# INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

**The quality of this reproduction is dependent upon the quality of the copy submitted.** Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

# Classification and Identification of Yeasts by Fourier Transform Infrared Spectroscopy

**Jianming Zhao**

**Department of Food Science and Agricultural Chemistry**
**McGill University, Montreal, Quebec, Canada**

**November, 2000**

**A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements of the degree of Master of Science.**

# Table of Contents

**Abstract**

Infrared spectra of microbial cells are highly specific, fingerprint-like signatures which can be used to differentiate microbial species and strains from each other. In this study, the potential applicability of Fourier transform infrared (FTIR) spectroscopy for the classification of yeast strains in terms of their biological taxonomy, their use in the production of wine, beer, and bread, and their sensitivity to killer yeast strains was investigated. Sample preparation, spectral data preprocessing methods and spectral classification techniques were also investigated. All yeast strains were grown on a single growth medium. The FTIR spectra were baseline corrected and the second derivative spectra were computed and employed in spectral analysis. The classification accuracy was improved when the principal component spectra (calculated from the second derivative spectra) were employed rather than the second derivative spectra or raw spectra alone. Artificial neural network (ANN) with 10 units in the input layer and 12 units in the hidden layer produced a robust prediction model for the identification of yeasts. Cluster analysis was employed for the classification of yeast strains in terms of their use in the production of wine, beer, and bread and in terms of their sensitivity to killer yeast strains. The optimum region for the classification in the former case was found to be between 1300 and 800 $cm^{-1}$ in the infrared spectrum whereas the optimum region for the classification of yeast strains in terms of their sensitivity was between 900 and 800 $cm^{-1}$. The results of this work demonstrated that FTIR spectroscopy could be successfully employed for the classification and identification of yeast strains with minimal sample preparation.

**Résumé**

Les spectres infrarouges des cellules microbiennes sont fortement spécifiques, des signatures comparables à des empreintes sont employés pour différencier les espèces et les types microbiens. Dans cette étude, l'application potentielle de la spectroscopie infrarouge par transformation Fourier (FTIR) pour la classification des types de levure dans la limite de leur taxonomie biologique, leur utilisation dans la production du vin, de la bière, et du pain et leur sensibilité aux types de levures destructrices a été étudiée. La préparation des échantillons et les méthodes de prétraitement de classification spectrales ont également été étudiées. Tous les types de levure ont été cultivées dans un milieu de croissance simple. Les spectres FTIR à base corrigée ainsi que leur seconde dérivée ont été utilisées dans l'analyse. La classification a été améliorée quand les spectres principaux (calculés à partir de la seconde dérivée des spectres) ont été utilisés plutôt que la seconde dérivée ou les spectres d'origine. Une classification a été réalisée par l'utilisation d'un réseau neurologique artificiel (ANN) avec une combinaison optimale de 10 unités de couches d'entrée et 12 unités de couches cachées. L'analyse multivariable a été utilisée lors de la classification des types de levure par leur utilisation dans la production du vin, de la bière, et du pain et par leur sensibilité aux levures destructrices. La région optimale pour la première méthode de classification s'est avérée être entre 1300 et 800 $cm^{-1}$ tandis que la région optimale pour la classification des types de levure par leur sensibilité était entre 900 et 800 $cm^{-1}$. Les résultats de ce travail ont démontré que la spectroscopie FTIR pourrait, avec succès, être utilisée pour différencier les types de levure avec une préparation minimale des échantillons.

## ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my graduate research supervisor Dr. Ashraf A. Ismail for his excellent guidance and financial support throughout this research and thesis preparation, his patience, encouragement and assistance are deeply appreciated. I also wish to thank Dr. F. R. van de Voort and Dr. Jacqueline Sedman for their invaluable advice.

I am also grateful to Dr. Alain Houde, Dr. Linda Saucier and Mrs. France Dussault from CRDA and Ms. Ann Dumount from Lallemand Inc. for supplying the bacteria strains and their wonderful cooperation.

I would like to thank all the faculty members and all fellow graduate students for their friendship and their help in the Department of Food Science and Agricultural Chemistry. Special thanks to Mrs. Lise Stiebel and Mrs. Barbara Laplaine for their nice help throughout my graduate study. Thanks Ziad Khoury for translating the abstract into French.

I want to take this opportunity to express my gratitude to my parents, my wife and my sister whose support and encouragement made this work possible.

## List of Figures

## List of Tables

# Chapter 1

## Introduction

The characterization of microorganisms (including detection, differentiation, identification, and susceptibility testing against antibiotics) is very important in a wide variety of industries. For example, in the pharmaceutical manufacturing industry microbes are either part of the manufacturing process or they interfere with or contaminate the process. Numerous analyses are regularly performed in the medical research institutions dedicated to the registration and epidemiological control of pathogens of both humans and animals and rigorous microbiological controls of raw material are also needed. Therefore, proper and rapid characterization of microorganisms is highly desirable. While morphological and biochemical techniques have been traditionally employed, in recent years, new methods such as PCR, have been adopted by the food industry. More recently, the application of GC/MS, pyrolysis for microorganism characterization also has been under active investigation.

In this context, Fourier transform infrared (FTIR) spectroscopy has the potential to become an important routine analytical tool as FTIR analysis can be performed rapidly with minimum sample preparation and without the use of reagent. It has been reported in the literature that FTIR spectroscopy can be employed in the classification and differentiation of microorganisms (Naumann et al., 1988; Helm et al., 1991; Goodacre et al., 1996b), to detect in situ intracellular compounds (Naumann, 1998a), to characterize growth dependent phenomena of microorganism (Reinstadler et al., 1997) and to monitor

chemical changes taking place during fermentation (Fayolle P. et al., 1997; Qiu, J. et al., 1999). Accordingly, a number of time-consuming morphological and biochemical tests may be replaced by FTIR spectroscopy. However, widespread application of FTIR spectroscopy for microorganism characterization will likely occur if the industry is provided with evidence that such an approach is both reliable and accurate. This thesis work addresses issues related to the development of a reliable and rapid method for the characterization of yeasts by FTIR spectroscopy.

# Chapter 2

## Literature Review

### 2.1 Biochemical Composition of Microorganisms

#### 2.1.1 Biochemical composition of bacteria cells

Most bacteria appear in variations of three different shapes: the rod (known as bacillus), the sphere (coccus) and the spiral (virions, spirilla, and spirochetes). To achieve motion, they utilize structures called flagella, which are composed of long, rigid strands of a protein called flagellin. Bacteria also possess appendages called pili that appear as short flagella. They are composed of proteins, which can anchor bacteria to surfaces or transfer genetic material among bacteria. The capsule (composed of polysaccharides and small proteins that adhere to the bacterial surface) serves as buffer between the cell and its external environment. Because of its high water content, the capsule protects the cell against dehydration while preventing nutrients from flowing away. The important chemical constituent of the bacterial cell wall is peptidoglycan. It is a very large molecule composed of alternating units of two amino-containing carbohydrates, N-acetylglucosamine and N-acetylmuramic acid, joined by cross-bridges of amino acids. Peptidoglycan occurs in multiple layers connected by side chains of four amino acids, and the many layers comprise one extremely large molecule.

The cell walls of Gram-positive and Gram-negative bacteria differ considerably (Figure 2.1). In Gram-positive bacteria, the peptidoglycan layer is about 25 nm wide and contains an additional polysaccharide called teichoic acid. About 60 to 90 percent of the

cell wall is peptidoglycan, and the material is so abundant that Gram-positive bacteria are able to retain the crystal violet-iodine complex in Gram staining. By contrast, Gram-negative bacteria have a peptidoglycan layer only 3 nm wide without any evidence of teichoic acid. The cell wall in these bacteria contains various polysaccharides, proteins, and lipids and so is much more complex than the cell wall of Gram-positive bacteria. Also, the cell wall is surrounded by an outer membrane barely separated from the cell wall by a so-called periplasmic space containing a gel-like material called

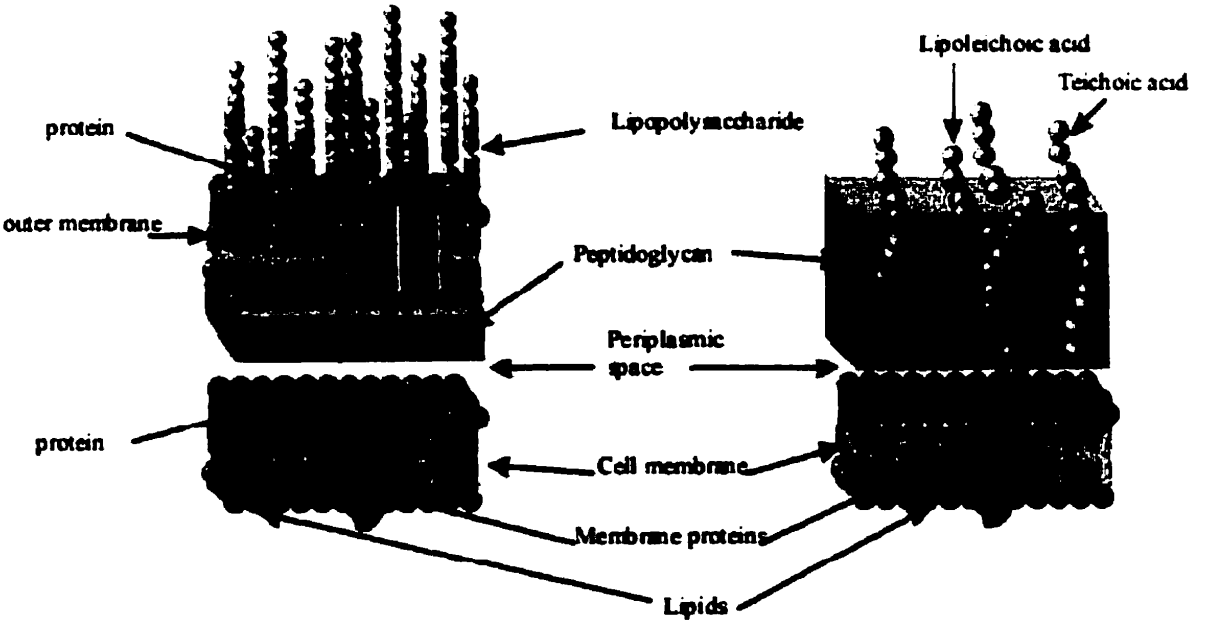


Figure2.1 Comparison of gram-positive and gram-negative bacterial cell wall structures ( adapted from Maier, Raina M. 2000 Environmental microbiology San Diego, Calif. Academic Press)

periplasm (which is made of proteins. In this compartment there are some metabolic activities - e.g. reactions dealing with toxic substances). On the inner side of the cell wall the periplasmic space is wider. Bacterial toxins and enzymes apparently remain in this

space and destroy antibacterial substances before they can affect the cell membrane, and other proteins.

The cell membrane (plasma membrane) is the boundary layer of the bacterial cell. Approximately 60 percent of it is composed of protein, and about 40 percent of lipid, mainly phospholipid. The phospholipid molecules are arranged in two parallel layers. Inside the cell membrane lies the cytoplasm, a gelatinous mass of proteins, carbohydrates, nucleic acids, salts, and inorganic ions. Certain Gram-positive bacteria are able to produce highly resistant structures called endospores or spores. Spores contain little water, however, they do have a large amount of dipicolinic acid which helps to stabilize their proteins.

### 2.1.2 Biochemical composition of yeast cells

Budding yeasts are true fungi of the phylum *Ascomycetes*, class *Hemiascomycetes*. The true yeasts belong to one main order *Saccharomycetales*, which includes at least ten families. Yeasts are heterotrophic, lack chlorophyll, and have a wide variety of natural habitats. Yeasts multiply as single cells that divide by budding or direct division (fission), or they may grow as simple irregular filaments (mycelia). In sexual reproduction most yeasts form asci, which contain up to eight haploid ascospores. These ascospores may fuse with adjoining nuclei and multiply through vegetative division or, as with certain yeasts, fuse with other ascospores (Kockova-Kratochvilova, 1990).

The chemical composition of yeasts such as brewer's, baker's, wine and fodder yeast differs widely (Table 2.1 and Table 2.2). These differences reflect differences in the yeast species, cultivation conditions and nutrient media (Kockova-Kratochvilova, 1990).

Table 2.1 *The chemical composition of brewer's yeast biomass:*

| Composition | Brewer's yeast biomass |
|---|---|
| carbon | 44 to 50% |
| hydrogen | 6 to 8% |
| nitrogen | 8 to 12% |
| oxygen | 30 to 36 % |

Table 2.2 *The main components of Baker's yeast:*

| Components | Baker's yeast |
|---|---|
| protein | 45 to 60% |
| saccharides | 25 to 35% |
| lipids | 4 to 7% |
| ash | 6 to 9% |

The guanine and cytosine (G+C) content of yeasts ranges from approximately 28 to 70 mol%. The G+C content of ascomycetous yeasts is generally less than 50%, whereas that of basidiomycetous yeasts is generally above 50% (Kurtzman et at., 1983). From the base composition, the taxonomic class of imperfect yeasts can be reliably

inferred. The range in G+C content among species within a genus is often 10% or less, with the exception of some obviously heterogenous genera. On a species level, the use of G+C content is only exclusionary in that a difference of 2 mol% or greater indicates strains belonging to different species (Price et al., 1978).

Polysaccharides in yeast cells fall topologically and functionally into two classes: cell wall polysaccharides (*e.g.* glucans and mannans) and intracellular polysaccharides. Glucan and mannan complexed with proteins represent about 80 to 90 % of the cell wall dry weight in S. *cerevisiae.* The rest is made up by chitin, proteins and lipids. The chemical structure of mannan (Figure 2.2) consists of mannose units bonded by $\alpha 1 \rightarrow 6$, $\alpha 1 \rightarrow 2$ and $\alpha 1 \rightarrow 3$ bonds. Isolated preparations of glucans from yeast cell walls are extremely heterogeneous (Manners et al., 1974). The major part is formed by insoluble β-1,3-glucan with a high relative molecular weight and a polymerization degree of about 1500, which contains 3 % of β- 1,6-glycosidic bonds inside the chain. A minor component, about 15 % of total glucans, is a soluble β-1,6-glucan with a polymerization degree of 130 to 140, containing about 14 % β- 1,3-glycosidic bonds inside the chain. Intracellular saccharides, mostly glycogen and trehalose, serve as reserve substances. Glycogen makes up 0.5 to 1.3 % of the yeast cell weight. Its properties are similar to those of the amylopectin starch fraction. It is composed of chains of glucose residues with predominantly $\alpha 1 \rightarrow 4$ type, bonds of the $\alpha 1 \rightarrow 6$ type being only localized at the chain branching points. Trehalose (α-D-glucopyranosyl-α-D-glucopyranoside, Figure 2.3) consists of two glucose units. The activity of trehalose is affected by cyclic AMP. Van Solingen and Van der Plaat (1975) found that the lag phase preceding the culture

growth is governed by a system including the action of cAMP and trehalose. In yeast cells cAMP acts as a regulator of protein-phosphorylating reactions. The activation of trehalose is associated with phosphorylation of the protein, which is controlled by cAMP.

Yeast intracellular lipids include neutral triacylglycerols and phospholipids (Table 2.3, Table 2.4). Yeast also produce lipid into the external medium or cultivation medium (extracellular lipids). There are four types of extracellular lipids: a) *esters of polyols and carboxylic acids* in which saturated and unsaturated hydroxycarboxylic acids are linked with five- to six-carbon polyols by an ester bond; *b) sophorosides of hydroxycarboxylic acids* in which saturated and unsaturated hydroxycarboxylic acids are linked by a glycosidic bond to the disaccharide sophorose; *c) acetylated sphingosines in* which hydroxy groups and amino groups of $C_{18}$-phytosphingosine and $C_{18}$-dihydrosphingosine are acetylated; *d)* $C_{22}$*-acids* in which tri- and dihydroxycarboxylic acid residues are acetylated.

Table 2.3: Composition of yeast lipids (adapted from Kockova-Kratochvilova, 1990)

| Carboxylic acid | Content in lipids of *C. utilis* [%] |
|---|---|
| Lauric | 0.5 |
| Myristic | 1.3 |
| Palmitic | 21.0 |
| Palmitoleic | 3.5 |
| Stearic | 2.9 |
| Oleic | 40.0 |
| Linoleic | 26.5 |
| Linolenic | 3.5 |

Table 2.4 Percent total phospholipid content in whole cells of *S.cerevisiae* (Cartledge et al., 1977)

| Phospholipid | Per cent of total phospholipid content in whole cells of *S. cerevisiae* |
|---|---|
| Phosphatidylethanolamine | 31.2 |
| Phosphatidylinositol | 29.7 |
| Phosphatidylcholine | 25.3 |
| Phosphatidylserine | 6.2 |
| Cardiolipin | 3.8 |
| Phosphatidic acid | 2.3 |

Yeast cells are multilayer systems in which membranes delineate separate reaction spaces. Membranes are assumed to serve as diffusion barriers between individual compartments. Yeast membranes include the dictyosomal membrane, nuclear membrane, ER-membrane, vacuolar tonoplast, mitochondrial membrane, and microsomal membrane. Yeast membranes contain a number of lipids and pigments that are not present in prokaryotic cells. These include sterols, sphingolipids, ergosterins, melanins, and some glycolipids. Culture conditions have a marked influence on the total lipid content and

lipid composition of yeasts. Factors controlling lipid content and composition are pH of the medium, temperature, and time of growth, and the ratio of N- and C-sources. Sterols occur both in free form and as esters with long-chain fatty acids. Both forms are interconvertible. Free sterols are associated with membrane functions, and sterol esters may fulfil a storage or "pool" function. Common sterol molecules of yeasts are ergosterol (Figure 2.4), lanosterol, and episterol, zymosterol, and fecosterol (Nes et al., 1978). Major structures of sphingolipids found in yeasts are the sphingosines (Figure2.5), cerebrins (ceramides), sphingomyelines, and cerebrosides (Kockova-Kratochvilova, 1990). A typical membrane lipid in yeast is ergosterin (Kockova-Kratochvilova, 1990). Its structure is similar to that of cholesterin and it belongs to the group of sterines. Further compounds frequently found in the membranes of yeasts are melanins, which are black pigments built up from tyrosine derivatives.

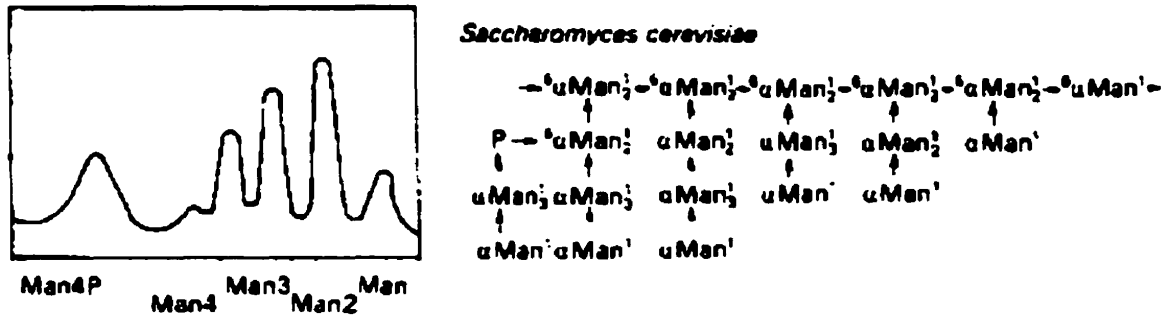


Figure 2. 2 The chemical structure of mannans from *Saccharomyces cerevisiae*

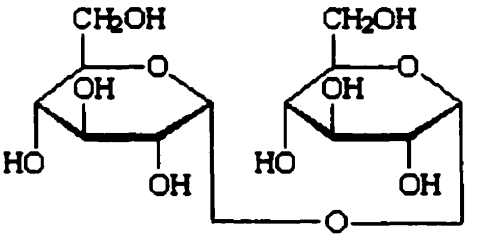


Figure2.3 Chemical structure of trehalose

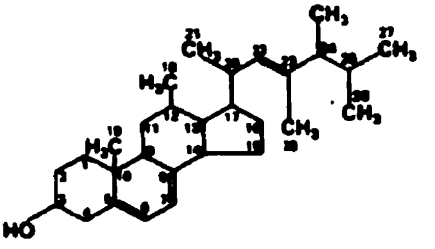


Figure 2.4 Chemical structure of ergosterol

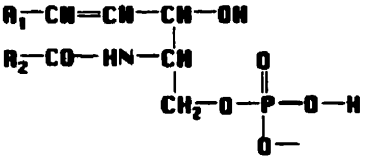


Figure 2.5 Chemical structure of sphingosines

## 2.2 Current techniques for microorganism characterization

Traditional microorganism identification procedures rely heavily upon the morphology of vegetative cells and sexual reproduction, including ultrastructural studies of cell walls, septae, and spores. Standard identification procedures include several physiological and biochemical tests to determine the ability of the isolate to ferment sugars and assimilate aerobic growth on various carbon and nitrogen compounds. In addition, conditions necessary for growth and demonstration of characteristics are important in the identification process. Standard physiological and biochemical tests are primarily used to determine the species of isolates. The traditional microorganism identification method is the only one acceptable for taxonomic purposes and requires considerable experience and skill in the performance and evaluation of a large number of specified and standardized tests (Bruno P. 2000).

Interest in the identification of clinically important microorganisms paved the way for the development of commercial ready-to-use systems in various microwell formats and provided a stimulus for further development of automated identification systems. Several miniaturized kits and systems have been developed and marketed over the past 20 years, such as e API 20A®, API 20C®, Minitek Anaerobe II, Rapid ID 32A, AN-Ident RapID ANA II, VITEK ANI Card, MicroScan Rapid Anaerobe Panel, Uni-Yeast Tec®, Abbott Quantum II®, Vitek ATB32®, and AutoMicrobic®. Some are designed to be used manually, while others are automated to various degrees (Table 2.5). Several systems require additional tests and/or morphological investigations. Stager and Davis (1992) provided an excellent overview and evaluations of commercial kits and automated

systems. Most commercially available systems provide accurate and reliable results, giving 90% or more agreement with data obtained by traditional methods. The API 20C system is probably the most widely used in yeast identification and has often been considered as a reference method for evaluating other systems (Polacheck et al., 1987; St.-Germain and Beauchesne, 1991). All of these commercially available identification systems were designed to meet the needs of the clinical microbiological laboratory. For this reason their databases are restricted to a limited number of species of clinical importance. The most reliable commercial systems could be used for the identification of large groups of microorganism if their databases were extended and certain additional tests were performed.

Phenotypic characterization of closely related organisms is not always a reliable method for microorganism differentiation. In the last two decades the application of molecular techniques has had a major impact on the classification of yeasts. The nuclear DNA relatedness has become the basis of species delineation. Molecular fingerprinting methods such as analysis of restriction fragment length polymorphisms, random amplified polymorphic DNA, PCR-amplified sequences and fragments, pulsed field gel electrophoresis of chromosome DNA and others allow intraspecies differentiation and typing. The most far-reaching method has been the sequencing of various parts of ribosomal DNA that has made it possible for the first time to assess the phylogenetic relationships among yeasts at different taxonomic levels. DNA fingerprinting techniques describe those procedures that provide a unique profile of the DNA of a given organism. Guanine + cytosine (G+C) ratios (relative to adenine + thymine [A+T]) is a good method

for microorganism identification, as is the base sequence of the chromosome method. Forbes and Hicks (1993) and Luk (1994) devised detection methods for *Mycobacterium tuberculosis* and *Salmonella typhi* based on the G+C ratio, respectively. Since 16S rRNA is derived from the DNA sequence, it can also be used for the differentiation of microorganisms (Gutell et al. 1994). Techniques revealing restriction fragment length polymorphism (RFLP) have proved useful in the taxonomic evaluation of yeast genera and species and can also be used to identify strains within the same yeast species (Pedersen, 1986). Degre et al. (1989) compared RFLP patterns with protein and fatty acid profiles as well as with chromosome karyotyping and found that DNA fingerprinting provided the most reliable method for characterizing wine yeast strains. Restriction endonuclease treatment of mitochondrial DNA (mtDNA) has also been used in the differentiation of yeast species (Vezinhet et al., 1990). The usefulness of DNA probes in conjunction with restriction analysis of DNA and/or chromosome karyotyping may lie not only in the possibility of identifying certain taxons (del Castillo Agudo et al., 1993 ) and detecting specific pathogenic biotypes (Scherer and Stevens, 1987) but also for identifying specific industrial bacteria strains to assure quality control or protect proprietary rights (Pretorius and van der Westhuizen, 1991). Upon developing the polymerase chain reaction (PCR) technique (Foster et al., 1993), new opportunities for the design of diagnostic procedures arose. The value of PCR in facilitating sequence analysis has been applied to taxonomic and phylogenetic analysis of yeasts (Barns et al., 1991; Molina et al., 1992). This technique also lends itself to species identification (Deak, T. 1999; Torriani, S. 1999).

Molecular probe technology is based on the binding of a molecule --the probe-- with a particular microorganism, virus, or an individual and unique component, and the detection of the probe-target complex. Most commonly, probes are nucleic acid molecules for the detection of either DNA or RNA, or are antibody molecules for the detection of proteins, carbohydrates, polysaccharides, or lipids. Nucleic acid probes are single-stranded DNA or RNA molecules. Detection is based on the formation of a "hybrid," between the nucleic acid probe and single-stranded DNA or RNA recovered from a microorganism or virus or a sample containing both (Macario and deMacario 1990). Antibodies are proteins produced by mammals that are capable of binding and forming complexes with different molecules, called antigens. Antigens can be proteins, polysaccharides, or lipids. Even molecules that are normally unable to elicit antibody formation in mammals can be made antigenic by coupling with another molecule, called a hapten. A variety of different antibodies, each able to bind and form a complex with a specific antigen, can be produced by a mammal. Since these antibody-forming cells cannot be propagated in a culture medium and a single reactive antibody molecule is preferred as a probe, a specific type of antibody-forming cell, a hybridoma, is employed for both the selection and production of the desired antigen-specific antibody probe (Harlow and Lane 1988).

The use of chemical analysis of microbial components (i.e., lipids, polysaccharides, proteins, and nucleic acids)--chemotaxonomy--has been increasingly applied to bacterial taxonomy (Brondz and Olsen, 1986). Analytical methodologies utilized included gel electrophoresis, orthogonal-field-gel electrophoresis,

spectrophotometry, proton magnetic resonance, high-performance liquid chromatography, gas chromatography, combined gas chromatography-mass spectrometry and pyrolysis-mass spectrometry techniques. Merz et al. (1988) used orthogonal-field-alternation gel electrophoresis to establish electrophoretic karyotypes for strains of *Candida albicatis*. They detected much greater strain variation than revealed by existing biotyping techniques, thus expanding the scope of epidemiological studies. Timmins et al. (1998a) used pyrolysis-mass spectrometry to analyze a group of 29 clinical and reference Candida isolates.

