Polysaccharide sequence reconstruction from digest patterns

Honours Project 2006
Jonathan Hunt

Contents

Acknowledgements **iii**

Abstract **iv**

1 Introduction **1**

1.1 Polymeric Information . . . 1

1.1.1 Models of copolymer creation . . . 1

1.2 Pectin . . . 2

1.2.1 Commercial Production of Pectin . . . 2

1.2.2 Structure and biosynthesis of pectin . . . 3

1.2.3 Experimental tools for the elucidation of homogalacturonan fine-structure . . . 4

1.2.4 Structure-function relationship . . . 6

1.3 High Performance Capillary Electrophoresis and its role in investigating pectin . . . 6

1.3.1 Electroendoosmosis . . . 7

1.3.2 Electrophoretic Mobility . . . 8

1.3.3 Investigation of homogalacturonan methyl-esterification patterns using HPCE . 9

1.4 Previous attempts at simulation . . . 10

1.4.1 Previous models of endo-PG II digestion . . . 10

Subsite Energy Binding Model . . . 11

2 Experimental Details **12**

2.1 Simulations of Homogalacturonan Digestion . . . 12

2.2 Simulation of Enzyme Degradation . . . 12

2.3 Simulation of unmethyl-esterified homogalacturonan and detailed comparison with experiment . . . 13

2.3.1 Experimental Details . . . 15

2.4 Extending the endo-PG II model to incorporate partial methyl-esterification . . . 17

2.5 Comparison of Reconstruction Strategies . . . 18

3 Results and Discussion **21**
3.1 Simulation of unmethyl-esterified homogalacturonan and detailed comparison with experiment . . . 21
3.2 Extending the endo-PG II model to incorporate partial methyl-esterification . . . 25
3.3 Comparison of Reconstruction Strategies . . . 28
3.4 Future Work . . . 31
3.4.1 Other Experimental Systems . . . 33
Microarrays and other ways of investigating pectin . . . 33
3.4.2 Binding Energy Calculations . . . 33
3.4.3 Web interface for reconstruction . . . 37
3.4.4 Optimisation of BJPS . . . 37

4 Conclusion **38**

Bibliography **39**

A Compiled Digestion Data **45**

B Documentation for BJPS **47**
B.1 Technologies Used . . . 47
B.2 Installation Instructions . . . 47
B.3 Using the Library . . . 48
B.3.1 `PolyBase` Object . . . 48
B.3.2 `PolySet` Object . . . 49
B.3.3 `IsomerSet` Object . . . 50
B.3.4 `Stats` Object . . . 51
B.3.5 `Matcher` Object . . . 53

C Publication **55**

D Associated Electronic Data **64**

Acknowledgements

To my mother and father who taught me almost everything I know.

Thanks to Dr. Bill (Boss) Williams for being my supervisor. It's been interesting. Thanks so much for all your help, chats etc. etc. etc. Thanks also to the rest of the Massey Institute of Fundamental Sciences, especially the lecturers, for all their help throughout my degree. Thanks also to my classmates, for going through a challenging year together and to the rest of the biopolymer research group for introducing me to the real life of a research student and particularly to Aurelie Cucheval for the valuable experimental work at which she proved much more capable than I. Thanks also for putting up with me for the 2^{nd} half of the year and thanks to the geek room comrades for putting up with me the 1^{st} half of the year.

Thanks to the many friends who have taught me so much throughout my Massey years. Thanks to my flatmates this year, for helping keep me sane (even if their best efforts ultimately proved unsuccessful).

Also, I feel I should put a note for my Aunty Herds, who always encouraged me to write and claimed that she would read any book I wrote, consider this my first book and I hope you enjoy it. Thanks to all my extended family for their support throughout my study.

I should also acknowledge Massey University and Industrial Research Limited for funding the Massey University Scholar and Applied Mathematics scholarships respectively, which made this year financially viable.

Soli De gloria.

Abstract

A simulation methodology for predicting the time-course of enzymatic digestions based solely on the enzymes subsite binding energies is described and shown to be capable of correctly predicting the digestion of unmethyl-esterified homogalacturonan by endo-polygalacturonase II from *Aspergillus niger*. A possible extension of the model to copolymer digestion of partially methyl-esterified homogalacturonan is shown to be inadequate, probably due to the assumption that differing residues in neighbouring subsites do not affect one another. A more extensible model to serve as a basis for future work is proposed. Various strategies for the reconstruction of unfragmented polymer fine-structure from fragmentation results based on such models is tested and a novel and extensible library for such simulation is described.

Chapter 1 - Introduction

Biopolymers are, quite literally, what hold us together. However, far more than being just glue, biopolymers are the information carriers on which the blueprints that make us who we are is stored, copied and passed on. Deoxyribose nucleic acid (DNA), first elucidated by Watson and Crick [1] has become the most publicly well-known biopolymer, with the informationally-related proteins coming a close second, but biopolymers are far more numerous.

The simplest polymers are a single continuous chain of identical monomers. Most polymers in biological systems are far more complex - incorporating branching and consisting of more than one type of monomer (copolymers) [2]. While sequencing of DNA has become routine due to its double-stranded structure, lack of branching and polymerase chain reactions (PCR) which make obtaining many identical copies of a given sequence possible and thus macroscopically observable, the situation with other biological polymers of interest is not as simple. Because polysaccharides are more diverse structurally and there is no known experimental techniques for obtaining identical copies of a given chain, attempts to obtain the fine-structure of polysaccharides have met with more limited success. However, because polysaccharides play a huge role in areas of biology from cell signalling to cell-wall structure and are of commercial interest in the food industry, better understanding of fine-structure of polysaccharides is a problem of long-standing interest.

This report outlines an attempt to use computational tools to verify proposed models of enzymatic action and to use these verified models to find ways to better elucidate the fine structure. It is part of a wider effort at Massey to obtain a better link between the fine structure of polysaccharides and their macroscopic properties.

1.1 Polymeric Information

One of the problems with understanding the best way to describe biological copolymers is that the mechanisms of synthesis are not usually known and post-synthetic modifications are often also important. It has been shown that a process whose synthesis kinetics are describable by a first-order Markovian model can appear much more complicated when the resulting sequence is considered [3]. A chain is best expressed in the basis discerned from the mechanism of polymer synthesis or post-synthetic modification, unfortunately often this is not known.

1.1.1 Models of copolymer creation

There are many differing models of copolymer synthesis, the majority of which were created for modelling synthetic polymer systems. Here only the stationary Markovian model commonly employed to approximate copolymer synthesis [4] will be discussed. Markov models can also be used to approximate non-Markovian systems, the approximation will improve with higher orders [3]. The model consist of a set of states. A process starts out in one of those states and moves from one state to another. The Markov parameters are the probability of transitioning from each state to the next [5]. When modelling copolymer creation, these states would correspond to different residues on the chain, as the chain is synthesised. For n$^{\text{th}}$ order stationary Markov processes describing an AB copolymer there are 2^n states so 2^n independent parameters are needed to specify the model. For instance a first order Markovian model with only two states (such as might be postulated for an AB copolymer) requires only two independent parameters. These parameters are often specified as a transition matrix:

$$P = \begin{pmatrix} P_{aa} & 1 - P_{aa} \\ P_{ba} & 1 - P_{ba} \end{pmatrix} \tag{1.1}$$

where the matrix elements P_{ij} are given as the probability of transitioning from state i to j. If the process is irreducible (no zero entries) then the stationary distribution giving the average properties of a long chain created with this model is the eigenvector associated with an eigenvalue of 1. Markovian models may also describe the results of post-synthetic modifications.

1.2 Pectin

Pectin is possibly the most complex of all biopolymers [6], but its pervasiveness throughout the plant kingdom makes it too important to avoid studying. Found abundantly in all land plants, pectin is known to constitute an important part of the cell wall. Pectin was first isolated and named by Henri Braconnot in 1824 [7], who coined the name from the Greek word $\pi\eta\chi\tau o\sigma$ 'pektikos' meaning 'to solidify.' Aside from the inherent interest of better understanding biology, pectin is an important commercial polymer, with the average man in the Western world consuming 4-5 grams of pectin daily. Pectin has managed to retain its perception as a "natural" food product, contributes to fibre intake, and is generally considered a healthy food additive and linked with several health claims (some more verified than others) [8]. For commercial food use it is mainly sourced from apples and citrus fruits. There has been fairly limited investment in chemically modifying pectin for the food industry. This is partially due to the high cost of certifying modified pectin as safe for human consumption and also a desire to retain the "natural" label currently associated with pectin. Pectin's main role in the food industry is as a gelling agent and for the stabilisation of acidified milk products. There are few non-food uses of pectin.

1.2.1 Commercial Production of Pectin

The use of crude extracts of pectin to aid gellation in jam-making may predate commerical jam making. Certainly, it has been around for well over a century, with jam-makers buying dried apple pomace as the starting material from which they extracted pectin for jam [9]. Improving the pectin extraction gave jam makers a more consistent additive which could be stored until needed. Pectin is commercially derived as a by-product of other industry, with no crop being grown commercially specifically for its pectin. A large amount of the starting material for pectin extract is waste product from juice manufacturing.

The commercial extraction of pectin from fruit uses heat and acid, this is known to destroy the structure of pectin, freeing it from the cellulose skeleton. Higher molecular weight pectin is more valuable, so the extraction processes are optimised to produce this. Usually the extraction is sheared to overcome physical resistance of extracting the viscous pectin.

1.2.2 Structure and biosynthesis of pectin

The exact structure of pectin is a topic of debate, not suprising for a branching polysaccharide consisting of as many as 17 different monosaccharides. Pectin is believed to have a backbone of 1,4-linked alpha-D-Gal*p*A residues that are subdivided into three classes: homogalacturonan, substituted galacturonans, and rhamnogalacturonan I. The structure of homogalacturonan is of primary interest to this report. Figure 1.1 shows the chemical structure of homogalacturonan. Homogalacturonan is a copolymer with each residue having the possibility of methyl-esterification or acetylation as shown in the figure. The methyl-esterification of HG has attracted significant interest because it is one of the main factors in determining the industrial usefulness of pectin [6] and it is known many plants and fungai have enzymes for modifying it. The overall structure of pectin is difficult to determine because of its complexity and because of difficulties in extracting pectin from the cell wall intact. Two proposed models of the structure of pectin [6, 10] are shown in figure 1.2.

Antibodies that recognise specified structural features have become an important means of determining the in situ distribution of polysaccharides in the cell wall. There are antibodies that are believed to recognise homogalacturonan with different degrees of methyl-esterification and also antibodies that recognise other parts of pectin. These have been used to better understand the role of pectin in cell walls, and although results are still not completely clear, it does appear that the degree of methyl-esterification of homogalacturonan varies with position in the cell wall [11]. These antibodies are also being used to better understand food matrices.

The fine structure of various constituents of pectin are far better understood than the overall structure. This report primarily focuses on better elucidating the fine-structure of homogalacturonan methyl-esterification. Acetylation, is also known to have an effect on functional properties and enzyme interactions [12], but all samples in this report were unacetylated and no effort was made to characterise acetylation effects.

With the structure of pectin still under debate, it is no suprise that the biosynthesis of pectin is a challenging area. It does appear that homogalacturonan is highly methyl-esterified when first synthesised and is believed to be post-synthetically modified by pectin methyl-esterases [13]. It is

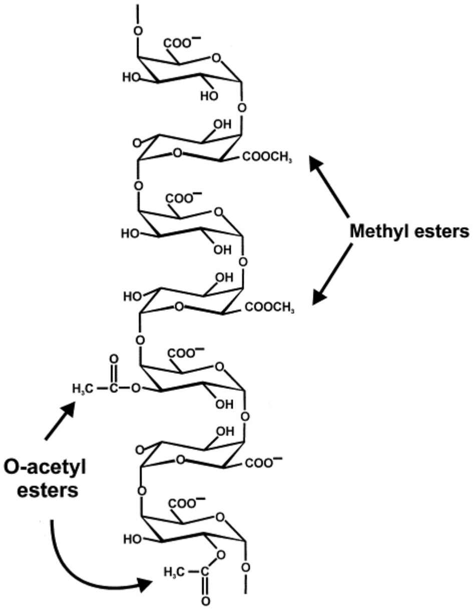
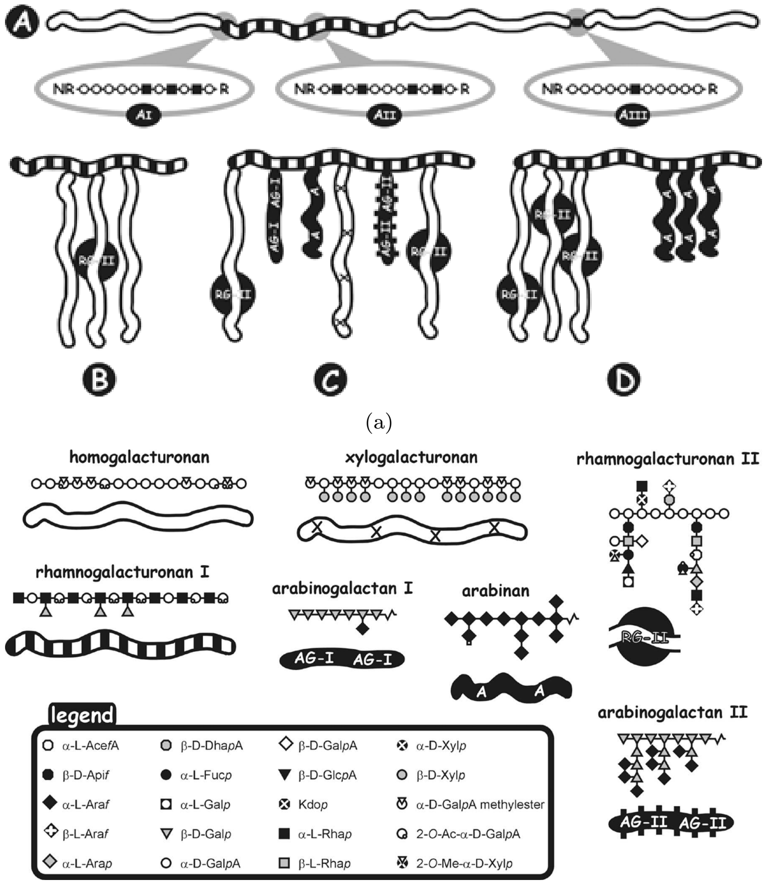


Figure 1.1: The primary structure of homogalacturonan with labels showing where the carboxylates can be methyl-esterified or acetylated. Figure from [11].

known that at least 53 enzymes are involved in its synthesis and that significant changes in the structure of pectin occur in golgi bodies [11]. Fungal pectin methyl-esterifases (PMEs) are known to randomly de-esterify while plant PMEs are known to result in blocky patterns [14,15]. Specific models for the methyl-esterification of pectin are not yet known. Commerically, low DM pectins are produced by acid de-esterification in alcohol [8].

1.2.3 Experimental tools for the elucidation of homogalacturonan fine-structure

There are huge number of experimental methods being used to elucidate the fine-structure of homogalacturonan. The most widely-used characterisation of the methyl-esterification pattern is the degree of blockiness (DB). Pectin is digested by an endo-PG enzyme and the fragments analysed. The DB is defined as the percentage of unmethyl-esterified GalA residues in the pectin sample fully digestable (to mono, di or trimer) by an endo-PG enzyme [16]. A further complication in the use of DB as a standard, is that there are many endo-PGs that are used to degrade the pectin, and the different enzymes used have different degrees of specificity. It is generally assumed that DB's determined using one enzyme are related linearly to DB's from another enyzme, although this is likely to be, at best, only approximately true. A previous attempt to improve on the characterisation of the methyl-esterification fine-structure of homogalacturonan in pectin used an endo-PG to degrade the pectin and



(a)

(b)

Figure 1.2: Proposed models of pectin structure. (a) shows the constituent polysaccharides of pectin. (b) shows the previously accepted model of pectin structure (A) and a proposed models with a rhamnogalacturonan I backbone and homogalacturonan as side-chains (B,C,D). Figures from [6].

the measure of unmethyl-esterified monomer, dimer and trimer released compared with the amount of methyle-esterified oligomers remaining was used as a measure of blockiness [17]. The authors propose a model for their enzyme, but simulation is not used as a means of verifying the proposal.

The primary method for characterising fragments outlined in this report is the use of High Performance Capillary Electrophoresis (HPCE) outlined further in section 1.3.3. Other commonly used techniques for fragment analysis are gel electrophoresis [18], sequencing using tandem mass spectometry [19] and chromatography techniques [17]. Methods used for analysising unfragmented (high DP) homogalacturonan include the use of Fourier-transform infrared spectrometry to determine DM distribution [16], NMR to measure triade frequencies [20] and antibody recognition of specific sequences [11].

1.2.4 Structure-function relationship

There is a close relationship between structure and function of pectin. It is hoped that in the future it will be possible to create 'designer pectins' with specific functionalities and even to modify pectin structure in the plant before extraction. The visco-elastic properties of two different pectins with the same DM but different patterning can be markedly different [16,21] with DB the more important factor in determining behaviour. Recent work has found that two different pectin samples whose functionality is markedly different, when separated into fractions based on DM and DB yielded samples from each pectin with similar properties [22, 23], showing that the DM and patterning of methyl-esterification are important factors in the chemico-physical properties. Another group used the ratio of released unmethyl-esterified mono-, di- and tri-galacutoronic acid to characterise the pectin and then used techniques borrowed from DNA to build a tree relating the various pectin samples to one another [24] and found, unsuprisingly, that pectins from the same organism tend to be more related to one another than pectins from other organisms.

1.3 High Performance Capillary Electrophoresis and its role in investigating pectin

High-Performance Capillary Electrophoresis (HPCE) is a commonly used technique, although one of the more recent chemical separation technologies. Electrophoresis, the separation of charged molecules in a fluid or gel, has been around for much longer than HPCE, usually using a gel medium for support and separation. HPCE has an advantage over gels of allowing higher voltages to be used because heat removal is much more efficient, and thus offers shorter separation times. Also, it more easily lends itself to automation than gel techniques. While CE has been in use since 1967, commercial instrumentation for HPCE only became available in 1988. A driving force behind the development of commerical CE has been its use in DNA sequencing, and the need for large-scale separations involved in the human genome project lead to the development of parallel HPCE machines, with many capillaries running at once [25]. An area of growing importance in many separation techniques is the use of micromachined

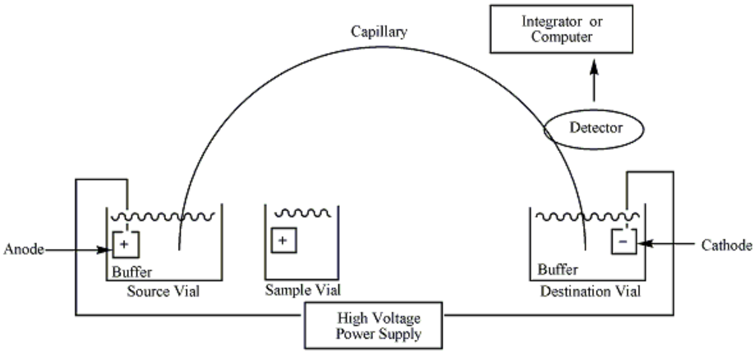


Figure 1.3: Schematic diagram of the basic components of a capillary electrophoresis machine. Figure from [26].

electrophoretic devices employing fabrication techniques similar to that for integrated circuits. They have the advantage of only requiring extremely small amounts of sample. Such devices are already in use for DNA sequencing and are expecting to become increasingly important, particularly in making separation techniques widely and cheaply available.

Figure 1.3 shows the essential components of a CE machine. It consists of a capillary connecting two buffer reservoirs, a high voltage power supply that generates a voltage difference between the buffers and a detector of some type to detect species as they migrate along the capillary. In almost all modern instruments, the detector output is recorded by a computer which can automatically control things such as the applied voltage and allowing scheduling of several samples. Injection of samples and cleaning of the capillary is usually performed by pressure-driven flows.

