)

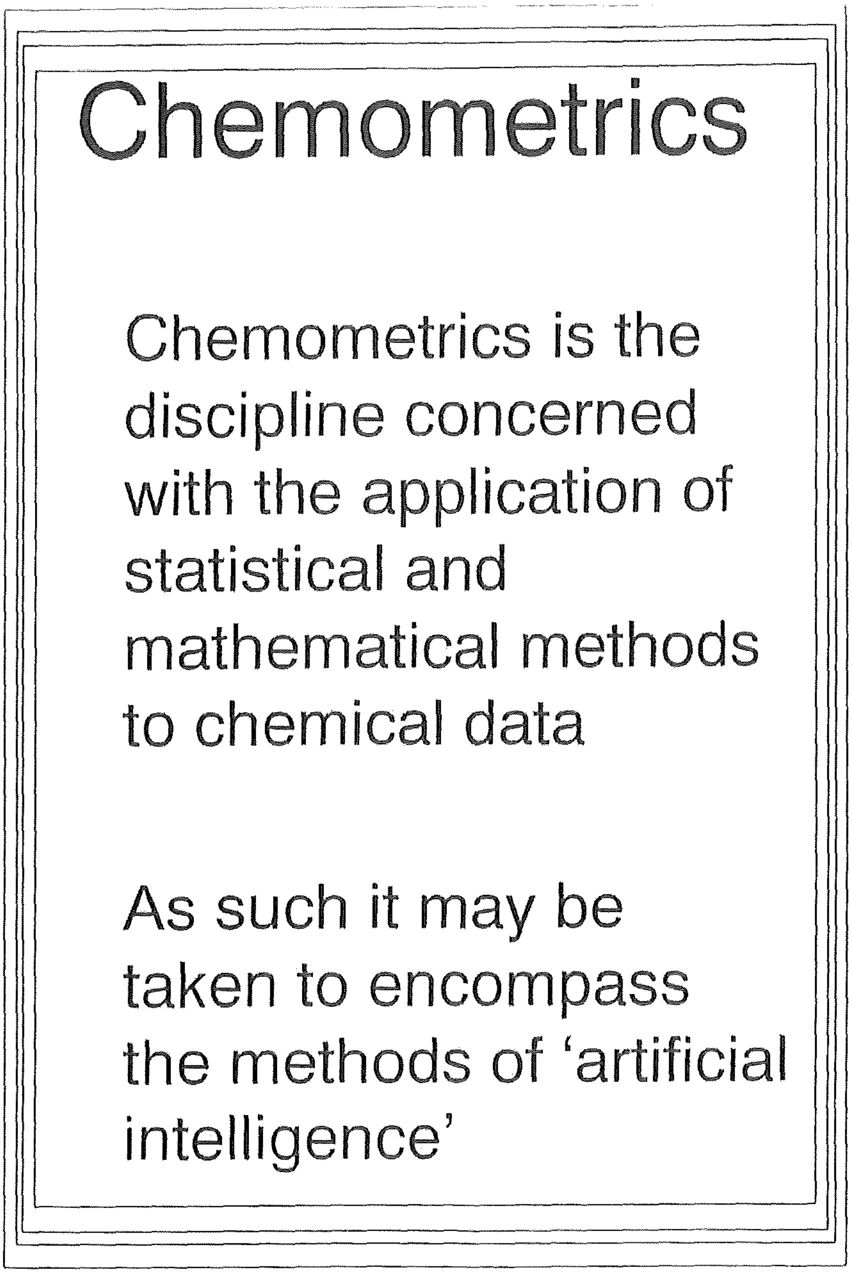
Figure 2.



The essential or take-home messages

- Many/most modern instrumentation-based techniques have had their origins in analytical *chemistry*
- Some (e.g. CAT/ MRI) innately require intensive signal processing, and have been exploited by biologists (especially in medicine)
- Many other such techniques which chemists routinely use for analysing purified material *might* be exploited for answering complex biological questions provided that they give *answers* and not just *data*
- Modern, and especially "*supervised*", chemometric methods when correctly applied to multivariate data can give these answers, rapidly and precisely, in a large number of areas of interest to the pharmaceuticals industry

Figure 3.

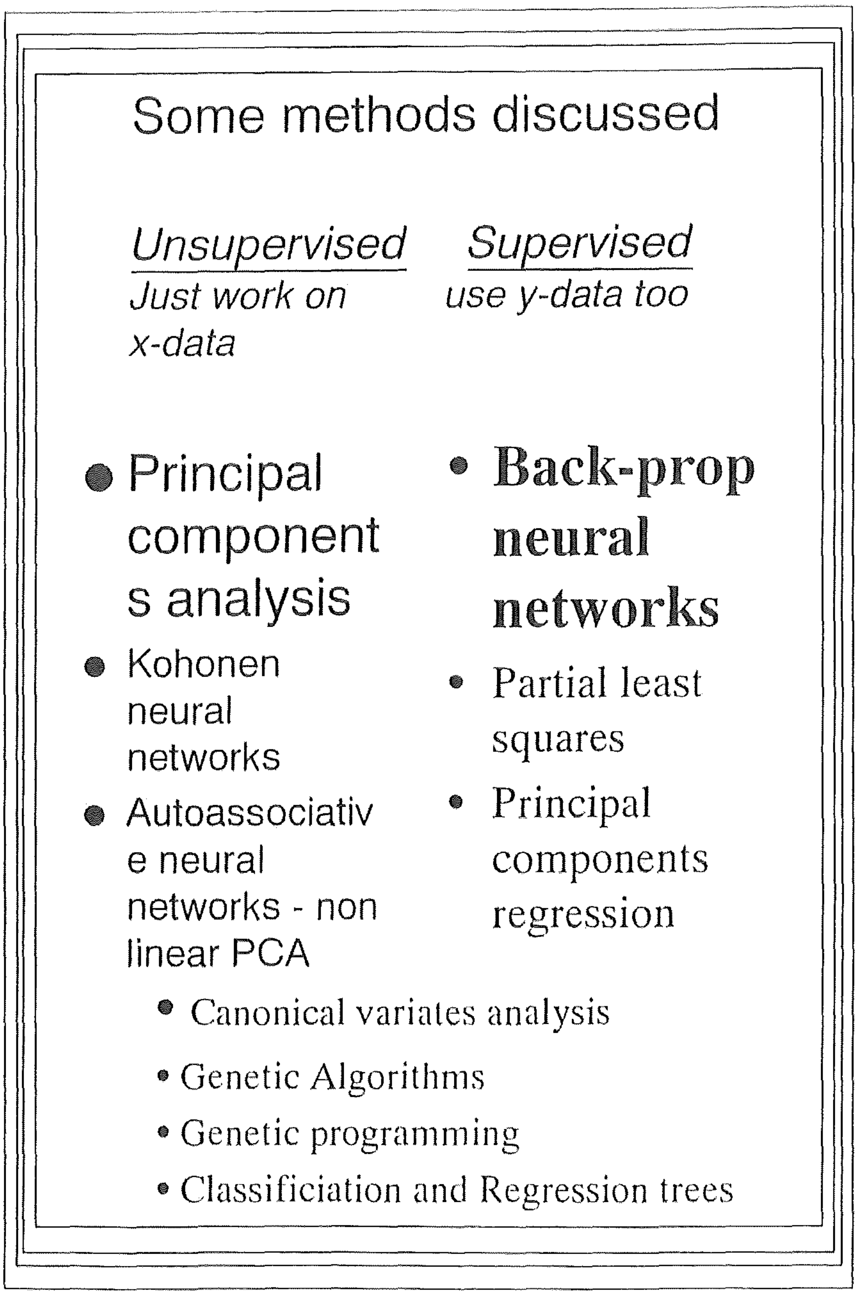


Chemometrics

Chemometrics is the discipline concerned with the application of statistical and mathematical methods to chemical data

As such it may be taken to encompass the methods of ‘artificial intelligence’

Figure 4.



Some methods discussed

Unsupervised
Just work on x-data

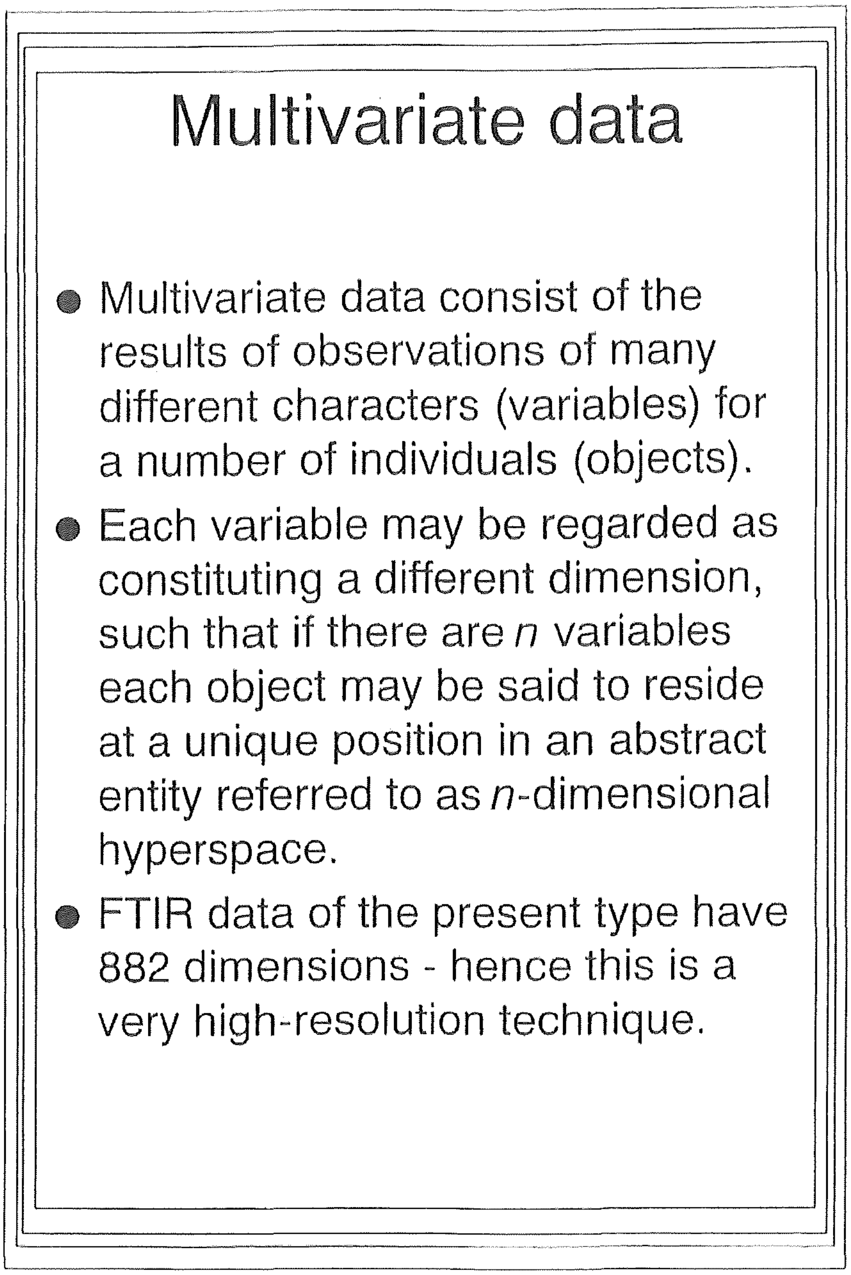
- Principal components analysis
- Kohonen neural networks
- Autoassociative neural networks - non linear PCA

Supervised
use y-data too

- **Back-prop neural networks**
- Partial least squares
- Principal components regression

- Canonical variates analysis
- Genetic Algorithms
- Genetic programming
- Classificiation and Regression trees

Figure 5.



Multivariate data

- Multivariate data consist of the results of observations of many different characters (variables) for a number of individuals (objects).
- Each variable may be regarded as constituting a different dimension, such that if there are *n* variables each object may be said to reside at a unique position in an abstract entity referred to as *n*-dimensional hyperspace.
- FTIR data of the present type have 882 dimensions - hence this is a very high-resolution technique.

Figure 6.

Unsupervised learning

The system is shown a set of inputs (spectra) and then left to cluster the spectra into groups. For multivariate analysis this optimization procedure is usually for simplification or dimensionality reduction. This means that a large body of data (the spectral inputs) are summarised by means of a few parameters with minimal loss of information. After clustering, the results then have to be interpreted.

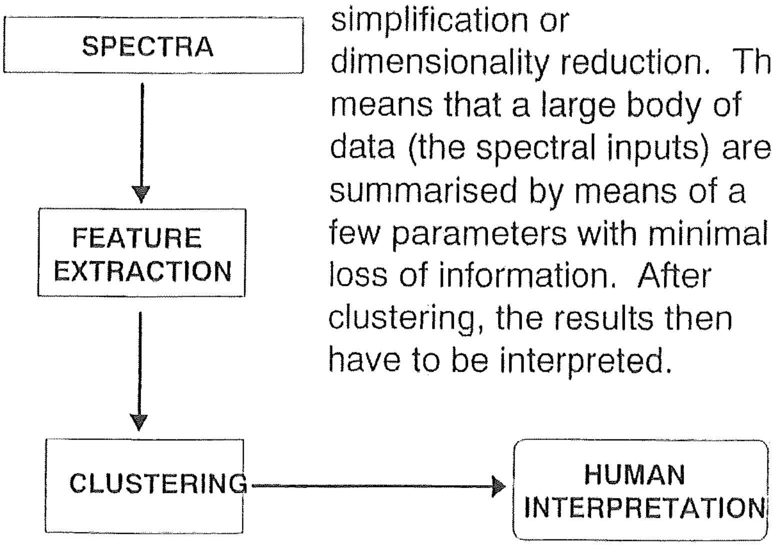
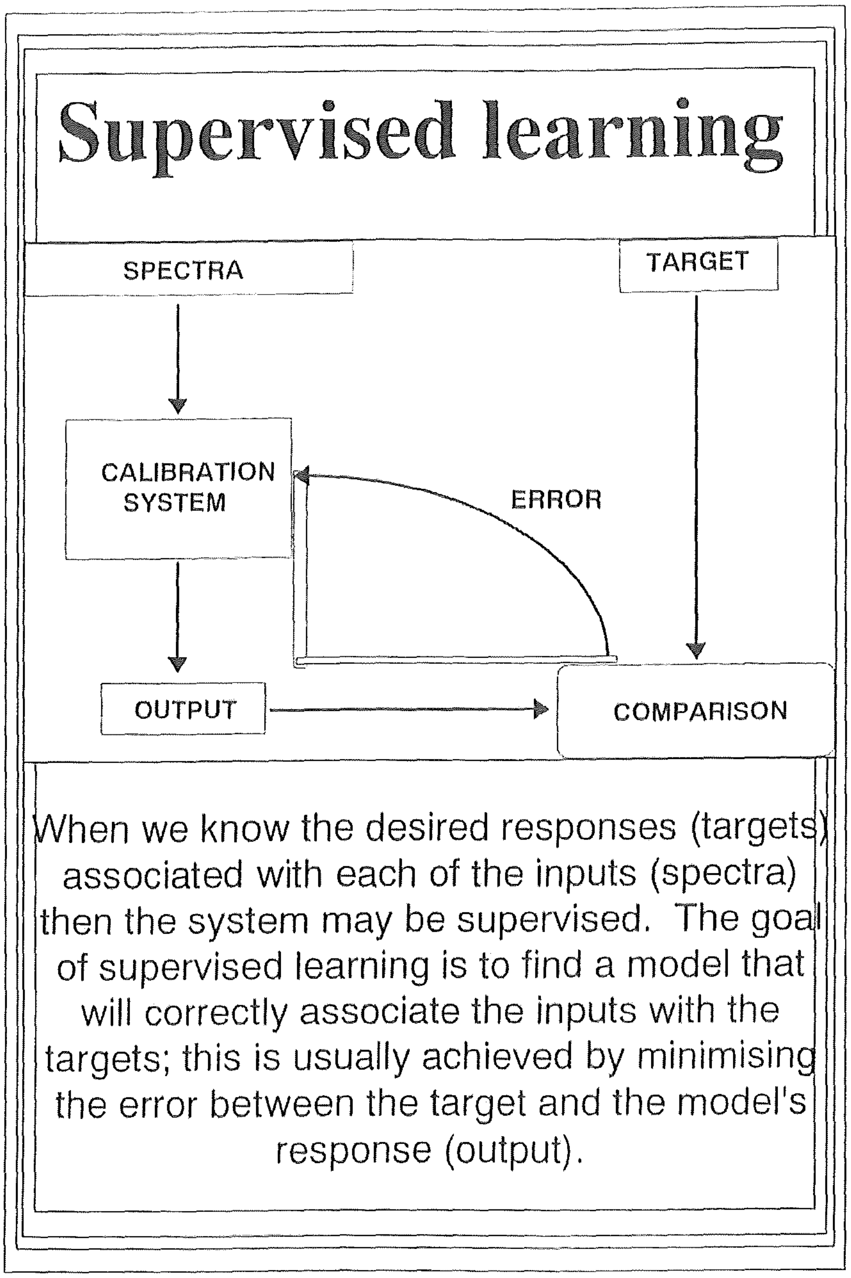


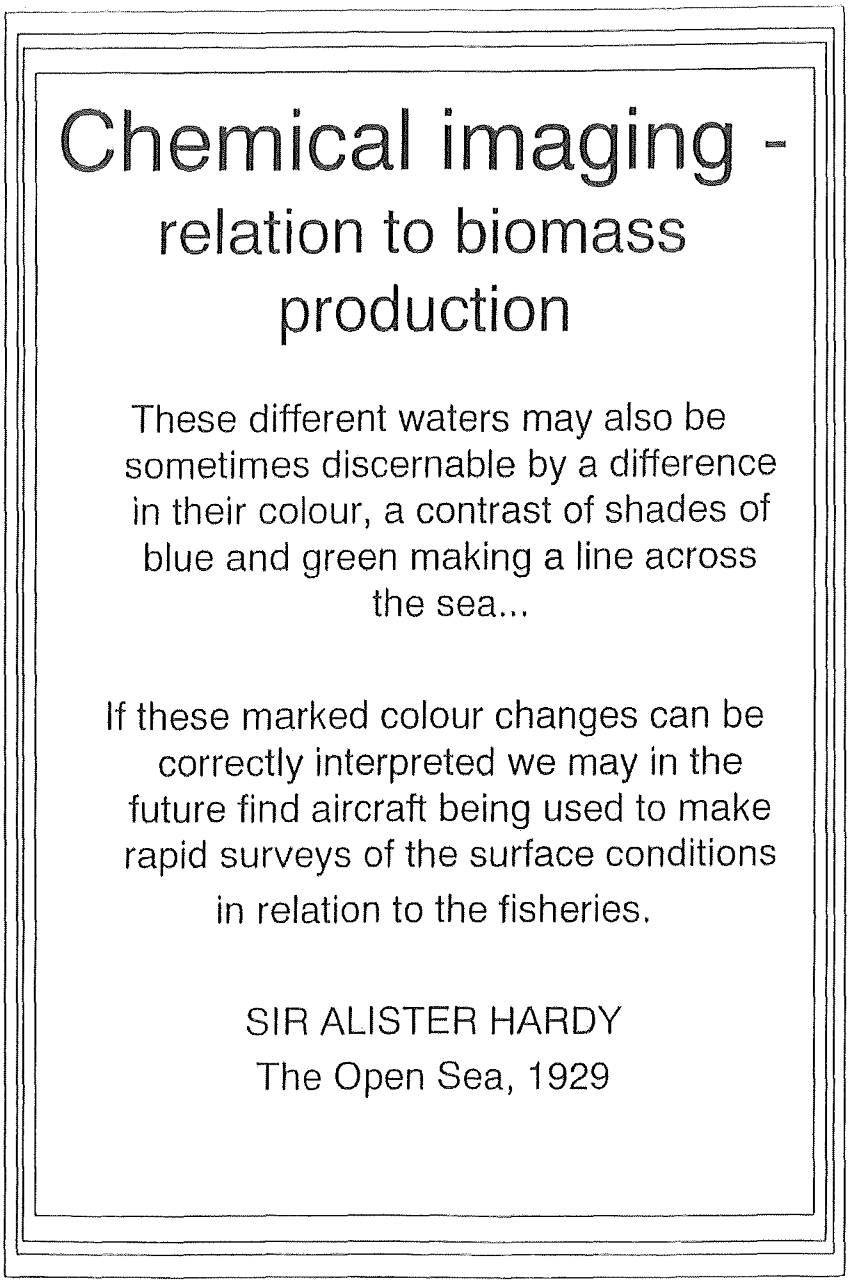
Figure 7.



