Modelling the effects of domestication in Wheat through novel computer vision techniques

Author:	Nathan Hughes (nah26@aber.ac.uk)
Supervisor:	Dr. Wayne Aubrey (waa2@aber.ac.uk)
Degree Scheme	G401 (Computer Science)

Date:	April 6, 2018
Revision:	0.1
Status:	Draft

This report was submitted as partial fulfilment
of a BSc degree in Computer Science (G401)

Declaration of originality

I confirm that:

- This submission is my own work, except where clearly indicated.
- I understand that there are severe penalties for Unacceptable Academic Practice, which can lead to loss of marks or even the withholding of a degree.
- I have read the regulations on Unacceptable Academic Practice from the University's Academic Quality and Records Office (AQRO) and the relevant sections of the current Student Handbook of the Department of Computer Science.
- In submitting this work I understand and agree to abide by the University's regulations governing these issues.

Name .....

Date .....

Consent to share this work

By including my name below, I hereby agree to this dissertation being made available to other students and academic staff of the Aberystwyth Computer Science Department.

Name .....

Date .....

Contents

List of Tables

List of Figures

List of Listings

Chapter 1

Introduction, Analysis and Objectives

This project aims to answer a biological research question through the use of computer science, whilst also creating a software suite which will enable further studies to be carried out with ease.

Primarily the focus has been on the data science elements of my degree, creating, cleaning and discerning meaning in it.

Additionally, as this is very much multi-disciplinary research, specific terms and definitions have been outlined in the *glossary* (table:7.3).

Background

Western society and agriculture has been dominated by the ability to create successful crops for the past 10,000 years [1]. Of these crops wheat is considered to be one of the most vital and is estimated to contribute to 20% of the total calories and proteins consumed worldwide, and accounts for roughly 53% of total harvested area (in China and Central Asia) [2].

During domestication, the main traits selected for breeding were most likely plant height and yield. This meant that important non-expressed traits such as disease resistance and drought tolerance were often neglected and lost overtime.

Whilst the choices made for selective breeding were successful, effects are now being felt as it is estimated that as much as a 5% dip is observed yearly on wheat production [2]. This decrease in efficiency is attributed to climate change bringing in more hostile conditions, which these elite and thoroughly domesticated genotypes are unprepared for.

Modern breeding programs have had some success in selecting primitive undomesticated genotypes and using them to breed back in useful alleles which would have been lost during domestication [3].

To create a better understanding of wheat and in particular how its traits have evolved over time, through domestication, a newly developed method of µ-CT image analysis [4] for crops was used to generate data.

These data were then used to test a number of hypothesis on the effects of domestication in the wheat genome.

Significance to Current Research

The biological interest in this area has been expressed in several areas of research [5], it is proposed that the key to unlocking diversity in the wheat genus lies in these ancestor, undomesticated species [6].

This research has the potential to be useful in several areas including: crop breeding; disease resistance; environmental stress.

The individual images in figure:1.1 show, at a glance, the diversity and also the difference in the wild and cultivated (domesticated) species. This work allows for these differences to be quantified and evaluated into useful metrics for answering research based questions.

