Chapter 1

Introduction

1.1. Background and scope of research

In recent times, the commercial use of fermentation has been rapidly growing, mainly due to the need for biopharmaceutical products, new biologics, cost effectiveness and improved microbial processes in food industry, and for alternative fuels such as ethanol, which could help to reduce carbon dioxide emissions and decrease oil imports. Ethanol can be produced from cellulosic biomass after its hydrolysis, which is followed by sugar fermentation processes. The biomass wastes contain a complex mixture of carbohydrate polymers from the plant cell walls known as cellulose, hemicellulose and lignin. In order to produce sugars from the biomass, the biomass is pre-treated with acids or enzymes to reduce the size of the feedstock and to open up the plant structure. The cellulose and the hemicellulose portions are hydrolysed by enzymes or diluted acids into cellulose and then fermented into ethanol. Instead of using acid to hydrolyze the biomass into cellulose, enzymes can be hydrolysed lignocelluloses to break down the biomass. The production of this enzyme called cellulase requires a microorganism that can secrete it for the hydrolysis of cellulosic substrates. The cellulase contains three major highly specific enzymes namely the endoglucanases (EC 3.2.1.4), the exoglucanases (EC 3.2.1.91) and *β*-glucosidases (EC 3.2.1.21). These enzymes are non constitutive and are produced by many microorganisms such as bacteria, actinomycete and fungi. The cellulase-system of fungal origin are the most abundant and widely studied. *Pycnoporus sanguineus* is one of the white-rot fungi which there are no report available for cellulase production. It

may presents an efficient organism in terms of cellulase production in submerged fermentation. In this study, the effect of selected media composition was examined for cellulase enzyme activity expressed by this fungus.

Subsequently, viscosity measurement of carboxymethylcellulose (CMC) substrate in the presence of the cellulase enzyme system is explored as a rapid method for indirect endoglucanases activity determination, which is one of the three enzymes in cellulase complex. Viscosity measurement of different construction can be used for determination of endoglucanases activity. There were some reports of viscometric method for endoglucanases activity from Skomarovsky in 2000. In this work, we further developed a mathematical method for the determination of endo-glucanases activity as a function of CMC viscosities.

1.1. Objectives of thesis

1. To determine the ability of *P. sanguineus* for cellulase production;

2. To study the effect of selected medium composition on the cellulase enzyme productivity by *P. sanguineus* and compare the cost effectiveness of the media formulation;

3. To study the relationship between endoglucanases activities and the reduction in CMC containing media's viscosities;

4. To propose a unique mathematical function relating endogluconases activities to media viscosities in each medium composition using non linear regression method.

Through the model, the relationship between viscosity and cellulase production can be express quantitatively.

1.2. Outline of the thesis

This thesis includes five chapters, which are presented as follows:

- The first chapter contains the introduction, objectives and thesis outline;

- The second chapter presents the literature review carried out on cellulose production *via* submerged liquid fermentation by fungal species and general introduction to regression for mathematical modelling;

- The third chapter presents the materials and methods for enzyme production in shake flasks and viscosity measurements;

- The fourth chapter discusses enzyme production from different selected media. It also presents a comparison between two types of mathematical models that relate the medium viscosity and cellulase activities;

- Finally, the fifth chapter outlines the general conclusion of this research and some recommendations for future work.

Chapter 2

Literature review

This chapter reviews literature relevant to the present research topic: cellulose structure, fungal cellulase, cellulose and its degradation to glucose, as well as a review of the mathematical modelling and curve fitting for biological data.

2.1. Biological part

2.1.1. White-rot Fungi

To understand the ability of fungi to degrade contaminants it is important to analyze the ecological niches of white-rot fungi. White-rot fungi belong to the wood-degrading *basidiomycetes* and are best known as the only micro-organisms responsible for the mineralization of all major wood polymers, including lignin, cellulose and hemicellulose (Crawford & Crawford 1980). The mechanisms for their wood biodegradation ability are dependent on the fungal species and conditions. The term white rot has been used to describe forms of wood decay in which the wood assumes a bleached appearance and where lignin, cellulose and hemicellulose are broken down (Sutherland and Crawford, 1981).

In nature, cellulolytic fungi inhabit the soil in the decaying substance of plant material and in the root, bark and surface of the wood of living plants where they play a major role in the conversion of the biomass of plant origin into soil.

Cellulose, the basic substrate of cellulase, is a complex polymer of glucose. Fungi, with the help of cellulase, degrade this glucose polymer into glucose, which not

only support the growth of the fungus but also supports the growth of other microbial populations in nature.

The cost of utilization of enzymatic hydrolysis is low compared to acid or alkaline hydrolysis because enzyme hydrolysis is usually conducted at mild conditions (pH 4.8 and temperature 45°C to 50°C) and does not present corrosion problems (Duff and Murray, 1996). Both bacteria and fungi can produce cellulases for the hydrolysis of lignocellulosic materials. These microorganisms can be aerobic or anaerobic, mesophilic or thermophilic.

Fungi that have been reported to produce cellulases include *Sclerotium rolfsii*, *Phanerochete chrysosporium* and species of *Trichoderma, Aspergillus*, *Schizophyllum* and *Penicillium pinophilum* (Sternberg, 1976; Fan *et al*., 1987; Duff and Murray, 1996). Of all these fungal genera, *Trichoderma reesei* has been most extensively studied for cellulase production (Sternberg, 1976).

In addition, white-rot fungi play a significant role in the recycling of lignin, capable of completely mineralizing lignin to CO_2 and H_2O due to their non-specific extracellular ligninolytic enzyme system. Degradation of the complex irregular aromatic structure of lignin polymer by white-rot fungi enables them to access cellulose and hemicellulose, which they then utilise as carbon and energy sources (Leatham 1986, Kirk & Farrell 1987).

Fungi play such a key role in human society that it could be readily argued that they are the most important biotechnologically useful organisms (Kurtzman, 1983). Fungi produce a vast array of enzymes such as amylase, amyloglucosidase, pullulanase, cellulase, β-glucosidase, invertase, lipase, laccase, catalase etc (Lambert 1983).

Fungal growth within the plant material is mainly responsible for the ultimate disintegration of the plant, producing huge masses of residue contributing to the major bulk of the soil. The fungi are classified based upon the characteristics of degradation of plant components e.g. cellulose, hemicelluloses and lignin (Chen and Chang, 1985). Cellulase can be produced under a wide variety of growth conditions of fungi. Although a variety of organisms have been tested for cellulase production, different species of *Trichoderma* are the most widely used organism for this purpose. Ahamed *et al.* (2008) suggested that species of *Trichoderma* are, by far, the best cellulase producers. Limited species of *Penicilium* and *Aspergillus* also produce significant amounts of cellulase.

Significant research efforts for the past three decades are focused on improving the economical production of cellulase. The high cost of enzyme production limits the widespread industrial use of the enzyme in the production of soluble sugars (Spano *et al*., 1978).

Table 2.1 shows that cellulase can be produced under a wide variety of growth conditions of fungi. Although a variety of organisms have been tested for cellulase production, different species of *Trichaderma* are the most widely used organism for this purpose.

Table 2.1. Distribution of cellulase in fungi.