Supervised learning

When we know the desired responses (targets) associated with each of the inputs (spectra) then the system may be supervised. The goal of supervised learning is to find a model that will correctly associate the inputs with the targets; this is usually achieved by minimising the error between the target and the model's response (output).

Figure 8.



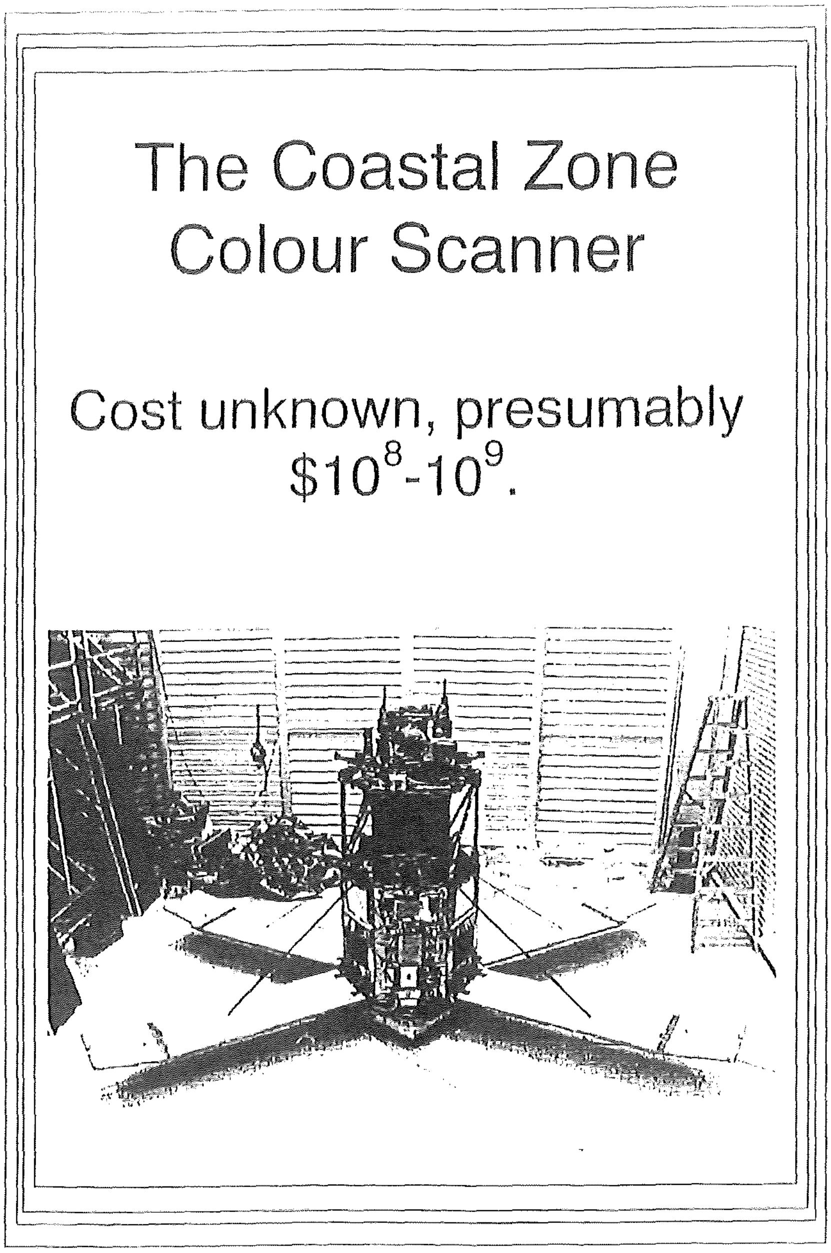
Chemical imaging - relation to biomass production

These different waters may also be sometimes discernable by a difference in their colour, a contrast of shades of blue and green making a line across the sea...

If these marked colour changes can be correctly interpreted we may in the future find aircraft being used to make rapid surveys of the surface conditions in relation to the fisheries.

SIR ALISTER HARDY
The Open Sea, 1929

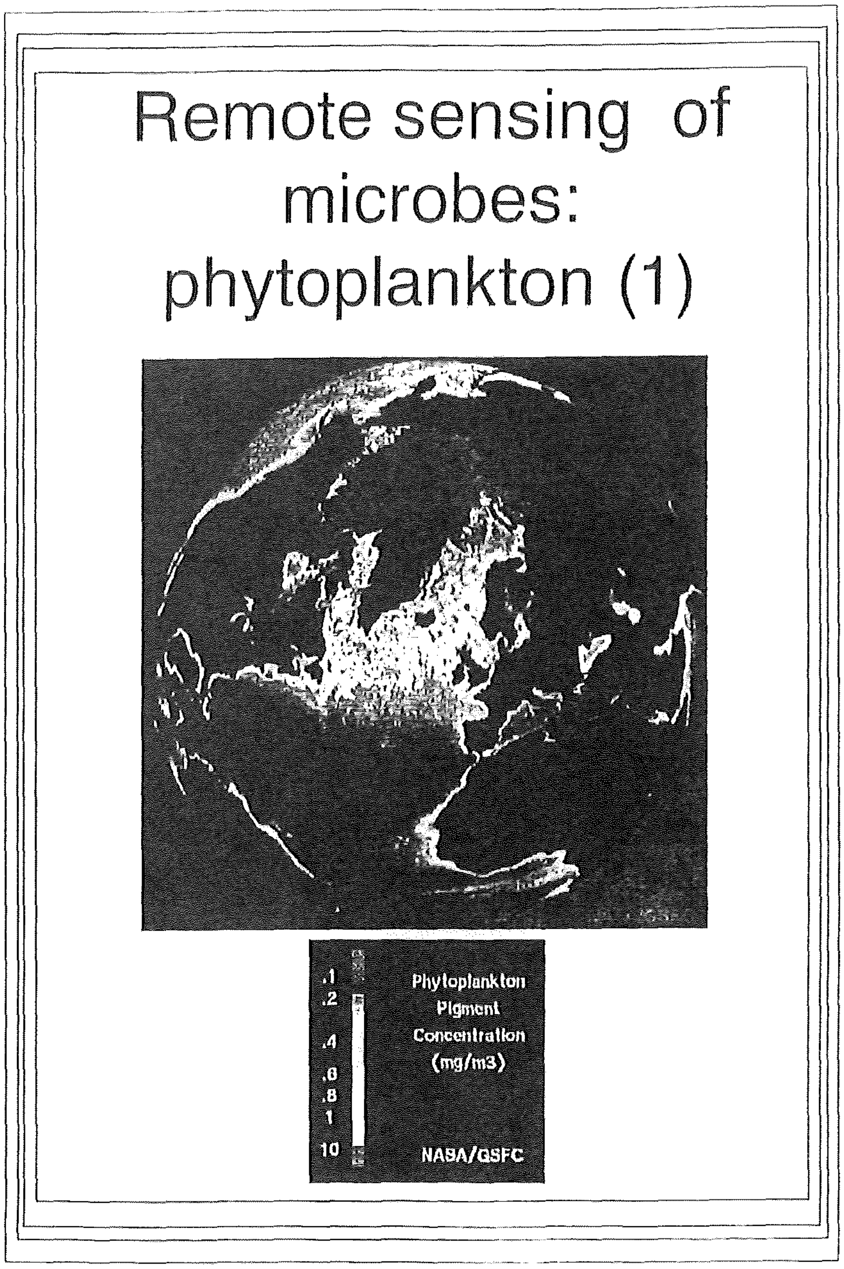
Figure 9.



The Coastal Zone Colour Scanner

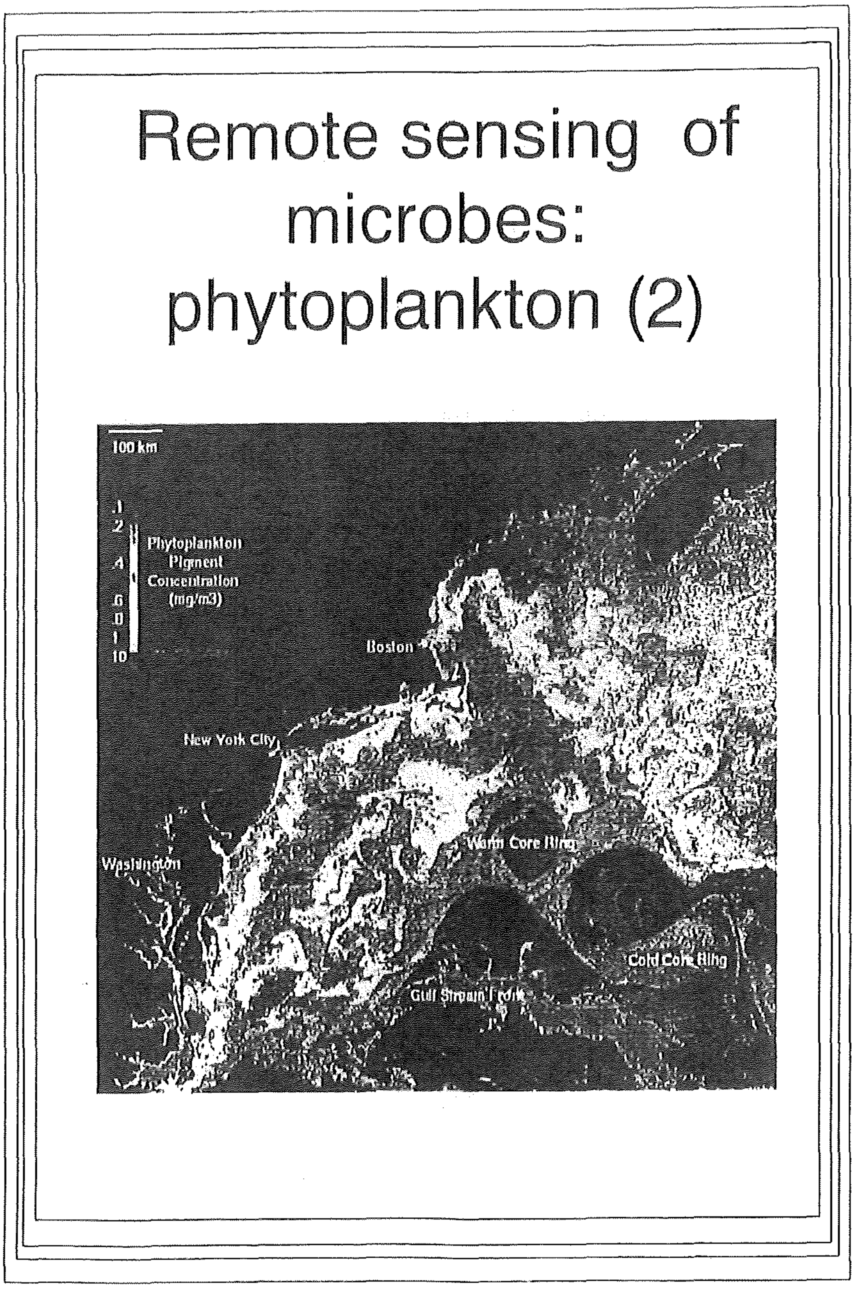
Cost unknown, presumably $\$10^8$-$10^9$.

Figure 10.



Remote sensing of microbes: phytoplankton (1)

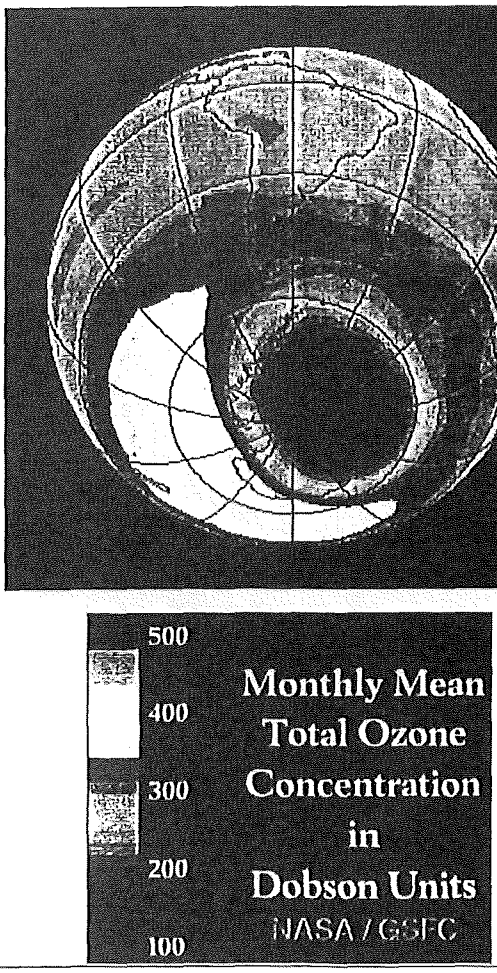
Figure 11.



Remote sensing of microbes: phytoplankton (2)

Figure 12.

