NTNU – Trondheim
Norwegian University of
Science and Technology

Genome analysis of four novel *Psychrobacter* spp. and characterisation of their six putative laccase-like multicopper oxidases using bioinformatics tools

Kjersti Rise

Biotechnology (5 year)	
Submission date:	August 2015
Supervisor:	Martin Frank Hohmann-Marriott, IBT
Co-supervisor:	Rahmi Lale, IBT

Norwegian University of Science and Technology
Department of Biotechnology

Get cape. Wear cape. Fly.

Summary

In May 2009, several *Psychrobacter* spp. were found at the bottom of the Norwegian sea outside of Svalbard. Four of these, P11F6, P2G3, P11G3 and P11G5, were selected and sequenced for further work, and form the basis for this thesis. This work began with automatic annotation using RAST. The four genomes were found to have between 3.2 - 3.4 million base pairs, a GC-content of 41.9 - 42.9 % and contained between 2674 and 2914 genes. RAST places genes in subsystems if it finds a gene that fits one of the 27 subsystems. With few exceptions, the results from RAST showed an equal distribution of genes when comparing the subsystem distributions of the four genomes. Mauve was used to analyse how evolution had changed the genomes, and how whole blocks of genes had changed positions compared to the other genomes. Visual observation by carefully zooming in on specific parts of the genome verified that large parts of sequence were fully conserved in all four genomes, as well as demonstrating that large stretches of sequence were close to fully conserved, with the only difference being a shift in position relative to the other genomes.

Further investigations were performed to figure out if the *Psychrobacter* spp. contained laccases. Six laccase-like multicopper oxidases (LMCO's) were found; two in each of P11G3 and P11G5, one in P2G3 and one in a plasmid of P11F6. Analyses showed that these protein sequences consisted of 565 - 568 aa's. The compositions of atoms and amino acids were determined using ExPASY's ProtParam. This showed great similarities, as well as finding the molecular weight (63.7 - 64.1 KDa) and theoretical isoelectric point (6.75 - 8.59). Half-life was determined to be "above 10 hours" and all proteins were found to be stable.

One of the most important features of the laccases are the copper binding residues, and the LMCO's were searched in hope of finding them. Using Phyre2, type 1 was found in complete, while type 2 and type 3 were only partially found. Manual searches were performed to find the remaining residues, and hence finding the complete Cu-binding sites. These sites were found in the so-called signature sequences; conserved sequences which were expected to be found in members of the multicopper oxidase family.

Further studies were done on visualizing the LMCO's in PyMOL, both separate and superpositioned, to see differences and similarities. The 3D models showed that the LMCO's that were expected to be similar based on the other analyses, turned out to have more different structures. PyMOL was also used to visualize the substrate pockets and compare them with regards to shape and size. Clustal was used to compare the sequences in alignments, and both signal sequences and the full protein sequences were aligned and compared. The phylogenetic trees made by Clustal showed the relationship between the LMCO's. The signal sequences were investigated with PSORT-B to determine their subcellular localization, which showed that all LMCO's were destined for the periplasm.

Finally, the *Psychrobacter* sp. P11F6 was grown on media containing 2-methoxy-phenols, in an attempt to alter the gene expression into transcribing LMCO's, with 2-methoxy-phenols being one of the many substrates of laccases and LMCO's. As it turned out, the LMCO's were not even on the list of upregulated genes. The promoter sequences for the top ten transcript list were still identified. To see if any of these ten upregulated genes were translated, the proteome was investigated. This showed only eight of the ten, although these were upregulated when compared to P11F6 grown on media which did not contain the substrate.

Sammendrag

I mai 2009 ble flere forskjellige stammer av *Psychrobacter* funnet på havbunnen ved Svalbard. Fire av disse stammene, P11F6, P2G3, P11G3 og P11G5, ble valgt ut og sekvensert for videre forskning og danner også grunnlaget for denne masteroppgaven. Arbeidet i masteroppgaven startet med at genomene ble annotert automatisk ved hjelp av RAST. De fire genomene besto av 3,2 - 3,4 millioner basepar, hadde et GC-innhold på 41,9 - 42,9 % og inneholdt mellom 2674 - 2914 gener. De genene RAST kunne plassere i et subsystem ble gruppert sammen med andre gener med tilsvarende kvaliteter i et av de 27 subsystemene, og med få unntak ble det funnet at de fire genomene hadde omtrent like mange gener i hvert subsystem. Mauve ble brukt for å visualisere hvordan evolusjon hadde endret på genomene og hvordan hele blokker av genomet hadde endret posisjon i forhold til de andre genomene. Ved å zoome inn på forskjellige deler av genomene viste Mauve både hvordan deler av sekvensene var konservert i alle de fire genomene, og hvordan større sekvenser var like mellom to genomer, dog forskjøvet i posisjon.

Det ble videre undersøkt om de fire utvalgte *Psychrobactene* inneholdt laccaser. Seks laccase-lignende multikobber oxidaser (LMCO) ble funnet; to hver i P11G3 og P11G5, en i P2G3 og en i et av plasmidene til P11F6. Det ble funnet at disse seks proteinsekvensene varierte mellom 565 og 568 aminosyrer. Videre ble ExPASY's ProtParam brukt for å bestemme sammensetning av både atomer og aminosyrer, som viste store likheter, samt molekylær vekt (63,7 - 64,1 KDa) og teoretisk isoelektrisk punkt (6,75 - 8,59). Halveringstiden ble funnet å være «mer enn 10 timer», og alle proteinene ble vurdert som stabile.

En av de viktigste egenskapene til laccaser er de kobber-bindende aminosyrene, og LMCO'ene ble også undersøkt for å finne disse. Phyre2 fant type 1 fullstendig, mens type 2 og 3 bare ble funnet delvis. Ved hjelp av manuelle søk ble også de resterende delene av type 2 og 3 funnet. Det ble også oppdaget at disse setene var plassert i de såkalte signatursekvensene; konserverte sekvenser som var forventet å finne hos medlemmer av multikobber oxidase-familien.

Videre ble LMCO'ene visualisert i PyMOL, både alene og sammen, for å se på forskjeller og likheter. Modellene viste at de strukturene som var forventet å være like, basert på tidligere undersøkelser, ikke var så like som antatt. Samtidig var det andre og mer ulike sekvenser som viste seg å ha mer lignende strukturer. PyMOL ble også brukt for å visualisere substratlommene og sammenligne dem med tanke på form og størrelse. ClustalW og Clustal Omega ble brukt for å sammenligne sekvensene i alignments, og både signalsekvensene og de fullstendige proteinsekvensene ble sammenlignet. Fylogenetiske trær viste hvordan LMCO'ene var beslektet. Ved å undersøke signalsekvensene med PSORT-B viste det seg at alle LMCO'ene var antatt å havne i periplasma.

I siste del av oppgaven ble *Psychrobacter* P11F6 dyrket på medium med 2-methoxy-fenol, for å se om dette kunne endre genuttrykket slik at *LMCO* ble uttrykt. 2-methoxy-fenol er et av de mange substratene som laccaser og dermed også LMCO bruker. Det viste seg derimot at LMCO'ene ikke var å finne på lista over oppregulerte gener. Promotersekvensene til de ti mest oppregulerte genene ble likevel funnet og kartlagt. For å se om disse ti genene ble translatert, ble de sjekket opp mot proteomet, som viste at åtte av de ti ble funnet og oppregulert i forhold til *Psychrobacter* P11F6 som vokste uten substratet.

Acknowledgements

This thesis is the final result of my six years at NTNU. It has been an adventure, and as all adventures, it had good times and bad times, beautiful sunrises, good music, classes that made me want to pull my hair out, classes that made me feel so inspired I could hardly stay in my seat, and perhaps the most important part; there was a whole bunch of amazing people. And loads of iced coffee!

I want to thank my supervisor Martin for letting me be a part of his research group, and for showing me that science is just as magical as I've always thought it would be. My co-supervisor Rahmi has showed me how to believe in myself and learn from every situation, as well as remembering to always have fun. Morteza included me in his thesis work, and showed me how to find the answers, and sometimes even the questions. To all of them, and everyone else in the PhotoSynLab and others who were included in this work in one way or another, thank you.

I want to say thanks to all friends and family for their support, and particularly to Einar Johan for all his technical support, proofreading of this thesis and most importantly, all the hugs. Another special thanks to Torje, for all the technical support and all the laughs. And finally, a great thanks to Pengvin, for simply being awesome.

Trondheim, August 2015
Kjersti Rise

Table of Contents

Summary i

Sammendrag iii

Acknowledgements v

Table of Contents viii

List of Tables ix

List of Figures xi

Abbreviations xiii

1 Introduction 1

1.1 *Psychrobacter* 2

1.2 Laccases 2

1.3 Promoters 5

1.4 Pre-work 6

1.5 Aim of study 6

2 Methods and programs 9

2.1 RAST annotation 9

2.2 Mauve 12

2.3 Laccase-like multicopper oxidases 13

2.3.1 ProtParam 13

2.3.2 Phyre2 14

2.3.3 PSORT-B 15

2.3.4 PyMOL 15

2.3.5 Clustal 17

2.4 Transcriptome, proteome and promoters 17

2.4.1 Transcriptome and proteome 17

2.4.2 Promoters 18

3 Results 19
3.1 RAST 19
3.1.1 Graphical distribution of genes 20
3.1.2 Comparison of the genomes 22
3.1.3 Comparison of the subsystems feature counts 22
3.2 Mauve 24
3.3 Laccase-like Multicopper Oxidases 28
3.3.1 Amino acids and half-life 28
3.3.2 Copper sites 31
3.3.3 Subcellular location 34
3.3.4 PyMOL 34
3.3.5 Alignments 42
3.4 Transcriptome, proteome and promoters 48
3.4.1 Transcriptomics and proteomics 48
3.4.2 Promoters 49

4 Discussion 53
4.1 RAST 53
4.2 Mauve 54
4.3 Laccase-like Multicopper Oxidases 55
4.3.1 Amino acids and half-life 55
4.3.2 Copper sites 57
4.3.3 Subcellular location 58
4.3.4 PyMOL 58
4.3.5 Alignments 62
4.4 Transcriptome, proteome and promoters 63
4.4.1 Transcriptomics and proteomics 63
4.4.2 Promoters 64

5 Further research 65
5.1 RAST 65
5.2 Mauve 65
5.3 Laccase-like Multicopper Oxidases 65
5.4 Transcriptome, proteome and promoters 66

6 Conclusion 67

Bibliography 69

Appendices 73
Appendix A: Substrate pocket residues 73

List of Tables

3.1 The first comparison of the four strains of *Psychrobacter* spp. 22
3.2 Comparison of genes found in each subsystem 23
3.3 Positions of the six *LMCO* sequences 28
3.4 Amino acid compositions in the six LMCO's 29
3.5 Compositions of atoms, theoretical pI and molecular weight in the LMCO's 30
3.6 Predicted half-life and stability for each of the six LMCO's 31
3.7 Cu binding sites found in each of the six LMCO's 32
3.8 Revised table of Cu sites 33
3.9 Subcellular location 34
3.10 Selected residues of the substrate pockets 40
3.11 Top ten list of upregulated genes in P11F6 48
3.12 The list of upregulated genes compared to proteome 49
3.13 Promoters 50
3.14 Ribosome binding sites 51

List of Figures

1.1 Example of copper site 4
1.2 Example of promoter 5

2.1 Settings for RAST step two 11
2.2 Settings for RAST step three 12

3.1 Graphical distribution of genes in all four genomes 21
3.2 Graphic view of the different gene blocks in the four genomes 24
3.3 Conserved regions of the *Psychrobacter* genomes 25
3.4 Closer look on the hidden blue box 26
3.5 Sliding parts of different genomes 26
3.6 Phylogenetic tree 27
3.7 T1 and trinuclear Cu sites 33
3.8 3D views of LMCO's from P11G3 35
3.9 3D views of LMCO1 and LMCO2 35
3.10 3D views of LMCO1 and LMCO3 36
3.11 3D views of LMCO's from P11G5 36
3.12 Substrate pocket analysis 38
3.13 Multiple alignment of the signal sequences 43
3.14 Phylogenetic tree of the signal sequences 43
3.15 Full alignment of LMCO's 46
3.16 Phylogenetic tree of the LMCO full sequence 47

Abbreviations

3D	=	Three-dimensional
aa	=	Amino acid(s)
API	=	Application Programming Interface
bp	=	Base pair(s)
CDD	=	Conserved Domain Database
Cu	=	Copper
Da	=	Dalton
DNA	=	Deoxyribonucleic acid
EMBL-EBI	=	The European Molecular Biology Laboratory - European Bioinformatics Institute
ExPASy	=	Expert Protein Analysis System
HMM	=	Hidden Markov Model
pI	=	Isoelectric point
KDa	=	Kilodalton
laccase	=	Refers to the gene coding for the laccase enzyme
laccase	=	Refers to the actual laccase enzyme
LMCO	=	Laccase-like multicopper oxidase
LCBs	=	Locally Collinear Blocks
LDF	=	Linear Discriminant Function
Multi-MUMS	=	Multiple Maximal Unique Matches
Mw	=	Molecular weight
ORF	=	Open Reading Frame
Phyre2	=	Protein Homology/analogY Recognition Engine v. 2.0
RAST	=	Rapid Annotation using Subsystem Technology
RBS	=	Ribosome Binding Site
RNA	=	Ribonucleic Acid
RPKM	=	Reads Per Kilobase per Million
sp.	=	Specie
spp.	=	Species
ss	=	Signal sequence
T1	=	Type 1 copper
T2	=	Type 2 copper
T3	=	Type 3 copper

Chapter 1

Introduction

The work in this thesis is based on the discovery of four *Psychrobacter* spp. at the bottom of the ocean outside Svalbard, which were named P11F6, P2G3, P11G3 and P11G5. On the way towards new discoveries in their genomes, we need to know what we are looking for. What are these *Psychrobacters*, what can we expect to learn about them? What is the function of a laccase, and can we expect to find them? What is a promoter, and why is this relevant for this thesis? All of these questions, and much more, will be part of this study.

The structure of this thesis is as follows: Each chapter is divided into three parts; genome studies, laccases/laccase-like multicopper oxidases (LMCO's) and finally transcriptomics/proteomics/promoters. Each of these have undercategories as needed, but the pattern remains the same. This means that in this introduction, the first part will start with the *Psychrobacter* itself. The ways of learning more about their genomes is explained in **section 2.1** and **section 2.2**, and the results are shown in **section 3.1** and **section 3.2**.

The next part of the introduction includes the laccases/LMCO's, the enzymes that are examined in various ways regarding chemical and physical properties, compositions, 3D models, substrate pocket analysis and alignments, along with findings of signal sequences and signature sequences in **section 2.3**, with results of these analyses shown in **section 3.3**.

In order to see how adding a specific substrate (2-methoxy-phenol) to the growth media would affect expression of laccases, transcriptomics and proteomics data were analyzed in part three. A general introduction to promoters is given in **section 1.3**. As transcriptomics and proteomics were performed by others, only parts of the results from these analyses are included, and focus here is more on the finding of promoters, which are shown in **section 2.4**. The top ten list of transcriptome, a comparison to proteomic data, and the found promoters are shown in **section 3.4**.

To give some more background information before we dive into what was actually done in this thesis, some of the work that was already done will be presented. The discovery of the

Psychrobacter spp., screening and sequencing and will be briefly covered in **section 1.4**, before the aim of the study is defined in **section 1.5**

1.1 *Psychrobacter*

The bacteria in question in this thesis are *Psychrobacter* spp., which were found at the bottom of the ocean outside of Svalbard in 2009. The name *Psychrobacter* was first used by Juni and Heym in 1986, when they suggested it as a name for the new strain of bacteria they had found (Juni and Heym, 1986). The name was rather descriptive, with *"Psychro"* from greek, meaning cold, and *"bacter"* meaning rod, it was the name of a rod growing at cold temperatures. *Psychrobacter* belong to the class of gammaproteobacteria and the family of Moraxellaceae. The strains of *Psychrobacter* were found to be aerobic, gram-negative and non-motile, halotolerant, being both nonpigmented and nonsporulating and being positive for both catalase and oxidase. The penicillin-suspectible coccobacilli were found to be 0.4 - 1.5 μm in diameter and 0.4 - 3.8 μm long (Juni and Heym, 1986; Bozal et al., 2003). Genome studies of various *Psychrobacter* spp. have shown a GC content ranging from 41 to 47 % (Bozal et al., 2003). Genome sequencing of *P. arcticus* 273-4 and *P. aquaticus* Strain CMS 56^T showed lengths ranging from 2.65 - 3.2 million bp (Ayala-del Río et al., 2010; Reddy et al., 2013)

1.2 Laccases

In general

Laccases (EC 1.10.3.2), or benzenediol oxygen oxidoreductases, are polyphenol oxidases. Laccases were first named "laccase" by Bertrand in 1894 when he studied the enzyme from the latex of the lac tree *Rhus succedanes* (Bertrand, 1984). Using molecular oxygen as electron acceptor, the enzymes catalyze oxidation of phenolic compounds (Sharma et al., 2007). These enzymes are the most numerous members of the multi-copper protein family, and can participate in cross-linking of monomers, degradation of polymers and ring cleavage of aromatic compounds (Kawai et al., 1988). Having a broad range of substrates, the laccases show many features that are interesting in a biotechnological point of view, such as decolorization of dyes (Baldrian et al., 2006) and lignin biosynthesis (O'Malley et al., 1993), as well as having potential applications in food industry, pulp/paper industry, nanobiotechnology, soil bioremediation and cosmetics (Couto and Herrera, 2006). One way of screening for laccase-producing microbes is using 2-methoxy-phenol (guaiacol) as substrate (Kiiskinen et al., 2004).

Laccases are commonly found in nature, and the first one found in a prokaryot was in *Azospirillium lipoferum* (Givaudan et al., 1993). The laccases found in plants and fungi are mostly extracellular, helping them avoid the problem of reactive species, while the ones found in bacteria are mostly found to be intracellular (Diamantidis et al., 2000). Studies have shown that the molecular weight of laccases can vary from 32 - 130 KDa, depending on the type of organism it came from (Morozova et al., 2007; Ihssen et al., 2015). Studies on laccases from plants and fungi have shown pI values ranging from 2.6 - 9.5, with $T_{1/2}$

ranging from 0.2 - 192 hours when measured at temperatures ranging between 40 - 80 °C (Morozova et al., 2007). Several bacterial laccases and laccase-like proteins were compared and found to have a length of 348 - 1662 aa (Sharma et al., 2007).

There are several other, similar enzymes that uses copper for oxidizing substrates, all being members of the multicopper-oxidase family. Examples of these are ferroxidase (EC 1.16.3.1) or ascorbate oxidase (EC 1.10.3.3). Other enzymes can be found that resemble these enzymes, such as laccase-like multicopper oxidases (LMCO). The common feature for all of these, along with oxidation using copper (Cu), are the Cu binding sites.

Cu binding sites

Cu can be bound in several places, giving name to binding sites of type 1 (T1), type 2 (T2) and type 3 (T3), where T2 and T3 can create a nuclear cluster (Fee, 1975; Colman et al., 1978). The T1 Cu binding site is created as a trigonal coordination, having two His and a Cys residue conserved plus one variable position, as can be seen in **figure 1.1**. The conserved positions create the equatorial ligands, while the variable one, usually a Met in bacteria and Leu or Phe in fungal laccases, creates an axial ligand (Claus, 2004). The T1 Cu site is the place of substrate oxidation, due to the high redox potential in this area, while the trinuclear cluster of T2 and T3 reduces molecular oxygen and releases water. This site is usually created by eight His residues; two in T2 and six in T3. Kumar et al did a multiple alignment of more than 100 laccases, in hope of finding sequence regions connected to the trinuclear Cu binding sites (Kumar et al., 2003). This showed four regions, which exist in four patterns of HXH, with X being a variable aa. In one of these regions, X is the Cys found in the T1 site. These results were confirming the genome comparisons done by Solomon et al in 1996, which used multiple multicopper oxidases from different species, trying to identify binding sites among others (Solomon et al., 1996).

As shown in **figure 1.1**, it takes eight His residues to create the trinuclear cluster, while the T1 site is made out of His - Cys - His - Met. In this case, Cys492 is surrounded by His491 and His493, which are used in each of the two T3's. The His residues come in HXH format, such as His105 - X - His107 and His422 - X - His424. As the figure shows, these His residues belong to different Cu's.

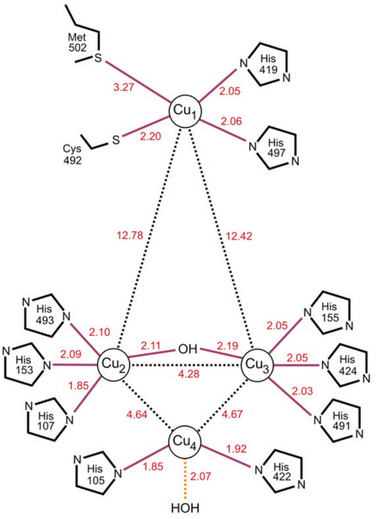


Figure 1.1: The figure shows how the copper sites are organized in *Bacillus subtilis*. Cu1 is the T1, made out of His - Cys - His - Met. Cu2 and Cu3 are of T3, with the belonging six His residues. Finally, Cu4 is T2, consisting of two His. Figure by Enguita et al. (2003)

It has been shown that these copper ligands are found in four specific patterns (Reiss et al., 2013):

HXH

HXHG

HXXHXH

HCHXXXHXXXXM/L/F

Signature sequences

Proteins that belong to the same family share sequences that distinguish them, known as signature sequences. It has been found that the laccases can contain several types of signature sequences, both sequences that are found in members of the multicopper oxidase family, and sequences that are specific to laccases (Ouzounis and Sander, 1991; Kumar et al., 2003).

The two following sequences have been reported as signature sequences in multicopper oxidases:

Type 1: G-X-[FYW]-X-[LIVMFYW]-X-[CST]-X_8-G-[LM]-X_3-[LIVMFW]

Type 2: H-C-H-X_3-H-X_3-[AG]-[LM]

Whereas the following four sequences have been observed as signature sequences in laccases in specific:

L1: H-W-H-G-X_9-D-G-X_5-Q-C-P-I

L2: G-T-X-W-Y-H-S-H-X_3-Q-Y-C-X-D-G-L-X-G-X-[FLIM]

L3: H-P-X-H-L-H-G-H

L4: G-[PA]-W-X-[LFV]-H-C-H-I-D-A-E-X-H-X_3-G-[LMF]-X_3-[FLM]

The X in the signature means any residue, while the [ ] indicates that any of the residues inside the brackets can be found in that position. As seen when comparing the sequences, L2 can conform into type 1, and L4 can conform into both type 1 and type 2. There are also similarities between L1 and L3, and L2 and L4. Although not statistically significant, there are some residues that are fully conserved within sequences of all laccases, such as the copper binding sites, which are found within these sequences (Kumar et al., 2003). When comparing these signature sequences to the patterns in the previous section on cu sites, it becomes clear that these are the same thing, and hence it is certain that the cu sites will be found within these sequences.