Organisms	Growth conditions	Enzymes	References
Tichoderma reesei QM	A. Cellulose B. Cellooligosaccharide C. Sophorose	A. EG — Conidia B. EG & BG— Mycelia C. EG & BG— Mycelia, conidia& Protoplast	Kubicek *et al*. (1993)
T. reesei	A. Sophorose B. Cellulose C. Lactose	A. EG - 2 forms B. EG - 5 forms C. EG - 4 forms	Kubicek *et al*. (1993)
T. reesei 6a	Sophorose	α- EG	Sternberg and Mandels (1979)
T. reesei	Sophorose	Cellulose complex, repression of β-glucosidase	Sternberg and Mandels (1980)
T. reesei QM9414	Continuous culture with cells recycle; slow addition of cellulose at high concentration, 0.5% glucose	High production of cellulase	Ghose and Sahai (1979)
T. reesei	ln stirred—tank fermenter on high cellulose (8 %) and using ammonium hydroxide as pH controller and nitrogen source	Cellulase	Sternberg Dorval (1979)
T. reesei Rui-C30	Avicel	EG BG BG FPase	Glenn *et al*. (1985)
T. reesei		EGIII	Saloheimo *et al*. (1988)
T. reesei M9414		EG CBH	Bhikhabhai 1984
T. reesei	β-glucan	CBH	Kubicek *et al*. (1993)
T. reesei		BG	Inglin *et al*. (1980)
T. reesei	Culture filtrate	EG (38 kDa); EGIII (50 kDa)	Saloheimo *et al*. (1988)
T. reesei		CBHI	Bhikhabhai and Pettersson (1984)
T. reesei			Fägerstam and

(Continued)

QM9414			Pettersson (1979)
T. reesei		EG, CBH, Avicelase	Reese and Mandels (1980)
T. reesei RUT-C30	Selected media	BG	Ahamed and Vermette (2008)
T. reesci QM 6a mutant		5 EG (20—36 %) 2 CBH (64430%), BG	Bissett (1979)
T. reesei RUT-C30	Selected media, inhibitor	BG	Ahamed and Vermette (2009)
T. reesei NG-MQM9414	Agar plate, screening media for mutant selection	Enhanced EG, BG, FPase	Montenecourt and Eveleigh (1977)
T. reesei		EG CBI	Penttilä *et al.* (1986)
T. reesei	Polyglycosidic substrate	Immunologically distinct enzymes; EG (43 kDa & 56-67 kDa)	Nummi *et al.* (1983)
T. viride 44	High nitrogen-nitrate, ammonium of organic form	Cellulase	Ostrikova and Konovalov (1983)
T. viride	Cellulose medium with pH controller	EG CBH BG	Sternberg (1976)
T. viride ITCC1433	Submerged culture with glucose and alkali treated cellulose powder	EG BG Avicelasc	Herr (1979)
T .viride QM9414	Grown in submerged fermentation with cellulose media	FPase (1.6 units/mL)	Sternberg (1976)
T. viride QM9414	Liquid medium	EG	Håkansson (1979)
T .viride QM9414	Culture filtrate	EG	Håkansson (1979)
T. viride	Commercial cellulase	EG CBH BG	Gong *et al*. (1977)
T. viride	Cellulase powder	BG	
T. viride	Commercial culture filtrate	CBH (48.4 kDa)	Gum and Brown (1976)

(Continued)

T. viride	Cellulase powder	EG 45 kDa CBH	Maguire (1977)
T. viride	Commercial cellulase from culture filtrate	EG	Berghem *et al*. (1976)
T. viride	Commercial cellulase	CBH	Okada *et al*. (1990)
T. viride	Commercial crude cellulases	EG Cellulase II: A 30 kDa B 43 kDa	Takahashi *et al*. (2002)
T. viride	CMC	EGI, II, III, IV isozymes	Okada *et al*. (1990)
T. viride	CMC	EGII, III, IV	Shoemaker and Brown (1978)
T. viride	Commercial cellulose	Cellulase acting on insoluble and amorphous cellulose	Huang (1975)
T. viride	CMC- yeast extract	EG	Gupta and Gupta (1979)
Aspergillus foetidus	CMC	EG CBH BG	Gusakov *et al*. (1984)
T. viride		Cellulase	Gong *et al*. (1979)
T. viride, *Basidiomycetes*	Shake flasks	EG, FPase cotton activity, high BG	Tan and Wahab (1997)
T. viride QM9414-4 mutant with UV	microcrystalline cellulose	BG	Farkas *et al*. (1981)
T. viride QM9123	Commercial glucose in submerged fermentation	Cellulase	Brown and Zainudeen (1977)
T. koningii		EG (produce short fiber from cel-luluse)	Halliwell and Vincen (1981)
Trichosporon longibrachiatum 7-26	Soil fungi, microscopic fungi produce more cellulase than yeast	Cellulase	Gracheva *et al*. (1978)
Soil fungi 19 selected-K1 and MNNG mutant KW7	Rose Bengal cellulase agar medium	EG	Vohra *et al*, (1980)
Trichosporon longibrachiatum	Cellulose powder in shake flasks	EG K1; 12,1 lU/mL; KW7; 3.7 IU/mL	Gracheva *et al*. (1978)
Trichosporon	Stationary culture better	Cellulase	Tan and Wahab

(Continued)

longibrachiatum	than submerged; CMC, (enzyme in culture filtrate)		(1997)
Myrothecium verrucaria	Noncellulosic substrate (with phenylethyl alcohol, a growth inhibitor)	High cellulase	Hurst *et al*. (1978)
Gliocladium sp.		Cellulase	Hurst *et al*. (1978)
A. aculatus	CMC & filter paper medium at pH 7.0 25 to 30°C	EG FPase	Fetzner *et al*. (2004)
Sparatrichum (*thermophile*)	Mineral salt medium, yeast extract, sodium nitrate or urea nitrogen, 45 °C incubation	EG	Coutts and Smith (1976)
Sporotrichum pulverulentum	Powder cellulose	EG (28 – 37 Da) CBH	Eriksson and Pettersson (1975)
A. terreus 17P	Wheat straw with possible stimulant e.g., Tween 80, oleic acid, vitamin C, acetic acid, etc.	Cellulase	Ismailova (1975)
A. zerreus 17P	Nutrient medium with 45% straw, 45 % wheat bran, 10% malt shoot	Cellulase (high)	Huang (1975)
A. niger	Commercial preparation of BG and EG	EG, BG (purified)	Hurst *et al*. (1977)
A. niger	Commercial cellulase	EG	Fetzner *et al*. (2004)
A. awamori		CBH, BG	Takahashi *et al*. (2002)

EG: Endoglucanase; BG, β-glycosidase; FPase, Filter paper hydrolysis activity; Avicelase, Avicel hydrolysis activity; CBH, Cellobiohydrolase; CMCase, Carboxymethyl cellulase; mol. wt., molecular weight.

2.1.2. Cellulose structure

Cellulose is a linear polymer of glucose units, which can be hydrolyzed by the action of glucosidases, cellobiohydrolases and endogluconases. Various substrates are able to induce secreted enzymes suited to degrade very precisely particular combination of polysaccharides and chemical bonds found in the carbon source. The high molecular weight substrates are not able to enter the cell and therefore, cannot themselves, generate such a precise response. It has been suggested that the regulatory systems respond to characteristic low molecular weight molecules liberated from a given

substrate through the action of small amount of constitutively secreted enzymes (Ilmen *et al*., 1996).

Cellulose is the major polymeric component of plant material and is the most abundant polysaccharide on Earth (Bayer *et al*., 1998). Cellulose consists of unbranched glucan polymer of linear chains of *β*-(1-4) linked D-glucose residues where every other glucose residue is rotated approximately 180 degrees (Fig. 2.1). Cellobiose (a two glucose unit linked *β*-(1-4) bond) is the basic unit of cellulose, rather than glucose, as cellobiose is the major product from cellulose hydrolysis.