The majority of commercial CE systems use absorption detection of some type. The problem with this, is that the path length of the detection will be short because of small diameter of the capillary. A commonly used solution is a “bubble,” an enlargement of the capillary at the detector which contributes to a longer path length for detection.

1.3.1 Electroendoosmosis

Electroendoosmosis provides the pumping mechanism in HPCE. The capillary walls are coated with a charged surface, in almost all cases fused-silica. The capillaries are conditioned with a strong base, usually sodium hydroxide, before being used for a run. The bases ionises silanol groups leaving a negatively charged coating that attracts cations and repels anions. In standard conditions, the inlet is anodic so the cations migrate towards the outlet. The cations drag fluid causing a net flow in the capillary. There is interest in this type of mechanism for use in microfluidic devices. The electroosmotic flow is given as:

$$\nu_{eo} = \frac{\varepsilon\zeta}{\eta}E \tag{1.2}$$

where ε is the dielectric constant of the buffer, ζ is the zeta potential of the liquid-solid interface and η is the viscocity of the buffer and E the electric field across the capillary. An advantage of the use of electroendosmotic flow, rather than pressure-driven flow as in many liquid chromatography systems, is that the flow generated is nearly uniform across the capillary except for very near the walls, rather than parabolic flow where velocity is greatest at the centre. This advantage can be offset somewhat, depending on setup, due to ohmic heating from the high electric fields. The temperature will be greater in the centre of the capillary, which for most fluids leads to a lower viscocity and faster flow rate.

1.3.2 Electrophoretic Mobility

The electroosmotic flow provides the overall force to drive species along the capillary. The separation in CE is due to the different time mobility of the specifies involved, which is primarily dependant on the charge/size ratio. The drag force of an approximately spherical species is given approximately by Stokes' law:

$$F = 6\pi\eta r\nu \tag{1.3}$$

where η is visoscity, r is ionic radius and ν is the velocity of species. Applying a voltage the force due to a charge q on a species will be:

$$F = qE \tag{1.4}$$

and defining the electrophoretic mobility as:

$$\mu = \nu/E \tag{1.5}$$

gives the electrophoretic mobility:

$$\mu = \frac{q}{F} = \frac{q}{6\pi\eta r} \tag{1.6}$$

In practice, what is measured is not the electrophoretic mobility but the mobility time, which is the time before the species is detected at the detector. Two lengths are important in describing a capillary, the actual capillary length L, which is the physical length between the two buffers and the effective length l, which is the distance from the injection side of the capillary to the detector. Using (1.5) and since the voltage applied is between the two buffers:

$$E = V/L \tag{1.7}$$

the mobility time of species is:

$$t = \frac{lL}{\mu V} \tag{1.8}$$

Since the actually mobilities are given by:

$$\mu_{obs} = \mu + \mu_{eof} \tag{1.9}$$

then the mobility of a species is calculated by:

$$\mu = \mu_{obs} - \mu_{eo} = \frac{lL}{V}\left(\frac{1}{t} - \frac{1}{t_0}\right) \tag{1.10}$$

There are many further complications and techniques used in practice. While very useful, HPCE also requires some effort to generate repeatable conditions, particularly in conditioning the capillary. In summary, HPCE provides an elegant experimental platform for the separation and detection of a wide range of molecules and avoids many disadvantages of other techniques. However, it is worth remembering:

> There are only three problems with capillary electrophoresis: injection, separation and detection.[1]

As with most separation techniques, the devil is in the details, [25] provides a good introduction to the practical details of using HPCE.

1.3.3 Investigation of homogalacturonan methyl-esterification patterns using HPCE

Capillary electrophoresis has been used as an important method for characterising homogalacturonan. For oligomers longer than about DP 15 CE can be used to obtain a DM distribution. This is due to the symmetrical scaling of charge and hydrodynamic friction which occurs above about DP 15. The polymers are detected using absorbance at 191 nm, and normalising the absorbance for the length of DP with an exception for monomer [27]. Another group claimed that blockiness of the homogalacturonan also had an effect on electrophoretic mobility [28], but further work [16, 29] showed that μ was related linearly to DM between about 25% to 80%.

For short oligomers of homogalacturonan, such as fragmentation results, CE has been used to investigate the methyl-esterification pattern, since the charge and DP both have a marked effect on low DP oligomers [18, 30]. Identification of particular isomers with particular peaks is not yet routine, and the technique does not separate oligomers with the same number of methyl-esterified residues but in different positions, although knowledge of the enzyme rules involved in the digest often allows the isomer to be determined.

[1] Quote from a short course by Dr. Olechno as cited in [25].

1.4 Previous attempts at simulation

The use of simulation based on enzyme rules to understand information about the results of fragmentation are not common, but have been used previously [31]. An example nearest to the efforts outlined in this report simulated fragmentation of galactomannans (another copolymer system) with an endo-mannase [32], although the calculation of enzyme binding was according to specific enzyme rules, not probabilistic. In this case, there was reason to believe the polymers were constructed according to a 2$^{\text{nd}}$ order Markovian process, so simulation was used to construct Markov chains in silico and fragment them according to a simple enzyme model. The comparison between simulation and experimental fragmentation results was used to assign 2$^{\text{nd}}$ order Markov parameters to galactomannans from several different plant species. It was notable that each species had characteristic Markov parameters. A later attempt to model alginate fragmentation by a lyase used a very different model leading to six coupled nonlinear differential equations whose parameters were obtainable from kinetic data [33] and this model had good agreement with digestion data. The use of subsite mapping to characterise enzymes has also been around for quite some time [34, 35], usually using the relative kinetics of different substrates to glean subsite energies, but the energies have not been previously applied to simulating digest timecourses as far as the author is aware.

1.4.1 Previous models of endo-PG II digestion

Most models of endo-PG II are subsite models, meaning that they model the interaction of the enzyme with a substrate as interacting with an array of subsites, each subsite binding an individual monomer according to certain rules. One of the earliest proposals of a subsite model for endo-PG II [36] (although the enzyme is at this time referred to as *A. Niger* extracellular endopolygalacturonase, this thought to have consisted primarily of endo-PG II), attempted to model the enyzme as consisting of 4 subsites with the enzyme postulated to cleave between the 3$^{\text{rd}}$ and 4$^{\text{th}}$ residues as counting from the non-reducing end of the substrate. This was based on early kinetic data and product analysis of short oligogalacturonides. No attempt was made at giving a quantitative answer to the relative affinities of different subsites. This was termed the 4,3 model.

Shortly after detailed structural information became available and in particular bond-cleavage frequencies of endo-PG II were accurately measured [37], an attempt was made to use computer simulation to predict digest results and allow a model of endo-PG II to be compared with experimental results [38]. This was a direct predecessor of the work outlined in this report. Two models were compared with experiment. The 4, 3 model just described and another model based on results which indicated a 7 subsite enzyme [37], with the enzyme cleaving between the 5$^{\text{th}}$ and 6$^{\text{th}}$ residue counting from the non-reducing end. The experimental data of cleavage frequencies was only for unmethyl-esterified homogalacturonan and so an ad-hoc method of dealing with methyl-esterified residues was made. These consisted of 3 rules, (i) subsites between which the incision occurred had to be unmethyl-esterified based on evidence from [39], (ii) three subsites out of the remaining 5 must contain unmethyl-esterified residues and (iii) the tetramer is an exception and can successfully bind to the enzyme. This model, did not attempt to account for the digestion of trigalacturonic acid which is known to be digested, but occurs on a timescales significantly longer than digestion of the tetramer so it can be treated separately. This work showed the 4,3 model was not a good description of the action of endo-PG II and also showed that the newly proposed model generated better agreement with experimental results for the amount of mono, di and trigalacturonic acid released on completed digestions and also for calculating digest patterns.

Subsite Energy Binding Model

The subsite binding model of enzymes is a commonly used model for enzyme binding [40]. It models an enzyme binding with substrate as an array of subsites, each with an individual binding energy that binds to a specify substrate monomer unit. Usually, the subsites are numbered with positive numbers to the non-reducing end (left) and negative numbers to the reducing end (right). Mapping the subsites of an enzyme is commonly performed experimentally by bond cleavage frequencies using kinetic methods or product analysis.

The calculation of the free energy difference between the binding of two substrates is given as:

$$\Delta G_{12} = -RT \ln \left(\frac{P_1}{P_2} \right) \tag{1.11}$$

where R is the gas constant and T is temperature, and P_1/P_2 is the ratio of the two different binding probabilities as calculated from the kinetics or cleavage frequencies. Thus using experimental results on bond cleavage frequencies and/or rates this can be used to calculate the various subsite energies.

Experimental work on endo-PG II digestion of oligomeric fragments of unmethyl-esterified homogalacturonan lead to a a -5 to +2 (7 subsite model) for endo-PG II [37], but these authors did not use the relative rates to fill in subsite energies for those that are unavailable from bond cleavage frequencies alone.

Chapter 2 - Experimental Details

The bulk of the work detailed in this report involved both computer simulation and experimental methods. There were three stages to the work outlined. The first, was modelling the digestion of unmethyl-esterified homogalacturonan with endopolygalacturonase II (endo-PG II) from *Aspergillus niger* using a subsite model for the enzyme, and quantification of the time-course of the digestion experimentally using HPCE [41]. It was then attempted to extend this model to describe digestion of partially methyl-esterified homogalacturonan, a copolymer system, and this was tested using known digest results. Finally, the the enzyme model was used to test algorithms for reconstructing the fine-structure of the homogalacturonan samples based on results from fragmentation with enzymes and analysis of the resulting digests using HPCE. While all work was related to this particular system, the problem of predicting enzymatic digests and using knowledge of the enzyme activity to aid in reconstructing the fine-structure is of more general interest.

2.1 Simulations of Homogalacturonan Digestion

Many biological polymers are copolymers with a variety of enzymes involved in modifying the polymers and, in most systems, the fine-structure is of interest, and not easily determinable. Therefore, although the system used in this report is homogalacturonan digested with endo-PG II, many of the results are expected to be generalisable to other systems.

Homogalacturonan was chosen because of it's availability, it's importance in the plant kingdom and the food industry, it is unbranched, consists of only two types of monomers and the availability of well-characterised enzymes which operate on it.

2.2 Simulation of Enzyme Degradation

Simulation is an all encompassing word, that in general, refers to the use of computers to execute an algorithm that in some way mimics a physical system of interest. Therefore it is necessary to describe what level of detail is meant by simulation of enzyme degradation in this report. The most detailed level would be to solve the Schrödinger equation for all particles (including explicit solvent molecules) of the enzyme and substrates. However, this would require computational power many orders of magnitude greater than even the most powerful supercomputers available and the simulation would need to be repeated many times to give an average digest pattern. Therefore it is necessary to choose a more grainy level of simulation. It is hoped (see section 3.4.2) that in the future more detailed calculations at a molecular level (although still not full ab-initio studies) might shed light on binding energies that are difficult to find experimentally.

The simulation outlined in this report is at what might be called the "informational" level. The simulation does not attempt to utilise knowledge about the physical structure of polymers, instead it treats a polymer as a mathematical string, with each character in the string representing a monomer of a certain type and whether that monomer is the non-reducing end, reducing end or middle of a polymer or is an unpolymerized monomer. It is important to note that many biological polymers have a direction associated with them.

Since, as far as the author is aware, there is no other published software working in this manner, a library named BJPS was developed to function as a "toolkit" for this work. The library is written using C++ and also has an interface with the Python scripting language commonly used in scientific research particularly bioinformatics. The GNU Scientific library was used to provide high-quality random numbers using the "Mersenne Twister" algorithm [42]. This random number generation was more computationally expensive than other, simpler random number generators, however, in tests it was shown to generate low probability events at the correct statistical frequency better than other algorithms. This is important as will be shown shortly. The library was tested using Intel's C++ compiler, although it should in principle be compatible with any modern C++ compiler/architecture. It was tested on both x86 and EM64 architectures. Riverbank's SIP was used to generate the Python interface. Futher information on the availability of these tools and user documentation of the library is provided in appendix B. Although the library was designed with performance issues in mind, it has not been extensively optimised. See also section 3.4.4.

2.3 Simulation of unmethyl-esterified homogalacturonan and detailed comparison with experiment

The model of the enzyme used for simulation was the binding energy subsite model previously described in section 1.4.1. The total binding energy for a substrate is determined by the individual binding energies of each subsite. Experimental data [37] lead to a -5 to +2 (7 subsite model) with subsites numbered from the non-reducing end where the enzyme cleaves between -1 and +1 as shown in figure 2.1. These results are also consistent with separately measured bond cleavage frequencies elsewhere [43, p. 54] (Note: Some authors were involved in both studies).

The distribution of starting material was used to populate an array representing the polymer as strings as described earlier. Enzyme encounters were assumed to occur at random at any point on the array. A position for subsite -1 was chosen and the binding energy for this encounter was calculating by summing the binding energies for all filled subsites. After a check to ensure +1 is filled (otherwise this is a pointless binding) this was then used to generate the probability of a successful binding by the inverse of (1.11):

$$P = \exp\left(-\frac{E_s - E_7}{RT}\right) \tag{2.1}$$

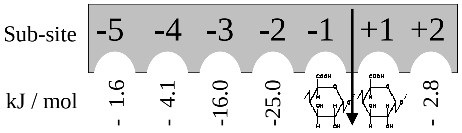


Figure 2.1: The model of endo-PG II subsite architecture showing the postulated binding energies associated with each position. Energies for subsites -5, -4 and +2 came from studies of bond cleavage frequencies of short oligomers and -3 from the relative kinetics of tetramer and trimer degradation and that of -2 chosen to reflect that the trimer always binds filling that position so it must be significantly preferred subsite over +2. Data used for calculations was [37,43], although neither source used relative kinetic rates to determine subsite energies. No attempt is made to find binding energies for subsites +/- 1 as this is unnecessary (see text). Figure previously published in [41].

where E_s is the total binding energies for all occupied subsites and E_7 is the energy for 7 filled subsites. This normalises the probability so that for 7 (all) subsites filled the probability of a successful binding is unity. Because of this normalisation there is no need to experimentally find the binding energies for subsites -1 and +1 (since they are always filled) or the absolute binding energies. Once the probability was determined, success of the binding was decided by comparing with the output of the random number generator. Upon successful bindings the array contents at subsites -1 and +1 were updated to indicate a scission. After a selected number of iterations the array elements were interrogated to produce a distribution of fragment lengths. The iteration count was decremented regardless of whether successful binding occurs, so that competitive inhibition is correctly simulated.

The first concern for the simulation of enyzmatic digestion was the repeatability of the simulation because it is a stochastic simulation. The initial input consisted of the subsite binding energies and the relative number of molecules of each DP for the starting material. The actual number of molecules generated in silico and simulated was varied between 10^2 and 10^8 and the simulation was carried out for a certain number of timesteps, chosen based on preliminary experimental data, such that the digest was nearing completed digestion. The number of timesteps was scaled linearly with the number of molecules. Each setup was simulated 10 times (with the random number generator reseeded each time based on the system time) and resulting pattern of oligomers of each DP existing near completion was obtained and normalized to the same total number of monomers. From this a mean pattern was calculated and variances were calculated between each member of the set and the mean to determine the average variance σ^2 in each set. The variance was used as a measure of the repeatability of the simulation.

The simulation was then performed and the simulation pattern was stored at regular intervals. After initial comparisons of experimental and simulated digest patterns by eye appeared to indicate good agreement, a simple search algorithm was performed to objectively match simulated and experimental digest patterns. The relative numbers of oligomers of each DP was normalised between both simulated and experimental results so that the total number of monomers of galacturonic acid was the same. Then the σ^2 distance between simulated and experimental results was calculated, allowing the experimental results to scale to minimise the distance to allow for the possibility that not all material had been experimentally detected. Using this metric, for each experimental timestep, a simple comprehensive search was made among the simulated patterns to find the best match.

It was then important to understand how sensitive to the binding energy parameters our model was. How much variation in the postulated binding energies of the various subsites is possible before it results in demonstratibly inferior results. Firstly, by comparing simulations with simulations the inherent algorithmic sensitivity was found. Twelve time steps in the simulation were chosen that represent experimental timesteps of interest. The binding energies were adjusted and the simulation repeated, this time calculating the digest pattern at regular intervals and matching the new simulation to the old as previously described for comparing experimental results. The distance between the best match of the base case to new simulation digest pattern was recorded. Each binding energy was adjusted in 1% increments between -50% and +50% of its calculated value to determine the individual sensitivities of each subsite.

Once the inherent sensitivity of the model was known, it was also important to find how clearly the binding energies could be verified from the experimentally obtained results. This is somewhat more complicated, as the CE digest patterns have experimental uncertainties on the order of 10% which tends to flatten or distort the variance curves. It was hoped to find if the calculated energy values were a minimum in comparison with experimental results. The first attempts involved the use of the Levenberg-Marquardt optimisation algorithm [44] and a foraging search algorithm [45] to search the entire combinatorial space in some efficient way for a best match, however, preliminary attempts proved unsuccessful due to the noise of the simulation and the effective quantisation of distance in this space (changing the binding energy of a subsite by a very small amount in order to calculate the gradient has no effect discernable from the noise). With five binding sites, the total combinatorial space is too vast to search completely, so a simple approach was chosen where each subsite energy was changed in isolation between -50% and +50% of the original value in 1% increments and the results compared with experiment.

2.3.1 Experimental Details

Because the unmethyl-esterified homogalacturonan contains only one type of monomer, the final digestion products contain relatively little information about the action of the enzyme. Therefore, it was necessary to generate a picture of the time-course of the enzyme digestion for detailed comparison with the model of enyzme digestion.

Randall Cameron, a collaborator in the USA kindly provided us with a well-characterised sample of oligogalacturonides with degree of polymerization between 2 and 17 generated by selective precipitation of partially digested polygalacturonic acid [46].

Briefly, a 2% (w/v) solution of the free acid of PGA in 50 mM lithium acetate at pH

> 4.7 was digested with 0.05 U mL^{-1} EPG (Lot 00801, Megazyme International Ireland Limited, Bray, Ireland) for 4.5 h at room temperature with constant stirring. The pH of the digested LiPGA was lowered to 2.0 with concentrated HCl and stored overnight at 4 ℃. The precipitate was pelleted by centrifugation at 23,500 xg for 30 min at 4 ℃. The pelleted precipitate, representing high DP oligomers was removed.The supernatant, containing the low- and medium-DP oligomers, was brought to 50 mM sodium acetate (NaOAc) and 22.5% EtOH and then placed at 4 ℃ overnight to precipitate the medium-DP fragments (DP 8-24). Following centrifugation as described above, the supernatant was decanted. The pelleted material was solubilized in 50 mM LiOAc and then re-precipitated by adjusting on the solution to 50 mM NaOAc and 22.5% EtOH. This material, representing the medium-DP oligomers, was centrifuged again and the pellets were solubilized in 50 mM LiOAc and stored at 4 ℃. Initial characterisation was carried out using HPAEC with an evaporative light scattering (ELS) detector and subsequently by capillary electrophoresis using UV detection. [41, p. 1697]

The endo-PG II from *A. niger*, previously prepared as described elsewhere [47], was used for the digestion. Digestions were performed by:

> Digests were carried out by incubating 1.0 mL of the oligogalacturonide mixture, at a total concentration of 3% galacturonic acid, and pH 4.2, with 20 μL of the enzyme solution, that was in turn generated by diluting 25 μL of a 7.5 mg mL^{-1} 1 protein stock into 2.0 mL of 50 mM acetate buffer at pH 4.2. All experiments were carried out at (301) ℃ by keeping the digest mixture in a waterbath. At various times aliquots were removed, the enzyme denatured by rapid heating to 95 ℃, and the current concentrations of the various oligomeric species recorded using CE. Thus a picture of the time-course of the digestion was recorded. [41, p. 1697]

Capillary electrophoresis was used along with the previously described HPAEC light scattering to identify the starting oligogalacturonides.