1.3 Promoters

Promoters are important in DNA transcription, as they are the ones that regulate binding of RNA polymerase. An example of a prokaryot promoter can be seen in **figure 1.2**. A prokaryot promoter contains a -10 region and a -35 region, which are binding sites/regulators for RNA polymerase. Transcription start site is in this case considered to be +1, meaning that -10 and -35 regions are 10 and 35 bp upstream of transcription start site, respectfully. -10 regions are usually around 8-10 bp long, while -35 regions are 5-6 bp long. In the section on promoters, Ribosome Binding Sites (RBS) will be searched for along with -10 and -35 regions. RBS' are expected to be found about six bases upstream of the start codon, although this is not included in the figure. These values are only approximations, and can vary (MendelUniversityBrno, 2015).

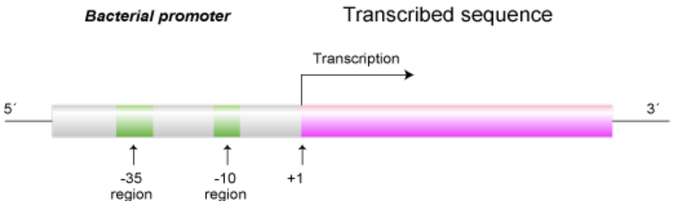


Figure 1.2: The -10 and -35 regions of a promoter shown in a simple figure. Figure by (MendelUniversityBrno, 2015)

1.4 Pre-work

As this thesis only involves applied bioinformatics, this section will explain in short what was done before the work in this thesis started. This work was done under supervision of associate professor Martin Hohmann-Marriott (NTNU), researcher Rahmi Lale (NTNU) and Alexander Wentzel (SINTEF), and is part of the Ph.D - thesis of Morteza Shojaei Moghadam. Parts of the work, such as the sequencing and the transcriptomics, was performed by Christian Rückert (Moghadam et al., 2015). The proteome data was given by Animesh Sharma.

It all started in May 2009: In collaboration with UiT (the Arctic University of Norway, Tromsø), the research cruise R/V Jan Mayen sampled biota, water and sediments in ten different locations in the region between and around the Svalbard archipelago and Bear Island in the Barents Sea. This collection of samples lead to establishment of a library containing 1448 single bacterial isolates, originating from biota (773), water (257) and sediments (418). Based on 16S rDNA sequences of 550 isolates, the library consists of at least 31 genera.

The *Psychrobacter* spp. used in this thesis were found at a depth of 20 m. Along with the rest of the bacteria from the sampling, they were cultivated and kept in 96-well plates, each well having a different strain. Bacteria from each well were grown on plates containing 2-methoxy-phenol, to screen the ones that showed typical laccase-activity of oxidizing these monophenols. Laccase-activity would be shown by brown colour zones around the colonies. The 13 colonies that showed this phenotype were further chosen and investigated, and 16S sequencing showed that the colonies were *Psychrobacter* spp.. This was done to determine which colonies to further work with. Four out of the thirteen were chosen based on a phylogenetic tree, choosing the ones that were most different from the others; P2G3, P11G3, P11G5 and P11F6. Using Illumnia MiSeq, the whole genomes were sequenced by Christian Rückert, trying to identify, among others, laccases. Six laccase-like multicopper oxidases were found and their signal sequences were determined by Morteza Shojaei Moghadam (Moghadam et al., 2015). And from here, the work presented in this thesis began.

1.5 Aim of study

The overall aim of this thesis is to get an overview of the genomes of the four *Psychrobacter* spp. and look at their differences and similarities, learn as much as possible about six putative laccase-like proteins, and finally comparing transcriptomics and proteomics of the top ten up-regulated genes when growing *Psychrobacter* P11F6 on media containing 2-methoxy-phenol and finding the promoters to these genes, all using nothing but bioinformatics tools of various kinds. The first part of the study involves the genomes. The specific aim here is to do an automatic annotation using RAST and figure out roughly what kind of genes these genomes contains, comparing sizes and contents. The second aim here is to see how the genomes have evolved and changed over time, which will be done by using the program Mauve.

The second part of the study involves laccase-like proteins. Even though laccases and other multicopper oxidases have been broadly studied from other organisms, such as fungi and plants, the ones with a bacterial origin have been left out a bit. The aim here is to see how these proteins are composed, how they differ from each other and try to discover more about their properties, along with trying to see if they actually are laccases. This includes finding compositions of aa/atoms, lengths, estimates of half-life and stability, finding substrate pocket residues and making 3D models, as well as doing multiple alignments to discover more on relationships and finding signature sequences.

The final part of the study involves transcriptomics, proteomics and promoters. The aim here is to see if growing the *Psychrobacter* P11F6 on media containing 2-methoxy-phenol will lead to an upregulation of *laccase*, seeing if it leads to an upregulation of laccase proteins and finally studying the promoters of the actually upregulated genes. Transcriptomics and proteomics were performed by Christian Rückert and Animesh Sharma, and for the promoters, these will be found by Softberry's BPROM and manual searching.

Chapter 2

Methods and programs

In this chapter, the genomes of four chosen species of *Psychrobacter* are analyzed in various ways, from whole genome analysis to specific genes, the laccase-like multicopper oxidases (LMCO's), and finally the chapter ends up with a part on promoters and transcriptomics and proteomics under certain conditions. This will be done by introducing the various types of programs used, and how they are used, including settings. These types of programs includes Rapid Annotation of Subsystem Technology (RAST) and Mauve for whole genomes, and Protein Homology/analogY Recognition Engine V 2.0 (Phyre2), PSORT-B, ProtParam, Clustal and PyMOL for the LMCO's. For promoters in the transcriptomics section, manual searching and Softberry's BPROM are used, while the transcriptomics and proteomics data was created by others using methods that will not be covered here.

2.1 RAST annotation

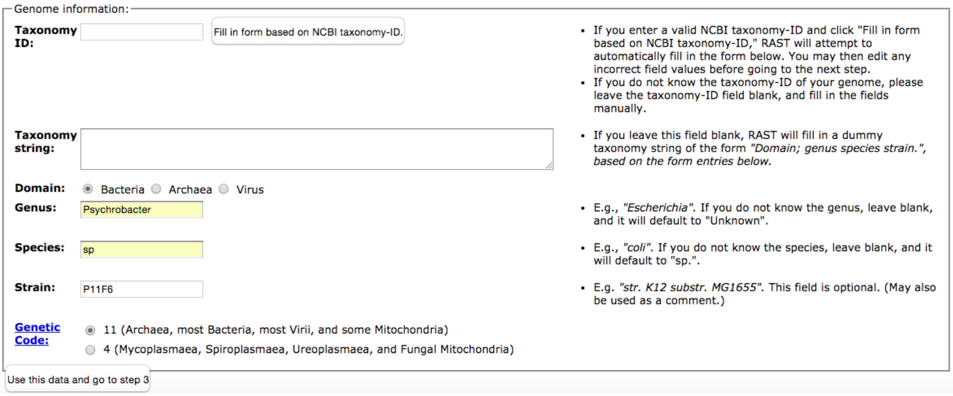
Manual annotation of a shorter DNA sequence is always possible, although time consuming. When it comes to full genomes however, they are not possible to annotate by hand, unless you have way too much spare time. As new technologies for sequencing are discovered and developed by the week, the need for automatic annotation is growing. Speed, although somewhat important, is less important than accuracy, completeness and consistency, and this is the focus of all the developers of automatic annotation software. One of the ways this has been solved is by using a growing library of subsystems, such as the ones used by RAST (Aziz et al., 2008). RAST produces two classes of asserted gene functions: subsystem-based assertions and nonsubsystem-based assertions. Subsystem-based assertions are based on recognition of the functional variants of subsystems, while the nonsubsystem-based assertions are filled using a number of other, more common tools. A subsystem is defined as a set of abstract functional roles, and all subsystems in RAST are manually curated. This means that proteins that do something similar, can be part of the same subsystem. RAST also uses another collection of protein families; FIGfams, which are derived from the subsystem technology. Each FIGfam consists of a set of proteins, which are isofunctional homologs, meaning that they're thought to have the same function and come from a common ancestor. FIGfams are not manually curated, but can be based upon both subsystems and non-subsystems (Parrello, 2015).

The pipeline in RAST in short is as follows (Aziz et al., 2008):

- Calling tRNA and rRNA genes
- Making an initial effort to call protein-encoding genes
- Establishing phylogenetic context
- Searching in the FIGfams of a "neighbour" genome
- Recalling protein-encoding genes
- Processing the remaining genes against the whole FIGfam collection
- Clean-up gene calls
- Processing the remaining proteins and finally constructing an initial metabolic model

Talking about RAST without including the SEED makes no sense, as these two are highly intertwined. The SEED (Overbeek et al., 2014) is a database for bioinformatics research, which integrates a genome database, Application Programming Interface (API), a web front end and server scripts, as well as housing the subsystems and FIGfams used in RAST. The database is constantly updated, to ensure the results are as good as possible. Basically, the SEED holds all the information, or links to where to find the information, used in RAST.

The RAST user interface is highly intuitive. Both FASTA format and genbank format files can be uploaded to RAST, in this thesis the genome sequences came in FASTA format. Uploading to RAST happens in three steps, where only step two and three have a number of settings. Step one has no settings, as it only involves choosing which file to upload. In step two, information about the genome is given, while settings for the annotation are set in step three. The settings chosen for step two are listed in **figure 2.1**, here using P11F6 as an example. Along with this info, it is possible to enter both taxonomy ID and taxonomy string if this kind of information is available. Entering a valid NCBI taxonomy ID leads to RAST attempting to fill in the form, which otherwise has to be filled in manually. These were left blank in uploading the genomes in this thesis, as a valid NCBI taxonomy ID was not available and is therefore not included here.



Genome information:

Taxonomy ID: [ ] Fill in form based on NCBI taxonomy-ID.

- If you enter a valid NCBI taxonomy-ID and click "Fill in form based on NCBI taxonomy-ID," RAST will attempt to automatically fill in the form below. You may then edit any incorrect field values before going to the next step.
- If you do not know the taxonomy-ID of your genome, please leave the taxonomy-ID field blank, and fill in the fields manually.

Taxonomy string: [ ]

- If you leave this field blank, RAST will fill in a dummy taxonomy string of the form *"Domain; genus species strain.", based on the form entries below.*

Domain: ● Bacteria ○ Archaea ○ Virus

Genus: Psychrobacter

- E.g., *"Escherichia"*. If you do not know the genus, leave blank, and it will default to "Unknown".

Species: sp

- E.g., *"coli"*. If you do not know the species, leave blank, and it will default to "sp.".

Strain: P11F6

- E.g. *"str. K12 substr. MG1655"*. This field is optional. (May also be used as a comment.)

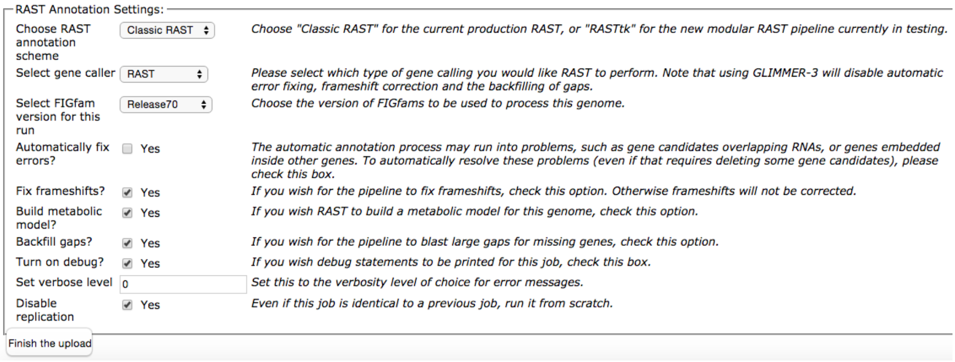
Genetic Code: ● 11 (Archaea, most Bacteria, most Virii, and some Mitochondria)
○ 4 (Mycoplasmaea, Spiroplasmaea, Ureoplasmaea, and Fungal Mitochondria)

Use this data and go to step 3

Figure 2.1: Options the user has when uploading P11F6 to RAST in step two. The settings used for uploading the four *Psychrobacter* spp. genomes are shown, as well as leaving the two fields on taxonomy blank. These settings were used for all genomes, not only P11F6.

As seen in **figure 2.1**, "taxonomy ID" and "taxonomy string" were not included, as there was no information on this area. Being bacteria, choosing "domain" was simple. Filling in *Psychrobacter* as "genus" was also a given. If this is left blank, it defaults to "unknown". Not knowing the name of the species, this was set to sp., which gave the same results as leaving it blank. The strain field is optional, and can be used as a comment field or some kind of ID, as it was here in the example for P11F6. "Genetic code" was set to "11", which corresponds to most bacteria.

There are several settings that can be altered in the final step of uploading the genome to RAST, which can be seen in **figure 2.2**. Choosing Classic RAST as "annotation scheme" means using the current production of RAST, which was chosen above RASTk, as this was currently in testing. Choosing RAST as "gene caller" instead of GLIMMER-3 was done to prevent disabling of automatic error fixing, frameshift corrections and backfilling of gaps, which is the default setting when using GLIMMER-3. FIGfam "version 70" was used, as it was the newest release at the time of uploading. "Automatically fixing errors" were turned off, as this could lead to deletions of gene candidates. The "fix frameshift" box was checked in order to have any problems with frameshifts fixed. The "build metabolic model" was also checked, as it at this point was not known if that would be useful for further work with the finished annotations. The "backfill gaps" box was also checked, as this would make the pipeline blast large gaps in the genome, and perhaps find some missing genes. Checking the debugger would create a list of debug statements, if any debugging was done along the way, which also would come in handy, leading to this being chosen. Verbosity level was left at the default value of 0, and by disabling replication, every job was run from scratch, even if it was identical to any other uploaded job.



RAST Annotation Settings:

Choose RAST annotation scheme	Classic RAST	*Choose "Classic RAST" for the current production RAST, or "RASTtk" for the new modular RAST pipeline currently in testing.*
Select gene caller	RAST	*Please select which type of gene calling you would like RAST to perform. Note that using GLIMMER-3 will disable automatic error fixing, frameshift correction and the backfilling of gaps.*
Select FIGfam version for this run	Release70	*Choose the version of FIGfams to be used to process this genome.*
Automatically fix errors?	□ Yes	*The automatic annotation process may run into problems, such as gene candidates overlapping RNAs, or genes embedded inside other genes. To automatically resolve these problems (even if that requires deleting some gene candidates), please check this box.*
Fix frameshifts?	☑ Yes	*If you wish for the pipeline to fix frameshifts, check this option. Otherwise frameshifts will not be corrected.*
Build metabolic model?	☑ Yes	*If you wish RAST to build a metabolic model for this genome, check this option.*
Backfill gaps?	☑ Yes	*If you wish for the pipeline to blast large gaps for missing genes, check this option.*
Turn on debug?	☑ Yes	*If you wish debug statements to be printed for this job, check this box.*
Set verbose level	0	*Set this to the verbosity level of choice for error messages.*
Disable replication	☑ Yes	*Even if this job is identical to a previous job, run it from scratch.*

Figure 2.2: Options used when uploading P11F6 to RAST in step three, along with the reasons RAST gives for why the user should choose the different settings.

When RAST is finished processing, the gene browser can be used to take a look at the annotation. It is possible to download the annotation in various formats, including comparing it to other annotations. The results from the annotations made here can be seen in **section 3.1**.

2.2 Mauve

As time goes by and genomes evolve, processes such as genome rearrangement, horizontal transfer, deletions and insertions all contribute to genomes becoming mosaics of specific gene segments. Mauve is a program that lets the user upload two or more genomes and compare them on an evolutionary level (Darling et al., 2004), by combining analysis of large-scale evolutionary events with the more traditional multiple sequence alignment. Where a multiple alignment will compare base by base or aa by aa, Mauve aligns blocks of genes, hereby identifying conserved regions, inversions, rearrangements and breakpoints across many genomes at once. Mauve is based on identifying and aligning locally collinear blocks (LCBs), which are homologous regions of sequence shared by at least two of the uploaded genomes. These blocks do not contain any rearrangement, and should therefore be identical in all the genomes that contain that specific block. Each block is weighted, providing a measure of confidence, and the user can choose minimum weight in order to ensure results that are more or less likely, depending on individual needs of specificity and sensitivity. The exactly matching boxes that are found in two or more of the genomes but occur only once in each genome, and is bounded by mismatched nucleotides on either side, is the secret to how Mauve works. The fact that they should occur only once is a part of the secret, as one of the major challenges is to figure out which of the regions to combine if there are many similar. These boxes are called Multiple Maximal Unique Matches (Multi-MUMS), and are used as anchors in determining which regions are actually homologous blocks. They all have a certain minimum length, and are exactly matching sequences, which reduces anchoring sensitivity.

The alignment algorithm can be summed up as follows (Darling et al., 2004):

- Find local alignments (multi-MUMs)
- Use them to create a phylogenetic tree
- Select a set of multi-MUMs to use as anchors in LCBs
- Use the anchors to identify alignment and finally perform alignments of each LCB by using the guide tree

The algorithm is made to identify both the matching regions and the regions which are specific for each of the genomes. Using the first genome as a reference, the boxes in the other genomes are oriented based on this.

Using Mauve only requires having genomes in FASTA format. The wanted number of genomes are uploaded at the same time, and compared. Settings in this case were left at default.

2.3 Laccase-like multicopper oxidases

Changing focus from genomes as a whole to only a specific protein, laccases were searched for. The first identification of possible genes was done by Morteza Shojaei Moghadam, where known sequences were used as queries to perform BLAST searches against the whole genomes. A total of six laccase-like multicopper oxidase (LMCO) genes were found; two in P11G5, two in P11G3, one in P2G3 and the final one in a plasmid of P11F6. The signal sequences were determined, along with properties such as length, amino acid composition and molecular weight. ProtParam was used for physiochemical properties, Phyre2's Investigator for Cu binding sites, PSORT-B for the subcellular location, PyMOL for visualizations, and ClustalW and Clustal Omega for alignments were all used in order to learn as much as possible about the proteins. All of these results are shown in **section 3.3**.

2.3.1 ProtParam

Starting with the basics, Expert Protein Analysis System (ExPASy)'s ProtParam and pI/Mw computing tool (Gasteiger et al., 2005) were used to compute various physical and chemical properties of the LMCO's. This included molecular weight (Mw), composition of amino acids, estimated half-life, isoelectric point (pI) and instability index. Mw, pI and composition is found by simply counting the contents, and then showing final counts or multiplying with e.g., weight. pI is calculated using pK values of amino acids at pH between 4.5 and 7.3, and temperature at 15 °C and 25 °C.

Estimation of half-life is a prediction of the time it takes for half the content of the protein to disappear after synthesis in the cell. Half-life estimation is based on the "N-terminal rule", relating the half-life to the residue in the N-terminal of the protein. Depending on species and residue compositions, the half-life can vary from minutes to hours. ProtParam estimates half-life for human, yeast and *E.coli*, and from these results it's possible to extrapolate the results to find predictions for similar organisms. This prediction is based only on the content in the N-terminal, and does not include variations in the environment.

Instability index gives an estimate of how stable the protein would be in a test tube. This is based on some specific dipeptides, who's presence makes a protein more stable. Weighted values of 400 dipeptides are used to compute instability index, where an index below 40 means the protein is considered stable.

Neither ProtParam nor the additional site for computing pI/Mw require any settings, only pasting the protein sequence into the assigned box and pressing "go". The results from these analyses are shown in **section 3.3.1**

2.3.2 Phyre2

Phyre2 is a program that can be used for prediction and analysis of protein structure, as well as showing function and mutations (Kelley et al., 2015). When uploading a protein sequence to Phyre2, the pipeline involves:

- Detecting sequence homology
- Predicting secondary structure and disorder
- Constructing a hidden Markov model (HMM) and scanning it against a library
- Constructing 3D models based on the HMM
- Modelling insertions/deletions
- Modelling aa side chains
- Submission of top model for binding site prediction and transmembrane helix and topology prediction

Phyre2 also uses other programs in order to give more information on the uploaded protein, such as fpocket for pocket detection (Le Guilloux et al., 2009).

When uploading using "intensive" modelling, the models are further investigated after construction. This is to ensure templates with maximum sequence coverage and confidence. Once the model prediction is done, it's possible to choose the best model and compare it further in the Investigator. Phyre2 shows a list of the top 20 similar proteins, called templates, which are ordered by similarity. From here, one can use the Investigator to compare the uploaded protein to the chosen template. The list of proteins compared to the uploaded one has two important fields for determining which protein to choose; confidence and % ID. Confidence indicates the probability (0-100) of the uploaded sequence and the template being true homologous. This value does not represent the accuracy of the model, although related. Having a confidence level >90 % means that you can be very confident that the uploaded sequence is similar to and adopts the overall folds of the template.

ID shows the percentage of identity between template and uploaded sequence. You want this to be above 30-40 % for extremely high accuracy, although models with ID >15 % are still useful if the confidence is high. In this case, all templates chosen were the top ones, having an ID of 24-25 % and a confidence of 100.0.

Once the Investigator has finished the comparison to the template model, it shows three tabs of analyses; quality, function and Conserved Domain Database (CDD), all containing a number of analyses that can be performed. In this thesis, Phyre2 was used to find T1 Cu binding site, trinuclear site, and the substrate pocket in the protein. Once the Investigator was finished investigating, “pocket detector” in the function tab was used for finding the pocket, while the CDD tab was used to find the T1 Cu binding site and the trinuclear Cu binding site. All of these results are shown in **section 3.3.2**.

2.3.3 PSORT-B

To further discover more about the signal sequences, they were analyzed for their final location. Prediction of subcellular location can gain insight in a whole bunch of things, such as function and detection of drug targets, and computational predictions provides a quick and inexpensive way of getting this information. One of the programs that can perform this prediction is PSORT, which was first introduced in 1991 (Nakai and Kanehisa, 1991), and the first version of the improvement PSORT-B came in 2003 (Gardy et al., 2003). PSORT and PSORT-B’s prediction is based upon the protein sequence of gram negative bacterial proteins, and compares it to protein sequences with known subcellular locations. By searching and comparing to known sequences, which compose structures such as signal peptides or transmembrane α helices involved in known subcellular location, the signal sequence can be put into one of five categories: extracellular, outer membrane, periplasmic, cytoplasmic membrane or cytoplasmic. The sequence in question is compared to the known sequences of all categories, and each get a score on how much it resembles. The highest score decides which category the protein belongs to.