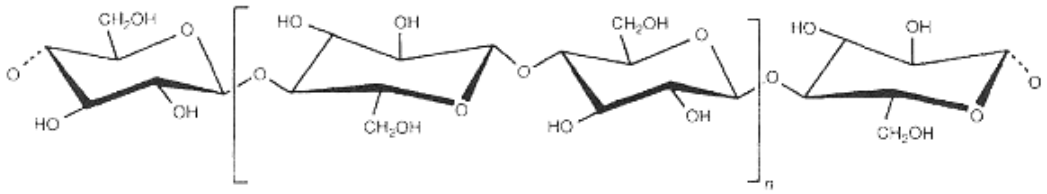

Figure 2.1. Diagram of the structural formula for the *β*-1,4-glucan polymer chain (cellulose). The repeating unit, cellobiose is indicated in brackets (Brown *et al*., 1996).

Native cellulose exists in the form of microfibrils (Fig. 2.2a), which are paracrystalline assemblies of several dozen (1→4) *β*-D-glucan chains with hydrogen bonds connected to one another (Carpita and McCann, 2000). The cellulose microfibrils are embedded in a matrix of noncellulosic polysaccharides, mainly hemicellulose and pectic substances (Fig. 2.2b).

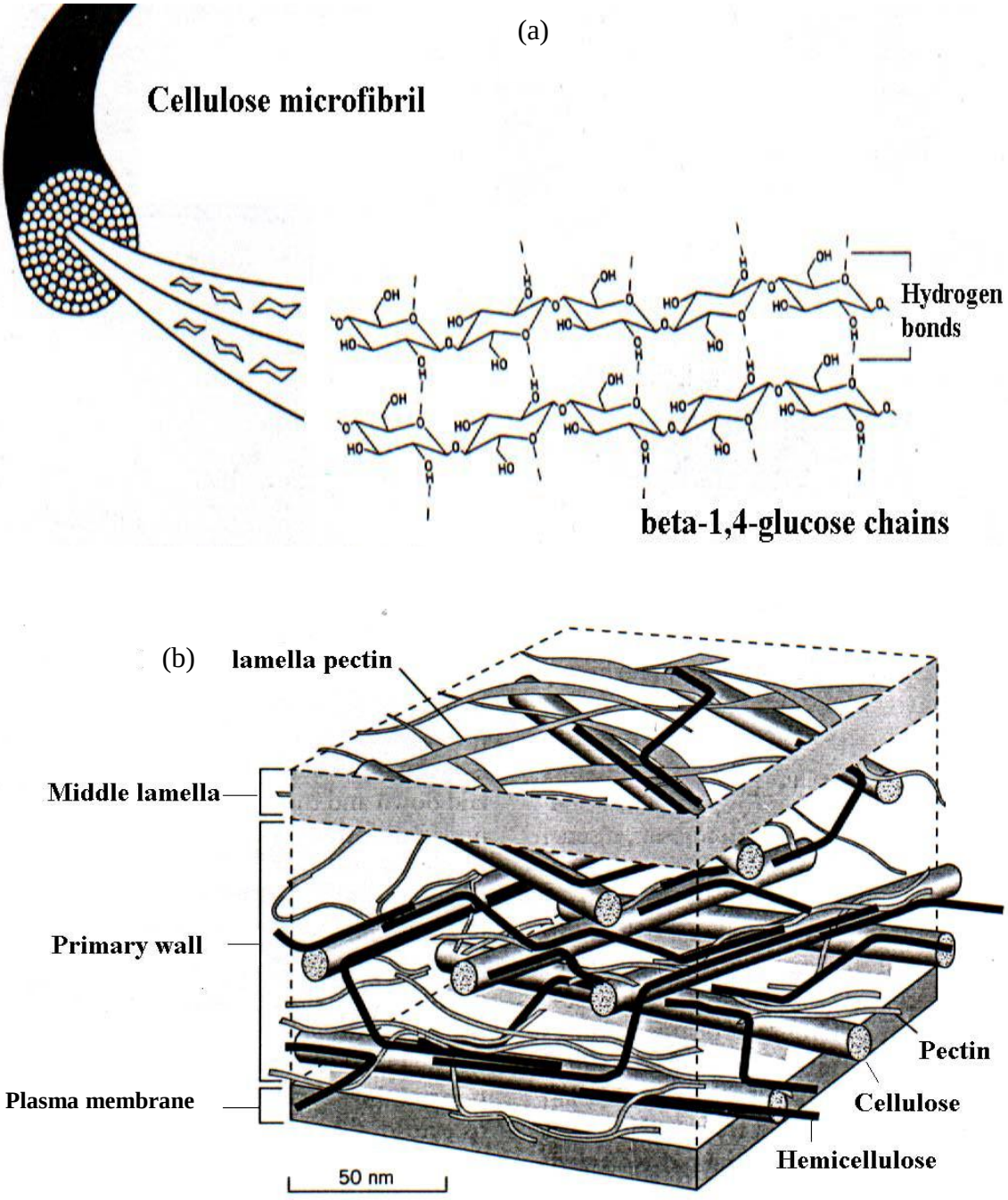

Figure 2.2. A simplified model to illustrate the cross-linking of cellulose microfibrils and hemicellulose in the lignocellulosic biomass (Hopkins 1995).

2.1.3. Cellulase enzyme system

At present, cellulase (EC 3.2.1.4) and related enzymes are used in food, brewery and wine, animal feed, textile and laundry, pharmaceutical, starch processing, pulp and paper industries, as well as in agriculture and for research purposes. Indeed, the demand for this enzyme is increasing significantly (Mach 2003).

Cellulase belongs to a group of enzymes which collectively hydrolyze cellulose. Cellulase is used commercially in industrial processes such as textile and paper production to prepare the surface of the cellulosic substrate for treatments such as printing and the application of resin finishes. There are at least three distinct enzymatic activities required for cellulase action (Fig. 2.3).

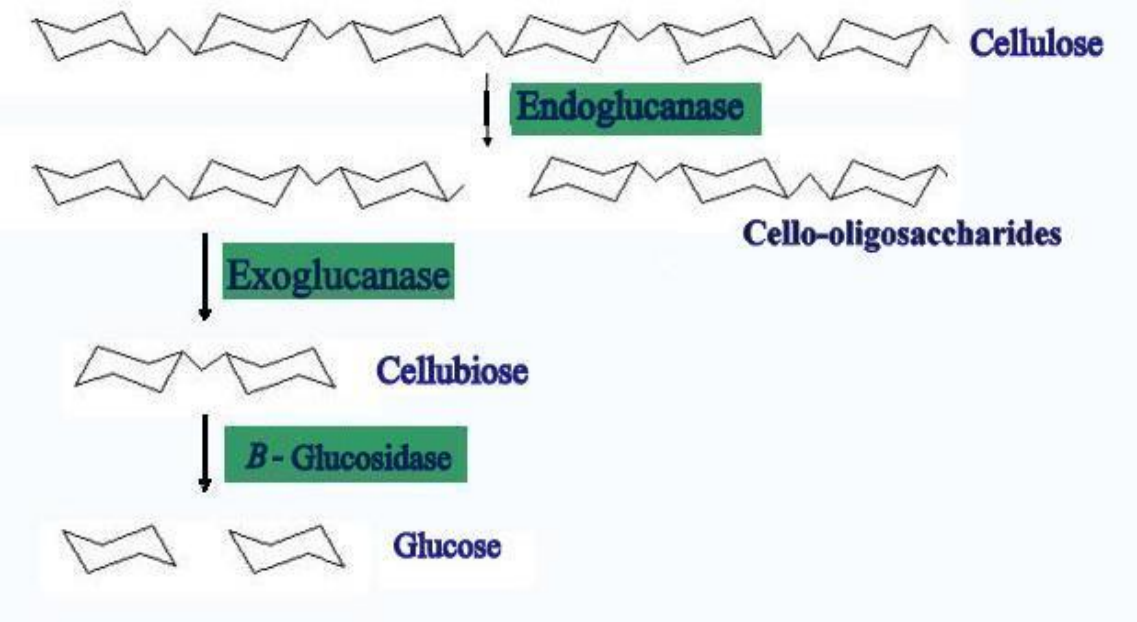

Figure 2.3. Enzymatic hydrolysis of cellulose to glucose.