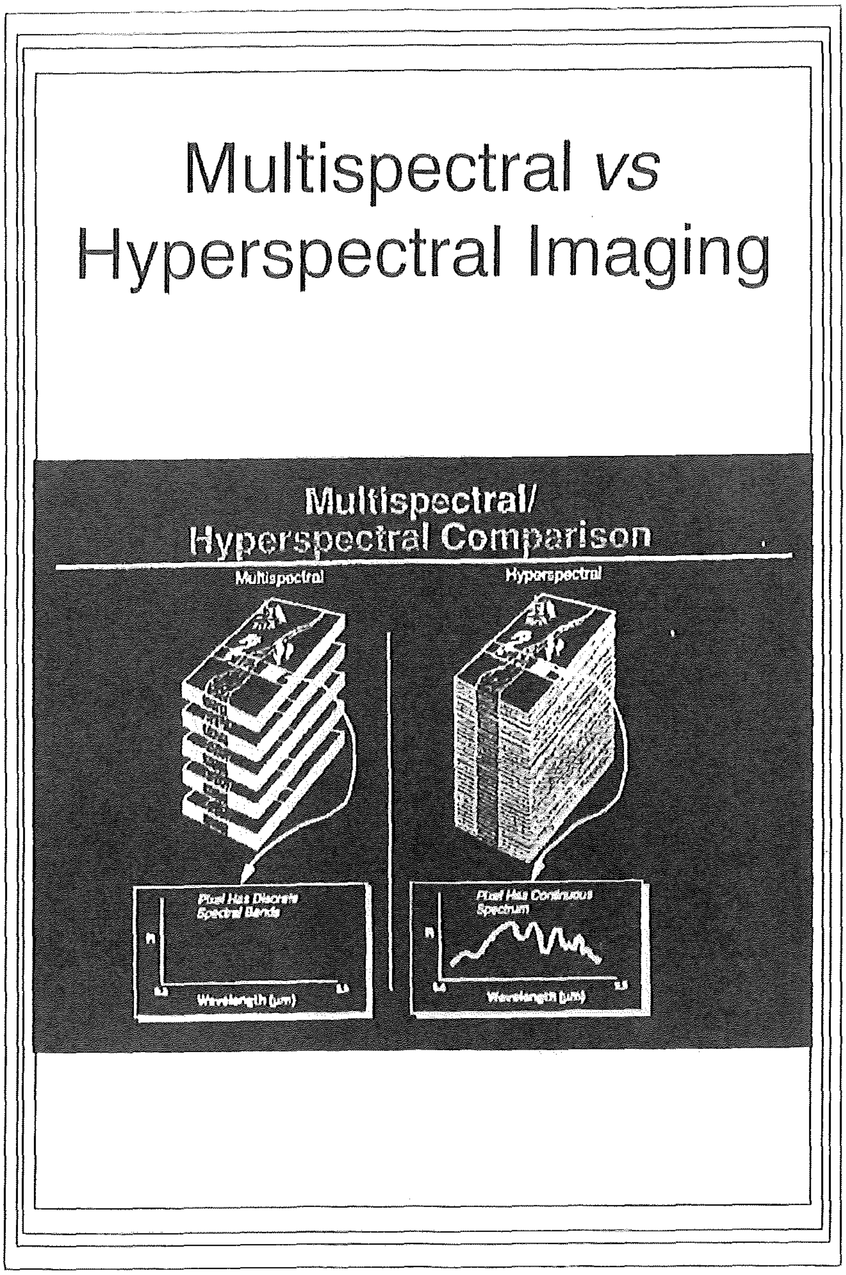
Ozone distributions
1992



Monthly Mean Total Ozone Concentration in Dobson Units

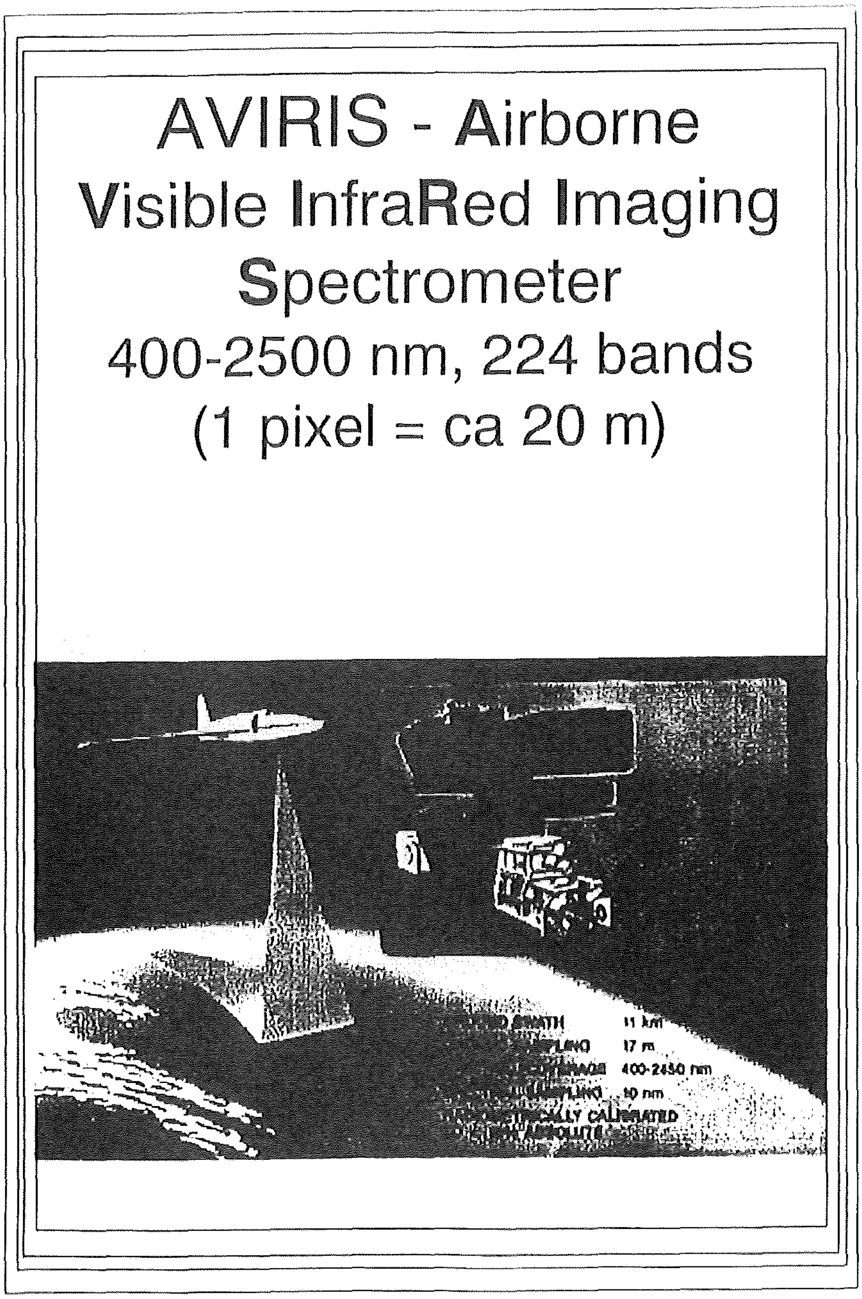
NASA / GSFC

Figure 13.



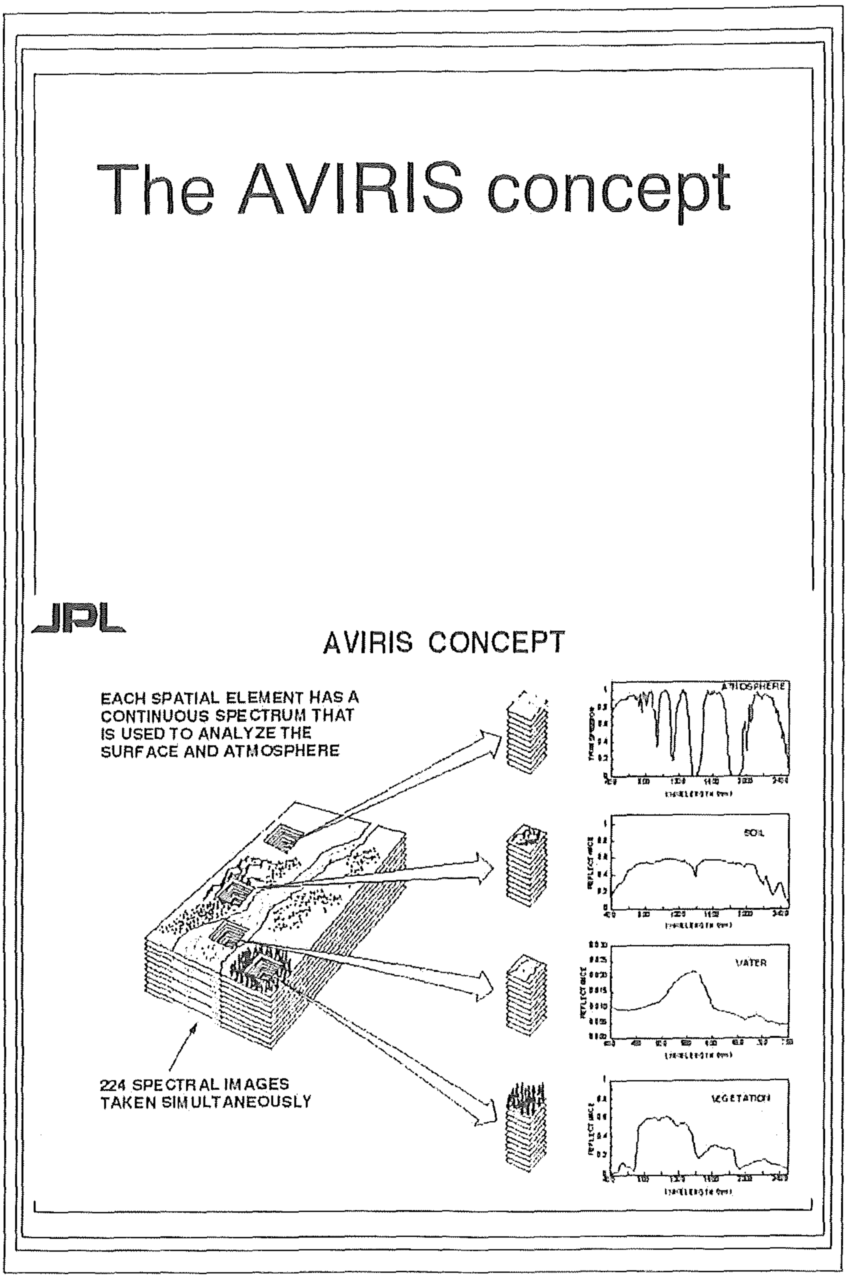
Multispectral *vs* Hyperspectral Imaging

Figure 14.



AVIRIS - Airborne
Visible InfraRed Imaging
Spectrometer
400-2500 nm, 224 bands
(1 pixel = ca 20 m)

Figure 15.



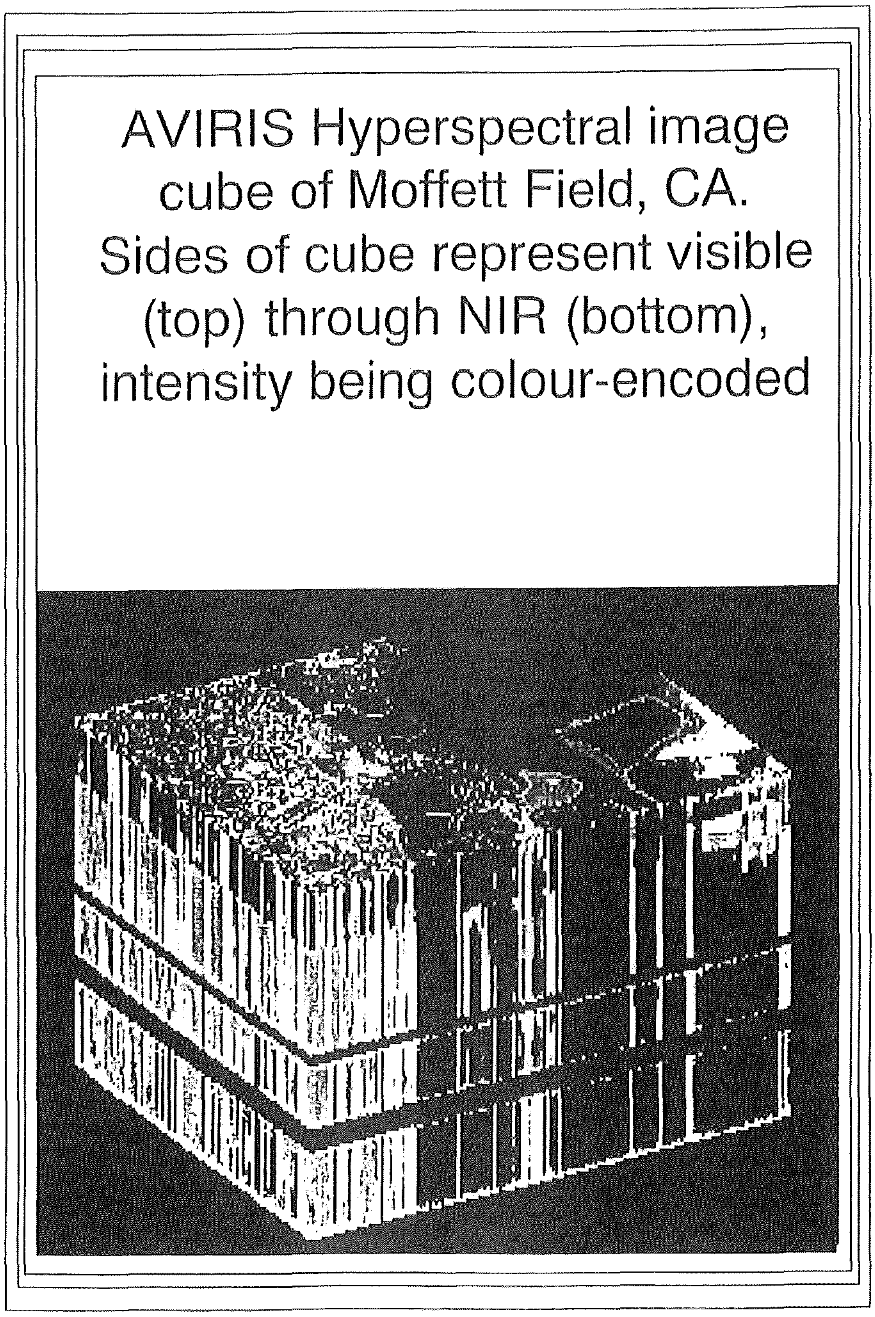
The AVIRIS concept

AVIRIS CONCEPT

EACH SPATIAL ELEMENT HAS A CONTINUOUS SPECTRUM THAT IS USED TO ANALYZE THE SURFACE AND ATMOSPHERE

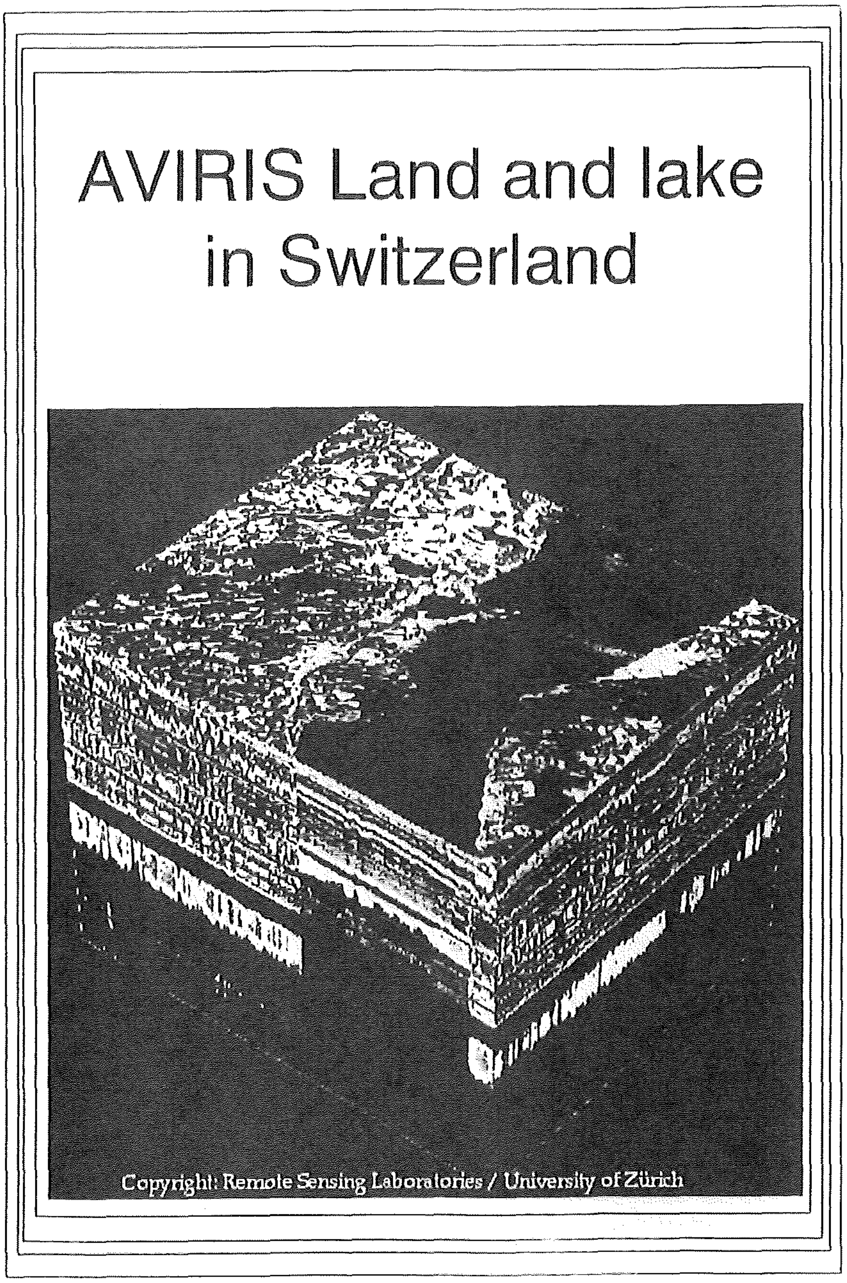
224 SPECTRAL IMAGES TAKEN SIMULTANEOUSLY

Figure 16.



AVIRIS Hyperspectral image cube of Moffett Field, CA. Sides of cube represent visible (top) through NIR (bottom), intensity being colour-encoded

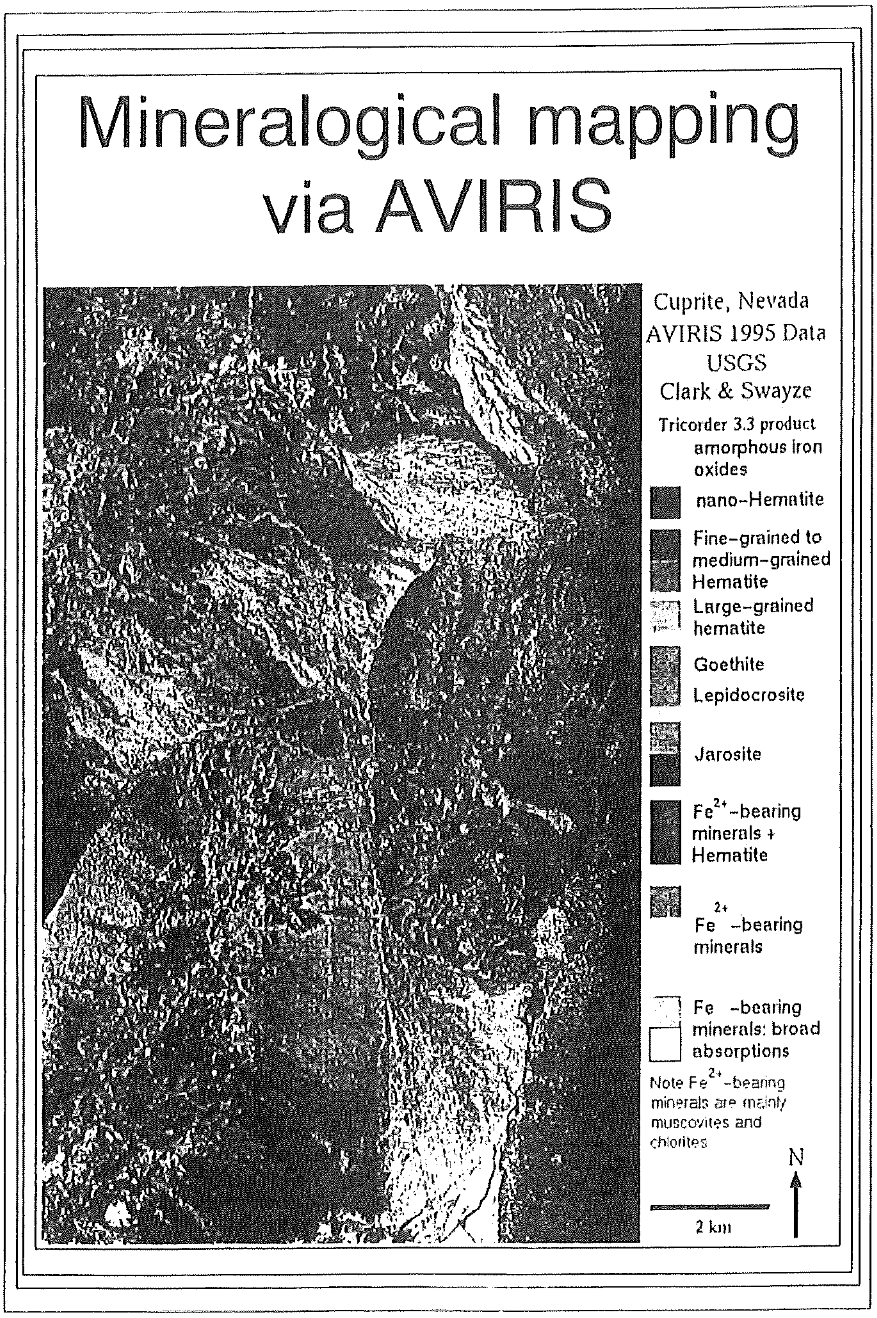
Figure 17.