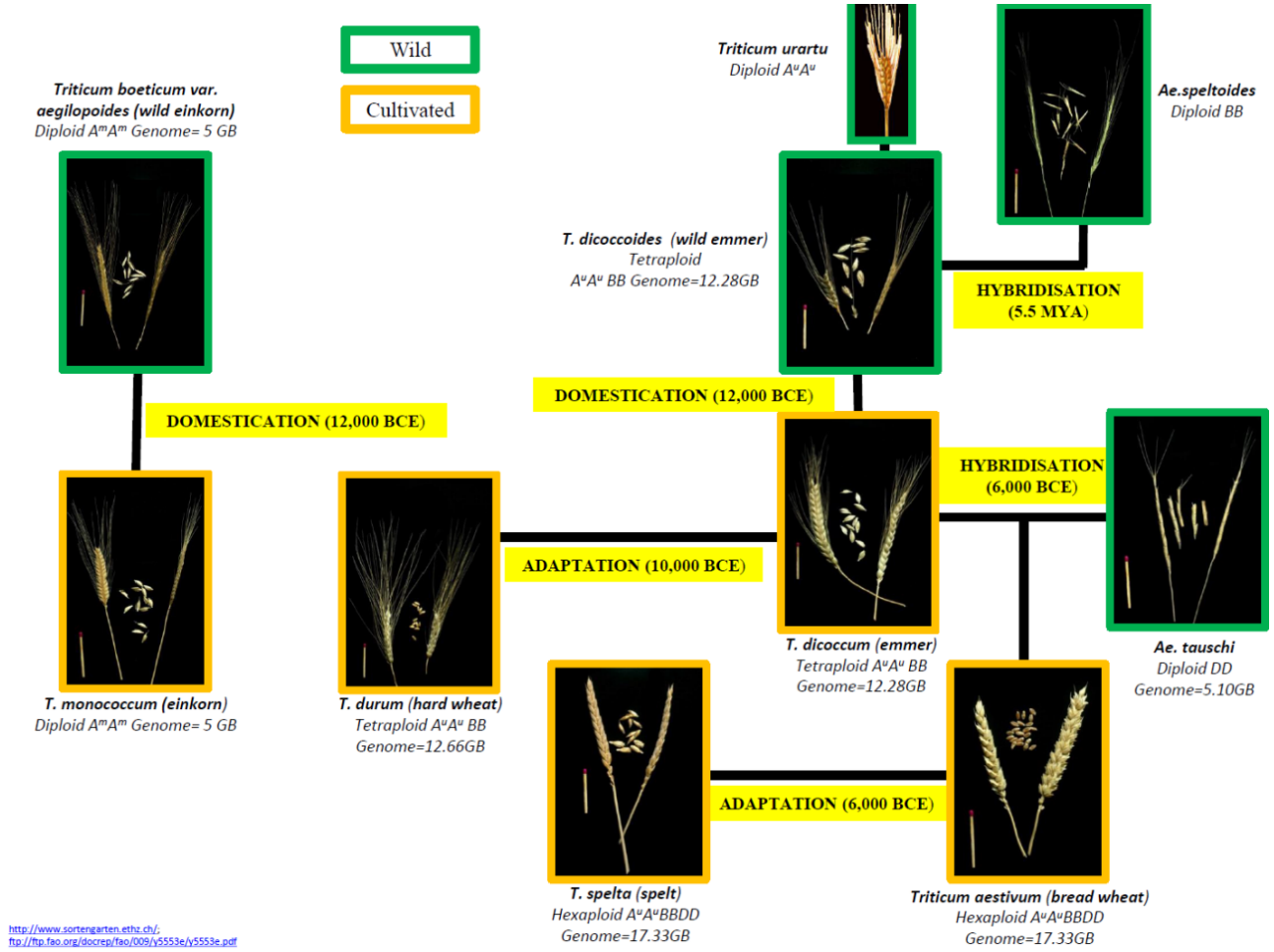

Figure 1.1: Phylogeny of wheat genotypes (Provided by Dr. Hugo Oliveira)

Hypothesis

To provide a full spectrum of analysis the null-hypothesis of this work is presented as investigating if there are morphometric differences in the seeds of several wheat varieties outlined in figure:1.1.

The comparison pairs are as follows:

1. Wild Einkorn and Domesticated Einkorn
2. Wild Emmer and Domesticated Emmer
3. Bread wheat and Spelta wheat
4. Domesticated Emmer and Durum
5. Wild Einkorn and Wild Emmer

Aim and Objectives

The overarching aim of this project has been to create several pieces of software which aid in answering the biologically significant questions outlined. As well as to prove/disprove the hypothesis stated.

The software created is robust in order to duplicate results and is flexible as to allow for further studies to be carried out and to use the same method.

Novel additions have been made to existing image analysis libraries in order to make them more flexible for this project.

Furthermore, the library written allows for easy data organisation and automation of otherwise difficult tasks such as concatenating data from multiple sources and graphing of information. Full documentation and integrated testing allows for a suite of tools which can be built upon in future and reduce the amount of effort required for similar studies to be carried out and analysed.

These aims have a focus on the phenotypic attributes generated from customised image analysis software [4] and can be seen in figure:1.2.