Endoglucanase (EG, endo-1,4-D-glucanohydrolase) or EC 3.2.1.4. is the enzyme nomenclature of International Union of Biochemistry and Molecular Biology

for endoglucanase that locate randomly on the amorphous site along the cellulose polysaccharide chain and insert a water molecule in the intramolecular *β*-1,4-glucosidic bonds, producing a new reducing or non-reducing oligosaccharides of variable size and consequently new chain ends (Atlas 2004).

Exoglucanases, including 1,4-*β*-D-glucan glucanohydrolases (also known as cellodextrinases) (EC 3.2.1.74) and 1,4-ß-D-glucan cellobiohydrolases (CBH, EC 3.2.1.91). Exoglucanases cut cellulose polysaccharide chain at the terminal end to release cellobiose (Barr *et al*., 1996). This enzyme can also act on microcrystalline cellulose structure (Teeri 1997).

β-glucosidases or β-glucoside glucohydrolases (EC 3.2.1.21) convert cellobiose to glucose (Doi *et al*., 2003). In addition to the three major groups of cellulase enzymes, there are also a number of ancillary enzymes that attack hemicellulose, such as glucuronidase, acetylesterase, xylanase, *β*-xylosidase, galactomannanase and glucomannanase (Duff and Murray, 1996). Following the enzymatic hydrolysis, cellulose degraded by the cellulases to reducing sugars can be fermented by yeasts or bacteria to ethanol.

Cellobiohydrolase acts on crystalline cellulose Avicel, filter paper, cotton, in addition to a derivative of amorphous soluble (carboxymethylcellulose), acid-swollen cellulose, polysaccharides and other such as substrate including xylan (Fig. 2.4 a).

Endoglucanase on the other hand hydrolyzes primarily amorphous cellulose (carboxymethylcellulose) producing reducing sugar and lowering the viscosity of carboxymethylcellulose (Fig. 2.4 b). It also releases fibers from the filter paper and acts

weakly on Avicel and cotton. The molecular weight of cellobiohydrolase (41 to 68 kDa) is greater than that of endoglucanase (12.5 to 50 kDa).

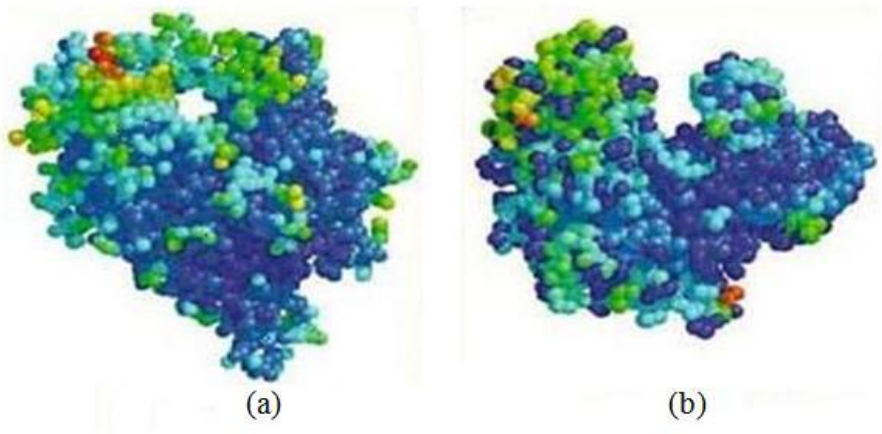

Figure 2.4. Typical cellulase structure. (a): Exogluconase, (b): Endogluconase (Bayer *et al*., 2006).

β-Glucosidase acts on sophorose and cellobiose to produce monosaccharide and on derivatives of substrates e.g., *p*-nitroso-*β*-D-glucoside producing colored compound. This enzyme has a high molecular weight (73 to 150 kDa). All these enzyme contain carbohydrates, but the relative quantities of sugar vary e.g., *β*-glucosidase and cellobiohydrolase. Both endogluconase and cellobiohydrolase rarely have been reported without any carbohydrate. It is possible that glycosylation may not be essential for the biosynthesis or secretion of these two enzymes. Cellobiohydrolase appears to contain acidic amino acids.

As depicted in Fig. 2.5, it is likely that enzymatic hydrolysis of oligomers is rapid relative to cellulose hydrolysis. Thus, in the absence of cells, oligomers would not be expected to accumulate and cellobiose would be the only apparent product of cellulose hydrolysis, as is commonly observed (Wells *et al*., 1995).

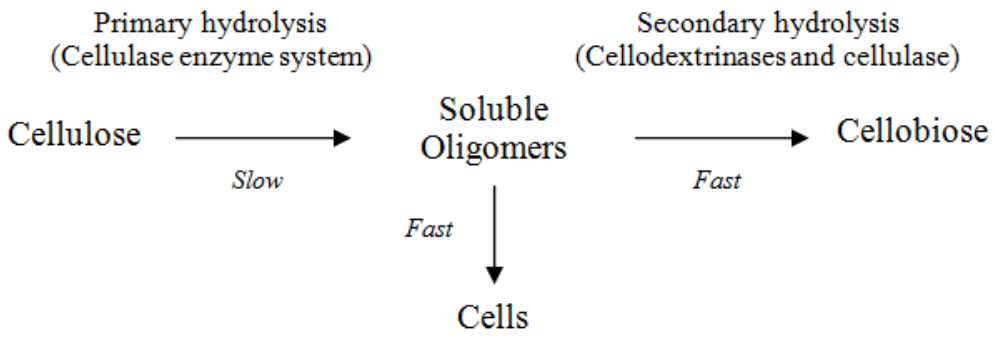

Figure 2.5. Hypothesis for the role of oligomers during microbially and enzymatically mediated cellulose hydrolysis (Wells *et al*., 1995).