The current version of PSORT-B is 3.0 (Yu et al., 2010), which was used to predict subcelluar location in the LMCO’s. This was done using organism type “bacteria” and gram stain “negative” in the upload. The results from this analysis are shown in **table 3.9**.

2.3.4 PyMOL

For visualizing the LMCO’s in 3D, PyMOL was used. PyMOL is a molecular visualization system that makes it possible to visualize both entire proteins and parts of proteins, as well as superpositioning them. PyMOL is open-source, and as the name suggests, written and extensible in the Python programming language (Schrödinger, 2010).

PyMOL was used to visualize the LMCO's and superpositioning them both as a whole and in regards to the pockets found in **section 2.3.2**. The output files from Phyre2 were used in PyMOL, simply uploading them when needed.

3D models

In choosing which of the LMCO's to superposition, this was first done on LMCO3 and LMCO4. These two were chosen rather arbitrarily, vaguely based on the analyses that had been performed and knowing that these two were supposed to be similar. As these two both came from P11G3, the structures were thought to be closely related. Following this, a superpositioning of LMCO1 and LMCO2 was made, to see if these two would show similar structures. Coming from P11F6 and P2G3 respectfully, there was no prediction of how similar these two would be. The results from this lead to a superpositioning of LMCO1 and LMCO3, to see if the predicted similarity was real. Then finally, a superpositioning was made for LMCO5 and LMCO6, just to see if they followed the patterns shown by the others.

When making the models and the superpositionings, all proteins were shown in "cartoon" mode, which included secondary structures such as α helices and β sheets. This was chosen to visualize the proteins in the best way possible. The models were also coloured by secondary structures, using Helix-Sheet-Loop for one of the models and Helix-Sheet-Loop for the other. In the superpositioning, these colours are mixed together if the sequences have a total match, which can be most easily seen in the β sheets, which in many places are both yellow and bright purple. To make the comparisons easier, the models were kept in the same position both for single and superpositioned models. All the 3D structures and superpositions are shown in **section 3.3.4**.

Pocket analysis

One of the many features in PyMOL is the possibility to show parts of proteins. Using the same models as before with the results from Phyre2, it was possible to manually mark each of the residues which were part of the pocket. With all the residues of the pockets marked, "sphere mode" was used to visualize the pocket residues. To further separate the pocket from the rest of the protein, all the spheres were coloured pink, to give a contrast to the green, red and yellow patches of the protein models. By hiding the rest of the protein, visualizing only the pockets made it possible to compare them. The pockets are shown in **section 3.3.4**.

2.3.5 Clustal

Multiple alignments are useful for visualizing similarity between parts of or complete sequences. The European Bioinformatics Institute (EMBL-EBI) provides many alignment tools, both for pairwise and multiple alignments in multiple versions (Goujon et al., 2010; McWilliam et al., 2013). Two of the most common ones for multiple alignment are Clustal Omega (Sievers et al., 2011) and ClustalW (Larkin et al., 2007). The classical versions like ClustalW and ClustalX are slowly being phased out and taken over by Clustal Omega, the newest member of the Clustal family. New and improved algorithms keep the programs up to date and makes alignments faster and more accurate. Some of the new improvements in Clustal Omega is the use of HMM-profiles and seeded guide trees, which means the program can align almost any number of protein sequences both quickly and accurately.

In this study, Clustal Omega was chosen for aligning the whole LMCO's sequences, and ClustalW for aligning the signal sequences. Even though the alignment for the whole sequence also includes the signal sequences, the overall alignment of the whole sequence did not give the signal sequences the attention they deserved. Therefore, a multiple alignment of the signal sequences was created in ClustalW, to see how similar they actually were.

All default settings were used in uploading the sequences to both ClustalW and Clustal Omega, including using a Gonnet matrix for scoring the alignment. The final alignment of the signal sequences can be seen in **figure 3.13**, along with a phylogenetic tree, also created by ClustalW, in **figure 3.14**. The alignment of the full sequences can be see in **figure 3.15**, along with the corresponding phylogenetic tree in **figure 3.16**.

2.4 Transcriptome, proteome and promoters

When all the research on the LMCO's was done, we wanted to see how the *LMCO* in P11F6 were regulated when the substrate 2-methoxy-phenol was added to the growth media. As previous research had shown in **section 1.4**, the medium around the colonies turned brown, as an indication of oxidation of substrate. It was therefore expected to see an up-regulation of LMCO.

2.4.1 Transcriptome and proteome

Transcriptome and proteome analyses were performed by Christian Rückert and Animesh Sharma, respectfully. The chosen strain of *Psychrobacter*, P11F6, was grown both with and without substrate, and the transcriptome and proteome were measured. This resulted in a list of up/down regulated transcripts and proteins. First, the top ten list of upregulated transcripts was picked out, studied, and compared to the proteome results for the same proteins. Not being a part of this thesis, the exact details on how these studies were performed are not discussed further. Only the results from these studies are used further.

Following the finding of the top ten upregulated genes, the promoters for these genes were identified.

2.4.2 Promoters

Finding the promoters was done partly manually. This was done by identifying the top ten proteins based on M-values after transcriptomics analysis. One by one, these ten protein sequences were picked and analyzed. The DNA sequence was identified and isolated along with approximately 500 bp upstream of start codon. For proteins on the reverse strand, the sequence was reversed and complemented, and this sequence was further used for analysis. The final results were still given in forward strand notation.

The start/stop codons were identified, along with any differences found by SnapGene Viewer. The sequence upstream of the start codon was then analyzed using Softberry's BPROM (Solovyev, 2011), which is a promoter prediction program. BPROM bases its predictions on genes regulated by sigma70 promoters, which is one of the major promoter classes in *E.coli*. Combining characteristics describing oligonucleotide composition and functional motifs, the linear discriminant function (LDF) is created as a "score" of how good the prediction is. This number is based on five motifs found in promoters, distance between the -10 and -35 boxes and the frequencies of certain octanucleotides which are overrepresented in the transcription start sites. Using this information, the score is approximated as:

$$\mathrm{LDF} = \log\left(\frac{\mathrm{P(is\ a\ promoter)}}{\mathrm{P(is\ not\ a\ promoter)}}\right)$$

This means getting a score of 0 will only be a neutral value, and there are no upper/lower limits on this logarithmic scale. The threshold for predicting a promoter is set to 0.20, and every promoter predicted gets its own score.

As said, approximately 500 bp upstream of the start codon were used to predict promoters, even if these are usually found closer. Uploading the sequence lead to prediction of -10 and -35 regions and transcription start site. All possible regions were marked. The final search for ribosome binding sites (RBS's) was done manually in the sequence. Searches were done for AGGAGG, AGGAGN and AGCA. As most of the searches done by BPROM resulted in two possible promoters, the final search for RBS's was used to determine which one was the most reliable, along with comparing LDF scores. In all of the cases, the promoter region closest to the start codon was chosen.

Chapter 3

Results

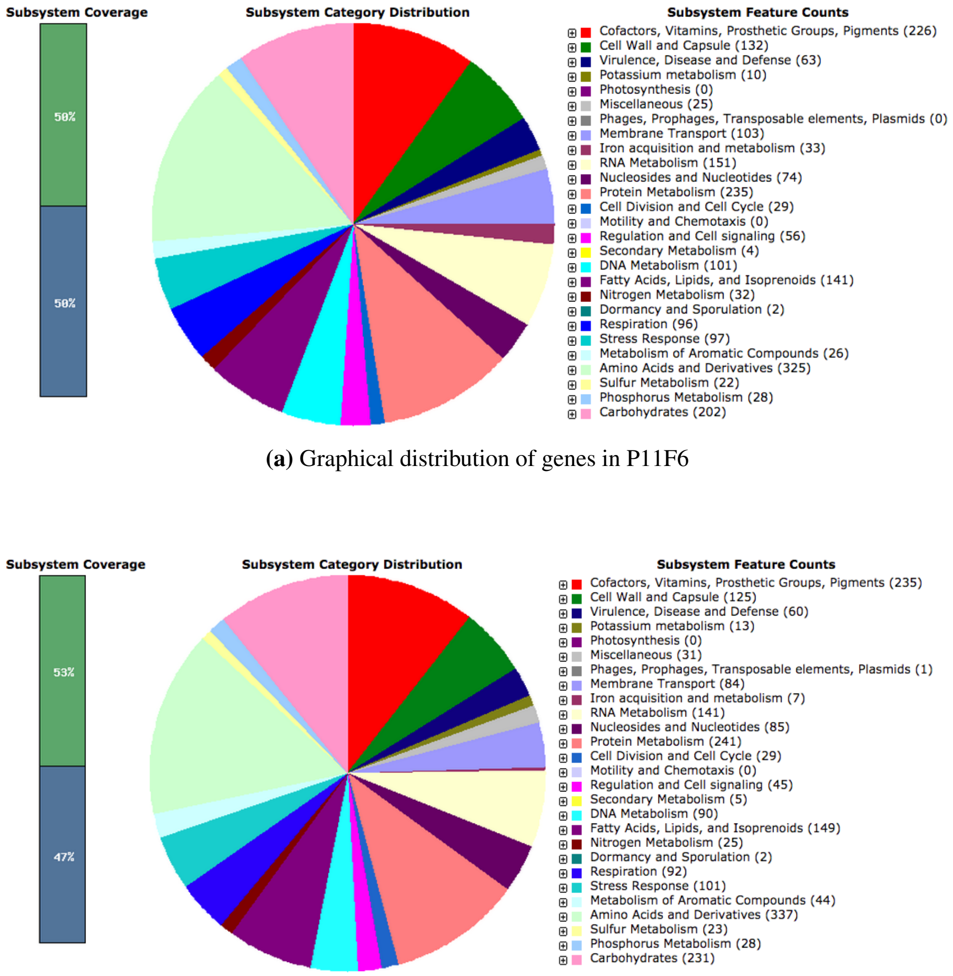
The goal in the first part of this study was to learn more about the four genomes of *Psychrobacter* spp., seeing their similarities and differences, and learning a bit about their evolutionary development. This was done using RAST for annotation, as seen in **section 2.1**, and Mauve for evolution in **section 2.2**. These results will be presented in **section 3.1** and **section 3.2**. Changing from genomes to proteins in part two, LMCO's were identified and investigated in various ways in **section 2.3**, with focus on size, amino acids, signal sequences and binding sites among others. The results from these analyses are all presented in **section 3.3**. The third and final part involves transcriptomics and proteomics of P11F6 grown on media containing 2-metoxy-phenol, where promoter sequences were found for the top ten up regulated genes, and transcription levels were compared for P11F6 growing on substrate/no substrate, as seen in **section 2.4**. These results are presented in **section 3.4**.

3.1 RAST

The first views of the results of a RAST annotation are shown in a graphic view, which include a percentage of subsystem coverage in a bar graph, a subsystem category distribution in a pie chart and a list of subsystem feature counts. The results from the annotations of the four novel *Psychrobacter* strains are shown in these graphical views in **section 3.1.1**, followed by a first comparison of the genomes in terms of size, CG-content and number of coding sequences, RNAs and subsystems in **section 3.1.2**. The final part of the RAST results includes a comparison of the distribution of genes into each subsystem in **section 3.1.3**.

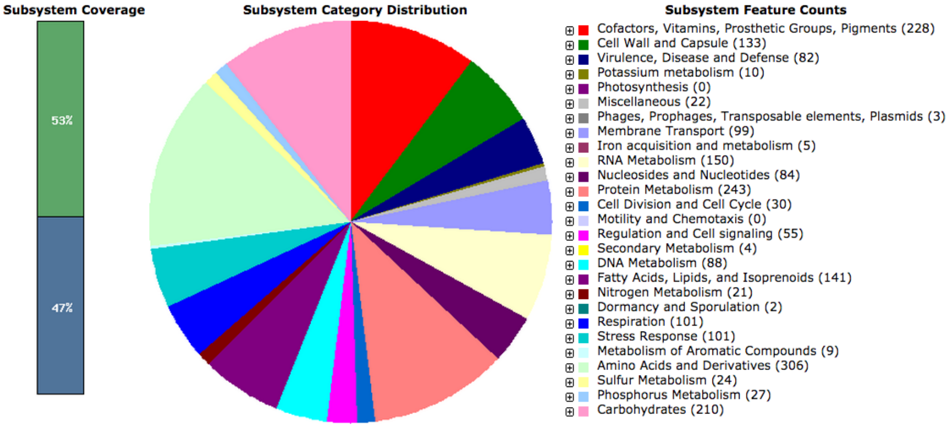
3.1.1 Graphical distribution of genes

The graphical view of the annotation of the four genomes of *Psychrobacter* spp. is shown in **figure 3.1**, divided into parts a-d. These graphical views include subsystem coverage, subsystem coverage distribution and subsystem features counts. As the figure shows, there are clear differences between the numbers subsystem coverage, the distribution in the pie charts are not identical and the corresponding numbers in the feature count are different when comparing the four genomes. These numbers will be revisited in **section 3.1.3**.

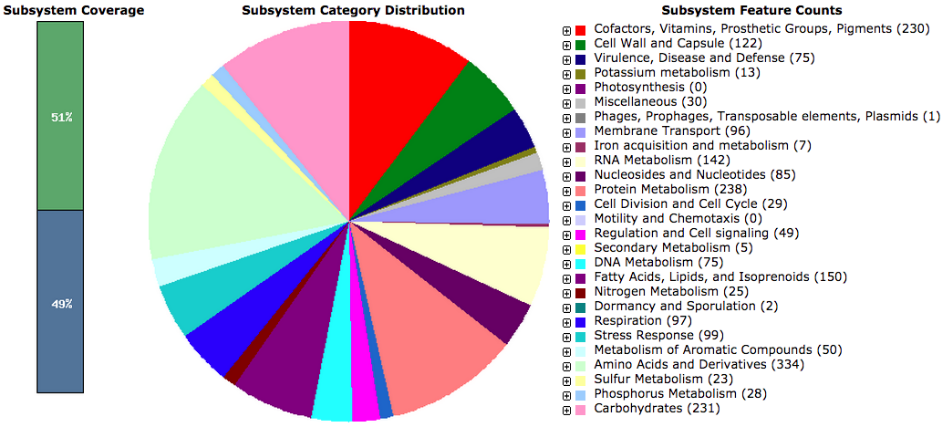


(a) Graphical distribution of genes in P11F6

(b) Graphical distribution of genes in P2G3



(c) Graphical distribution of genes in P11G3



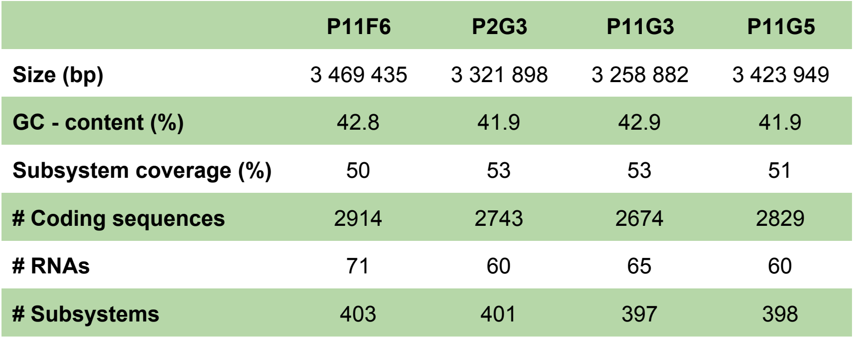
(d) Graphical distribution of genes in P11G5

Figure 3.1: All the graphical distributions of **a)** P11F6, **b)** P2G3, **c)** P11G3 and **d)** P11G5. This includes the subsystem coverage bar graph to the left, showing how many of the found genes could be placed in a subsystem. In the middle, a pie chart shows the distribution of genes in each subsystem. To the right is the same distribution of genes in subsystems as in the pie chart, only represented as the numbers that were used to create the pie chart.

3.1.2 Comparison of the genomes

The first comparison of the four strains of *Psychrobacter* spp., with comparisons of genome size, percentage of guanine and cytosine (GC), subsystem coverage, number of coding sequences, number of RNA's, and number of subsystems, is shown in **table 3.1**. The size of the genomes in the four strains varies from 3 258 882 to 3 469 435 base pairs, and the content of GC varies from 41.9 to 42.9 %. The number of coding sequences varies from 2674 to 2914 and the number of RNA's varies from 60 to 71. The number of subsystems varies from 397 to 403. It can be seen that P11F6, which has the highest number of base pairs (3 469 435), also has the highest number of coding sequences (2914), RNA's (71) and subsystems (403). This despite the fact that P11F6 has the lowest subsystem coverage of all four strains (50%).

Table 3.1: The first comparison of the four strains of *Psychrobacter* spp., focusing on size, GC-content, subsystem coverage and numbers of coding sequences, RNAs and subsystems found by RAST.



	P11F6	**P2G3**	**P11G3**	**P11G5**
Size (bp)	3 469 435	3 321 898	3 258 882	3 423 949
GC - content (%)	42.8	41.9	42.9	41.9
Subsystem coverage (%)	50	53	53	51
# Coding sequences	2914	2743	2674	2829
# RNAs	71	60	65	60
# Subsystems	403	401	397	398

3.1.3 Comparison of the subsystems feature counts

RAST divides the found genes into 27 main subsystems, each with various numbers of subgroups. The overview of the main subsystems and the genes found within each of them is shown in **table 3.2**. Variation between number of genes is found in almost all of the subsystems, which is as expected. Some of the numbers vary more than the others, such as in “Phages, prophages, transposable elements and plasmids”, where no genes are found in P11F6, while the others have at least one. In “Iron aquisition and metabolism” on the other hand, P11F6 has a total of 33, whereas the others have five-seven. Finally, in “Metabolism of aromatic compounds”, P11G3 has only nine genes in total, against 26 - 50 in the others.

Some of the numbers are perfectly similar throughout all the strains, such as “Photosynthesis”, “Motility and chemotaxis”, and “Dormancy and sporulation”, with all strains having zero, zero and two genes, respectfully. These variations and similarities might be due to the system coverage being low, as seen in **table 3.1**, and hence be natural. This will not be examined further in this thesis.

Table 3.2: Comparison of the total number of genes in each of the subsystems found by RAST, in each of the four genomes.

Subsystems	**P11F6**	**P2G3**	**P11G3**	**P11G5**
Cofactors, vitamins, prosthetic groups and pigments	226	235	228	230
Cell wall and capsule	132	125	133	122
Virulence, disease and defence	63	60	82	75
Potassium metabolism	10	13	10	13
Photosynthesis	0	0	0	0
Miscellaneous	25	31	22	30
Phages, prophages, transposable elements and plasmids	0	1	3	1
Membrane transport	103	84	99	96
Iron acquisition and metabolism	33	7	5	7
RNA metabolism	151	141	150	142
Nucleosides and nucleotides	74	85	84	85
Protein metabolism	235	241	243	238
Cell division and cell cycle	29	29	30	29
Motility and chemotaxis	0	0	0	0
Regulation and cell signaling	56	45	55	49
Secondary metabolism	4	5	4	5
DNA metabolism	101	90	88	75
Fatty acids, lipids and isoprenoids	141	149	141	150
Nitrogen metabolism	32	25	21	25
Dormancy and sporulation	2	2	2	2
Respiration	96	92	101	97
Stress response	97	101	101	99
Metabolism of aromatic compounds	26	44	9	50
Amino acids and derivates	325	337	306	334
Sulfur metabolism	22	23	24	23
Phosphorus metabolism	28	28	27	28
Carbohydrates	202	231	210	231
Total number of genes	**2213**	**2224**	**2178**	**2236**

3.2 Mauve

As the four genomes originate from four different species of *Psychrobacter*, it was natural to expect some genetic variations. This was already confirmed in **section 3.1**, where the various annotations showed genomic differences. Using Mauve, these variations were possible to visualize. As seen in **figure 3.2**, the full genomes are divided into blocks, which are located in various positions of the different genomes. The blocks of genes can move around within the genomes of different species. Mauve finds any similar blocks in the compared genomes, and draws a line between them. By only looking at the figure, it seems like P2G3 and P11G5 are the most similar genomes.

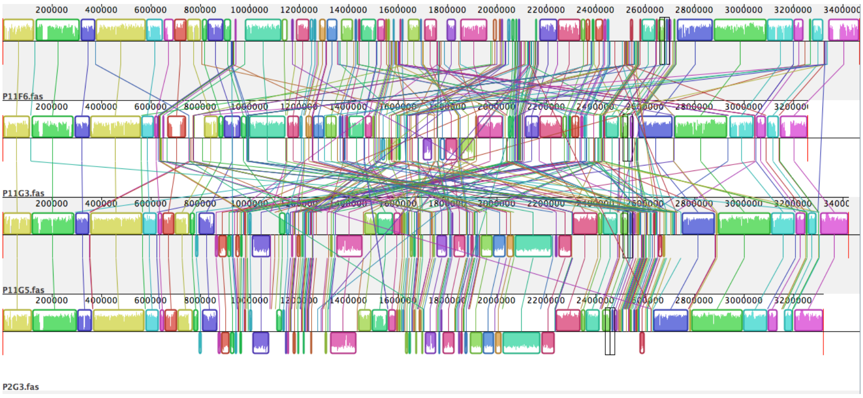


Figure 3.2: Graphic view of the different gene blocks in the four genomes, having different locations. Some of the blocks, such as the three furthest to the left, seem to be conserved in all four genomes, while others are moved around

Having certain genes in mind, it was possible to zoom in and follow the line from a block in one genome to the equivalent block in another genome. It was also possible to see that some of the blocks were close to conserved in all of the genomes, such as the pattern of yellow, green, purple, yellow, light blue and so forth in the far left end of the alignment seen in **figure 3.3**. It was possible to see how this region of the genomes are almost fully conserved in the pattern of yellow, green, purple and yellow blocks in the N-terminal part of the protein. As the figure shows, there are parts that are not fully conserved, in smaller blocks, such as the light blue line in the left of the picture, between the first yellow and the green block.