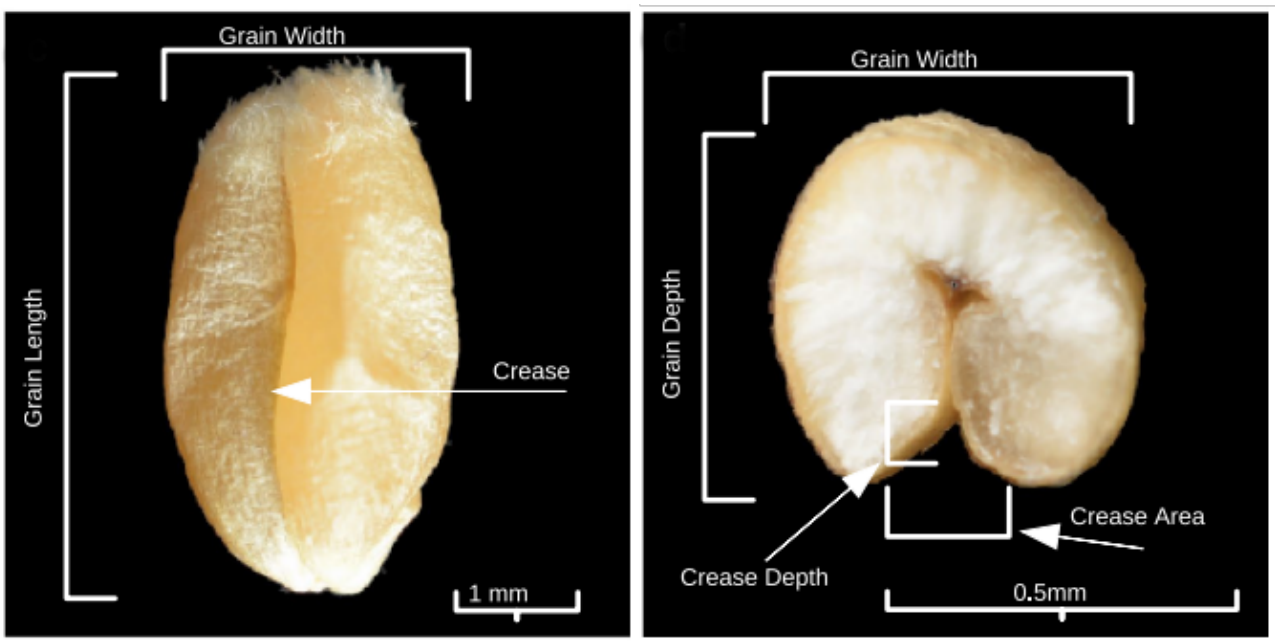

Figure 1.2: wheat grain labelled (*left*), wheat grain cut in half (*right*)

Deliverables

This project provides three final deliveribles:

1. A flexible software suite written in *Python* that provides a standardised method for analysing and interpreting µ-CT data output.
2. A Graphical User Interface (GUI) which offers a point and click method for data gathering, graphing and manipulating µ-CT data, using the library from deliverable 1 as a backend.
3. Answers to the proposed questions (hypothesis), the *Results* and *Discussion* sections of this report provides this.