AVIRIS Land and lake in Switzerland

Copyright: Remote Sensing Laboratories / University of Zürich

Figure 18.



Mineralogical mapping via AVIRIS

Cuprite, Nevada
AVIRIS 1995 Data
USGS
Clark & Swayze
Tricorder 3.3 product
amorphous iron oxides
nano-Hematite
Fine-grained to medium-grained Hematite
Large-grained hematite
Goethite
Lepidocrosite
Jarosite
Fe^{2+}-bearing minerals + Hematite
Fe^{2+}-bearing minerals
Fe -bearing minerals: broad absorptions
Note Fe^{2+}-bearing minerals are mainly muscovites and chlorites
N
2 km

Figure 19.

Vegetation mapping

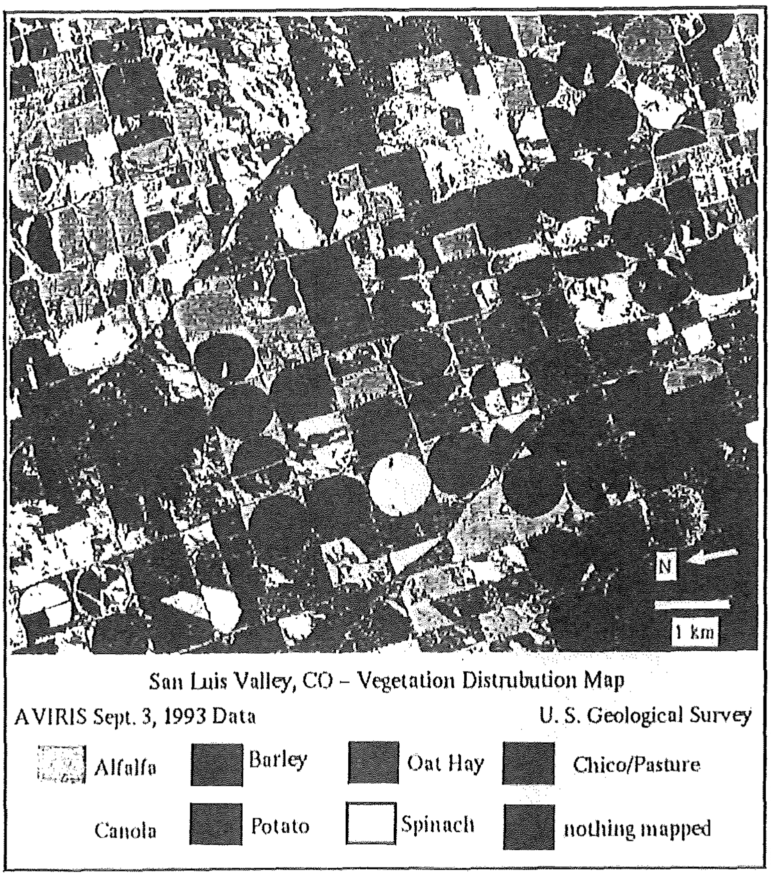
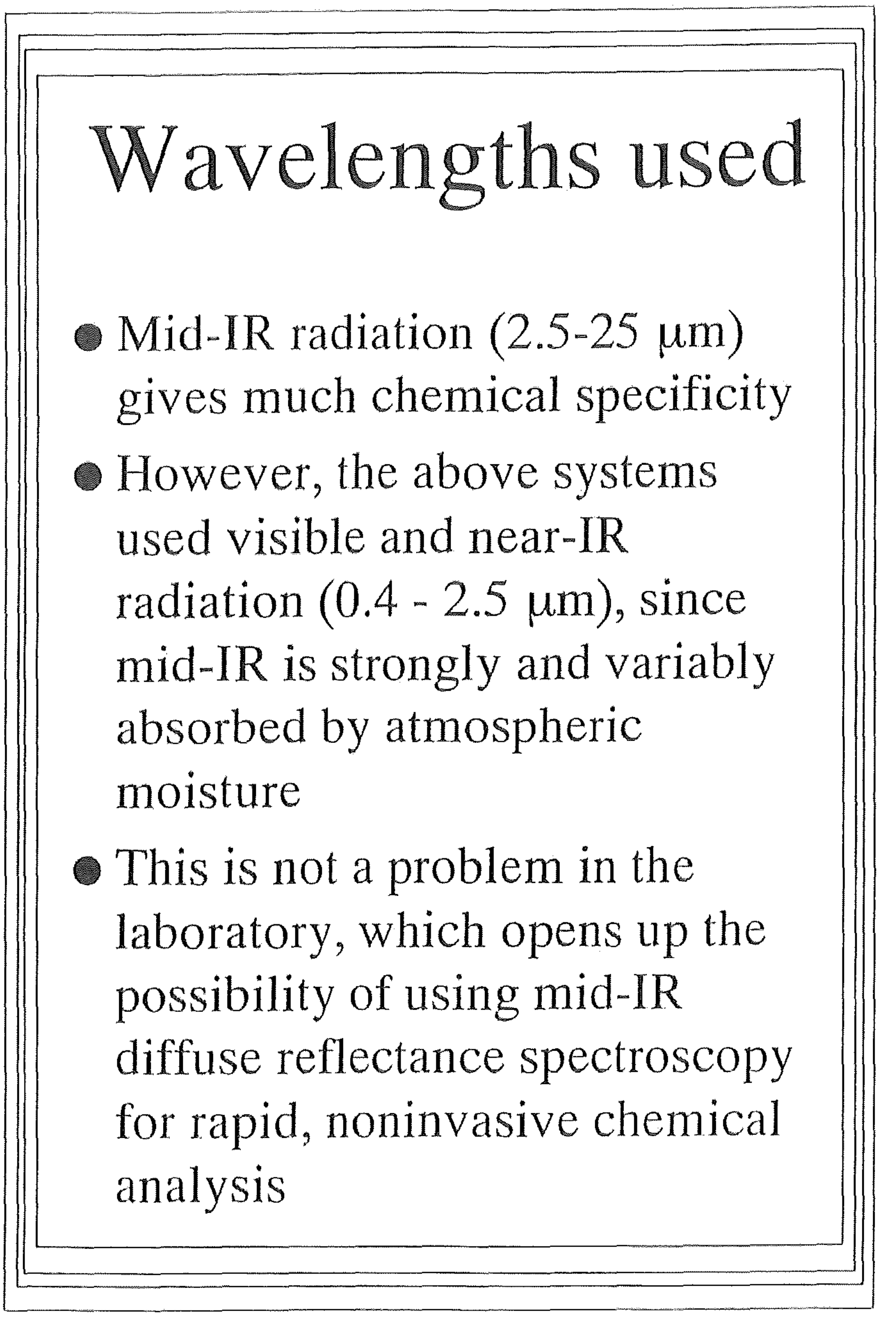


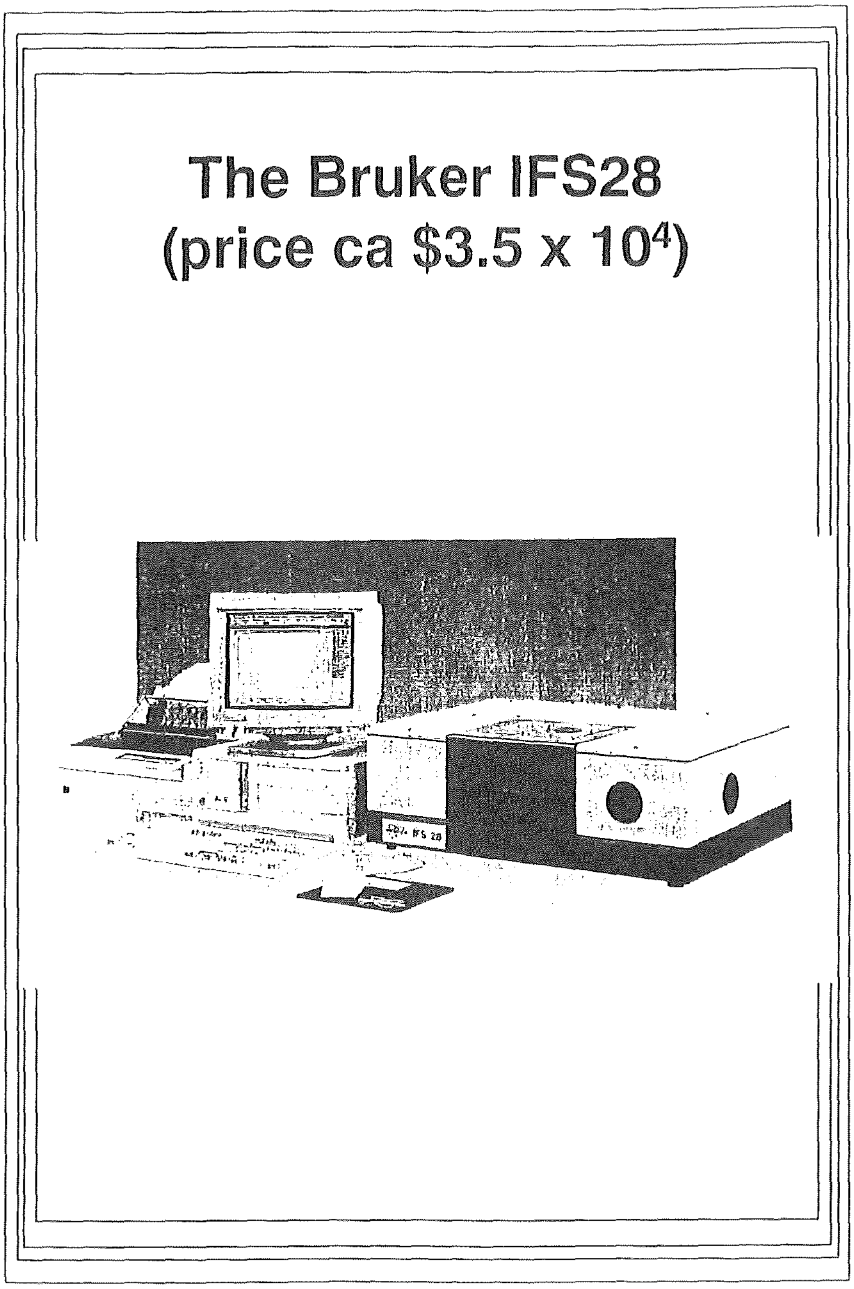
Figure 20.



Wavelengths used

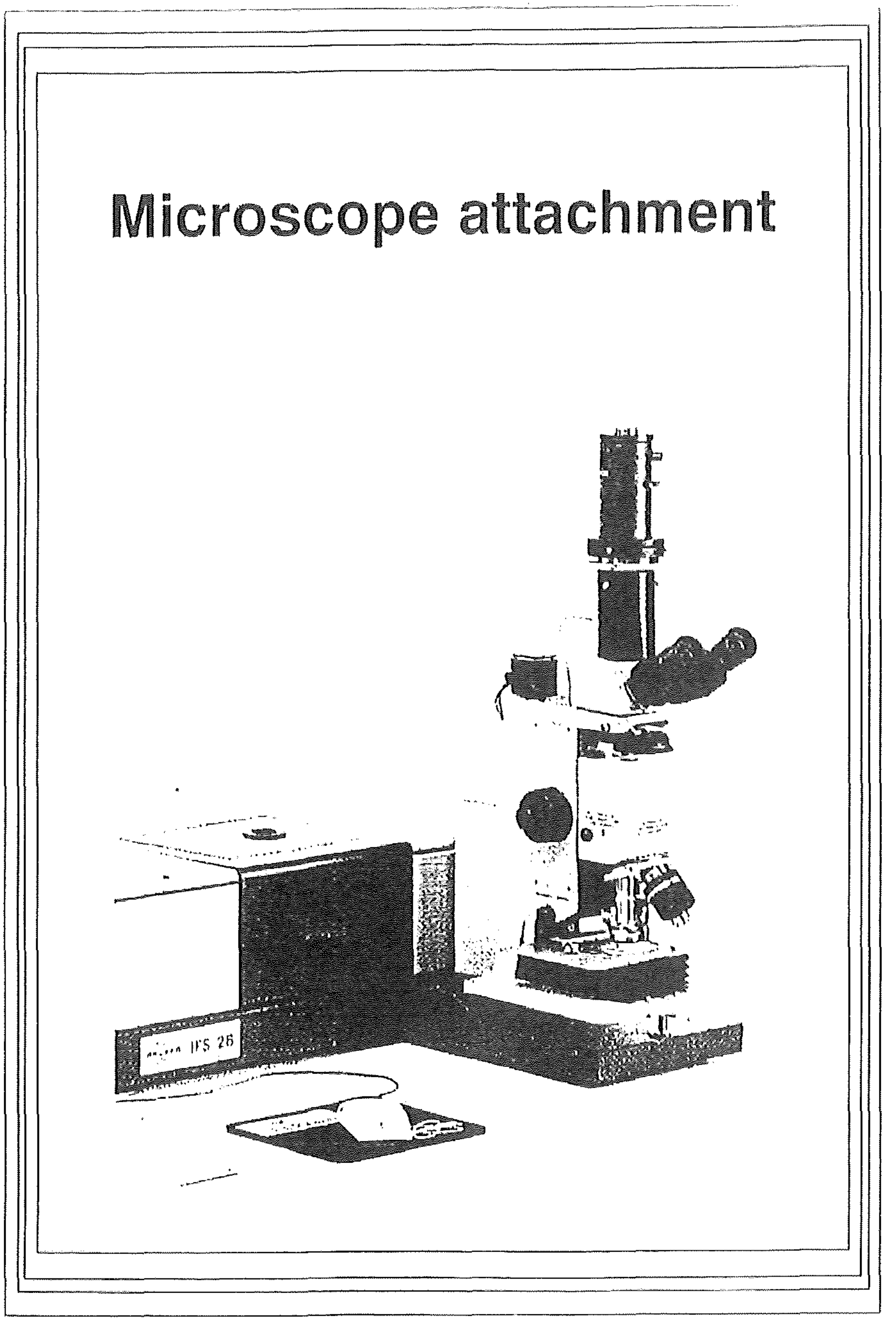
- Mid-IR radiation (2.5-25 μm) gives much chemical specificity
- However, the above systems used visible and near-IR radiation (0.4 - 2.5 μm), since mid-IR is strongly and variably absorbed by atmospheric moisture
- This is not a problem in the laboratory, which opens up the possibility of using mid-IR diffuse reflectance spectroscopy for rapid, noninvasive chemical analysis

Figure 21.



The Bruker IFS28
(price ca $\$3.5 \times 10^4$)

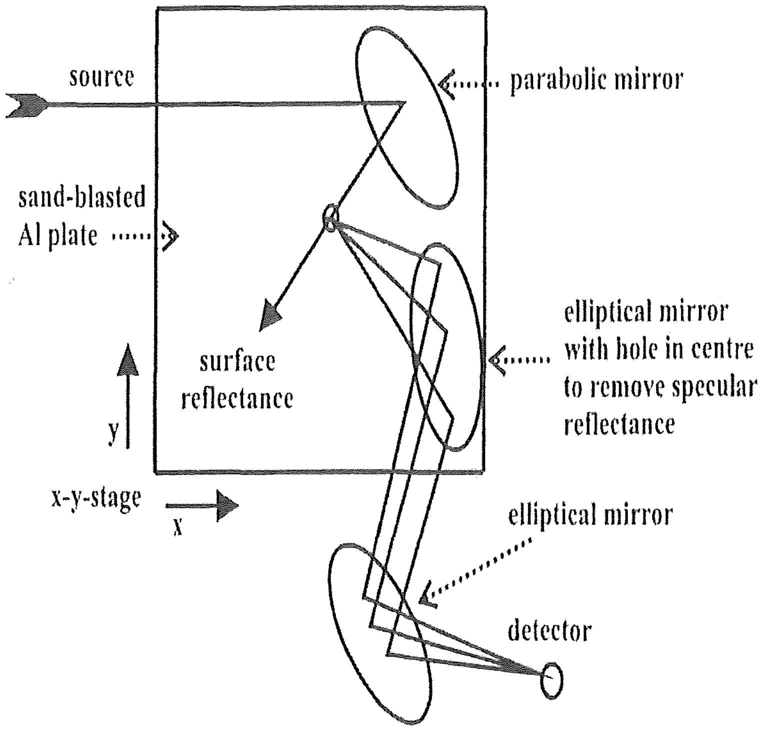
Figure 22.



Microscope attachment

Figure 23.

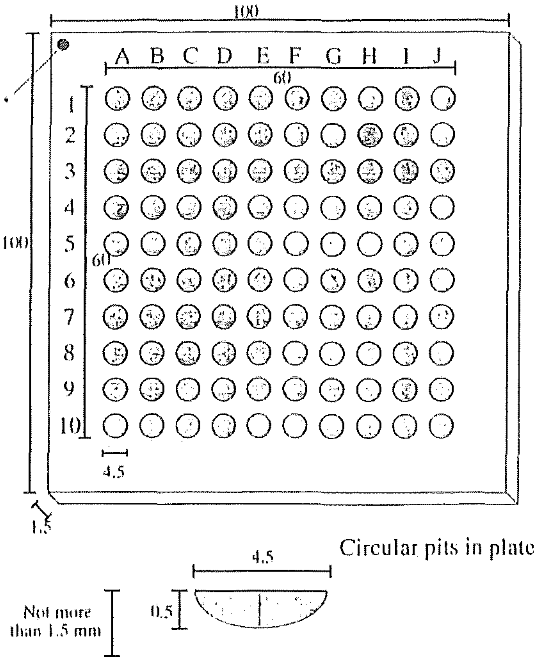
Scheme of the TLC unit for diffuse reflectance



Adapted from Glauniger, G., Kovar, K.-A. & Hoffmann, V. (1990) Possibilities and limits of an on-line coupling of thin-layer chromatography and FT-IR spectroscopy. *Fresenius Journal of Analytical Chemistry* 338, 710-716.