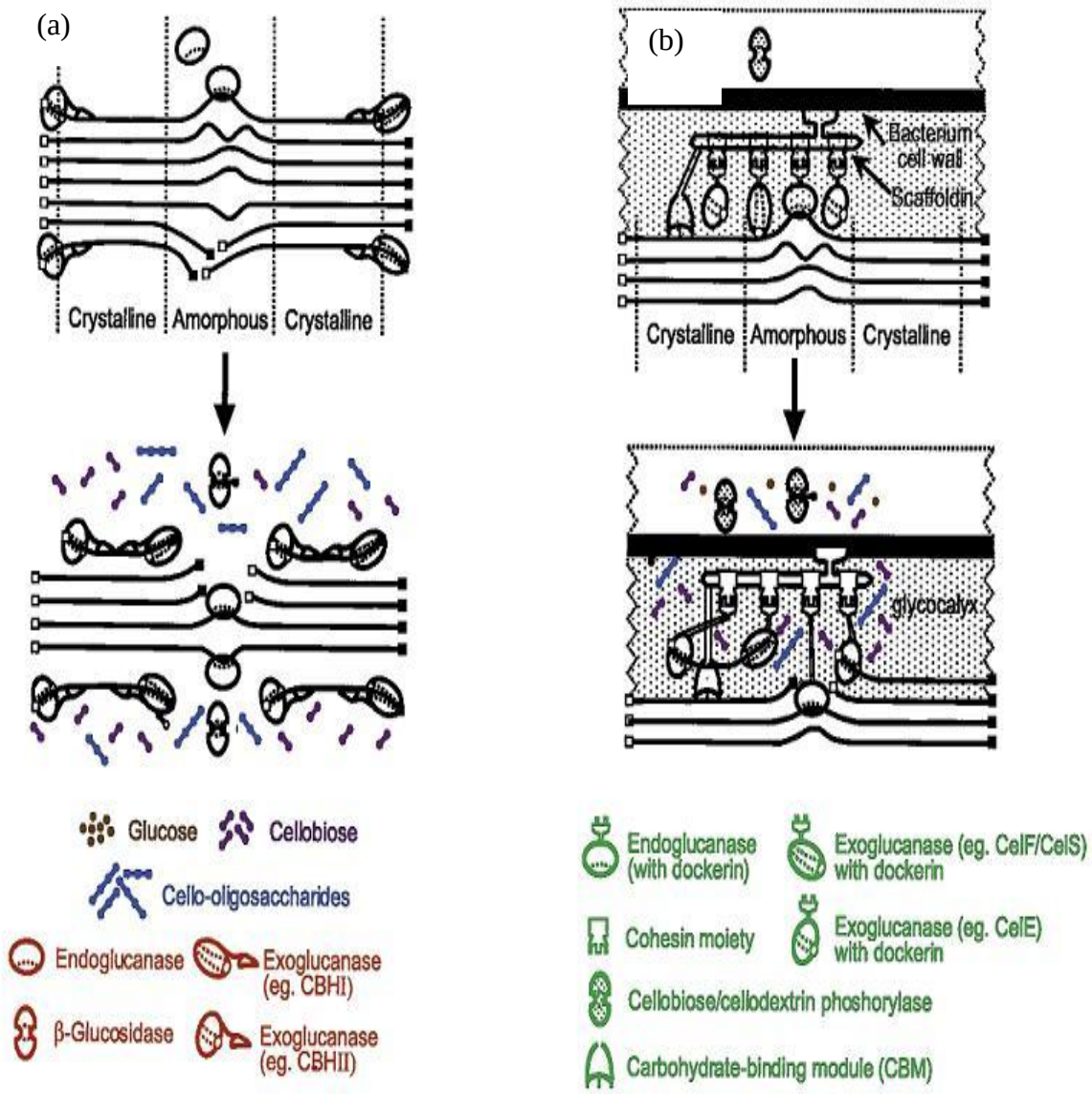

Figure 2.6. Schematic representation of the hydrolysis of amorphous and microcrystalline cellulose by non-complexed (a) and complexed (b) cellulase systems. The solid squares represent reducing ends, and the open squares represent nonreducing ends. Amorphous and crystalline regions which are two different types of cellulose are indicated in this figure. Cellulose, enzymes, and hydrolytic products are not shown to scale (Lynd *et al*., 2002).

Enzymatic hydrolysis of cellulose consists of three steps: adsorption of cellulase enzymes onto the surface of the cellulose, the biodegradation of cellulose to fermentable sugars, and desorption of cellulases. Cellulase activity decreases during the hydrolysis. The irreversible adsorption of cellulase on cellulose is partially responsible for this deactivation (Converse *et al*., 1988).

The two different structural types of cellulase systems found in bacteria and fungi are designated non-complexed and complexed. Some anaerobes are known to produce an extracellular multi-enzyme complex called a cellulosome (Schwarz, 2001) (Fig. 2.6a). The cellulosome comprises cellulases organised on a non-catalytic scaffolding protein which mediates binding to the cellulose. In contrast, cellulases from the majority of aerobes are not produced as complexes but bind directly to the cellulose (Zhang and Lynd, 2004). These non-complexed cellulases can have a modular structure with non-catalytic carbohydrate binding domains (CBD) linked to catalytic domains by flexible linkers. The enzymes in this cellulase system do not form stable high-molecular weight complexes and therefore are called non-complexed system that is produced from aerobic fungi such as *T. ressei* (Fig. 2.6b) (Sheehan and Himmel, 1999).

In the near future one of the important potential applications of cellulases and *β*-glucosidases will be the production of fuel ethanol from lignocellulosic biomass (Olsson *et al*., 2003), which is a good substitute for gasoline in internal combustion engines. The most promising technology for conversion of lignocellulosic biomass to fuel ethanol is based on the enzymatic breakdown of cellulose using cellulase enzyme (Holker *et al*., 2004, Juhasz *et al*., 2004).

Enzymatic saccharification of cellulose to glucose will likely be a central reaction in the conversion of lignocellulosic biomass to value-added products. Unfortunately, the rate of this reaction is notoriously slow. Presently, even under ideal conditions, a week or more is required to approach cellulose conversions of 90% (Roche *et al*., 2009a, Roche *et al*., 2009b).

However, numerous studies on the potential of additives such as surfactants, polymers, and proteins to affect the kinetics of enzymatic saccharification have been completed. For example, Kristensen and co-workers observed that the presence of polyethylene glycol (PEG) can significantly improve cellulose conversions (Kristensen *et al*., 2007). Several other surfactants such as Tween-20 (Polysorbate 20), Tween-80 (Polysorbate 80), Triton X-100 and nonylphenol ethoxylates improved the reaction kinetics, possibly by allowing enzymes to desorb more readily from the substrate (Ballesteros *et al*., 1998, Kim *et al*., 2007).

A main class of non-ionic surfactant widely used in the biotechnology industry is Tweens. Their popularity is basically as the result of their effectiveness at low concentrations and relatively low toxicities. Furthermore, they usually do not strongly intercalate with active ingredients. These characteristics have made them one of the key components in the industries in the past two decades. In addition, they have been widely used to prevent and/or inhibit enzyme surface adsorption (Gombotz *et al*., 1996) and aggregation under different processing situation, such as refolding (Bam *et al*., 1996), mixing (Katakam *et al*., 1995a,b), freeze–thawing (Chang *et al*., 1996), freeze-drying (Sarciaux *et al*., 1999), and reconstitution (Zhang et al., 1995; Zhang et al., 1996).

Others investigated the binding of ionic surfactants such as cetyl trimethylammoniumbromide (CTAB), cetyl pyridinium chloride (CPCl), and sodium dodecylbenzene sulfonate (NaDBS) to cellulosic substrates, which in turn may affect cellulase binding isotherms (Paria *et al*., 2004). Polymers such as polyethylene glycol also enhance cellulose conversion (Borjesson *et al*., 2007, Kristensen *et al*., 2007) as has bovine serum albumin (Yang and Wyman 2006).

The cellulase is an inducible enzyme system in which several carbon sources have been investigated as find the inducers (Mandels and Reese 1957, Muthuvelayudham 2005). Cellulose itself has been recognized as one of the best inducers for the complete cellulase complex. The other important inducers include saphorose and lactose (Mandels 1957, Mandels *et al*., 1962, Harikrishna *et al.,* 2000). Ryu and Mandels (1980) have stated that cellulose, cellobiose and lactose are effective inducers only at high concentrations. Muthuvelayudham *et al*. (2006) have demonstrated that the biosynthesis of cellulase in *T. reesei* QM9414 increased by using a mixture of cellulose and lactose in the culture medium. Similarly, a mixture of lactose (0.5%) and lactobionic acid (0.5%) has been proved to be a good inducer for cellulase production in *T. reesei* M7 (Janas 2002). Some studies indicated that the cellulase biosynthesis is repressed by a glucose catabolite (Suto and Tomita 2001), when glucose is pulse-fed to the culture in which cellulase biosynthesis is in progress, until glucose is exhausted or its residual concentration falls below a critical level of 0.1 g L^{-1} (Peitersen 1977).