> Experiments to separate, identify and quantify oligogalacturonides of varying degrees of polymerization were carried out using an automated CE system (HP 3D), equipped with a diode array detector. Electrophoresis was carried out in a fused silica capillary of internal diameter 50 μm and a total length of 46.5 cm (40 cm from inlet to detector). The capillary incorporated an extended light-path detection window (150 μm) and was thermostatically controlled at 25 ℃. Phosphate buffer at pH 7.0 was used as a CE background electrolyte (BGE) and was prepared by mixing 0.2M Na2HPO4 and 0.2M NaH2PO4 in appropriate ratios and subsequently reducing the ionic strength to 90 mM. At pH 7.0 galacturonic acid residues are fully charged and while the oligomers are susceptible to base-catalysed beta-elimination above pH 4.5, no problems were encountered during the CE runs of some 20

> min at room temperature. All new capillaries were conditioned by rinsing for 30 min with 1 M NaOH, 30 min with a 0.1 M NaOH solution, 15 min with water and 30 min with BGE. It was found that for the samples used in this study similar harsh washing of the capillary was also required between runs. Detection was carried out using UV absorbance at 191 nm with a bandwidth of 2 nm. Samples were loaded hydrodynamically (various injection times at 5000 Pa, typically giving injection volumes of the order of 10 nL), and typically electrophoresed across a potential difference of 20 kV. All experiments were carried out at normal polarity (inlet anodic) unless otherwise stated. Samples of mono-, di-, and tri-galacturonic acid, used as standards, were obtained from Sigma-Aldrich Corp., St. Louis, MO, USA. [41, p. 1697]

Raw CE absorbances were processed to quantify the amount of different oligomers present as described elsewhere [27]. The peak area was divided by the migration time as described in section 2.3.1. Peaks were identified by spiking with commercially available samples of dimer and trimer. CE has been previously used in the study of oligogalacturonides and it is known that above a certain DP the hydrodynamic friction and charge scale symmetrically rendering a loss of resolution at higher DPs. The exact at which DP this occurs is part of ongoing work, but is predicted to be in the DP 15-20 range [48] hence the choice of the starting substrate just under this range, ensuring it would be possible to record the full time-course.

2.4 Extending the endo-PG II model to incorporate partial methyl-esterification

The first step to simulating a copolymer digestion pattern is the creation of a digital representation of the starting polymer material. The degree and width of methyl-esterification distribution can be measured using CE as outlined in section 1.3.3 and the degree and width of polymerization is usually also known. Usually both the DM and DP distributions are assumed to be Gaussian (although other distributions could easily be incorporated). As a trade-off between consistent results and computational time, somewhere between 1000 to 10000 polymers were made in silico for the copolymer digestions described here. Gaussian distributions of the desired DM and DP distributions and with the desired total number of polymers are calculated and each DP entry is randomly paired with a DM entry. The chain is then created with this DM and DP. At the moment, primarily for lack of a better understanding of the creation of pectin, the only method of chain creation implemented is by specifying Markov parameters (0$^{\text{th}}$ order Markov can be thought of as Bernoullian). Markovian statistics were described earlier in section 1.1.1. While it is known that galactomannans are synthesised in a manner that is well described as a 2$^{\text{nd}}$ order Markov process [32], there is no particular reason that pectin synthesis should be a Markov process, particularly as it is known that post-synthetic modifications occur, but it was chosen because Markov processes are well understood, easily characterised and approximate many physical processes. It is expected that results should be generalisable to other models.

The most obvious extension to the subsite model for incorporating methyl-esterified residues will be termed the extended subsite model, and consists of associating two subsites energies with each subsite as shown in figure 2.2, one for when the subsite is occupied by an unmethyl-esterified residue and one for occupation by a methyl-esterified residue except for subsites +1 and -1 which are known to require unmethyl-esterified residues for binding to occur [37]. In principle, the most straightforward way to test this model would be to synthesize oligogalacturonides with DPs 1 to 7 and with various specific methyl-esterification patterns and to use kinetic and product analysis of the oligomers to calculate the methyl-esterified binding energies similar to that done for the unmethyl-esterified version [37]. However, due to immense difficulties in synthesis of these oligomers, this path appears to be infeasible with current methodologies. One group has synthesised patterned trimer and produced kinetic data [39] for endo-PG II digestion of these trimers. These results can be used to give constraints (assuming the extended subsite model is correct) using (2.1) for the methyl-esterified energy values of subsites +2 and -2 relative to unmethyl-esterified subsite -2. This still leaves 3 unconstrained parameters.

The experimental data for partially methyl-esterified homogalacturonan digestion is limited to completed digestion data except for some recent timecourse data (A. Cucheval, unpublished data) which is not included in this report, but is expected to be of use in the future. An attempt was made to find constraints for the remaining 3 methyl-esterified subsite energies using obtained digestion data which is shown in appendix A. The primary data used for comparison was the fragmentation results of 30% randomly methyl-esterified pectin (A. Cucheval, unpublished data). Initial attempts to find these energies consisted of educated guesses (based on what was known of the amino acids involved in the enzyme subsites).

Since previous schemes for searching the parameter space according to some global search algorithm had proved unsuccessful, and it was considered likely that there was no good solution for this simple model, a grid search was performed where the methyl-esterified binding energy parameters for subsites -4 and -3 were ranged over the range -50 to +50 kJ/mol in 1 kJ/mol increments and subsite -5 was ranged over -10 to 10 kJ/mol. For each position, a simulation was performed with these energy values and the scaled variance between the simulated isomer fragments and experimental digest results (A. Cucheval, unpublished data) was calculated in a similar manner to the previously described DP variances. The grid search also had the advantage of allowing the surface of the parameter space to be visualised.

2.5 Comparison of Reconstruction Strategies

Because a completely verified model of endo-PG II digestion of partially methyl-esterified homogalacturonan was not fully completed, the extended subsite model with the energies shown in figure 2.2 was chosen as the basis for study of reconstruction. Although this model is known not to correctly reproduce all experimental results, it generates the correct type of end fragments and was used as the

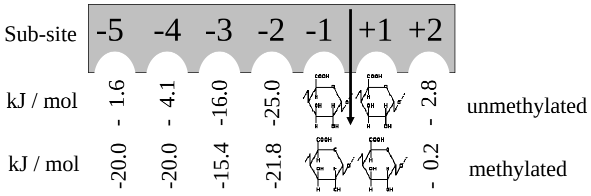


Figure 2.2: The extended subsite model, where each subsite (except -1 and +1 which must always be occupied by unmethyl-esterified residues for a successful binding) has two energy levels associated with them, one for the binding energy contribution when occupied by an unmethyl-esterified residue and one for when occupied with a methyl-esterified residue. This gives the model a total of 10 independent parameters. The methyl-esterified binding energies for subsites -5,-4 and -3 are estimates.

basis for reconstruction techniques.

The goal of reconstruction, is to generate maximal information about the original copolymer fine-structure based on the final digestion pattern and a model of the enzyme digestion. For the purposes of verifying the reconstruction techniques, no attempt was made to use experimental results, rather various chains of specified Markovian composition were generated and fragmented using simulation. The reconstruction of the chains proceeded from the simulated fragments and the reconstructions were compared with the original chains. The metric used for comparison was the σ^2 distance between diad frequencies - which correspond to the 1^{st} order Markovian parameters. Although the enzyme model used in this reconstruction was intended to describe endo-PG II, is it believed that these results may be of more general interest, and the generalisability of each construction strategy will be noted as it is discussed. In particular, it is likely that the results could be used for other endo-PG enzymes.

Four different reconstruction strategies were devised and compared with one another. The baseline strategy, designated random, simply reconstructed by randomly reassembling the fragments into chains of the appropriate length. The random strategy, requires no information about the mode of fragmentation and therefore is a completely generalisable and computationally efficient reconstruction strategy. For the reconstruction attempt, the algorithms were given the DP distribution and DM distribution of the original polymer set and the final digestion pattern. Experimentally all this information would be available via CE or other means. However, experimentally the final digestion pattern is only discernable for the low DP fragments, so reconstruction algorithms were tested under the condition of being allowed access to the full set of digest oligomers, only those less than DP 25, and only those less than DP 15, corresponding roughly to the ideal situation, the best resolution that seems achievable with improvement on current techniques and what is at the limits of current routine experimental methods.

Another strategy, designated overlap, reconstructed chains by randomly picking the first fragment and placing it to the far left of the desired reconstructed chain, and thence forth finding chains with the most overlap between the growing (right) edge of the chain and appending the remainder of the overlapping chain onto the end. The thought behind this strategy, grandly named the pseudo-carbonome idea[1], postulates that since each pectin chain in a given specimen can be considered a sample from the same statistical generator, if the generator followed simple rules likely to give repeating sequences then it may make sense to consider the resulting fragments as overlapping, even though the fragments come from many different chains.

The endo-PG II reconstruction strategy was the most direct approach in incorporating knowledge of the enzyme digestion model to aid reconstruction efforts. After randomly picking the first fragment and placing to the left of the intended reconstructed chain, the chain is subsequently extended by randomly choosing the next fragment and calculating the Boltzmann probability of cleavage occurring at the join according to the extended enzyme model. The output from the random number generator is then used to decide if this join is accepted and the processes is continued until the correct length is obtained. It is worth noting at this point, that this model will not (and does not attempt to) generate the time-order reverse of the digestion sequence, because during the time-course of digestion the most probable bindings will occur early on and as the simulation progresses successively less probable bindings will begin to form part of the digest course. This reconstruction strategy favours probable bindings occuring early in the reconstruction, corresponding to late in the digestion, which means the reverse sequence will not be generated. It is not clear exactly how best to mimic the reverse digestion.

Finally, an approach called matching[2] was used. In this, chains with similar DP and DM distributions over the entire range of Markovian input parameters were generated and their digestion simulated. These results were stored and compared with the digestion patterns of the test chains by means of a comprehensive search. The best match digest was then considered the reconstructed polymer. This method differs from the others in that the actual fragments of the test case are not directly used for reassembly, but only to match seperately generated chains.

[1] I am pleased to report this terminology is not my own, but attributable to Dr. M. Williams, my supervisor.

[2] Unfortunately, this is my own imaginative terminology.

Chapter 3 - Results and Discussion

3.1 Simulation of unmethyl-esterified homogalacturonan and detailed comparison with experiment

Figure 3.1 shows the quantification of the starting distribution of oligogalacturonides. There is good agreement between the characterisation by HPCE and HPAEC. This distribution of relative numbers of oligomers of different lengths was used as input to the simulation.

Figure 3.2 shows the results of determining the variance for identical starting conditions to find simulation sensitivity. It is clear the variance scales roughly inversely to the number of chains involved. Note that the memory and processor time requirements scale roughly linearly with the number of chains. From these results it was decided that 10^4 chains provided a convenient compromise between repeatability and computational resources. With 10^4 chains a single simulation of complete digestion took on the order of seconds, however, as previously described, many simulations had to be carried out for certain results.

Experimental data was obtained for the enyzmatic digestion of the sample at 0, 2, 4, 6, 8, 10, 15, 30, 45, 266 and 13000 min after addition of the enzyme, as described in section 2.3.1. A digest simulation was performed, using the quantisation of the starting material. The digest pattern was stored after each step of 200 iterations in the simulation up to a total of 6×10^6 iterations. At this point in the digest there is only trimer, dimer and monomer remaining, with the trimer being slowly digested. Experimentally, it is known the trimer should be slowly digested, but at a rate 20 times slower than tetramer degradation [36] so this allows the trimer degradation to be uncoupled and treated separately. In many experimental studies the ratio of monomer, dimer and trimer is taken to be the practical endpoint of the digestion. It is noteworthy that the simple binding energy parameters faithfully reproduce this result as a natural consequence of the difference in binding energies.

Figure 3.3 shows the results from matching the experimental data to simulated data, along with a least-squares linear fit line. A clear linear relationship up to 45 min is clearly seen which shows that despite the lack of sophistication and molecular details in the model, the subsite energies do well-describe the time dependence of the real reaction. Matching later time steps precisely is problematic because the digestion pattern is changing very slowly and experimental data contains noise. Figure 3.4 shows the variance between experimental data and simulated timesteps. It is clear it that is possible to narrowly match early timesteps where the digest patterns are distinctive but later timesteps are problematic. The slope of this fit gives timesteps per minute, allowing simulated timesteps to be mapped onto experimental time.

Figure 3.5 shows the comparison of experimental and simulated data at each recorded experimen-

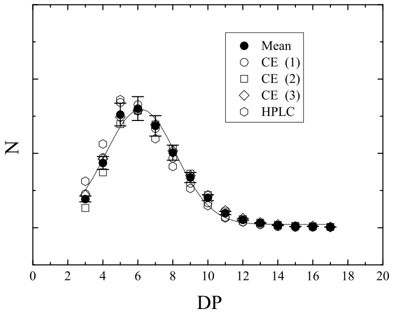


Figure 3.1: The relative number of molecules of oligogalacturonides of different degree of polymerization for the starting distribution. Measured using CE and HPAEC. Figure previously published in [41].

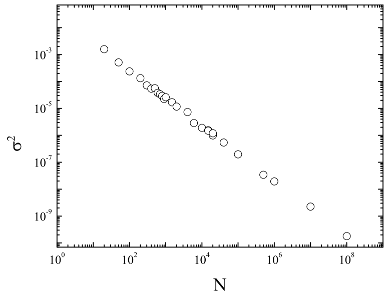


Figure 3.2: The variance within 10 repeated simulations from identical starting conditions (except, of course, random number generator seed) and carried out with varying number of starting chains. Figure previously published in [41].

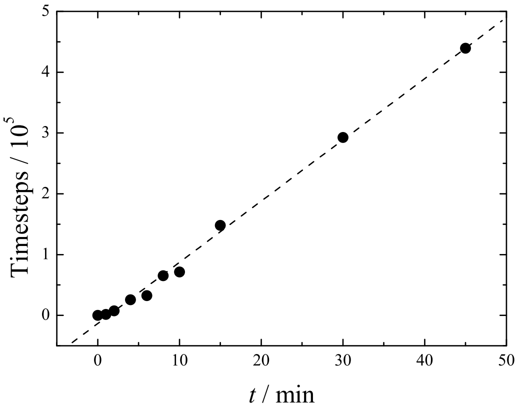
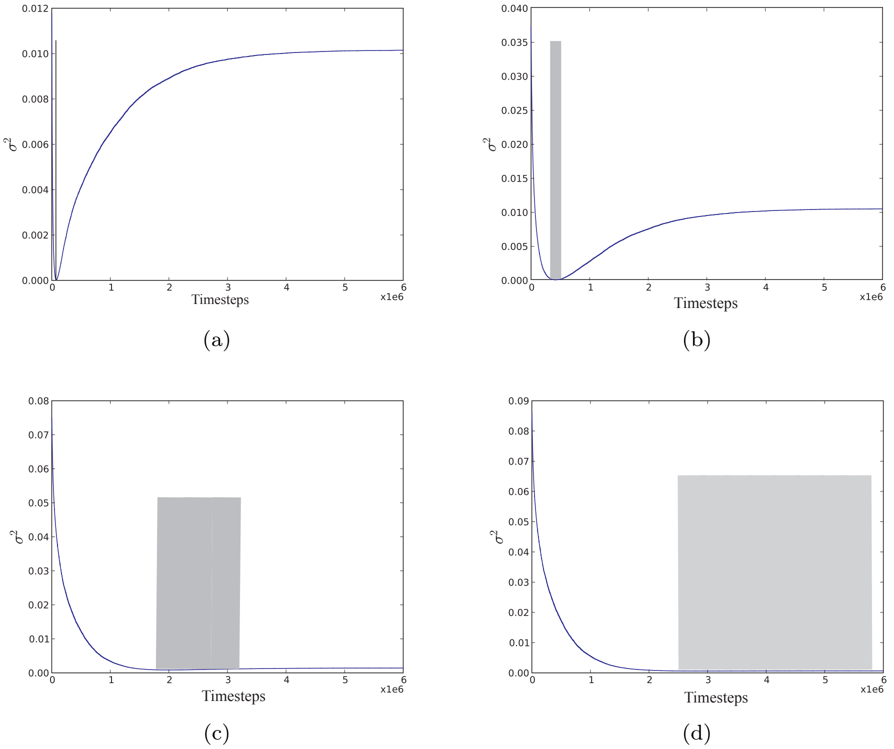


Figure 3.3: Number of simulated iterations versus experimental timestep it best matches. Searches for best matches were performed as described in text. Figure previously published in [41].

tal timestep and it's best-matched simulated counterpart. This shows good quantitative agreement between the two results. Since number of molecules is what is shown in the agreements, it is important to remember that experimentally what is actually recorded is absorbances which, with the exception of the monomer, is proportional to the degree of polymerization, meaning small amounts of low DP oligomers are experimentally difficult to detect.

The variance between digest patterns generated with varying subsite energies is shown in figure 3.6. It is clear that early on in the digestion, changing the subsite energies makes little difference to the obtained digest patterns. This is because at the early stages of digestion, most attempted bindings fill the subsites completely and so the relative energies have no effect. However, by 10 min into the simulation, it is clear that the energies for subsites -5, -4 and +2 are having a marked effect on the digest pattern. In order to generate significant change in the pattern the subsite energies must be shifted by about 10%. The inner subsites -2 and -3 don't play a major role in determining the digest pattern during this time, as expected, because most substrates are larger than tetramer at this stage. Nearer the end of the digestion the outer subsite energies become less important, while subsite -3 becomes more sensitive as only low DP oligomers are remaining. Since the digestion is stopped before the binding of trimer becomes important, the binding energy of subsite -2 never has much effect.

Figure 3.7 shows the predicted digest match with experimental results of 10 min. This plot shows that the three outer subsites are well-constrained by the experimental data and were they to differ by more than 10% from optimum value then the match would be detectably worse. As expected from the simulation sensitivity experiments, the outer subsites are the ones most constrainable from matching with digest patterns.



(a)

(b)

(c)

(d)

Figure 3.4: The variance σ^2 between experimental digest pattern at specific times compared with each recorded step of the simulated digest pattern. Results are shown for (a) 10 min, (b) 45 min, (c) 266 min and (d) 13000 min. The gray areas in each graph indicate the portions of simulations steps that can be considered a good match. It is clear that while a clear minimum is found for 10 and 45 min, due to the slow change in digest patterns in the later timesteps matching becomes problematic.

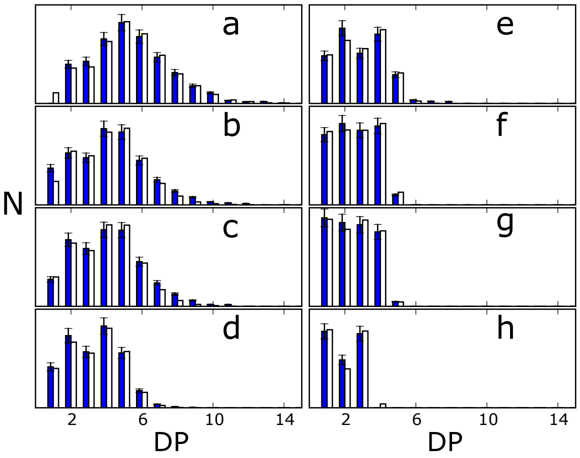


Figure 3.5: Comparisons of experimental and simulated digest patterns for select times during the digestion of oligalacturonides with endo-PG II. The experimental data is the filled bars, and estimated uncertainties are shown. Times shown are (a) 2, (b) 4, (c) 6, (d) 10, (e) 15, (f) 30, (g) 45, (h) 13,000 min. Figure previously published in [41].

3.2 Extending the endo-PG II model to incorporate partial methyl-esterification

The global search over of the methyl-esterification energies of subsites -5, -4 and -3 was at the limits of what is feasible using a single PC for optimisation, taking several weeks of computer time. However, it provided quite conclusive results. The experimental results used for matching are included in appendix A. Figure 3.8 show isosurface plots of the variance at different limits. It is clear that there exists a minimum within the parameter spaced searched because as the isosurface is shown for lower variances, it converges on one area of the parameter space which is not at the boundary of the search space. It would seem unlikely that a better fit would exist outside the parameter range searched as it was quite extensive and the optimum was found well away from the edge of the search space. The minimum was found with the subsite energies being 0, 16, -28 kJ/mol for the methyl-esterified binding energies of subsites -5, -4 and -3 respectively. Figure 3.9 shows a slice of the space taken where subsite -5 has methyl-esterified energy 0.