The development of PFGE (pulsed field gel electrophoresis) techniques has led to descriptions of electrophoretic karyotypes for several microorganism species (Boekhout et al., 1993; Vaughan-Martini and Martini, 1993). PFGE data obtained by various techniques have revealed that variability in chromosome size among strains of the same species is common and that chromosome polymorphism can be used for differentiating and/or identifying industrial microorganisms such as wine yeast (Yamamoto et al., 1991; Vezinhet et al., 1992). To test polymorphism and evaluate electrophoretic karyotypes more effectively, the PFGE technique is usually combined with DNA-DNA hybridization. Electrophoretically separated bands are blotted onto membranes and hybridized by labeled probes (Torok et al. (1992, 1993)). The main advantage of PFGE is its discriminatory power and relatively simple banding patterns, but long and laborious DNA isolation procedures and digestion of the samples mean that results may take from a few days up to a week to obtain (Maslow et al., 1994; Matushek et al., 1996).

Bruneau and Guinet (1989) applied electrophoretic protein patterns (polyacrylamide gel electrophoresis (PAGE), with or without sodium dodecyl sulfate (SDS) technique) for the identification of medically important yeasts and concluded that the method allowed good species discrimination, but preparation of extracts was time-consuming. However, Degre et al. (1989) indicated that the drawback of protein electrophoresis is that it depends on growth conditions. Gas chromatography of cellular volatile fatty acids requires relatively expensive instrumentation and lengthy preparatory work. Under standardized cultivation and analytical conditions, however, volatile fatty acid analysis (VFAA) can be a reliable method for characterization of microorganisms. (Botha and Kock, 1993).

Flow cytometry measures physical and chemical characteristics of cells that are suspended in a liquid and pass singly by one or more optical sensors. It is now in common use for classifying normal and tumor cells, blood cells, and cells from the reticuloendothelial system; it has been applied by researchers in a wide range of other fields, including bacteriology, protozoology, microbial ecology, and pharmacology. Flow cytometric techniques have become increasingly important in diagnostic procedures (Kleine et al. 1990). Bassoe and Bjerknes (1985) and Bassoe et al. (1983) described the use of flow cytometry in measurement of the phagocytosis of bacteria by leukocytes and proposed that such measurements could prove useful in clinical studies for the assessment of cell-mediated immune function of patients suffering from the effects of severe burns or chronic infections. Flow cytometry was also investigated by Pinder et al. (1990) as a rapid detection and counting method for bacteria in pure cultures.

All the methods described above have certain advantages and drawbacks. Some have high sensitivity (e.g., pulsed field gel electrophoresis and PCR) but are time-consuming and expensive and require trained professionals. Other methods are simpler but cannot differentiate between microorganisms down to the strain level. There is an ongoing effort to develop new methods that are sensitive, reliable, rapid and cost-effective. At present, a number of groups are working on spectroscopy-based methods including infrared spectroscopy, which will be discussed in the next section.

Table 2.5 Characteristics of some commercially available yeast identification systems (Adapted from TiBor Deak et al., "Handbook of Food Spoilage Yeasts", CRC Press, 1996)

| Principle | System | Method | No. of tests | No. of species in database | Time (h) required for result | Accuracy (%) |
|---|---|---|---|---|---|---|
| Growth based | API 20C® | Manual | 20 | 42 | 72 | 99 |
| | ATB 32 ID® | Manual/automated | 32 | 63 | 48 | 91 |
| | AutoMicrobic® | Automated | 30 | 62 | 24 | 83 |
| | Microring YT® | Manual | 6 | 18 | 48 | 53 |
| | Minitek® | Manual | 12 | 28 | 72 | 97 |
| | Quantum II® | Automated | 20 | 34 | 24 | 82 |
| | Uni-Yeast-Tek® | Manual | 15 | 42 | 48 | 40 |
| Enzyme based | MicroScan® | Manual/automated | 27 | 42 | 4 | 85 |
| | YeastIdent® | Manual | 20 | 42 | 4 | 55 |

## 2.3 Basic principle and application of FTIR spectrometers

There are two basic types of infrared spectrophotometers, characterized by the manner in which the infrared frequencies are handled. In the first type, infrared light is separated into its individual frequencies by dispersion, using a grating monochromator, whereas in the second type the infrared frequencies are modulated to produce an interference pattern. A Fourier transform infrared spectrometer based on the latter principle provides improved speed and sensitivity and unparalleled wavelength precision and accuracy relative to a grating spectrometer (Borman, S. A. 1983).

The basic components of a FTIR spectrometer are a source, an interferometer, a detector, and a laser. A computer is required for controlling optical components, collecting and storing data, performing signal averaging, carrying out the Fourier transformations and displaying spectra. The heated source gives off infrared radiation, which is deflected off a mirror into the interferometer where the spectral encoding takes place. The detector is the device which produces an electrical signal in response to the encoded radiation striking it. The most commonly used detector material in the mid-infrared is deuterated triglycine sulfate (DTGS). The DTGS detector is known as a pyroelectric bolometer. The advantages of DTGS detectors are that they are simple, inexpensive and robust. The vast majority of FTIR spectrometers employ DTGS detectors. The major drawback of DTGS detectors is that they are less sensitive than other detectors available. The more sensitive detectors cooled by liquid $N_2$ are the mercury cadmium telluride (HgCdTe) or "MCT" detectors. The MCT element consists of an alloy of these three elements, and it is a semiconductor. The major advantage of MCT

detectors is their sensitivity. They are up to 10-50 times more sensitive than DTGS detectors. Unfortunately, there is a tradeoff between bandwidth and sensitivity with MCT detectors. The most sensitive detectors are the narrow band ones, which are useful from 4000 to 700 $cm^{-1}$. Wide band MCTs go down to 400 $cm^{-1}$, but are 5-10 times noisier than the narrow band MCT detector. In many applications, the wide band MCT represents only a modest improvement in sensitivity over a conventional DTGS detector. Another advantage of MCT detectors is that they are fast. As a result, one can scan many times faster than with a DTGS detector, and obtain spectra with high signal-to-noise ratio (SNR) faster. A drawback to MCT detectors is that they must be cooled. Without this cooling, heat given off by the detector element itself is detected, giving rise to a large noise signal.

The techniques used to acquire and analyze infrared spectra continue to grow. Attenuated total reflectance (ATR) accessories, using single bounce and multiple bounces, have been widely used in acquiring IR spectra of biological and chemical samples (Banwell, 1983). ATR is now used extensively in the study of tissues, microbial and human cells, and body fluids and in investigations of isolated components such as proteins and peptides involved in pathologic disorders. ATR-based fiber optic probes have also been developed and are useful for on-line monitoring of chemical reactions. More recently, FTIR spectrometers based on photoacoustic measurements (Drapcho et al., 1997) have been developed for depth profiling of samples. The combination of infrared spectrometers with optical microscopes is surely the most significant advance in the field of biomedical application of FTIR spectroscopy. Infrared microscopes using

single detectors or array detectors operating in transmission or reflection are of great value in the study of cell populations, for example in histological section, because they allow the focus of the IR beam on specific areas of interest. The advances of step-scan instruments have allowed improvements in the time and spatial resolution capabilities of infrared spectroscopy. For example, step-scanning FTIR photoacoustic spectroscopy has been used to perform depth profiling studies on polymeric multilayers (Urban et al., 1998; Jiang, 1998), single particles and fibers (Jiang, 1999), and organic reactions and catalysts (Frei, 1998). The beam qualities associated with synchrotron light sources also allow for improvements in spatial resolution beyond the current capabilities using standard sources. It is the beam attributes of low thermal noise, brightness, low divergence, and excellent signal-to-noise ratio that make these IR sources unique. These attributes make it possible to perform experiments where small aperturing is important or sample scattering normally precludes IR spectroscopy. Synchroton light sources have been used in studies of inorganic-organic interaction at the bacterial-mineral interface (Holman et al., 1998). IR imaging techniques enable mapping of chemical functionality. The combination of infrared microscopy instrumentation with confocal plane array detectors produces what are known as infrared imaging systems. Spatial chemical functionality information has been available through mapping techniques using motorized translation stages and infrared spectroscopy. The advent of focal-plane array (FPA) MCT detectors dramatically reduced the analytical time. Marcott et al., (1997) employed an FPA/FTIR imaging system to examine the cross section of a laminated polymer film and human tissue. The major advantage of this new instrumentation is the coupling of noninvasive infrared chemical analysis and visualization. The latter is

extended beyond the visualization of the infrared image when a CCD camera is added to the system, allowing the simultaneous visualization of the physical image.

## 2.4 Infrared Sampling Techniques

In general, the acquisition of infrared spectra of biological samples can be problematic. In order to obtain reproducible spectra, sampling conditions have to be controlled and standardized rigidly. There is no simple and universally applicable technique to meet these requirements. However, depending on the nature of the sample, these requirements can be fulfilled by using traditional transmission, reflectance, diffusion or attenuated total reflection techniques.

Transmission Spectroscopy:

The transmission technique is the simplest sampling technique in optical spectroscopy and is recommended for routine spectral measurements for all kinds of samples. In this technique the sample is placed in the light beam of a spectrometer and the intensity of the incident beam is compared with the intensity transmitted by the sample. For an incident beam of intensity $I_0$ the transmitted intensity I is given by

$$I = I_0 * 10^{-abc} \qquad \text{Eq. 2.1}$$

where $a$ is the absorptivity, $b$ is the sample thickness, and $c$ is the concentration. Equation 2.1 assumes that there is no loss of intensity due to light scattering or reflection.

In all cases sample thickness must be adjusted. Liquid samples require short optical pathlengths (0.025-1 mm), because organic molecules have strong infrared

absorption. Spacers are used to control the pathlength. Transmission spectral measurements are more complex for solid samples than for liquids. A thin film or section must be obtained from the sample before the spectrum can be acquired. A Brewster's angle accessory usually is used to reduce the energy losses due to reflection from the sample surface and interference fringes. For powdered samples, different particle size and optical properties can cause the Christiansen effect. The KBr pellet method or mineral oil (Nujol mull) method can be good choices, but both of them have the disadvantage of destroying the sample (Harrick Scientific Corporation ,1987 Optical Spectroscopy: Sampling Techniques Manual).

Internal Reflection Spectroscopy

Internal reflection spectroscopy, also referred to as attenuated total reflectance (ATR) or multiple internal reflectance (MIR), was developed in the 1960's (Harrick Scientific Corporation, 1987 Optical Spectroscopy: Sampling Techniques Manual). In an ATR / MIR measurement, the IR beam from the spectrometer is directed onto a prism at an angle which exceeds the critical angle. As the beam is directed into the crystal at an angle that exceeds the critical angle, internal reflections take place. When a sample is placed in optical contact with the prism at the point at which an internal reflection occurs, the sample absorbs IR energy at wavelengths equivalent to those that would be noted in a transmission experiment.

It has been proposed that the internal reflection generates an evanescent wave which extends beyond the surface of the crystal into a sample held in contact with the

surface. The penetration depth of the electromagnetic wave into the rarer medium is defined by the wavelength-dependent ratio $n_2/n_1$ of the refractive indices of the denser ($n_2$) and the

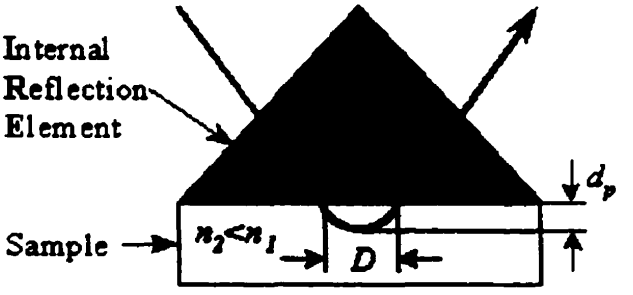


Figure 2.6 Schematic of an Attenuated Total Reflectance accessory

rarer ($n_1$) media and by the angle of incidence $\alpha$ and is on the order of a few micrometers (Figure 2.6). The penetration depth is calculated according the following equation (Eq. 2.2):

$$d_p = \lambda / \{2\pi n_1 [\sin^2(\alpha) - (n_2/n_1)^2]^{1/2}\} \qquad \text{Eq. 2.2}$$

where $\lambda$ is the wavelength of the infrared radiation, $n_1$ is the refractive index of the IRE material, $n_2$ is the refractive index of the sample and $\alpha$ is the angle at which the infrared radiation strikes the IRE interface (Smith, 1996). A property of this wave, which makes ATR such a useful technique, is the exponential decay of the intensity of the wave with the distance from the surface. This makes ATR measurements generally insensitive to sample thickness. Hence, the technique is readily applied to the analysis of strongly absorbing samples.

Solid samples must be mechanically pressed into contact with the crystal to achieve optical contact. The intensity of the bands in the ATR spectrum are a function of optical contact, and the highest degree of reproducibility is achieved when samples are in intimate optical contact with the ATR crystal; that is, when the contact efficiency approaches 100%. When liquids are analyzed by the ATR technique, intimate optical contact is achieved readily. When solids and powders are analyzed, the spectral intensity will be largely governed by optical contact.

A comparison of an internal reflection spectrum and a transmission spectrum reveals the intensities of the bands at high wavenumbers are lower than in an equivalent transmission spectrum. Most spectrometers offer an ATR correction, which increases the intensity of the absorbance by a defined value. This makes it easier to compare the spectrum with libraries of transmission data.

To date, ATR accessories have been successfully used to acquire the spectra of many kinds of biological samples: Borel et al., (1993) examined intact living bacterial cells by ATR/FTIR spectroscopy. They found that typical samples, including both gram-positive and gram-negative bacteria, can be classified and differentiated by this technique. Schmitt et al., (1998) studied different FTIR techniques as a means to investigate microorganisms in biofilms. They reported that the ATR technique could be used for the observation of biofilms forming directly on the surface of a germanium ATR crystal. These crystals can be coated to obtain a surface more relevant to the study of interfacial processes. Spectra can be acquired nondestructively, in situ and in real time.

Suci et al. (1998) reviewed the capabilities of ATR/FTIR spectroscopy to provide information on both transport of an antimicrobial agent to bacteria embedded in the biofilm and interactions between an antimicrobial agent and biofilm components. Doak et al. (1999) used a diamond ATR probe to study *Escherichia coli* fermentation in situ. The probe showed excellent stability over a 6-month operating period and was unaffected by either agitation or aeration.

## 2.5 IR bands assignments of chemical constituents in microorganisms

FTIR spectroscopy provides information not only on the chemical composition of a given bacterial strain but also on the secondary and even tertiary structures of proteins. All of the information can be obtained from the number, relative intensities and band contours of the bands in the IR spectra. Since 1926, many groups (Helm et al., 1991; Naumann et al., 1991a,b, 1998a) have recorded the spectra of microorganisms and published tentative band assignments. Some of the more important assignments are summarized below.

1. The region between 3000 and 2800 $cm^{-1}$ is dominated by C-H stretching vibrations of $-CH_3$, $>CH_2$, and $\equiv CH$ and, hence, by the fatty acids of the various membrane amphiphiles. Some complementary information can be deduced from the region between 1500 and 1400 $cm^{-1}$, where the various deformation modes of the same functional groups are observed, and bands near 1740 $cm^{-1}$ can be assigned to $>C=O$ stretching of the ester functional groups.

2. The region between 1700 and 1500 $cm^{-1}$ is dominated by the so-called amide I and amide II bands of proteins, which are the most intense bands in nearly all bacterial spectra so far tested. Since the characteristic IR absorptions resulting from the DNA-RNA base-ring structures are not as intense as the amide I and II bands, the spectral features observed in this spectral domain are almost completely defined by the protein absorption.

3. In the region between 1500 and 1200 $cm^{-1}$, complex absorption profiles are observed between 1300 and 1500 $cm^{-1}$ arising predominately from $>CH_2$ and $->CH_3$ bending modes of lipids and proteins. A characteristic, but weak feature is often observed near 1400 $cm^{-1}$, which may be caused by the symmetric stretching vibration of $-COO^-$ (functional groups of amino acid side chains or free fatty acids). Around 1230 $cm^{-1}$ superimposed bands typical of different >P=O double bond asymmetric stretching vibrations of phosphodiester, free phosphate and monoester phosphate functional groups are observed.

4. The region between 1200 and 1250 $cm^{-1}$ is "dominated" by different >P=O double bond asymmetric stretching frequencies resulting from the various phosphodiester functional groups. The band near 1220 $cm^{-1}$ is most probably due to the phosphodiester functional groups of DNA/RNA polysaccharide backbone structures. Other >P=O double-bond stretching frequencies are due to head group vibrations of phospholipids or phosphorus–containing carbohydrates such as "teichoic acids" and "lipoteichoic acids".

5. The spectral region between 1200 and ~900 $cm^{-1}$ is dominated by a complex sequence of peaks essentially resulting from C-O-C and C-O-P stretching vibrations of, predominantly, oligo- and polysaccharidic nature. Selective assignments are not yet available because of the extensive superpositions of the characteristic absorptions of various polysaccharides. This region, in particular, turned out to be abundantly endowed with discriminating spectral traits and, thus, represents one of the most sensitive and selective spectral regions for differentiation of microorganisms down to the strain and even serotype level.

6. The region between 900 and 600 $cm^{-1}$ exhibits a variety of weak but extremely characteristic features superimposed on an underlying, rather broad contour. With the exception of only a few peaks (e.g., a band near 720 $cm^{-1}$, resulting from the $>CH_2$ rocking modes of the fatty acid chains present in amphiphilic compounds), valid assignments can hardly be achieved.

## 2.6 Chemometric techniques employed in analyzing infrared spectra

Chemometrics is the discipline concerned with the application of statistical and mathematical methods to chemical data (Massart et al., 1988; Martens, 1999). A variety of powerful methods have been applied to the "unsupervised" and "supervised" analysis of multivariate data. Cluster analysis (CA), principal component analysis (PCA), factor analysis (FA), discriminant analysis (DA), partial-least-squares regression (PLS) and artificial neural networks (ANNs) are most widely used in infrared spectroscopy for quantitative analysis and sample identification.

Based on Fisher's method and incorporated two important validation stages: (1) full leave-one-observation-out cross-validation and (2) randomized permutation distribution testing. Jonathan et al. (1996) developed a computationally efficient approach to perform two-group linear discriminant analysis. The resulting algorithm and software are known as CREDIT (cross-validated random-permutation-tested efficient discrimination based on an adjusted generalized inverse for the sample total covariance matrix). Li et al. (1999) used a real genetic algorithm to develop a high-breakdown method for linear discriminant analysis (LDA). Their algorithm is capable of locating the global optimal solution with high probability and acceptable computational burden. Kemsley (1996) compared partial least squares (PLS) and principal component analysis (PCA) in terms of their data compression ability. He found that PLS had considerably better class separation and discriminant ability. In general, few compressed dimensions are required to give the same level of prediction successes as the full spectrum, and for some data sets, PLS methods yield higher prediction success rates than those obtainable using PCA scores. Wentzell et al. (1997) established a new PCA algorithm: maximum likelihood principal component analysis (MLPCA). The theoretical foundations of MLPCA were initially established using a regression model and extended to the framework of PCA and singular value decomposition (SVD). Generalization of the algorithm allows its adaptation to cases of correlated errors provided that the error covariance matrix is known. Models with intercept terms can also be accommodated.

Several groups offered new PLS algorithms. Cummins and Andrews (1995) introduced iteratively reweighted PLS as a robust method for calibration and

demonstrated its resistance to the effects of outliers with a Monte Carlo study. Zhu and Barnes (1995) reported an iterative version of PLS algorithm that was faster and less memory intensive than PLS implemented with the nonlinear iterative partial least squares (NIPALS) algorithm. Gil and Romera (1998) reported the development of a robust and more efficient PLS algorithm. He stabilized the covariance matrix using the well-known Stahel-Donoho estimator. The prediction error in PLS can be minimized through judicious wavelength selection. Heise and Bitter (1997) demonstrated that multiple linear regression analysis could perform as well as PLS when improved variable selection procedures were used. Spiegelman et al., (1998) developed a theoretical justification for wavelength selection in PLS. Stork and Kowalski (1999) demonstrated the utility of sample weighting for lowering prediction error. Schemes that employ leverage-based criterion for selecting weights and new calibration samples have been described. Thus, fewer samples describing a new source of variation will be needed to update a model. Achievement of a satisfactory calibration model is usually not the final step in the practical application of PLS or any other multivariate calibration method. Once a calibration model is developed, it must be transferred to other instruments, so the calibration can be used at the point of application. Hoffmann and Zanier-Szydlowski (1999) used a Shenk-Westerhaus correction to take into account changes in sample temperature and the field of view of the instrument for PLS models to predict various properties of hydro-treated gas oils. Brown and Wentzell (1998) used a different approach to standardize multivariate calibration models for near-IR FTIR spectrometers equipped with fiber-optic probes. Calibration transfer across instruments and probes was

studied by employing calibration models built on one instrument to predict properties from spectra measured on the other.

The goal of pattern recognition is classification. Developing a classifier from spectral data may be desirable for any number of reasons, including strain identification, presence or absence of disease in an animal or person from which the sample was taken, and food quality testing. During the past several years, some new classification methods were reported in the literature. Smit et al. (1993) noted that drift, which may cause neural networks to misclassify objects when the class clusters lie relatively close to each other, can be corrected using the amount of drift as an extra input variable in the neural network. Radomski et al. (1994) showed that feed-forward neural networks could unambiguously recognize spectra at a signal-to-noise ratio significantly below that needed for by-eye interpretation. Meyer et al. (1993) showed that network architecture could be minimized without a concomitant reduction in prediction performance when the principal component scores of the training and prediction set spectra are input elements for the network. Li and van Espen (1994) observed that neural nets performed better than conventional methods for classification of spectra when network parameters are optimized, e.g., scaling and learning mode, range of the initial weights, and transfer function. Li et al., (1999) developed a robust linear discriminant analysis routine, which has a high breakdown value for outliers. Lavine et al., (1998) developed a genetic algorithm (GA) for pattern recognition analysis of spectroscopic data. The GA selects features that optimize the separation of the classes in a plot of the two largest principal components (PCs) of the data.

### 2.6.1 Hierarchical Clustering

Hierarchical clustering is a widely used algorithm for classification. Its aim is the fusion of N data points into groups. Given a set of N items to be clustered, and an N × N distance (or similarity) matrix, the basic process of Johnson's (1967) hierarchical clustering is: 1, Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain. 2, Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster. 3, Compute distances (similarities) between the new cluster and each of the old clusters. 4, Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. As to the definition of the distance, there are many methods available; in the definitions below, $x_{ij}$ is the value of variable j for object i.

Euclidean: The distance between objects i and k is defined as

$$d_{ik} = \sqrt{\sum_j [x_{ij} - x_{kj}]^2}$$

Eq. 2.3

Pearson: The distance between objects i and k is defined as

$$d_{ik} = \sqrt{\sum_j [x_{ij} - x_{kj}]^2} / S_j$$

Eq. 2.4

$S_j$ = the standard derivation of variable j

### 2.6.2 Discriminant analysis

Another powerful clustering method is discriminant analysis. It is a parametric method that models each class of samples by its centroid and covariance matrix and assigns each object to the closest class. Different discriminant analysis methods are available, such as nearest means classification (NMC), linear discriminant analysis (LDA), which assumes the same covariance structure in each class, quadratic discriminant analysis (QDA), and regularized discriminant analysis (RDA). The methods differ in the technique that is used to calculate the object-class distances, i.e., under what assumption the class covariance matrices are calculated. The Mahalanobis distance is a measure of the distance of a sample from the mean of a set of standards, represented by the following equation:

$$D(g, x) = (X-c_g)'S^{-1}(X-c_g) \qquad \text{Eq. 2.5}$$

where S is an estimate of the common covariance matrix, $c_g$ is an estimate of the centroid for class g, and X is an object.

### 2.6.3 Principal component analysis

Principal component analysis (PCA) is an extremely useful method for data compression and information extraction. PCA finds combinations of variables, or factors, that describe major trends in the data. Mathematically, PCA relies upon an eigenvector decomposition of the covariance or correlation matrix of the process variables. For a given data matrix X with m rows and n columns, with each variable being a column and

each sample a row, the covariance matrix of X is defined as

$$\mathrm{cov}(X) = \frac{X^T X}{m-1} \qquad \text{Eq. 2.6}$$

provided that the columns of X have been "mean centered," i.e. adjusted to have a zero mean by subtracting the original mean of each column. If the columns of X have been autoscaled, i.e. adjusted to zero mean and unit variance by dividing each column by its standard deviation, the equation above gives the correlation matrix of X. (Unless otherwise noted, it is assumed that data is either mean centered or autoscaled prior to analysis.) PCA decomposes the data matrix X as the sum of the outer product of vectors $t_i$ and $p_i$ plus a residual matrix E:

$$X = t_1 P^T_1 + t_2 P^T_2 + \ldots + t_k P^T_k + E \qquad \text{Eq. 2.7}$$

Here k must be less than or equal to the smaller dimension of X, *i.e.* k<=5 min(m,n). The $t_i$ vectors are known as score and contain information on how the samples relate to each other. The $p_i$ vectors eigenvectors of the covariance matrix, *i.e.* for each $p_i$

$$\mathrm{cov}(X)p_i = \lambda_i P_i \qquad \text{Eq. 2.8}$$

where $\lambda_i$ is the eigenvalue associated with the eigenvector $p_i$. In PCA the $p_i$ are known as loadings and contain information on how the variables relate to each other. The $t_i$ form an orthogonal set ($t_i^T t_j = 0$ for $i \neq j$), while the $p_i$ are orthonormal ($p_i^T p_i = 0$ for $i \neq i$, $P_i^T p_j = I$ for $i = j$ ). Note that for X and any $t_i$ , $p_i$ pair

$$XP_i = t_i \qquad \text{Eq. 2.9}$$

i.e. the score vector $t_i$ is the linear combination of the original X data defined by $p_i$. (Another way to look at this is that the $t_i$ are the projections of X onto the $p_i$.) The $t_i$ , $p_i$ pairs are arranged in descending order according to the associated $\lambda_i$ The $\lambda_i$ are a measure

of the amount of variance described by the $t_i$ , $p_i$ pair. In this context, we can think of variance as information. Because the $t_i$, $p_i$ pairs are in descending order of $\lambda_i$, the first pair captures the largest amount of information of any pair in the decomposition. In fact, it can be shown that the $t_i$, $p_i$ pair captures the greatest amount of variation in the data that it is possible to capture with a linear factor, and each subsequent pair captures the greatest possible amount of variance remaining after subtracting $t_i p^T_i$ from X.

Generally, it is found (and it is usually the objective) that the data can be adequately described using far fewer factors than original variables. Thus, the data overload often experienced can be solved by observing fewer scores (weighted sums of the original variables) than original variables, with no significant loss of information. It is also often found that PCA turns up combinations of variables that are useful descriptions, or even predictors, of particular events or phenomena. These combinations of variables are often more robust indicators of laboratory sample or process conditions than individual variables due to the signal averaging aspects of PCA.