Figure 24.

Design for metal plates for diffuse reflectance

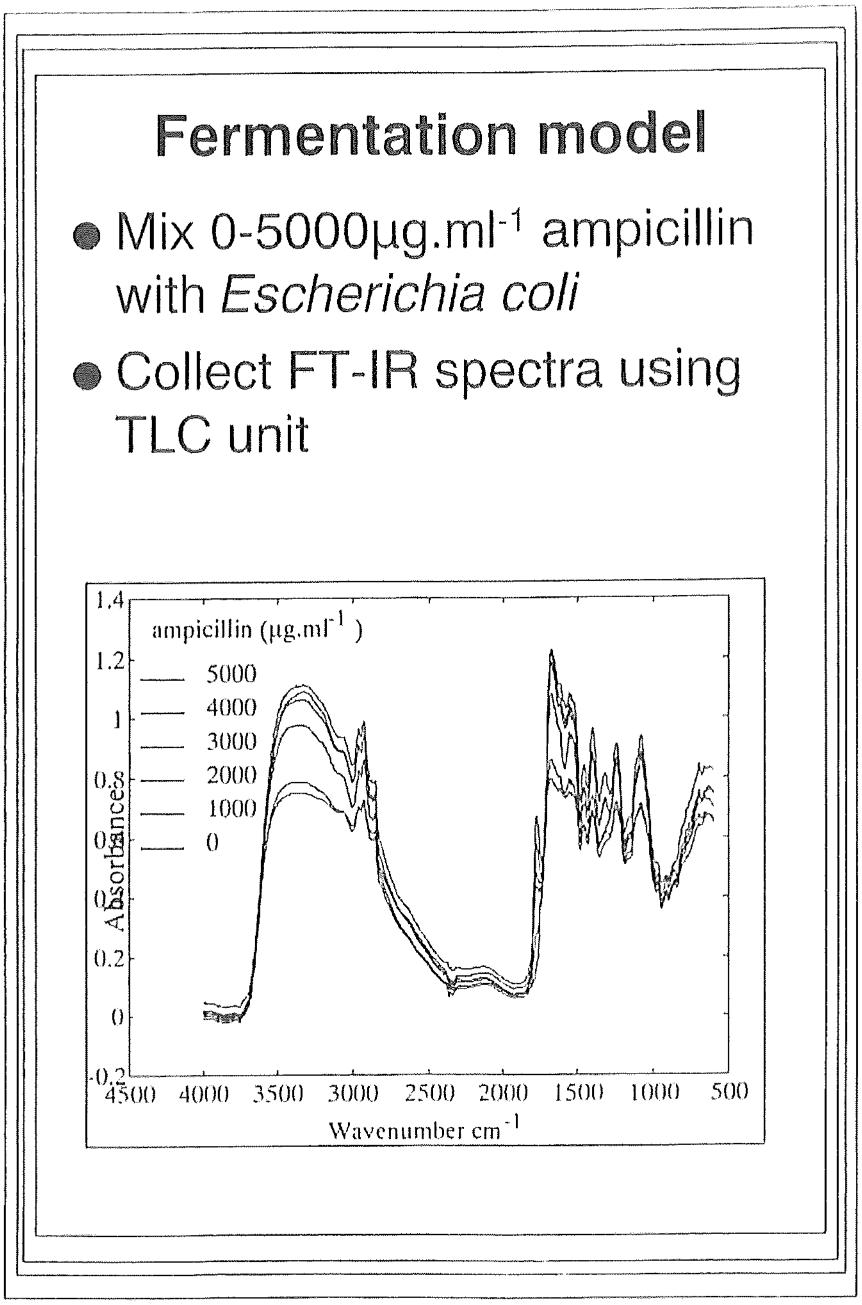


Circular pits in plate

* Small locating hole

Not to Scale. All measurements are mm.

Figure 25.

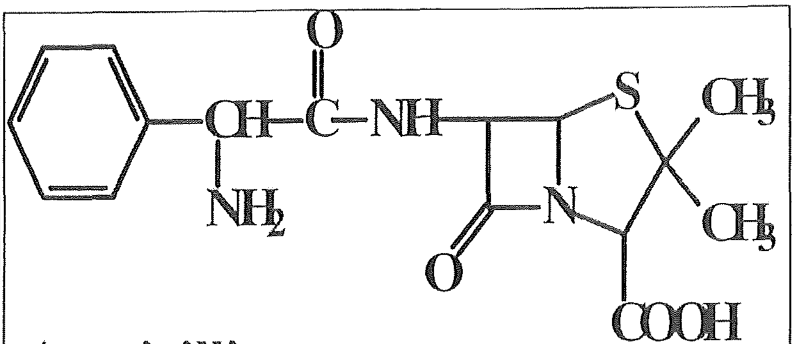


Fermentation model

- Mix 0-5000μg.ml^{-1} ampicillin with *Escherichia coli*
- Collect FT-IR spectra using TLC unit

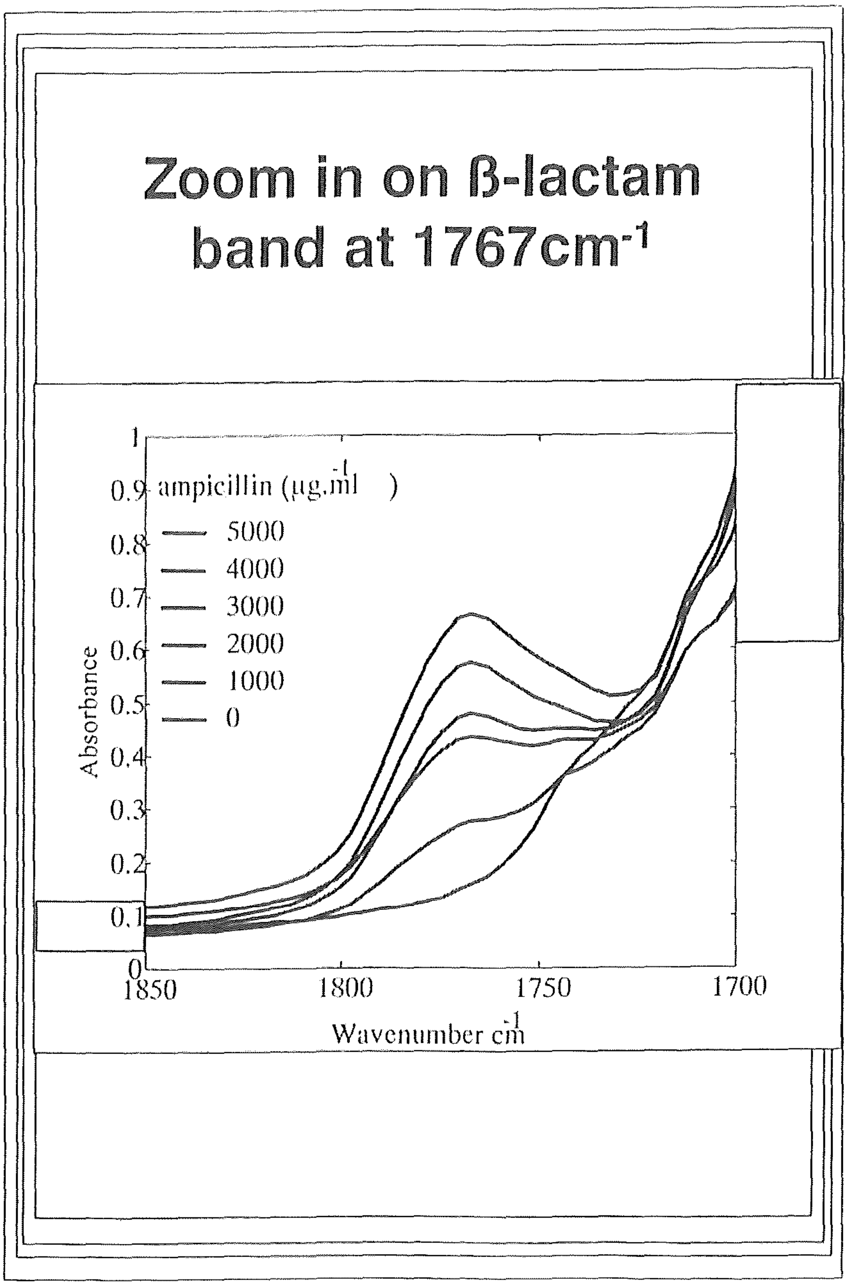
Figure 26.

The band at 1767 cm-1 in FT-IR spectra is characteristic of the ß- lactam moiety



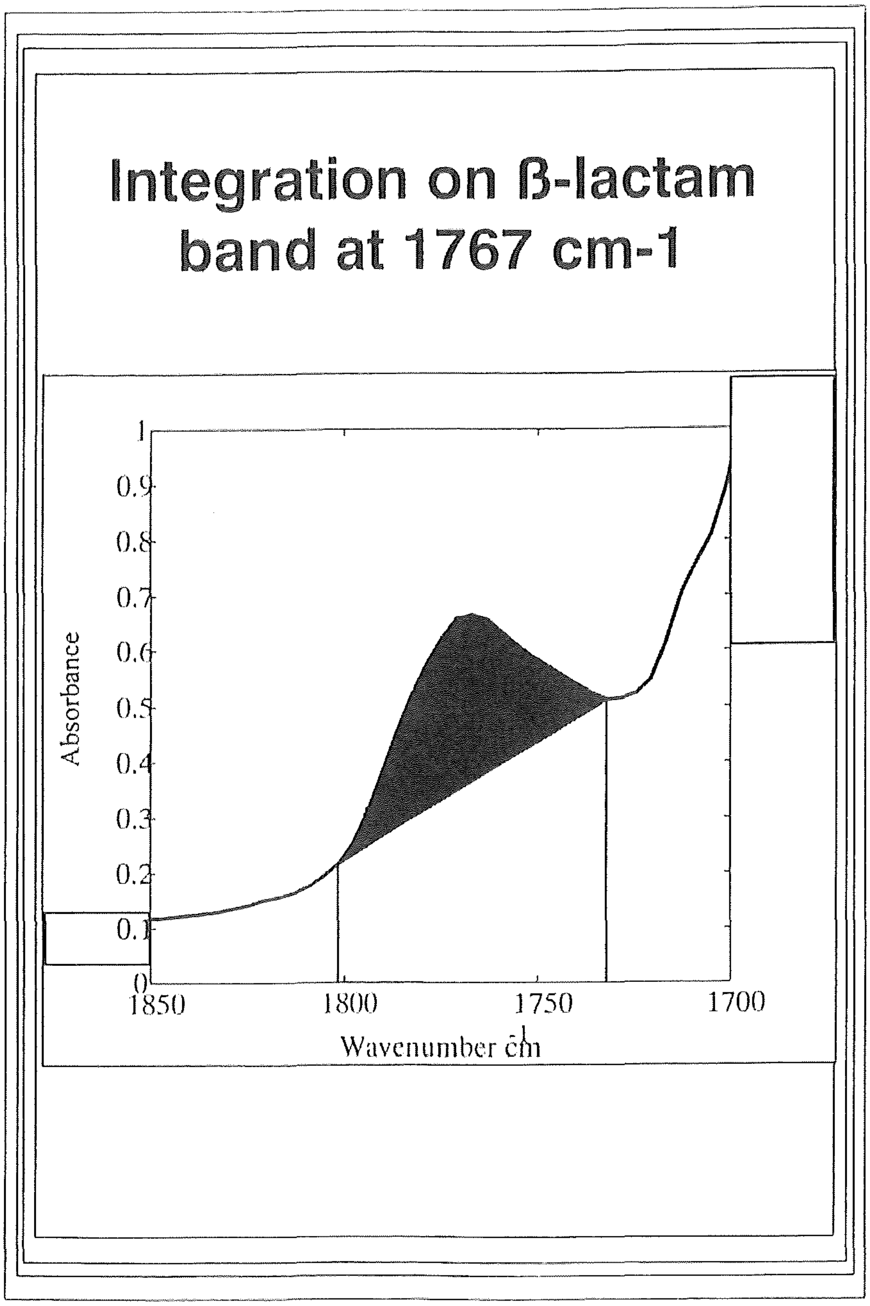
Ampicillin

Figure 27.



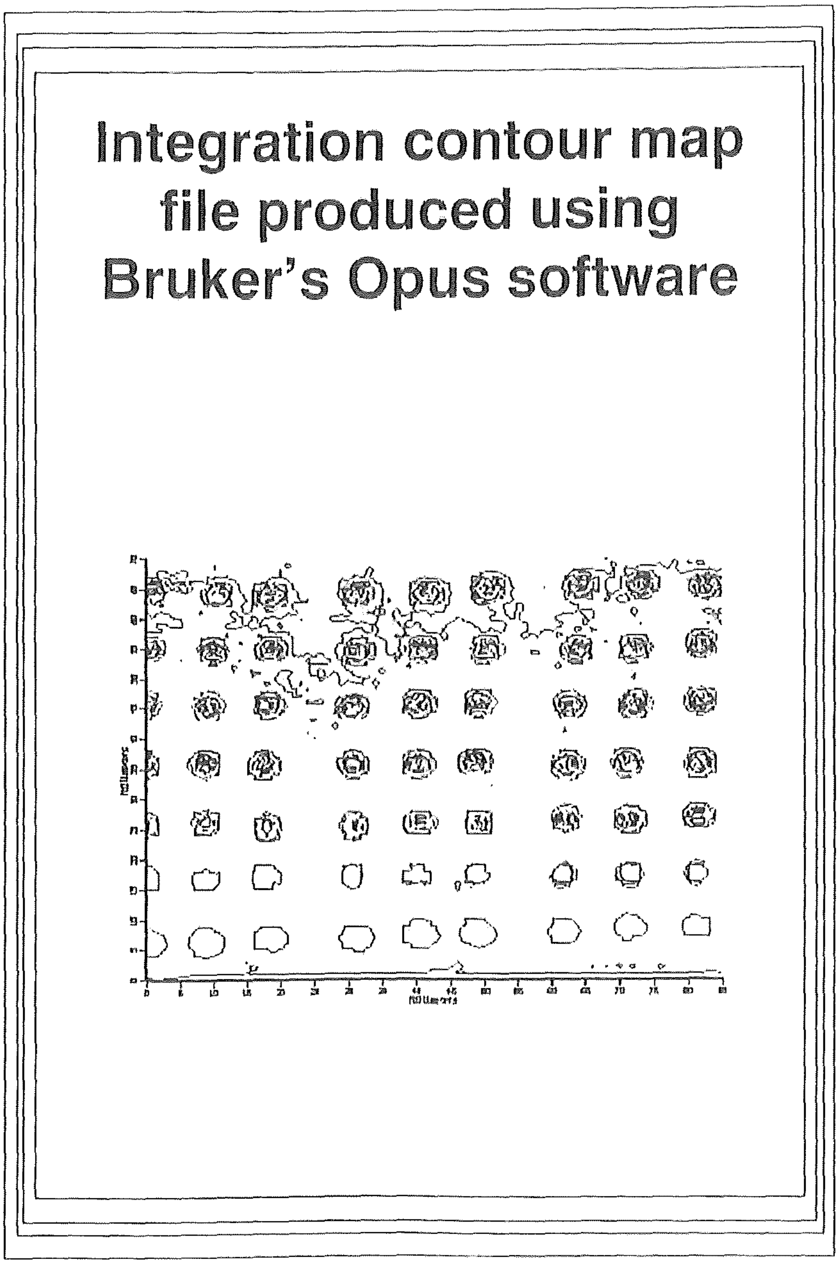
Zoom in on ß-lactam band at $1767 cm^{-1}$

Figure 28.



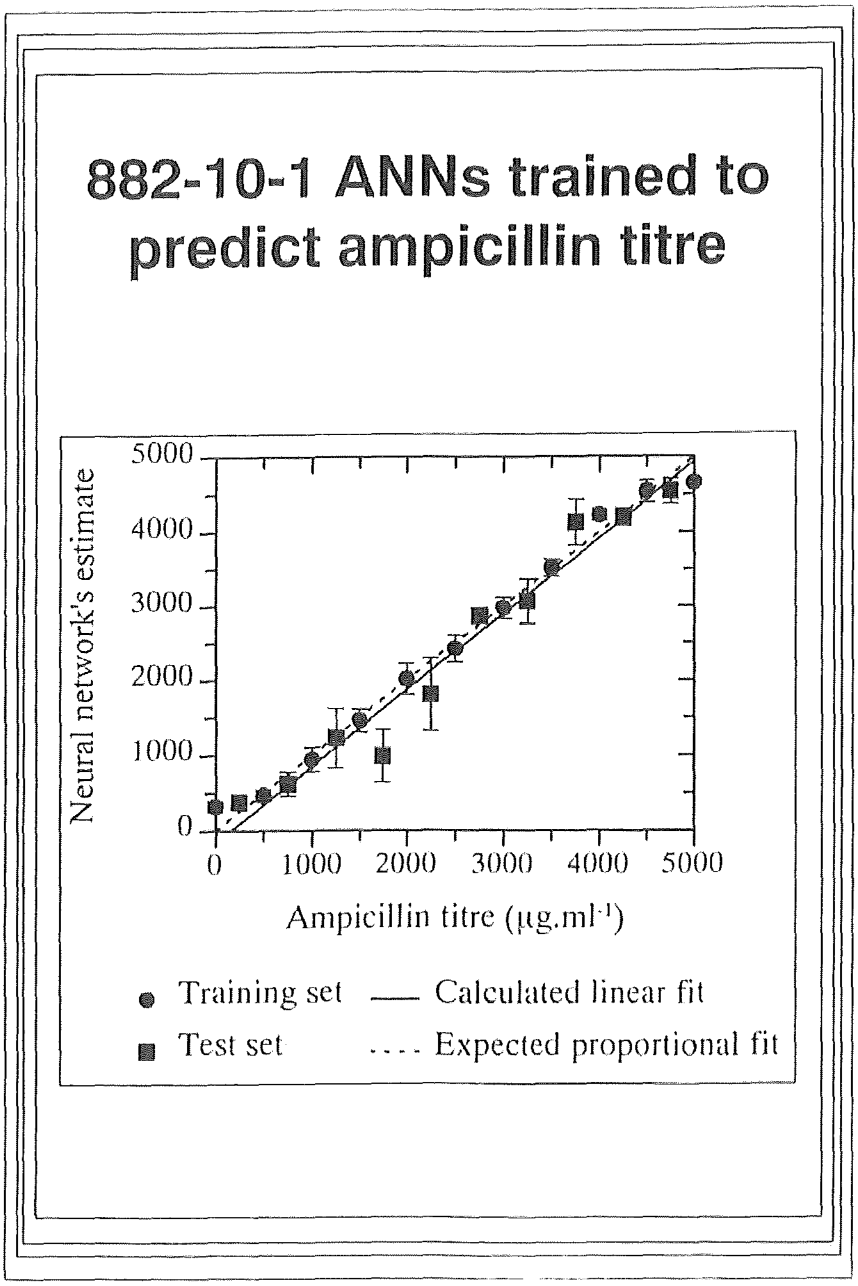
Integration on ß-lactam band at 1767 cm-1

Figure 29.