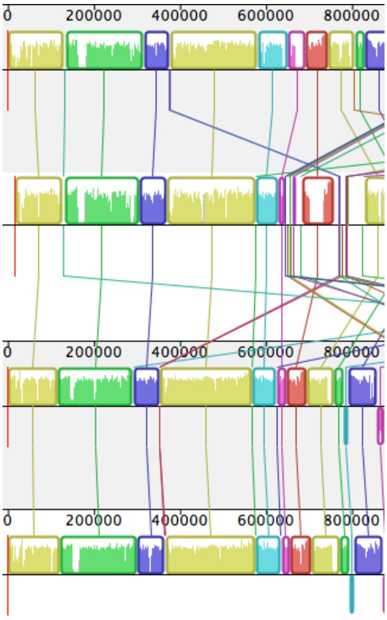


Figure 3.3: Closing in on the genomes it was possible to see conserved regions such as this one. The figure shows the N-terminal of the alignment, and how the blocks are almost identical in a pattern of yellow, green, purple and yellow. A tiny, hidden blue box can be seen only as a blue line in the two top genomes

By zooming in on the sequence around the blue line, it showed a tiny blue box, which can be seen in **figure 3.4**. The blue box is located between the conserved parts in the N-terminal of P11F6 and P11G3, but is in another location in P11G5 and P2G3. This shows how the lines follow the various blocks, and how the user can track down where the blocks are found in other genomes.

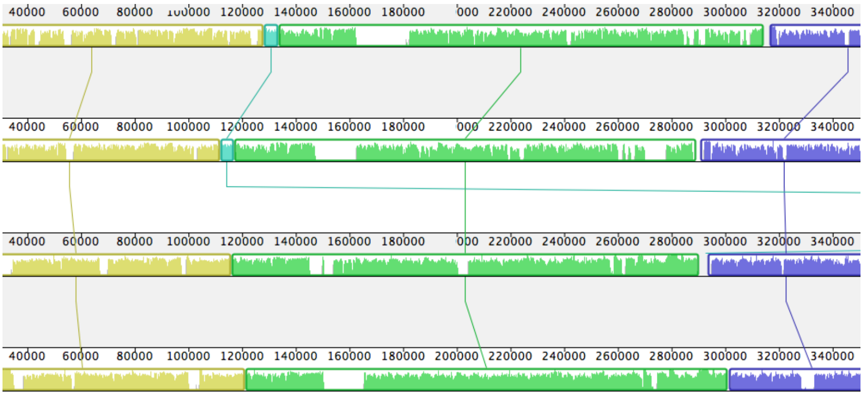


Figure 3.4: This figure is a close up of **figure 3.3**, showing the tiny hidden blue box, and how big it is once zoomed in on the region. The lines between the yellow, green and purple boxes show that the regions are conserved, and the line from the blue box, which is conserved only in the top two genomes, leads to another part of the remaining genomes

A particularly fun comparison of P11G5 (top line) and P2G3 (bottom line) is shown in **figure 3.5**. The two genomes are quite similar, only in different positions. It seems like the whole part of ∼ 20 000 bp is the same in the two genomes, only 15 000 bp's further down in P11G5 genome. This means that the genomes are similar, most likely from a common ancestor, and that they've evolved into something slightly different over time.

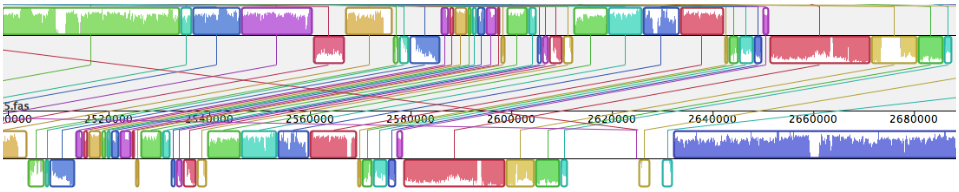


Figure 3.5: Comparing a specific part of P11G5 and P2G3, it's possible to see that the genomes are very similar, although not in the same position. All the oblique lines shows how the sequence is shifted, and that this is maintained throughout the whole region

As known from **figure 3.2**, the area covered in **figure 3.5** was not conserved in all genomes, although parts were similar. This shows how some parts of the genomes are similar and different depending on which genomes are compared.

Even if the Mauve comparison showed differences, it was not possible to determine the real relationship between the four *Psychrobacter* spp. based on this. In order to see how closely related these four actually were, a phylogenetic tree was made by Christian Rückert, placing these four (in bold) with 39 other *Psychrobacter* spp. and one specie of *Moraxella*. This is shown in **figure 3.6**, where it is possible to see how P11G3 was furthest away from the other *Psychrobacter* spp. in this thesis. The tree shows how P2G3 and P11G5 are closest related of these four, which was predicted from **figure 3.2**.

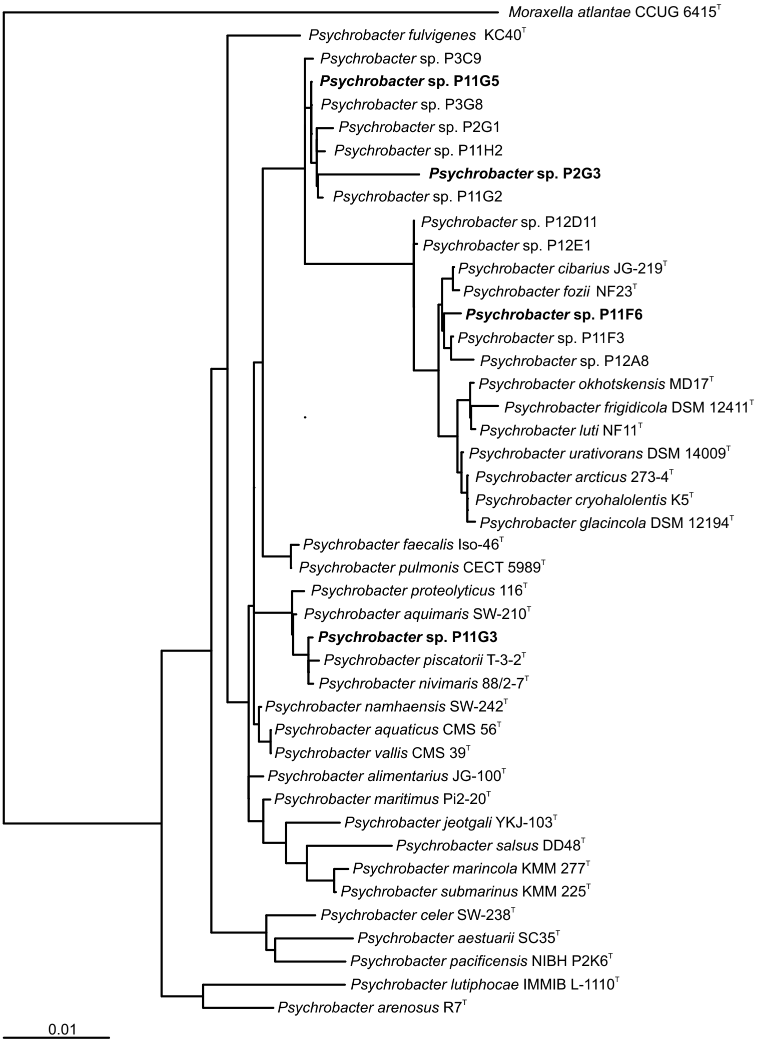

Moraxella atlantae CCUG 6415T
Psychrobacter fulvigenes KC40T
Psychrobacter sp. P3C9
Psychrobacter sp. P11G5
Psychrobacter sp. P3G8
Psychrobacter sp. P2G1
Psychrobacter sp. P11H2
Psychrobacter sp. P2G3
Psychrobacter sp. P11G2
Psychrobacter sp. P12D11
Psychrobacter sp. P12E1
Psychrobacter cibarius JG-219T
Psychrobacter fozii NF23T
Psychrobacter sp. P11F6
Psychrobacter sp. P11F3
Psychrobacter sp. P12A8
Psychrobacter okhotskensis MD17T
Psychrobacter frigidicola DSM 12411T
Psychrobacter luti NF11T
Psychrobacter urativorans DSM 14009T
Psychrobacter arcticus 273-4T
Psychrobacter cryohalolentis K5T
Psychrobacter glacincola DSM 12194T
Psychrobacter faecalis Iso-46T
Psychrobacter pulmonis CECT 5989T
Psychrobacter proteolyticus 116T
Psychrobacter aquimaris SW-210T
Psychrobacter sp. P11G3
Psychrobacter piscatorii T-3-2T
Psychrobacter nivimaris 88/2-7T
Psychrobacter namhaensis SW-242T
Psychrobacter aquaticus CMS 56T
Psychrobacter vallis CMS 39T
Psychrobacter alimentarius JG-100T
Psychrobacter maritimus Pi2-20T
Psychrobacter jeotgali YKJ-103T
Psychrobacter salsus DD48T
Psychrobacter marincola KMM 277T
Psychrobacter submarinus KMM 225T
Psychrobacter celer SW-238T
Psychrobacter aestuarii SC35T
Psychrobacter pacificensis NIBH P2K6T
Psychrobacter lutiphocae IMMIB L-1110T
Psychrobacter arenosus R7T
0.01

Figure 3.6: The four *Psychrobacter* spp. used in this studied placed in a phylogenetic tree (in bold) along with 39 other *Psychrobacter* spp. and one *Moraxella* sp., to see the relationship between the species. Figure by Christian Rückert.

3.3 Laccase-like Multicopper Oxidases

Turning focus on to a specific protein, LMCO's were found in the genomes of P2G3, P11G3 and P11G5 and in a plasmid of P11F6. As it turned out, both P11G3 and P11G5 had two copies of LMCO's; one on the forward strand and one on the reverse strand. The six LMCO's sequences are listed in **table 3.3**, along with direction, start/stop codons and lengths of open reading frames (ORFs) given both in bp and aa. The six LMCO sequences are of similar size, as expected, being the same protein. The table shows which strain the proteins come from, and which ID they will have for the remainder of this thesis. The main part of the LMCO's were found on the forward strand, and only the second proteins of P11G3 and P11G5 were found on the reverse strand. It also shows the positions and the length of the proteins, which vary from 565 to 568 aa's.

Table 3.3: The six *LMCO* sequences found within the four genomes, including start/stop, direction and length of ORF (bp/aa). The table also shows which ID's the proteins have for the rest of the thesis.

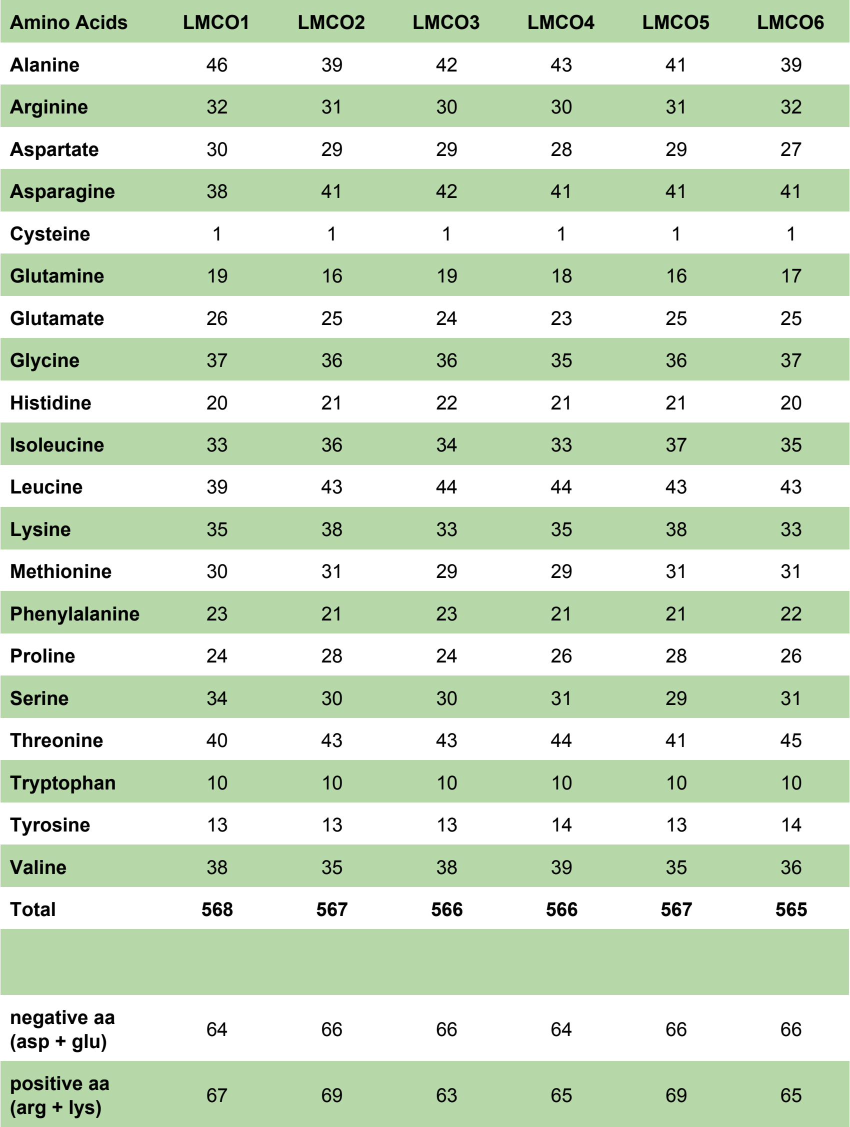
Strain	ID	Direction/strand	Start	Stop	ORF (bp/aa)
P11F6	LMCO1	Forward	17 811	19 517	1707/568
P2G3	LMCO2	Forward	981 591	983 294	1704/567
P11G3 - 1	LMCO3	Forward	634 765	636 465	1701/566
P11G3 - 2	LMCO4	Reverse	1 635 112	1 633 412	1701/566
P11G5 - 1	LMCO5	Forward	982 489	984 192	1704/567
P11G5 - 2	LMCO6	Reverse	2 716 692	2 714 995	1698/565

Once the LMCO's were found, their placement was compared to the results from RAST. It was expected to find the LMCO's in the subsystem of "metabolism of aromatic compounds", with the LMCO's oxidizing polyphenols and all. It was found, however, that RAST placed the LMCO's in the subsystem "virulence, disease and defence", sub-subsystem of "resistance to antibiotics and toxic compounds" and sub-sub-subsystem of "copper homeostasis". The nature of LMCO's, being multicopper oxidases, makes this as natural as placing them in the subsystem of "metabolism of aromatic compounds".

3.3.1 Amino acids and half-life

The LMCO's were further analysed with emphasis on the composition of aa, atoms and half-life of the six proteins, using ExPASy's ProtParam. These analyses were made in order to get a wider understanding of the protein's compositions. **Table 3.4** shows the composition of amino acids, which varies slightly. The only two that are identical in all six, are cystein and tryptophan. All six have only one cysteine, and also contains 10 residues of tryptophan. LMCO1 is found to be the protein with the highest number of aa's with 568, with LMCO2 and LMCO5 on second with 567, LMCO3 and LMCO4 on third with 566 and finally LMCO6 with 565 on fourth.

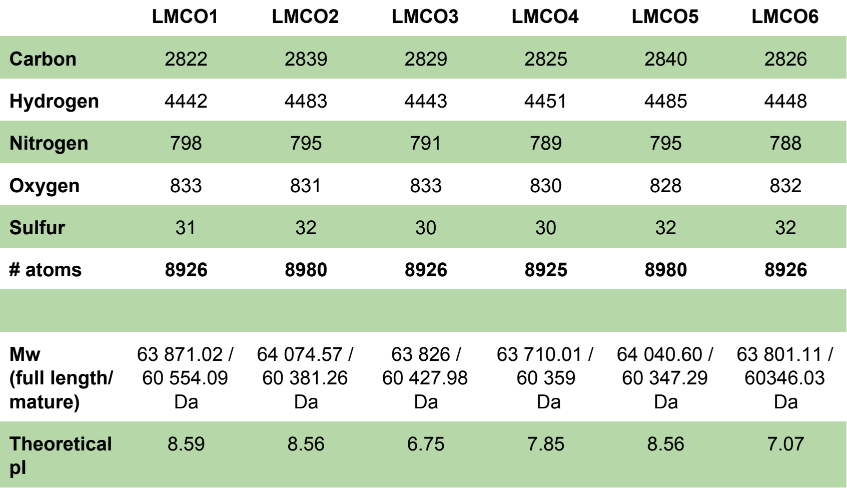
Table 3.4: The results of using ExPASy's ProtParam, showing the composition of amino acids found in all six LMCO's, including the number of positively and negatively charged residues.



Amino Acids	LMCO1	LMCO2	LMCO3	LMCO4	LMCO5	LMCO6
Alanine	46	39	42	43	41	39
Arginine	32	31	30	30	31	32
Aspartate	30	29	29	28	29	27
Asparagine	38	41	42	41	41	41
Cysteine	1	1	1	1	1	1
Glutamine	19	16	19	18	16	17
Glutamate	26	25	24	23	25	25
Glycine	37	36	36	35	36	37
Histidine	20	21	22	21	21	20
Isoleucine	33	36	34	33	37	35
Leucine	39	43	44	44	43	43
Lysine	35	38	33	35	38	33
Methionine	30	31	29	29	31	31
Phenylalanine	23	21	23	21	21	22
Proline	24	28	24	26	28	26
Serine	34	30	30	31	29	31
Threonine	40	43	43	44	41	45
Tryptophan	10	10	10	10	10	10
Tyrosine	13	13	13	14	13	14
Valine	38	35	38	39	35	36
Total	**568**	**567**	**566**	**566**	**567**	**565**
negative aa (asp + glu)	64	66	66	64	66	66
positive aa (arg + lys)	67	69	63	65	69	65

As **table 3.5** shows, the variations in composition of atoms are not that great, as the numbers are found to follow a pattern. LMCO2 and LMCO5 both contain a slightly higher level of carbon of around 30 extra atoms, and a more significantly raised level of hydrogen, with approximately 40 atoms more. This is, naturally, reflected in the total number of atoms, with these two proteins having 8980 vs. the rest having 8925-8926. Having the highest amount of atoms is reflected in the Mw, as these two proteins are the ones that are above 64 KDa. Not surprisingly, these two also have the exact same pI.

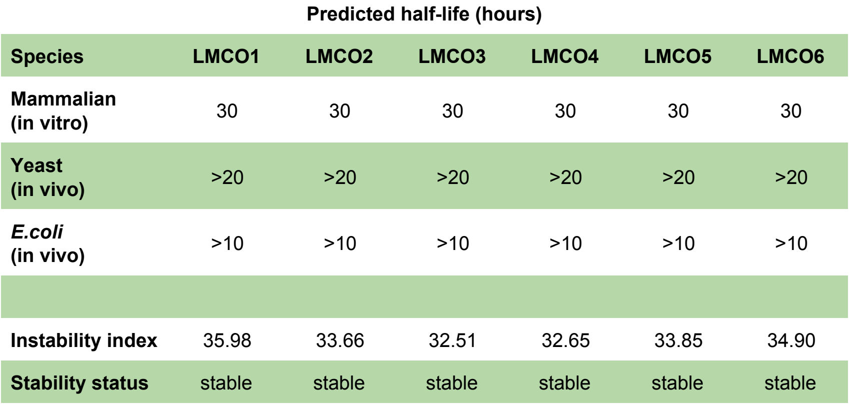
Table 3.5: The table shows the results from using ExPASy's ProtParam, giving the number and composition of atoms (C, H, N, O and S), pI and Mw for all LMCO's. Molecular weight was found for both full and mature sequence (Da).



	LMCO1	**LMCO2**	**LMCO3**	**LMCO4**	**LMCO5**	**LMCO6**
Carbon	2822	2839	2829	2825	2840	2826
Hydrogen	4442	4483	4443	4451	4485	4448
Nitrogen	798	795	791	789	795	788
Oxygen	833	831	833	830	828	832
Sulfur	31	32	30	30	32	32
# atoms	**8926**	**8980**	**8926**	**8925**	**8980**	**8926**
Mw (full length/ mature)	63 871.02 / 60 554.09 Da	64 074.57 / 60 381.26 Da	63 826 / 60 427.98 Da	63 710.01 / 60 359 Da	64 040.60 / 60 347.29 Da	63 801.11 / 60346.03 Da
Theoretical pI	8.59	8.56	6.75	7.85	8.56	7.07

As seen in **table 3.6**, the prediction of half-life in hours is given by three organisms; mammalian, yeast and *E.coli*. For mammals, the prediction is "in vitro", while it's "in vivo" for yeast and *E.coli*. As there are no predictions for *Psychrobacter* in specific, the closest predictions are the ones for *E.coli*. As the value of half-life is not a very precise one, here being ">10 h", it can only be suggested that the half-life for LMCO's found in *Psychrobacter* spp. is above 10 hours. It was also predicted that all LMCO's would be stable, as their values on the instability index were below 40.

Table 3.6: Using ExPASy's ProtParam, the predicted half-life and stability for all six LMCO's were found. The prediction defines half-life both in vivo and in vitro.



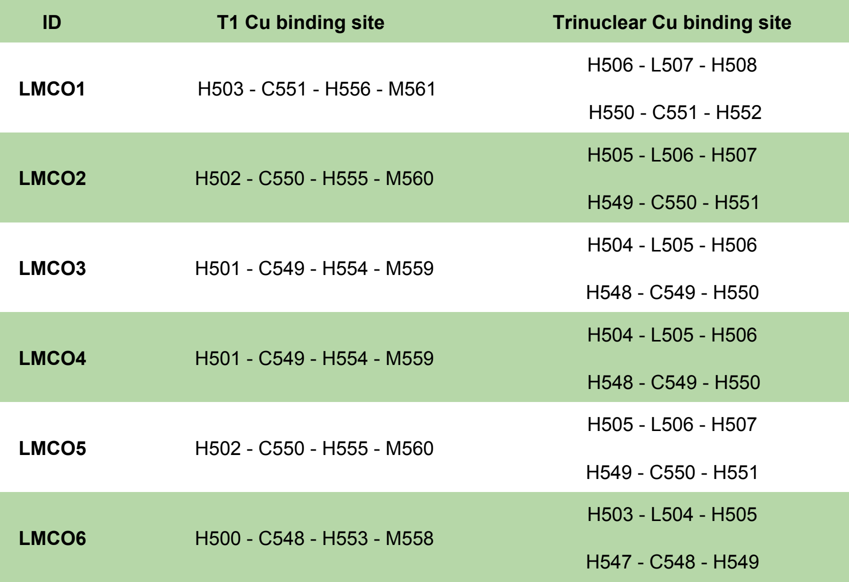
Predicted half-life (hours)

Species	**LMCO1**	**LMCO2**	**LMCO3**	**LMCO4**	**LMCO5**	**LMCO6**
Mammalian (in vitro)	30	30	30	30	30	30
Yeast (in vivo)	>20	>20	>20	>20	>20	>20
***E.coli* (in vivo)**	>10	>10	>10	>10	>10	>10
Instability index	35.98	33.66	32.51	32.65	33.85	34.90
Stability status	stable	stable	stable	stable	stable	stable

3.3.2 Copper sites

Using Phyre2's Investigator, it was possible to find the copper binding sites of the various LMCO's. As described in **section 2.3.2**, when using the Investigator, one has to choose one of the template models to compare the uploaded sequence to. All template models chosen here had an identity of 24-25 % and a confidence of 100.0. The T1 copper site was found for each protein, as well as the trinuclear sites. All of the findings are compared in **table 3.7**, where the prediction of copper sites were found to consist of His - Cys - His - Met and two times HXH, respectfully. The positions of the residues follow a pattern, as in the T1 site, where the Cys residue is placed 48 aa downstream of the first His, followed by another His five residues downstream and finally the Met residue five residues further downstream. This pattern is constant, even if the start varies between the LMCO's. For the trinuclear site, the prediction shows the first HXH being positioned three residues downstream of the first His in the T1 site, and the final HXH is surrounding the Cys residue of T1.