2.6.4 Artificial neural network

Inspired by the structure of the brain, a neural network consists of a set of highly

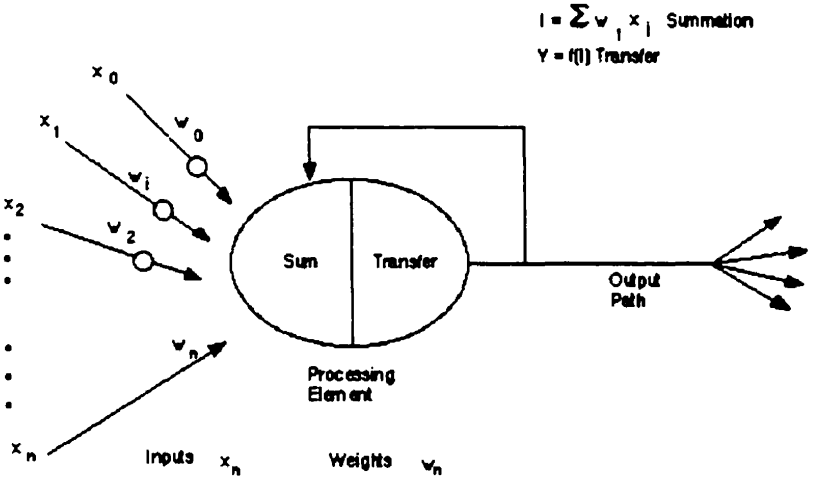


Figure 2.7 A typical single unit in an Artificial Neural Network (ANN).

interconnected entities, called *nodes* or *units*. Each unit is designed to mimic its biological counterpart, the neuron. Each accepts a weighted set of inputs and responds with an output. Figure 2.7 presents a picture of one unit in a neural network.

Let $X = (x_1, x_2, ..., x_n)$, where the $x_i$ $(1 \leq i \leq n)$ are real numbers, represent the set of inputs presented to the unit U. Each input has an associated weight that represents the strength of that particular connection. Let $W = (w_1, w_2, ..., w_n)$, with $w_i$ $(1 \leq i \leq n)$ real, represent the weight vector corresponding to the input vector X. Applied to U, these weighted inputs produce a net sum at U given by

$$S = SUM(w_i * x_i) = W.V. \qquad \text{Eq.2.10}$$

Learning rules will allow the weights to be modified dynamically. The state of a unit U is represented by a numerical value A, the *activation value* of U. An activation function *f* determines the new activation value of a unit from the net sum to the unit and the current activation value. In the simplest case, *f* is a function of only the net sum, so $A = f(S)$. The following are some other transfer functions that are often used. Figure 2.8 shows the plot of function *f*=1/(1+Exp[-sum], it can be found that when sum=0, f(sum)=0.5:

logistic -- f(x)=1/(1+exp(-x)) Eq. 2.11

linear -- f(x)=x Eq. 2.12

tanh -- f(x)=tanh(x) Eq. 2.13

tanh15 -- tanh(1.5x) Eq. 2.14

sine -- sin(x) Eq. 2.15

symmetric_logistic -- 2/(1+exp(-x))-1 Eq. 2.16

Gaussian -- exp(-x^2) Eq. 2.17

Gaussian-complement -- 1 - exp(-x^2) Eq. 2.18

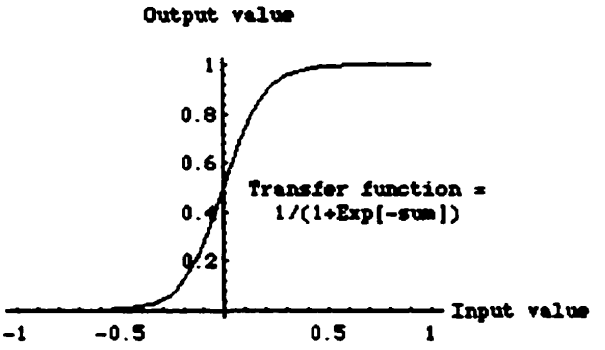


Figure 2.8 The plot of a typical transfer function (*f*=1/(1+Exp[-sum] ) used in Artificial Neural Network

A neural network is composed of such units and weighted unidirectional connections between them. In some neural nets, the number of units may be in the thousands. The

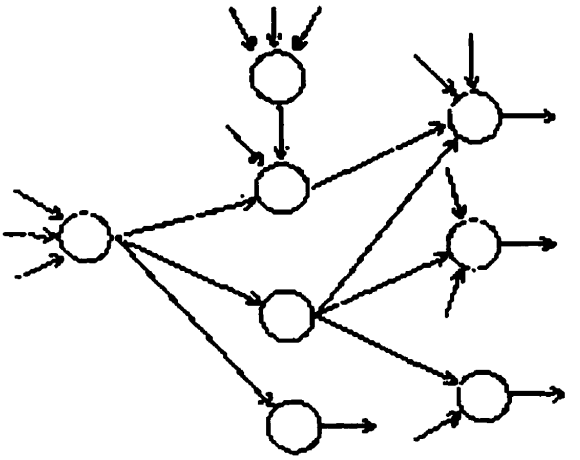

Figure 2.9 An example of an artificial neural network architecture

output of one unit typically becomes an input for another. There may also be units with external inputs and/or outputs. Figure 2.9 shows one example of a possible neural network structure.

Once a network has been structured for a particular application, that network is ready to be trained. To start this process the initial weights are chosen randomly. Then, the training, or learning, begins. There are two approaches to training - supervised and unsupervised. Supervised training involves a mechanism of providing the network with the desired output either by manually "grading" the network's performance or by providing the desired outputs with the inputs. Unsupervised training is where the network has to make sense of the inputs without outside help.

In supervised training, both the inputs and the outputs are provided. The network then processes the inputs and compares its resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the weights which control the network. This process occurs over and over as the weights are continually tweaked. The set of data which enables the training is called the "training set." During the training of a network, the same set of data is processed many times as the connection weights are ever refined.

Another important part is the rules of training. There are many algorithms used to implement the adaptive feedback required to adjust the weights during training. The most common technique is backward-error propagation, more commonly known as back-propagation.

When finally the system has been correctly trained, and no further learning is needed, the weights can be "frozen." This trained system is then tested with unknown sample data.

## 2.7 Classification of microorganism by FTIR spectroscopy

The differences in the biochemical composition of microorganism account for their diversity. Because infrared spectroscopy provides detailed information on biochemical composition, it can potentially serve as a valuable tool for the classification of microorganisms. The study of microorganism classification by infrared spectrophotometry arose almost half a century ago (Thomas and Greenstreet, 1954). The

constraints associated with the use of dispersive instruments and the unavailability of computers caused interest in this approach to vanish by the mid-1960s. However, the subsequent development of FTIR spectroscopy and of powerful classification algorithms that can be performed on personal computers resulted in renewed interest in this research.

The advantages associated with FTIR spectroscopy have allowed detailed studies of the potential of infrared spectroscopy as a means of microorganism classification. Helm et al. (1988) discriminated enteropathogenic *Escherichia coli* isolates by applying FTIR spectroscopy. This particular grouping was achieved by using IR bands in the region between 900-1200 $cm^{-1}$ where the O-specific side chains of lipopolysaccharides are the predominant spectral features. Hedrick et al. (1991) used diffuse reflectance spectroscopy of lipid extracts to distinguish between eubacteria and archaebacteria, the two main groups of bacteria. Within eubacteria, differentiation between gram-positive and gram-negative strains was performed on the basis of whole-cell spectroscopy (Naumann et al., 1988, 1991a). This differentiation is based on the fact that gram-negative bacteria have an outer membrane, which leads to distinct spectral differences between gram-negative and gram-positive bacteria in the spectral region between 2800 and 3000 $cm^{-1}$ (fatty acid region) and less significant differences between 1600 and 1700 $cm^{-1}$ (protein region). FTIR spectroscopic classification of bacteria agreed well with conventional grouping schemes and gave some valuable complementary results. Good classifications were obtained for different genera (e.g., Staphylococcus, Clostridium, Streptococcus, and Legionella) (Helm et al., 1991). Classification studies on oral streptococci also produced good results (Van der Mei et al., 1993).

There are many exciting developments in the mathematical discrimination techniques employed for the classification of microorganisms. Lipkus and colleagues (1990) investigated the reproducibility of the infrared spectra of microorganisms and its implications for microorganism identification. They reported that in an attempt to use the spectral information in the region 1190-980 $cm^{-1}$ to build an identification system based on a spectral library search, successful identification was obtained for cells grown in one batch. To obtain a quantitative basis for identification, classification, or differentiation, Naumann (1991b) suggested the use of the spectral distance or "D" value as an index. The spectral distance (D) can be considered as a measure of the difference between two IR spectra. The D value is defined by the equation: $D=(1-\alpha)*1000$, where $\alpha$ is Pearson's correlation coefficient. In order to obtain a classification that can be correlated with conventional taxonomy, Helm et al. (1991) resorted to systematically varying the spectral treatment parameters and selecting spectral windows prior to performing cluster analysis with the measurements of correlation calculated between those treated spectra. The results of classification of bacteria from their FTIR spectra showed that even when grown on different media, a strain of bacteria could be classified in one cluster with a 96.8% similarity level. Van der Mei and colleages (1993) classified 40 *Streptococcal* species by cluster analysis employing the first derivative of the infrared spectra and selected regions. Holt et al. (1995) were the first to use PCA to study microorganism classification; since their work, the number of publications in this field has multiplied dramatically and they are mostly focused on the development of mathematical techniques for the treatment of the spectral information. Timmins et al. (1998b) applied PCA and discriminant function

analysis to differentiate 22 brewing yeast strains. Goodacre et al. (1996b) and Timmins et al. (1998a) reported that artificial neural networks have good discriminating capabilities. Schmitt et al. (1998) evaluated six different neural network architectures with respect to their capability to build spectral libraries for different bacteria and yeasts. After developing these libraries, the networks were connected to a large library. These "multilayered neural networks" allowed for an optimal differentiation based on specific strains. Alsberg et al. (1998) studied Eubacterium species by FTIR spectroscopy. To identify important wavenumber regions for the classification of the bacterial isolates, they investigated three rule induction methods and various spectral preprocessing regimes. They found that the FuRES (fuzzy multivariate rule-building expert system) method was superior in terms of prediction, whereas the rules proposed by the univariate CART method (Classification and Regression Trees) were easier to interpret in terms of which wavenumbers in the IR spectra were important for bacterial class separation. Scaling and normalization of FTIR spectra as preprocessing steps were necessary to obtain optimal classification models. McNaughton et al., (1999) applied the multivariate statistical techniques of PCA, soft independent modeling by class analogy (SIMCA), K-Nearest Neighbors (KNN), and artificial neural networks (ANN) to IR spectra of several cyanobacterial species and successfully classified the bacteria. Employing the first-derivative IR spectra of bacteria as input resulted in reduction of baseline variability and minimized intra-class variation.

Sockalingum et al. (1998) used the ATR sample-handling technique to obtain FTIR spectra of bacteria and demonstrated that ATR/FTIR spectroscopy can

discriminate and classify bacterial strains. The combination of infrared spectrometers and optical microscopes is probably the most significant advance in the field of microorganism classification. Naumann et al. (1998b) studied the use of FTIR microscopy to characterize microorganisms. Dubois (1999) collected the spectra of bacteria deposited on a polyethylene substrate and reported satisfactory results employing cluster analysis and ANN algorithm for bacteria differentiation.

The fundamental work carried out to date on the potential application of infrared spectroscopy has provided valuable information about the limitations and critical factors that must be considered prior to the final elaboration of an automated identification system based on FTIR spectroscopy. However, widespread application of FTIR spectroscopy for the characterization of microorganisms will likely occur only if a reliable, stable and automated method is available. The work described in this thesis will focus on the potential utility of an automated sampling system in combination with controlled growth condition and the use of numerical analysis for the classification of yeast strains based on their FTIR spectra. The potential use of FTIR spectroscopy to classify yeast in terms of their function and sensitivity will also be undertaken.

# Chapter 3

# Classification and Identification of Yeasts by Combined Use of Infrared Spectroscopy and Chemometric Techniques

## 3.1 Introduction

Yeasts are heterotrophic, lack chlorophyll, and have a wide array of natural habitats. They have not only provided us with fermented food products such as wine, bread, and yogurt but are also responsible for food spoilage, and some species are of health concern. Therefore, yeast identification is of practical importance. To fulfill this task, many different methods have been developed. Conventional yeast differentiation systems use morphological characteristics as well as patterns of assimilation and fermentation of carbon sources. These methods are tedious and time-consuming, and their capacity is limited since many species are distinguished from one another by a single physiological reaction controlled by only one mutable marker. New techniques such as fatty acid analysis, electrophoretic karyotyping, restriction fragment length polymorphism, DNA fingerprinting, restriction enzyme analysis of PCR-amplified rDNA, randomly amplified polymorphic DNA, and nucleic acid hybridization with oligonucleotide probes have also been used for this purpose (Olson, 1995). While some of these techniques do provide satisfactory results, molecular methods in general are still

difficult to perform on a routine basis in laboratories of the food industry.

For routine purposes, the ideal method for yeast characterization would require minimal sample preparation, would analyze samples directly (i.e. would not require reagents), and would be rapid, automated and (at least relatively) inexpensive. With recent developments in analytical instrumentation, these requirements are being fulfilled by spectroscopic methods. One of the most commonly investigated methods is Fourier transform infrared (FTIR) spectroscopy (Helm et al., 1991; Naumann et al., 1991a, b).

FTIR spectroscopy measures dominantly vibrations of functional groups and highly polar bonds. Thus, IR 'fingerprints' are made up of the vibrational features of all the chemical compounds in the sample. For microbial samples, these will include DNA/RNA, proteins, and membrane and cell-wall components. The interpretation of the spectra of microorganisms has conventionally been done by the application of 'unsupervised' pattern recognition methods such as hierarchical cluster analysis (HCA). With 'unsupervised learning' the algorithms seek 'clusters' among the spectral data, which allows the investigator to group objects on the basis of their perceived similarity. More recently, more powerful supervised methods have been employed to analyze the spectral data (Goodacre et al., 1996 a, b).

Within microbiology, FTIR spectroscopy has been shown to allow the chemically-based discrimination of intact microbial cells, without their destruction, and produces complex biochemical fingerprints which are reproducible and distinct for

different bacteria. In particular Helm et al., (1991) and Naumann et al., (1991a,b) have shown that FTIR spectroscopy (in the mid-IR range of 4000-400 $cm^{-1}$) provides a powerful tool with sufficient resolving power to distinguish microbes at the strain level.

The aims of this study are to differentiate 56 yeast strains based on their FTIR spectra and investigate the differentiation performance of different chemometric techniques.

### 3.2 Materials and Methods

Fifty-six yeast strains representing 20 species of 7 genera were obtained from Lallemand Inc. (Montreal, Canada). All the strain codes and related information can be found in Appendix 1. All strains were stored at -45°C.

#### 3.2.1 Growth Conditions and Sample Preparation

To recover possible injured cells, strains were thawed on Universal Growth Medium (Quelab Laboratories, Montreal) and incubated at 37±2°C for 24±1 hours twice to ensure the acquisition of pure cultures. Subsequently, a sample of the culture was taken with a platinum loop (3-mm-diameter platinum loop) and reinoculated on Universal Growth Medium (Quelab Laboratories, Montreal). This time the Universal medium was used to standardize the contribution of the growth medium to the yeast infrared spectra. After incubation at 37±2°C for 24±1 hours, a sample of a confluent colony was carefully taken with a calibrated sterile platinum loop (3mm-diameter platinum loop) in the third quadrant of the growth media surface and deposited into 100 µl of distilled water. The yeast suspension was centrifuged at 6000 RPM for 2 minutes, the supernatant was

decanted and the pelleted yeast was resuspended in 100 µl of distilled water. This process was repeated twice to remove any remaining metabolic by-products. A 30-µl aliquot of the resuspended yeast cell was deposited onto a ZnSe window of an autosampler wheel containing eight ZnSe windows, and the wheel was dried at 37±2°C for 2 hours to yield transparent yeast films suitable for transmission FTIR measurements. The films were kept in constant humidity prior to recording of the infrared spectra of the yeast films.

### 3.2.2 Spectral Acquisition

All the infrared spectra were recorded between 4000 and 400 $cm^{-1}$ employing a Michelson FTIR spectrometer (ABB Bomem, Inc.). The spectrometer was purged with pure $N_2$. Spectra were acquired by coadding 32 scans at a resolution of 4 $cm^{-1}$. The sample wheel was controlled by Bacteria ID1.0 software (running under Win98, obtained from Quelab Inc., Montreal, Canada). All samples were run in triplicate. A typical yeast FTIR spectrum is shown in Figure 3.1.

### 3.2.3 Preprocessing

To minimize problems arising from baseline shifts and differences in film thickness and to enhance the resolution of superimposed bands, the following procedures were performed: 1) all spectra were baseline corrected from 4000 to 400 $cm^{-1}$; 2) all spectra were then normalized so that the highest absorbance was set to 1; 3) second

derivatives of all the spectra were calculated; 4) the second derivative spectra were smoothed with a 9-point smoothing function (Savitzky and Golay, 1964).

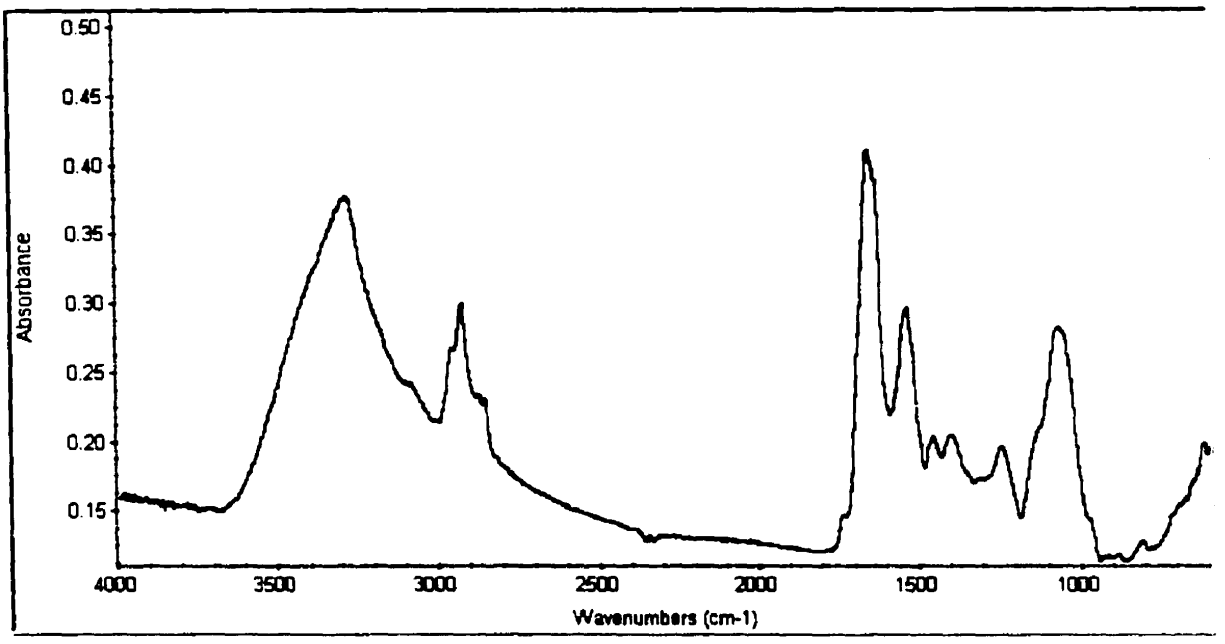


Figure 3.1 A typical FTIR spectrum of a *Saccharomyces italicus* strain (strain 6074)

### 3.3 Spectral Analysis by Chemometrics

The possibility of identifying yeasts based on their FTIR spectra was evaluated using cluster analysis, principal component analysis (PCA), discriminant analysis and artificial neural network methods. The hierarchical clustering employed centroid linkage and measurement of the squared Eulidean distance between the points to provide an unsupervised grouping, while the discriminant analysis employed Mahalanobis distance for supervised classification. PCA was performed according to the nonlinear iterative partial least squares (NIPALS) algorithm (Wold, 1966). All these three algorithms were part of the SCAN software (Minitab Inc,. State College, PA).

The artificial neural network (ANN) analysis was carried out by NeuroShell software (Ward Systems Inc., Frederick, MD). The ANN employed consists of three layers. The first layer has two options: 1) the entire spectra and 2) the first 10 PC values. The second layer is the hidden layer and the last layer is the binary-coded output layer (for yeast classification, the output layer was encoded as follows: *Saccharomyces cerevisiae* was coded as 1000, *Saccharomyces chevalieri* was coded as 0100, *Saccharomyces capensis* was coded as 0010, *Saccharomyces italicus* was coded as 0001; for classification of yeasts according to the type of fermentation process in which they are employed, the output layer was encoded as follows: wine was coded as 100, beer was coded as 010, bread was coded as 001; for yeast classification in terms of their sensitivity killer yeast strains, the output layer was encoded as follows: sensitive was coded as 100, possess was coded as 010, neutral was coded as 001.

The optimization of the ANN employed the following procedures: 1) Standardization of the input variables: standardizing the inputs can make training faster and reduce the chances of getting stuck in local optima. Standardizing inputs removes the problem of scale dependence on the initial weights. In particular, scaling the inputs to [-1,1] will work better than [0,1], although any scaling that sets to zero the mean or median or other measure of central tendency is likely to be as good (Iglewicz, 1983); 2) in standard backpropagation, too low a learning rate makes the network learn very slowly. Too high a learning rate makes the weights and objective function diverge, so there is no learning at all. Trying to train an ANN using a constant learning rate is usually a tedious process requiring too much trial and error. Here, batch training was selected since it does not require a constant learning rate. (Fahlman 1989; Riedmiller and Braun 1993); 3) Activation functions for the hidden units are needed to introduce nonlinearity into the network. Neural networks can be made more powerful by adding the hidden units than just plain perceptions (which do not have any hidden units, just input and output units). Functions such as *tanh* that produce both positive and negative values tend to yield faster training than functions that produce only positive values such as logistic, because of better numerical conditioning (Jordan, 1995). The *tanh* function was chosen to be the activation function for hidden units. For the output units, the binary (0/1) outputs were selected (Jordan, 1995).

The network was presented with input and corresponding outputs and was trained by adjusting the connections between input, hidden and output layers; training was stopped after 40 generations without improvement greater than 0.5% in the external test

set, which was randomly extracted from the input matrix. After training, the relationship of all the yeast spectra was encoded in the weights.

Library search routines were employed to assess spectral reproducibility. Library search routines compare the unknown sample spectrum with each reference spectrum in the selected libraries and find the spectra that most closely match the unknown. Most library search algorithms involve a point-by-point evaluation with an overall closeness of match being determined by some form of similarity metric. The match value is between 0 and 100 and indicates how well the library spectrum matches the unknown. A match value of 100 indicates a perfect match. The closer the value is to 100, the better is the match. To evaluate the spectral reproducibility, one of the three preprocessed spectra for each yeast strain was stored in a spectral library as a standard spectrum, and the other two were compared to it by the application of spectral library search algorithms.

## 3.4 Results and Discussions

### 3.4.1 Spectral Reproducibility

A major issue in the differentiation or identification of yeasts by FTIR spectroscopy is the spectral reproducibility; that is, differentiation and identification can only be achieved if reproducible spectra can be recorded for each yeast strain. FTIR spectra of yeasts are influenced by many factors, such as the composition of the growth medium, growth temperature, incubation time, the washing method and drying method. For a high level of reproducibility it was necessary to develop a standardized sample preparation procedure as described above.

Since different kinds of yeast cells have relatively similar biochemical composition, it is obvious that all the yeast spectra showed fairly similar patterns (Figure 3.2). Thus, it is very important that the spectral variability introduced by growth and sample preparation conditions be minimized to allow the subtle inherent differences between the spectra of different yeast strains to be detected. In this study a single growth medium was employed to reduce the sources of spectral variability, and the temperature and incubation time were kept constant for all strains. The sample preparation protocol was also standardized.

The easiest way to check spectral reproducibility is to overlay replicate spectra and check if they are completely superimposed (Figure 3.3). Kenner et al. (1958) reported that changes in the region 1200-830 $cm^{-1}$ region (related to the polysaccharide content in the microorganisms) were correlated to temperature variations. Naumann (1991b) reported that the bands at 2960, 2922, 2873 and 2852 $cm^{-1}$ (corresponding to the symmetric and asymmetric vibrations of methyl and methylene groups, in the membrane phospholipids) exhibited variations with temperature owing to phase transitions. The replicate spectra in Figure 3.3 exhibit some band shifts in these regions; however, because they are less than 0.1 $cm^{-1}$, the differences due to temperature variations will not significantly affect the discriminate ability of IR spectroscopy.

From the statistical point of view, the reproducibility of the FTIR spectra can be evaluated by calculating their spectral average, standard deviation and range. The spectral

software Omnic (Nicolet Inc., Madison, WI) was employed to calculate the arithmetic mean of the absorbance values at each wavenumber, the standard deviation of the absorbance values for each data point; and the range of absorbance for each data point (the lowest absorbance value for a data point is subtracted from the highest absorbance value for that point). Figures 3.4, 3.5, and 3.6 show the results of these calculations for strain 6071 (a strain from *Saccharomyces italicus*); in the region of 1800-800 $cm^{-1}$ the variance is <0.004 and the range is < 0.008.

A library search algorithm was employed to investigate the reproducibility of yeast spectra. Table 3.1 lists the average percent similarity for all the spectra recorded from *Sacharomyces cerevisiae* employed. In the region of 1800-800 $cm^{-1}$, the average percent similarity is >95.

The Pearson correlation coefficient was also employed to evaluate the spectral reproducibility. The value of the coefficient (Eq.3.1) typically ranges from –1, indicating a perfect negative correlation, to +1, indicating a perfect positive correlation with a coefficient of zero indicating absence of correlation between the variables.

$$r_{ij} = [\sum (x_{ki} - \overline{x_i})(x_{kj} - \overline{x_j})]/(\sqrt{\sum (x_{ki} - \overline{x_i})^2} \sqrt{\sum (x_{kj} - \overline{x_j})^2}$$

Eq. 3.1

In equation 3.1, $x_{ki}$ and $x_{kj}$ are variables, and $x_i$ bar and $x_j$ bar are the means for variable $x_{ki}$ and $x_{kj}$ respectively. The correlation coefficients are calculated pairwise to evaluate the similarities between two individual spectra. In the region of 1800-800 $cm^{-1}$, the average correlation coefficient values for each strain are >0.92.