Using these best-fit parameters the extended subsite model was tested against known results. Figure 3.10 shows a comparison of the isomers generated from simulation and experimental fragmentation. It is clear that although the extended subsite model can reproduce the correct isomers, a

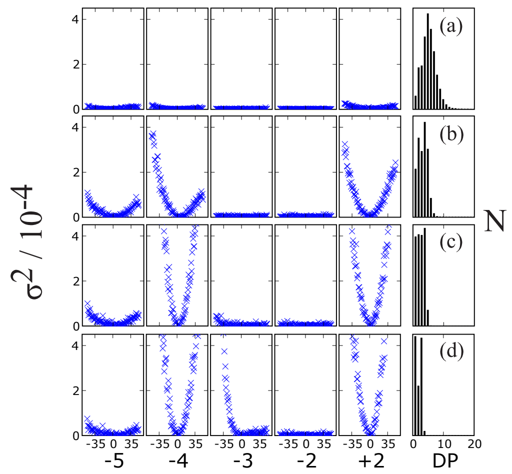


Figure 3.6: The variance between digest patterns generated with the best estimates of subsite energies (figure 2.1) compared with variations between -50% to +50% of the value. Note: simulation is being compared with simulation in this figure, not experimental results. (a) 2, (b) 10, (c) 30, (d) 13,000 min. Figure previously published in [41].

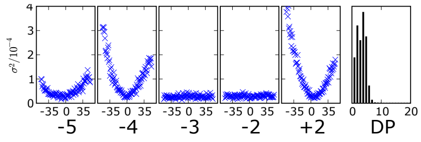
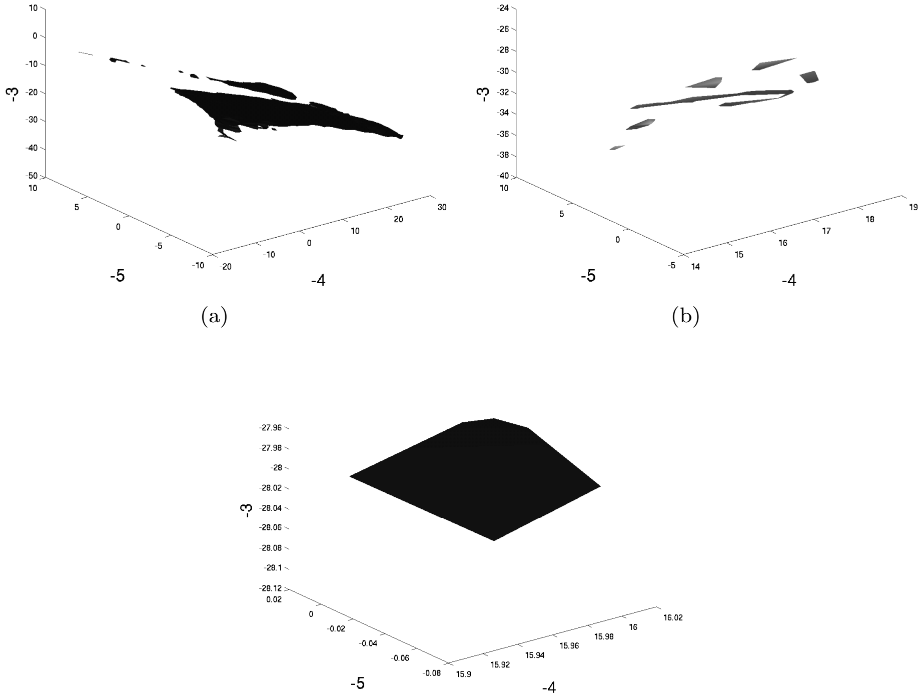


Figure 3.7: The variance of predicted digest patterns, calculated with varying binding energies as described in the text, compared with the experimental digest pattern observed at 10 min. The plot shows the match with experimental results would be detectably worse if the binding energies of sites -5, -4 and +2 were modified more than about 10%. Figure previously published in [41].

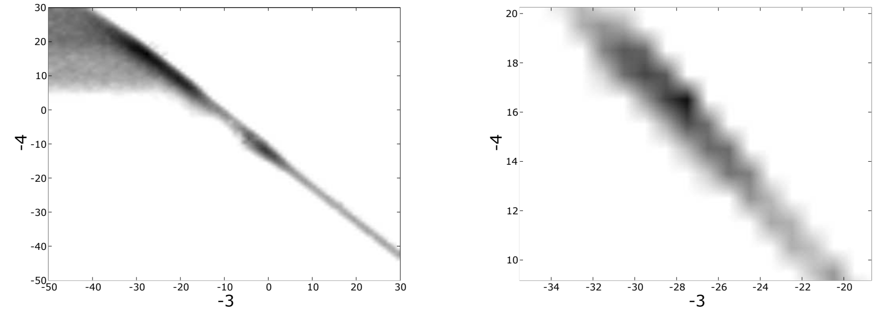


(a)

(b)

(c)

Figure 3.8: Isosurfaces of the variance between digest results and simulation at (a) 0.010, (b) 0.090 and (c) 0.087 variance (units arbitrary) as the methyl-esterified subsite energies of sites -5, -4 and -3 are adjusted over the range indicated. Units of the subsites energies are in kJ/mol. These make it clear that there is a consistent global optimum is found in the search, as the isosurface shrinks to a single, isolated area as the variance is reduced.



(a) (b)

Figure 3.9: Pseudo-colour plots of the variance σ^2 between digest results and simulation as the methyl-esterified subsite energies of sites -4 and -3 are adjusted. The slice is taken with subsite -5 methyl-esterified binding energy set to 0. Darker areas indicate better matches. (a) shows the total slice and (b) shows the area nearest the global minimum, the figure colouring is renormalised between figures. This slice indicates that a clear minima over the total area search by the grid search is found. All units of subsite energies are in kJ/mol. The difference in σ^2 between the global minimum and the next, separate valley on the this surface is 50% of the minimum value.

notable achievement in its own right, it does not reproduce their relative frequencies. Because it was possible to find what are arguably the best parameters for the extended subsite model, and these do not correctly reproduce isomer quantities, this is considered good evidence that the extended subsite model is not a sufficient extension for methyl-esterfied homogalacturonan. This is unsurprising for such a simple model, the success of the subsite model for the unmethyl-esterified homogalacturonan case surprised at least one reviewer. A probable explanation is that the methyl-esterification of a residue would tend to have an effect not only on the affinity with a specific subsite, but with the overall conformation of the bound oligomer and affect nearby subsite energies. Molecular docking simulations on a similar endo-PG [49], although only giving qualitative results, do seem to indicate that the binding energy differences between various oligomers cannot be decomposed into the extended subsite model because the total length of the substrate seemed to be of importance also. Unfortunately, experimental insight into a better way to model the interaction does not appear forthcoming due to difficulties in synthesizing patterned oligomers, but it is hoped that better molecular docking simulations may give insight into possible extensions of the subsite model, such as calculating a binding energy for each isomer, see section 3.4.2.

3.3 Comparison of Reconstruction Strategies

All reconstruction attempts were tested with chains with a distribution of DM centred at 50%. Three sets of first-order Markov test chains were generated with the Markov parameter P_{uu} set at 0.3 (anti-

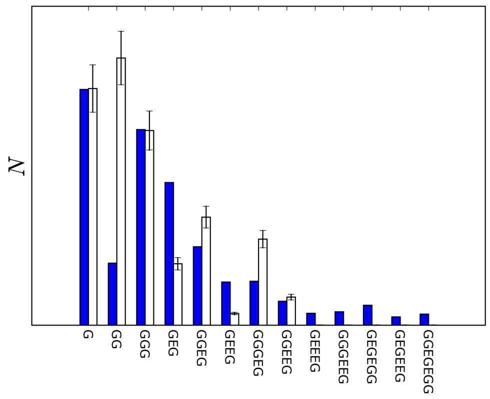


Figure 3.10: The best values for the subsites -5, -4 and -3 were found using the search described in text. This figure shows a comparison of the simulated isomers after fragmentation of 30% randomly methyl-esterified chain compared with experimental results (Aurelie Cucheval, unpublished data) see also appendix A. A cutoff was applied, simulated isomers occurring less than 200 times in the results are not shown. It is clear that although the correct isomers are generated with this model, the relative amounts are quantitatively incorrect and do not fit experiment.

blocky), 0.5 (random) and 0.7 (blocky) and $P_{mu} = 1 - P_{uu}$ (to give 50% DM), each consisting of a 1000 chains with a DP distribution centred at 200. Each set underwent simulated digestion for 3×10^7 iterations of the extended subsite model, this being chosen as the minimum number of iterations needed to reach a near-completed digest pattern that changed only minimally with a further 3×10^7 iterations. All the reconstruction algorithms generated chains of the correct DP pattern by design. Since none of the algorithms specifically tried to reconstruct correctly the DM distribution, it was of interest how well the DM distribution was matched in the reconstructed chains, both when all final fragments were retained and when only shorter fragments were retained. Figure 3.11 shows an example of typical DP distribution of fragments. Figure 3.12 shows the reconstructed DM distributions of the various algorithms.

It is clear that the match algorithm is the only one that reconstructs a DM distribution to within experimentally detectable limits consistently. The random algorithm performs well when all fragments are retained, however when only shorter fragments are retained the missing fragments will tend to be of higher methyl-esterification since any unmethyl-esterified areas have been fragmented due to the enzyme rules, so when choosing from this skewed sample the reconstructed DM distribution is shifted lower. Any method that attempts to directly utilise the final fragments will suffer from this, since the experimental limitations of isomer sequencing are not likely to improve in the near future. Although it would be possible to devise some method of attempting to model the higher DP fragments this would require assuming a statistical model of the fine-structure and this is akin to simply using the

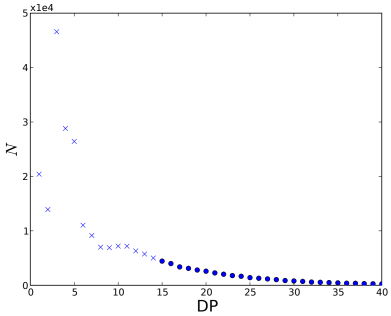
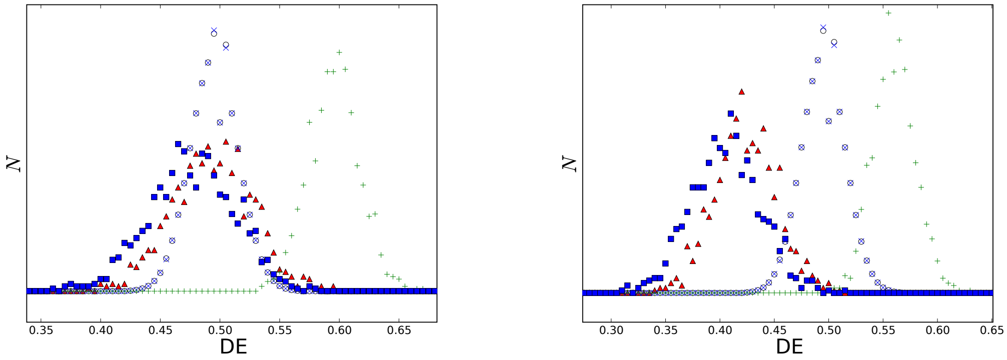


Figure 3.11: The DP distribution of fragments after completely digestion of the randomly 50% methyl-esterified polymer. For comparison with N, the starting distribution had a total of 1000 molecules. It is clear that while the majority of the fragments are less than DP 15 a significant number are not and this leads to difficulties with reconstruction techniques. Experimentally, it is not feasible yet to identify fragments of length greater than about DP 15 although progress is being made.

match algorithm. It is this result that leads the author to conclude that all reconstruction attempts involving direct use of the final fragments is unlikely to be successful without significant experimental improvements. For the purpose of realistic DM reconstruction, the algorithm incorporating knowledge of the endo-PG II model performs similarly to random. It can also be seen that the overlap algorithm skews the DM distribution in the positive direction. This is because due to the enzyme rules, most fragments (all except end-chain fragments), will always have an unmethyl-esterified residue on either end and so when overlap is performed at least one unmethyl-esterified residue will be lost, but this is not the case with the inner residues. Finally, it is clear that the match algorithm performs very successfully here, generating DM distributions very similar to the original.

Table 3.1 shows the 1$^{\text{st}}$ order Markov parameters for reconstructed chains. The higher order Markov statistics are not shown as they did not reveal any features not seen at 1$^{\text{st}}$ order, this is unsurprising as the test chains were 1$^{\text{st}}$ order. It is clear that endo-PG II and overlap are not viable algorithms at present. Both consistently shift the Markov parameters in reconstructing the chain. The random algorithm performs remarkably well considering no knowledge of enzyme kinetics is used, however, as soon as a cutoff to the retained fragments is applied, it is no longer useful, since the missing fragments act to misrepresent the totality of the original chain. The match algorithm consistently returns parameters near the original under all conditions. This algorithm has the advantage of easily being modified for a different set of enzyme rules or random fragmentation as the fragmentation method is not integral to the algorithm. While it is computationally intensive and forces a statistical model of the polymer fine-structure to be assumed, it seems the only viable candidate for further work. It was preliminarily tested on reconstructing 2$^{\text{nd}}$ order chains and seemed to perform well, however, the computational resources needed for this algorithm scale as N^2 with the Markov order employed. Other non-Markovian statistical models could be used as well, and the problem of how many sets of different



(a)

(b)

Figure 3.12: The degree of methyl-esterification distributions for the reconstructed polymers using several different algorithms. (a) result when all fragments are retained, (b) only fragments less than DP 15 are retained. The results are x, original distribution; triangles, random; boxes, pgII; o, match; +, overlap. Both figures show the clear success of the match algorithm under both ideal and realistic conditions.

composition are needed to effectively cover the space of all possible physical digests results is not yet resolved. This algorithm is easily modified to usefully compare with other available data such as the timecourse of the digestion, although it is too time-consuming to obtain experimental time course data for this to become a routine characterisation tool. Other extensions would be de-esterifying the fragments of the completed digestion to experimentally obtain a fragment DP distribution for comparison, fragmentation by several different enzymes or randomly fragmenting the polymers by shearing, all of which could contribute to improved knowledge of the fine-structure. Unfortunately, current resolutions on CE and other techniques mean that the vast quantity of the different fragments generated by random fragmentation would be unresolvable, however this may change in the future. There is also the difficult question of determining at what level is the fine-structure well-described. It may be that 2^{nd} order Markovian parameters are all that is necessary to completely characterise the functional behaviour of homogalacturonan, but this is difficult to determine, certainly the full sequence information is likely to only be unobtainable but unnecessary.

3.4 Future Work

As with almost all research, there is much still to be explored. This section outlines some ideas of where future work could proceeded profitably.

(a)

Algorithm	No cutoff P_{uu}	No cutoff P_{um}	Cutoff 15 P_{uu}	Cutoff 15 P_{um}
original	0.31	0.71	0.31	0.71
random	0.31	0.71	0.41	0.76
endo-PG II	0.37	0.71	0.49	0.77
overlap	0.14	0.69	0.12	0.76
match	0.33	0.69	0.33	0.69

(b)

Algorithm	No cutoff P_{uu}	No cutoff P_{mu}	Cutoff 15 P_{uu}	Cutoff 15 P_{mu}
original	0.50	0.51	0.50	0.51
random	0.50	0.51	0.58	0.58
endo-PG II	0.54	0.51	0.61	0.58
overlap	0.30	0.49	0.26	0.58
match	0.51	0.50	0.51	0.50

(c)

Algorithm	No cutoff P_{uu}	No cutoff P_{mu}	Cutoff 15 P_{uu}	Cutoff 15 P_{mu}
original	0.70	0.31	0.70	0.31
random	0.70	0.31	0.76	0.41
endo-PG II	0.68	0.31	0.72	0.41
overlap	0.45	0.30	0.48	0.43
match	0.70	0.30	0.70	0.31

Table 3.1: This table shows the first order Markovian parameters (or diad frequencies) of the original test and reconstructed chains. P_{uu} is the frequency of unmethyl-esterified residues followed by a methyl-esterified residue and P_{mu} is the frequency of methyl-esterified residues followed by unmethyl-esterified residues. (a) for anti-blocky $P_{uu} = 0.3$, (b) for random chain $P_{uu} = 0.5$ and (c) for blocky $P_{uu} = 0.7$. The degree to which the reconstructed Markov parameters agree with the original is considered a rough measure of the goodness of the reconstruction algorithm. The columns labelled Cutoff 15 indicate the reconstructed parameters when only digestion of fragments DP 15 or less were passed to the reconstruction algorithm.

3.4.1 Other Experimental Systems

There are other systems of copolymers with well-characterised enzymes that are of interest. The most obviously applicable areas are other well-known polysacchararides. Similar tools have been used for galactomannananans [32] and there are many other copolymers where the patterning is considered important and there are limited existing tools for determining the fine structure. If molecular docking calculations were successful, this would make it much easier to extended to copolymer systems with more than 2 types of residues. It would also be of interest to extend the simulation to branched polymers, although this is likely to be non-trivial but one can envision simple structures that might be able to reproduce at least qualitative results. Such work could be of use in better determining the full structure of pectin or other branched polysaccharides but would be quite ambitious at this stage. Kinetic data is available for several other another less-specific endo-PGs [50, 51] which seem amenable to modelling and the results of molecular docking simulations for one of the endo-PGs [49] may be of use if the binding energies were scaled to make them realistic.

Microarrays and other ways of investigating pectin

The use of enzymes to fragment and HPCE to investigate the fragments is far from the only experimental insight into homogalacturonan methyl-esterification patterns. However, all current methods have limitations on the amount of detail they are able to provide. Other techniques include mass spectrometry or other chromatography methods to determine fragment isomers (although the quantification of this technique is limited), microarray technologies of polyslides to which only certain patterns will bind and the use of antibodies that bind only to specific methyl-esterification patterns. The use of chemometrics to combine data from a variety of experimental sources is also beginning to be used [13]. One possible avenue of future work would be extending BJPS to derive the maximum possible structural information utilising data from a spectrum of these techniques, for instance models of antibody binding could be developed and used to predict experimental results from generated structures, similar to what has been done for fragmented structures. The constraints from results of several techniques utilised on pectin from a similar source could provide further insight into the fine structure. It is of ongoing interest to use models of experimental techniques to determine the optimum combination of techniques to give insight into fine-structure features of interest. Long-term possibilities include modifying an enzyme through mutagenesis to give a “designer” hydrolysis enzyme that has properties most amenable for fragment reconstruction. There are already known mutants of endo-PG II with different properties including a single amino acid mutation which changes the precessionally behaviour of the enzyme [43, 52, 53].

3.4.2 Binding Energy Calculations

With high-speed computers becoming commodity products, now is a good time to be involved in any work involving computation. Ab-initio quantum calculations promise to solve from first-principles almost any question asked of them, however, unfortunately for any molecule of more than a few atoms this is infeasible on today's computer. However, improvements are being made every day to more practical force-field calculations and Monte-Carlo simulations over states.

Protein-substrate docking is now a standard tool for elucidating a complexes structure. As detailed previously, due to difficulties in synthesizing homogalacturonan oligomers with specified methyl-esterification patterning, experimental access to the binding energies of variously methyl-esterified oligomers with endo-PG II (and other enzymes) is not expected to be available in the near future. However, groups have already attempted using computer simulation of docking to calculate these values for similar enzymes. These binding energies could then be used to understand how to extend the previously detailed subsite-model in some way that still retains the granularity that allows macroscopic simulations.