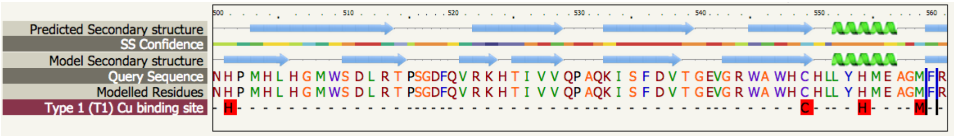
Table 3.7: The table shows the Cu sites found by Phyre2. This includes both T1 and trinuclear sites, and shows how the T1 site was made out of His - Cys - His - Met, while the trinuclear site consists of two times HXH.



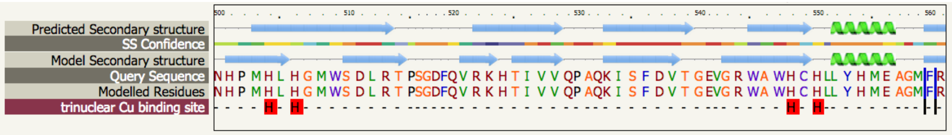
ID	T1 Cu binding site	Trinuclear Cu binding site
LMCO1	H503 - C551 - H556 - M561	H506 - L507 - H508 H550 - C551 - H552
LMCO2	H502 - C550 - H555 - M560	H505 - L506 - H507 H549 - C550 - H551
LMCO3	H501 - C549 - H554 - M559	H504 - L505 - H506 H548 - C549 - H550
LMCO4	H501 - C549 - H554 - M559	H504 - L505 - H506 H548 - C549 - H550
LMCO5	H502 - C550 - H555 - M560	H505 - L506 - H507 H549 - C550 - H551
LMCO6	H500 - C548 - H553 - M558	H503 - L504 - H505 H547 - C548 - H549

Using LMCO4 as an example, both T1 and trinuclear sites are shown as seen when using Phyre2 in **figure 3.7**. The figure shows six rows of information, and above them is a line defining each residue's number. The bottom three rows are interesting in this case, as they show the query sequence, the model's residues and the binding sites. In other words, they show the uploaded sequence, the modelled sequence which the uploaded sequence is compared against, and finally the found sites. The binding sites are marked as red boxes in the bottom row. The top three rows are for prediction of secondary structure, which is covered in **section 3.3.4**.

As **figure 3.7** shows, the residues the binding site predictions are composed of are marked with red boxes. These figures sum up the results from **table 3.7** and, as one can see in the figure, Phyre2 found the T1 Cu site in His501 - Cys549 - His554 - Met559, and trinuclear sites in His504 - His506 and His548 - His550. The line of numbers at the top of the figures indicates the positions of the residues, which can be traced back to the ones mentioned in the table. The prediction of secondary structure in both the model and the uploaded sequence, here LMCO4, are not taken into concern in this part, but will be further investigated in **section 3.3.4**.


500
510
520
530
540
550
560
Predicted Secondary structure
SS Confidence
Model Secondary structure
Query Sequence
Modelled Residues
Type 1 (T1) Cu binding site
NHPMHLHGMWSDLRTPSGDFQVRKHTIVVQPAQKISFDVTGEVGRWAWHCHLLYHMEAGMFR
NHPMHLHGMWSDLRTPSGDFQVRKHTIVVQPAQKISFDVTGEVGRWAWHCHLLYHMEAGMFR
H
C
H
M

(a) T1 Cu binding sites found in LMCO4, marked with red boxes in the bottom row.


500
510
520
530
540
550
560
Predicted Secondary structure
SS Confidence
Model Secondary structure
Query Sequence
Modelled Residues
trinuclear Cu binding site
NHPMHLHGMWSDLRTPSGDFQVRKHTIVVQPAQKISFDVTGEVGRWAWHCHLLYHMEAGMFR
NHPMHLHGMWSDLRTPSGDFQVRKHTIVVQPAQKISFDVTGEVGRWAWHCHLLYHMEAGMFR
H
H
H
H

(b) Trinuclear Cu binding sites found in LMCO4, marked with red boxes in the bottom row.

Figure 3.7: Showing how Phyre2 showed its findings of T1 and trinuclear Cu sites, here marked as red boxes. This shows how T1 was predicted as His - Met - His - Cys, while the trinuclear site was predicted as two times HXH.

As seen in **section 1.2**, the trinuclear site is supposed to consist of eight His residues, and not only four, as was found by Phyre2. A manual search for the remaining four was therefore performed, using the alignment of the full protein sequences, which will be presented in **section 3.3.5**. This lead to the findings presented in **table 3.8**, where the two remaining His residues are shown, meaning that all eight His residues of the trinuclear site were identified. Using the signature sequences for confirming the conservation of the residues, the residues in **table 3.8** were considered the ones responsible for Cu binding to the LMCOs.

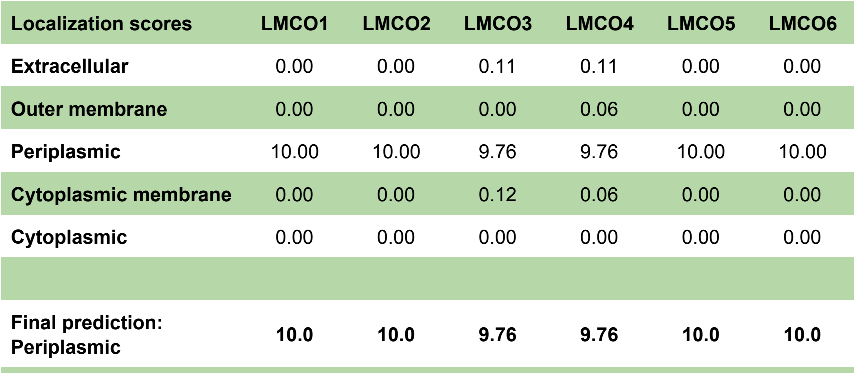
Table 3.8: After a manual search, the table of Cu sites based on Phyre2 had to be revised, to include the remaining His residues that Phyre2 could not find. This table shows how all four HXH were possible to find, along with the T1 site.

ID	T1 Cu binding site	Trinuclear Cu binding site	
LMCO1	H503 - C551- H556 - M561	H506 - L507- H508 H116 - W117 - H118	H550 - C551 - H552 H158 - S159 - H160
LMCO2	H502 - C550 - H555 - M560	H505 - L506 - H507 H115 - W116 - H117	H549 - C550 - H551 H157 - S518 - H159
LMCO3	H501 - C549 - H554 - M559	H504 - L505 - H506 H114 - W115 - H116	H548 - C549 - H550 H156 - S157 - H158
LMCO4	H501 - C549 - H554 - M559	H504 - L505 - H506 H114 - W115 - H116	H548 - C549 - H550 H156 - S157 - H158
LMCO5	H502 - C550 - H555 - M560	H505 - L506 - H507 H115 - W116 - H117	H549 - C550 - H551 H157 - S158 - H159
LMCO6	H500 - C548 - H553 - M558	H503 - L504 - H505 H113 - W114 - H115	H547 - C548 - H549 H155 - S156 - H157

3.3.3 Subcellular location

Using PSORT-B it was possible to identify the location of the protein based on the signal sequence, which showed that all LMCO's were destined to end up in the periplasm, as seen in **table 3.9**. All of the five possible location were tested on got a score based on this, and the highest score was found to be the destination.

Table 3.9: Using PSORT-B, all of the possible subcellular locations got their own score, deciding which of the categories the proteins belonged to. This showed that all of the LMCO's were destined for the periplasm.



Localization scores	LMCO1	LMCO2	LMCO3	LMCO4	LMCO5	LMCO6
Extracellular	0.00	0.00	0.11	0.11	0.00	0.00
Outer membrane	0.00	0.00	0.00	0.06	0.00	0.00
Periplasmic	10.00	10.00	9.76	9.76	10.00	10.00
Cytoplasmic membrane	0.00	0.00	0.12	0.06	0.00	0.00
Cytoplasmic	0.00	0.00	0.00	0.00	0.00	0.00
Final prediction: Periplasmic	**10.0**	**10.0**	**9.76**	**9.76**	**10.0**	**10.0**

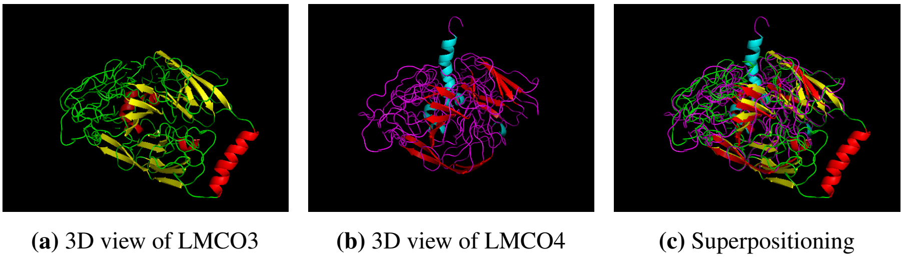
3.3.4 PyMOL

PyMOL is an excellent tool for visualizing proteins, and was in this thesis used to visualize the 3D structures of the LMCO's, including superpositioning the sequences, and identifying the substrate pockets. By using different colours for different secondary structures or pockets, the structures became more clear.

3D models and superpositioning

When it came to showing the 3D models and superpositioning them, each of the models were first shown by themselves and then together as superpositioned, showing how the structures were different and similar.

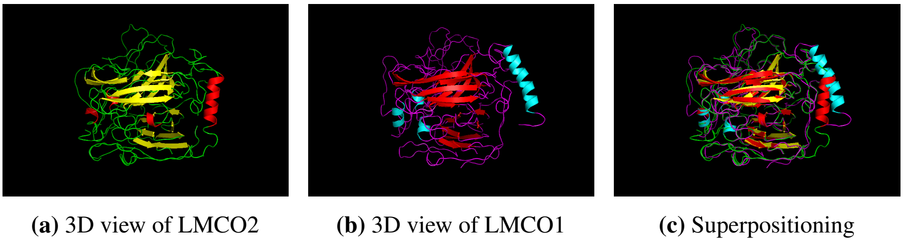
As seen in **figure 3.8**, figure a and figure b have a few differences. In figure c, figures a and b seem to be simply laid on top of each other, although the sequences were in fact aligned. Following the strands, it is possible to see how these two proteins are similar and different, such as with the α helices, marked in red (**3.8a**) and light blue (**3.8b**), which does not have the same orientation at all. Parts of the β sheets are also in different places, showing the proteins to be even more different.



(a) 3D view of LMCO3 **(b)** 3D view of LMCO4 **(c)** Superpositioning

Figure 3.8: The two LMCO's from P11G3; LMCO3 (left) and LMCO4 (middle), in 3D view made with PyMOL. The picture on the right is a superpositioning of the two proteins. Even without the superpositioning, the models shows great differences, such as with the major α helix being in a very different place. The β sheets are partly similar, as seen when the β sheet arrows are both yellow and red, although there are parts of β sheet that are only red or yellow.

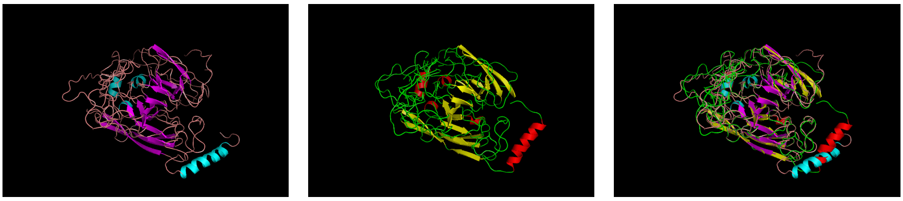
Just as with the proteins in **figure 3.8**, LMCO1 and LMCO2 were shown both separately and together in **figure 3.9**, showing similarities and differences. Seeing the α helix in **figure 3.9b**, this helix seemed more similar to the one in **figure 3.8a**, although their sizes seemed to be different. This lead to the comparison of LMCO1 and LMCO3, as it was suspected that these two would be more similar than the alignments so far. Models and superpositionings are shown in **figure 3.10**.



(a) 3D view of LMCO2 **(b)** 3D view of LMCO1 **(c)** Superpositioning

Figure 3.9: The two LMCO's; LMCO2 (left) and LMCO1 (middle), in 3D view made with PyMOL. The picture to the right is a superpositioning of the two proteins, showing similarities and differences. The models by themselves show more similarity than in the comparisons seen so far, having similar structures of β sheet and having the major α helix in a more similar position. The minor α helices are not in the same positions however, which can be seen in the superpositioning.

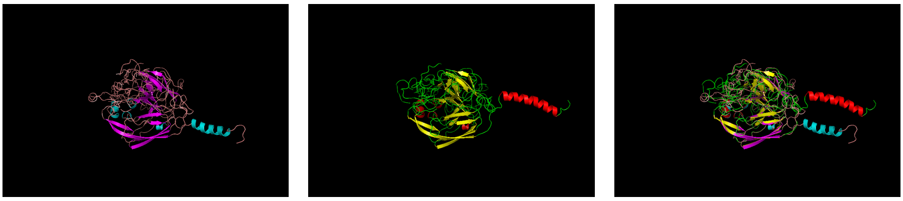
As seen in **figure 3.10**, the alignment of LMCO1 and LMCO3 does seem more similar than the other alignments. This can for instance be seen in the two major α helices, seemingly having the same length and being in the same place. The arrangement of β sheets and the smaller α helices also seem to have a more similar alignment than the other alignments, which might indicate that the sequences of LMCO1 and LMCO3 are closer related than LMCO1 and LMCO2, or LMCO3 and LMCO4.



(a) 3D view of LMCO1 **(b)** 3D view of LMCO3 **(c)** Superpositioning

Figure 3.10: The two LMCO's; LMCO1 (left) and LMCO3 (middle), in 3D view made with PyMOL. These two were chosen for comparison due to expected similarity, which is clear in both the models by themselves and in the superpositioning. Both the β sheets and the α helices, both minor and major, are found in similar positions. The one thing that is different is the turns and stretches of sequences which are not α or β structures, here marked in green or pink. The superpositioning shows how these parts are partly same and partly different, as well as showing how the α helices are in close positions but not identical.

As all the other LMCO's had been investigated and superpositioned, it was only fair to see how the 3D structures of LMCO's from P11G5 looked, which can be seen in **figure 3.11**. The 3D structure of the LMCO's from P11G5 look a bit different than the other LMCO's, especially with regards to the α helix pointing straight out from the rest of the protein. The overall superpositioning of LMCO5 and LMCO6 shows structures of high similarity, perhaps indicating a close relationship. All these indications of the proteins' relationships between each other will be further considered in **section 3.3.5**.



(a) 3D view of LMCO5 **(b)** 3D view of LMCO6 **(c)** Superpositioning

Figure 3.11: The two LMCO's from P11G5; LMCO5 (left) and LMCO6 (middle), in 3D view made with PyMOL. The models show how the β sheets are similar, especially in the bottom part of the models, as well as having the major α helix point straight out to the right. In the superpositioning, these two α helices are shown to not point in the exact same direction, although both pointing out on the right hand side. With the β sheets being both yellow and purple, this indicates them being in the same place.

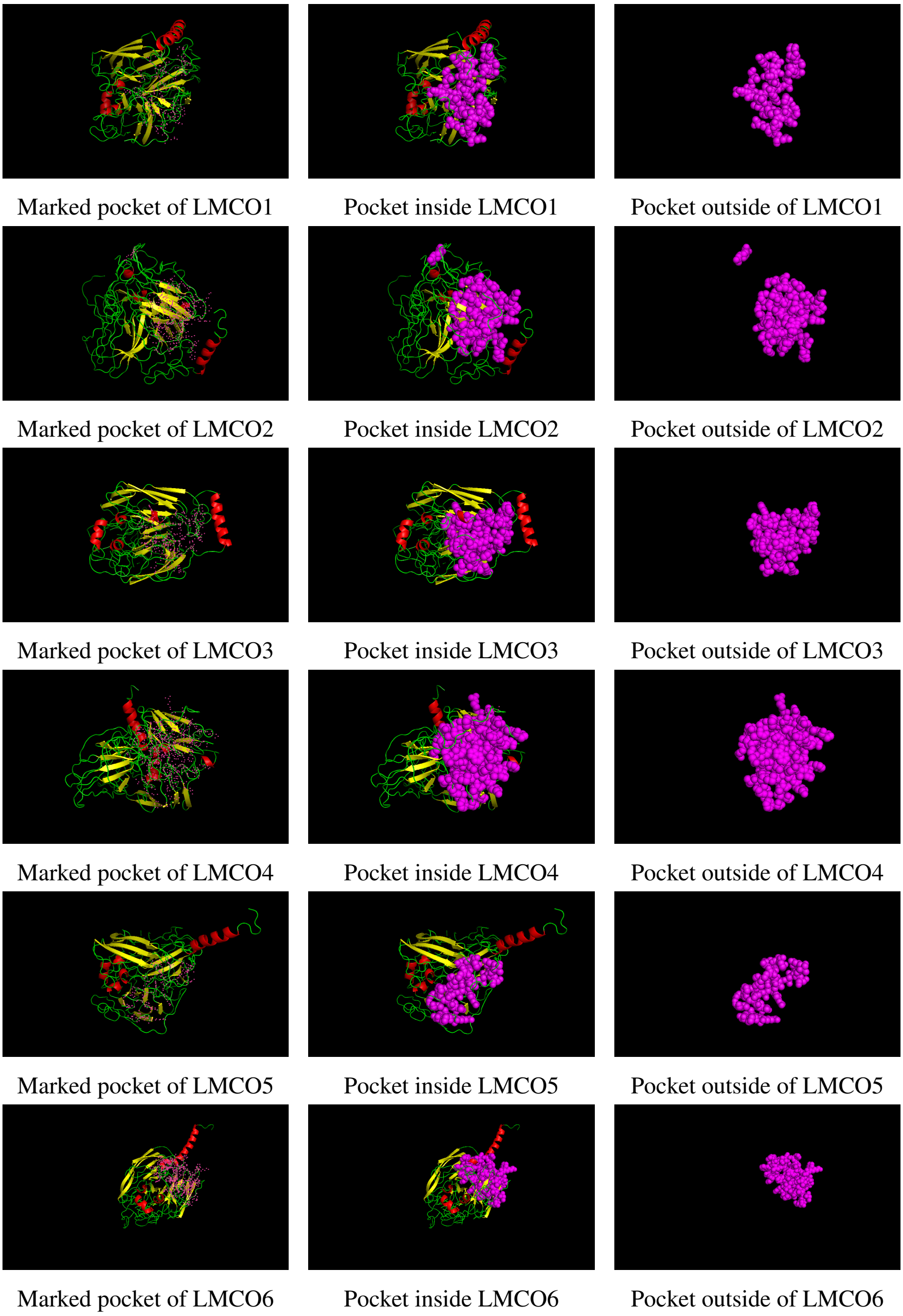
Pocket analysis

Following the structure comparisons, the pocket analysis from Phyre2 was also visualized in PyMOL. The point here was to see how the substrate pockets were oriented in the overall folding of the proteins, both marked as the residues involved, having the residues shown as spheres, and extracted pocket spheres.

By putting all of the pockets next to each other, as done in **figure 3.12**, it's possible to visualize and compare the pocket structures. The pockets here are partly revised in **table 3.10**, and the full pocket residue list is shown in **Appendix A**.

Each of the LMCO's has its own line of three figures. The first column shows the overall 3D structures, much as seen in the superpositionings, only this time with tiny pink markings, indicating the position for each of the residues that are part of the pockets. There are several pink marks for each of the residues, as they resembles the residue in structure more than just the position. In the mid column, the tiny pink marks have been transformed into pink spheres, showing the overall pocket. In the final column, the rest of the proteins have been removed, leaving only the pockets for comparisons.

By only looking at the pockets, there are a few things that stand out. One example of this is the pocket of LMCO2, having a tiny part of the pocket which is not connected to the rest of the pocket. The pockets are not made out of one continuous sequence, although all the other pockets are connected enough to create one whole figure in sphere mode. Another example of pockets that stands out is LMCO3, clearly being bigger than the others. LMCO6 is seemingly smaller than the others, although this might have something to do with the model being smaller.



Marked pocket of LMCO1 Pocket inside LMCO1 Pocket outside of LMCO1

Marked pocket of LMCO2 Pocket inside LMCO2 Pocket outside of LMCO2

Marked pocket of LMCO3 Pocket inside LMCO3 Pocket outside of LMCO3

Marked pocket of LMCO4 Pocket inside LMCO4 Pocket outside of LMCO4

Marked pocket of LMCO5 Pocket inside LMCO5 Pocket outside of LMCO5

Marked pocket of LMCO6 Pocket inside LMCO6 Pocket outside of LMCO6

Figure 3.12: Visualization of the substrates pockets in all six LMCO's. Each row belongs to a specific LMCO, all lines showing three "states" of pocket: Marked in the structure with tiny pink residues, shown with the pink markings replaced with pink spheres, and finally the pockets alone.

Looking at something more solid, **table 3.10** shows a shortened version of the residues that make the pockets. As the pockets in general consist of a large amount of residues, the full list can be seen in **Appendix A**. The table here on the other hand, includes the 13 parts where the pockets are similar, including places where one would expect that the pockets were similar. This table also includes the total number of residues in each pocket, shown in the head line.

As the first line in **table 3.10** shows, the number of residues in the pockets vary from 32 to 83, which is a fairy large variation considering these proteins have similar sequences. Some of the green boxes show specifically interesting features regarding conserved residues, such as number two from the top, showing that all pockets have a Glu residue at $\sim$96. The positions are still similar enough to be grouped as the same residue. Similar is green box number three, where three of the pockets have similar sequences of multiple residues. The remaining three have something similar, as in LMCO2's case, where three of the residues are similar, or as in the case for LMCO1 and LMCO5, which only have a Leu residue at $\sim$120. This corresponds to similar Leu residues in all of the other LMCO's. Seeing the known HXH of the trinuclear site in three of the pockets made this an interesting part.