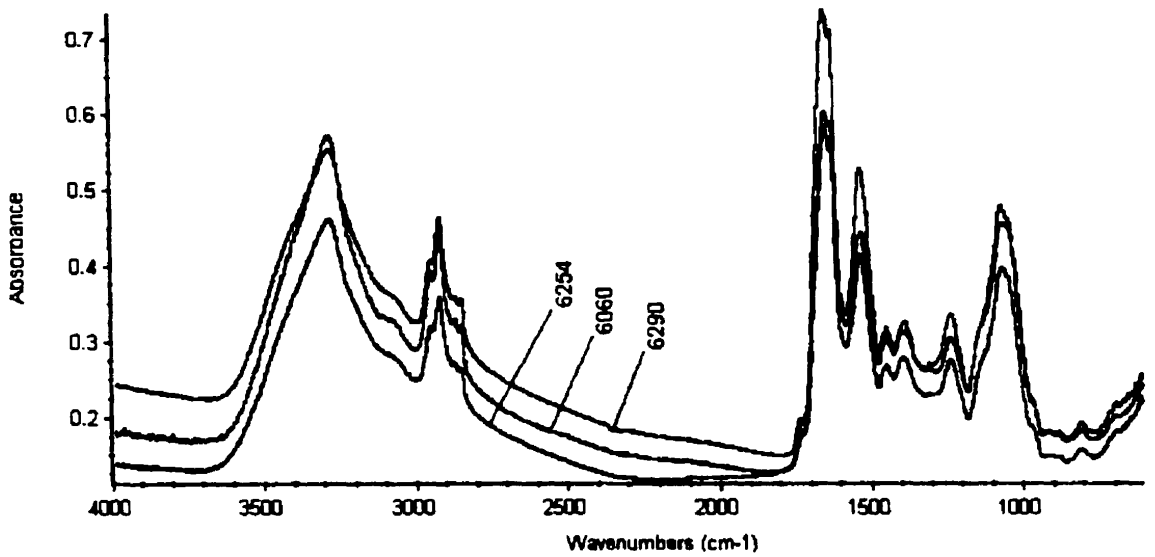


Figure 3.2 Overlaid FTIR spectra from *Saccharomyces chevalieri* (strain 6254), *Saccharomyces cerevisiae* (strain 6060) and *Saccharomyces capensis* (strain 6290) yeast species

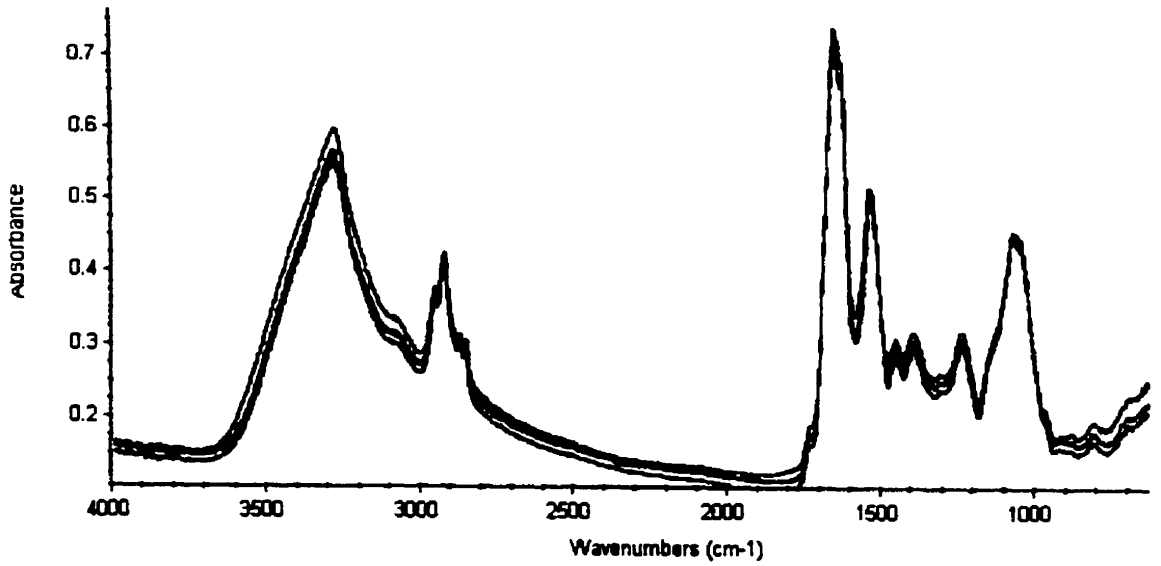


Figure 3.3 Overlaid FTIR spectra from three different batches of a *Saccharomyces cerevisiae* strain (strain 6060)

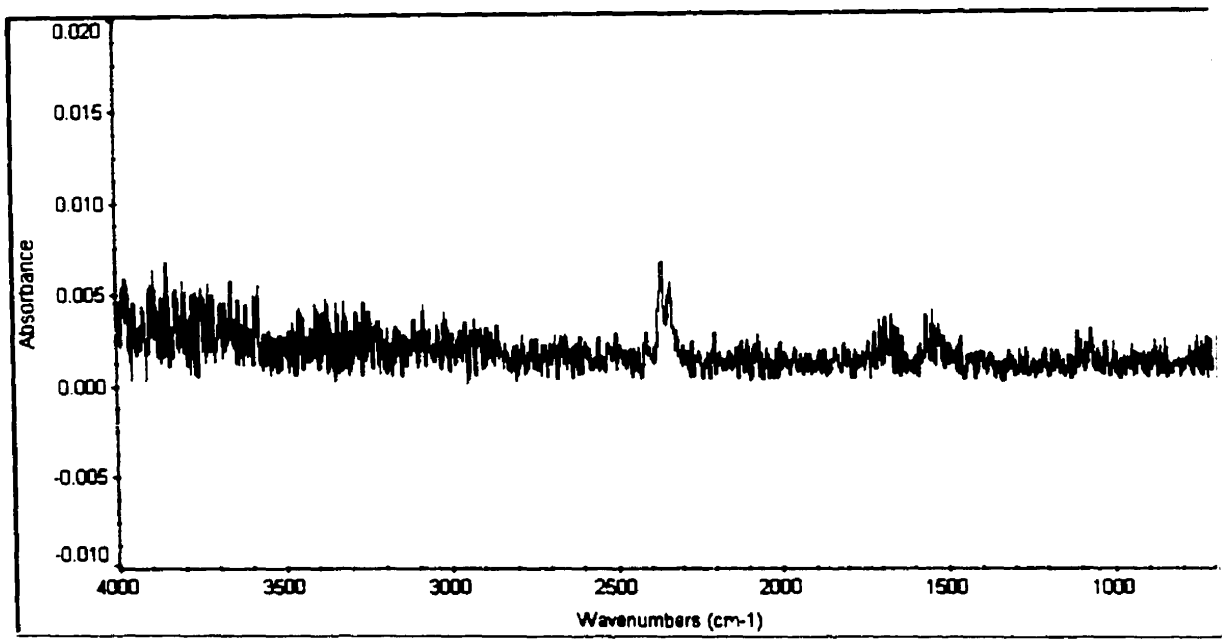


Figure 3.4 Absorbance variability in the FTIR spectra of strain 6071 recorded from three different batches

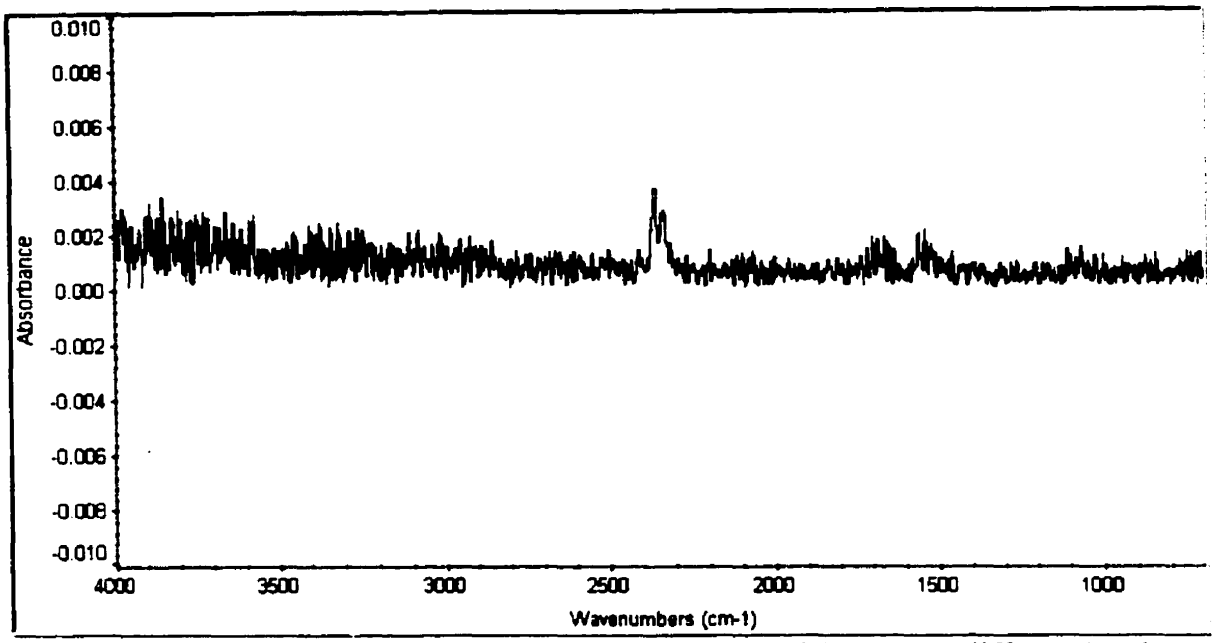


Figure 3.5 The variance of spectra of strain 6071 recorded from three different batches

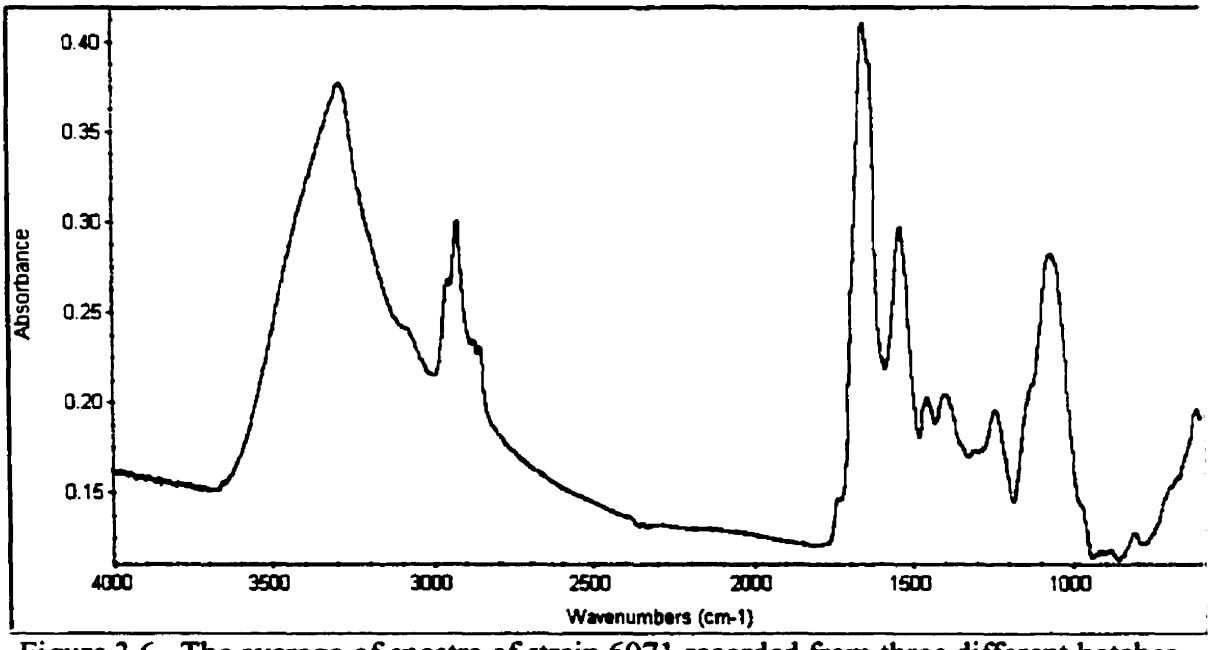


Figure 3.6 The average of spectra of strain 6071 recorded from three different batches

Table 3.1 The average percent similarity between the infrared spectrum of a yeast strain from *Saccharomyces cerevisiae* compared to the infrared spectra of the same strain in a spectral database recorded from different batches (spectral region: 1800-800 $cm^{-1}$)

| Strain number | Average percent similarity |
|---|---|
| 6050 | 99.05 |
| 6467 | 99.03 |
| 6400 | 97.38 |
| 6287 | 95.98 |
| 6649 | 99.02 |
| 6648 | 98.45 |
| 6348 | 95.80 |
| 6032 | 96.13 |
| 6014 | 97.53 |
| 6061 | 95.47 |
| 6562 | 97.80 |
| 6058 | 95.24 |
| 6060 | 98.57 |
| 6163 | 95.98 |
| 6412 | 95.80 |
| 6422 | 97.10 |
| 6101 | 96.24 |
| 6059 | 98.69 |
| 6301 | 96.01 |
| 6100 | 97.63 |
| 6652 | 98.01 |

### 3.4.2 Identification of Yeast Strains in Terms of Their Taxonomic Characteristics by FTIR Spectroscopy

#### 3.4.2.1 Unsupervised Analysis

As exemplified by the typical yeast infrared spectra shown in Figure 3.2, all the infrared spectra of yeast strains showed complex and broad contours, and there was very little qualitative difference between them that can be discerned easily. Accordingly, it was appropriate to consider the use of multivariate analysis to extract the differences between the IR spectra of different yeast strains. Multivariate pattern recognition methods are divided into 'unsupervised' and 'supervised' categories. The former, such as

Hierarchical Clustering or Fuzzy Clustering, classify spectra based upon the degree of their overall similarity and require no training. The latter, such as the use of artificial neural networks, train the classifier based on the obtained class identities and then use it to predict the class identity of unknown samples (Helm et al., 1991; Naumann, 1998a, b; Goodacre, 1998c).

Among modern yeast identification techniques, DNA fingerprinting is one of the most important ones. It uses the unique profiles of the DNA of known yeast strains to identify the unknown yeast strains. Accordingly, identification of yeasts by using regions of their IR spectra in which absorptions due to DNA are observed was considered. The most useful IR region employed in the study of DNA is the region between 1080 and 1240 $cm^{-1}$, in which bands arising from the stretching vibrations of phosphodiester groups are observed. This region of the spectrum is very informative, as it is dominated by absorptions from both polysaccharides and triacylglycerols, in addition to the contribution of cellular DNA. The classification results for the 56 yeast strains from their infrared spectra in the region between 1080 and 1240 $cm^{-1}$ are shown in Figure 3.7, while the results of classification employing the broader region of 1800-800 $cm^{-1}$ are shown in Figure 3.8. Although some improvement was achieved by employing the DNA spectral region, the overall classification accuracy is not adequate. It is not surprising given the fact that there are so many chemical components in the yeast cell, some composition other than DNA or RNA may also contribute some absorption in the region of 1080-1240 $cm^{-1}$, which makes the DNA or RNA signal weak, and lead to the misclassification.

One way to solve this problem is to use weighting factors to amplify the contribution of a weak signal to be used in the classification. Selected regions combined with proper weighting factors may provide better classification results. These problems have been highlighted in a number of studies, which have attempted to classify microorganisms based on their infrared spectra (Helm et al., 1991; Naumann et al., 1991b; Kummerle et al., 1998). Three regions (3030-2830 $cm^{-1}$, 1350-1200 $cm^{-1}$, and 900-700 $cm^{-1}$; all weighting factors were 1) were selected to carry out the classification (Kummerle et al., 1998). In addition, three sets of spectral data were employed to evaluate the utility of derivatization as a means of resolving overlapping bands: 1) spectral data obtained after baseline correction and normalization of the selected spectral region without derivatization, 2) spectral data obtained after baseline correction, normalization, computation of the $1^{st}$ derivative and smoothing of the selected spectral region, 3) spectral data obtained after baseline correction, normalization, computation of the $2^{nd}$ derivative and smoothing of the selected spectral region.

A centroid linkage and squared Euclidean distance was employed in this unsupervised cluster analysis. In centroid linkage each cluster is represented by its centroid; the distance between two clusters is the distance between their centroids. This method does not distort the cluster space. Figures 3.9, 3.10 and 3.11 show dendrograms obtained with the three spectral preprocessing techniques described above. It can be observed that better results were obtained after derivatization of the spectral data and that the results from the $2^{nd}$ derivative spectral data were better than those from the $1^{st}$ derivative data. Because derivatization makes the absorption bands sharper, it increases

the discriminate ability of the clustering algorithm, dramatically increases the difference between spectra of different strains and gives more multidimensional space to find the subtle differences between the spectra of different strains. Derivatization can also partially reduce baseline variation. Overall, derivatization can improve the extraction of the classification information inherent in the spectra of yeast. Compared to the $1^{st}$ derivative, the $2^{nd}$ derivative has a more refined band contour and contains more bands, and that is why it yields more promising results (Figure 3.12).

Principal component analysis (PCA) is a well-known technique for reducing the dimensionality of multivariate data while preserving most of the variance. To reduce the number of variables and to detect relationships between variables, principal component analysis according to the NIPALS algorithm (Wold, 1966) was performed on the preprocessed data (spectral data obtained after baseline correction, normalization and computation of the $2^{nd}$ derivative in the 1800-800 $cm^{-1}$ region), Figure 3.13 shows a plot of eigenvalues against principal component number. It is clear that most of the variation lies in the first principal component. The first 10 principal components, which account for over 99% of the data variance, were selected and employed for cluster analysis. The results shown in Figure 3.14 clearly demonstrated that PCA alone cannot be used to cluster these yeast strains, because different strains from the same species do not fall into the same cluster.

Since collinear variables cannot be employed in discriminant analysis, discriminant analysis cannot be used to analyze the original spectra data directly. To

enhance the PCA performance and expand its discriminate ability, we combined the PCA, discriminant analysis and cluster analysis together. The classification procedure is as follows: first reduce the dimension of yeast IR spectra by PCA, then use discriminant analysis to distinguish groups on the basis of the retained principal components (PCs) and the priori knowledge of which spectra were replicates, and last, the square of the Euclidean distance between priori group centres can be used to construct a similarity measure and cluster analysis is then employed to construct the dendrogram. Figure 3.15 shows that while there was some improvement, the unsupervised learning methods employed here cannot be used to discriminate between the yeast strains employed in this study.

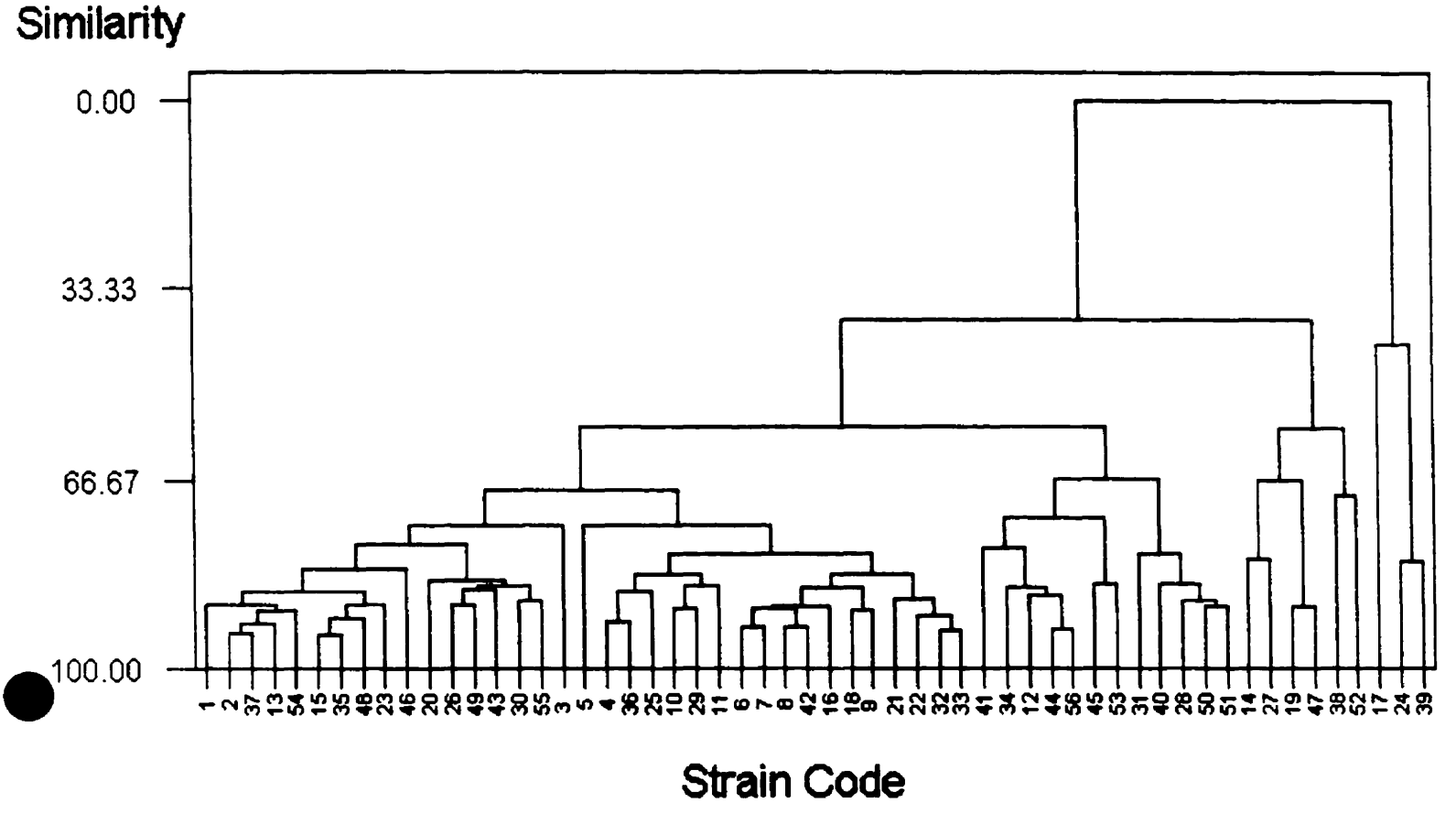


Figure 3.7 A plot of a dendrogram generated from the cluster analysis of 56 different yeast strains based on the changes in infrared spectral region between 1240-1080 $cm^{-1}$ after baseline correction and normalization of the FTIR spectra of the yeast strains (please refer to Appendix 1 for the strain identity)

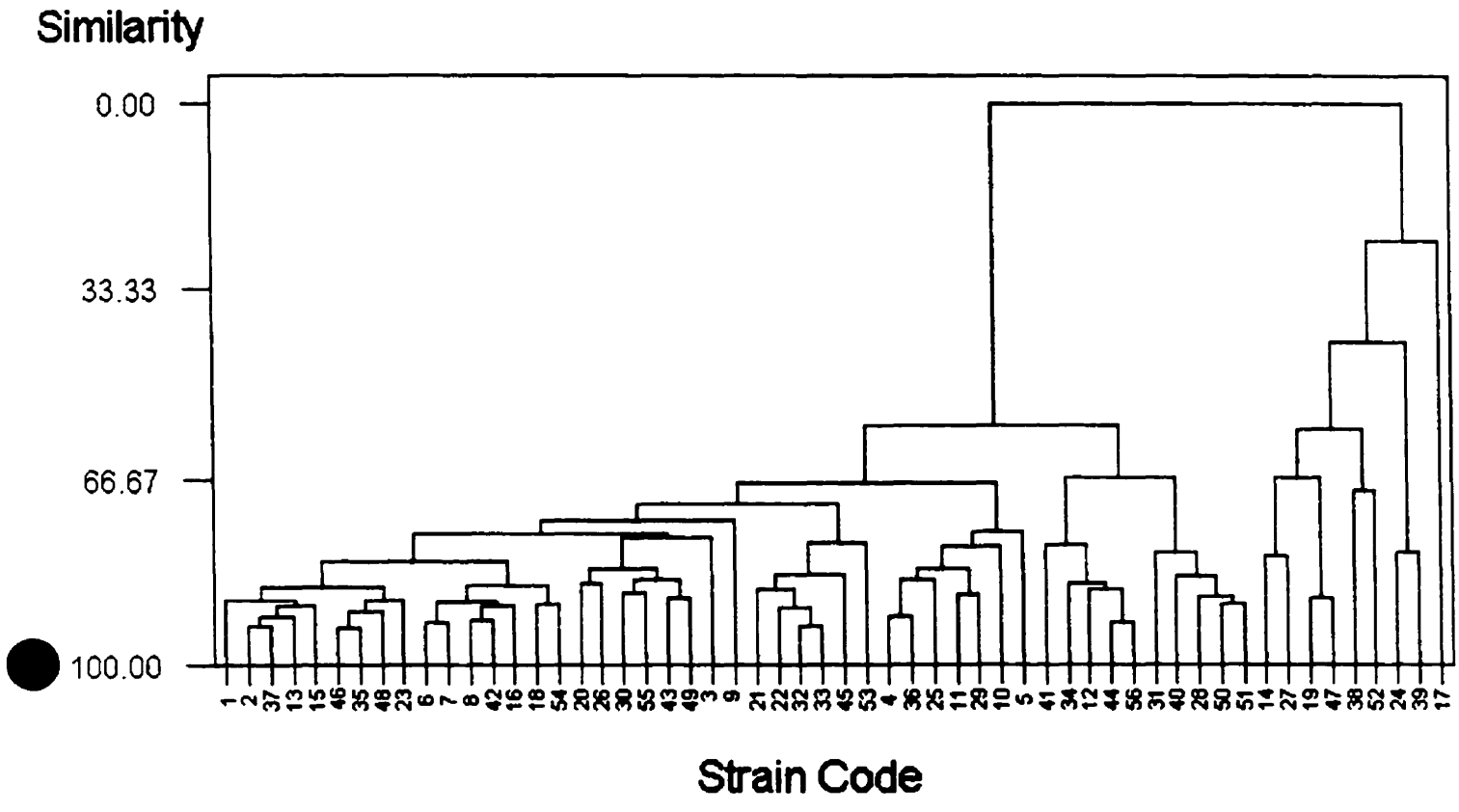


Figure 3.8 A plot of a dendrogram generated from the cluster analysis of 56 different yeast strains based on the changes in infrared spectral region between 1800-800 $cm^{-1}$ after baseline correction and normalization of the FTIR spectra of the yeast strains (please refer to Appendix 1 for the strain identity)

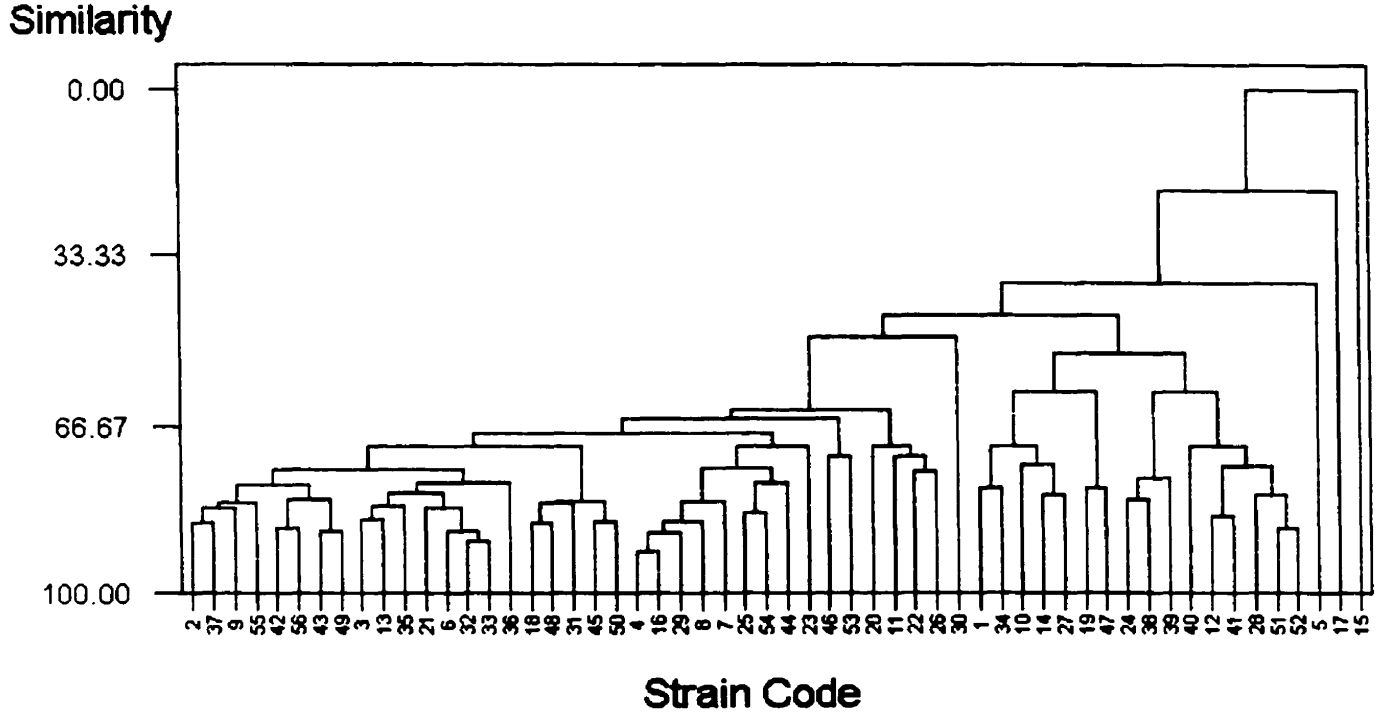


Figure 3.9 A plot of a dendrogram generated from the cluster analysis of 56 different yeast strains based on the changes in the infrared spectral region between 3030-2830 $cm^{-1}$, 1350-1200 $cm^{-1}$ and 900-700 $cm^{-1}$ (all weighting factor were 1) after baseline correction and normalization of the FTIR spectra of the yeast strains (please refer to Appendix 1 for the strain identity)