The rate-enhancing mechanisms of these additives appear to correlate with the reduction of unproductive binding of cellulase to lignin. Lignin is chemically inert in saccharification processes, and irreversible adsorption to lignin effectively reduces the concentration of active enzymes (Kumar and Wyman 2009).

2.1.4. Endogluconases activity using viscometric method

Selective methods for endogluconase activity determination in multi enzyme samples are based on measurements of the viscosity of carboxymethylcellulose (CMC) solution in the presence of the enzyme. While exogluconase removing small fragments from the ends of polymers molecules do not significantly reduce the viscosity of the polymer solution, endogluconases break down internal bonds in CMC molecules resulting in a rapid decrease in the viscosity. Viscometers of different construction can be used for determination of endogluconases activity (Skomarovsky *et al*., 2000).

The viscometric method is considered feasible for evaluating endoglucanases activity since it is highly sensitive for reactions that hydrolyze internal bonds within a polymer molecule (Almin and Eriksson 1967a, Manning 1981). The change in viscosity is primarily due to the change in the degree of polymerization of CMC that is the result of cleavage of the glucosidic linkages remote from the chain end. In contrast, exoglucanases, which act on CMC near the chain end, give little change in viscosity while releasing significant amounts of reducing end-groups (Esterbauer *et al*., 1985).

Endocgluconases activity is usually determined by measuring the amount of reducing cello-oligosaccharides liberated from carboxymethylcellulose (CMC) into the reaction mixture using colorimetric methods such as the Somogyi–Nelson method (Somogyi 1952). However, longer-chain CMC fragments that are actually inert with the Somogyi–Nelson reagent may be also produced as a result of endodegradation. Thus, the colorimetric method cannot be used effectively to evaluate the production of longer-

chain nonreducing sugars as a result of endocellulase activity. On the other hand, viscometric analysis can be applied to evaluate endocellulase activity because the hydrolysis of internal bonds within polymer molecules alters the viscosity of a solution (Almin and Eriksson 1967a, b). Endocellulase activity determined by viscometric methods is most often expressed in terms of arbitrary viscometric units based on the initial rate of decline in specific viscosity or the initial rate of increase in specific fluidity (Ishihara *et al*., 2005).

Unlike viscometric measurements, methods based on determination of reducing end-groups (reductometric methods) are sensitive to both endo- and exo- action patterns of CMC hydrolysis and provide the activity of endoglucanases as the number of glucosidic bonds hydrolyzed per unit time (Esterbauer et al., 1985, Sharrock 1988).

The activity in this case is often called the CMCase activity. Various procedures for determination of the CMCase activity have been reported (Beldman *et al*., 1985, Esterbauer *et a*l., 1985). Since the reductometric methods measure all reducing groups in solution independent of the position within the polymer chain where it is released, it is clear that for crude enzyme preparations containing both endo- and exoglucanases, CMCase activity is not entirely synonymous with endoglucanase activity (Canevascini and Gattlen 1981).

The experimental measurement of viscosity is relatively simple, but uncertainty exists in the mathematical calculation that converts viscosity of a polymer in solution to the enzymatic activity. In the current study, the correlation between viscosity and

enzyme activity was investigated and the mathematical method for inter-conversion of these two parameters was formulated.

2.2. Mathematical modeling

Mathematical modeling in biological systems is an important task aimed to increase and efficiently use data structures and communication tools. It involves the application of computer simulation of cellular systems (such as the networks of metabolites and enzymes which contain metabolism, signal transduction pathways and gene regulatory system) to both analyze and imagine the multipart connections of the biological processes. Usual study of biological systems requires reductive methods in which quantities of data are gathered by category, such as concentration over time in response to a certain motivation. The goal of modeling is to make accurate real-time models of a system's response to environmental and internal stimuli (Barabasi and Oltvai 2004).

A monograph on mathematical modeling summarize a wide amount of published research in this area up to 1987, including topics in the following areas: computer modeling in biology and medicine, arterial system models, neuron models, biochemical networks, quantum automata, quantum computers in molecular biology and genetics, cancer modeling, neural nets, genetic networks, abstract relational biology, metabolic-replication systems, category theory applications in biology and medicine, automata theory, cellular automata and complete self-reproduction, chaotic systems in organisms, relational biology and organism theories (Bonneau 2008).

2.2.1. General Regression Theory

Regression analysis gives us the ability to summarize a collection of sampled data by fitting it to a model that will accurately describe the data. Each regression model has adjustable parameters, or variables, which can be adjusted in order to achieve close agreement between values of the regression model and the sampled data. These model parameters typically come from derived scientific or statistical theory that the data is supposed to satisfy. Regression analysis can turn the sampled data points into a smooth continuous function that may be used analytically or utilized by a computer program to return expected values at certain values of the independent variable. The user may decide to fit all of the variables in the regression model, or constrain some of them in order to satisfy some known conditions (Rashidian *et al*., 2006).

For example, equation 2.1 is commonly used model to describe exponential growth:

$$y = ae^{bx} \qquad \text{Eq. (2.1)}$$

This model contains two parameters, or variables, *a* and *b*, as well as the independent, or predictor variable *x*. The variable *y* is a dependent or response variable.

The basic idea behind regression analysis is to choose a method of measuring the agreement between your data and a regression model with a particular choice of variables. This measurement of agreement is called the merit function, and is arranged so that small values represent close agreement between the collected data and the regression model. The variables are then adjusted iteratively (in the case of nonlinear regression) in order to minimize the merit function. Once the merit function has been minimized, it is possible to determine how well the model describes the data.

There are more issues to consider in addition to just finding the optimal model parameters:

1. Which model to choose? Usually the model is known prior to performing the regression, but there are times when it may not be;
2. How well does the selected model describe the data? There should be a way to numerically determine the goodness of fit. It is sometimes not good enough just 'eyeballing' the plot of model vs. data;
3. Are there measurement errors involved? It is very rare for a model to exactly match the measurements made, so there should be a way to model the measurement errors and determine whether or not the model is statistically valid;
4. How accurate are the optimized parameters? There should be a way to statistically determine the likelihood of errors in the optimized parameters.

There are two widely used and accepted methods for performing regression analysis. The first, and easiest to implement, is linear regression. The second more general method is called nonlinear regression (Ott and Longnecker, 2008).

2.2.1.1. Linear Regression

$$y = a + a_2 x + a_3 x^2 + \cdots + a_M x^{M-1} \qquad \text{Eq. (2.2)}$$

This equation 2.2 represents a polynomial of degree M-1. Linear regression is not limited to polynomial functions only. The general form for a Least Squares linear regression model is presented in equation 2.3:

$$y(x) = \sum_{k=1}^{M} a_k f_k(x) \qquad \text{Eq. (2.3)}$$

where $f_k(x)$ is any arbitrary function of x. In regression modelling, the term 'linear' does not mean that the function of x itself is linear (in other words, a straight line), but that the models dependence on its parameters a_k is linear.

The merit function can take the form of equation 2.4:

$$\chi^2 = \sum_{i=1}^{n} \left\{ \frac{y_i - \sum_{k=1}^{M} a_k f_k(x_i)}{\sigma_i} \right\}^2 \qquad \text{Eq. (2.4)}$$

where σ_i is the measurement error, or standard deviation of the ith data point. From the merit function, the sum of the squares of the distances between the actual data points and the regression line are minimized. For regression models whose dependence on its parameters is linear, this is a straightforward process. However, some models that appear nonlinear may be re-arranged as to appear linear. For example, the equation 2.5:

$$y = ax^b \qquad \text{Eq. (2.5)}$$

may be 'linearized' by taking the natural log of both sides of the equation 2.6 and re-arranging it:

$$\ln\,(y) = \ln\,\left(ax^b\right) = \ln\,(a) + \ln\,\left(x^b\right) = \ln\,(a) + \mathrm{b}\ln\,(x) \qquad \text{Eq. (2.6)}$$

Once the model is linearized, it has the form of the general least squares regression model shown above. The fitting algorithm and merit function is now operating on ln (y_i) as opposed to y_i, which may in some cases have a significant effect on the accuracy of the estimated parameters. Also, there are only a few useful regression models that can be linearized in this fashion. To circumvent these issues, we must turn to nonlinear regression modelling.