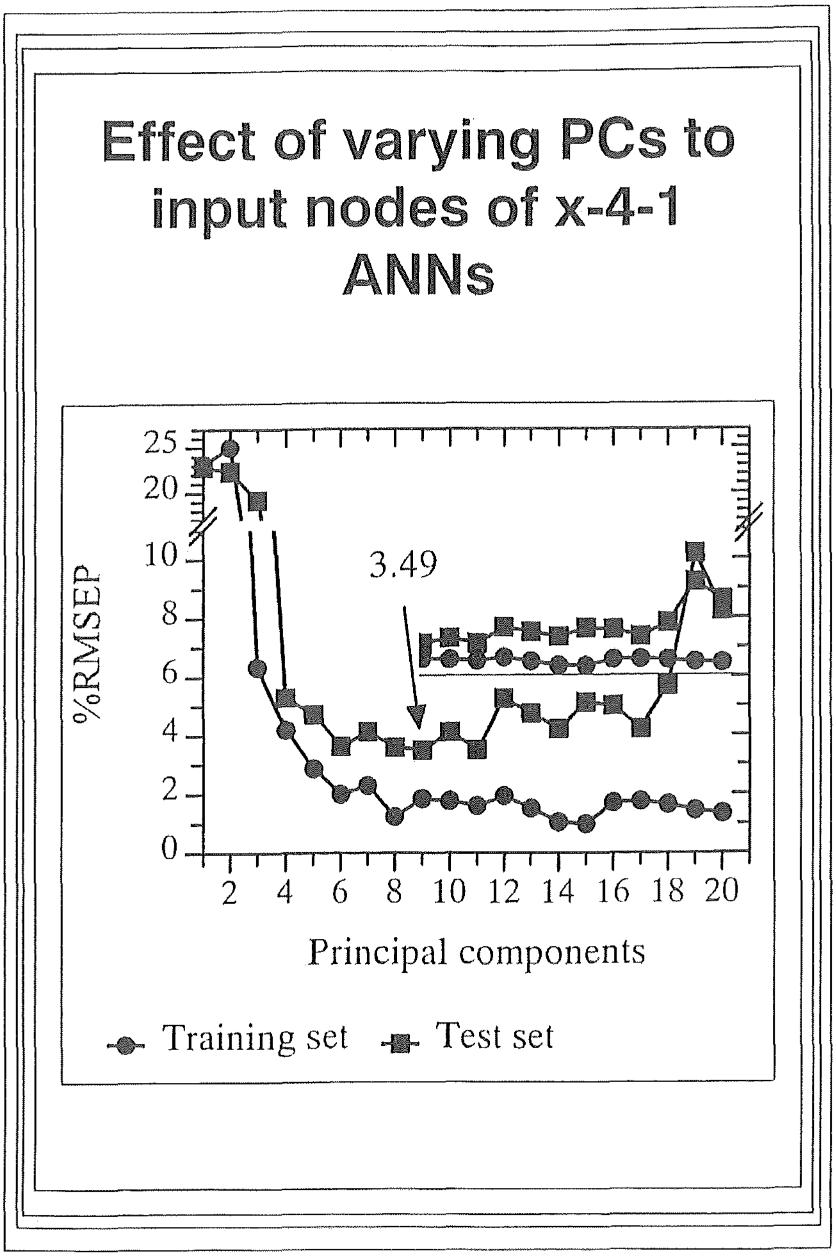
Integration contour map file produced using Bruker's Opus software

Figure 30.



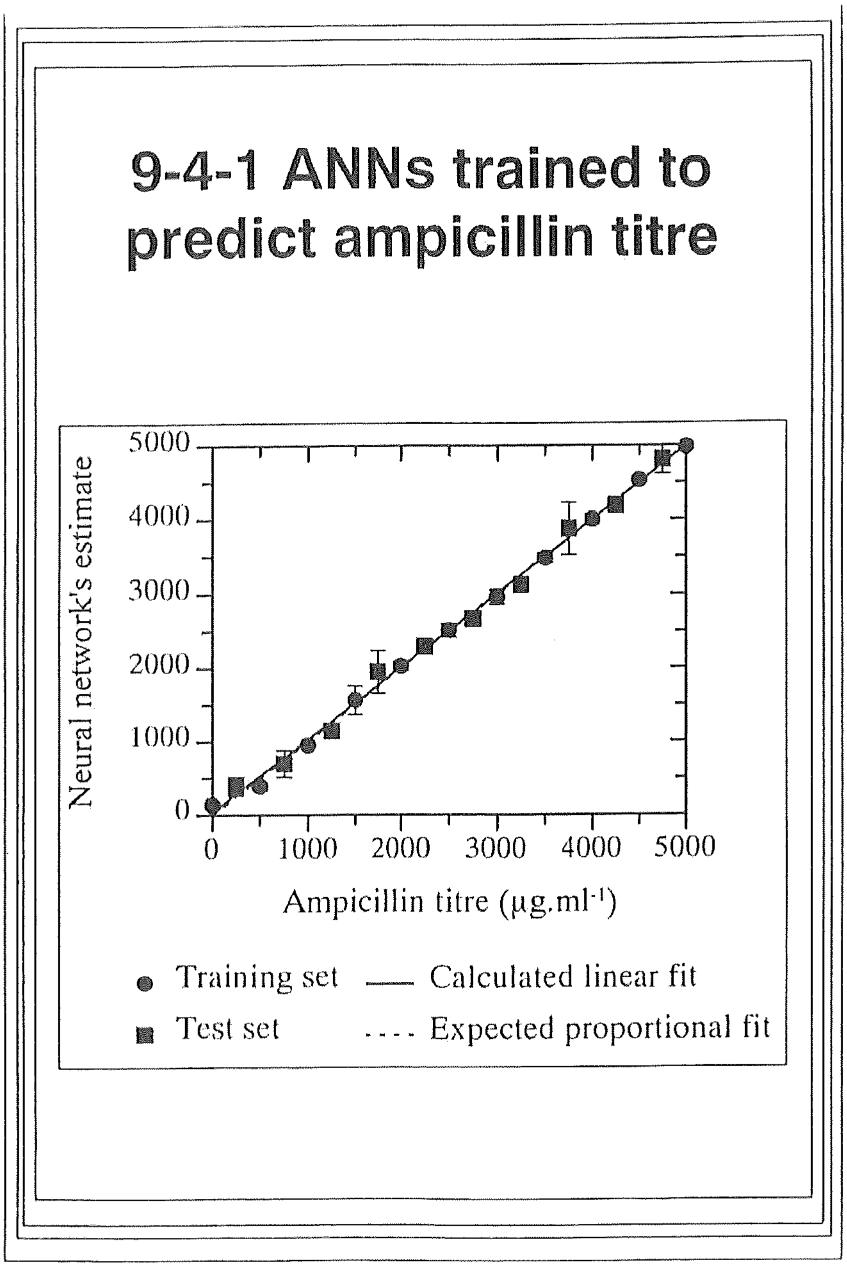
882-10-1 ANNs trained to predict ampicillin titre

Figure 31.



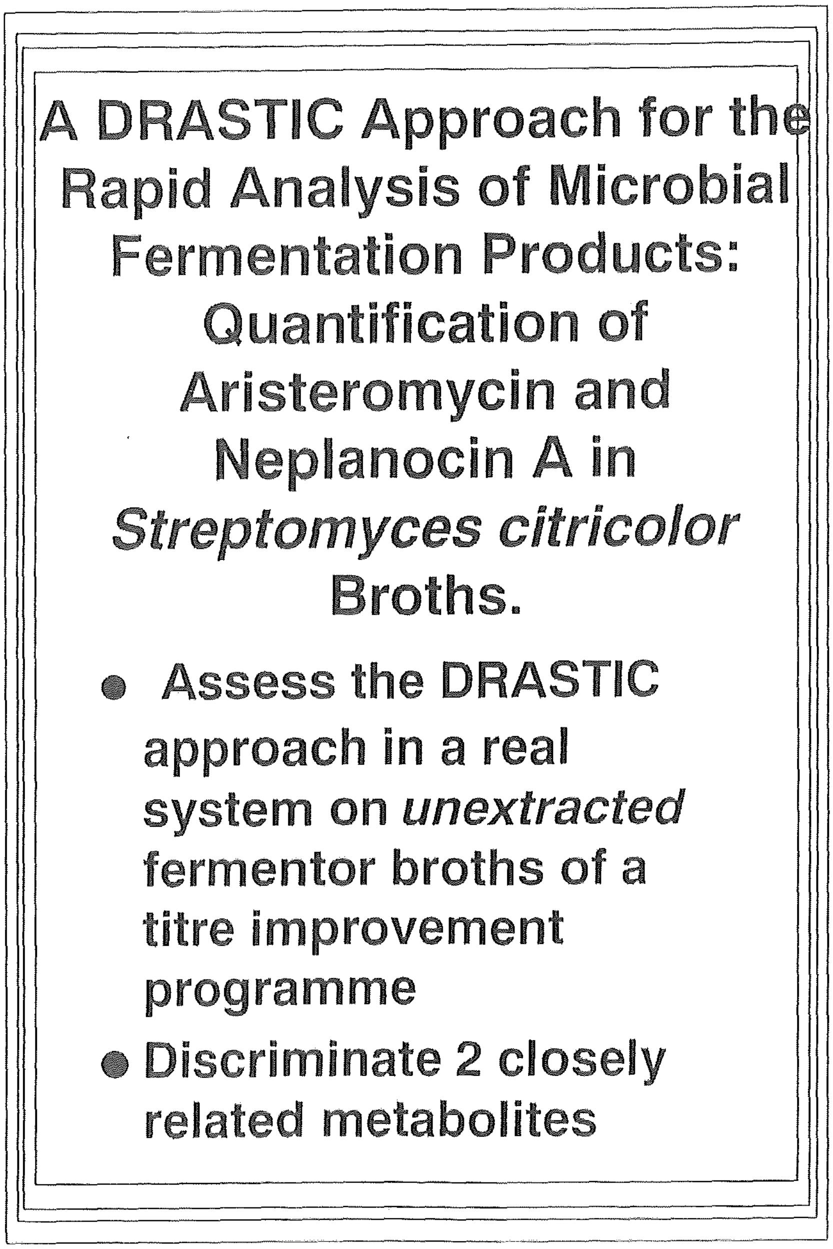
Effect of varying PCs to input nodes of x-4-1 ANNs

Figure 32.



9-4-1 ANNs trained to predict ampicillin titre

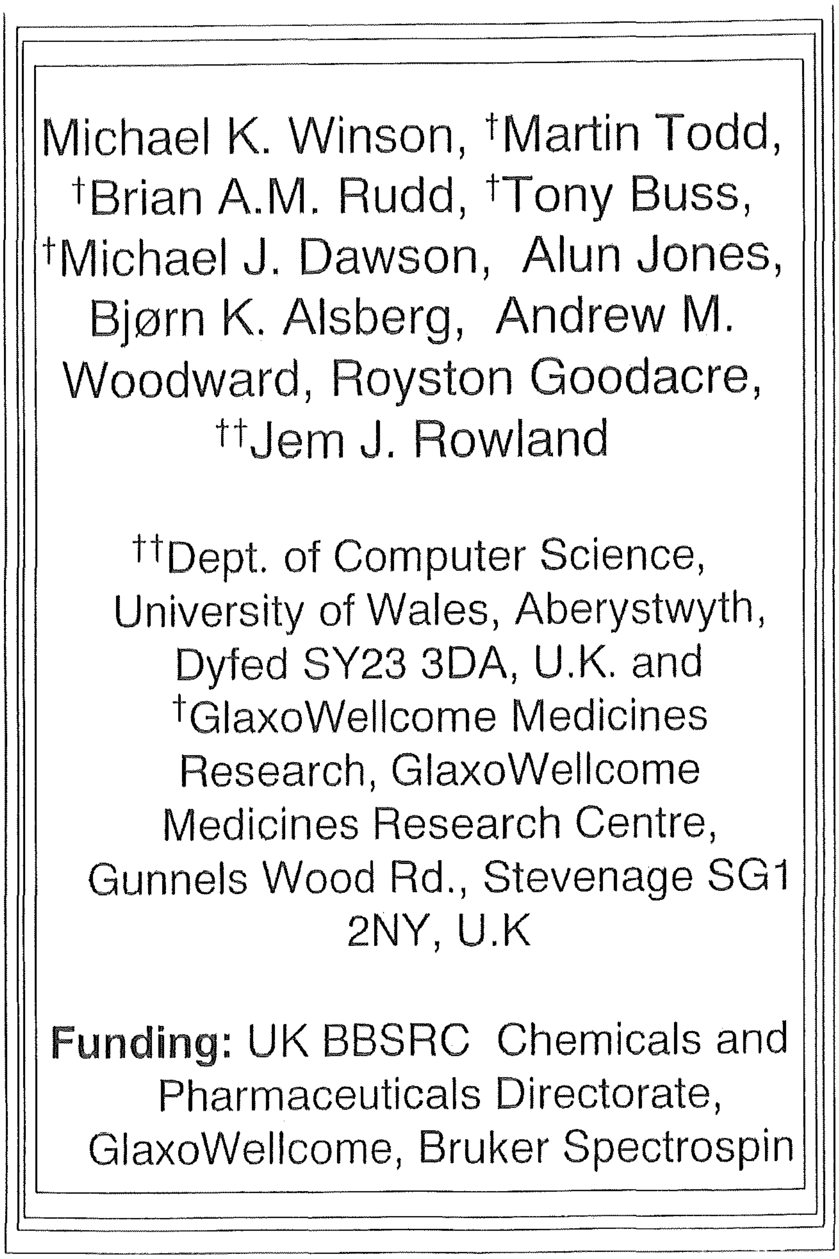
Figure 33.



A DRASTIC Approach for the Rapid Analysis of Microbial Fermentation Products: Quantification of Aristeromycin and Neplanocin A in *Streptomyces citricolor* Broths.

- Assess the DRASTIC approach in a real system on *unextracted* fermentor broths of a titre improvement programme
- Discriminate 2 closely related metabolites

Figure 34.



Michael K. Winson, [†]Martin Todd, [†]Brian A.M. Rudd, [†]Tony Buss, [†]Michael J. Dawson, Alun Jones, Bjørn K. Alsberg, Andrew M. Woodward, Royston Goodacre, [††]Jem J. Rowland

[††]Dept. of Computer Science, University of Wales, Aberystwyth, Dyfed SY23 3DA, U.K. and [†]GlaxoWellcome Medicines Research, GlaxoWellcome Medicines Research Centre, Gunnels Wood Rd., Stevenage SG1 2NY, U.K

Funding: UK BBSRC Chemicals and Pharmaceuticals Directorate, GlaxoWellcome, Bruker Spectrospin

Figure 35.

Aristeromycin and Neplanocin A

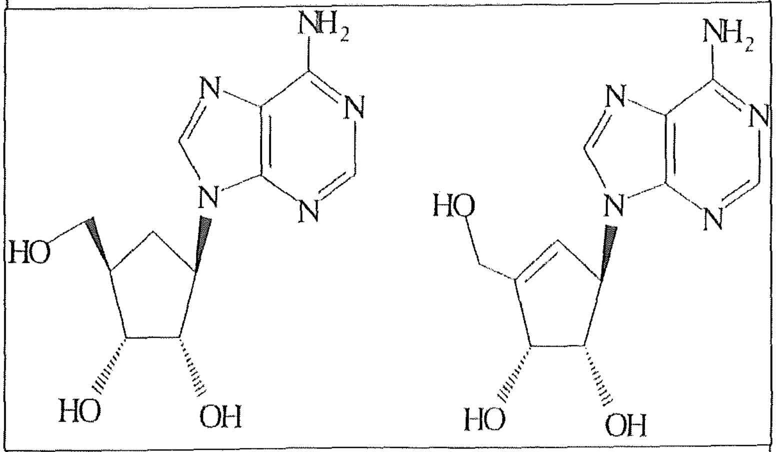
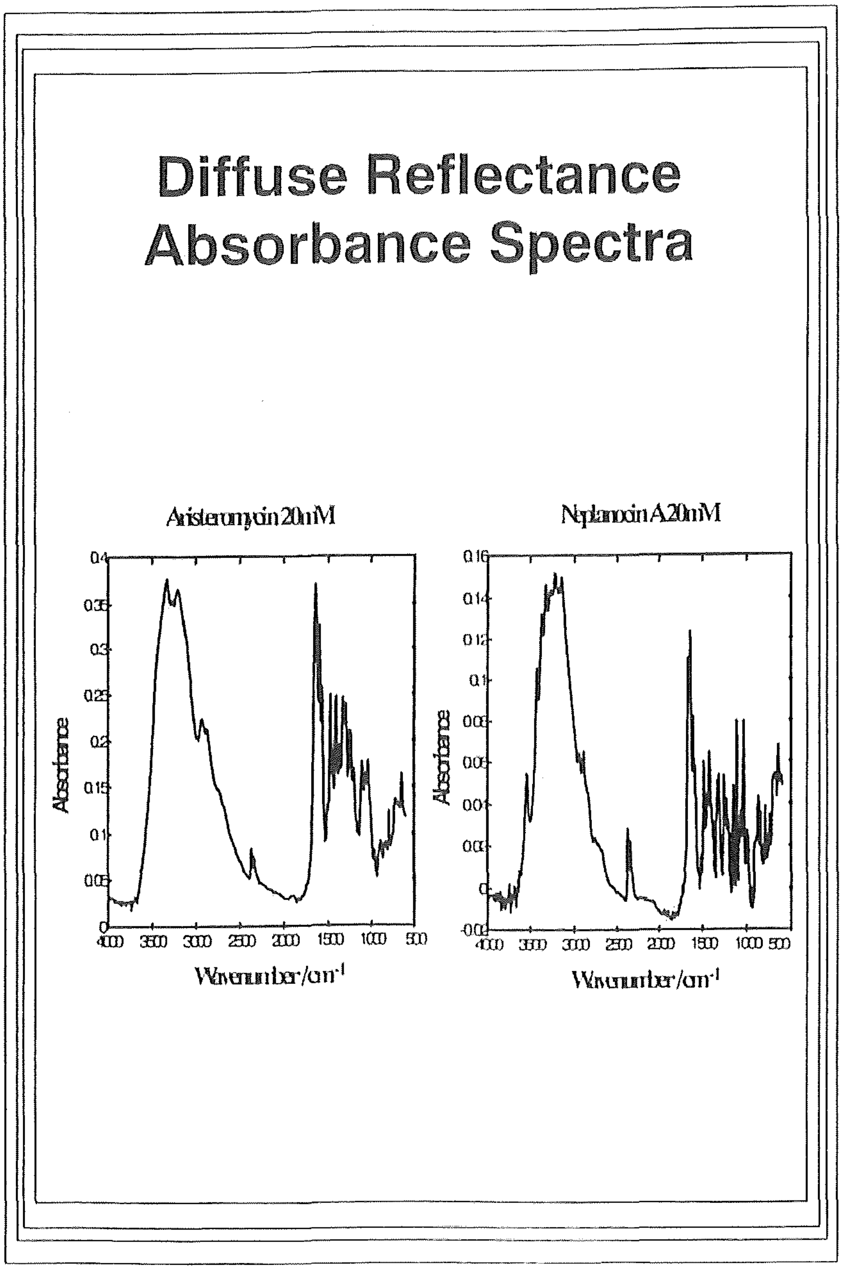


Figure 36.