In green box number five, a highly conserved section of Gly, Asn/Ile and Asp are found in almost exactly the same position in all pockets, being the only pocket residues in this part of the sequence. Green box number eight resembles the second green box, with three pockets having a certain sequence of multiple residues, followed by two pockets which has almost the same sequence, and finally one pocket which has only one of them. In this case, the last one is LMCO1, having only the Arg490 similar to the others. The ninth green box includes another HXH site, which is conserved in three of the pockets. Once again, a Leu is found to be partly conserved. In the eleventh field, a promising conserved start of Gly - Met - Trp is found in all LMCO's, and by comparing the pocket further, this sequences seem to be followed by Ser - Asp - Leu, although these are just partly conserved among the different pockets.

The twelfth green field contains a fully conserved part of Val - Thr - Gly - Glu, with some surrounding residues being partly conserved. Five out of the six pockets have Val - Val as their two last residues, except from LMCO1, which again has no residues in this area.

Table 3.10: Parts of the pockets that were particularly similar, in an attempt to find conserved parts of the pockets. Each green field contains a number of residues that are found to be in at least three of the pockets, and in some cases in all of them.

LMCO1 (36)	**LMCO2 (57)**	**LMCO3 (48)**	**LMCO4 (83)**	**LMCO5 (32)**	**LMCO6 (72)**
Glu97	Glu96	Glu95	Glu95	Glu96	Glu94
		His114	His114		His113
		Trp115	Trp115		Trp114
		His116	His116		His115
		Gly117	Gly117		Gly116
	Leu119	Leu118	Leu118		Leu117
Leu121	Leu120	Leu119	Leu119	Leu120	Leu118
	Val121		Val120		Val119
		Leu147	Leu147		Ile147
Lys150	Lys149	Val148	Lys148	Leu148	Gln148
Gln151	Gln150	Gln149	Gln149	Lys149	Ser149
Ser152	Ser151	Ser150	Ser150	Gln150	
			Gly151	Ser151	
Lys178		Lys176	Lys176	Lys177	Lys175
Gly315	Gly314	Gly313	Gly313	Gly314	Gly312
Asn316	Ile315	Asn314	Ile314	Ile315	Ile313
Asp317	Asp316	Asp315	Asp315	Asp316	Asp314
Pro415	Pro414		Pro413		Pro412
Arg416					
Met417	Met416			Met416	Met414
	Asn424		Asn423		Asn422
	Arg426		Arg425		Arg424
	Val482		Val481		Val480
	Ile484	Ile483	Ile483	Ile484	Ile482
	Lys485				Lys483
	Pro486	Pro485	Pro485	Pro486	Pro484
Arg490	Arg489	Arg488	Arg488	Arg489	Arg487
	Val490	Val489	Val489	Val490	Val488
	Ile492	Ile491	Ile491	Ile492	Ile490
	Thr493		Thr492		Thr491
	Met504		Met503		Met502
		His504	His504		His503
	Leu506	Leu505	Leu505		Leu504
		His506	His506		His505

Gly509	Gly508	Gly507	Gly507	Gly508	Gly506
Met510	Met509	Met508	Met508	Met509	Met507
Trp511	Trp510	Trp509	Trp509	Trp510	Trp508
Ser512	Ser511		Ser510		Ser509
Asp513					Asp510
	Leu513		Leu512		Leu511

Gln522					Gln519
	Val522		Val521		Val520
Arg524	Arg523	Arg522		Arg523	Arg521
	Lys524		Lys523		Lys522
	His525	His524	His524		His523
			Thr525		Thr524
	Ile527		Ile526		Ile525
			Ile526		

	Phe537		Phe536		Phe535
Asp539	Asp538		Asp537		Asp536
Val540	Val539	Val538	Val538	Val539	Val537
Thr541	Thr540	Thr539	Thr539	Thr540	Thr538
Gly542	Gly541	Gly540	Gly540	Gly541	Gly539
Glu543	Glu542	Glu541	Glu541	Glu542	Glu540
	Ala543				Ala541
	Trp546		Trp545		Trp544
	Trp548	Trp545	Trp547	Trp546	Trp546

	Arg562		Arg561		Arg560
	Glu563		Glu562		Glu561
	Val564	Val563	Val563	Val564	Val562
	Val566	Val565	Val565	Val566	Val564

To sum up **table 3.10**, there are certain areas of the pockets which are highly similar in all, or nearly all, of the pockets. Having long stretches of similar residues in the same position for many of the pockets, this seemed to be too similar to accept as a coincidence. Seeing that LMCO6 had the second highest number of residues makes it a puzzle why the model of the pocket turned out to be so much smaller than the others, and the fact that the model of the protein itself is a bit smaller than the other ones should not be responsible for all of that.

3.3.5 Alignments

Multiple alignments were made for both the signal sequences and the full sequences of the LMCO's, in order to compare them and discover the relationship between them. ClustalW was used for the signal sequences, and Clustal Omega for the full sequences.

Reading the alignments should be done based on the signs below each position and the colour of the residues. There are three consensus symbols:

***** indicates all residues in the position are fully conserved

: indicates partly conserved residues, having strongly similar properties

. indicates partly conserved residues, having weakly similar properties

Further, the colouring of the residues are based on their properties:

Red: Small and hydrophobic aa's, including aromatic (not Y); AVFPMILW

Blue: Acidic aa's; DE

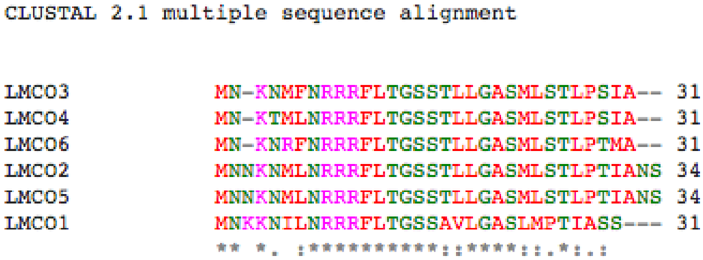
Magenta: Basic aa's (not H); RK

Green: Hydroxyl + sulhydryl + amine + G; STYHCNGQ

This means that fully conserved residue will have an * below, and all the residues above it will be the same, and hence have the same colour. Residues that have similar properties of varying degree will have : or . below, and may have either similar or different colour on their residues.

Signal sequences

The signal sequences vary from 31 to 34 aa, and an alignment of the six sequences can be found in **figure 3.13**. The signal sequences are highly conserved in most of the residues, with the sequences of LMCO6 and LMCO2 being identical, and LMCO3, LMCO4 and LMCO5 being quite similar. LMCO1 is shown to be the most different from the other signal sequences.



```
CLUSTAL 2.1 multiple sequence alignment


LMCO3           MN-KNMFNRRRRFLTGSSTLLGASMLSTLPSIA-- 31
LMCO4           MN-KTMLNRRRRFLTGSSTLLGASMLSTLPSIA-- 31
LMCO6           MN-KNRFNRRRRFLTGSSTLLGASMLSTLPTMA-- 31
LMCO2           MNNKNMLNRRRRFLTGSSTLLGASMLSTLPTIANS 34
LMCO5           MNNKNMLNRRRRFLTGSSTLLGASMLSTLPTIANS 34
LMCO1           MNKKNILNRRRRFLTGSSAVLGASLMPTIASS--- 31
                ** *. :***********::****::.*:.:
```

Figure 3.13: Multiple alignment by ClustalW, of the six signal sequences of the chosen LMCO's

This is also shown in **figure 3.14**, which shows a phylogenetic tree based on the alignment. The signal sequences of LMCO2 and LMCO5 are found to be identical, with LMCO1 being furthest away from the others. The numbers indicates how big the difference is between the sequences. The fact that the sequence from LMCO1 is most different from the others can be due to the fact that this gene is found in a plasmid.

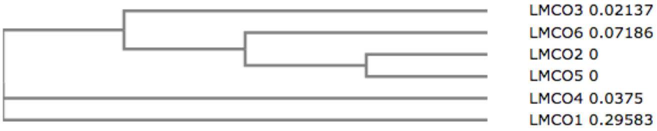


Figure 3.14: When Clustal aligns sequences, it also creates a phylogenetic tree. This tree is for the signal sequences only, showing how LMCO1 is furthest away from the others along with LMCO4, and LMCO2 and LMCO5 are the closest related proteins, which is as expected since they were found to be identical in the alignment

Full alignment

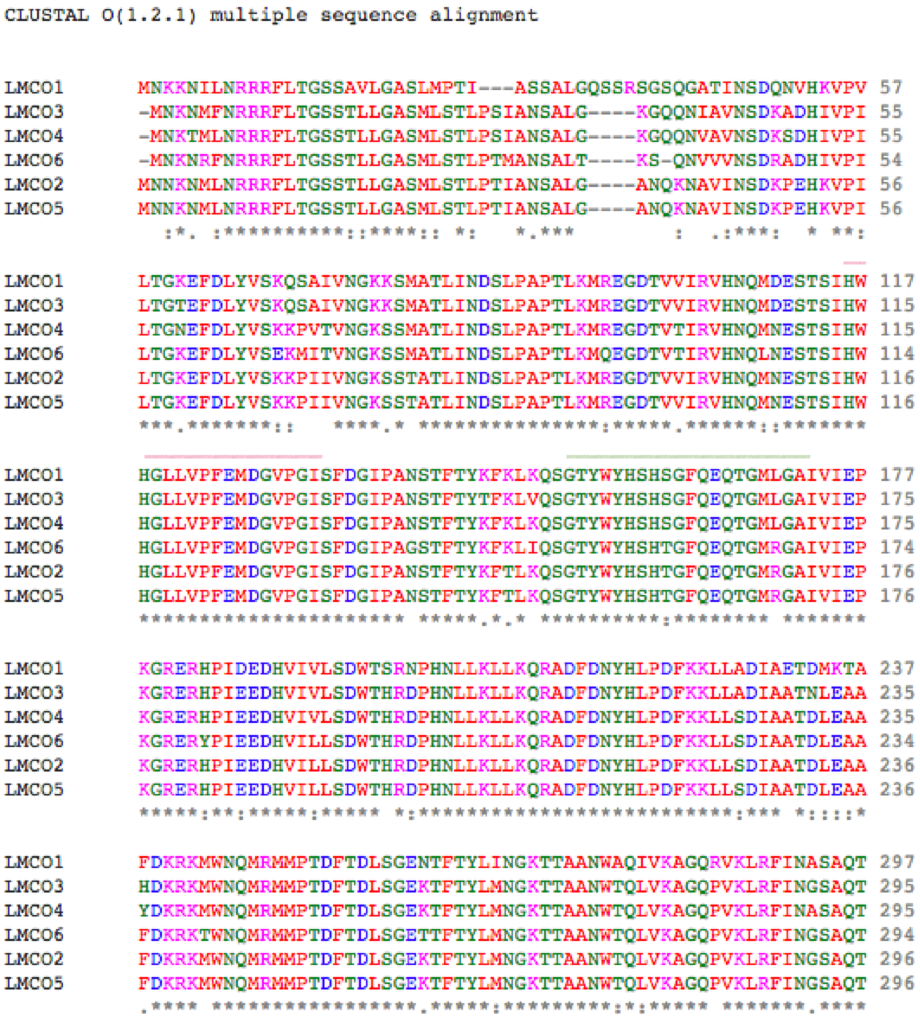
A multiple alignment was also made for the full length LMCO's, as seen in **figure 3.15**. This alignment includes the signal sequences, and as seen, the first $\sim$ 32 aa of the alignment are identical to the ones in **figure 3.13**, although the alignment itself is done differently.

As the number of * in the alignment in **figure 3.15** tells, the amount of residues that are fully conserved in all of the LMCO's is high. There are a few places with variations, but as indicated by the colours, most of the variations are made by residues with similar chemical properties. There are naturally residues with very different chemical properties, but these seem to not affect the overall folding in tertiary structure, as was seen in **section 3.3.4**.

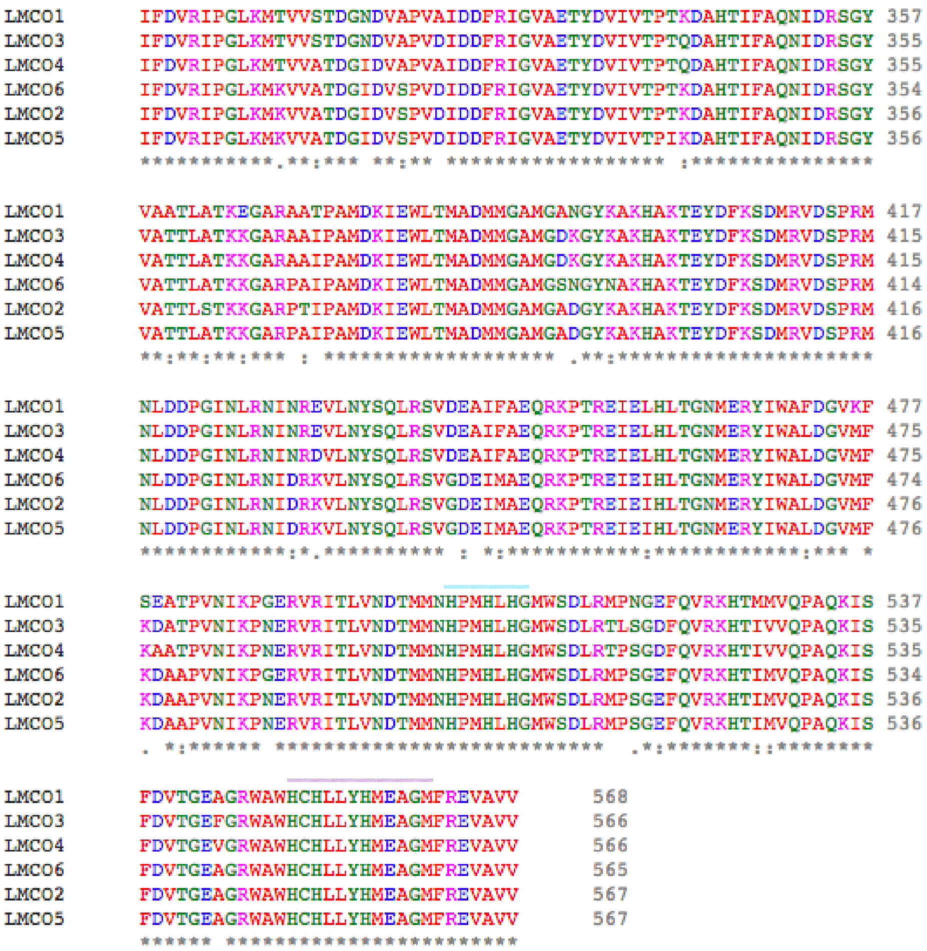
Another fascinating thing that can be observed are the areas that are marked by coloured lines. In part one of the figure, there are two markings; one in pink and one in green. The green has the exact same sequence as the signature sequence type 1, which also makes it partly L2, being that these two are partly similar. When comparing to L2, the beginning is conserved, starting with G-T-Y-W-Y-H-S-H-S-G-F-Q, with S-G-F being X_3. The rest of the sequences does not follow the pattern of L2.

The pink marking is an attempt to find L1, where the beginning is conserved, starting with H-W-H-G, not counting X_9, as these can vary. The following sequences does touch a few places that would fit into L1, such as -D-G-, although they're not in the correct position.

In part two of **figure 3.15**, there are also two markings; one turquoise and one purple. The turquoise is an attempt to find L3, which actually has a good match apart from the last residue, which is an M in stead of an H. The purple is the exact sequence as the one of type 2, which also makes it a bit similar to L4. There are residues that fits into L4, although not really enough to make it noteworthy.



```
CLUSTAL O(1.2.1) multiple sequence alignment


LMCO1           MNKKNILNRRRFLTGSSAVLGASLMPTI---ASSALGQSSRSGSQGATINSDQNVHKVPV 57
LMCO3           -MNKNMFNRRRFLTGSSTLLGASMLSTLPSIANSALG----KGQQNIAVNSDKADHIVPI 55
LMCO4           -MNKTMLNRRRFLTGSSTLLGASMLSTLPSIANSALG----KGQQNVAVNSDKSDHIVPI 55
LMCO6           -MNKNRFNRRRFLTGSSTLLGASMLSTLPTMANSALT----KS-QNVVVNSDRADHIVPI 54
LMCO2           MNNKNMLNRRRFLTGSSTLLGASMLSTLPTIANSALG----ANQKNAVINSDKPEHKVPI 56
LMCO5           MNNKNMLNRRRFLTGSSTLLGASMLSTLPTIANSALG----ANQKNAVINSDKPEHKVPI 56
                  :*. :**********::****:: *:   *.***        :  .:***:  * **:

LMCO1           LTGKEFDLYVSKQSAIVNGKKSMATLINDSLPAPTLKMREGDTVVIRVHNQMDESTSIHW 117
LMCO3           LTGTEFDLYVSKQSAIVNGKKSMATLINDSLPAPTLKMREGDTVVIRVHNQMDESTSIHW 115
LMCO4           LTGNEFDLYVSKKPVTVNGKSSMATLINDSLPAPTLKMREGDTVTIRVHNQMNESTSIHW 115
LMCO6           LTGKEFDLYVSEKMITVNGKSSMATLINDSLPAPTLKMQEGDTVTIRVHNQLNESTSIHW 114
LMCO2           LTGKEFDLYVSKKPIIVNGKSSTATLINDSLPAPTLKMREGDTVVIRVHNQMNESTSIHW 116
LMCO5           LTGKEFDLYVSKKPIIVNGKSSTATLINDSLPAPTLKMREGDTVVIRVHNQMNESTSIHW 116
                ***.*******::   ****.* ***************:*****.******::*******

LMCO1           HGLLVPFEMDGVPGISFDGIPANSTFTYKFKLKQSGTYWYHSHSGFQEQTGMLGAIVIEP 177
LMCO3           HGLLVPFEMDGVPGISFDGIPANSTFTYTFKLVQSGTYWYHSHSGFQEQTGMLGAIVIEP 175
LMCO4           HGLLVPFEMDGVPGISFDGIPANSTFTYKFKLKQSGTYWYHSHSGFQEQTGMLGAIVIEP 175
LMCO6           HGLLVPFEMDGVPGISFDGIPAGSTFTYKFKLIQSGTYWYHSHTGFQEQTGMRGAIVIEP 174
LMCO2           HGLLVPFEMDGVPGISFDGIPANSTFTYKFTLKQSGTYWYHSHTGFQEQTGMRGAIVIEP 176
LMCO5           HGLLVPFEMDGVPGISFDGIPANSTFTYKFTLKQSGTYWYHSHTGFQEQTGMRGAIVIEP 176
                ********************** *****.*.* **********:******** *******

LMCO1           KGRERHPIDEDHVIVLSDWTSRNPHNLLKLLKQRADFDNYHLPDFKKLLADIAETDMKTA 237
LMCO3           KGRERHPIEEDHVIVLSDWTHRDPHNLLKLLKQRADFDNYHLPDFKKLLADIAATNLEAA 235
LMCO4           KGRERHPIEEDHVIVLSDWTHRDPHNLLKLLKQRADFDNYHLPDFKKLLSDIAATDLEAA 235
LMCO6           KGRERYPIEEDHVILLSDWTHRDPHNLLKLLKQRADFDNYHLPDFKKLLSDIAATDLEAA 234
LMCO2           KGRERHPIEEDHVILLSDWTHRDPHNLLKLLKQRADFDNYHLPDFKKLLSDIAATDLEAA 236
LMCO5           KGRERHPIEEDHVILLSDWTHRDPHNLLKLLKQRADFDNYHLPDFKKLLSDIAATDLEAA 236
                *****:**:*****:***** *:**************************:*** *::::*

LMCO1           FDKRKMWNQMRMMPTDFTDLSGENTFTYLINGKTTAANWAQIVKAGQRVKLRFINASAQT 297
LMCO3           HDKRKMWNQMRMMPTDFTDLSGEKTFTYLMNGKTTAANWTQLVKAGQPVKLRFINGSAQT 295
LMCO4           YDKRKMWNQMRMMPTDFTDLSGEKTFTYLMNGKTTAANWTQLVKAGQPVKLRFINASAQT 295
LMCO6           FDKRKTWNQMRMMPTDFTDLSGETTFTYLMNGKTTAANWTQLVKAGQPVKLRFINGSAQT 294
LMCO2           FDKRKMWNQMRMMPTDFTDLSGEKTFTYLMNGKTTAANWTQLVKAGQPVKLRFINGSAQT 296
LMCO5           FDKRKMWNQMRMMPTDFTDLSGEKTFTYLMNGKTTAANWTQLVKAGQPVKLRFINGSAQT 296
                .**** *****************.*****:*********:*:***** *******.****
```



```
LMCO1      IFDVRIPGLKMTVVSTDGNDVAPVAIDDFRIGVAETYDVIVTPTKDAHTIFAQNIDRSGY 357
LMCO3      IFDVRIPGLKMTVVSTDGNDVAPVDIDDFRIGVAETYDVIVTPTQDAHTIFAQNIDRSGY 355
LMCO4      IFDVRIPGLKMTVVATDGIDVAPVAIDDFRIGVAETYDVIVTPTQDAHTIFAQNIDRSGY 355
LMCO6      IFDVRIPGLKMKVVATDGIDVSPVDIDDFRIGVAETYDVIVTPTKDAHTIFAQNIDRSGY 354
LMCO2      IFDVRIPGLKMKVVATDGIDVSPVDIDDFRIGVAETYDVIVTPTKDAHTIFAQNIDRSGY 356
LMCO5      IFDVRIPGLKMKVVATDGIDVSPVDIDDFRIGVAETYDVIVTPIKDAHTIFAQNIDRSGY 356
           ***********.**:*** **:** ****************** :***************

LMCO1      VAATLATKEGARAATPAMDKIEWLTMADMMGAMGANGYKAKHAKTEYDFKSDMRVDSPRM 417
LMCO3      VATTLATKKGARAAIPAMDKIEWLTMADMMGAMGDKGYKAKHAKTEYDFKSDMRVDSPRM 415
LMCO4      VATTLATKKGARAAIPAMDKIEWLTMADMMGAMGDKGYKAKHAKTEYDFKSDMRVDSPRM 415
LMCO6      VATTLATKKGARPAIPAMDKIEWLTMADMMGAMGSNGYNAKHAKTEYDFKSDMRVDSPRM 414
LMCO2      VATTLSTKKGARPTIPAMDKIEWLTMADMMGAMGADGYKAKHAKTEYDFKSDMRVDSPRM 416
LMCO5      VATTLATKKGARPAIPAMDKIEWLTMADMMGAMGADGYKAKHAKTEYDFKSDMRVDSPRM 416
           **:**:**:*** : ******************* .**:*********************

LMCO1      NLDDPGINLRNINREVLNYSQLRSVDEAIFAEQRKPTREIELHLTGNMERYIWAFDGVKF 477
LMCO3      NLDDPGINLRNINREVLNYSQLRSVDEAIFAEQRKPTREIELHLTGNMERYIWALDGVMF 475
LMCO4      NLDDPGINLRNINRDVLNYSQLRSVDEAIFAEQRKPTREIELHLTGNMERYIWALDGVMF 475
LMCO6      NLDDPGINLRNIDRKVLNYSQLRSVGDEIMAEQRKPTREIEIHLTGNMERYIWALDGVMF 474
LMCO2      NLDDPGINLRNIDRKVLNYSQLRSVGDEIMAEQRKPTREIEIHLTGNMERYIWALDGVMF 476
LMCO5      NLDDPGINLRNIDRKVLNYSQLRSVGDEIMAEQRKPTREIEIHLTGNMERYIWALDGVMF 476
           ************:*.********** : *:***********:************:*** *

LMCO1      SEATPVNIKPGERVRITLVNDTMMNHPMHLHGMWSDLRMPNGEFQVRKHTMMVQPAQKIS 537
LMCO3      KDATPVNIKPNERVRITLVNDTMMNHPMHLHGMWSDLRTLSGDFQVRKHTIVVQPAQKIS 535
LMCO4      KAATPVNIKPNERVRITLVNDTMMNHPMHLHGMWSDLRTPSGDFQVRKHTIVVQPAQKIS 535
LMCO6      KDAAPVNIKPGERVRITLVNDTMMNHPMHLHGMWSDLRMPSGEFQVRKHTIMVQPAQKIS 534
LMCO2      KDAAPVNIKPNERVRITLVNDTMMNHPMHLHGMWSDLRMPSGEFQVRKHTIMVQPAQKIS 536
LMCO5      KDAAPVNIKPNERVRITLVNDTMMNHPMHLHGMWSDLRMPSGEFQVRKHTIMVQPAQKIS 536
           . *:****** ***************************  .*:*******::********

LMCO1      FDVTGEAGRWAWHCHLLYHMEAGMFREVAVV      568
LMCO3      FDVTGEFGRWAWHCHLLYHMEAGMFREVAVV      566
LMCO4      FDVTGEVGRWAWHCHLLYHMEAGMFREVAVV      566
LMCO6      FDVTGEAGRWAWHCHLLYHMEAGMFREVAVV      565
LMCO2      FDVTGEAGRWAWHCHLLYHMEAGMFREVAVV      567
LMCO5      FDVTGEAGRWAWHCHLLYHMEAGMFREVAVV      567
           ****** ************************
```

Figure 3.15: The full alignment of all six LMCO's, including the marked sequences, which refers to signature sequences or attempts to find signature sequences. The pink line is an attempt of L1, while the turquoise is an attempt of finding L3. The green one is the sequence of signature sequence type 1, while the purple is type 2.