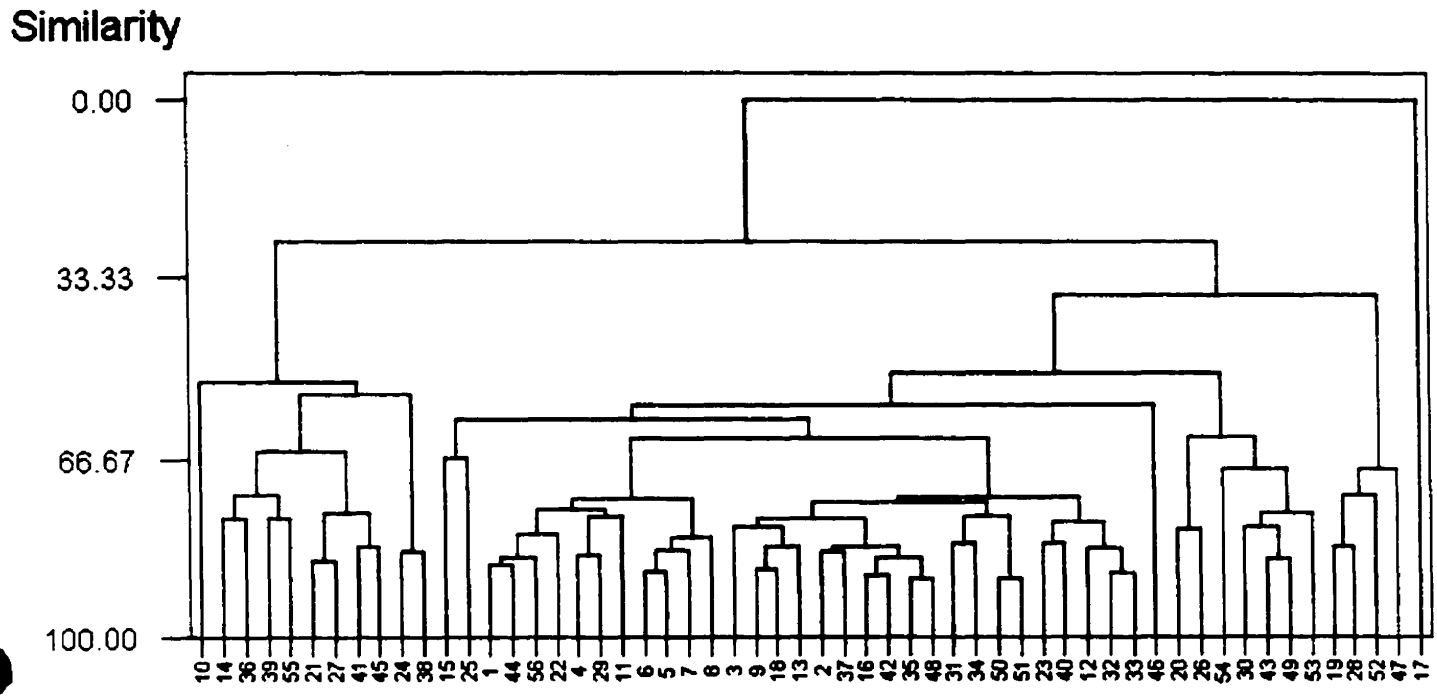


Figure 3.10 A plot of a dendrogram generated from the cluster analysis of 56 different yeast strains based on the changes in the infrared spectral region between 3030-2830 $cm^{-1}$, 1350-1200 $cm^{-1}$ and 900-700 $cm^{-1}$ (all weighting factor were 1) after baseline correction, normalization and computation of the first derivative data with 9 point smoothing of the FTIR spectra of the yeast strains
(please refer to Appendix 1 for the strain identity)

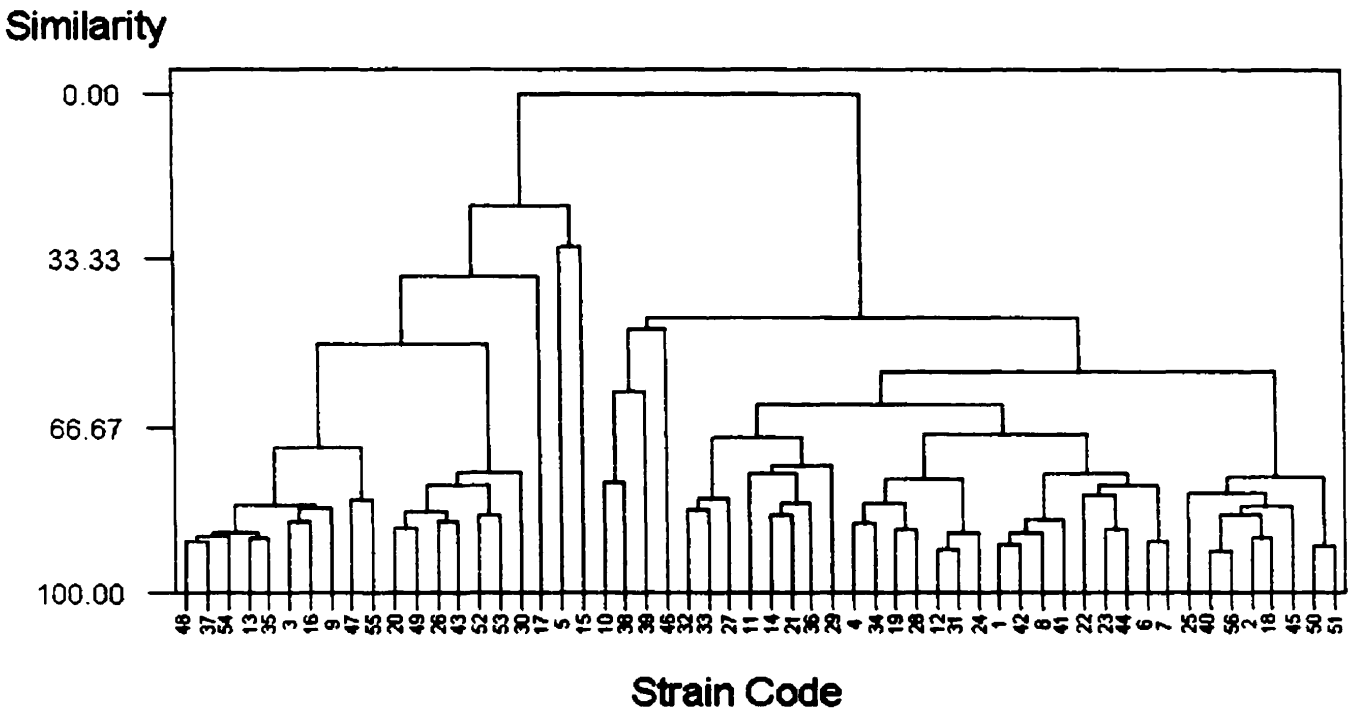


Figure 3.11 A plot of a dendrogram generated from the cluster analysis of 56 different yeast strains based on the changes in the infrared spectral region between 3030-2830 $cm^{-1}$, 1350-1200 $cm^{-1}$ and 900-700 $cm^{-1}$ (all weighting factor were 1) after baseline correction, normalization and computation of the second derivative data with 9 point smoothing of the FTIR spectra of the yeast strains (please refer to Appendix 1 for the strain identity)

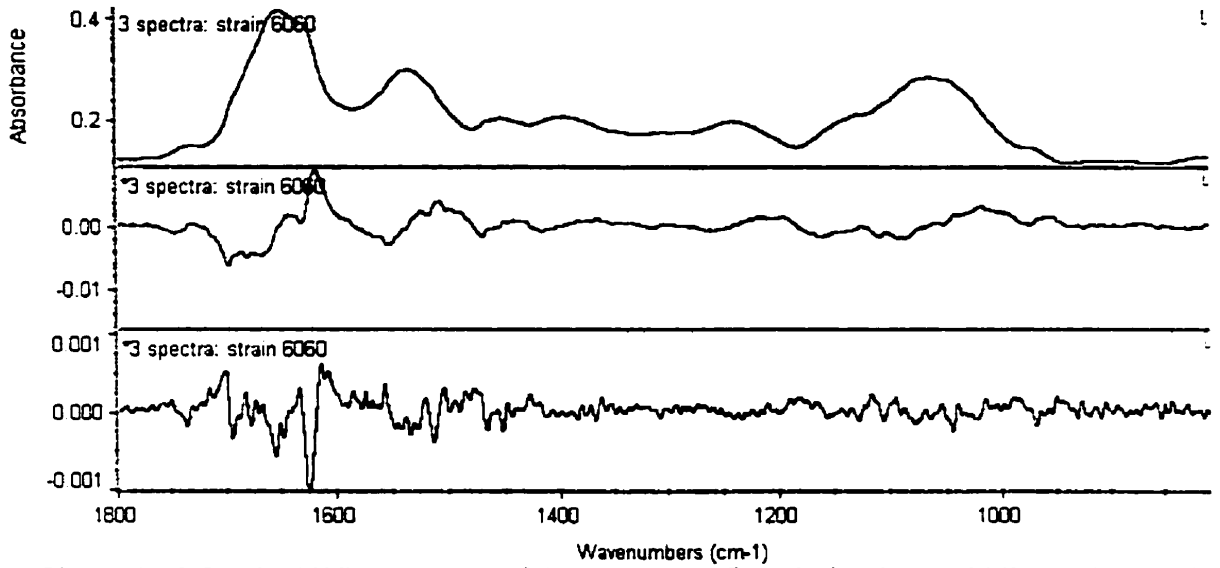


Figure 3.12 Stacked FTIR spectra of the raw (top), first-derivative (middle) and second-derivative (bottom) spectra of a *Saccharomyces cerevisiae* strain (strain 6060)

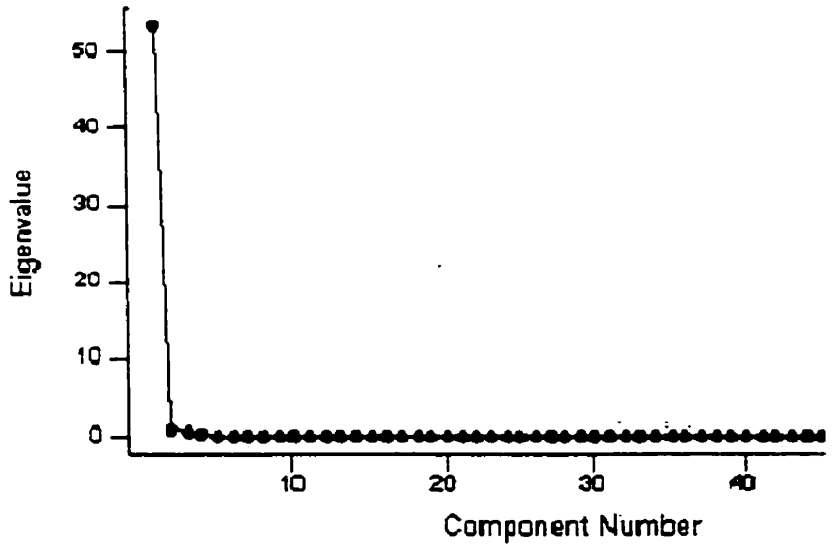


Figure 3.13 The plot of the eigenvalues against the principal component number

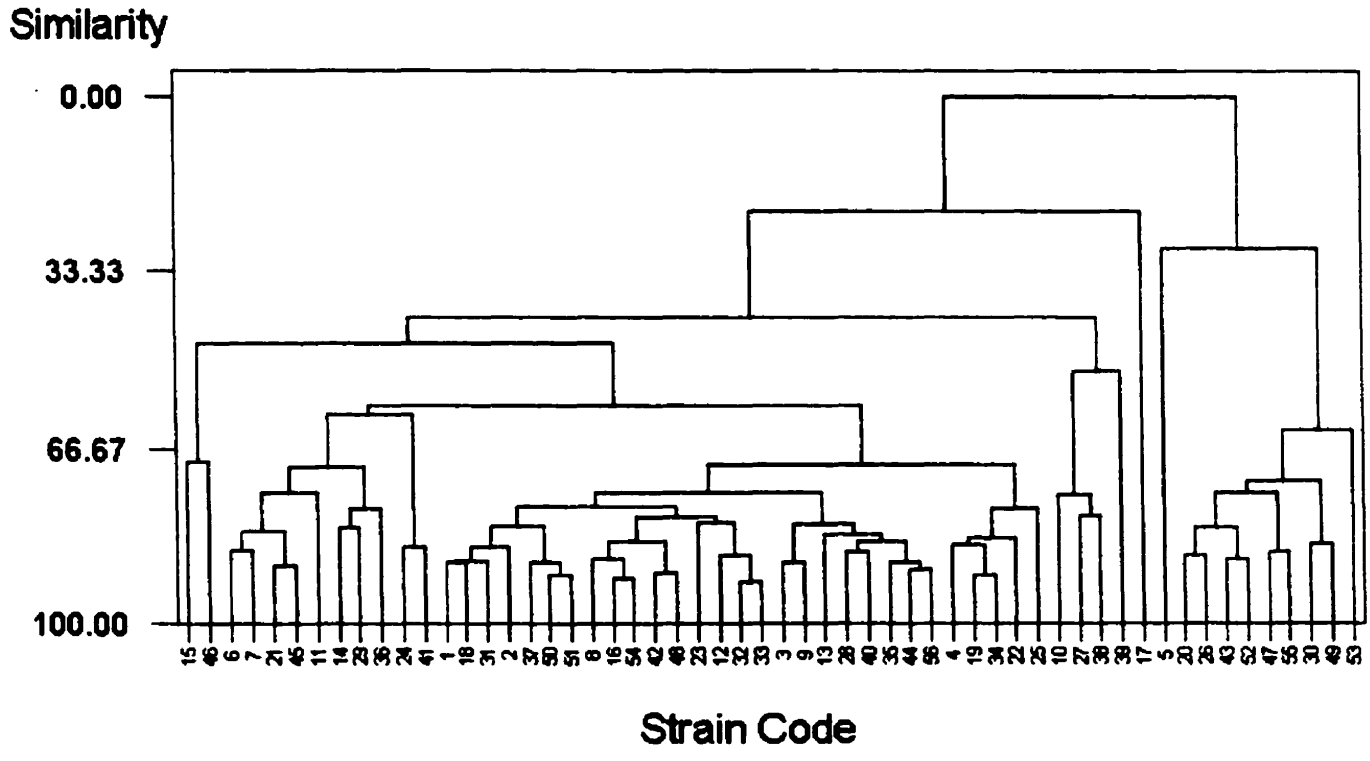


Figure 3.14 A plot of a dendrogram of 56 different yeast strains employing cluster analysis on the first 10 principal components values from the infrared spectral region between 1800-800 $cm^{-1}$
(please refer to Appendix 1 for the strain identity)

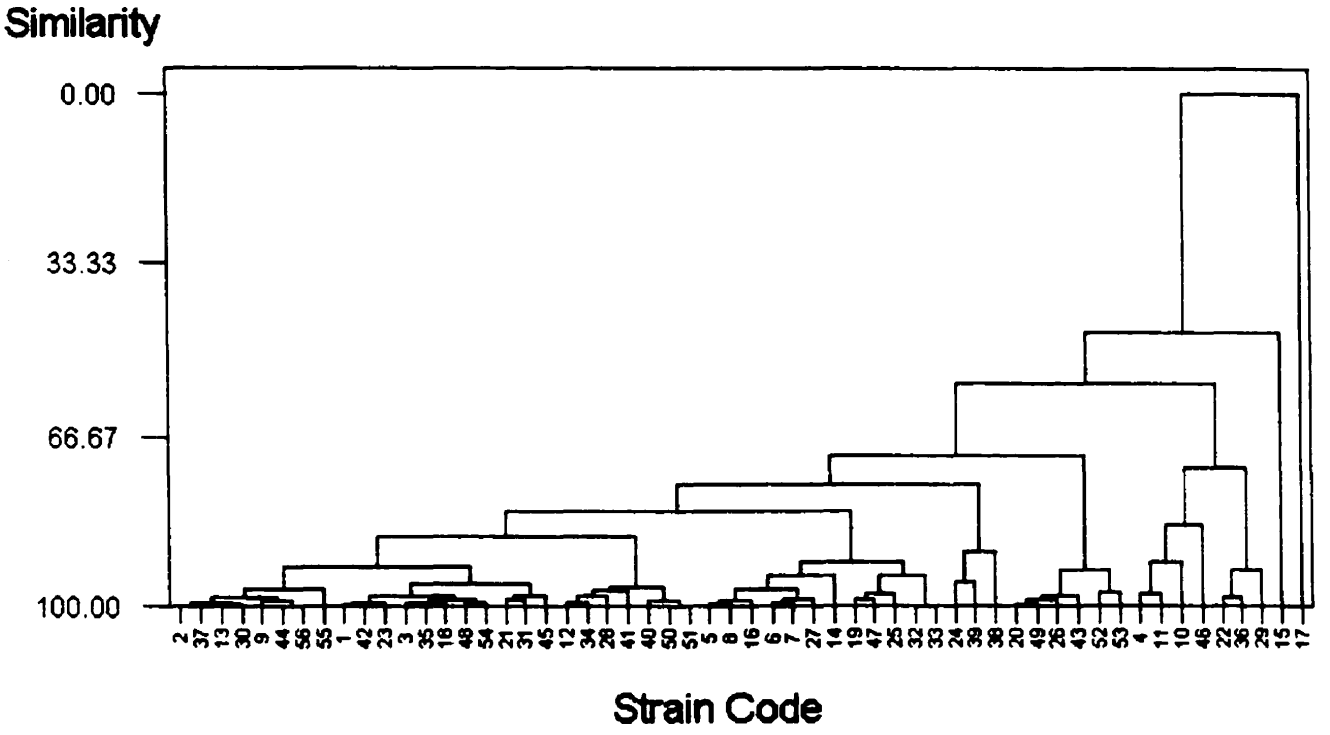


Figure 3.15 A plot of a dendrogram of 56 different yeast strains employing the combination of PCA, discriminate analysis and cluster analysis (please refer to Appendix 1 for strain identity)

3.4.2.2 Supervised analysis

The most important conclusion to be drawn from the above analysis is that the 'unsupervised' learning methods fail to classify the yeast strains correctly and therefore cannot be used to identify them. Accordingly, the use of methods based on 'supervised learning' to identify yeasts from their infrared spectra was investigated. The approach employed was to supervise the analysis using an artificial neural network (ANN)-based expert system. ANNs have recently been successfully employed in the identification of bacteria based on their infrared spectra (Goodacre et al., 1996a; Naumann et al., 1998a; Schmitt et al., 1998).

ANNs are based on a very complex algorithm. Before using this approach to carry out the classification work, several questions had to be addressed. First, in order to obtain the maximum structure and composition information and the minimum noise, how should the spectral data be preprocessed? Four different preprocessed spectral data sets were tested: 1) spectral data obtained after baseline correction, normalization without derivatization or PCA calculation, 2) spectral data obtained after baseline correction, normalization, and computation of the $2^{nd}$ derivative but without the use of PCA calculation, 3) spectral data obtained after baseline correction, normalization without derivatization but with PCA calculation performed, and 4) spectral data obtained after baseline correction, normalization with computation of the $2^{nd}$ derivative and PCA calculation performed. Second, how many input neural units should be used, and how many hidden neural units should be used? If too few hidden units are used, the training error and generalization error will be high due to underfitting and high statistical bias. If

too many hidden units are used, the training error should be low but generalization error will still be high due to overfitting and high variance. The simple and efficient way to solve this problem is to try many networks with different numbers of hidden units, estimate the generalization error for each network, and choose the network with minimum generalization error.

The results obtained show that an ANN provides good classification accuracy compared to classical cluster analysis methods (Table 3.2). This result indicates that ANN may be a more appropriate method for the classification of complex biological systems.

Figure 3.13 shows that the selection of 10 PCs as the input units is satisfactory. When too few PCs are used (e.g. only one or two PCs), not enough information is present, and when too many PCs are employed, the later PCs contribute only noise to the model, thus increasing the probability of chance correlation between input and output data.

The results in Table 3.2 show that training ANNs with all the spectral data (preprocess method A) to develop a classifier does not work. That is because the architecture of this model is too complex and may fit the noise. The best way to avoid overfitting is to use a lot of training data. If at least 30 times as many training cases are used as there are weights in the network, the ANN is unlikely to suffer from much overfitting (Smith, 1996). This means that if all the spectral data points are employed to

train the ANNs, we should have at least 500 (number of input units) × number of hidden units × 4 (number of output units) × 30 training cases, which is quite unpractical. Another way to improve the ANN is to train the network with jitter. Jitter is artificial noise deliberately added to the inputs during training. Training with jitter works because the functions that NNs learns are mostly smooth. NNs can learn functions with discontinuities, but the functions must be piecewise continuous in a finite number of regions if the network is restricted to a finite number of hidden units. In other words, if we have two cases with similar inputs, the desired outputs will usually be similar. That means we can take any training case and generate new training cases by adding small amounts of jitter to the inputs. As long as the amount of jitter is sufficiently small, we can assume that the desired output will not change enough to be of any consequence, so we can just use the same target value (Koistinen and Holmstrom, 1992). Compared to these two methods, PCA combined with ANN is definitely an efficient method to carry out the classification work, because PCA can reduce the number of the input units and at the same time separate the useful information and noise information, it also simplifies the ANN architecture and reduces the number of samples or the need to add jitter. Combined with derivative spectroscopy to reduce the baseline shift and resolve the absorption bands, an increase in the discriminate ability of the classification method can be achieved.

Table 3.2 shows that the spectral preprocessing method which combines derivatization and PCA analysis is the best method for the classification of the yeast spectral data set, with a predictive accuracy of 93.8%. Using the optimum preprocessing

protocol, the optimization of the number of hidden units can be undertaken. Table 3.3 shows that with the increment of the number of hidden units, the predictive accuracy of the neural network will be improved. However, above a certain value, the network's prediction ability will drop due to overfitting. Based on the results in Table 3.3, the optimum number of hidden units was 12. Table 3.3 also shown that for a certain network architecture, if one unit is deleted from its hidden layer, the prediction accuracy of the network will not change too much; for example, the rate of correct prediction with 10 hidden units is the same as that with 11 units. This means that the neural network has a certain fault tolerance, such that the damage of a certain unit will not result in abnormal prediction of the whole network. The weight adjustment process of the network has a kind of auto-repair function so that it can adjust for a certain amount of interruption.

This study clearly showed that FTIR spectroscopy could be used to obtain reproducible biochemical fingerprints from yeast cells. Although classical cluster analysis could not be used to characterize the taxonomic properties of yeasts, an artificial neural network could be trained to identify these yeast strains successfully.

Table 3.2 Effect of different spectral preprocessing techniques on the predictive accuracy of artificial neural networks.

| | Preprocess A | Preprocess B | Preprocess C | Preprocess D |
|---|---|---|---|---|
| Number of correct assignments | 21 out of 32 | 22 out of 32 | 29 out of 32 | 30 out of 32 |
| Percent accuracy | 65.6% | 68.8% | 90.6% | 93.8% |

*Preprocess A*: spectral data (between 1800-800 $cm^{-1}$) obtained after baseline correction, normalization without derivatization or PCA calculation.
*Preprocess B*: spectral data (between 1800-800 $cm^{-1}$) obtained after baseline correction, normalization, and computation of $2^{nd}$ derivative but without the use of PCA calculation.
*Preprocess C*: spectral data (between 1800-800 $cm^{-1}$) obtained after baseline correction, normalization without derivatization but with PCA calculation performed.
*Preprocess D*: spectral data (between 1800-800 $cm^{-1}$) obtained after baseline correction, normalization with computation of $2^{nd}$ derivation and PCA calculation performed.

Note: the default number of hidden units (9), which is set by the software, is used for this test

Table 3.3 Effect of varying the number of hidden units in the hidden layer on the predictive accuracy of the ANN

| | Number of hidden units in the hidden layer | Number of correct assignments | Prediction accuracy |
|---|---|---|---|
| 1 | 3 | 25 out of 32 | 78.1% |
| 2 | 4 | 25 out of 32 | 78.1% |
| 3 | 5 | 29 out of 32 | 90.6% |
| 4 | 6 | 29 out of 32 | 90.6% |
| 5 | 7 | 29 out of 32 | 90.6% |
| 6 | 8 | 29 out of 32 | 90.6% |
| 7 | 9 | 30 out of 32 | 93.8% |
| 8 | 10 | 30 out of 32 | 93.8% |
| 9 | 11 | 30 out of 32 | 93.8% |
| 10 | 12 | 32 out of 32 | 100% |
| 11 | 13 | 32 out of 32 | 100% |
| 12 | 14 | 32 out of 32 | 100% |
| 13 | 15 | 29 out of 32 | 90.6% |

### 3.4.3 Classification of Yeast Strains in Terms of Their Use in Food Production by FTIR Spectroscopy

While the application of genetic engineering to the production of new yeast strains (with desirable features such as the capacity to produce good flavor and aroma; the ability to ferment wort rapidly until fructose, glucose, sucrose, maltose, and maltotriose have been used; the propensity to grow in wort rapidly) has been successful, there have been few instances of induced hybridization to produce commercial brewing yeast strains. Mutation and transformation have also been suggested for producing brewing strains with new properties. With all these new genetically modified yeast strains at hand, the most common question one encounters is: "How can we predict the result of the genetic modification on the function of the microorganism?" The functions and activities of yeast strains are determined by their encoded biological and chemical information. FTIR spectroscopy has the capability to measure the fingerprint of all the biochemical compounds within a microbial cell, such as DNA, RNA, proteins, membrane and cell wall components. Accordingly, it may be a useful technique to predict the effects of genetic modifications on the function of microorganisms.