There are several different techniques available to minimize the merit function for linear least squares models. The technique used in this study is Singular Value Decomposition (Wall *et al*., 2003) because of its exceptional ability to handle singular matrices common in least squares solutions.

2.2.1.2. Nonlinear Regression

Similar to linear regression, the goal of nonlinear regression is to determine the best-fit parameters for a model by minimizing a chosen merit function. The difference is that the nonlinear regression model has a nonlinear dependence on the unknown parameters, and the process of merit function minimization is an iterative approach. The process is to start with some initial estimates and incorporates algorithms to improve the estimates iteratively. The new estimates then become a starting point for the next iteration. These iterations continue until the merit function effectively stops decreasing (Blum and Francois, 2010).

The nonlinear model to be fitted can be represented by equation 2.7:

$$y = y\,(x;a) \qquad \text{Eq. (2.7)}$$

The merit function minimized in performing nonlinear regression is as the following equation 2.8:

$$\chi^2(a) = \sum_{i=1}^{N} \left\{ \frac{y_i - y(x_i; a)}{\sigma_i} \right\}^2 \qquad \text{Eq. (2.8)}$$

where σ_i is the measurement error, or standard deviation of the *i*th data point. As with linear regression, the sum of the squares of the distances between the actual data points and the regression line are minimized.

Nonlinear regression iterations proceed as follows:

1. Obtain initial estimates for all of the variables being fitted for in the model. These initial estimates can be obtained from linear regression, rules, or by examining the curve generated by the data points. For models pre-defined in regression processes, linear regression is used to obtain the initial estimates. For user defined models, either rules need to be created or the user must specify the initial estimates;
2. Using the initial estimates, the merit function is computed;
3. Then an algorithm is used to adjust the variables in order to improve the fit of the model to the data points. In this study, the Levenberg-Marquardt method was utilized and models use analytical and/or numerical derivatives during the optimization process;
4. Again, the merit function is computed and compared to the previous iteration;
5. Repeat steps 3 and 4 until there is essentially no change in the merit function, then cease the iterations;
6. Calculated the goodness of fit statistics.

Why is this better method than linear regression? To start with, it is a much more general procedure. There are a very limited number of models that can be expressed in linear form without transforming the data. Also, remember that transforming the data means that the fitting routine will be minimizing the merit function on the transformed data, not the actual data. This makes nonlinear regression more accurate. Nonlinear regression can also be applied to essentially any equation that defines the independent variable *Y* as a function of the independent variable(s) *X* and at least one parameter.

2.2.2. Derivatives and Nonlinear Regression

Most of the overhead associated with the Levenberg-Marquardt algorithm lies with calculating derivatives. The derivative of the model with respect to every parameter being fitted must be calculated at every data point a number of times during the solution process. It is easy to see how this can be time consuming, especially if there are a large number of data points (Ralston and Jennrich 1978). The derivatives are calculated iteratively in the following manner:

1. The function is evaluated, offsetting the parameter by some (rather significant) small positive number by equation 2.9:

$$f(x+h) \qquad \text{Eq. (2.9)}$$

2. Again, the function is evaluated, offsetting the parameter variable by the same small number negated equation 2.10:

$$f(x-h) \qquad \text{Eq. (2.10)}$$

3. Then, the derivative approximation is evaluated by equation 2.11:

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} \qquad \text{Eq. (2.11)}$$

4. The extrapolation error is subsequently estimated;
5. Value is decreased, and steps 1 through 4 are repeated. When the error reaches an acceptable level, stop and return the approximated derivative.

2.2.3. Understanding and Interpreting Regression Results

This section first explains how the results are calculated in data fitting procedure (DataFit), and then how to interpret the results.

Predicted Value

The i^{th} predicted, or fitted value of the dependent variable Y, is denoted by $Y_{\mathbf{i}}$. This value is obtained by evaluating the regression model $\hat{Y} = f(X, \hat{\beta}_j)$, where $\hat{\beta}_j$ are the regression parameters, or variables.

Residuals

$$i\text{th residual} = (Y_i - \hat{Y}_i) \qquad \text{Eq. (2.12)}$$

Residuals are the vertical difference between the actual data points Y_i and the curve generated from the predicted values $\widehat{Y}_i$. If a residual is positive, it means that the

actual data point lies above the curve in equation 2.12. A negative residual means that the actual data point lies below the curve. If the residual is zero, the actual data point lies on the curve. The larger the residual, the further the data point lies from the curve.

Sum of Residuals

$$\text{Sum of Residuals} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i) \qquad \text{Eq. (2.13)}$$

where *n* is the number of data points, or observations. This is the total sum of the residuals for all data points. According to equation 2.13 if the curve passed through each data point, the sum of residuals would be zero. It must be remembered, though, that a regression model can have large positive and negative residuals and still sum to a small number.

Average Residual

$$\text{Average Residuals} = \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)}{n} \qquad \text{Eq. (2.14)}$$

This is the average value of the residuals in equation 2.14.

Residual Sum of Squares

The absolute residual sum of squares is the sum of the squares of the differences between the actual data points and the predicted values (Eq. 15), where the relative residual sum of squares is the *weighted* sum of the squares of the differences between the actual data points and the predicted values (Eq. 16). If no standard deviation information was entered with the input data (the regression was not weighted), these two quantities will be equal.

$$\textbf{SSE} = \text{Residual or Error Sum of Squares (Absolute)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (Y_i - \hat{Y}_i)^2 \qquad \text{Eq. (2.15)}$$

$$\boldsymbol{SSER} = \textit{Residual or Error Sum of Squares (Relative)} \qquad \text{Eq. (2.16)}$$

$$= \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} [(Y_i - \hat{Y}_i)^2 * W_i]$$

where $W_i = \frac{1.0}{{\sigma_i}^2}$ normalized so that $\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} W_i = n$, σ_i= the standard deviation of the i^{th} data point Y_i, n is the number of data points, or observations.

The principle behind nonlinear regression is to minimize the residual sum of squares by adjusting the parameters $\widehat{\beta}_j$ in the regression model to bring the curve close to the data points. This parameter is also referred to as the error sum of squares, or ***SSE***. If the residual sum of squares is equal to 0.0, the curve passes through every data point.

Error Variance

$$\hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-p} = \frac{SSE}{n-p} \qquad \text{Eq. (2.17)}$$

where p is the number of parameters or variables in the regression model in equation 2.17, then, n is the number of data points, or observations. This is an estimate of the error variance σ^2.

Standard Error of the Estimate

$$\sigma^2 = S = \sqrt{\hat{\sigma}^2} \qquad \text{Eq. (2.18)}$$

This is the standard deviation of the residuals in equation 2.18.