Successful calculation of experimental binding energies from computer simulation is not yet trivial, although it is improving constantly. It is computationally difficult to simulate full-flexibility of the large molecules involved, particularly while including solvent effects, so in general the structures are considered rigid for generating the first set of possible bound structures. If the substrate docking involves large protein conformational change this will produce flawed results. Fortunately, from other members of the endopolygalacturonase family, there is reason to believe the endo PG enzyme conformation changes less in bound state for non-precessive enzymes so this assumption should hold [54].

Another group [49] has attempted to compute the binding energies of various oligogalacturonides with a *F. moniliforme* endo-PG. While interesting, their results appear to be only qualitative, with predicted binding energy differences between experimentally observed positions of binding in the order of hundreds of kJ/mol such that if these were accurate only one position would ever be experimentally observed. Their results, if at least qualitatively correct, do seem to indicate that the simple subsite model described earlier would need to be extended to account for all the features of the binding energies. In calculating their energies they did not take explicit solvent effects into account.

It is hoped that these results could be improved upon, and physically accurate binding energies could be calculated. A well-resolved x-ray diffraction structure of *A. niger* endo-PG II is available [55], along with similar endopolygalacturonases in a bound state. Furthermore, mutagenesis experiments [43, 52, 53] and mass spectrometry [56] provide good evidence of the amino acids involved at each subsite. This provides ample experimental evidence of the active subsite, which minimises the conformational freedom (and therefore computational time) that needs to be given for accurate binding energy calculations.

Unfortunately, the vast majority of work on docking has been driven more by interest in the docked state geometry than by a desire to generate an accurate binding energy difference. An advantage of the probability normalisation described earlier, is that only the free energy differences between various binding positions are needed, so the absolute free binding energy is not important. This eliminates one source of difficultly. Also, most work to date has concentrated on protein-protein or protein-RNA interactions, so well-calibrated force-fields for carbohydrate work are not so widely available. However this situation is changing. Two different sets of tools are proposed for future work. HADDOCK [57], while originally designed for protein-protein interactions, has been extended to carbohydrate interactions. HADDOCK generates an initial set of structures based on rigid interactions and then refines these structures through simulated annealing and in the final step adds explicit solvent molecules. An alternative set of tools is the use of AutoDock [58] to generate an initial set of structures based on rigid interactions. Then SLICK [59] would be used to generate binding energies for each structure. SLICK is a scoring function specifically optimised for protein-carbohydrate interactions. Solvent effects are only included implicity in SLICK. SLICK claimed to, after calibration, produce binding energies with a mean absolute error compared with experiment of 1.0 kJ/mol for test cases.

Both methods of computing binding energies could probably be improved by calibrating the energy scoring functions specifically for endopolygalacturonase binding. This requires experimental results to compare with. As detailed earlier results are available for all fully unmethyl-esterified binding positions and all trimer bindings. If binding energies were consistently calculable to within 1 kJ/mol this would probably be accurate enough for reasonable simulation of digestion patterns based on the previously listed results of the sensitivities of the simulation to changes in binding energy. One option would be rather than trying to find an extension to the subsite model, simply calculating the binding energy of all possible bindings. There a five subsites involved in successful binding that are not required to always be filled with an unmethyl-esterified residue, and each of them may be filled with a methyl group, filled with an unmethylated group or unfilled giving $5^3 - 1 = 124$ possibilities, with the one subtracted for the dimer which is known not to bind. This is a low enough number of possiblities to be computational feasible to calculate binding energies for all of them, eliminating the need for decomposing the binding energies into a specific subsite model. For generating substrate 3-D structures for docking once again the DNA/RNA/protein field is ahead of carbohydrate research. However, there are two tools that it is expected should be suitable for generating the various oligomers needed easily. POLYS [60] which is designed specifically for generating pectin structures or the easily accessible via WWW interface SWEET-II [61,62] for rapid 3-D construction of oligo and polysaccharides. Both tools are expected to be capable of generating the galacturonan fragments need, with SWEET-II already having been tested. Both accept a 1-D monosaccharide sequence (in ASCII, and may include branching side-chains) and generate a 3-D structure suitable for input to molecular docking using known sugar geometries and further force-field based optimisation.

It is hoped that the use of docking tools could also make extending this work to new copolymer/enzymes systems much easier, and limit the need for the same amount of extensive experimental input. In particular, it would seem likely that if this approach is useful for completing the model for endo-PG II it would should also be helpful for the enzymes in the endo PG family, although the model would also need to be extended to account for precessional effects which are commonly seen in the endo PG family [52]. A rather ambitious proposal for how, if successful, this work could be applied to new copolymeric systems with a minimal amount of experimental input is shown in figure 3.13.

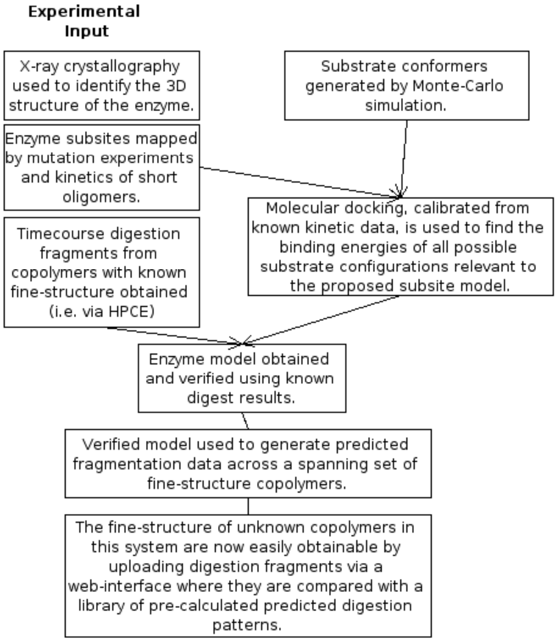

Experimental Input
X-ray crystallography used to identify the 3D structure of the enzyme.
Substrate conformers generated by Monte-Carlo simulation.
Enzyme subsites mapped by mutation experiments and kinetics of short oligomers.
Molecular docking, calibrated from known kinetic data, is used to find the binding energies of all possible substrate configurations relevant to the proposed subsite model.
Timecourse digestion fragments from copolymers with known fine-structure obtained (i.e. via HPCE)
Enzyme model obtained and verified using known digest results.
Verified model used to generate predicted fragmentation data across a spanning set of fine-structure copolymers.
The fine-structure of unknown copolymers in this system are now easily obtainable by uploading digestion fragments via a web-interface where they are compared with a library of pre-calculated predicted digestion patterns.

Figure 3.13: A flow diagram outlining the proposed way characterisation of polysaccharide fine-structure using enzymes could become more routine.

3.4.3 Web interface for reconstruction

One of the technologies that has speeded the advancement of genome research has been the wide availability of free and easy to use tools, particularly web interfaces that require no local installation to utilise. Python is a widely used language for web interfacing using the common gateway interface standard and since BJPS is scriptable by python, there is interest in developing a low learning curve, publically available interface which would allow someone to upload fragmentation data about a particular digest and retrieve the best attempt at a reconstructed polymer. To be of general interest, enzymes other than endo-PG II would need to be modelled.

Protein and DNA/RNA bioinformatics are far ahead of the carbohydrate field in terms of routines for easy comparison of structures, and generation of 3-D structure. It is believed by this author, that there is much fertile ground for making certain types of carbohydrate research currently only used by a few groups with specialised tools, much more routine, particularly via the use of low-learning curve web interfaces. See [63] for a review on the current state of carbohydrate chemistry on the web.

3.4.4 Optimisation of BJPS

BJPS was designed with the goal of being extensible, easy to maintain and to have competitive performance. Although the author has no plans to in the near future, one possible project would to maximise the performance of the library. BJPS has reasonably fast performance, however it has never been written to take advantage of particular hardware features. One source of improved performance would be the creation of a version optimised specifically for a certain hardware set.

Microprocessors have been significantly faster than main memory of an average computer system for quite some time and this situations looks likely to stay the case [64]. This means, in order to utilise the full performance of a processor, BJPS needs to be written in a way that maximises the use of on-processor cache. At the moment, using the random attack mechanism, memory anywhere in the polymer set is accessed at random, which if the polymer set size exceeds the processor's cache size, is the worst possible scenario for cache usage. It would seem that no significant changes in results are likely to be found if the set is broken into subsets of a size smaller than the cache size, and each of these subsets digested separately, allowing better use of the cache. This also provides an obvious mechanism for taking advantage of the increasing prevalence of shared-memory multiprocessor systems.

Chapter 4 - Conclusion

Biopolymer research is an active area of innovation. The use of computational informatic tools has played a huge part in advancing our understanding of genomics, and looks set to play a major role in elucidating the structure-function relationship of other biopolymers such as polysaccharides.

This report has outlined the development of an extensible, general library for simulation of polymer fragmentation and reconstruction at the informational level. This library has been used to verify a subsite model for endo-PG II digestion of unmethyl-esterified homogalacturonan, and show that the subsite binding energies calculated from product analysis and relative kinetics correctly reproduces the time course of an experimental digestion. A straightforward attempt to extend this model to the copolymer system of partially methyl-esterified homogalacturonan was shown to be overly simplistic and a global optimisation of the parameter space showed conclusively that this model is unfeasible, although it is qualitatively correct. The assumption that a methyl-esterified residue in a subsite will not affect nearby subsite bindings is likely the reason this model does not work. A way forward, using molecular docking simulations to avoid the difficulties of oligomer synthesis, was proposed for finding an extension of the endo-PG II subsite model to a copolymer system and avoiding the need to decompose the binding energies into a subsite model. It is proposed that if successful, these techniques should be generalisable to many other systems of interest.

Several algorithms were proposed for the reconstruction of copolymer fine-structure based on fragmentation data and tested using the extended subsite model for fragmentation on three 1st order Markov chains. The match algorithm, which involves generating polymers across the entire range of Markov parameters and comparing the predicted digest patterns with the test digest was the only algorithm able to perform well when not all experimental results were given. It is easily generalisable to other enzyme models or statistical systems but is computationally intensive.

Bibliography

[1] J. D. Watson and F. H. C. Crick. A structure for deoxyribose nucleic acid. *Nature*, 171:737, 1953.

[2] M. Rubinstein and R. H. Colby. *Polymer Physics*. Oxford University Press, 2003.

[3] G. Monaco. On the microstructural analysis of (pseudo) copolymers. *Macromolecule Theory and Simulation*, 11:84–92, 2002.

[4] I. R. Herbert. *NMR Spectroscopy of Polymers*, chapter Statistical analysis of copolymer sequence distribution, pages 50–79. Springer, 1993.

[5] J. L. Snell and C. M. Grinstead. *Introduction to Probability*, chapter Markov Chains, pages 405–470. American Mathematical Society, 1997.

[6] J.-P. Vincken, H. A. Schols, R. J. F. J. Oomen, M. C. McCann, P. Ulvskov, A. G. J. Voragen, and R. G. F. Visser. If homogalacturanan were a side chain of rhamnogalacturonan i. implications for cell wall architecture. *Plant Physiology*, 132:1781–1789, 2003.

[7] H. Braconnot. *Ann. Chim. Phys.*, 28(2):173–178, 1825.

[8] W. Pilnik. *Gums and Stabilisiers for the Food Industry 5*, chapter Pectin - a many splendoured thing, pages 209–221. Oxford University Press, 1990.

[9] C. D. May. *Gums and Stabilisiers for the Food Industry 5*, chapter Commerical sources and production of pectins, pages 223–241. Oxford University Press, 1990.

[10] S. Pérez, K. Mazeau, and C. H. du Penhoat. The three-dimensional structures of pectic polysaccharides. *Plant Physiology and Biochemistry*, 38:37–55, 2000.

[11] B. L. Ridley, M. A. O'Neill, and D. Mohnen. Pectins: structure, biosynthesis, and oligogalacturonide-related signaling. *Phytochemistry*, 51:929–967, 2001.

[12] E. Bonnin, A. Le Goff, G.-J. W. M. van Alebeek, A. G. J Voragen, and J.-F. Thibault. Mode of action of *Fusarium moniliforme* endopolygalacturonase towards acetylated pectin. *Carbohydrate Polymers*, 52:381–388, 2003.

[13] W. G. T Willats, P. Knox, and J. D. Mikkelsen. Pectin: new insights into an old polymer are starting to gel. *Trends in Food Science & Technology*, 17:97–104, 2006.

[14] G. Limberg, R. Körner, H. C. Buchholt, T. M. I. E. Christensen, P. Roepstorff, and J. D. Mikkelsen. Analysis of different de-esterification mechanisms for pectin by enzymatic fingerprinting using endopectin lyase and endopolygalacturonase ii from *A. niger*. *Carbohydrate Research*, 327:293–307, 2000.

[15] G. J. W. M. van Alebeek, K. van Scherpenzeel, G. Beldman, H. A. Schols, and A. G. J. Voragen. Partially esterified oligogalacturonides are the preferred substrates for pectin methylesterase of *Aspergillus niger*. *Biochemical Journal*, 372:211–218, 2003.

[16] A. H. E. Ström. *Characterisation of pectin fine-structure and its effect on supramolecular properties*. PhD thesis, National University of Ireland, Cork, March 2006.

[17] P. J. H. Daas, A. G. J. Voragen, and H. A. Schols. Study of the methyl ester distribution in pectin with *endo*-polygalacturonase and high performance size-exclusion chromatography. *Biopolymers*, 58:195–203, 2000.

[18] F. Goubet, A. Ström, P. Dupree, and M. A. K. Williams. An investigation of pectin methylesterification patterns by two independant methods: capillary electrophoresis and polysaccharide analysis using carbohydrate gel electrophoresis. *Carbohydrate Research*, 340:1193–1199, 2005.

[19] R. Körner, G. Limberg, T. M. I. E. Christensen, J. D. Mikelsen, and P. Roepstorff. Sequencing of partially methyl-esterified oligogalacturonates by tandem mass spectrometry and its use to determine pectinase specificities. *Analytical Chemistry*, 71:1421–1427, 1999.

[20] T. G. Neiss, H. N. Cheng, P. J. H. Daas, and H. A. Schols. Compositional heterogeneity in pectic polysaccharides: NMR studies and statistical analysis. *Macromolecular Symposium*, 140:165–178, 1999.

[21] E. Bonnin, E. Dolo, A. L. Goff, and J.-F. Thibault. Characterisation of pectin subunits released by an optimised combination of enzymes. *Carbohydrate Research*, 337:1687–1696, 2002.

[22] S. E. Guillotin, E. J. Bakx, P. Boulenguer, J. Mazoyer, H. A. Schols, and A. G. J. Voragen. Populations having different GalA blocks characteristics are present in commercial pectins which are chemically similar but have different functionalities. *Carbohydrate Polymers*, 60:391–8, 2005.

[23] P. Hellín, M.-C. Ralet, E. Bonnin, and J.-F. Thibault. Homogalacturonans from lime pectins exhibit homogenous charge density and molar mass distributions. *Carbohydrate Polymers*, 60:307–315, 2005.

[24] P. J. H. Daas, B. Boxma, A. M. C. P. Hopman, A. G. J. Voragen, and H. A. Schols. Nonesterified galacturonic acid sequence homology of pectins. *Biopolymers*, 58:1–8, 2001.

[25] R. Weinberger. *Practical Capillary Electrophoresis*. Academic Press, 2000.

[26] Capillary electrophoresis. In *Wikipedia*.

[27] A. Ström and M. A. K. Williams. On the seperation, detection and quantification of pectin derived oligosaccharides by capillary electrophoresis. *Carbohydrate Research*, 339:1711–1716, 2004.

[28] C. M. Jiang, M. C. Wu, W. H. Chang, and H. M. Chang. Determination of random- and blockwise-type de-esterified pectins by capillary zone electrophoresis. *Journal of Agricultural and Food Chemistry*, 49:5584–5588, 2001.

[29] A. Ström, M.-C. Ralet, J.-F. Thibault, and M. A. K. Williams. Capillary electrophoresis of homogenous pectin fractions. *Carbohydrate Polymers*, 60:467–473, 2005.

[30] M. A. K. Williams, G. M. C. Buffet, and T. J. Foster. Analysis of partially methyl-esterified galacturonic acid oligmers by capillary electrophoresis. *Analytical Biochemistry*, 301:117–122, 2002.

[31] T. Suganuma, R. Matsuno, M. Ohnishi, and K. Hiromi. A study of the mechanism of action of taka-amylase a on linear oligosaccharides by product analysis and computer simulation. *Journal of Biochemistry*, 84:293–316, 1978.

[32] J. S. Grant Reid, M. Edwards, M. J. Gidley, and A. H. Clark. Enzyme specificity in galactomannan biosynthesis. *Planta*, 195:489–95, 1995.

[33] K. Østgaard, B. Stokke, and B. Larsen. Numerical model for alginate block specificity of mannuronate lyase from *Haliotis*. *Carbohydrate Research*, 260:83–98, 1994.

[34] J. A. Thoma, C. Brothers, and J. Spradlin. Subsite mapping of enzymes. studies on *Bacillus subtilis* amylase. *Biochemistry*, 9:1768–1776, 1970.

[35] J. A. Thoma, G. V. K. Rao, C. Brothers, and J. Spradlin. Subsite mapping of enzymes. *Journal of Biological Chemistry*, 246:5621–5635, 1971.

[36] L. Rexova-Benkova. The size of the substrate-binding site of an *Aspergillus niger* extracellular endopolygalacturonase. *European Journal of Biochemistry*, 39:109–115, 1973.

[37] J. A. E. Benen, H. C. M. Kester, and J. Visser. Kinetic characterization of *Aspergillus niger* N400 endopolygalacturonase i, ii and c. *European Journal of Biochemistry*, 259:577–585, 1999.

[38] M. A. K. Williams, G. M.C. Buffet, T. J. Foster, and I. T. Norton. Simulation of endo-pg digest patterns and implications for the determination of pectin fine structure. *Carbohydrate Research*, 334:243–250, 2001.

[39] H. C. M. Kester, D. Magaud, C. Roy, D. Anker, A. Doutheau, V. Shevchik, N. Hugouvieux-Cotte-Pattat, J. A. E. Benen, and J. Visser. Performance of selected microbial pectinases on synthetic monomethyl-esterified di- and trigalacturonates. *Journal of Biological Chemistry*, 274(52):37053–37059, 1999.

[40] G. Géymánt, G. Hovánszki, and L. Kandra. Subsite mapping of the binding region of $\alpha -$ amylases with a computer program. *European Journal of Biochemistry*, 269:5157–5162, 2002.

[41] J. J. Hunt, R. Cameron, and M. A. K. Williams. On the simulation of enzymatic digest patterns: the fragmentation of oligomeric and polymeric galacturonides by endo-polygalacturonase II. *Biochimica Biophysica Acta*, 1760:1696–1703, 2006.

[42] M. Matsumoto and T. Nishimur. Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modeling and Computer Simulation*, 8(1):3–30, January 1998.

[43] Upgrading of sugar beet pectin by enzymatic modification and molecular breeding. Technical report, EuroPectin Project, March 2002.

[44] M.I.A. Lourakis. levmar: Levenberg-marquardt nonlinear least squares algorithms in C/C++. [web page] `http://www.ics.forth.gr/~lourakis/levmar/`, Jul. 2004. [Accessed on 31 Jan. 2005.].

[45] K. M. Passino. *Biomimicry for optimization, control, and automation.* Springer, 2004.

[46] R. G. Cameron, G. Luzio, E. B., J. N., and A. Plotto. Production of narrow-range size-classes of polygalacturonic acid oligomers. *Proceedings of the Florida State Horticulture Society*, 118:406–409, 2005.

[47] H. J. D. Bussink, H. C. M Kester, and J. Visser. Molecular cloning, nucleotide sequence and expression of the gene encoding prepro-polygalacturonase ii of *Aspergillus niger*. *FEBS Letters*, 273:127–130, 1990.

[48] M. A. K. Williams, S. J. Williams, and D. K Lloyd. Quantitative aspects of capillary electrophoresis. *Trends in Analytic Chemistry*, 10:272–279, 1991.