Chapter 2

Software Design, Implementation and Testing

Software Development Methodology

Designing Process

Implementation

Testing

Chapter 3

Methods

Data Pipeline

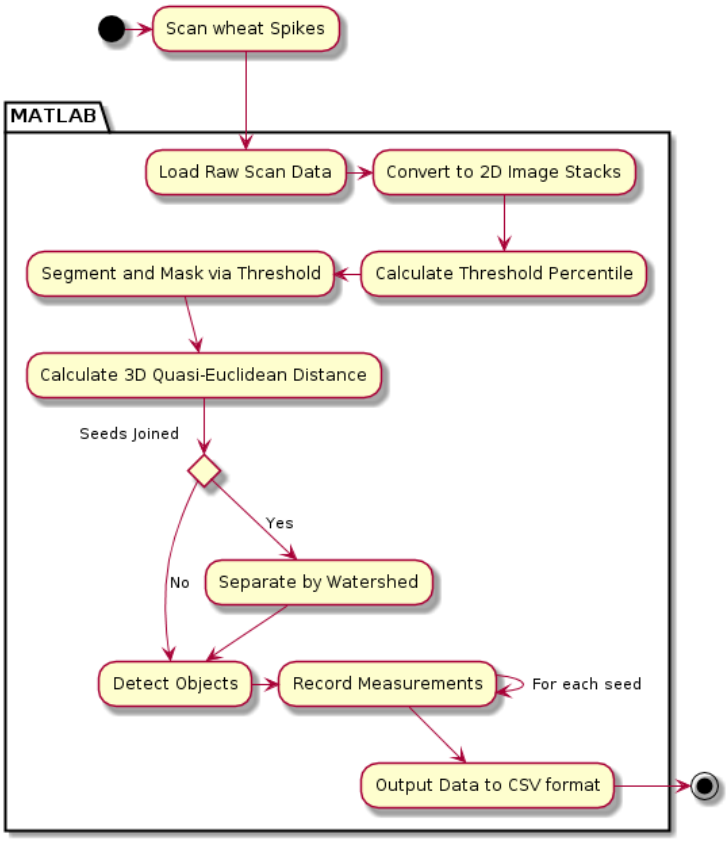

Figure 3.1: Image Processing Pipeline

Improvements to 3D imaging software

New Watershed Algorithm

In order to solve the problem of misidentified and joint seeds, from the primitive collection, a *quasi-euclidean* distance transform was implemented into the analysis pipeline (figure:3.1). This provided much

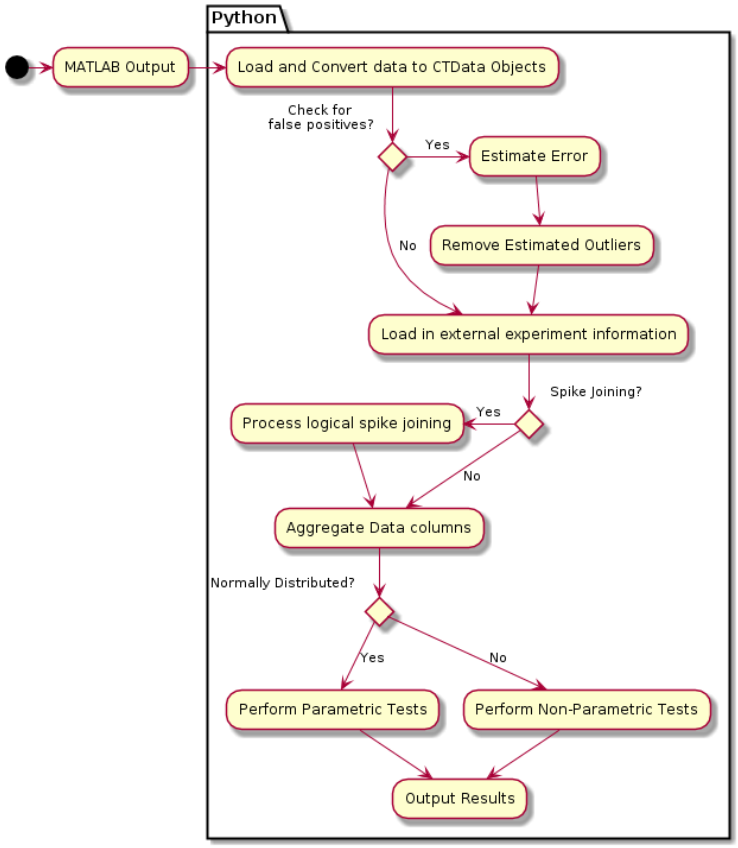

Figure 3.2: Data Pipeline and Information Flow

better results than the previous *chessboard* transform which had been successful on more uniform data in previous studies [4].

Quasi-Euclidean algorithm

This algorithm measures the total euclidean distance along a set of horizontal, vertical and diagonal line segments [7].

$$|x_1 - x_2| + (\sqrt{2} - 1), |x_1 - x_2| > |y_1 - y_2| \, (\sqrt{2} - 1) \, |x_1 - x_2| \, , \text{otherwise} \tag{3.1}$$

In order to apply this to a 3D space Kleinberg's method is used [8]. This allows for nearest neighbour pixels to be sorted by k-dimensional trees and enabling fast distance transforms via Rosenfeld and Pfaltz's *quasi-euclidean* method stated in equation:3.1.

Effect of Enhanced Watershed algorithm

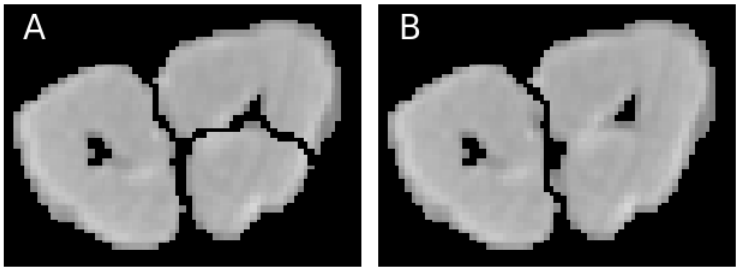

Figure 3.3: *A* showing the chessboard method, *B* improved quasi-euclidean method