#### 3.4.3.1 Unsupervised Method

A total of 31 yeast strains used to produce wine, beer and bread were obtained from Lallemand Inc. Their FTIR spectra were collected and preprocessed by the procedure described in Section 3.2. Figure 3.6 shows that most of the spectral information is in the region of 1800-800 $cm^{-1}$. In order to find the region that can be used for yeast classification in terms of their use in the production of wine, beer or bread, one

end of the target spectral region was held constant and the other end narrowed in a stepwise fashion in 100 $cm^{-1}$ increments toward the constant end. This 'scan' action was repeated each time the fixed end was narrowed in 100 $cm^{-1}$ increments in order to identify the best spectral region. In this analysis a centroid linkage and the squared Euclidean distance were employed to carry out the cluster analysis. In centroid linkage each cluster is represented by its centroid; the distance between two clusters is the distance between their centroids. This method does not distort the cluster space.

The results of the analysis of each spectral region are shown in Table 3.4. It is clear that when the region of [x, 800] was used, where x was varied in 100-$cm^{-1}$ increments between 1800 and 1700 $cm^{-1}$, all of the wine strains were identified correctly except two, but they were not classified as a single group; instead, they were separated by the beer group or the bread group into two isolated groups (Figure 3.17). The same spectral region could also be employed to separate bread and beer yeasts from each other with a small error. When the region of [x, 800] was employed, where x was varied in 100-$cm^{-1}$ increments between 1600 and 1400 $cm^{-1}$, the cluster analysis algorithm grouped all the wine strains correctly; it also grouped all the beer strains correctly except three and all the bread strains correctly except one. When the region of [x, 800] was employed, where x was varied in 100-$cm^{-1}$ increments between 1300 and 1100 $cm^{-1}$, the algorithm grouped all the wine strains into one single group correctly; it grouped all the beer strains correctly except three; the algorithm also correctly grouped all the bread strains (Figure 3.18). When the region of 1000-800 $cm^{-1}$ was employed, the algorithm correctly grouped all the wine strains, all the beer strains except three; and all the bread strains except two.

When the region of 900-800 $cm^{-1}$ was employed, the algorithm had difficulties in classifying all the strains.

Based on the results from Table 3.4, the best region to group the wine strains alone is [x, 800] where x is varied in 100-$cm^{-1}$ increments between 1300 and 1100 $cm^{-1}$. Absorptions in this region include the amide III band components of proteins (1310-1240 $cm^{-1}$), the P=O stretching vibration of $>PO_2^-$ in phosphordiesters (1250-1220 $cm^{-1}$) and phosphodiesters (1088-1084 $cm^{-1}$), and the ring vibrations of carbohydrates (1200-900 $cm^{-1}$). The best region to identify beer strains alone is [x, 800] and [x, 900] where x is varied in 100-$cm^{-1}$ increments between 1800 and 1700 $cm^{-1}$ and between 1800 and 1600 $cm^{-1}$, respectively. The best region to identify bread strains alone is [x, 800] where x is varied in 100-$cm^{-1}$ increments between 1700 and 1600 $cm^{-1}$ and [x, 800] where x is varied in 100-$cm^{-1}$ increments between 1300 and 1100 $cm^{-1}$. Finally, the best region to separate these three kinds of yeast strains is [x, 800] where x is varied in 100-$cm^{-1}$ increments between 1300 and 1100 $cm^{-1}$; the classification results using cluster analysis and the spectral data from this region are presented in Figure 3.18.

It can be concluded from this study that unsupervised analysis methods can be used for classification of yeast strains in terms of the type of fermentation process in which they are employed and that this method yield >90% correct classification.

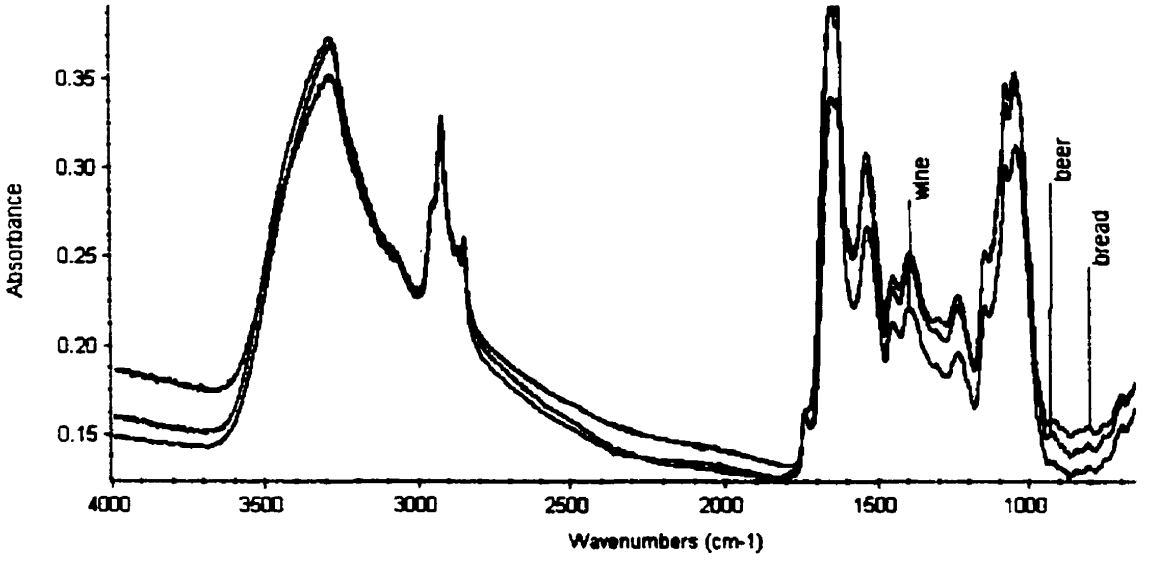


Figure 3.16 Comparison between the FTIR spectra of a wine, a beer and a bread yeast strains

Table 3.4 Effect of selection of the infrared spectral region between 1800 and 800 $cm^{-1}$ on the predictive accuracy of yeast classification in terms of their use in the production of wine, beer, and bread by cluster analysis.

| Region ($cm^{-1}$) | number of incorrect wine strain assignments | number of incorrect beer strain assignments | number of incorrect bread strain assignments |
|---|---|---|---|
| 1800/1700-800 | 2* out of 16 | 1 out of 12 | 0 out of 3 |
| 1600/1400-800 | 0* out of 16 | 3 out of 12 | 1 out of 12 |
| 1300/1100-800 | 0 out of 16 | 3 out of 12 | 0 out of 12 |
| 1000 -800 | 0* out of 16 | 3 out of 12 | 2 out of 12 |
| 900 -800 | Inadequate discrimination | Inadequate discrimination | Inadequate discrimination |
| 1800/1600-900 | 0* out of 16 | 1 out of 12 | 2 out of 12 |
| 1500/1400-900 | 2* out of 16 | 3 out of 12 | 1 out of 12 |
| 1300/1100-900 | 2* out of 16 | 2 out of 12 | 2 out of 12 |
| 1000 -900 | Inadequate discrimination | Inadequate discrimination | Inadequate discrimination |
| 1800/1700-1000 | 0* out of 16 | 2 out of 12 | 2 out of 12 |
| 1600 -1000 | 2* out of 16 | 2 out of 12 | 2 out of 12 |
| 1500/1300-1000 | 0* out of 16 | 3 out of 12 | 1 out of 12 |
| 1200/1100-1000 | Inadequate discrimination | Inadequate discrimination | Inadequate discrimination |
| 1800/1700-1100 | 0* out of 16 | 2 out of 12 | 2 out of 12 |
| 1600 -1100 | 2* out of 16 | 3 out of 12 | 2 out of 12 |
| 1500/1300-1100 | 1* out of 16 | 3 out of 12 | 2 out of 12 |
| 1200 -1100 | Inadequate discrimination | Inadequate discrimination | Inadequate discrimination |
| 1800/1600-1200 | 0* out of 16 | 3 out of 12 | 2 out of 12 |
| 1500/1400-1200 | 2* out of 16 | 2 out of 12 | 2 out of 12 |
| 1300 -1200 | Inadequate discrimination | Inadequate discrimination | Inadequate discrimination |
| 1800 -1300 | 2* out of 16 | 2 out of 12 | 2 out of 12 |
| 1700/1500-1300 | 2* out of 16 | 2 out of 12 | 2 out of 12 |
| 1400 -1300 | Inadequate discrimination | Inadequate discrimination | Inadequate discrimination |
| 1800/1600-1400 | 3* out of 16 | 3 out of 12 | 2 out of 12 |
| 1500 -1400 | Inadequate discrimination | Inadequate discrimination | Inadequate discrimination |
| 1800/1700-1500 | 3* out of 16 | 2 out of 12 | 2 out of 12 |
| 1600 -1500 | Inadequate discrimination | Inadequate discrimination | Inadequate discrimination |
| 1800 -1600 | Inadequate discrimination | Inadequate discrimination | Inadequate discrimination |
| 1700 -1600 | Inadequate discrimination | Inadequate discrimination | Inadequate discrimination |
| 1800 -1700 | Inadequate discrimination | Inadequate discrimination | Inadequate discrimination |

*note: the wine group is divided into two separate groups

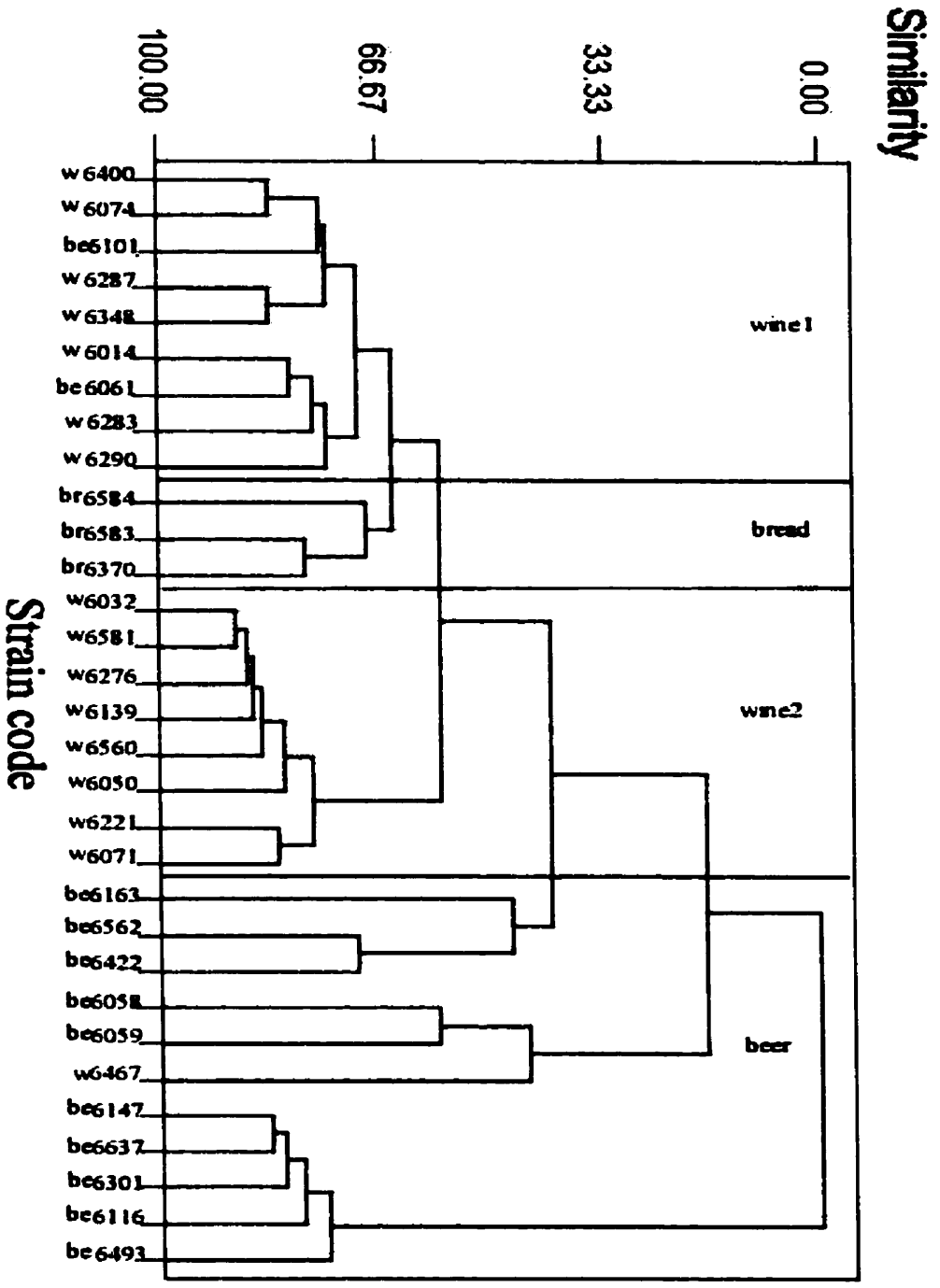


Figure 3.17 A plot of a dendrogram generated from the cluster analysis of 31 different yeast strains (employed in the production of wine, beer and bread) based on the changes in infrared spectral region between 1700-800 $cm^{-1}$ after baseline correction, normalization and computation of the second derivative data and 9 point smoothing of the FTIR spectra of the yeast strains
('w' refers to wine, 'be' refers to beer and 'br' refers to bread)

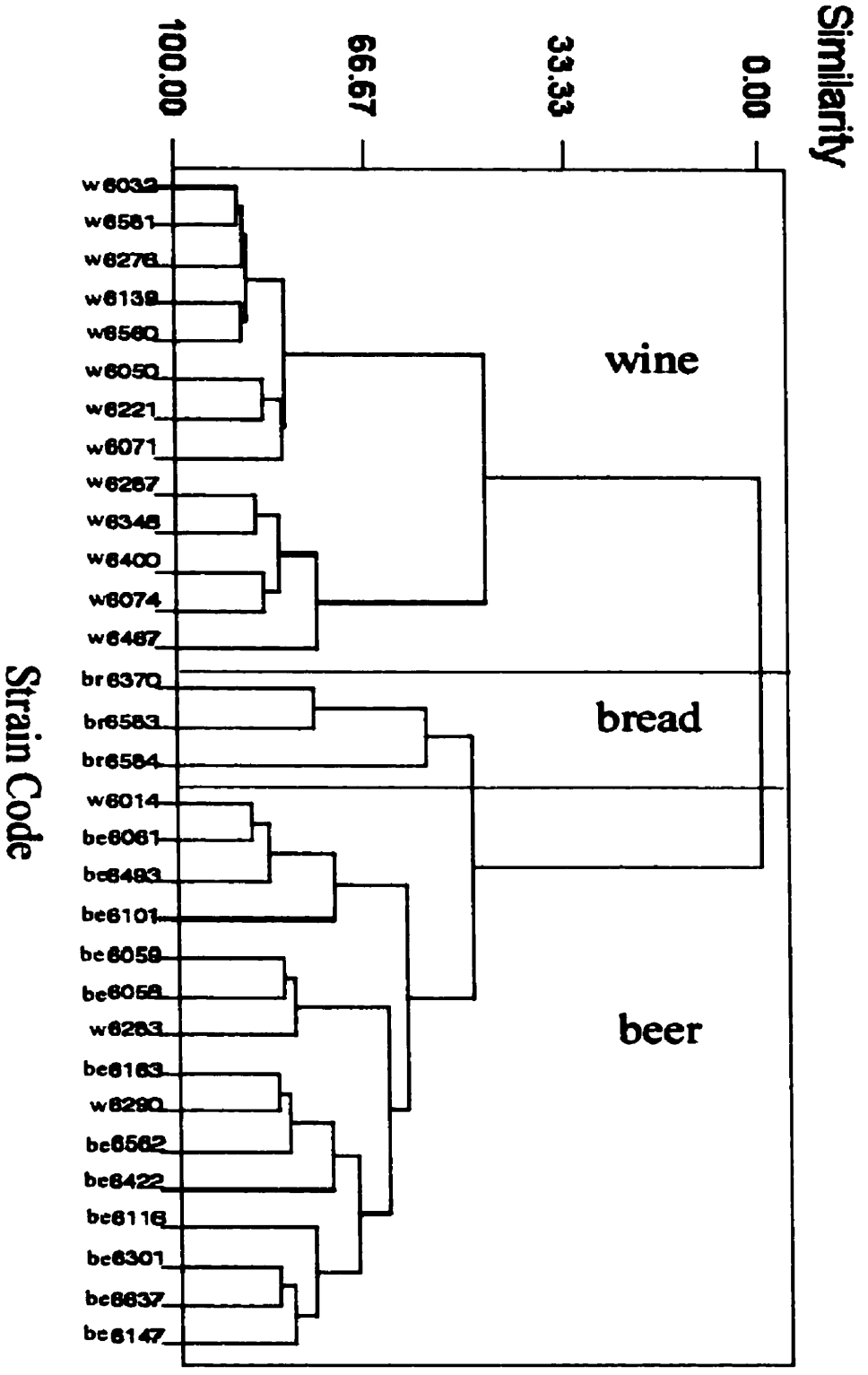


Figure 3.18 A plot of a dendrogram generated from the cluster analysis of 31 different yeast strains (employed in the production of wine, beer and bread) based on the changes in infrared spectral region between 1300-800 $cm^{-1}$ after baseline correction, normalization and computation of the second derivative data and 9 point smoothing of the FTIR spectra of the yeast strains. ('w' refers to wine, 'be' refers to beer and 'br' refers to bread)

3.4.3.2 Supervised Method

The discussion above shows that classical cluster analysis can work. However, the process of searching for the best spectral region for yeast classification is tedious. The possibility of utilizing supervised pattern recognition methods such as neural networks, which mimic the human brain in its learning process and further apply the knowledge to solving problems, may simplify the analysis.

The following procedure was employed in the training of an ANN to classify yeasts in terms of their use in the production of wine, beer, or bread. Principal component analysis (PCA) using the NIPALS algorithm was carried out using SCAN (Minitab Inc., State College, PA, USA) on the preprocessed spectral data (spectral data obtained after baseline correction, normalization, computation of $2^{nd}$ derivative and 9-point smoothing in the region of 1800-800 $cm^{-1}$). The first 8 PCs were used as input to the network. The ANN was built using Neuroshell 2 (Ward System Group Inc, Frederick, MD, USA). The structure of the ANN used in this study consisted of three layers: one input layer (containing 8 units), one output layer (containing 3 units) and one hidden layer (containing 10 units). To train the ANN, each of the inputs was normalized and paired with each of the desired outputs (the output layer was binary encoded such that wine is represented by 100, beer by 010 and bread by 001). Before training commenced, the connection weights were set to small random values. TurboProp algorithm was employed to train the neural network TurboProp algorithm is a training method faster in the "batch" mode than other backpropagation methods, and it is not sensitive to learning rate and momentum, so learning rate and momentum are not required to be set during the training

procedure. In this algorithm, training proceeds through an entire epoch (one complete calculation in the network is called an epoch) before the weights are updated. It adds all of the weight changes and at the end of an epoch modifies the weights. The TurboProp method utilizes an independent weight update size for each different weight, rather than the usual method of having a single learning rate and momentum that applies to all weights. Furthermore, the step sizes are adaptively adjusted as learning progresses. Before training commences, the connection weights are set to small random values, including the weights connecting the bias to the hidden and output layers. Next, the input values are applied to the network, which is allowed to run until an output is produced at each output unit. The differences between the actual output and that expected, taken over the entire set of patterns, are fed back through the network in the reverse direction to signal flow (hence backpropagation) modifying the weights as they go. This process is repeated until a suitable level of error is achieved. To prevent overfitting: 1) the test set and production set were extracted from the original spectral data set by the software. The test set is used with calibration, during training the network was interrogated with patterns in the test set, the errors between the output seen and that expected were calculated, thus allowing a learning curve for the test set to be drawn. Training is stopped when the RMS error on the test set is lowest. The production set is used to test the network's results with data the network has never "seen" before. 2) the ANNs were trained five times to determine whether they converged reproducibly.

When training had ceased, the network was interrogated. As expected, the network's estimate of the identities of the yeasts in the calibration set were the same as

their known identities. The results of the network's final analysis of the unknown production set is shown in Table 3.5; the difference between the ANN's estimates and the output that was expected is also given. It can be seen in Table 3.5 that all the unknowns were correctly identified.

This study clearly showed that FTIR spectroscopy could discriminate between different yeast strains in terms of their use in the production of wine, beer, or bread. Artificial neural networks were also successfully trained to fulfill this objective. We conclude that the combination of FTIR spectroscopy and ANNs provides a rapid and accurate discriminatory technique.

Table 3.5 Artificial neural network classification results for 31 yeast strains

| | | Strain Code | Act1 | Act2 | Act3 | | Net 1 | Net 2 | Net 3 | | Act-Net 1 | Act-Net 2 | Act-Net 3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Training Set | Wine | 6467 | 1 | 0 | 0 | | **0.87** | 0.09 | 0.39 | | 0.13 | -0.1 | -0.4 |
| | | 6074 | 1 | 0 | 0 | | **0.98** | 0.02 | 0.20 | | 0.02 | 0 | -0.2 |
| | | 6400 | 1 | 0 | 0 | | **0.77** | 0 | 0.09 | | 0.23 | 0 | -0.1 |
| | | 6287 | 1 | 0 | 0 | | **0.99** | 0.01 | 0.05 | | 0.01 | 0 | -0.1 |
| | | 6071 | 1 | 0 | 0 | | **0.83** | 0.43 | 0.11 | | 0.17 | -0.4 | -0.1 |
| | | 6050 | 1 | 0 | 0 | | **0.69** | 0.00 | 0.09 | | 0.31 | 0 | -0.1 |
| | | 6139 | 1 | 0 | 0 | | **0.85** | 0 | 0.08 | | 0.15 | 0 | -0.1 |
| | | 6276 | 1 | 0 | 0 | | **0.68** | 0 | 0.2 | | 0.32 | 0 | -0.2 |
| | | 6032 | 1 | 0 | 0 | | **0.93** | 0 | 0.16 | | 0.07 | 0 | -0.2 |
| | | 6014 | 1 | 0 | 0 | | **0.66** | 0 | 0.01 | | 0.34 | 0 | |
| | | 6283 | 1 | 0 | 0 | | **0.99** | 0 | 0.22 | | 0.01 | 0 | -0.2 |
| | | 6290 | 1 | 0 | 0 | | **0.59** | 0 | 0 | | 0.41 | 0 | 0 |
| | | | | | | | | | | | | | |
| | Beer | 6061 | 0 | 1 | 0 | | 0.13 | **0.98** | 0.32 | | -0.1 | 0.02 | -0.3 |
| | | 6493 | 0 | 1 | 0 | | 0.17 | **0.65** | 0.15 | | -0.2 | 0.35 | -0.2 |
| | | 6101 | 0 | 1 | 0 | | 0.12 | **0.85** | 0.27 | | -0.1 | 0.15 | -0.3 |
| | | 6059 | 0 | 1 | 0 | | 0.05 | **0.86** | 0.03 | | -0.1 | 0.14 | 0 |
| | | 6562 | 0 | 1 | 0 | | 0.32 | **0.80** | 0.06 | | -0.3 | 0.2 | -0.1 |
| | | 6422 | 0 | 1 | 0 | | 0.21 | **0.81** | 0.05 | | -0.2 | 0.19 | -0.1 |
| | | 6116 | 0 | 1 | 0 | | 0.25 | **0.96** | 0.01 | | -0.3 | 0.04 | 0 |
| | | 6637 | 0 | 1 | 0 | | 0.26 | **0.54** | 0.08 | | -0.3 | 0.46 | -0.1 |
| | | | | | | | | | | | | | |
| | Bread | 6370 | 0 | 0 | 1 | | 0.22 | 0.05 | **0.65** | | -0.2 | -0.1 | 0.35 |
| | | 6583 | 0 | 0 | 1 | | 0.07 | 0 | **0.75** | | -0.1 | 0 | 0.25 |
| Test Set | Wine | 6221 | 1 | 0 | 0 | | **0.55** | 0.28 | 0.07 | | 0.45 | -0.3 | -0.1 |
| | | 6581 | 1 | 0 | 0 | | **0.85** | 0.24 | 0.25 | | 0.15 | -0.2 | -0.3 |
| | | | | | | | | | | | | | |
| | Beer | 6147 | 0 | 1 | 0 | | 0.21 | **0.97** | 0.27 | | -0.2 | 0.03 | -0.3 |
| | | 6058 | 0 | 1 | 0 | | 0.25 | **0.62** | 0.40 | | -0.3 | 0.38 | -0.4 |
| | | | | | | | | | | | | | |
| | Bread | 6584 | 0 | 0 | 1 | | 0.36 | 0 | **0.89** | | -0.4 | 0 | 0.11 |
| Production Set | Wine | 6560 | 1 | 0 | 0 | | **0.65** | 0.34 | 0.48 | | 0.35 | -0.3 | -0.4 |
| | | 6348 | 1 | 0 | 0 | | **0.94** | 0.35 | 0.05 | | 0.06 | -0.4 | -0.1 |
| | | | | | | | | | | | | | |
| | Beer | 6163 | 0 | 1 | 0 | | 0.17 | **0.93** | 0.07 | | -0.2 | 0.07 | -0.1 |
| | | 6301 | 0 | 1 | 0 | | 0.01 | **0.78** | 0.18 | | 0 | 0.22 | -0.2 |

Note: all of the neural network output values given are the averages from training the network five times; the bold values indicate the correct class.

Act I ($1 \leq I \leq 3$ ) refers to the expected output

Net I ($1 \leq I \leq 3$ ) refers to the actual output

### 3.4.4 Classification of Yeast Strains in Terms of Their Sensitivity to Killer Yeast Strains by FTIR Spectroscopy

Killer yeast strains (phenotype K+R+) produce an extracellular toxin that kills sensitive yeast strains (phenotype K-R-). There also exist neutral yeast strains (phenotype K+R-) that are resistant to killer toxin but do not produce it. Exotoxins (generally proteins or glycoproteins) that are able to kill susceptible cells belonging to the same or congeneric species have been defined as killer toxins. Killer yeast strains are toxin-producing fungi that are immune to the activity of their own killer toxins. The killer phenomenon was discovered in yeast by Bevan and Mackower (1963). The most thoroughly investigated yeast killer system is that of *S. cereviviae* (Bussey, 1991; Tipper, et al., 1991;Wickner, 1992, 1996). Currently, the killer yeasts belonging to this species have been classified into three main groups (K1, K2, and K28) on the basis of the molecular characteristics of the secreted toxins, their killing profiles, the lack of cross-immunity, and the encoding genetic determinants. They are constituted by strains producing toxins encoded by dsRNA. Other killer yeasts producing toxins named KHR and KHS, which are encoded on chromosomal DNA, have also been identified (Goto et al., 1990; 1991). The K1, K2, and K28 toxins are encoded by different cytoplasmically inherited satellite dsRNAs (M1, M2, and M28), encapsidated in virus-like particles (VLPs) and dependent on another group of helper yeast viruses (L-A) for their replication and encapsidation. The M dsRNAs are responsible for either killer activity or self-immunity, a phenotype that is characteristic of yeast killer toxin-producing strains.