[49] G. Andre-Leroux, D. Tessier, and E. Bonnin. Action pattern of *Fusarium moniliforme* endopolygalacturonase towards pectin fragments: Comprehension and prediction. *Biochimica et. Biophysica Acta*, 1749:53–64, 2005.

[50] E. Bonnin, A. Le Goff, R. Korner, G.-J. W. M. Van Alebeek, T. M. I. E. Christensen, A. G. J. Voragen, P. Roepstorff, C. Caprari, and J.-F. Thibault. Study of the mode of action of endopolygalacturonase from *Fusarium moniliforme*. *Biochimica et. Biophysica Acta*, 1526:301–309, 2001.

[51] E. Bonnin, A. Le Goff, R. Korner, J. Vigouroux, P. Roepstorff, and J.-F. Thibault. Hydrolysis of pectins with different degrees and patterns of methylation by the endopolygalacturonase of *Fusarium moniliforme*. *Biochimica et. Biophysica Acta*, 1596:83–94, 2002.

[52] S. Pages, W. H. M. Heijne, H. C. M. Kester, J. Visser, and J. A. E. Benen. Subsite mapping of *Aspergillus niger* endopolygalacturonase ii by site-direct mutagenesis. *Journal of Biological Chemistry*, 275:29348–29353, 2000.

[53] S. Armand, M. J. M. Wagemaker, P. Sanchez-Torres, H. C. M. Kester, Y. Santen, B. W. Dijkstra, J. Visser, and J. A. E. Benen. The active site topology of *Aspergillus niger* endopolygalacturonase ii as studied by site-direct mutagenesis. *Journal of Biological Chemistry*, 275:691–696, 2000.

[54] G. van Pouderoyen, H. J. Snijder, J. A. E. Benen, and B. W. Dijkstra. Structural insights into the processivity of endopolygalacturonase i from *Aspergillus niger*. *FEBS Letters*, 554:462–466, 2003.

[55] Y. van Santen, J. A. E. Benen, K.-H. Schroter, K. H. Kalk, S. Armand, J. Visser, and B. W. Dijkstra. 1.68-Å crystal structure of endopolygalacturonase ii from *Aspergillus niger* and identification of active site residues by site-directed mutagenesis. *Journal of Biological Chemistry*, 274:30474–30480, 1999.

[56] D. King, M. Lumpkin, C. Begmann, and R. Orlando. Studying protein-carbohydrate interactions by amide hydrogen/deuterium exchange mass spectrometry. *Rapid Communications in Mass Spectrometry*, 16:1569–1574, 2002.

[57] C. Dominguez, R. Boelens, and A. M.J.J. Bonvin. HADDOCK: a protein-protein docking approach based on biochemical and/or biophysical information. *Journal of the American Chemical Society*, 125:1731–1737, 2003.

[58] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a lamarckian genetic algorithm and and empirical binding free energy function. *Journal of Computional Chemistry*, 19:1639–1662, 1998.

[59] A. Kerzmann, D. Neumann, and O. Kohlbacher. SLICK - scoring and energy function for protein-carbohydrate interactions. *Journal of Chemical Information and Modeling*, 46:1635–1642, 2006.

[60] S. B. Engelsen, S. Cros, W. Mackie, and S. Pérez. A molecular builder for carbohydrates: Application to polysaccharides and complex carbohydrates. *Biopolymers*, 39:417–433, 1996.

[61] A. Bohne, E. Lang, and C.-W. von der Lieth. W3-SWEET: Carbohydrate modeling by internet. *Journal of Molecular Modelling*, 4:33–43, 1998.

[62] A. Bohne, E. Lang, and C. W. von der Lieth. SWEET - WWW-based rapid 3D construction of oligo- and polysaccharides. *Bioinformatics Applications Note*, 15:767–768, 1999.

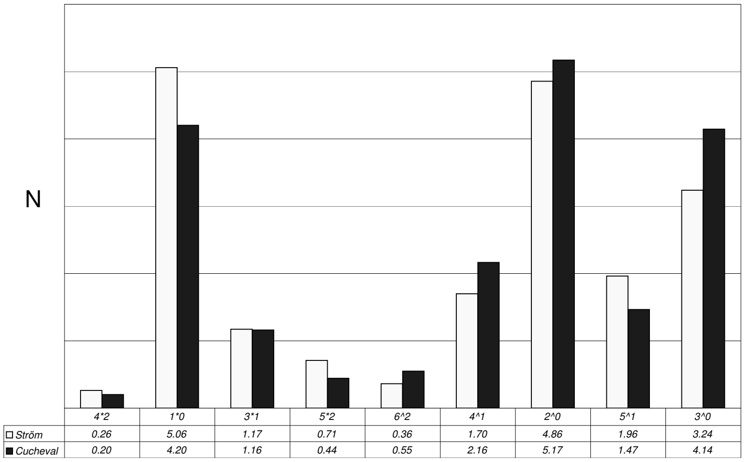
[63] O. Berteau and R. Stenutz. Web resources for the carbohydrate chemist. *Carbohydrate Research*, 339:929–936, 2004.

[64] R. van der Pas. Memory hierarchy in cache-based systems. Technical report, Sun Microsystems, 2002.

Appendix A - Compiled Digestion Data

The isomer digest data used in this report is that shown in figure 3.10 from A. Cucheval, unpublished data. It was obtained using from HPCE at several different ionic strengths performed on the digest of a 30% DM pectin that is believed to be randomly methyl-esterified. It is shown again in figure A.1 compared with similar data [16] (which is also published in [18]).

Since HPCE does not separate isomers of the same length with the same number of methyl-esterified residues it is necessary to use knowledge of the enzyme rules to determine the isomers, assuming the starting pectin is of high enough DP that end fragments make up an insignificant portion of the resulting isomers. Since endo-PG II is known to only cleave between unmethyl-esterified residues any proposed isomer must begin and end with an unmethyl-esterified residue. Additionally, mass spectrometry of similar digestions [14] has been used to sequence the final fragments, although this method has the disadvantage of not being able to quantify the relative amounts of each isomer. Another source that attempted to identify and quantify the isomers resulting from endo-PG II digestion (M-C Ralet, unpublished data), used anion exchange chromatography and mass spectrometry on the fragments from a 43% DM pectin digestion. Ralet's method did give quantifiable results but because the starting pectin was of higher DM, this was not compared in the earlier figure. Table A.1 shows a comparison of the results of these two techniques. These results allow each of the detected isomers in figure A.1, except for 6^2, to be uniquely assigned a sequence.



	4*2	1*0	3*1	5*2	6^2	4^1	2^0	5^1	3^0
□ Ström	0.26	5.06	1.17	0.71	0.36	1.70	4.86	1.96	3.24
■ Cucheval	0.20	4.20	1.16	0.44	0.55	2.16	5.17	1.47	4.14

Figure A.1: Comparison of two quantifications of the low DP isomers resulting from digestion of 30% randomly methyl-esterified pectin with endo-PG II. Notation for the labelling is $\texttt{a} \bigwedge \texttt{b}$ where a is the DP of the fragment and b is the number of esterified residues. Agreement is reasonable given the limitations of HPCE.

	Isomer	Identified in A	Identified in B	Identified in HPCE results
1^0	`G`	y	y	y*
2^0	`GG`	y	y	y*
3^0	`GGG`	y		y*
3^1	`GEG`	y	y	y*
4^1	`GGEG`	y	y	y*
4^2	`GEEG`	y	y	y*
5^1	`GGGEG`	y		y*
5^2	`GGEEG`	y	y	y*
5^3	`GEEEG`		y	
6^1	`GGGGEG`	y	y*	
6^2	`GGGEEG`	y	y*	y*
	`GGEGEG`	y	y*	y*
	`GGEEGG`	y	y*	y*

Table A.1: The detected isomers of DP 6 or less in the results of pectin digestion by endo-PG II. Column A shows isomers detected by [14], column B by M-C Ralet (unpublished data) and the final column those by the HPCE results. Isomers are written non-reducing end first and labelled with 'G' for unmethyl-esterified residues and 'E' for methyl-esterified residues. y* is used to indicate residues whose length and number of esterified residues were identified but the specific isomer was not confirmed. The differences between columns A and B can all be explained by postulating the B was allowed to digest more thoroughly, which is confirmed by the absence of any unmethyl-esterified trimer.

Appendix B - Documentation for BJPS

BJPS aims to be a suite of high-performance, extensible libraries useful for the high-level simulation of enzyme digestion and other mechanical means of fragmenting biopolymers. Although, currently targeted towards simulation of digestion and reconstruction of partially methyl-esterified homogalacturan it is easily extensible to other co-polymer systems and was made with generality and extensibility in mind.

A version of BJPS is included on CD-ROM with this report in source form. It is expected that future updates and further documentation should be available from `http://bjps.sourceforge.net`.

B.1 Technologies Used

BJPS is written as a C++ library using the Standard Template Library (STL), it aims to be cross-platform and compiler-independent, although has only been tested Intel's C++ compiler[1] running on Linux. It has been tested on both x86 and AMD64 architectures (also known by Intel as EM64) and is able to take advantage of 64-bit address space when running on 64-bit platforms. A Python[2] binding was also developed using SIP[3], and the library is equally accessible from python or C++. The simulations in this software were done using the python interface with all graphical results being generated using Matplotlib[4], however, there are many alternative possibilities. The GNU Scientific Library[5] is used for it's high-quality random number generator that is consistent across platforms.

B.2 Installation Instructions

Installation of BJPS varies between machine type and distribution. The prerequisites are a functioning C++ compiler (including Standard Template Libraries) and make, GNU Scientific Library (tested with version 1.4). SIP (tested with version 4.3.2) is necessary to compile the python bindings. Firstly, edit `config` to set the C++ Compiler and flags to use. Then running `make` should build both the library and python bindings in the `lib` folder. Installation of the library and bindings are distribution specific, and no provision has been made yet for automatic installation.

[1] Available from `http://www.intel.com/cd/software/products/asmo-na/eng/compilers/284132.htm`

[2] Python Scripting Language available from `http://www.python.org/`

[3] SIP available from `http://www.riverbankcomputing.co.uk/sip/`

[4] A Python 2D plotting library available from `http://matplotlib.sourceforge.net/`

[5] Available from `http://www.gnu.org/software/gsl/`

B.3 Using the Library

The BJPS library is fully object-oriented and all services the library provides are accessed using objects. The library should be thread-safe, provided the user ensures that only one thread is ever manipulating the same object at any given time. All code should compile in both 64-bit and 32-bit systems. Compiling with `NDEBUG` defined will make the library significantly faster as all debugging code will be removed.

The full documentation is provided as source-code comments which are extractable using Doxygen[6] to generate either a PDF or HTML reference documentation. This appendix attempts only to describe the most commonly used objects and functions, and the internals of the library are not discussed. The definitions are given in C++, but it is usually clear what the python equivalent should look like. The STL types are not prefixed with `std` namespace, but this is assumed. Anytime a STL list object is used as a parameter, a python `List` object is expected from python, and conversion code between the two types are part of the library. Also calls not explicitly listed as static also involve an implicit `this` pointer to an instance of the object. The library throws a exception with an informational message for invalid input, although input verification is not extensive yet.

As detailed earlier, the code is for simulation of polymer fragmentation at the 'information' level, it has been written for AB copolymer digestion but provides a framework that would be extensible to ABC copolymers.

B.3.1 `PolyBase` Object

This is a virtual parent object for both `PolySet` and `IsomerSet` and contains functions in common between the two classes. All normal use of the library involves substantiating a `PolySet` object and then performing creation, digestion or statistics gathering operations using the set.

`void generate_specified_isomers(const vector<double> &`*`n`*`, const vector<string> &`*`iso`*`)`

This function provides the finest control over the creation of polymers. It allows the user to generate polymers in any manner they wish and specify exactly their creation.

n An array listing the amount of each isomer to create. For PolySet these must be postive integers, while IsomerSet allows any positive number.

iso An array of strings specifying the isomers where 'G' designates unmethyl-esterified residues and 'E' specifies methyl-esterified residues and any other characters will give an error.

`void clear()`

A loaded polymer set may consume a large amount of memory which is freed when the object is insubstantiated or via this function, which will empty the set.

`void save(const char *`*`filename`*`)`

[6] Available from `http://www.stack.nl/~dimitri/doxygen/`

It may be useful to retain polymer sets between processes etc. This function provides a means of serializing the polymer data to disk.

filename A string specifying the filename to save to. The process must have write permission to the file and will overwrite any existing file.

`void load(const char *`*`filename`*`)`

Counterpart to `save`. Clears any existing contents in the set before loading.

filename String specifying the filename to load.

`Stats *get_stats()`

Returns a `Stats` object containg the statistics of this polymer set.

returns A pointer to the Stats object. The caller is responsible for freeing the object.

`void set_stats_mask(Stats_Mask `*`m`*`)`

Generating the statistics over a polymer set can become the computional bottlenock in certain cases. If only a specific statistic is needed, setting a mask using this function means that only the unmasked statistics are generated so as to be more efficient.

m The mask to be used. There is a bit associated with each Stats function.

`Stats_Mask get_stats_mask()`

Counterpart to `set_stats_mask` .

returns The current statistics mask on this polymer set.

B.3.2 `PolySet` Object

`PolySet` inherits all `PolyBase` functions. It implements the main digestion algorithm as outlined in this report.

`void generate(Algorithm `*`alg`*` , int `*`numchains`*` , int `*`avglen`*` , int `*`dpwidth`*` , double `*`avgde`*` , double `*`dewidth`*` , vector<double> &`*`params`*` , const PolySet `*`build_from`*` = NULL)`

This function not only is the means of accessing the generation algorithms but also the reconstruction algorithms, which can be thought of as generation algorithms with additional input.

alg The algorithm to employ in the generation, options include Markovian and the reconstruction algorithms outlined in section 2.5.

numchains The number of chains to reconstruct. The actual number constructed may differ slighty from this due to way the Gaussian distributions are constructed.

avglen The DP on which to centre the Gaussian distribution of DPs of the chains being made.

dpwidth The width of the Gaussian distribution for DP.

avgde The average degree of esterification to centre the Gaussian distribution on. Argument is given as a fraction.

dewidth The width of the Gaussian distribution for DM.

params Additonal parameters that vary between construction algorithm used.

build_from When a reconstruction algorithm is used, this argument should point to another `PolymerSet` containg the fragments from which to reconstruct.

`void generate_specified_dist(const vector<int> &dist)`

This provides a quick means of filling with entirely unmethyl-esterified polymers.

dist This is an array with each position in the array counting from 1 specifying an integer value of the number of polymers of this length.

`void endopgII_digest(int numattacks, double T, const vector<double> &unmethylenergies, const vector<double> &methylenergies)`

The heart of the simulation methodology, this performs the digestion simulation according to the extended subsite model described in section 2.4. Although the proposed model postulates 7 subsites between -5 and +2 this function allows for 9 subsites between -6 and +3, one can set unused subsite binding energies to 0 as is done when simulating only unmethyl-esterified homogalacturonan.

numattacks The number of attempted bindings to perform.

T The physical temperature (in Kelvin) to perform the Boltzmann calculations at.

unmethylenergies The subsite binding energies for unmethyl-esterified residues between -6 and +3 in kJ/mol.

methylenergies The subsite binding energies for methyl-esterified residues between -6 and +3 in kJ/mol.

`void random_fragment(int numattacks)`

Randomly makes a specified number of incissions on the polymer set. A true approximation of shearing would need to account for the differing effects of different length chains and perhaps also the differing response of methyl-esterified chains.

`PolySet *filter_dp(int dpmin, int dpmax)`

Returns a newly allocated `PolySet` containing only the fragments on the current set with DP between *dpmin* and *dpmax* inclusive.

returns The newly created `PolySet`. The caller is responsible for its deallocation.

B.3.3 `IsomerSet` Object

`IsomerSet` uses a recursive, deterministic algorithm, that scales with the length of the polymer involved as $n!$ for calculation of digestion patterns. This poor scaling means that it is not usable for most calculations and `PolySet` should be used instead. However, it is retained for testing on short oligomers because it is deterministic and also can be used as a check for other algorithms. It inherents all functions of `PolyBase`.

`IsomerSet` internally stores the amount of each isomer (of every length) as a floating point. To calculate digestion it takes each given isomer and calculates the Boltzmann factor for enzyme binding at each position on the chain using the normalization as in (2.1). It uses the sum of probabilities across the whole fragment to renormalize the values to a probability of fragmentation for each position on the isomer and generates new isomers according to:

$$n = Npt \tag{B.1}$$

where n is the amount of the two new isomers created, N is the amount of the original isomer, p is the probability of fragmentation and t is the time specified to simulate to. It also associates with the newly created isomers the probability p of their creation and recursively performs fragmentation of the newly created isomers only with all the probabilities of the new isomers multiplied by p, since the two fragmentations are independant events.

`void set_time(double t)`

The algorithm calculates the digest based on a timescale where 0 is the beginning with no fragmentation occuring and 1 specifies the time where every possible first incission filling all subsites with unmethyl-esterified residues has been made. This function must be called before using `get_stats` on this set.

`void set_energy_levels(const vector<double> &um, const vector<double> &me, double T)`

The subsite energies and the temperature used in the digestion calculations must be specified. The arguments to this function are identical to their counterparts in the `PolySet` function `endopgII_digest`.

`double get_time()`

Counterpart to `set_time`.

`void set_cutoff(double c)`

In order to conserve resources, amounts of isomers calculated to occur only in small amounts of discarded. This functions sets the minimal level of polymer before it is discarded.

`double get_cutoff()`

The counterpart to `set_cutoff`.

B.3.4 Stats Object

The `Stats` object is used to store, display and generate statistical information about the polymer sets.

`Stats_Mark get_stats_mask()`

Returns the statistics mask in use when these Statistics were created. Any statistics that are masked may return unpredictable results from this set.

`void save(const char *filename)`

Often it is desired to retain statistics between sessions. This provides functionality to save to a file.

`void load(const char *filename)`

Counterpart to `save`, empties any existing statistics.

`const vector<int> & get_dp_dist()`

Returns a reference to the array containg the DP distribution, with each entry in the array (counting from 1) indicating the number of chains of this DP.

`const vector<int> & get_de_dist ()`

The DM statistics are discretised into 0.5% bins. This returns a reference to a 200 item array, where each entry represents the number of polymers with a DM between $n/2$ to $n/2 + 0.5\%$.

`list<Isomer> get_isomer_dist_n(int n)`

Returns a list of the isomers of DP *n* and the number of each isomer.

`vector<double> &get_markov(int n)`

Retrieves the Markov statistics of order *n*.

returns An array of the Markov parameters. The array is the list of transition frequencies ordered with 'G' before 'E' in dictionary fashion.

`double compare_markov_stats(const Stats &s, int ord)`

Find the σ^2 distance between the Markov parameters of order *ord* between this and another `Stats` object.

`double compare_dp_dist(const Stats &s, int maxnum)`

Returns the σ^2 comparison between the DP distributions of the two sets with no scaling. The DPs are compared up to DP *maxnum*.

`compare_t compare_dp_dist(const Stats &s, int maxnum)`

Returns the σ^2 comparison between the DP distributions of the two sets allowing scaling. The DPs are compared up to DP *maxnum*.

returns A structure containg the σ^2 result and the scaling factor.

`compare_t compare_isomer_dist(const Stats &s, int maxlen, bool scale=true)`

Returns the σ^2 distance between the isomer distributions of two `Stats` objects.

maxlen The maximum DP isomers to compare up to.

scale Set to true to allow the amounts to be scaled for a better match.

returns A structure containg the σ^2 of the match and the scaling factor employed.

B.3.5 `Matcher` Object

There are many times when it is desired to match results within a large set. Particularly when scripted from python, it is vastly more efficient if the matching can be performed by a library function to circumvent a large amount of inefficient passing between python and C++. That is the purpose of this class.

`void clear()`

Often a large amount of memory is associated with this object. This frees all associated memory and returns the object to a fresh state.