Chapter 4

Results

Chapter 5

Discussion

Chapter 6

Critical Evaluation

Organisational Methods

Relevance to Degree

Time Management

Collaborative Work

Other Issues

Chapter 7

Appendix

Software Packages Used

Libraries

Table 7.1: Software libraries used

MATLAB Image Processing Toolbox	Numpy	Matplotlib
Seaborn	Scipy	Sklearn
Statsmodels	Pymc3	Xlrd
PyQt5		

Tools

Table 7.2: Software tools used

MATLAB	Python Debugger (PDB)	IPython
Emacs	git	org-mode
Tomviz	ImageJ	

Glossary

Table 7.3: Dictionary for Terms and acronyms

Term	**Definition**
µ-CT	Micro Computed Tomography
Genotype	A genetically distinct individual or group
Phenotype	A physical/measurable trait
Alleles	A variant of a gene
Genus	
Genome	
Morphometric	
GUI	Graphical User Interface

Code Segments and Examples

MATLAB Watershedding

```
function [W] = watershedSplit3D(A)
  % Takes image stack A and splits it into stack W
  % Convert to BW
  bw = logical(A);
  % Create variable for opening and closing
  se = strel('disk', 5);
  % Minimise object missshapen-ness
  bw = imerode(bw, se);
  bw = imdilate(bw, se);
  % Fill in any left over holes
  bw = imfill(bw,4,'holes');
  % Use chessboard for distance calculation for more refined splitting
  chessboard = -bwdist(~bw, 'quasi-euclidean');
  % Modify the intensity of our bwdist to produce chessboard2
  mask = imextendedmin(chessboard, 2);
  chessboard2 = imimposemin(chessboard, mask);
  % Calculate watershed based on the modified chessboard
  Ld2 = watershed(chessboard2);
  % Take original image and add on the lines calculated for splitting
  W = A;
  W(Ld2 == 0) = 0;
end
```

Listing 1: MATLAB Watershedding function

References

[1] H. Özkan, A. Brandolini, R. Schäfer-Pregl, and F. Salamini, "AFLP Analysis of a Collection of Tetraploid Wheats Indicates the Origin of Emmer and Hard Wheat Domestication in Southeast Turkey," *Molecular biology and evolution*, vol. 19, no. 10, pp. 1797–1801, oct 2002. [Online]. Available: http://academic.oup.com/mbe/article/19/10/1797/1259152

This paper discusses the origin of wheat domestication and provides evidence for it's origin in society

[2] B. Shiferaw, M. Smale, H.-J. Braun, E. Duveiller, M. Reynolds, and G. Muricho, "Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security," *Food Security*, vol. 5, no. 3, pp. 291–317, jun 2013. [Online]. Available: http://link.springer.com/10.1007/s12571-013-0263-y

[3] G. Charmet, "Wheat domestication: Lessons for the future," *Comptes Rendus - Biologies*, vol. 334, no. 3, pp. 212–220, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.crvi.2010.12.013

[4] N. Hughes, K. Askew, C. P. Scotson, K. Williams, C. Sauze, F. Corke, J. H. Doonan, and C. Nibau, "Non-destructive, high-content analysis of wheat grain traits using X-ray micro computed tomography," *Plant Methods*, vol. 13, 2017.

[5] F. J. Leigh, I. Mackay, H. R. Oliveira, N. E. Gosman, R. A. Horsnell, H. Jones, J. White, W. Powell, and T. A. Brown, "Using diversity of the chloroplast genome to examine evolutionary history of wheat species," *Genetic Resources and Crop Evolution*, vol. 60, no. 6, pp. 1831–1842, 2013.

[6] J. Cockram, H. Jones, F. J. Leigh, D. O'Sullivan, W. Powell, D. A. Laurie, and A. J. Greenland, "Control of flowering time in temperate cereals: Genes, domestication, and sustainable productivity," *Journal of Experimental Botany*, vol. 58, no. 6, pp. 1231–1244, 2007.

[7] J. L. Pfaltz, "Sequential Operations in Digital Picture Processing," *Journal of the ACM*, vol. 13, no. 4, pp. 471–494, oct 1966. [Online]. Available: http://portal.acm.org/citation.cfm?doid=321356.321357

[8] J. M. Kleinberg, "Two algorithms for nearest-neighbor search in high dimensions," in *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing - STOC '97*. New York, New York, USA: ACM Press, 1997, pp. 599–608. [Online]. Available: http://dl.acm.org/citation.cfm?id=258533.258653