Diffuse Reflectance Absorbance Spectra

Figure 37.

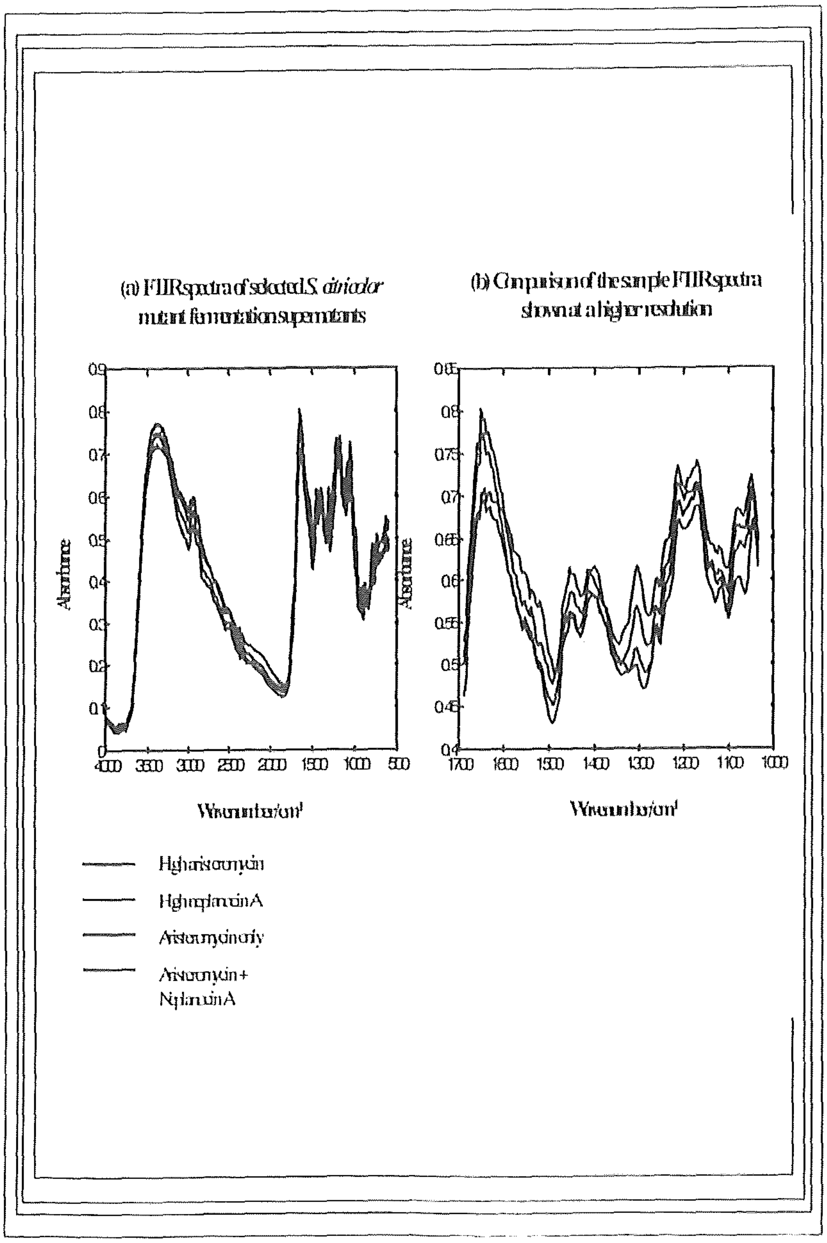
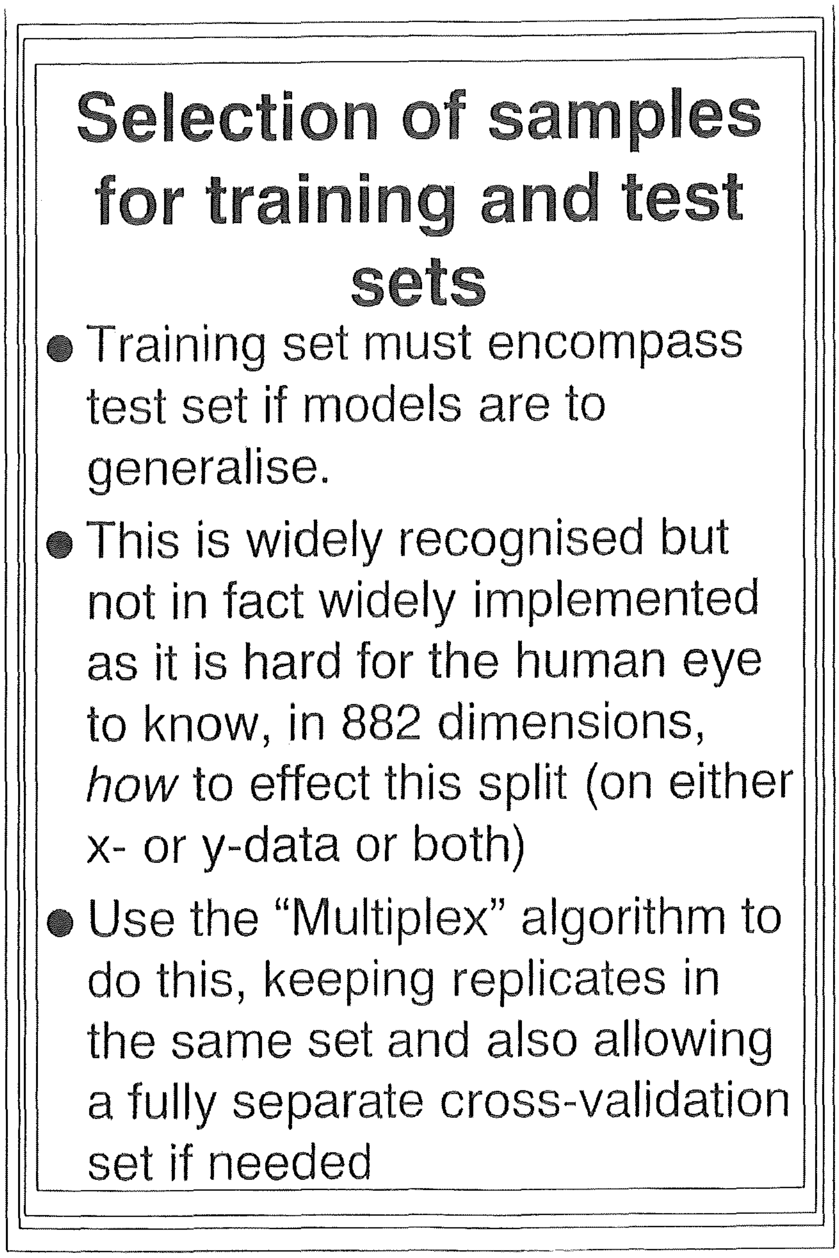

(a) FTIR spectra of selected S. *citricolor* mutant fermentation supernatants
(b) Comparison of the sample FTIR spectra shown at a higher resolution
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Absorbance
4000
3500
3000
2500
2000
1500
1000
500
Wavenumber/cm-1
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
Absorbance
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumber/cm-1
High aristeromycin
High neplanocin A
Aristeromycin only
Aristeromycin + Neplanocin A

Figure 38.



Selection of samples for training and test sets

- Training set must encompass test set if models are to generalise.
- This is widely recognised but not in fact widely implemented as it is hard for the human eye to know, in 882 dimensions, *how* to effect this split (on either x- or y-data or both)
- Use the "Multiplex" algorithm to do this, keeping replicates in the same set and also allowing a fully separate cross-validation set if needed

Figure 39.

PLS predictions - Aristeromycin

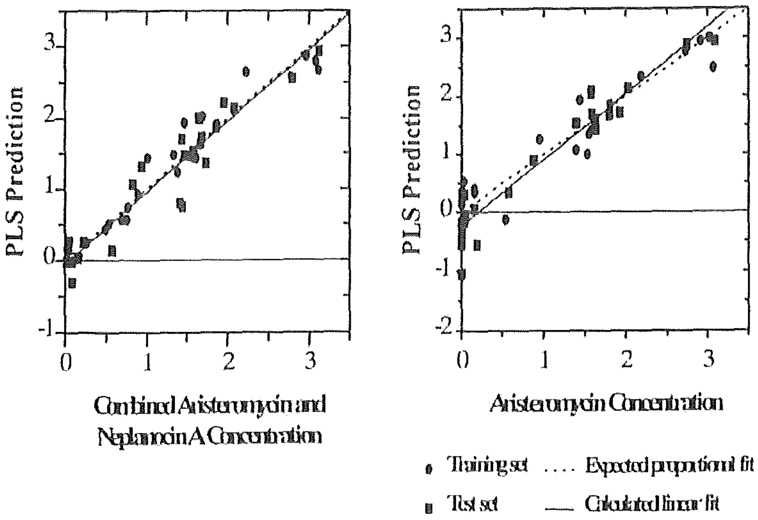
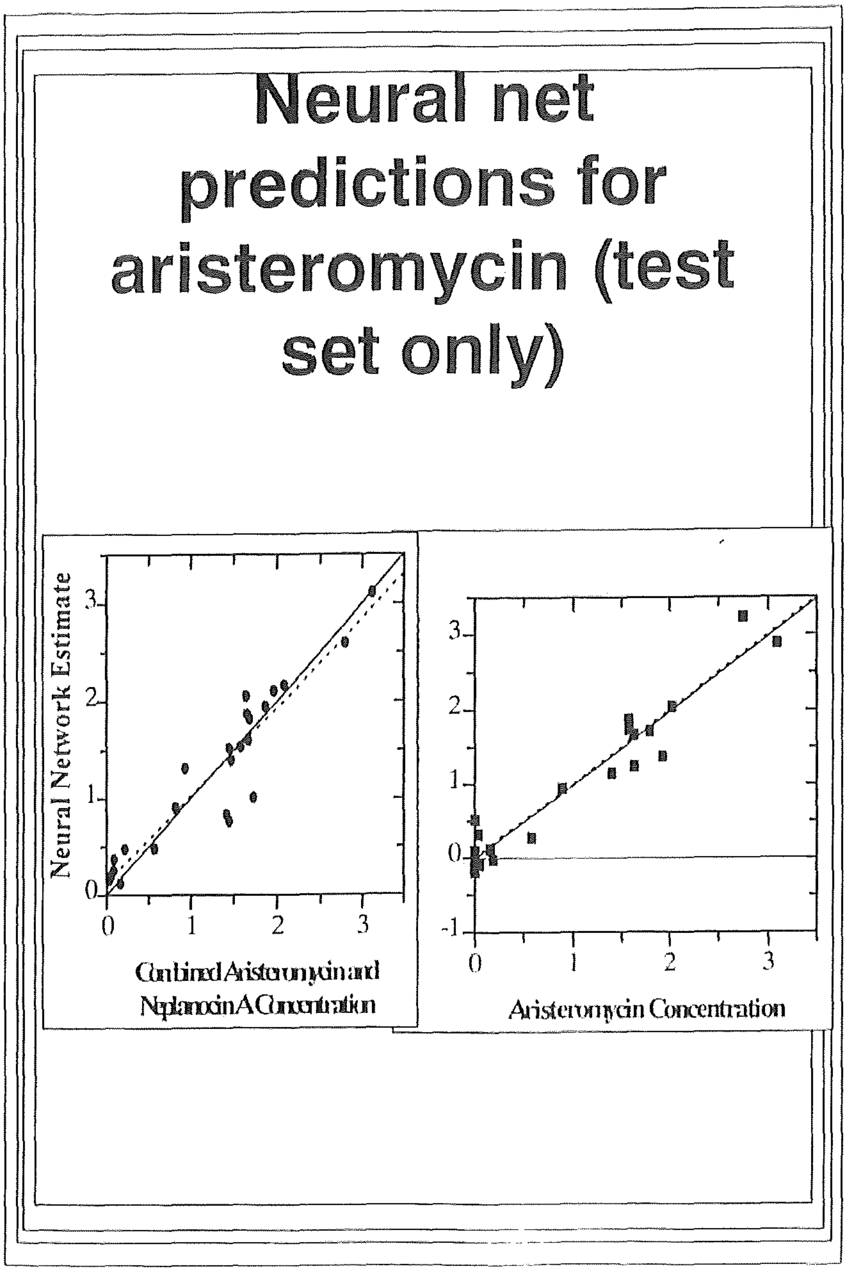
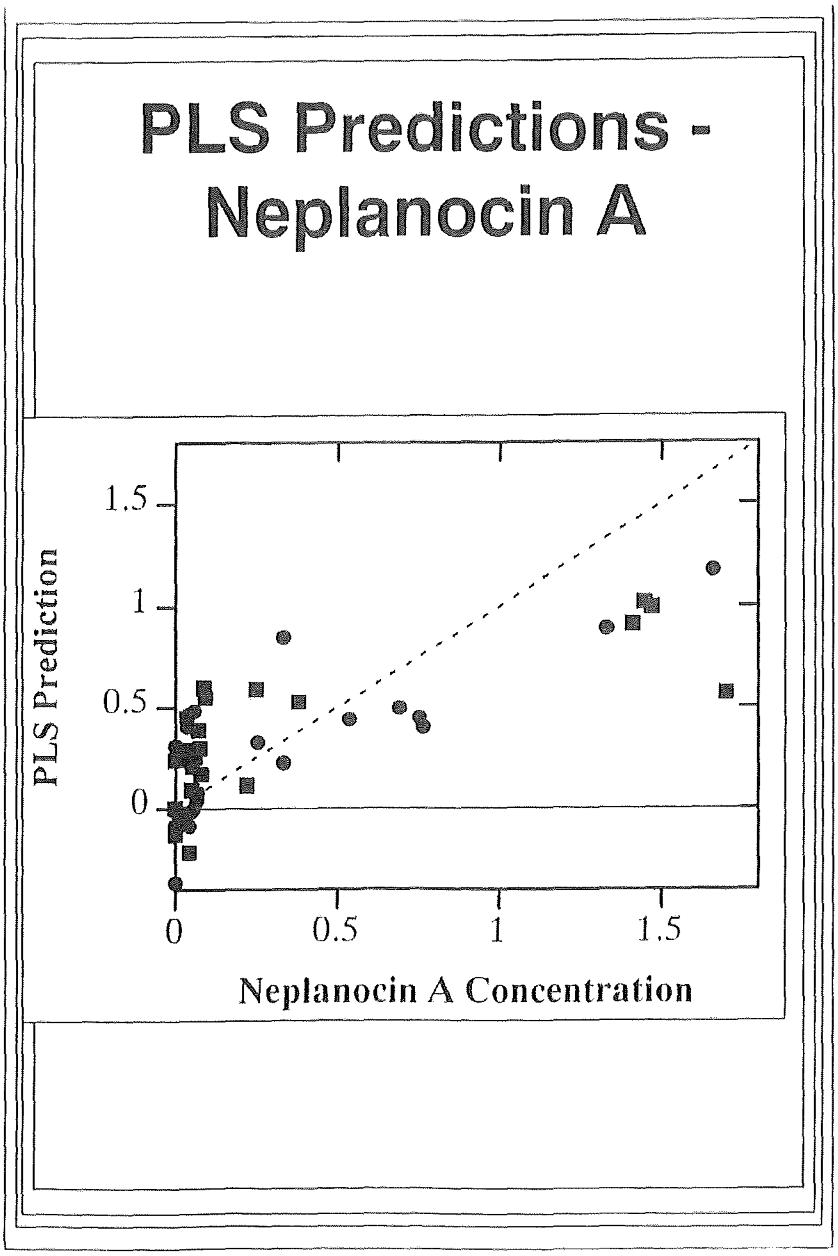


Figure 40.



Neural net predictions for aristeromycin (test set only)

Figure 41.



PLS Predictions - Neplanocin A

Figure 42.