`void generate_match_pgII(Stats_Mask` *`mask`*`, int` *`startpos`*`, int` *`numpos`*`, int` *`stepsize`*`, PolySet` *`p`*`, double` *`T`*`, vector<double> &`*`unmethylenergies`*`, const vector<double> &`*`methylenergies`*`)`

Generates the digest statistics over a range of times for matching with results.

mask The `Stats_Mask` used in generating the digest statistics. This should be set with only the statistics required, as generating statistics is the most computationally expensive part of the process.

startpos The number of iterations to perform before recording timesteps.

numpos The number of separate steps to record.

stepsize The number of steps between each timestep recorded.

p The polymer set to use a starting material for the simulated digestion.

T Argument passed to `endopgII_digest`.

unmethylenergies Argument passed to `endopgII_digest`.

methylenergies Argument passed to `endopgII_digest`.

`bool is_filled()`

Returns true when `generate_match_pgII` has been filled and the object has a generated a set of patterns to match.

`match_res_t find_match_dp(const Stats &`*`s`*`, int` *`maxcompare`*`)`

Find the best match of DP distribution.

s Statistics to compare against.

maxcompare The maximum DP to compare up to.

returns A structure containing the σ^2 value of the best match, the step number it occurred at and a pointer to the `Stats` class that was the best match.

`match_res_t find_match_dp_scale(const Stats &`*`s`*`, int` *`maxcompare`*`)`

Find the best match of DP distribution while allowing scaling.

s Statistics to compare against.

maxcompare The maximum DP to compare to.

returns A structure containing the σ^2 value of the best match, the step number it occurred at, the scaling that provided the best match and a pointer to the `Stats` class that was the best match.

Appendix C - Publication

There is currently one accepted peer-reviewed manuscript published based on the results described in this report.

J. J. Hunt, R. G. Cameron, M. A. K. Williams, On the simulation of enzymatic digest patterns: the fragmentation of oligomeric and polymeric galacturonides by endo-polygalacturonase II. *Biochimica et Biophysica Acta*, 1760:1696-1703, 2006.

A copy of this publication is included for convenience.

Available online at www.sciencedirect.com
ScienceDirect

Biochimica et Biophysica Acta 1760 (2006) 1696 – 1703

BIOCHIMICA ET BIOPHYSICA ACTA
BBA
www.elsevier.com/locate/bbagen

On the simulation of enzymatic digest patterns: The fragmentation of oligomeric and polymeric galacturonides by endo-polygalacturonase II

Jonathan J. Hunt [a], Randall Cameron [b], Martin A.K. Williams [a,*]

[a] Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand
[b] USDA, Citrus and Subtropical Products Laboratory, 600 Ave S, N.W. Winter Haven, FL 33881, USA

Received 9 June 2006; received in revised form 8 August 2006; accepted 19 August 2006
Available online 3 September 2006

Abstract

A simulation methodology for predicting the time-course of enzymatic digestions is described. The model is based solely on the enzyme's subsite architecture and concomitant binding energies. This allows subsite binding energies to be used to predict the evolution of the relative amounts of different products during the digestion of arbitrary mixtures of oligomeric or polymeric substrates. The methodology has been specifically demonstrated by studying the fragmentation of a population of oligogalacturonides of varying degrees of polymerization, when digested by endo-polygalacturonase II (endo-PG II) from *Aspergillus niger*.

Keywords: Enzyme; Polysaccharide; Endo-polygalacturonase; Oligogalacturonides; Oligomeric and polymeric substrate

1. Introduction

The importance of degradative enzymes across the whole of Nature can hardly be over-emphasised. All biopolymers are, at some point, disassembled during recycling or digestion. Additionally, the routine remodelling of biopolymer fine structures in order to maximise their *in-vivo* functionality often involves carrying out modifications to their degree of polymerization (DP). Furthermore, coupling a thorough knowledge of the action of degrading enzymes with the measurement of the structures of released fragments has the potential to yield information on the fine structure of pre-digested substrates. Such an approach still holds particular promise for polysaccharide systems, where the molecular biology tools that have so advanced nucleotide and protein sequencing are simply not available. Hence, polysaccharide degrading enzymes are not only of fundamental biochemical interest but also, owing to the substrate sequence sensitivity of their enzyme–substrate interactions, offer the hope of developing a tangible link between the structures of released digest fragments and polymeric fine structure [1–9].

Models of polysaccharide–enzyme interactions routinely begin with the protein being described by a series of subsites, each capable of binding a sugar residue [10]. These subsites consist of groups of amino-acids, with their chemical nature giving rise to differential affinities between such sites and the substrate. Subsites labelled +1 and −1 indicate the position of the active site, while surrounding subsites, extending from the active site to the proximities of the binding cleft, play a role in binding the substrate [11]. By observing the preferred cutting position of substrates that are smaller than the number of subsites in the enzyme binding site, relative affinities and hence binding energies of different subsites can be obtained [12–17]. Such maps can be used to give support to molecular models of the enzyme–substrate interaction by providing experimental values that can be compared with chemical intuition based on the nature of the moieties within the amino acids of each designated subsite. Once the relative importance of hydrophobic and charge interactions is inferred the effects of solution properties might also be rationalised. Such experiments also provide data that can be compared with attempts to directly model the enzyme–substrate interaction [18]. However, despite the many positive aspects of subsite mapping procedures such results have not, as far as we are aware, been used to predict the general time-course of the enzyme digestion of arbitrary mixtures of oligomeric or polymeric substrates.

[*] Corresponding author. Tel.: +64 6 350 5799x3543.
E-mail address: m.williams@massey.ac.nz (M.A.K. Williams).

0304-4165/$–see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbagen.2006.08.022

The study of the enzymatic digestion of biopolymer substrates has also been a fertile area of research in its own right. In particular, experimental techniques for obtaining data on biopolymer digest fragments and studying the structure of complexes continue to improve [19–23]. Modelling approaches meanwhile have focussed largely on solving differential equations, [24] developing recursion schemes [25] and using methods that assume statistical models of chain structure and a strict set of prescribed enzymatic rules [1]. In this work we show that a simulation of substrate digestion, derived simply from the enzyme's subsite architecture and attendant binding energies, can successfully describe the experimental time-course of digestion of distributions of substrate molecules of varying lengths. We demonstrate this specifically by taking a well defined starting mixture consisting of substrates of various degrees of polymerization (DP), measuring the evolution of the concentration of all species during enzymatic processing, and comparing the results with a simulation. Further, we examine the sensitivity of the form of the predicted digest pattern to variations in the binding energies associated with the different subsites.

Oligogalacturonide substrates digested with an important pectin degrading enzyme, endo-polygalacturonse II (endo-PG II) from *Aspergillus niger*, has been selected as a model system, so that the generic digestion problem can be discussed while producing specific results of importance for many areas of plant science. The ubiquitous occurrence of the polysaccharide pectin in the cell walls of land plants, taken together with the large number of pectin-modifying enzymes encoded in the genomes of plants [26] and their pathogens, [27] clearly points to the importance of pectin, and pectin derived compounds, in diverse aspects of plant physiology. Indeed, the products of pectin digestion, the oligogalacturonides, play a key role in triggering plant defence [28]. Endo-PG II has been classified as a family 28 glycosyl hydrolase and proceeds by an inverting mechanism [29,30]. Its preferred substrate is polygalacturonic acid but it can also digest partially methylesterified galacturonides and co-polymers of galacturonic acid and its methylesterified counterpart (pectin), although there is an absolute requirement for subsites −1 and +1, which straddle the scission point, to be unesterified [31]. Its sequence and crystal structure have been determined and a number of mutants have been studied [32–35]. Bond cleavage frequencies have also been obtained for the degradation of galacturonic acid oligomers up to the hexamer, allowing subsite binding energies to be mapped [31]. In addition, while tetramers and trimers each have only one productive binding mode that spans the active site, the relative kinetics of fragmentation [36] can provide differential binding energy estimates and hence probe the affinity of the extra subsite involved in tetramer degradation. Fig. 1 shows the current best model of the endo-PG II subsite architecture.

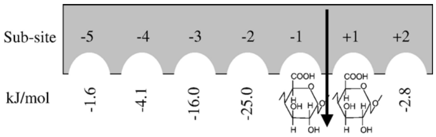


Fig. 1. The current best model of the endo-PG II subsite architecture, showing the position of the active site and the binding energies associated with the different subsites that have been used in this work. Subsite binding energies for sites −5, −4 and +2 have been taken from studies of bond cleavage frequencies; that for −3 from the relative kinetics of tetramer and trimer degradation; and that for −2 arbitrarily selected.

2. Materials and methods

2.1. Oligomeric substrate

A set of oligogalacturonides containing degrees of polymerization between 2 and 17 was generated by selective precipitation of a partially digested polygalacturonic acid (PGA) as described previously [37]. Briefly, a 2% (w/v) solution of the free acid of PGA in 50 mM lithium acetate at pH 4.7 was digested with 0.05 U mL^{-1} EPG (Lot 00801, Megazyme International Ireland Limited, Bray, Ireland) for 4.5 h at room temperature with constant stirring. The pH of the digested LiPGA was lowered to 2.0 with concentrated HCl and stored overnight at 4 °C. The precipitate was pelleted by centrifugation at 23,500×*g* for 30 min at 4 °C. The pelleted precipitate, representing high DP oligomers was removed. The supernatant, containing the low- and medium-DP oligomers, was brought to 50 mM sodium acetate (NaOAc) and 22.5% EtOH and then placed at 4 °C overnight to precipitate the medium-DP fragments (DP 8–24). Following centrifugation as described above, the supernatant was decanted. The pelleted material was solubilized in 50 mM LiOAc and then re-precipitated by adjusting the solution to 50 mM NaOAc and 22.5% EtOH. This material, representing the medium-DP oligomers, was centrifuged again and the pellets were solubilized in 50 mM LiOAc and stored at 4 °C. Initial characterisation was carried out using HPAEC with an evaporative light scattering (ELS) detector [38,39] and subsequently by capillary electrophoresis using UV detection [20,22].

2.2. Enzymes

Endo-PG II (EC 3.2.1.15) from *Aspergillus niger* was prepared as described previously [40]. Digests were carried out by incubating 1.0 mL of the oligogalacturonide mixture, at a total concentration of 3% galacturonic acid, and pH 4.2, with 20 μL of the enzyme solution, that was in turn generated by diluting 25 μL of a 7.5 mg mL^{-1} protein stock into 2.0 mL of 50 mM acetate buffer at pH 4.2. All experiments were carried out at (30±1) °C by keeping the digest mixture in a waterbath. At various times aliquots were removed, the enzyme denatured by rapid heating to 95 °C, and the current concentrations of the various oligomeric species recorded using CE. Thus a picture of the time-course of the digestion was recorded.

2.3. Capillary electrophoresis

Experiments to separate, identify and quantify oligogalacturonides of varying degrees of polymerization were carried out using an automated CE system (HP 3D), equipped with a diode array detector. Electrophoresis was carried out in a fused silica capillary of internal diameter 50 μm and a total length of 46.5 cm (40 cm from inlet to detector). The capillary incorporated an extended light-path detection window (150 μm) and was thermostatically controlled at 25 °C. Phosphate buffer at pH 7.0 was used as a CE background electrolyte (BGE) and was prepared by mixing 0.2M Na_2HPO_4 and 0.2M NaH_2PO_4 in appropriate ratios and subsequently reducing the ionic strength to 90 mM. At pH 7.0 galacturonic acid residues are fully charged and while the oligomers are susceptible to base-catalysed β-elimination above pH 4.5, no problems were encountered during the CE runs of some 20 min at room temperature. All new capillaries were conditioned by rinsing for 30 min with 1 M NaOH, 30 min with a 0.1 M NaOH solution, 15 min with water and 30 min with BGE. It was found that for the samples used in this study similar harsh washing of the capillary was also required between runs. Detection was carried out using UV absorbance at 191 nm with a bandwidth of 2 nm. Samples were loaded hydrodynamically (various injection times at 5000 Pa, typically giving injection volumes of the order of 10 nL), and typically electrophoresed across a potential difference of 20 kV. All experiments were carried out at normal polarity (inlet anodic) unless otherwise stated. Samples of mono-, di-, and trigalacturonic acid, used as standards, were obtained from Sigma-Aldrich Corp., St. Louis, MO, USA.

2.4. Simulation

Oligomeric distributions of starting material were modelled by a simple one dimensional array in which the elements were assigned to a character, L, M or R, denoting the left (non-reducing end), middle, and right (reducing end) of chains. This basic representation is similar to that used previously in work where more ad-hoc representations of the enzymes action were investigated [25]. Enzymatic encounters were presumed to occur at random. Each encounter consisted of selecting a substrate binding position for subsite-1 and summing the binding energies for all filled subsites in this position. It was assumed that each subsite binding energy was unaffected by the presence or not of substrate at other sites. This summed energy was then used to generate a Boltzmann factor that determined the probability of successful binding. This conversion to probability included a normalisation of the binding energy to that of 7 filled subsites (the accepted extent of the endo-PG subsite architecture, as shown in Fig. 1), so that such a contact always results in a successful binding event. Whether the binding was successful or not was determined by comparing this probability with the output of a random number generator from GNU Scientific Library [41]. Using this random number generator ensured that the frequency of low-probability events was simulated accurately. In successful binding cases, if subsites +1 and −1 were both covered, then a scission was made and the array elements reassigned accordingly in order to denote the presence of new chain ends. After a selected number of iterations the array was interrogated to produce a distribution of fragment lengths, and it was this simulated data that was compared with the experimentally measured time-course. The simulation was carried out on an Intel P4 computer and takes only seconds to complete a simulated digest evolution starting with 10^4 substrate molecules. The code was written in C+, compiled with Intel C Compiler (Version 9), and was executed using the scripting language Python within a Linux environment. Initially, solely tetramers, pentamers, or hexamers were fed to the enzyme model to ensure that the relevant number of different product possibilities (i.e. the known bond cleavage frequencies) were reproduced.

3. Results and discussion

Fig. 2 shows the quantification of the starting oligogalacturonide distribution as measured by both HPAEC with ELS detection and CE using UV absorption. The agreement between these two quite different methodologies is remarkably good and provides further confidence in both techniques. The CE data were processed in order to quantify the amount of different oligomers present, as described in detail elsewhere [22], ensuring that the peak area was divided by the migration time in order to account for the different time that the separated species spend travelling past the detection window [42]. The peaks were unequivocally identified by spiking with commercially available samples of dimer and trimer. CE has been used previously in the study of oligogalacturonides and it is known that above a certain molecular length the hydrodynamic friction and charge scale symmetrically with the introduction of further sugar residues, leading to a loss of resolution. While the exact DP at which this occurs and the modelling of the mobility forms part of ongoing work it has previously been predicted to occur at around DP 15–20 [43]. Bearing this in mind the current oligomer mixture was selected as the starting substrate in order to ensure it would be possible to record the time-course of all starting oligomers.

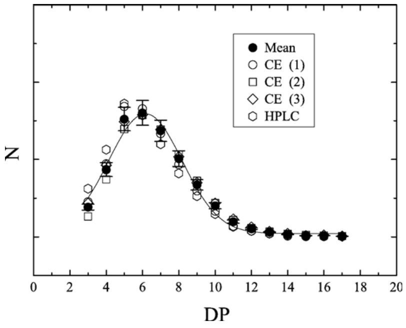


Fig. 2. The relative number, N, of starting oligogalacturonides with varying degrees of polymerization, DP, as measured by both HPAEC and CE, as described in the text.

The initial input for the enzymatic fragmentation model was the relative number of molecules of each oligomer, generated from the average of the HPAEC data and three CE datasets. The only other parameters used are a best estimate set of the binding energies for the endo-PG II subsites, shown in Fig. 1. Preliminarily, the number of starting substrate chains that could be used in order for the simulation to generate reproducible results was determined. For a chosen initial number of chains a digest simulation was run for a fixed number of time-steps, the value of which was selected based on preliminary experimental data, and the resulting pattern recorded. This was then repeated with the same number of chains so that 10 estimates of the resulting pattern (the relative number of the different DP oligomers existing at a particular time) were obtained. From these a mean pattern was obtained, and the variance, σ^2, of the 10 repeats around this mean determined. This process was repeated for sets of solutions carried out with different numbers of starting chains, (keeping the number of iterations per chain constant), and the results are shown in Fig. 3. It is clear that the reproducibility of the calculation becomes better as more chains are used, as expected. An optimal starting value, taking into account the extra resources required to perform the calculations for a greater number of chains, was deemed to be 10^4 substrate molecules, and this was used for all subsequent simulations.

Experimental data were obtained for the evolution of the enzyme digest pattern at 0, 2, 4, 6, 8, 10, 15, 30, 45, 266, and 13000 min as described in Materials and methods. A simulation of the digest process was also run. The simulated digest pattern data were typically stored after each additional 200 iterations. The calculation was terminated after 6×10^6 iterations at which time the changes in the digest pattern consisted solely of trimer being slowly fragmented to monomer and dimer, with a further 10^6 iterations generating concentration changes of a fraction of a percent. It is worth commenting at this point that it is experimentally observed that the trimer is indeed digested to the dimer and monomer but at a rate some 20 times slower than the tetramer degradation [36]. This rate is sufficiently slow as to essentially decouple the

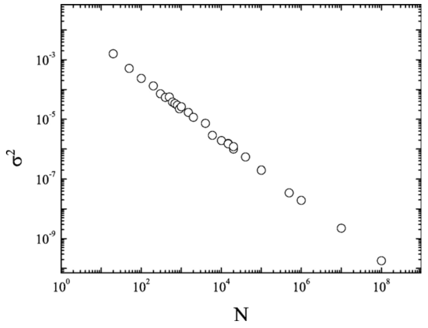


Fig. 3. The variance within sets of 10 repeat simulations carried out with different numbers of starting chains.

trimer processing from the rest of the digestion and therefore, in many previous studies, the relatively long-lived situation with a similar ratio of monomer, dimer and trimer has been taken practically to be the end of the digestion. In the model described here these two regimes emerge naturally as a consequence of the differential binding energies of the species of various lengths.

Initially the experimental and simulated digests were compared by eye in order to assess whether there were calculated digest pattern solutions during the simulated fragmentation process that manifested the experimentally found ratios of different DP oligomers. Encouraged by a reasonable agreement in the general time-course of the digest, a simple searching algorithm was subsequently written that took each experimentally determined digest pattern and searched through the calculated data in order to find the particular simulation (corresponding to a certain number of iterations) that best matched, as determined by a simple minimisation of the standard deviation of the experimental and simulated datasets. As a prelude to the formation of the standard deviation the patterns concerned were both normalised so that the total amount of galacturonic acid was the same in both cases. A scaling of the experimental data total amount was permitted during the pattern matching to allow for the possibility that not all the material was detected, but was found to not be significantly different from one when the matches were made.

Fig. 4 shows the comparison of the experimental and calculated digest patterns for these best fits. In general there is excellent agreement, particularly when experimental uncertainties in the relative amounts are taken into account. It should also be born in mind that while it is the number of molecules that is represented in these plots it is the absorbance that is actually detected in the CE experiments reported herein, which is proportional to the degree of polymerization (with a minor exception for the monomer, as described previously elsewhere [22]). This means that small amounts of the dimer and monomer, as predicted at the beginning of the simulation, are difficult to detect experimentally. Nevertheless, the agreement is good. It is interesting to use the correspondence of the simulated and experimental data to map the iteration number onto real-time, as shown in Fig. 5. The plot can be seen to show an excellent linear relationship up to 45 min, implying that despite the relatively simplistic nature of the model including the random nature of the encounters and the lack of dynamics and molecular detail, nevertheless, the previously measured subsite

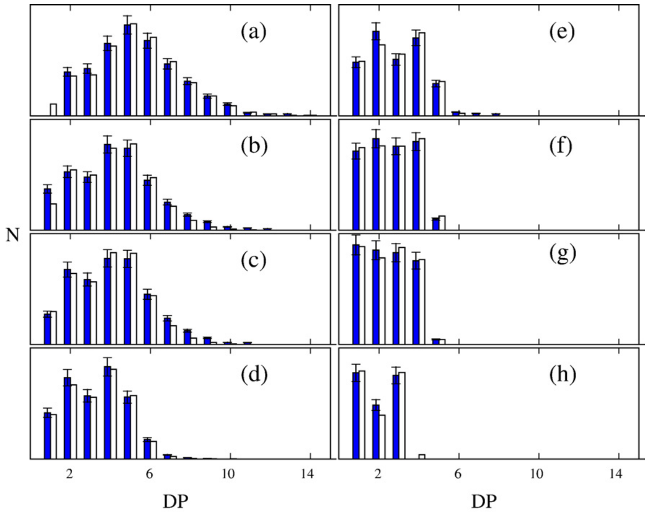


Fig. 4. The comparison of the experimental and calculated digest patterns at selected times during the fragmentation of a starting oligogalacturonide solution with endo-PG II. Experimental data is shown with estimated uncertainties. (a) 2, (b) 4, (c) 6, (d) 10, (e) 15, (f) 30, (g) 45, (h) 13,000 min.