The study of killer yeast strains is very important and useful. For example, the yeast killer system has been proved to be fruitful not only in the differentiation of important slowly growing pathogenic bacteria, such as the mycobacteria, but also in the differentiation of faster-growing gram-positive and gram-negative bacteria (Wickner, 1992). When used to investigate the serotypes of bacterial isolates, the yeast killer system was able to differentiate isolates of *Neisseria meningitidis* group C (Morace et al., 1989). The yeast killer system, when properly used, has been proved to be of great value in the identification of the species and varieties of heterogeneous microorganisms (Morace et al., 1988). Stuck wine fermentation is one of the most important problems in the wine industry (Lafon-Lafourcade et al., 1984; Kunkee, 1991). Several causes of stuck and sluggish wine fermentation have been described (Ribereau-Gayon et al., 1975; Rosini, 1983). As expected, killer toxins can inhibit wine fermentation by sensitive yeasts (Van and Wingfield, 1986).

In this study, we investigated the use of FTIR spectroscopy for the classification of yeasts in terms of their sensitivity to killer yeast strains. Due to the limited number of strains, here we tried to classify all the available yeast strains into two groups: sensitive strains and non-sensitive strains (which include possess and neutral strains).

All the 25 yeast strains (19 sensitive strains and 6 non-sensitive strains) were obtained from Lallemand Inc. Their infrared spectra were collected and preprocessed by procedures described in Section 3.2. Figure 3.19 shows that most of the spectral information is in the region between 1800 and 800 $cm^{-1}$. In order to find a region that can

be used for the yeast classification in terms of their sensitivity to the killer yeast strains, various spectral regions were tried as described in the previous section.

The results of the analysis of each spectral region are shown in Table 3.6. When a spectral region of [x, 800], where x is varied in 100-$cm^{-1}$ increments between 1800 and 1700 $cm^{-1}$, was employed, the cluster analysis algorithm divided the 25 strains into two groups: one a sensitive group, and the other a non-sensitive group. Within the sensitive group, there are 21 strains in total; 19 strains indeed belong to sensitive group and 2 strains do not. The sensitive group is divided into two separate groups by the non-sensitive group. In the non-sensitive group there are 4 strains in total, all of which are non-sensitive strains. When the region of 1600-800 $cm^{-1}$ was employed, the algorithm classified 18 sensitive strains and 2 non-sensitive strains in the sensitive group, and 4 non-sensitive strains and 1 sensitive strain in the non-sensitive group. When the region of [x, 800], where x is varied in 100-$cm^{-1}$ increments between 1500 and 1000 $cm^{-1}$, was employed, the algorithm classified 19 sensitive strains and 2 non-sensitive strains in the sensitive group and 4 non-sensitive strains in the non-sensitive group. When the region of 900-800 $cm^{-1}$ was employed, the algorithm classified 18 sensitive strains in the sensitive group and 6 non-sensitive strains and one sensitive strain in the non-sensitive group. Other details can be found in Table 3.6.

Based on the results from Table 3.6, the optimum region to classify yeast strains in terms of their sensitivity to killer yeast strains is 900-800$cm^{-1}$. The dendrogram produced by cluster analysis using the spectral data from this region is shown in Figure

3.20. This spectral region is dominated by C-O-C, C-O and ring vibrations of carbohydrates and C-H rocking of >$CH_2$ methylene groups.

It can be concluded from this study that unsupervised analysis methods can be used as an approach of classification of yeast strains in terms of their sensitivity to killer yeast strains and this method yields > 90% correct classification.

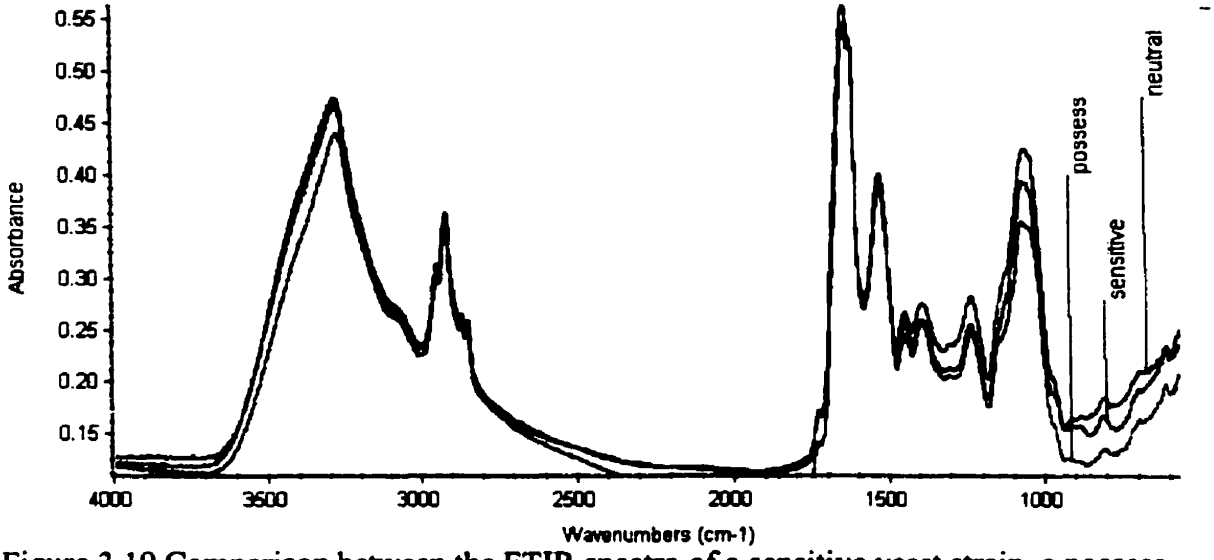


Figure 3.19 Comparison between the FTIR spectra of a sensitive yeast strain, a possess yeast strain and a neutral yeast strain

Table 3.6 Effect of selection of infrared spectral regions between 1800 and 800 $cm^{-1}$ on the predictive accuracy of yeast classification in terms of their sensitivity by cluster analysis.

| | Sensitivity | Non-sensitivity |
|---|---|---|
| 1800/1700-800 cm-1 | 19+2- * | 4+ |
| 1600 -800 cm-1 | 18+2- | 4+1- |
| 1500/1000-800 cm-1 | 19+2- * | 4+ |
| 900 -800 cm-1 | 18+ | 6+1- |
| 1800/1500-900 cm-1 | 18+2- | 4+1- |
| 1400/1100-900 cm-1 | 19+2- * | 4+ |
| 1000 -900 cm-1 | 19+2- * | 4+ |
| 1800/1600-1000 cm-1 | 17+ | 6+2- |
| 1500 -1000 cm-1 | 17+1- | 5+2- |
| 1400 -1000 cm-1 | 19+2- | 4+ |
| 1300/1200-1000 cm-1 | 18+2- | 4+1- |
| 1100 -1000 cm-1 | 18+2- * | 4+1- |
| 1800/1700-1100 cm-1 | 18+1- | 5+1- |
| 1600 -1100 cm-1 | 17+ | 6+2- |
| 1500 -1100 cm-1 | 17+1- | 5+2- |
| 1400 -1100 cm-1 | 18+1- | 5+1- |
| 1300 -1100 cm-1 | 17+1- | 5+2- |
| 1200 -1100 cm-1 | 17+1- * | 5+2- |
| 1800/1700-1200 cm-1 | 17+1- | 5+2- |
| 1600 -1200 cm-1 | 16+ | 6+2- |
| 1500 -1200 cm-1 | 17+1- | 5+2- |
| 1400 -1200 cm-1 | 15+ | 6+4- |
| 1300 -1200 cm-1 | 17+2- | 4+2- |
| 1800/1700-1300 cm-1 | 17+1- | 5+2- |
| 1600 -1300 cm-1 | 17+ | 6+2- |
| 1500 -1300 cm-1 | 14+ | 6+5- |
| 1400 -1300 cm-1 | Inadequate discriminant | Inadequate discriminant |
| 1800/1600-1400 cm-1 | 17+ | 6+2- |
| 1500 -1400 cm-1 | 15+ | 6+4- |
| 1800/1700-1500 cm-1 | Inadequate discriminant | Inadequate discriminant |
| 1600 -1500 cm-1 | Inadequate discriminant | Inadequate discriminant |
| 1800 -1600 cm-1 | Inadequate discriminant | Inadequate discriminant |
| 1700 -1600 cm-1 | Inadequate discriminant | Inadequate discriminant |
| 1800 -1700 cm-1 | Inadequate discriminant | Inadequate discriminant |

Note * means the whole group was divided into two separated groups
+ means the correct assignments
- means the incorrect assignments

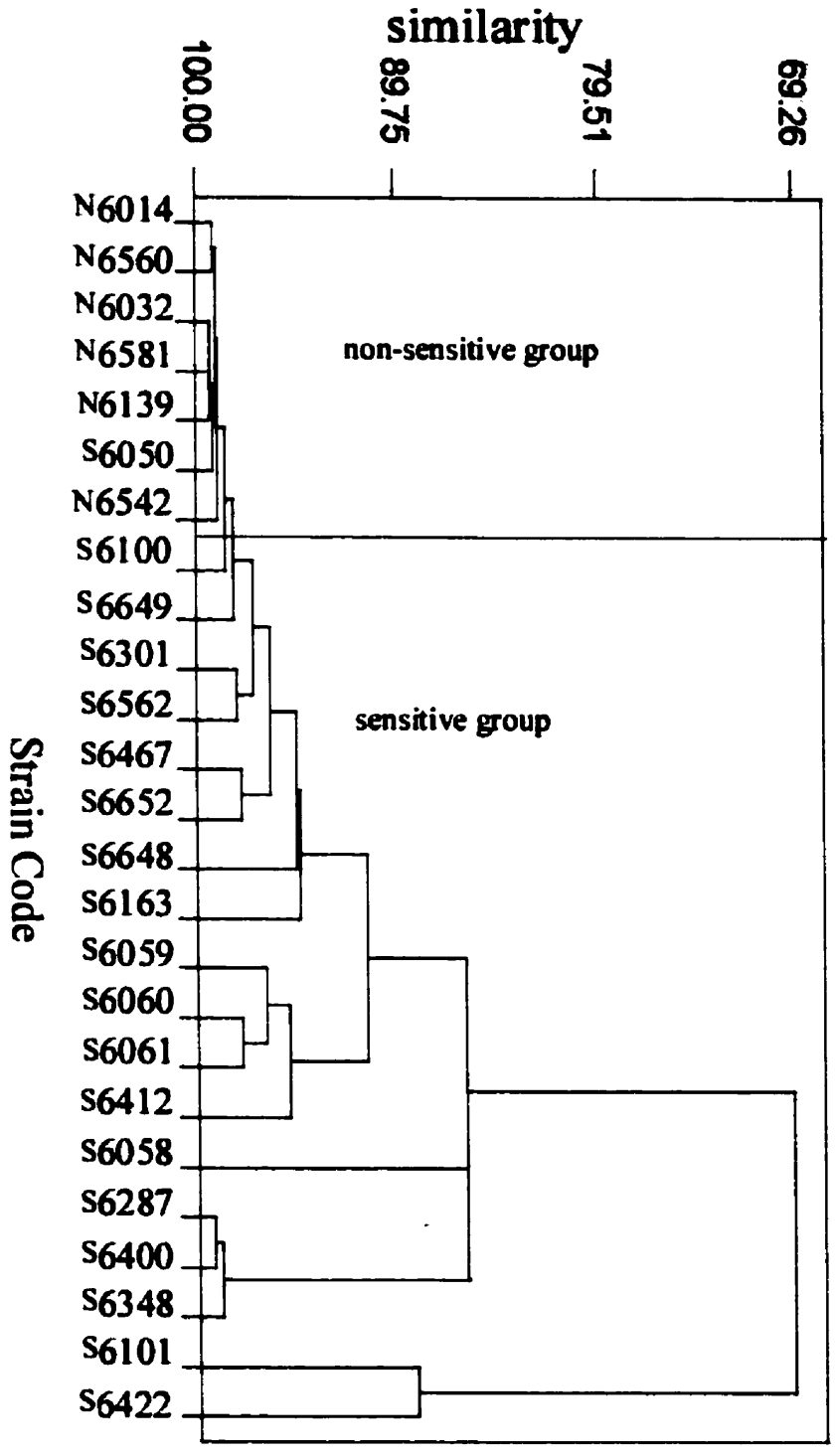


Figure 3.20 A plot of a dendrogram showing the results of the cluster analysis classification of 25 yeast strains in terms of their sensitivity to killer yeast strains employing infrared spectra of the yeast strains in the region between 900-800 $cm^{-1}$ ('N' refers to non-sensitive and 'S' refers to sensitive)

## Chapter 4 Conclusion

The objective of the research presented in this thesis was to investigate the feasibility of employing FTIR spectroscopy for the classification of yeast in terms of taxonomy, use in production of wine, beer, and bread, and sensitivity to killer strains.

In order to obtain reproducible spectra of yeasts, a strict control of growth conditions, isolation protocol, and sampling methodology was required. The spectral reproducibility was then evaluated by a spectral library search approach and the use of Pearson's correlation coefficient. All the results showed that the methodology of spectral acquisition and sampling protocol developed in this work produced highly reproducible spectra from different batches of the same strain.

Different classical classification approaches based on hierarchical clustering (including region selection, region selection combined with weighting factor, PCA combined with hierarchical clustering, PCA and discriminate analysis combined with hierarchical clustering) were evaluated. It was found that these unsupervised classification approaches had difficulties in assigning all the spectra of the yeast strains to the correct groups. This may be attributed to the similarity of the strains in terms of their biochemical composition. Supervised learning methods employing an artificial neural network (ANN) were then evaluated in combination with different spectral preprocessing techniques. It was found that baseline correction compensated for some of the light scattering from yeast samples deposited as films on a ZnSe window, while spectral normalization compensated for some of the variability in film thickness. The use of

second-derivative spectra also reduced the effect of baseline variation and resolved the absorption peaks. Because of the large amount of spectral information, the use of principal component spectra in place of the raw spectral data results in the reduction of the dimensionality of the information. Thus, the combined advantages of spectral data processing have been found to improve the performance of the ANN models.

Classification of yeasts in terms of the type of fermentation process in which they serve is a new approach in the application of FTIR spectroscopy to microbiology. In this study, both cluster analysis and ANN were equally effective in the classification of yeast strains. Cluster analysis was effective when the spectral region was narrowed between 1300 and 800 $cm^{-1}$. An ANN successfully predicted 100% of the test set.

Classification of yeasts in terms of their sensitivity to killer yeast strains has economic implications in relation to the efficacy of the production process. Because of the limited amount of strains available, only the cluster analysis algorithm was employed. The separation of the yeast trains into two distinct groups, sensitive and non-sensitive, was accomplished by employing the spectral information between 800 and 900 $cm^{-1}$.

It can be concluded from the results obtained from this work that FTIR spectroscopy in combination with suitable chemometric techniques can be of potential utility for the rapid identification and classification of yeast strains. Future work should be directed toward increasing the size of the spectral database and carrying out extensive

validation studies of production samples, with emphasis placed on employing supervised training for spectral analysis.

## References


Alsberg, B.K., Wade, W.G. & Goodacre, R. 1998. Chemometric analysis of diffuse reflectance-absorbance Fourier transform infrared spectra using rule induction methods: application to the classification of *Eubacterium* species. Applied Spectroscopy 52, 823-832.

Banwell, C. N. 1983. Fundamentals of molecular spectroscopy, 3rd Ed. McGraw-Hill MaidenHead, England, 338p.

Barns, S. M., Lane. D.J., Sogin, M. L., Bibeau, C., and Weisburg, W.G. 1991 Evolutionary relationships among pathogenic Candida species and relatives. J. Bacteriol. 173:2250-2255

Bassoe, C.F. , O.D. Laerum, J. Glette, G. Hopen, G. Haneburg, and C. O. Solberg. 1983. Simultaneous measurement of phagocytosis and phagosomal pH by flow cytometry: Role of polymorphonuclear neutrophilic leukocyte granules in phagosome acidification. Cytometry 4:254-262

Bassoe, C. F., and R. Bjerknes. 1985. Phagocytosis by human leukocytes, phagosomal pH and degradation of seven species of bacteria measured by flow cytometry. J. Med. Microbiol. 19:15-125

Bevan, E. A., and M. Makower. 1963 The inheritance of a killer character in yeast (*Saccharomyces Cerevisiae*). Proc. Int. Congr. Genet. 1:202-203

Bishop, C.M. 1995. Neural Networks for Pattern Recognition, Oxford: Oxford University Press.

Boekhout, T., Fonseca, A., Sampaio, J. P., and golubev, W. I. 1993 Classification of heterobasidiomycetous yeasts: characteristics and affiliation to higher taxa of Heterobasidiomycetes. Can. J. Microbiol. 39:276-290

Borel, M.; Lynch, B. M. 1993 A study of differentiation of living bacteria by ATR/FTIR spectroscopy. Can. J. Appl. Spectrosc. 38(1), 18-21.

Borman, S.A., 1983. Fouier Transform IR: Are the older grating instruments going the way of dinosaurs? Anal. Chem. 55, 1054A.

Botha,A. and Kock, J.F. L. 1993 Application of fatty acid profiles in the identification of yeasts. Int. J. Food Microbiol. 19:39-51

Brondz, I. and I. Olsen 1986. Review. Microbial chemotaxonomy. Chromatography, electrophoresis, and relevant profiling techniques. J. Chromatogr. 379:367-411

Brown, C.D. and Wentzell, P.D. 1999. J. Chemom. 13(2), 133-152

Bruneau,S. and Guinet, R. 1989 Rapid identification of medically important yeasts by electrophoretic protein patterns. FEMS Microbiol. Lett. 58:329-339

Bruno P. 2000 Future trends in identification of microorganisms, a brilliant future also for the food industry? Innovations Food Technol. 8, 49-52.

Bussey, H. 1981. Physiology of killer factor in yeast. Adv. Microb. Physiol. 22:93-122

Carledge, T.G. Rose, A. H., Belk, D. M., Goodall, A. H. 1977 J. Bacteriol. 126, 426-433

Cummins, D. J. and Andrews, C. W. 1995. Iteratively reweighted partial least squares. A performance analysis by Monte Carlo simulation. J. Chemom. 9(6), 489-507.

Darken, C. and Moody, J. 1992. Towards faster stochastic gradient search. in Moody, J.E., Hanson, S.J., and Lippmann, R.P., eds. Advances in Neural Information Processing Systems 4, San Mateo, CA: Morgan Kaufmann Publishers, pp. 1009-1016.

Deak, T.. 1999 Molecular taxonomy of yeasts. Acta Microbiol. Immunol. Hung. 46 (2-3), 181-186.

Degre, R., Thomas, D. Y., Ash, J., Mailhot, K., Morin, A., and Dubord, C. 1989. Wine yeast strain identification. Am. J. Enol. Vitic. 40:309-315

Del Castillo Agudo, L. 1992. Lipid content of Saccharomyces cerevisiae strains with different degrees of Candida albicans with DNA probes. Curr. Microbiol. 26:57-60

DeVore, R.A., Howard, R., and Micchelli, C.A. 1989. Optimal nonlinear approximation. Manuscripta Mathematica, 63, 469-478.

Doak, D. L.; Phillips, J. A. 1999 In situ monitoring of an Escherichia coli fermentation using a diamond composition ATR probe and mid-infrared spectroscopy. Biotechnol. Prog. 15(3), 529-539.

Drapcho, D. L.; Crocombe, R. A.; Seebode, J. 1997. Advances in photoacoustic step-scan FT-IR spectroscopy. Mikrochim. Acta, Suppl. 14(Progress in Fourier Transform Spectroscopy), 585-588.

Dubois, J., 1999. Selected applications of Fourier Transform Infrared Spectroscopy to the study of cells and cellular components Ph.D. Thesis McGill University

Fahlman, S.E. 1989 Faster-Learning Variations on Back-Propagation: An Empirical Study, in Touretzky, D., Hinton, G, and Sejnowski, T., eds., Proceedings of the 1988 Connectionist Models Summer School, Morgan Kaufmann, 38-51

Fayolle, P.; Picque, D.; Corrieu, G.. 1997. Monitoring of fermentation processes producing lactic acid bacteria by mid-infrared spectroscopy. Vib. Spectrosc. 14(2), 247-252.

Forbes, B. A., and Hicks, K.E.S. 1993. Direct detection of Mycobacterium tuberculosis in respiratory specimens in a clinical laboratory by polymerase chain reaction. J. Clin. Microbiol. 31:1688-1694.

Foster, L. M., Kozak, K. R., Loftus, M. G. Stevens, J.J., and Ross, I.K. 1993 The polymerase chain reaction and its application to filamentous fungi. Mycol. Res. 97:769-781

Frei, H. 1998. Nanosecond step-scan FT-infrared absorption spectroscopy in photochemistry and catalysis. AIP Conf. Proc. 430(Fourier Transform Spectroscopy), 28-39.

Geman, S., Bienenstock, E. and Doursat, R. 1992. Neural Networks and the Bias/Variance Dilemma. Neural Computation, 4, 1-58.

Gil, J. A. and Romera, R. 1998. On robust partial least squares (PLS) methods. J. Chemom12(6), 365-378.

Gilbert, R.J., Goodacre, R., Woodward, A.M. & Kell, D.B. 1997 Genetic programming : a novel method for thequantitative analysis of pyrolysis mass spectral data. Analytical Chemistry 69, 4381-4389

Goodacre, R. & Kell, D.B. 1996a Correction of mass spectral drift using artificial neural networks. Analytical Chemistry 68, 271-280

Goodacre, R. Timmins, É.M., Rooney, P.J., Rowland, J.J. & Kell, D.B. 1996b Rapid identification of *Streptococcus* and *Enterococcus* species using diffuse reflectance-absorbance Fourier transform infrared spectroscopy and artificial neural networks. FEMS Microbiology Letters 140, 233-239.

Goodacre, R., Hiom, S.J., Cheeseman, S.L. Murdoch, D., Weightman, A.J. & Wade, W.G. 1996c Identification and discrimination of oral asaccharolytic Eubacterium spp. by pyrolysis mass spectrometry and artificial neural networks. Current Microbiology 32, 77-84.

Goodacre, R., Timmins, É.M., Burton, R., Kaderbhai, N., Woodward, A., Kell, D.B. & Rooney, P.J. 1998a Rapid identification of urinary tract infection bacteria using hyperspectral, whole organism fingerprinting and artificial neural networks. Microbiology 144, 1157-1170.

Goodacre, R., Shann, B., Gilbert, R. J.; Timmins, E. M., McGovern, A. C., Alsberg, B. K., Logan, N. A. and Kell, D. B. 1998b. Rapid characterization of Bacills species from PyMS and FT-IR data. Proc. ERDEC Sci. Conf. Chem. Biol. Def. Res. 257-265

Goodacre, R., Timmins, E. M., Burton, R., Kaderbhai, N., Woodward, A. M., Kell, D. B., and Rooney, P. J. 1998c. Rapid identification of urinary tract infection bacteria using

hyperspectral whole-organism fingerprinting and artificial neural networks. Microbiology (Reading, U. K.) 144(5), 1157-1170.

Goto, K., Y. Iwase, K. Kichise, K. Kitano, A. Totuka, T. Obata, and S. Hara. 1990. Isolation and properties of a chromosome-deendent KHR killer toxin in Saccharomyces cerevisiae. Agric. Biol. Chem. 54:505-509

Goto, K., H. Fukuda, K. Kichise, K. Kitano, and S. Hara. 1991. Cloning and nucleotide sequence of the KHS killer gene of *Saccharomyces cerevisiae.* Agric. Biol. Chem. 54:979-984

Gutell, R.R., N. Larsen, and C.R. Woese. 1994. Lessons from an evolving rRNA:16S and 23S rRNA structures from a comparative perspective. Microbiol. Rev. 58:10-26.

Harlow, E., and Lane D. 1988. DNA probe method for detection of specific microorganisms in the soil bacterial community. Appl. Environ. Microbiol. 54:703-711.

Harrick Scientific Corporation 1987 Optical Spectroscopy: Sampling techniques Manual.

Hedrick, D. B., Nivens, D. E., Stafford, C., White, D. C. 1991. Rapid differentiation of archaebacteria from eubacteria by diffuse reflectance Fourier-transform IR spectroscopic analysis of lipid preparations. J. Microbiol. Methods 13(1), 67-73.

Heise, H. M. and Bittner, A. 1997. Rapid and reliable spectral variable selection for statistical calibrations based on PLS-regression vector choices. Fresenius' J. Anal. Chem. 359(1), 93-99.

Helm D., H. Labischinski, G. Schallehn, and D. Naumann, 1991, Classification and identification of bacteria by FTIR spectroscopy. J. Microbiol. 173, 69-79.

Helm, D. and Naumann, D. 1995. Identification of some bacterial cell components by FT-IR spectroscopy. FEMS Microbiol. Lett. 126(1), 75-80.

Hoffmann, U. and Zanier-Szydlowski, N. 1999. Portability of near infrared spectroscopic calibrations for petrochemical parameters. J. Near Infrared Spectrosc. 7(1), 33-45.

Holman, Hoi-Ying N., Perry, D. L., Martin, M. C., and McKinney, W. R. 1998. Applications of synchrotron infrared microspectroscopy to the study of inorganic-organic interactions at the bacterial-mineral interface. Mater. Res. Soc. Symp. Proc. 524(Applications of Synchrotron Radiation Techniques to Materials Science IV), 17-23.

Holmstrom, L. and Koistinen, P. 1992. Using additive noise in back-propagation training. IEEE Transaction on Neural Networks, 3, 24-38.

Holt, C., Hirst, D., Sutherland, A. and Macdonald, F. 1995 Discrimination of species in the genus Listeria by Fourier transform infrared spectroscopy and canonical variate analysis. Appl. Env. Microbiol., 61(1):377-8

Hornik, K., Stinchcombe, M. and White, H. 1989. Multilayer feedforward networks are universal approximators. Neural Networks, 2, 359-366.

Hornik, K. 1993. Some new results on neural network approximation. Neural Networks, 6, 1069-1072.

Iglewicz, B. 1983 Robust scale estimators and confidence intervals for location, in Hoaglin, D.C., Mosteller, M. and Tukey, J.W., eds., Understanding Robust and Exploratory Data Analysis, NY: Wiley

Jiang, E. Y., Drapcho, D. L., McCarthy, W. J., and Crocombe, R. A. 1998. Frequency-resolved, phase-resolved and time-resolved step-scan Fourier transform infrared photoacoustic spectroscopy. AIP Conf. Proc. 430(Fourier Transform Spectroscopy), 381-384.

Jiang, E. Y., Wang, H. Y., and Ma, Z. W. 1999. The effect of Co addition on the saturation magnetization of Fe16N2. J. Appl. Phys. 85, 4488-4490.

Johnson S. C. 1967 Hierarchical clustering schemes. Psychometrika 32(3), 241-54.

Jonathan, P., McCarthy, W. V., and Adrian M. I. 1996. Discriminant analysis with singular covariance matrixes. A method incorporating cross-validation and efficient randomized permutation tests. J. Chemom. 10(3), 189-213.

Jordan, M. I. (1995), "Why the logistic function? A tutorial discussion on probabilities and neural networks", MIT Computational Cognitive Science Report 9503, http://www.cs.berkeley.edu/~jordan/papers/uai.ps.Z.

Kemsley, E. K. 1996. Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. Chemom. Intell. Lab. Syst. 33(1), 47-61.