R^2, or Coefficient of Multiple Determination

$$R^2 = \frac{SSR}{SST} = 1.0 - \frac{SSE}{SST} \qquad (0 \le R^2 \le 1)$$

$$\textbf{SSR} = \text{regression sum of squares} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \qquad \text{Eq. (2.19)}$$

$$\textbf{SSE} = \text{residual or error sum or squares} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad \text{Eq. (2.20)}$$

$$\textbf{SST} = \text{total sum of squares} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad \text{Eq. (2.21)}$$

and **SST** = **SSE** + **SSR**

SST measures the variation in the observed response. **SSR** measures the "explained" variation. **SSE** measures the "unexplained" variation. Therefore, R^2 measures the proportion of variation in the data points Y_i which is explained by the regression model. For example, if R^2 = 0.95, the 95% of the variation in the dependent, or response variable Y is explained by the regression model. A value of R^2 = 1.0 means that the curve passes through every data point. A value of R^2 = 0.0 means that the regression model does not describe the data any better than a horizontal line passing through the average of the data points (Nagelkerke 1991).

R_a^2 or Adjusted Coefficient of Multiple Determinations

$$R_a{}^2 = \frac{(n-1)R^2 - k}{n-1-k} \qquad R_a{}^2 \leq R^2 \qquad \text{Eq. (2.22)}$$

where k = number of regression parameters in the model and n = number of data points. $R_a{}^2$ is used to balance the cost of using a model with more parameters against the increase in R^2.

2.2.4. Regression Variable (parameter) Results Value

The data input in the Value column are estimated for fitted parameter values represented mathematically by $\widehat{\beta}_j$.

Standard Error

The data in the Standard Error column are the estimates of the standard deviations of the fitted regression parameters, represented mathematically by $S_{\widehat{\beta}_j}$.

t-ratio

$$t - \text{ratio} = \frac{\widehat{\beta}_j}{S_{\widehat{\beta}_j}} \qquad \text{Eq. (2.23)}$$

where $\widehat{\beta}_j$ is the estimated, or fitted parameter value. $S_{\widehat{\beta}_j}$ is the estimated standard deviation of $\widehat{\beta}_j$.

Prob(*t*) or *p*-value

This is used to test the null hypothesis $H_0: \beta_j = 0$ for each parameter. The smaller the value of Prob(*t*), the less likely the parameter is actually zero. For example, if Prob(*t*) = 0.01, there is a 1% chance that the actual parameter is zero. If Prob(*t*) = 0.95, there is a 95% chance that the actual parameter value is zero. In cases like the latter, the parameter in question can usually be removed from the model without affecting the regression accuracy.

2.2.5. Confidence Intervals

It is not entirely correct to say that, for example, a 90% confidence interval means that there is a 90% chance that the actual value of the parameter lies within the confidence interval. It is correct, however, to say that if an experiment is performed many times (sample data and perform regression analysis on it), 95% of the computed

confidence intervals will contain the actual value of the parameter. On the other hand, 5% of the computed confidence intervals will fail to contain the value of the actual parameter.

2.2.6. Variance Inflation Factors (VIF)

The Variance Inflation Factors are calculated only when performing variable selection. They are used to determine the level of multi-collinearity between the independent variables.

The variance inflation factors (VIF) are a measure of how well each independent variable can be predicted from all of the others (excluding the dependent variable). The VIF for a particular independent variable is calculated from the R^2 values determined by creating and fitting a regression model comprised of the remaining variables:

The individual R^2 in the above equation is not to be confused with the overall R^2 of the regression model. The overall R^2 is the goodness of fit measure of the entire regression model. In fact, it is derived that the overall R^2 to be high (indicating a good fit for the entire model), and the individual R^2's to be low (indicating minimal collinearity between variables). If an individual R^2 is high (indicating substantial collinearity between variables), the VIF will be greater than 1.0. If an individual R^2 is low, the VIF will approach 1.0. In summary, the effect of collinearity on the regression model is that it will increase the width of the confidence intervals for the equation coefficients by a factor of the square root of the VIF (hence the name variance inflation factor).

2.2.7. Determining the goodness of fit

When determining the goodness of fit of the models, the following points should be examined:

The solution convergence for nonlinear models should be checked. Each iterative step of the nonlinear solver returns the best estimate found so far in the solution process. After each iteration, the merit function is compared to that from the previous iteration. Since the solver returns the best estimates reached so far, the newly computed merit function will either be better (lower) or unchanged. So as to not run on indefinitely, we stop the process if the percentage difference in the merit function between iterations reaches a reasonable specified Regression Tolerance, a Maximum Number of Iterations, or a Maximum Number of Unchanged Iterations. If the solution reached the Maximum Number of Iterations, it is worth checking to see if the merit function was steadily decreasing and increase the allowable number of iterations (Straume and Johnson 1992).

The residual scatter plot should be examined. The residuals should be randomly scattered around zero and show no discernable pattern, i.e., they should have no relationship to the value of the independent variable. If there are groups of residuals with like signs, or the residuals increase or decrease as a function of the independent variable, it is probable that another functional approximation exists that would better describe the data.

The residuals should be checked that they are normally distributed by looking at the residual probability plot. The residual probability plot shows a plot of the

normalized residuals on the vertical axis and the normal quantiles on the horizontal axis. If the residuals are normally distributed around zero, the plot should be a straight line with a 45-degree slope passing through the origin. You can compare this to the reference line which has a slope of one and an intercept of zero.

Plot of the regression model and the data points should be examined. The data points should be randomly distributed above and below the curve.

Check to see how well the regression model describes the actual data. This information can be obtained by the following calculated parameters: measures the proportion of variation in the data points Y_i which is explained by the regression model. A value of $R^2 = 1.0$ means that the curve passes through every data point. A value of $R^2 = 0.0$ means that the regression model does not describe the data any better than a horizontal line passing through the average of the data points.

The Residual Sum of Squares (RSS) is the sum of the squares of the differences between the entered data and the curve generated from the fitted regression model. A perfect fit would yield a residual sum of squares of 0.0.

The Standard Error of the Estimate is the standard deviation of the differences between the entered data and the curve generated from the fitted model. This gives an idea about how scattered the residuals are around the average. As the standard error approaches 0.0, you can be more certain that the regression model accurately describes the data. A perfect fit would yield a standard error of 0.0.

The % Error is the percentage of error in the estimated dependent variable value as compared to the actual value. An error percentage of 0% means that the estimated value is equal to the actual value. The larger the percent error (positively or negatively), the farther away the estimated data point is from the actual point.

The results should be check if they are scientifically or statistically meaningful. Does the fitted value of any of the variables violate a possible physical reality? For example, suppose a model is fitted in which one of the parameters represents electrical resistance and returns a negative value. This probably means that the model selected is not the correct one.

The confidence intervals should also be examined. The confidence intervals for each variable are reported at levels of 68%, 90%, 95% and 99%. If the confidence is very wide, the fit is not unique, meaning that different values chosen for the variables would result in nearly as good a result. Data containing a lot of scattering, or not collecting a sufficient amount of data would cause the confidence intervals to be excessive. However, the most common reason is fitting the data to a model with variable redundancy. In the equation $Y = \log(a \times b \times x)$, the variables a and b are indistinguishable. There is no way for the algorithm to determine how to distribute values (the product of a and b) between these two variables.

It is possible to converge on a false minimum in the merit function. This is a problem inherent in any iterative optimization procedure. Nonlinear regression will

assure that once a solution has been obtained, small changes in the parameters will worsen the fit. It is rare but possible, however, that some very large change may actually converge to a better fit. This problem is rare, except in cases where the data is widely scattered, there were too few data points collected or the model chosen is completely wrong for the data. Determining good initial estimates is important for highly nonlinear models, which is why they are calculated for use in DataFit when performing nonlinear regression. If it is suspected that a false minimum was reached, the Range feature may be used when specifying the initial estimates. This feature will solve the model with a range of initial estimates, returning the best-fit parameters from the enumerated initial conditions specified.