Just like ClustalW, Clustal Omega creates a phylogenetic tree when it creates the alignment, as shown in **figure 3.16**. The indications made in **section 3.3.4** on relationships were then put on a test. As the phylogenetic tree shows, LMCO1 is furthest apart from the others, and has LMCO3 as its closest neighbour, although it's far away. LMCO3 and LMCO4 are closely related, as are LMCO2 and LMCO5.

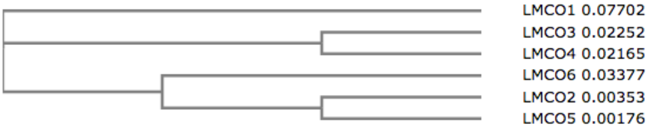


Figure 3.16: A phylogenetic tree based on the alignment of the full sequences of the six LMCO's

Looking at the phylogenetic tree in **figure 3.16**, the predictions from comparing the 3D models were both right and wrong. LMCO3 and LMCO4 are more closely related than expected, especially when LMCO3 seemed more similar to LMCO1 than LMCO4, when they are actually further apart according to the phylogenetic tree. LMCO1 were predicted to be closer to LMCO3 than LMCO2, which works well with the findings in the phylogenetic tree, and LMCO5 and LMCO6 were further apart than predicted.

3.4 Transcriptome, proteome and promoters

In this part of the thesis, the focus of the study was narrowed down to only using P11F6 for transcriptomics and proteomics. As both transcriptomics and proteomics were done by others, these results were just delivered, analysed, and compared. Further work was based on these results, such as the promoters chosen were based on the top ten list of up-regulated genes.

3.4.1 Transcriptomics and proteomics

The top ten upregulated genes were chosen from the transcriptomics results list. These all had M-values above 4, which was chosen as the limit in the first place. These ten can be found in **table 3.11**. None of the top ten upregulated genes turned out to be a LMCO, which came as a surprise, as 2-methoxy-phenols are a substrate of LMCO's. It did however show several phenol hydroxylases, which could be capable of doing the same oxidation as the LMCO's. When looking at the start/stop positions for the phenol hydroxylases, these are found in the same area of the genome.

Table 3.11: When adding 2-methoxy-phenol, the gene expression changed. The table shows the top ten upregulated genes, where to find them, their length (bp), which strand they were found on and their m-value. Data provided by Christian Rückert

Protein	Start	Stop	Length (bp)	Strand	M-value
TonB-dependent receptor	2 042 339	2 044 519	2181	Fwd	9.28
Phenol hydroxylase, assembly protein DmpK	2 022 932	2 022 618	315	Rev	8.08
Phenol hydroxylase P1 protein	2 022 536	2 021 526	1011	Rev	7.83
Hemin-degrading family protein	2 041 858	2 040 725	1134	Rev	7.55
DoxX protein	2 520 602	2 520 261	342	Rev	7.43
Phenol hydroxylase, P2 protein	2 021 472	2 021 206	267	Rev	7.36
Phenol hydroxylase P3 protein	2 021 133	2 019 616	1518	Rev	7.26
Hypothetical membrane protein	974 142	973 846	297	Rev	7.18
Conserved hypothetical secreted protein	2 044 622	2 045 887	1266	Fwd	6.96
Phenol hydroxylase P4 protein	2 019 469	2 019 107	363	Rev	6.84

To see if the top ten upregulated genes were translated, the proteomics data for these ten genes were searched for. This lead to the results shown in **table 3.12**. One of the most surprising findings here are that there are only eight proteins here. This is because the phenol hydroxylase assembly protein and the hypothetical membrane protein in **table 3.11** were not found in the proteomics study. **Table 3.12** shows, in a 2-log manner, how much these proteins were up-regulated compared to growing without substrate. Measurements were made for both pellet and supernatant, and the table also include RPKM values, which is a measure on “reads per kilobase per million”. This shows that, apart from the two which could not be found, the transcripts found in **table 3.11** were actually transcribed into proteins.

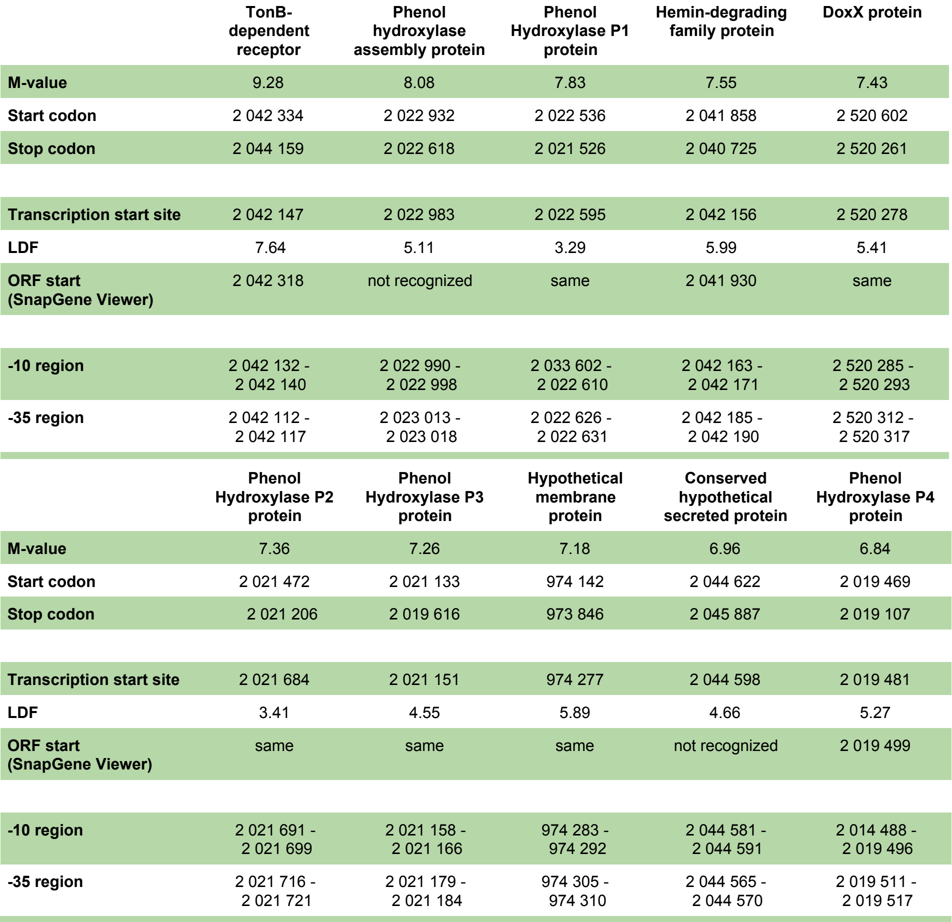
Table 3.12: The top ten list of genes were searched for in the proteome, findings these eight proteins. The proteome here was from P11F6 growing on media containing 2-methoxy-phenol, shown in a 2-log manner compared to P11F6 growing without substrate. Measures were made for both pellet and supernatant, as well as calculating the RPKM value. Data provided by Animesh Sharma

Protein	Pellet	Supernatant	RPKM
TonB-dependent receptor	22.64	0	9.92
Phenol hydroxylase P1 protein	7.56	4.92	7.89
Hemin-degrading family protein	6.15	0.81	7.67
DoxX protein	0	0	7.67
Phenol hydroxylase, P2 protein	4.44	2.95	7.43
Phenol hydroxylase P3 protein	8.87	6.48	7.32
Conserved hypothetical secreted protein	3.86	1.38	7.08
Phenol hydroxylase P4 protein	6.71	31.37	6.91

3.4.2 Promoters

Using Softberry’s BPROM and manual searching, the top 10 upregulated genes’ promoters were identified. **Figure 1.2** shows an approximate view of a promoter, including the regions that were searched for. These regions included -10 and -35 sequences, as well as first bases in start/stop codons and RBS’s. **Table 3.13** shows where the -10 and -35 regions were found for each of the genes. The table also includes m-values (a value on upregulation), transcription starting site, LDF value and ORF start. This table is divided into two parts, each with five promoters. All the promoters are sorted on m-value, which indicates how much the gene has been upregulated when compared to growing the P11F6 strain on media not containing 2-methoxy-phenol. The proteins are sorted by m-values from highest at the top left to lowest at the bottom right. All the m-values came from the transcriptome analysis, which also included the start/stop codon. Transcription start site was found by BPROM, along with LDF values, -10 region and -35 region. The LDF values give a value of how good the prediction is, as seen in **section 2.4.2**. As there are no upper or lower limit to these, they are hard to interpret. Having positive values are still considered as a sign of actual promoters though. As the genomes were also viewed in SnapGene viewer, this row tells if the genes in question were found by SnapGene Viewer, and if they were, if they had the same ORF start.

Table 3.13: The promoter regions for the top ten upregulated genes when growing the P11F6 strain on media containing 2-methoxy-phenols. The promoter sequences are defined on forward strand, and the table includes start/stop, transcription start site, LDF values, -10 sequence and -35 sequence. The row of "ORF" start indicates if SnapGene Viewer gave a different transcription start side than the one BPROM found



	TonB-dependent receptor	Phenol hydroxylase assembly protein	Phenol Hydroxylase P1 protein	Hemin-degrading family protein	DoxX protein
M-value	9.28	8.08	7.83	7.55	7.43
Start codon	2 042 334	2 022 932	2 022 536	2 041 858	2 520 602
Stop codon	2 044 159	2 022 618	2 021 526	2 040 725	2 520 261
Transcription start site	2 042 147	2 022 983	2 022 595	2 042 156	2 520 278
LDF	7.64	5.11	3.29	5.99	5.41
ORF start (SnapGene Viewer)	2 042 318	not recognized	same	2 041 930	same
-10 region	2 042 132 - 2 042 140	2 022 990 - 2 022 998	2 033 602 - 2 022 610	2 042 163 - 2 042 171	2 520 285 - 2 520 293
-35 region	2 042 112 - 2 042 117	2 023 013 - 2 023 018	2 022 626 - 2 022 631	2 042 185 - 2 042 190	2 520 312 - 2 520 317

	Phenol Hydroxylase P2 protein	Phenol Hydroxylase P3 protein	Hypothetical membrane protein	Conserved hypothetical secreted protein	Phenol Hydroxylase P4 protein
M-value	7.36	7.26	7.18	6.96	6.84
Start codon	2 021 472	2 021 133	974 142	2 044 622	2 019 469
Stop codon	2 021 206	2 019 616	973 846	2 045 887	2 019 107
Transcription start site	2 021 684	2 021 151	974 277	2 044 598	2 019 481
LDF	3.41	4.55	5.89	4.66	5.27
ORF start (SnapGene Viewer)	same	same	same	not recognized	2 019 499
-10 region	2 021 691 - 2 021 699	2 021 158 - 2 021 166	974 283 - 974 292	2 044 581 - 2 044 591	2 014 488 - 2 019 496
-35 region	2 021 716 - 2 021 721	2 021 179 - 2 021 184	974 305 - 974 310	2 044 565 - 2 044 570	2 019 511 - 2 019 517

The RBS's were found manually, searching the entire isolated sequence for AGGAGG, AGGAGN and AGCA. No cases of AGGAGG were found, but the cases of AGGAGN and AGCA that seemed plausible are listed in **table 3.14**. Some of the promoters were found to have more than one possible RBS, and all of these are included in the table. Four of the promoters were found to have only one possible RBS. Even though all of these sites were found, and in some cases the only ones found, they are not all likely to be actual RBS's. The ones found for DoxX, hypothetical membrane protein and the conserved hypothetical secreted protein are probably not real, even if they were the only possibilities found for these sequences. As these were the only ones that were found, they were still included.

Table 3.14: The found RBS's for the top ten upregulated genes when growing the P11F6 strain on media containing 2-methoxy-phenols. Two types of RBS were found; AGCA and AGGAGN. Some of the promoters contained only one RBS, while others contained two or three.

	AGGAN	**AGCA**		
TonB-dependent receptor		2 042 269 - 2 042 272	2 042 285 - 2 042 288	
Phenol hydroxylase assembly protein	2 023 240 - 2 023 245			
Phenol Hydroxylase P1 protein	2 022 546 - 2 022 550	2 022 538 - 2 022 541		
Hemin-degrading family protein		2 041 988 - 2 041 991		
DoxX protein		2 520 322 - 2 520 325		
Phenol Hydroxylase P2 protein		2 021 511 - 2 021 514	2 021 586 - 2 021 598	2 021 644 - 2 021 647
Phenol Hydroxylase P3 protein	2 021 142 - 2 021 147	2 021 154 - 2 021 157		
Hypothetical membrane protein	974 548 - 974 552	974 483 - 974 486		
Conserved hypothetical secreted protein		2 044 554 - 2 044 557		
Phenol Hydroxylase P4 protein		2 019 507 - 2 019 510	2 019 493 - 2 019 496	

Chapter 4

Discussion

In this chapter, all of the results from **chapter 3** will be discussed, along with some of the decisions made in **chapter 2** when doing the research. Following the order created in the other sections, we'll start with RAST and Mauve, then the LMCO's and end up with the transcriptomics/proteomics and the promoters.

4.1 RAST

RAST was chosen for automatic annotation based on two reasons: its user friendly interface and the high quality of the results. When doing the automatic annotation, RAST is a very intuitive program. It does not require a lot of settings, and running it in default mode will usually give a good result. Some choices were made though, such as the settings in **figure 2.1** and **figure 2.2**. There were not any real choices to make in **figure 2.1**, as these do not really have any influence on the annotation. Using another genetic code might be possible, although using anything but 11, which represents bacteria, makes no sense.

Choosing the settings in **figure 2.2** were based on the explanations of each setting given in the figure, as they made the most sense for the annotation at the moment. This included wanting the current version of RAST instead of the one in testing, using RAST over GLIMMER-3 to avoid disabling of many settings, and using the latest version of FIGfam. Automatically fixing errors were left "off" to avoid deleted gene candidates, while fixing frameshifts, building a metabolic model, BLASTing large gaps, printing debug statements and running each job from scratch were all wanted features. Leaving the verbose level to default 0 was made to get any error message that might come.

As seen in **figure 3.1** and **table 3.1**, the subsystem coverage was around 50 % for all four genomes. This should ideally have been closer to at least 90 %, to ensure that all genes were found and put in a subsystem group. However, these values do not mean that the annotation only found half of the genes, but simply that only about half of the found ones were possible to put in a subsystem group. These numbers were considered good enough for further use, and all the *LMCO* in the next section were found by both RAST and other annotation programs used.

In addition to the figure, **table 3.1** shows more information given by RAST. The GC-content is similar in all four genomes, and the variations in both number of coding sequences, number of RNA's and number of subsystems are similar, as expected from four closely related genomes of similar size. The genome sizes of 3.2 - 3.4 million bp shows that these strains are more similar to *P. aquaticus* than to *P. arcticus* when it comes to size (Reddy et al., 2013; Ayala-del Río et al., 2010). GC content being 41.9 - 42.9 % fits with the results found by Bozal et al. (2003)

As shown in **table 3.2**, the total number of genes here does not match the number of coding sequences found in **table 3.1**, but is actually a lot smaller. This is most likely due to the fact that not all coding sequences/proteins could be placed in a subsystem. From the table, it's possible to see that even though most of the columns have similar values, within a reasonable range, some of the values differ a lot, such as in "Phages, prophages, transposable elements and plasmids" or "iron aquisition and metabolism". The reasons for these variations are not known, but these differences are possible to study further, as will be discussed further in **chapter 5**.

4.2 Mauve

Mauve gave a nice view of the evolutionary changes in **figure 3.2**, and showed how some blocks of genes had changed places, while others were conserved. Mauve is clearly a program that makes more sense when used directly rather than showing pictures of it, and showing pictures does not serve the program the respect it deserves. The close up examples in **figure 3.3** and **figure 3.4** show only parts of the possibilities of tracking changes in Mauve, even though the real tracking options with zooming and following lines does not show very clearly in still pictures. The point of these figures were to show how using the features in the program could reveal parts that were not visible in the first place. The best example of this was "the hidden, blue box", which was not visible in **figure 3.3**, apart from a line, seemingly appearing from nowhere. When closing in on the sequence in **figure 3.4**, the tiny blue line turned out to belong to a blue box, which appeared in the two top genomes but not in the two lower ones, at least not in that particular position. As the blue line headed towards the right end of the picture and disappeared, it seemed plausible to believe that there were other blue boxes in the two bottom genomes, although this was not investigated further.

What still pictures of Mauve did show was that the relationship between the four genomes of *Psychrobacter* was as clear as one would expect. The beginning and the end of the genomes contained the same coloured boxes in mostly the same pattern, as well as in the middle section, even if the boxes were a bit rearranged, they could be found in the other genomes by following the lines. As seen in the annotation, these genomes were expected to be quite similar, and this was confirmed by Mauve.

The findings in **figure 3.5** also proved that the genomes were similar, although somewhat different. Seeing how two sequences of $\sim$ 20 000 bp were so similar, but in different positions, was quite interesting. The sequences were perfectly slided about 15 000 bp downstream in P2G3 when compared to P11G5, showing that the sequences were in fact very similar, even if they were not identical.

When comparing the Mauve results to the phylogenetic tree in **figure 3.6**, P2G3 and P11G5 were found to be the closest related, which might explain why these two showed the features in **figure 3.5**, while the remaining genomes did not.

4.3 Laccase-like Multicopper Oxidases

The first interesting thing about the *LMCO's*, was the fact that the number and position of them varied between the four *Psychrobacter* genomes. P2G3 had only one, P11G3 and P11G5 had two copies, and P11F6 was the real outsider, having only one copy, and having it in a plasmid! P11G3 and P11G5 both had one copy on each strand; one on the forward and one on the reverse. All of these sequence details were summed up in **table 3.3**, along with the start/stop positions. This table also showed that the length of the proteins varied from 565 to 568 aa, which showed both high similarity between the LMCO's from the *Psychrobacter* spp. and also fits with the findings by (Sharma et al., 2007).

4.3.1 Amino acids and half-life

In these first analyses, ExPASy's ProtParam and pI/Mw computing tool were chosen for the task. ExPASy is a known resource portal for bioinformatical tools, both external ones and their own, such as ProtParam. ProtParam calculates a various number of physical and chemical properties, and their pI/Mw computing tool is capable of calculating these values.

Starting with the basic compositions of the proteins, **table 3.4** showed that the composition of amino acids was quite similar in all of the LMCO's. The two that stands out were tryptophan and cytosine, for being the only two that were exactly the same in all six LMCO's. The cytosine residue was also the same one that was found in the T1 copper site. The number of Cys residues seem to vary, and having only one Cys residue means no formation of disulfide bridges. Studies on CotA from *Bacillus subtilis*, *B. pumilus* and *B. licheniformis* showed two additional Cys residues creating a disulfide bridge. Studies on *B. clausii* had only one additional Cys residue, while *B. coagulans* contained only the one Cys residue from the Cu binding site, and neither of these would create any disulfide bridges (Enguita et al., 2003; Ihssen et al., 2015).

A part from the Trp and Cys content, all the other aa residues were different in all the LMCO's. Alanine was the one with the most different contents, varying from 39 to 46. Apart from this specific aa, the variation was generally lower between the genomes.