Kenner, B.A. , J.R. Riddle, S.W. Rockwood and B. H. Bordner, 1958, Bacteria identification by infrared spectrophotometry, J. Bacteriol., 75, 16-20.

Kirschner C., N.A. Ngo Thi, and D. Naumann 1999 FT-IR spectroscopic investigations of antibiotic sensitive and resistant microorganisms Spectroscopy of Biological Molecules: New Directions 8th European Conference on the Spectroscopy of Biological Molecules, 2.9.8-2.9.9

Kleine, T. O., R. Hackler, and H. Meyer-Rienecker. 1990. Classical and modern methods for cerebrospinal fluid analysis. Eur. J. Clin. Chem. Clin. Biochem. 29(10):705-714

Kockova-Kratochvilova, A. 1990. Yeast and Yeast-like Organisms, NY: VCH

Koistinen, P. and Holmstrom, L. 1992. Kernel regression and backpropagation training with noise. NIPS4, 1033-1039.

Kummerle, M., Scherer, S., and Seiler, H. 1998 Rapid and reliable identification of food-borne yeasts by Fouier-transform infrared spectroscopy., Appl. Environ. Micro. 64(6): 2207-14.

Kunkee, R.E. 1991 Selection and modification of yeast and lactic acid bacteria for wine fermentation. Food Microbiol. 1:315-332

Kurtzman, C. P., Wicherham, L.J., and Hesseltine, C.W. 1970 Yeasts from wheat and flour. Mycologia 62:542-547

Kushner, H.J., and Yin, G. 1997. Stochastic Approximation Algorithms and Applications, NY: Springer-Verlag.

Lafon-Lafoureade, S., C. Geneix, and P. Ribereau-Gayon. 1984. Inhibition of alcoholic fermentation of grape must by fatty acids produced by yeasts and their elimination by yeast ghosts. Appl. Environ. Microbiol. 47:1246-1249

Lang, K.J. and Witbrock, M.J. 1988. Learning to tell two spirals apart. in Touretzky, D., Hinton, G., and Sejnowski, T., eds. Procedings of the 1988 Connectionist Models Summer School, San Mateo, CA: Morgan Kaufmann.

Lavine, Barry K., Moores, A., and Helfend, L. K. 1999. A genetic algorithm for pattern recognition analysis of pyrolysis gas chromatographic data. J. Anal. Appl. Pyrolysis 50(1), 47-62.

Lawrence, S., Giles, C.L., and Tsoi, A.C. 1996. What size neural network gives optimal generalization? Convergence properties of backpropagation. Technical Report UMIACS-TR-96-22 and CS-TR-3617, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, http://www.neci.nj.nec.com/homepages/lawrence/papers/minima-tr96/minima-tr96.html

Lawrence, S., Giles, C.L., and Tsoi, A.C. 1997. Lessons in Neural Network Training: Overfitting May be Harder than Expected. Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI-97, AAAI Press, Menlo Park, California, pp. 540-545

LeCun, Y., Simard, P.Y., and Pearlmetter, B. 1993. Automatic learning rate maximization by on-line estimation of the Hessian's eigenvectors. in Hanson, S.J., Cowan, J.D., and Giles, C.L. (eds.), Advances in Neural Information Processing Systems 5, San Mateo, CA: Morgan Kaufmann, pp. 156-163.

Li, Y.; Jiang, J. H., Chen, Z. P., Xu, C. J. and Yu, R. Q. 1999. Robust linear discriminant analysis for chemical pattern recognition. J. Chemom. 13(1), 3-13.

Li, Y. and van Espen, P. 1994. Study of the influence of neural network parameters on the performance characteristics in pattern recognition. Chemom. Intell. Lab. Syst. 25(2), 241-8.

Lipkus, A.H., Chittur, K.H., Vesper,S.J., Robinson, J.B. and Pierce,G.E. 1990. Evaluation of infrared spectroscopy as a bacterial identification mehtod., J. Ind. Microbiol., 6:71-75.

Luk, J.M.C. 1994. A PCR enzyme immunoassay for detection of Salmonella typhi. Biotechniques 17:1038-1042.

Ma, L., Sukuta, S., Bruch, R. F., Afanasyeva, N. I., and Looney, Carl G. 1998. Tumor diagnosis using backpropagation neural network method. Proc. SPIE-Int. Soc. Opt. Eng. 3257(Infrared Spectroscopy: New Tool in Medicine), 273-283.

Macario, A. J. L., and E.C. deMacario. 1990. Gene probes for bacteria. New York: Academic Press, Inc.

Maier, Raina M. 2000. Environmental microbiology San Diego, Calif. Academic Press

Manchester, L.N., Toole, A. & Goodacre, R. 1995 Characterisation of Carnobacterium species by pyrolysis mass spectrometry. Journal of Applied Bacteriology 78, 88- 96.

Manners, D. J., Masson, A. J., Patterson, J. C. 1974. Heterogeneity of glucan preparations from the walls of various yeasts. J. Gen. Microbiol. 80(2) 411-17.

Marcott, G., Story, G.M., Dowrey, A.E., Reeder, R.cG., and Noda, I. 1997. Photoacoustic depth profiling, dynamic rheo-optics, and spectroscopic imaging microscopy of polymers by step-scanning FTIR spectroscopy. Mikrochimica Acta, Suppl. 14:157-63

Martens, H., Martens, M., and Jacobsen, C. 1999. Multivariate data analysis for more effective R&D and better quality control in the laboratory. Managing Mod. Lab. 4(1), 9-17.

Maslow J. N., Mulligan M. E., and Arbeit R D 1994 Recurrent Escherichia coli bacteremia. Journal of Clinical Microbiolgy 32(3), 710-4.

Massart, D. L. and Buydens, L. Chemometrics in pharmaceutical analysis. 1988. J. Pharm. Biomed. Anal. Volume Date 1987, 6(6-8), 535-45.

Matushek, M. G. Bonten, M. J. M., and Hayden, Mary K. 1996 Rapid preparation of bacterial DNA for pulsed-field gel electrophoresis. J. Clin. Microbiol. 34(10), 2598-2600.

McNaughton, D., Romeo, M., Kanzis, M., Wood, B., Heraud, P., Burden, F., and Beardall, J. 1999. Infrared spectroscopy and multivariate statistics applied to medical and biological problems. Spectrosc. Biol. Mol.: New Dir., Eur. Conf., 8th 475-478.

Mendelsohn, R., Hassankhani, A., Dicarlo, E., Boskey, A. 1989 FTIR microscopy of endochondral ossification at 20μm spatial resolution. Calcified Tissue International. 44:20-4.

Merz, W.G., C. Connelly, and P. Hieter. 1988. Variation of electrophoretic karyotypes among clinical isolates of Candida albicans. J. Clin. Microbiol. 26:842-845

Meyer, M.; Meyer, K., and Hobert, H. 1993. Neural networks for interpretation of infrared spectra using extremely reduced spectral data. Anal. Chim. Acta 282(2), 407-15.

Miichael, K., Siegfried.S., and Herbert S. 1998. Rapid and Reliable Identification of Food-Borne Yeasts by Fourier-Transform Infrared Spectroscopy, Appl. Envir. Micro. 64 (6):2207-2214

Molina, F. I., Inone, T., and Jong, S.C. 1991. Ribosomal DNA restriction analysis reveals genetic heterogeneity in Saccharomyces cerevisiae Meyen ex Hansen. Int. J. Syst. Bacteriol. 42:499-502

Moody, J.E. 1992. The Effective Number of Parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Systems. in Moody, J.E., Hanson, S.J., and Lippmann, R.P., Advances in Neural Information Processing Systems 4, 847-854.

Morace, G., G. Dettori, M. Sanguinetti, S. Manzara, and L. Polonelli. 1988. Biotyping of aerobic actinomycetes by modified killer system. Eur. J. Epidemiol. 4:99-103

Morace, G., S. Manzara, G. Dettori, F. Fanti, S.Conti, L.Campani, L. Polonelli, and C. Chezzi. 1989 Biotyping of bacterial isolates using the yeast killer system. Eur. J. Epidemiol. 5:303-310

Naumann, D., Barnickel, G., Bradaczek, H., Labischinski, H., and Giesbrecht, P. 1982 Infrared spectroscopy, a tool for probing bacterial peptidoglycan. Potentialities of infrared spectrosocopy for cell wall analytical studies and rejection of models based on crystalline chitin. Eur. J Biochem. 125:50515.

Naumann, D., Labischinski, H., Ronspeck, W., Barnickel, G., BradacZek, H. 1987a Vibrational spectroscopic analysis of LD-sequential, bacterial cell wall peptides: An IR and Raman study. Biopolymers 26:795-817.

Naumann, D., Schultz, C., Born, J., Labischinski, H., Brandenburg, K., Busse, G.V., Brade, H., and Seydel, U. 1987b Investigations into the polymorphism of lipid A from lipopolysaccharides of Escherichia coli and Salmonella minnesota by Fourier transform infrared spectroscopy. Eur. J Biochem. 164:159-69.

Naumann, D., Fijala, V., Labischinski, H., Giebrecht, P. 1988. The rapid differentiation and identification of pathogenic bacteria using Fourier transform infrared spectroscopic and multivariate statistical analysis., J.MoL Struct., 174:165-70.

Naumann, D., Helm, D. and Labischinski, H. 1991a. Microbiological characterizations by FT-IR spectroscopy., Nature, 351:81-2.

Naumann, D., Helm, D., Labischinski, H., Giebrecht, P. 1991b. The characterization of microorganisms by Fourier transform infrared spectroscopy (FTIR)., In Modern techniques for rapid microbiological analysis, Ed Nelson, New-York, p 43-96.

Naumann D. 1998a. FT-IR and FT-NIR Raman spectroscopy in biomedical research In: J.A. de Haseth (ed.): "Fourier Transform Spectroscopy: 11th International Conference" AIP Conference Proceedings 430, 96-109, Woodbury, New York

Naumann, D, 1998b. Infrared and NIR Raman spectrocopy in medical microbiology.In: Infrared Spectroscopy: New tool in medicine, SPIE, vol 3257: 245-57

Nes, W. R., Sekula, B. C., Nes, W. D. and Adler, J. H. 1978 J. Biol. Chem. 253, 6218-6225

Olsen, J.E., Aabo, S., Rasmussen, O.F. and Rossen, L. 1995 Oligonucleotide probes specific for the genus Salmonella and for Salmonella typhimurium. Lett. Appl. Microbiol. 20(3), 160-3.

Pedersen, M. B. 1986 DNA sequence polymorphisms in the genus Saccharomyces. III. Restriction endonuclease fragment patterns of chromosomal regions in brewing and other yeast strains. Carlsberg Res. Commun. 48:485-503

Pinder, A.C., P. W. Purdy, S.A. Poulter, and D.C. Clark. 1990. Validation of flow cytometry for rapid enumeration of bacterial concentrations in pure cultures. J. Appl. Bacterial. 69(1):92-100

Polacheck, I., Melamed, M., Bercovier, H., and Salkin, I. F. 1987. β-Glucosidase in Candida albicans and its application in yeast identification. J. Clin. Microbiol. 25(5), 907-10.

Pretorius, I.S. and van der Westhuizen, T.J. 1991 The impact of yeast genetics and recombinant DNA technology on the wine industry: a review. S. Afr. Enol. Vitic. 12:3-31

Price, C. W., Fuson, G. B., and Phaff, H. J. 1978 Genome comparison in yeast systematics: delimitation of speciecs within the genera *Schwanniomyces*, *Saccharomyces*, *Debaryomyces* and *Pichia*. Microbiol. Rev. 42: 161-193.

Qiu, J., Pan, H., Han, C., Ye, Q. and Zhang, S. 1999 Monitoring glucoamylase fermentation with infrared spectroscopy. Guangpuxue Yu Guangpu Fenxi 19(6), 831-833.

Radomski, J. P., van Halbeek, H., and Meyer, B. 1994. Neural network-based recognition of oligosaccharide 1H-NMR spectra. Nat. Struct. Biol. 1(4), 217-18.

Reed, R.D., and Marks, R.J, II 1999. Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks, Cambridge, MA: The MIT Press, ISBN 0-262-18190-8.

Ribereau-Gayon, P., S. Lafon-Lafourcade, and A. Bertrand. 1975. Vigne Vin 9:117-139

Riedmiller, M. and Braun, H. 1993 A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm, Proceedings of the IEEE International Conference on Neural Networks 1993, San Francisco: IEEE

Ripley, B.D. 1996 Pattern Recognition and Neural Networks, Cambridge: Cambridge University Press

Rosini, G. 1983. The occurrence of killer character in yeasts. Can. J. Microbiol. 29:1462-1464

Sarle, W.S. 1995. Stopped Training and Other Remedies for Overfitting. Proceedings of the 27th Symposium on the Interface of Computing Science and Statistics, 352-360

Savitzky, A., Golay, M. J. E. 1964 Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627-1639

Schalkoff, R.J. 1997 Artificial neural networks. McGraw-Hill, New-York, 422p.

Scherer, S. and Stevens, D.A. 1987 Application of DNA typing methods to epidemiology and taxonomy of Candida species. J. Clin. Microbiol. 25:675-679

Schmitt, J., Udelhoven, T., Naumann, D. and Flemming, H.C. 1998 Stacked spectral data processing and artificial neural networks applied to FTIR and FT-Raman spectra in biomedical applications. IN: Infrared spectrocopy: New Tool in Medicine, SPIE, vol. 3257: 236-44.

Schultz, C.and Naumann, D. 1989 In vivo characterization of the membranes of Gram-negative bacteria by FT-IR. Proceedings of 7th European Conference on the Spectroscopy of Biological Molecules (Italy).

Schultz, C. & Naumann, D. 1991 In vivo study of the state of order of the membranes of Gram-negative bacteria by Fourier transform infrared specterocopy (FT-IR). FEBS 294:43-6.

Schultz, C. P., Liu, K.Z., Jonston, J.B., & Mantsch, H.H. 1996. Study of chronic lymphocytic leukemia cells by FTIR spectroscopy and cluster analysis. Leukemia Research, 20(8) :649-55

Scott, D.W. 1992 Multivariate Density Estimation, Wiley.

Shaw, A. D., Winson, M. K., Woodward, A. M., McGovern, A. C., Davey, H. M., Kaderbhai, N., Broadhurst, D., Gilbert, R. J., Taylor, J., Timmins, E. M., Goodacre, R. and Kell, D. B. 2000. Rapid analysis of high-dimensional bioprocesses using multivariate spectroscopies and advanced chemometrics. Adv. Biochem. Eng./Biotechnol. 66(Bioanalysis and Biosensors for Bioprocess Monitoring), 83-113.

Smit, J. R. M., Melssen, W. J., Rolf, G. H. and Kateman, G. 1993.Two-dimensional mapping of IR spectra using a parallel implemented self-organizing feature map. Chemom. Intell. Lab. Syst. 18(2), 195-20

Smith, M. 1996. Neural Networks for Statistical Modeling, Boston: International Thomson Computer Press, ISBN 1-850-32842-0.

Sockalingum, G.D., Bouhedja, W., Pina, P., Allouch, P., Bloy, C. and Manfait, M. 1998 FTIR spectroscopy as an emerging method for rapid characterization of microorganisms. Cell. Mol. Biol., 44(1):261-9

Sontag, E.D. 1992. Feedback stabilization using two-hidden-layer nets. IEEE Transactions on Neural Networks, 3, 981-990.

Spiegelman, C. H., McShane, M. J., Cote, G. L., Goetz, Marcel J., Motamedi, M., and Yue, Q. L. 1998. Theoretical Justification of Wavelength Selection in PLS Calibration: Development of a New Algorithm. Anal. Chem. 70(1), 35-44.

Stager, C. E., and J. R. Davis. 1992. Automated systems for identification of microorganisms. Clin. Microbiol. Rev. 5:302-327

St.-Germain, G. and Beauchesne, D. 1991 Evaluation of the MicroScan rapid yeast identification panel. J. Clin. Microbiol. 29:2296-2299

Stork, Chris L. and Kowalski, B. R., 1999. Weighting schemes for updating regression models - a theoretical approach. Chemom. Intell. Lab. Syst. 48(2), 151-166.

Suci, P. A., Vrany, J. D. and Mittelman, M. W. 1998 Investigation of interactions between antimicrobial agents and bacterial biofilms using attenuated total reflection Fourier transform infrared spectroscopy. Biomaterials 19(4-5), 327-339.

Swingler, K. 1996. Applying Neural Networks: A Practical Guide, London: Academic Press.

Tetko, I.V., Livingstone, D.J., and Luik, A.I. 1995. Neural Network Studies. 1. Comparison of Overfitting and Overtraining. J. Chem. Info. Comp. Sci., 35, 826-833.

Thomas, L. C. and Greenstreet, J.E.S. 1954 The identification of microorganisms by infrared spectrophotometry., Spectrochim. Acta, 6:302-19

TiBor D. et al., 1996 "Handbook of Food Spoilage Yeasts", CRC Press,

Timmins, É.M., Howell, S.A., Alsberg, B.K., Noble, W.C. & Goodacre, R. 1998a Rapid differentiation of closely related *Candida* species and strains by pyrolysis mass spectrometry and Fourier transform infrared spectroscopy. Journal of Clinical Microbiology 36, 367-374.

Timmins, É.M., Quain, D.E. & Goodacre, R. 1998b Differentiation of brewing yeast strains by pyrolysis mass spectrometry and Fourier transform infrared spectroscopy. Yeast 14, 885-893.

Timmins, E. M., Howell, S. A., Alsberg, B. K., Noble, W. C. and Goodacre, R. 1998c. Rapid differentiation of closely related Candida species and strains by pyrolysis-mass spectrometry and Fourier transform-infrared spectroscopy. J. Clin. Microbiol. 36(2), 367-374.

Tipper, D. J. , and M. J. Schmitt. 1991. Yeast dsRNA viruses: replication and killer phenotypes. Mol. Microbiol. 5:2331-2338

Torok, T., Royer,C., Rockhold,D., and King, A.D. 1992 Electrophoretic karyotyping of yeasts, and Southern blotting using whole chromosomes as templates for the probe preparation. J. Gen. Appl. Microbiol. 38:313-325

Torok, T., Rockhold, D., and King, A.D. 1993 Use of electrophoretic karyotyping and DNA-DNA hybridization in yeast identification. Int. J. Food Microbiol. 19:63-80

Torriani, S., Zapparoli, G. and Dellaglio, F. 1999 Use of PCR-based methods for rapid differentiation of Lactobacillus delbrueckii subsp. bulgaricus and L. delbrueckii subsp. lactis. Appl. Environ. Microbiol. 65(10), 4351-4356.

Udelhoven, T., Naumann, D., Schmitt, J. 2000.Development of a hierarchical classification system with artificial neural networks and FT-IR spectra for the identification of bacteria. Appl. Spectrosc. 54(10), 1471-1479.

Urban, M. W. 1998. Multi-dimensional surface and interfacial analysis of polymers and coatings: ATR, step-scan photoacoustic, FT-IR/FT-Raman imaging. Polym. Mater. Sci. Eng. 78 18-19.

Van der Mei, H.C., Naumann, D. and Busscher, H.J. 1993. Grouping of oral streptococcal species using Fourier transform infrared spectroscopy in comparison with classical microbiological identification., Mol. Biol., vol. 38(11):1013-9

Van Solingen, P. and Van der Plaat, J. B. 1975. Partial purification of the protein system controlling the breakdown of trehalose in baker's yeast. Biochem. Biophys. Res. Commun. 62(3), 553-60.

Van V., and B.S.Wingfield. 1986. killer yeast –cause of stuck fermentations in a wine cellar. S. Afr. J. Enol. Vitic. 7:102-107.

Vaughan-Martini, A. and Martini, A. 1993 A taxonomic key for the genus Saccharomyces. Syst. Appl. Microbiol. 16:113-119

Vezinhet,F., Blondin, B., and Hallet, J. N. 1990. Chromosomal DNA banding patterns and mitochondrial DNA polymorphism as tools for identification of enological strains of Saccharomyces cerevisiae. Appl. Microbiol. Biotechnol. 32:568-571

Vezinhet, F., Hallet, J. N., Valade, M., and Poulard, A. 1992. Ecological survey of wine yeast strains by molecular methods of identification. Am. J. Enol. Vitic. 43:83-86

Vinod, H.D. and Ullah, A. 1981. Recent Advances in Regression Methods, NY: Marcel-Dekker.

Wayne P. Olson 1999. Automated Microbial Identification and Quantitation – Technologies for the 2000s, Interpharm Press, Inc. Buffalo Grove, IL

Weigend, A. 1994. On overfitting and the effective number of hidden units. Proceedings of the 1993 Connectionist Models Summer School, 335-342.

Wentzell, P. D., Andrews, D. T., Kowalski, and Bruce. R. 1997. Maximum Likelihood Multivariate Calibration. Anal. Chem. 69(13), 2299-2311.

Wickner, R.B. 1992. Double-stranded and single-stranded RNA viruses of *Saccharomyces cerevisiae*. Annu.Rev. Microbiol. 46:347-375

Wickner, R.B. 1996. Double-stranded RNA viruses of *Saccharomyces cerevisiae*. Microbiol. Rev. 60:250-265

Wold, H. 1966. Estimation of principal components and related models by iterative least squares, in Multivariate Analysis,P.R. Krishnaiah, ed. Academic Press, New York, 391-420

Yamamoto, N., Amemiya, H., Yokomori, Y., Shimizu, K., and Totsuka, A. 1991. Electrophoretic karyotypes of wine yeasts Am. J. Enol. Vitic. 42:358-363

Zhu, E. and Barnes, R. M. 1995. A simple iteration algorithm for PLS regression. J. Chemom. 9(5), 363-72.

# Appendix 1 Reference number of yeast strains

| *Saccharomyces cerevisiae* | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Wine** | **Beer** | **Bread** | **Distillers** | **Probiotics** | **Animal nutrition** | **Other** |
| **Sensitive** | 6050 (3)<br>6467 (30)<br>6400 (26)<br>6287 (20)<br>6649 (42)<br>6648 (41)<br>6348 (49) | 6061 (8)<br>6562 (36)<br>6058 (5)<br>6060 (7)<br>6163 (14)<br>6412 (27)<br>6422 (29)<br>6101 (10)<br>6059 (6)<br>6301 (22) | | | 6100 (44)<br>6652 (55) | | |
| **Possess** | 6032 (2)<br>6014 (1) | | | | | | |
| **Neutral** | | | | | | | |
| *Saccharomyces chevalieri* | | | | | | | |
| | **Wine** | **Beer** | **Bread** | **Distillers** | **Probiotics** | **Animal nutrition** | **Other** |
| **Sensitive** | 6254 (48) | | | | | | 6196 (46) (Chocolate)<br>6165 (15) (Lab strain) |
| **Possess** | 6581 (37)<br>6139 (13) | | | | | | |
| **Neutral** | 6542(54)<br>6560(35) | | | | | | |
| *Saccharomyces capensis* | | | | | | | |
| | **Wine** | **Beer** | **Bread** | **Distillers** | **Probiotics** | **Animal nutrition** | **Other** |
| **Sensitive** | 6276 (18)<br>6290 (21) | | | | | | |
| **Possess** | | | | | | | |
| **Neutral** | 6221 (16) | | | | | | 6375 (25) (Unknown) |

*Saccharomyces italicus*

| | Wine | Beer | Bread | Distillers | Probiotics | Animal nutrition | Other |
|---|---|---|---|---|---|---|---|
| **Sensitive** | 6071 (9)<br>6074 (43) | | | | | | |
| **Possess** | 6302 (23) | | | | | | |
| **Neutral** | | | | | | | |

*Saccharomyces diastaticus*

| | Wine | Beer | Bread | Distillers | Probiotics | Animal nutrition | Other |
|---|---|---|---|---|---|---|---|
| **Sensitive** | | | | | | | |
| **Possess** | | | | | | | |
| **Neutral** | 6123 (12) | 6637 (40) | | | | | |

*Saccharomyces delbrueckii*

| | Wine | Beer | Bread | Distillers | Probiotics | Animal nutrition | Other |
|---|---|---|---|---|---|---|---|
| **Sensitive** | | 6116 (11)<br>6056<br>6493(31) | | | | | |
| **Possess** | | | | | | | |
| **Neutral** | | 6147 (45) | | | | | |

*Saccharomyces bayanus*

| | Wine | Beer | Bread | Distillers | Probiotics | Animal nutrition | Other |
|---|---|---|---|---|---|---|---|
| **Sensitive** | | | | | 6653 (56) | | |
| **Possess** | | | | | | | |
| **Neutral** | | | | | | | 6352 (Unknown) |

*Saccharomyces rosei*

| | Wine | Beer | Bread | Distillers | Probiotics | Animal nutrition | Other |
|---|---|---|---|---|---|---|---|
| **Sensitive** | | | | | | | |
| **Possess** | | | | | | | |
| **Neutral** | | | | | | 6242 (47) | |

*Schizosaccharomyces pombe*

| | Wine | Beer | Bread | Distillers | Probiotics | Animal nutrition | Other |
|---|---|---|---|---|---|---|---|
| **Sensitive** | 6262 | | | | | | |
| **Possess** | | | | | | | |
| **Neutral** | 6265 (17) | | | 6514<br>6515 | | | |

*Saccharomyces cerevisiae / Schizosaccharomyces pombe*

| | Wine | Beer | Bread | Distillers | Probiotics | Animal nutrition | Other |
|---|---|---|---|---|---|---|---|
| **Sensitive** | | | | | | | |
| **Possess** | | | | | | 6527<br>6528 (32) | |
| **Neutral** | 6578 | | | | | 6529 (33) | |

*Kluyveromyces marxianus*

| | Wine | Beer | Bread | Distillers | Probiotics | Animal nutrition | Other |
|---|---|---|---|---|---|---|---|
| **Sensitive** | | | | | | | |
| **Possess** | | | | | | | |
| **Neutral** | | | | | | | 6425 (52)<br>(Lactoserum)<br>6349 (50)<br>(Lactose) |

*Hansenula valbyensis*

| | Wine | Beer | Bread | Distillers | Probiotics | Animal nutrition | Other |
|---|---|---|---|---|---|---|---|
| **Sensitive** | | | | | | | |
| **Possess** | | | | | | | |
| **Neutral** | 6533 (34) | | | | | | |

*Candida utilis*

| | Wine | Beer | Bread | Distillers | Probiotics | Animal nutrition | Other |
|---|---|---|---|---|---|---|---|
| **Sensitive** | | | | | | | |
| **Possess** | | | | | | | |
| **Neutral** | 6283 (19) | | | | | | 6504 (53)<br>(Wood fermentation) |

| *Zygosaccharomyces cidrii* | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Wine** | **Beer** | **Bread** | **Distillers** | **Probiotics** | **Animal nutrition** | **Other** |
| **Sensitive** | | | | | | | |
| **Possess** | | | | | | | |
| **Neutral** | | | | | | | 6414 (28) (Bioingredients) |
| Unknown | | | | | | | |
| | **Wine** | **Beer** | **Bread** | **Distillers** | **Probiotics** | **Animal nutrition** | **Other** |
| **Sensitive** | | | 6583 (38)<br>6584 (39) | | | | |
| **Possess** | | | | | | | |
| **Neutral** | | | 6370 (24) | | | | |

Note : The four digits reference number is the reference number from Lallemand Inc.
The number in the bracket is the corresponding reference number we used