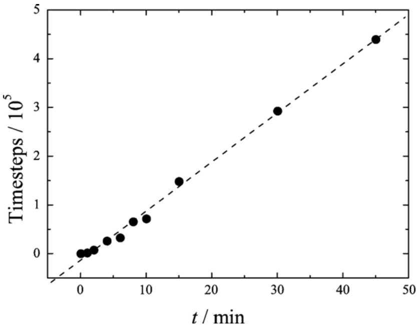


Fig. 5. The number of iterations required in the simulation in order to generate a simulated digest pattern that best matches that found experimentally at the corresponding time.

binding energies do describe the time dependence of the real reaction well. The last two points (266, 13,000 min) were more difficult to match to a specific time owing to the slowness of the evolution, but simulations terminated at the number of iterations that correspond to such long times according to the data in Fig. 5 were indeed found to be consistent with the experimental data.

While in detail the value of the gradient of such a plot will depend, among other things, on enzyme and substrate concentrations it could be argued that performing one time-step in the simulation might be thought of as modelling a substrate–enzyme contact. The mapping of real-time onto equivalent simulation time-steps then, as shown in Fig. 5 suggests that there is roughly one cut attempt every 6 ms, that is; there is an encounter between enzyme and substrate on the order of milliseconds. A simple order of magnitude calculation of the collision frequency can be made by distributing the enzyme molecules homogeneously throughout the solution and calculating the time it would take the oligogalacturonides to move between them (the diffusion coefficient of the protein is expected to be at least ten times less than the sugar oligomers). Taking the enzyme concentration as 100 nM and an average substrate diffusion coefficient of 10^{-10} ms^{-2} gives an average encounter frequency in the tenths of milliseconds regime that agrees reasonably for such a simple model.

We have shown then that the binding energies inferred from bond cleavage and kinetic data on small substrates ($\mathrm{DP} < 7$) can be utilised in a physically realistic model in order to reproduce the temporal evolution of the digest patterns of arbitrary mixtures of considerably longer starting substrates. This implies that, for this system at least, the binding energy of the subsites is not significantly modified by the presence or not of substrate in neighbouring binding sites, and further, that they are not significantly altered by any straining of the substrate.

Further simulations have also been carried out in order to assess the sensitivity of the predicted digest patterns to the values of the binding energies and hence address the question—how much variation could you have in the binding energy values before you obviously modify the form of the simulation and cannot match the experiments? The basic philosophy of these was the following. Twelve “times” were chosen and at each of these a digest pattern was simulated with the base, best estimate values of the endo-PG II subsite energies as given in Fig. 1. These “times” were in fact the number of time-steps that gave digest patterns in the simulation that corresponded to experimental data recorded, some of which were shown in Fig 4. Subsequently the binding energies were altered in a known manner and the simulation was re-run, this time calculating the digest pattern at each 200 iterations. A searching algorithm then found the best match pattern generated by the simulation using the modified binding energies to the base case. The difference between this digest pattern, generated with modified energy variants and that calculated with the best binding energies, was then quantified by means of a simple variance, as described above.

Fig. 6 shows the results for a representative number of these calculations. Owing to the complexity of the combinatorial space of five binding sites, a simple approach was taken in which the binding energy of each subsite was changed in isolation, between the limits of the original value minus 50% to the original value plus 50%, at 1% increments. It can be seen that at very short times there is, as expected, little change in the calculated digest pattern even with large differences in the binding energies, simply reflecting the fact that there has been little change in the original distribution at this point. However, by 6 min it is soon apparent that changing the binding energies of subsites +2, −4 and −5 leads to a substantial change in the digest pattern. In order to generate a significant change however, alterations in the binding energies of the relevant subsites need to be of the order of 10–20%. It is also clear that subsites −2 and −3 do not play a significant role in determining the form of the digest pattern in this region, as expected, at a time when no degradation of tetramers and trimers will have occurred. This trend essentially continues for the simulations run to a number of time-steps equivalent to 10 min. Still later in the time-course of the simulated digest, by 30 min, it becomes apparent that there has been such a substantial time period involving the fragmentation of the smaller species that binding energy value of subsite −5 becomes less important than +2 and −4 in terms of generating the final pattern. Towards the end of the digestion process the fragmentation pattern can be seen additionally to be sensitive to increases in the affinity at site −3.

In addition to carrying out these simulations, performed in order to show that the calculated form of the digest pattern is sensitive to the input values used for the binding energies, the ability to detect variances of the predicted magnitudes experimentally has also been examined. Fig. 7 shows the variances of predicted digest patterns, calculated with different binding energy variants, from that experimentally observed at 10 min. The plot clearly shows that if the binding energies of sites −5, −4 and +2 were modified by more than around 10% then the match *to the experimental data* would be detectably worse. In conclusion then, the fact that the simulated enzymatic fragmentation calculated based on subsite binding energies matches experimental data is indeed significant; changes in the affinities of individual binding sites of greater

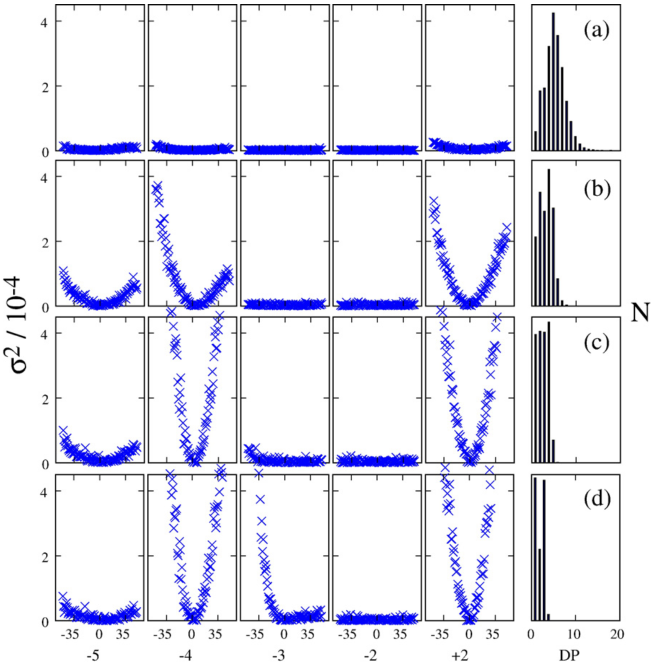


Fig. 6. The variance between digest patterns generated with the best estimate subsite binding energies and those calculated with the indicated subsite energy specifically modified in isolation. The variation is carried out between the limits of the original value minus 50% to the original value plus 50%, at 1% increments. (a) 2 min, (b) 10 min, (c) 30 min, (d) 13,000 min.

than 10–20% would have produced detectable discrepancies in the ability of the simulated data to successfully match the experimental time-course.

4. Conclusion

A model for predicting the time-course of enzymatic digestions of arbitrary mixtures of oligomeric or polymeric substrates has been described. The success of such a simulation has been demonstrated by modelling the fragmentation of a population of oligogalacturonides of varying degrees of polymerization, when digested by endo-polygalacturonse II (endo-PG II) from *Aspergillus niger*. This allows subsite binding energies obtained from the results of bond cleavage and kinetic experiments, carried out on oligomers with lengths less than the number of subsites in the binding cleft, to be used to predict the evolution of the relative amounts of different products during the digestion of mixtures of arbitrarily long oligomeric or polymeric substrates. The different phases of digestion within the fragmentation process naturally evolve from the differential binding energies of different DP substrates. This simple homo-polymeric case

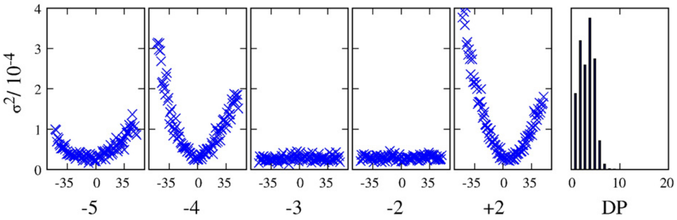


Fig. 7. The variances of predicted digest patterns, calculated with different binding energy variants, from that experimentally observed at 10 min. The plot clearly shows that if the binding energies of sites −5, −4 and +2 were modified by more than around 10% then the match to the experimental data would be detectably worse.

serves to demonstrate the promise of the technique and current work is involved in extending the modelling to the fragmentation of co-polymers, in which the different residue types can have different binding energies to the enzyme subsites.

References

[1] B.V. McCleary, A.H Clark, I.C.M. Dea, D.A Rees, The fine structure of carob and guar galactomannans, Carbohydr. Res. 139 (1985) 237–260.

[2] I.C.M. Dea, A.H. Clark, B.V. McCleary, Effect of galactose-substitution patterns on the interaction properties of galactomannans, Carbohydr. Res. 147 (1986) 275–294.

[3] J.S. Grant Reid, M. Edwards, M.J. Gidley, A.H. Clark, Enzyme specificity in galactomannan biosynthesis, Planta 195 (1995) 489–495.

[4] P.J.H. Daas, K. Meyer-Hansen, H.A. Schols, G.A. DeRuiter, A.G.J. Voragen, Investigation of the non-esterified galacturonic acid distribution in pectin with endopolygalacturonase, Carbohydr. Res. 318 (1999) 135–145.

[5] G. Limberg, R. Körner, H.C. Buchholt, T.M.I.E. Christensen, P. Roepstorff, J.D. Mikkelsen, Analysis of pectin structure part 1—Analysis of different de-esterification mechanisms for pectin by enzymatic fingerprinting using endopectin lyase and endopolygalacturonase II from A-niger, Carbohydr. Res. 327 (2000) 293–307.

[6] G. Limberg, R. Körner, H.C. Buchholt, T.M.I.E. Christensen, P. Roepstorff, J.D. Mikkelsen, Analysis of pectin structure part 3—Quantification of the amount of galacturonic acid residues in block sequences in pectin homogalacturonan by enzymatic fingerprinting with exo- and endo-polygalacturonase II from *Aspergillus niger*, Carbohydr. Res. 327 (2000) 321–332.

[7] P.J.H. Daas, A.G.J. Voragen, H.A. Schols, Investigation of the galacturonic acid distribution of pectin with enzymes part 2—Characterization of non-esterified galacturonic acid sequences in pectin with endopolygalacturonase, Carbohydr. Res. 326 (2000) 120–129.

[8] P.J.H. Daas, G.J.W.M. vanAlebeek, A.G.J. Voragen, H.A. Schols, Determination of the distribution of non-esterified galacturonic acid in pectin with endo-polygalacturonase, in: P.A. Williams, G.O. Philips (Eds.), Gums and Stabilisers for the Food Industry, 10, ed, The Royal Society of Chemistry, Cambridge, 2000, pp. 3–19.

[9] P.J.H. Daas, B. Boxma, A.M.C.P. Hopman, A.G.J Voragen, H.A. Schols, Nonesterified galacturonic acid sequence homology of pectins, Biopolymers 58 (2001) 1–8.

[10] J.A. Thoma, C. Brothers, J. Spradlin, Subsite mapping of enzymes. Studies on *Bacillus subtilis* amylase, Biochemistry 9 (1970) 1768–1775.

[11] G.J. Davies, K.S. Wilson, B. Henrissat, Nomenclature for sugar-binding subsites in glycosyl hydrolases, Biochem. J. 321 (1997) 557–559.

[12] J.A. Thoma, G.V.K. Rao, C. Brothers, J. Spradlin, L.H. Li, Subsite mapping of enzymes, J. Biol. Chem. 246 (1971) 5621–5635.

[13] Y. Nitta, M. Mizushima, K. Hiroma, S. Ono, Influence of molecular structures of substrates and analogues on Taka-amylase-A catalyzed hydrolyses .1. Effect of chain length of linear substrates, J. Biochem. 69 (1971) 567–576.

[14] J.D. Allen, J.A. Thoma, Subsite mapping of enzymes. Depolymerase computer modelling, Biochem. J. 159 (1976) 105–120.

[15] T. Suganuma, R. Matsuno, M. Ohnishi, K. Hiromi, A study of the mechanism of action of Taka-amylase A on the linear oligosaccharides by product analysis and computer simulation, J. Biochem. 84 (1978) 293–316.

[16] E. Bonnin, A. Le Gof, R. Körner, G.-J.W.M. Van Alebeek, T.M.I.E. Christensen, A.G.J. Voragen, P. Roepstorff, C. Caprari, J.-F. Thibault, Study of the mode of endopolygalacturonase from *Fusarium moniliforme*, Biochim. Biophys. Acta 1526 (2001) 301–309.

[17] G. Gyöngyi, G. Hovánski, L. Kandra, Subsite mapping of the binding region of α-amylases with a computer program, Eur. J. Biochem. 269 (2002) 5157–5162.

[18] G. André-Leroux, D. Tessier, E. Bonnin, Action pattern of *Fusarium moniliforme* endopolygalacturonase towards pectin fragments: comprehension and prediction, Biochim. Biophys. Acta 1749 (2005) 53–64.

[19] R. Körner, G. Limberg, T.M.I.E. Christensen, J.D. Mikkelson, P. Roepstorff, Sequencing of partially methyl-esterified oligogalacturonates by tandem mass spectrometry and its use to determine pectinase specificities, Anal. Chem. 71 (1999) 1421–1427.

[20] M.A.K. Williams, G.M.C. Buffet, T.J. Foster, Analysis of partially methyl-esterified galacturonic acid oligomers by capillary electrophoresis, Anal. Biochem. 301 (2002) 117–122.

[21] D. King, M. Lumpkin, C. Bergmann, R. Orlando, Studying protein–carbohydrate interactions by amide hydrogen/deuterium exchange mass spectroscopy, Rapid Commun. Mass Spectrom. 16 (2002) 1569–1574.

[22] A. Ström, M.A.K. Williams, On the separation, detection and quantification of pectin derived oligosaccharides by capillary electrophoresis, Carbohydr. Res. 339 (2004) 1711–1716.

[23] F. Goubet, A. Ström, P. Dupree, M.A.K. Williams, An investigation of pectin methylesterification patterns by two independent methods: capillary electrophoresis and polysaccharide analysis using carbohydrate gel electrophoresis, Carbohydr. Res. 340 (2005) 1193–1199.

[24] K. Ostgaard, B.T. Stokke, B. Larson, Numerical model for alginate block specificity of mannuronate lyase from Haliotis, Carbohydr. Res. 260 (1994) 83–98.

[25] M.A.K. Williams, G.M.C. Buffet, T.J. Foster, I.T. Norton, Simulation of endo-PG digest patterns and implications for the determination of pectin fine structure, Carbohydr. Res. 334 (2001) 243–250.

[26] B. Henrissat, P.M. Coutinho, G.J. Davies, A census of carbohydrate-active enzymes in the genome of *Arabidopsis thaliana*, Plant Mol. Biol. 47 (2001) 55–72.

[27] R.P. DeVries, J. Visser, Aspergillus enzymes involved in the degradation of plant cell wall polysaccharides, Microbiol. Mol. Biol. Rev. 65 (2001) 497–522.

[28] B.L. Ridley, M.A. O'Neill, D. Mohnen, Pectins: structure, biosynthesis and oligogalacturonide related signaling, Phytochemistry 57 (2001) 929–967.

[29] B. Henrissat, G.J. Davies, Structural and sequence-based classification of glycoside hydrolases, Curr. Opin. Struct. Biol. 7 (1997) 637–644.

[30] B. Henrissat, Enzymology of cell-wall degradation, Biochem. Soc. Trans. 26 (1998) 153–156.

[31] J.A.E Benen, H.C.M. Kester, J. Visser, Kinetic characterization of *Aspergillus niger* endopolygalacturonase I, II and C, Eur. J. Biochem. 259 (1999) 577–585.

[32] Y. Van Santen, J.A.E. Benen, K-H. Schröter, K.H. Kalk, S. Armand, J. Visser, B.W. Dijkstra, 1.68-A Crystal structure of endopolygalacturonase II from *Aspergillus niger* and identification of active site residues by site-directed mutagenesis, J. Biol. Chem. 274 (1999) 30474–30480.

[33] S. Pagès, W.H.M. Heijne, H.C.M. Kester, J. Visser, J.A.E. Benen, Subsite mapping of *Aspergillus niger* endopolygalacturonase II by site-directed mutagenesis, J. Biol. Chem. 275 (2000) 29348–29353.

[34] S. Armand, M.J.M. Wagemaker, P. Sánchez-Torres, H.C.M. Kester, Y. van Santen, B.W. Dijkstra, J. Visser, J.A.E. Benen, The active site topology of *Aspergillus niger* endopolygalacturonase II as studied by site-directed mutagenesis, J. Biol. Chem. 275 (2000) 691–696.

[35] S. Pagès, H.C.M. Kester, J. Visser, J.A.E. Benen, Changing a single amino acid residue switches processive and non-processive behaviour of *Aspergillus niger* endopolygalacturonase I and II, J. Biol. Chem. 276 (2001) 33652–33656.

[36] L. Rexová-Benková, The size of the substrate binding site of an *Aspergillus niger extracellular* endopolygalacturonase, Eur. J. Biochem. 39 (1973) 109–115.

[37] R.G. Cameron, G. Luzio, E. Baldwin, J. Narciso, A. Plotto, Production of narrow-range size-classes of polygalacturonic acid oligomers, Proc. Fla. State Hort. Soc. 118 (2005) 406–409.

[38] R.G. Cameron, K. Grohmann, Separation, detection and quantification of galacturonic acid oligomers with a degree of polymerization greater than 50, J. Liq. Chromatogr. Relat. Technol. 28 (2005) 559–570.

[39] R.G. Cameron, A.T. Hotchkiss, S.W. Kauffman, K. Grohmann, Utilisation of an evaporative light scattering high-performance size-exclusion chromatography acid oligomers, J. Chromatogr., A 1011 (2003) 227–231.

[40] H.J.D. Bussink, H.C.M. Kester, J. Visser, Molecular cloning, nucleotide-

sequence and expression of the gene encoding prepro-polygalacturonase II of *Aspergillus niger*, FEBS Lett. 273 (1990) 127–130.

[41] M. Matsumoto, T. Nishimura, Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator, ACM Trans. Model. Comput. Simul. 8 (1998) 3–30.

[42] D.M. Goodall, S.J. Williams, D.K. Lloyd, Quantitative aspects of capillary electrophoresis, Trends Anal. Chem. 10 (1991) 272–279.

[43] M.A.K. Williams, T.J. Foster, H.A. Schols, Elucidation of pectin methylester distributions by capillary electrophoresis, J. Agric. Food Chem. 51 (2003) 1777–1782.

Appendix D - Associated Electronic Data

Below you should find attached a CD-ROM containing the following content along with manifest files with more details on contents:

- Electronic version of this report as a Adobe Portable Document Format (pdf).
- A video clip (several different formats) showing a comparison of experimental and simulated results for the digestion over time of unmethyl-esterified homogalacturonan.
- Full LaTeXsourcecode for this report.
- Full source code for BJPS library.
- Full supporting python code and raw data for the simulations described in this report.
- A Subversion[1] repository containing the change history for the development of BJPS.

[1] Subversion is a open-source version control system for which full source server and client are available from `http://subversion.tigris.org/`.