Neural net predictions for neplanocin (test set only)

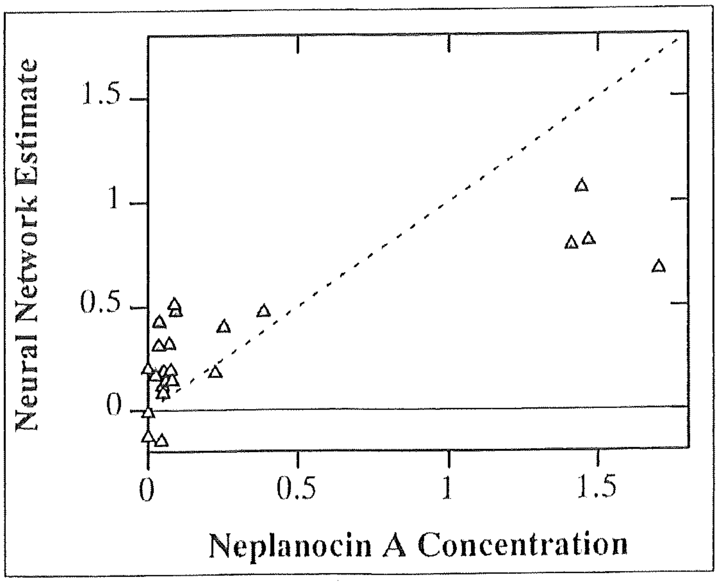
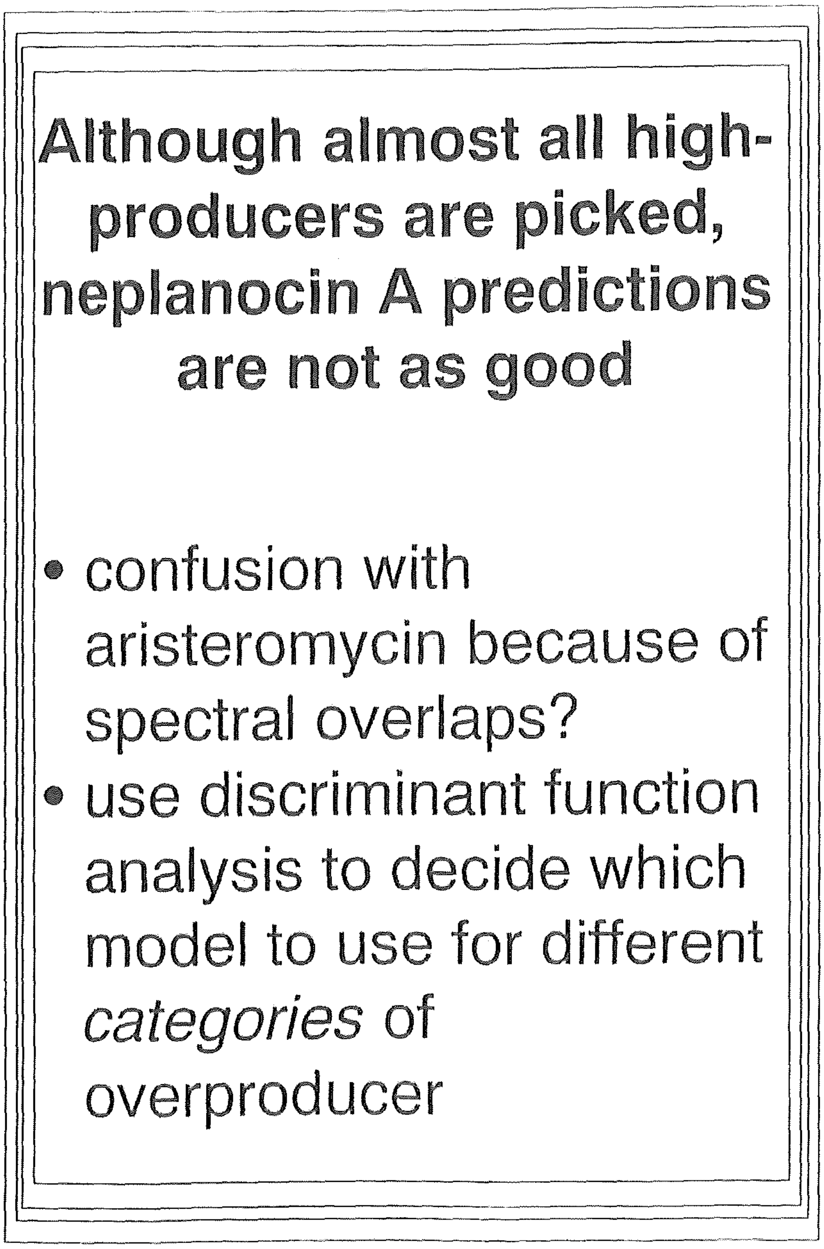


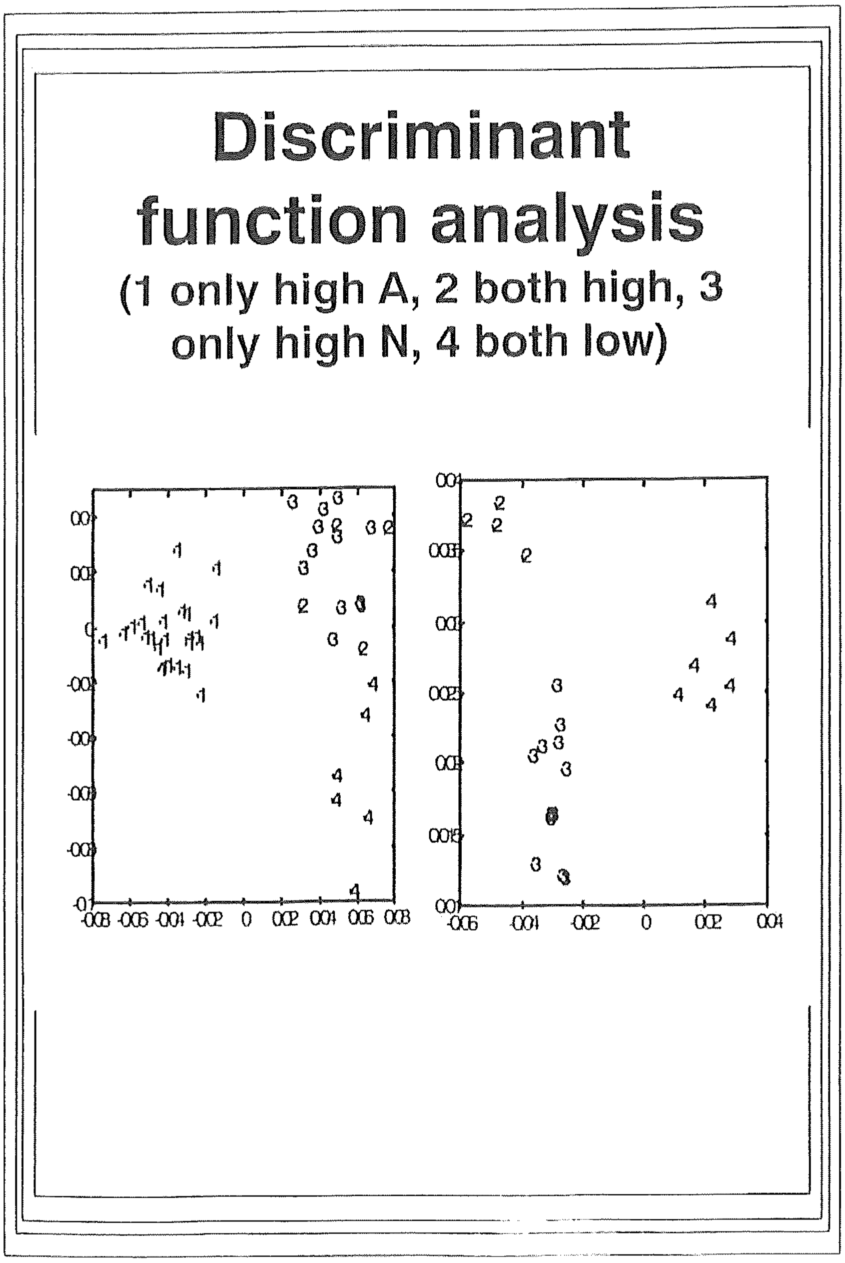
Figure 43.



Although almost all high-producers are picked, neplanocin A predictions are not as good

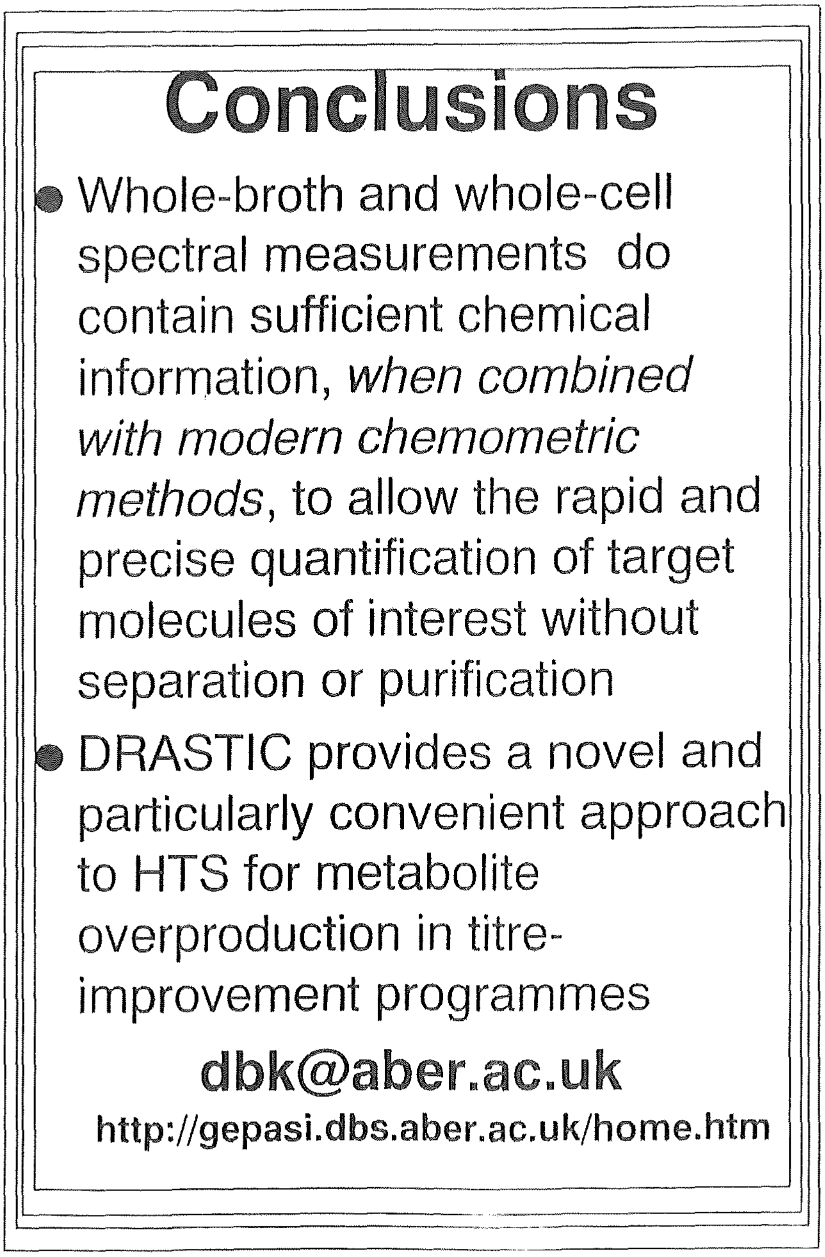
- confusion with aristeromycin because of spectral overlaps?
- use discriminant function analysis to decide which model to use for different *categories* of overproducer

Figure 44.



Discriminant function analysis
(1 only high A, 2 both high, 3 only high N, 4 both low)

Figure 45.



Conclusions

- Whole-broth and whole-cell spectral measurements do contain sufficient chemical information, *when combined with modern chemometric methods*, to allow the rapid and precise quantification of target molecules of interest without separation or purification
- DRASTIC provides a novel and particularly convenient approach to HTS for metabolite overproduction in titre-improvement programmes

dbk@aber.ac.uk

http://gepasi.dbs.aber.ac.uk/home.htm

Figure 46.

BIBLIOGRAPHY

Many more details appear on our Web pages at **http://gepasi.dbs.aber.ac.uk/home.html** but an introduction to some of our approaches to both spectroscopy and chemometrics to complex biological systems may be found in the following:

Alsberg, B.K., Goodacre, R., Rowland, J.J., and Kell, D.B. Classification of pyrolysis mass spectra by fuzzy multivariate rule induction-comparison with regression, K-nearest neighbour, neural and decision-tree methods. *Analytica Chimica Acta* 348:389-407, 1997.

Alsberg, B.K., Woodward, A.M., Winson, M.K., Rowland, J., and Kell, D.B. Wavelet denoising of infrared spectra. *Analyst* 122:645-652, 1997.

Bianchi, G., Giansante, L., Goodacre, R., Kell, D.B., Lazzari, M., Salter, G.J., and Surricchio, G. Rapid determination of geographical origin of extra virgin olive oil using Curie point pyrolysis mass spectrometry and artificial neural networks. *J. Anal. Appl. Pyrol.* In press, 1997.

Broadhurst, D., Goodacre, R., Jones, A., Rowland, J.J., and Kell, D.B. 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* 348:71-86, 1997.

Davey, H.M., and Kell, D.B. Flow cytometry and cell sorting of heterogeneous microbial populations: the importance of single-cell analysis. *Microbiol. Rev.* 60:641-696, 1996.

Goodacre, R., Edmonds, A.N., and Kell, D.B. Quantitative analysis of the pyrolysis mass spectra of complex mixtures using artificial neural networks. Application to amino acids in glycogen. *Journal of Analytical and Applied Pyrolysis* 26:93-114, 1993.

Goodacre, R., Karim, A., Kaderbhai, M.A., and Kell, D.B. Rapid and quantitative analysis of recombinant protein expression using pyrolysis mass spectrometry and artificial neural networks - application to mammalian cytochrome b5 in Escherichia coli. *Journal of Biotechnology* 34:185-193, 1994.

Goodacre, R., and Kell, D.B. Rapid and quantitative analysis of bioprocesses using pyrolysis mass spectrometry and neural networks. Application to indole production. *Analytica Chimica Acta* 279:17-26, 1993.

Goodacre, R., and Kell, D.B. Correction of mass spectral drift using artificial neural networks. *Anal. Chem.* 68:271-280, 1996.

Goodacre, R., Kell, D.B., and Bianchi, G. Neural networks and olive oil. *Nature* 359:594, 1992.

Goodacre, R., Kell, D.B., and Bianchi, G. Rapid assessment of the adulteration of virgin olive oils by other seed oils using pyrolysis mass spectrometry and artificial neural networks. *Journal of the Science of Food and Agriculture* 63:297-307, 1993.

Goodacre, R., Neal, M.J., and Kell, D.B. Rapid and quantitative analysis of the pyrolysis mass spectra of complex binary and tertiary mixtures using multivariate calibration and artificial neural networks. *Analytical Chemistry* 66:1070-1085, 1994.

Goodacre, R., Neal, M.J., and Kell, D.B. Quantitative analysis of multivariate data using artificial neural networks: a tutorial review and applications to the deconvolution of pyrolysis mass spectra. *Zentralblatt. für Bakteriologie* 284:516-539, 1996.

Goodacre, R., Neal, M.J., Kell, D.B., Greenham, L.W., Noble, W.C., and Harvey, R.G. Rapid identification using pyrolysis mass spectrometry and artificial neural networks of Propionibacterium acnes isolated from dogs. *J. Appl. Bacteriol.* 76:124-134, 1993.

Goodacre, R., Neal, M.J., Kell, D.B., Greenham, L.W., Noble, W.C., and Harvey, R.G. Rapid identification using pyrolysis mass spectrometry and artificial neural networks of Propionibacterium acnes isolated from dogs. *J. Appl. Bacteriol.* 76:124-134, 1994.

Goodacre, R., Timmins, E.M., Jones, A., Kell, D.B., Maddock, J., Heginbothom, M.L., and Magee, J.T. On mass spectrometer instrument standardization and interlaboratory calibration transfer using neural networks. *Analytica Chimica Acta* 348:511-532, 1997.

Goodacre, R., Timmins, É.M., Rooney, P.J., Rowland, J.J., and Kell, D.B. Rapid identification of Streptococcus species using diffuse reflectance-absorbence Fourier transform infrared spectroscopy and artificial neural networks. *FEMS Microbiol. Lett.* 140:233-239, 1996.

Goodacre, R., Trew, S., Wrigley-Jones, C., Neal, M.J., Maddock, J., Ottley, T.W., Porter, N., and Kell, D.B. Rapid screening for metabolite overproduction in fermentor broths, using pyrolysis mass spectrometry with multivariate calibration and artificial neural networks. *Biotechnol. Bioeng.* 44:1205-1216, 1994.

Kell, D.B., and Davey, C.L. On fitting dielectric spectra using artificial neural networks. *Bioelectrochemistry and Bioenergetics* 28:425-434, 1992.

Kell, D.B., and Sonnleitner, B. GMP - Good Modelling Practice: an essential component of good manufacturing practice. *Trends Biotechnol.* 13:481-492, 1995.

Mendes, P., and Kell, D.B. On the analysis of the inverse problem of metabolic pathways using artificial neural networks. *Biosystems* 38:15-28, 1996.

Neal, M.J., Goodacre, R., and Kell, D.B. On the analysis of pyrolysis mass spectra using artificial neural networks. Individual input scaling leads to rapid learning. In: Proceedings of the World Congress on Neural Networks, vol. I, pp. 318-323, International Neural Network Society, San Diego, 1994.

Nicholson, D.J., Kell, D.B., and Davey, C.L. Deconvolution of the dielectric spectra of microbial cell suspensions using multivariate calibration and artificial neural networks. *Bioelectrochemistry and Bioenergetics* 39:185-193, 1996.

Shaw, A.D., diCamillo, A., Vlahov, G., Jones, A., Bianchi, G., Rowland, J., and Kell, D.B. Discrimination of the variety and region of origin of extra virgin olive oils using C-13 NMR and multivariate calibration with variable reduction. *Analytica Chimica Acta* 348:357-374, 1997.

Winson, M.K., Goodacre, R., Timmins, É.M., Jones, A., Alsberg, B.K., Woodward, A.M., Rowland, J.J., and Kell, D.B. Diffuse reflectance absorbence spectroscopy taking in chemometrics (DRASTIC). A hyperspectral FT-IR-based approach to rapid screening for metabolite overproduction. *Analytica Chimica Acta* 348:273-282, 1997.

Woodward, A.M., Jones, A., Zhang, X., Rowland, J., and Kell, D.B. Rapid and non-invasive quantification of metabolic substrates in biological cell suspensions using nonlinear dielectric spectroscopy with multivariate calibration and artificial neural networks. Principles and applications. *Bioelectrochem. Bioenerg.* 40:99-132, 1996.