Further investigations on the basic compositions lead to the results shown in **table 3.5**, which showed that LMCO2 and LMCO5 had the highest number of atoms with 8980, while the remaining had 8925 - 8926. Naturally, this was reflected on these two having the highest molecular weight, and also the exact same pI. Even if the amount of each and every aa was not the same in these two, the total number of positively and negatively charged residues was exactly the same in both proteins, which naturally effected the pI. Regarding pI and positive/negative aa's, it seemed that the LMCO's with a higher level of negatively charged residues against positively charged residues had a lower pI. LMCO3 had 66 negative residues, against 63 positive. This LMCO had the lowest pI, with 6.75. LMCO1 had an almost identical pI as LMCO2 and LMCO5; 8.59 vs. 8.56. When looking at their composition of aa's, LMCO1 had 64 negative and 67 positive, while LMCO2 and LMCO5 both have 66 negative and 69 positive. Even if the numbers were different, the ratio between negative and positive was still the same. LMCO4 and LMCO6 both had 65 positive residues, but where LMCO4 had 64 negative residues, and hence was more positively charged, LMCO6 had 66 negatively charged residues, and hence was a bit more negative. This was reflected in their pI; LMCO4's of 7.85 vs. LMCO6's of 7.07. Comparing to the pI's found by (Morozova et al., 2007), which ranged from 2.6 - 9.5, the estimated pI's found for the LMCO's seems to fit nicely, even if the comparing pI's were found for laccases in plant and fungi, and not bacterial.

The results found in **table 3.5** also showed the calculated Mw for all of the LMCO's, which varied from 63.7 to 64.1 KDa. These findings fits the Mw found by (Morozova et al., 2007; Ihssen et al., 2015), which showed how laccases from various organisms had a Mw of 32-130 KDa. The various bacterial laccases studied by (Ihssen et al., 2015) showed a smaller range of 32.6-59.7 KDa, which shows that the LMCO's from these four strains of *Psychrobacter* were slightly bigger.

Regarding half-life and instability, the values on half-life in hours were computed for mammalian, *E.coli* and yeast. The *E.coli* might be the best to compare with, being that they're both bacteria, at least. The half-life calculation is based on "N-terminal rule", meaning that the first residue at the N-terminal determines the half-life. As known from the multiple alignment of the signal sequences, which consists of roughly the first 30-ish residues of the N-terminal, the sequences were quite similar and all started with a Met residue. This indicated that the expected half-life, as long as it was based on the N-terminal, was expected to be equally similar for all sequences. As ">10", which was the expected half-life in *E.coli*, was not a very easy number to compare to, it had to be assumed that the value was quite similar to all of the LMCO's, having similar sequences and all starting on the same residue. The half-life being measured based on sequences did not take different temperatures or other conditions into account, only making a prediction based on the N-terminal sequences. Reiss et al. did a study on half-life in *B. pumilus*, showing half-life above 10 hours being possible in conditions below 45 °C in deionized water. At other conditions, such as proteins being kept in McIllvain buffer or potassium phosphate buffer both at pH 7 at a range of temperatures (4-65 °C), the half-life was one hour or lower (Reiss et al., 2011). This indicates that the half-life estimation based on N-terminal rule alone might not be the most reliable, although can work as a starting point. Another point in this case is the fact that all proteins have an initiator Met residue in the mRNA, which is removed after translation (Xiao et al., 2010). If the Met residue is removed straight after translation, the new N-terminal residue would be Asn, as known from **figure 3.13**. If *E.coli* is the organism used for comparison, the estimated half-life would still be ">10 h", although for mammalian it's down to 1.4 h, and 3 min for yeast. As it still makes most sense to use *E.coli*'s value, this does probably not change anything.

The instability index is based on dipeptides of known stability, and a value below 40 means the protein is considered stable. As **table 3.6** shows, all six LMCO's had instability values below 40, and were therefore considered stable. As this was also a prediction based on the sequence, it did not consider temperature or any other conditions. Naturally, this would most likely vary a lot if these tests were performed in a test tube or in vivo. Ihssen et.al tested stability of LMCO's from a number of *Bacillus sp.* under several conditions, showing an overall stability. Even if the stability in this thesis is estimated based on dipeptides, this still might be comparable, as both studies concluded with an overall stability (Ihssen et al., 2015).

4.3.2 Copper sites

Phyre2 had two functions in this thesis; predicting pockets and Cu binding sites, and creating output files for PyMOL. Having a program who did both of these tasks was a lucky coincidence, Phyre2 would have been used for prediction of pockets/Cu binding sites anyway. Being able to use the output files in PyMOL just made it possible to skip a step.

The LMCO's should contain a T1 copper site and a trinuclear copper site, as known form **section 1.2**. These were possible to detect with Phyre2, which showed that these T1 sites were made out of a repeating pattern of His - Cys - His - Met and two HXH, as shown in **table 3.7** and **figure 3.7**. The problem with the predictions made by Phyre2, was that there was only two HXH's, when it was supposed to be four, according to literature. As seen in **figure 1.1**, the HXH's were involved in different Cu binding sites, such as His493 belonging to Cu2, while His491 belonged to Cu3. The numbers on the residues were not directly applicable to the Cu sites in the LMCO's here, as they would vary at least slightly between different LMCO's, but the positions of HXH in regards to histidines were still the same. This meant that the two HXH Phyre2 could find were just parts of the trinuclear site, technically half of it.

According to (Reiss et al., 2013), the four pairs of his residues should be found in four specific patterns, and the ones that were found by Phyre2 corresponded to HXXHXH and HCHXXXHXXXXM/L/F. As shown in **section 1.2**, there were certain conserved sequences, which could be found in members of the multicopper oxidase family and in the laccases in specific. The Cu binding sites were supposed to be found within these sequences, which the patterns were part of, and while doing manual searches in **figure 3.15**, these were used as a confirmation that the found alternatives for HXH were actual candidates. Only finding two candidates made the decision far more easy than expected, although being able to confirm that the found candidates actually did exist within the signature sequences did strengthen the theory on these being the missing two HXH's. The candidates, here found for LMCO1, were positioned at His116 (HWH) and His157 (HSH), which were parts of L1 and L2, respectfully. The numbers on the His residues must be adjusted naturally, but since it was found inside a conserved sequence, they could all be found in **figure 3.15**.

With the new found HXH's, this made the whole trinuclear site visible, as seen in the revised **table 3.8**. It is still not known why Phyre2 only found two of the four HXH's, although it seems reasonable to believe that the two that where found manually actually were the remaining two. This lead to having the complete T1 site, and a full trinuclear site of four HXH's, just as expected. The signature sequences will be further discussed in **section 4.3.5**. All in all, this means that the predictions based on reading (Reiss et al., 2013) turned out to be real. The findings here also fits findings by (Colman et al., 1978; Solomon et al., 1996; Enguita et al., 2003; Claus, 2004), among others.

4.3.3 Subcellular location

As known from (Diamantidis et al., 2000), laccases are mostly found to be intracellular in bacteria and extracellular in fungi and plants. According to **table 3.9**, all of the LMCO's were found to be intracellular and destined for the periplasm. LMCO3 and LMCO4 had minor scores on extracellular, which might be due to relationships to fungal or plant laccases. LMCO3 had a minor score on cytoplasmic membrane, while LMCO4 had an even smaller score for both outer membrane and cytoplasmic membrane. It was interesting to see how only LMCO3 and LMCO4 had scores on anything except periplasmic, while the four others had so clear results.

The findings might conclude with LMCO3 and LMCO4 being a bit different from the others, and it also might suggest that the LMCO's from P11G3 were closer related to fungi or plants than the other strains investigated here. It was however clear that all of the LMCO's from all species were destined for the periplasm, regardless of any other minor scores.

4.3.4 PyMOL

When time came for visualization, PyMOL was an easy choice, as it makes nice 3D models which can be turned and twisted for viewing from all angles, colours can be added to separate secondary structures or domains, and various types of models can be created. The only downside to using it in a thesis is the fact that this program works best live, not as still pictures. When making the models, "cartoon" mode was used on the proteins, in order to get the best views of the details in the secondary structures. Other modes were possible, such as "lines", which showed a stick model with all atoms, "sticks", which showed the same as "lines" only with thicker lines, or "ribbon", which showed the sequences as a ribbon without secondary structures. The two first were found to be useless for this use, while the third one was ok, but not detailed enough. Using the "cartoon" mode along with colouring based on secondary structures gave a detailed model which clearly showed the predicted structure of the proteins, and also made it possible to compare them in a sensible matter.

3D models and superpositioning

When choosing LMCO3 and LMCO4 to make the first models and comparisons, this was a rather arbitrary choice, only slightly based on the results from earlier analyses of the LMCO's. All research done up to that point showed that these two LMCO's had quite similar properties. Coming from the same strain, P11G3, and having shown similar properties, it was expected that these two would have similar structures. Finding that they in fact did not, was therefore quite surprising. They had similarities, of course, as seen with the β sheets and following the loops of the structure, although the largest α helices were not in the same place at all.

When it came to picking the two next LMCO's to model and compare, LMCO1 and LMCO2 were chosen based on being more different. As LMCO3 and LMCO4 did not have identical structures despite being so similar up till that point, the next two were chosen for being different, hoping this would make some interesting results. Combining the models showed that they did in fact seem a lot more similar than the first comparison, which was found interesting, as they were chosen based on the idea that they were more different. As with the first comparison, the β sheet area seemed to be similar, as well as parts of looping structure. Both LMCO1 and LMCO2 had the large α helix in (almost) the same position, although their length seemed to be different. In fact, the model of LMCO1 resembled the model of LMCO3, and like that, the third pair of models to compare was found.

Comparing LMCO1 and LMCO3 showed that they had their similarities and differences, just as expected. The large α helix, that so far had been the easiest difference to spot, were more similar here compared to the other cases. Their position and length were both close in resemblance, and the β sheets in the middle were also similar.

As there were only two LMCO's left, these were compared to each other. LMCO5 and LMCO6 both came from P11G5, and had so far not shown any specific resemblance. It was therefore another surprise when the models turned out to be quite similar, both having the large α helix pointing straight out to the side. The bottom part of the β sheet showed some interesting and very similar features, which had not been visible to that extent in the other models, despite rotating them. As these models were so similar to each other, and at the same time so different from the others, no further comparisons were done.

All in all, comparing the models and superpositioning the sequences showed that all ideas on which structures would be similar were wrong. LMCO3 and LMCO4, which had shown very similar properties in all analyses done so far, were not as similar as expected. LMCO1 and LMCO2, and LMCO5 and LMCO6, showing different properties, were much more similar than expected. In general, the β sheets showed the most resemblance, while the major α helix gave the best view of difference.

Pockets

In comparing the pockets in PyMOL, the ability to add different colours to different parts really came in handy. PyMOL has a whole bunch of settings regarding colouring the various selections. By colouring the pockets purple, they could really stand out from the secondary structure colouring of red, green and yellow. One of the challenges here was to get the models as similar as possible, which as can be seen for LMCO6, is not always possible. Finding the exact same size/zoom for this protein turned out to be really hard, and ended up being impossible. This meant that the pocket looked a lot smaller than the others, when it in fact was not. One of the things that the pocket analysis showed, was that the pockets had different orientations, giving some of them a more compact finish than the others. When looking at the pockets, the angle of the model and the zoom were all involved in how the pocket looked. The proteins were twisted around to find the individual view that showed the most of the pocket. It would have been possible to arrange all the proteins in the same view, although this would lead to a lot of the pockets not being visible, and this was considered a bad way of showing off the pockets.

LMCO1 had one of the "loosest" pockets, meaning the residues were distributed in an order that created more air between the spheres. The size seemed to be fairly average, and all the spheres seemed connected.

LMCO2 was the only pocket with part of the pocket being separate from the rest of the spheres. When using sphere mode, all the other spheres touched at least one other sphere, and hence were connected to the others. In LMCO2 however, one of the residues was all alone. Keeping in mind that each sphere did not indicate one residue, the spheres that were separate from the others in LMCO2 were in fact just one residue by itself. Despite twisting and turning, this was not shown in any of the other pockets.

LMCO3 seemed to have one of the smallest pockets, but apart from that it did not stand out. LMCO4 seemed to have the biggest pocket, having a partly "loose" distribution, but no separate residues. LMCO5 seemed to have one of the smallest pockets, and having a "loose" distribution. LMCO6 had the smallest pocket when compared directly. For unknown reasons, this model refused to be in the same scale as the rest, and hence the pocket seemed smaller.

From only looking at the models in **figure 3.12**, all the assumptions on the size and structure of the pockets were highly objective. Looking at **table 3.10** however, there were actual numbers to consider. This first line of this table showed that LMCO4 actually was the biggest pocket, having a total amount of 83 residues, followed by LMCO6 with 72, which was surprising, as this seemed to be the smallest pocket. Again, this could be due to differences in zoom. LMCO2 followed with 57 residues, then LMCO3 with 48, LMCO1 with 36 and finally LMCO5 with 32. LMCO5 was believed to be the smallest pocket, although it seemed like the "loose" pocket of LMCO1 was actually smaller than LMCO3, which was surprising. Again, trying to decide the number of residues in each pocket based on the 3D models were not the most reliable assumption.

The rest of **table 3.10** showed the areas of the pocket which were thought to be conserved. These areas, all separated and marked in green fields, can be anything from a single residue to a whole stretch of sequence, being fully conserved in all of the LMCO's or just a few. The limit for selecting these were set to three, meaning that at least three of the LMCO's had to show conserved residues for the region to count. Green field number one clearly showed how a Glu residue was conserved around position 95. Despite this being interesting, field number two was even more interesting, as both LMCO3, LMCO4 and LMCO6 had a conserved stretch of at least six residues, which included one of the HXH's. LMCO1, LMCO2 and LMCO5 showed only partly conservation in this area, which was probably a sign of the pocket detector missing out on some of the residues belonging to the pocket. The HXH in this case belonged to the trinuclear site, according to **table 3.8**, which as known from **section 1.2** is involved in reducing molecular oxygen and releasing water. It was therefore suspected that these residues would be involved in the pocket. It is possible that the Leu, which was conserved in all of the pockets, was enough to keep the substrate close to the Cu sites. This was also the case for the ninth green field, where four of the LMCO's had a conserved Leu in the area that holds an HXH, even if only three of the pockets actually included the HXH. Having an almost conserved Leu here as well might have been an indication on Leu residues being involved in keeping the substrate close to the trinuclear sites, although these are simply suggestions for further research.

Green field number five was one of the few areas that were close to fully conserved in all the pockets. All LMCO's had Gly-Asn/Ile-Asp around position 313, which also was the only group of pocket residues in this area at all. As seen in **Appendix A**, these three residues were the only ones between roughly positions 180 and 413.

It is not known if the lack of residues in LMCO1 and LMCO5, and the general diversity among the pockets were caused by actual smaller pockets or due to a bug in the pocket detector. In order to see if Phyre2 and Fpocket were reliable, sequences of studied multicopper oxidases from various fungi and one plant were run through Phyre2 and the pocket detector (work not shown). These sequences were found using the protein data bank (PDB) codes in (Larrondo et al., 2003), and laccases from *Coprinus cinereus* (PDB code 1A65), *Melanocarpus albomyces* (PDB 1GW0) and *Trametes versicolor* laccase 1 (PDB 1GYC), and ascorbate oxidase from *Curcubita pepo* (PDB 1AOZ). As Larrondo et al. had found the residues involved with substrate binding sites, the results from the article could be compared to the results from Phyre2. Two of the comparisons here, *M. albomyces* and *C. pepo*, showed similarity, and the variations here were found to be natural. In the two other cases, *C. cinereus* and *T. versicolor*, the results did not have any matching residues. In all four cases, T1 copper sites were found as predicted, and partly/full trinuclear sites were found. Based on these comparisons, it seems like Phyre2 works perfectly when it comes to finding T1/trinuclear sites, although the results for substrate binding site should be confirmed with another program and/or further lab work. This is one of the things that are revisited in **chapter 5**.

4.3.5 Alignments

When making a multiple alignment of the full sequences of the LMCO's, it would have been possible to just compare the signal sequences in the full alignment. This did however not show the full alignment of the signal sequences as it should be, due to the differences in length. Comparing them within the full sequence made the C-terminals of the signal sequences align with the rest of the protein, and this did not seem right.

The choice of using two different versions of Clustal in the two alignments was based on how the results came out when aligning the signal sequences. Usually, with the improvements in Clustal Omega and the fact that the older versions such as ClustalW are being faced out and Clustal Omega taking over, Clustal Omega would have been the first choice for all alignments made. For smaller sequences such as the signal sequences however, it's easy to see how the different scoring matrices aligned the sequences. Using Clustal Omega's alignment, the different lengths of the signal sequences were considered in a way that created holes in the middle, in stead of at the ends like the algorithm of ClustalW did. The Clustal Omega alignment ended up with fewer conserved residues, and holes that seemed unlikely, and therefore the older version was chosen above the newer. There were no signs of effects like these when using ClustalW vs. Clustal Omega for the full sequences, and hence ClustalW was used.

Signal sequences

The signal sequences in each protein was made out of 31 or 34 aa, not depending on the length of the protein itself, as LMCO's from P11F6, which had the longest chain of aa's, only had a signal sequence of 31 aa's. The tendency seemed to be that the proteins of 567 aa's, LMCO2 and LMCO5, were the ones with a signal sequence of 34 aa. The remaining proteins of 565, 566 and 568 aa, all had 31 aa in their signal sequence. Looking at the alignment in **figure 3.13**, the signal sequences contains multiple conserved residues (18), and in the remaining positions, the residues have similar properties. There were few positions which were not conserved in any way, and apart from the ends and the varying lengths, the signal sequences were found to be quite similar. Along with the 18 fully conserved residues, the first 31 residues contains seven strongly similar and three weakly similar residues.

The similarity between the signal sequences could also be seen in **figure 3.14**, which showed their close relationship. The phylogenetic tree clearly showed how LMCO1 was the most different signal sequence, and that LMCO2 and LMCO5 are identical. It is possible that the reason LMCO1 is so different is that it comes from a plasmid.

Full sequence

Looking at the alignment of the full protein structure in **figure 3.15**, the first notable thing was that Clustal Omega aligned the signal sequences different than ClustalW, which lead to fewer conserved residues. The full alignment showed in the first 31 positions, 16 fully conserved residues, seven strongly similar and two weakly similar. This was due to the way Clustal Omega handled gaps, here by creating end gaps in stead of gaps inside the alignment.

Apart from the signal sequences in the beginning, the overall alignment showed a lot of conserved residues, both fully conserved and residues with strong/weak similarities. As the figure showed, there were four parts of the alignment marked with different coloured lines. These showed signature sequences, or attempts of identifying signature sequences. The green line was found to be the exact sequence of type 1, and the purple was the exact sequence of type 2, which confirmed the LMCO's as actual members of the multicopper oxidase family. With type 1 and type 2 resembling L2 and L4, these were partly covered. There were two attempts of finding signature sequences for laccases, here marked as a pink and a turquoise line, responding to L1 and L3, respectfully. These had residues in common, although they did not match completely. The parts of the laccase signatures which involved the Cu binding sites were fully conserved though, as expected.

Looking at the phylogenetic tree in **figure 3.16**, LMCO1 was still the farthest away from the others. The closer relationship between LMCO2 and LMCO5 was further confirmed though, along with their neighbour LMCO6. The biggest surprise here, when comparing to the signal sequences' tree, was LMCO3 becoming close to LMCO4, even though they as a couple were farther away from the others. When the relationships here were first predicted with the PyMOL models in **section 3.3.4**, it was found that basing predictions on relationship on models did not really work out, and the alignments and phylogenetic trees were necessary to finally determine the relations between the LMCO's.

4.4 Transcriptome, proteome and promoters

This section took a step away from the topics covered so far, and focused on something a bit different. The choice of stepping into transcriptomics/proteomics was done in order to see if it was possible to alter the gene expression of LMCO's, by growing them on media containing 2-methoxy-phenol. For this thesis, the main goal was finding the promoter sequences for these upregulated genes.

4.4.1 Transcriptomics and proteomics

As seen in **table 3.11**, none of the top ten upregulated genes in P11F6 were *LMCO's*. The protein at the top was TonB-dependent receptor, which is found in gram negative bacteria (Ferguson and Deisenhofer, 2002). These proteins are located in the outer membrane, transporting substrates into the bacteria. A majority of the genes that were upregulated were phenol hydroxylases, which was not very surprising giving the substrate added to the media was 2-methoxy-phenol. Phenol hydroxylases are classified as oxidoreductases, capable of reducing molecular oxygen, making them capable of using the substrate and perhaps in this way being favoured over the LMCO's (Kegg, 2015).

To see if the top ten list of upregulated genes actually became translated, the genes were checked against the proteome data. As seen in **table 3.12**, only eight of these genes were actually found in the proteome. It is not known why not all of the top ten upregulated genes were translated, or if they were translated and simply not detected. All that can be done here are speculations, and this topic will be discussed further in **chapter 5**.

4.4.2 Promoters

Most of the work done in this section was based on identifying the promoter sequences for the top ten upregulated genes. Seeing the promoters in **table 3.13**, they were all sorted on m-values from highest to lowest. Using BPROM to find the -10 and -35 regions, it was always kept in mind that this search engine was made for sigma70 promoters in *E.coli* and might not be completely suitable to find these regions in the *Psychrobacter* spp.. The results from this search did however show promoter regions that made sense, the various regions were placed in reasonable areas for them to be actual promoters, and the overall value of the LDF's were good. Being that LDF is a measure with neither upper nor lower limits, they were only used for comparison if BPROM found more than one possible promoter. Using 500 bp upstream of the start codon might have been a bit too much, seeing that all of the promoters found were closer than that. In many cases, BPROM found more than one possible promoter, and in all of these cases, the one closest to the start codon had both the highest LDF value and seemed more promising. As all possible RBS's were found and put in the sequence along with all possible promoters before the decision on which to go for were made, all possibilities were accounted for when choosing the ones to go on with.

ORF starts by SnapGene viewer were included just to show that some of the genes were not recognized by SnapGene Viewers ORF finder, or in some cases the ORF was different. This does not have to have any actual meaning, it was just added to show how things look different using different programs, and also as way of confirming that in most cases, the ORF's were confirmed by more than one program.

When looking at the RBS's found in **table 3.14**, most of the found cases were of AGCA. Three of the promoters also had an extra RBS in AGGAGN. None were found having AGGAGG. Most of the RBS's in the table were included because they seem plausible, when compared to the start codons and the found promoters. For three of the RBS's however, this was not the case. DoxX, hypothetical membrane protein and conserved hypothetical secreted protein all have only one RBS, and this is found outside the part that will become the mRNA. This means that these mRNA's would not be recognized and bound by any ribosome, and hence not translated. As two out of these three were found in the proteome, and the third one was most likely there just not identified, the found RBS's must be wrong. As the three sequences searched for, AGGAGG, AGGAGN and AGCA, are consensus sequences, it should be possible to find sequences that are closely related to these, which was possible in all cases. These were not included in the results here, but could be part of further studies, as discussed in **chapter 5**.

Chapter 5

Further research

5.1 RAST

As seen in **section 3.1 and 4.1**, there is still a lot of work possible to do here. As known from **section 2.1**, all the annotations done here were automatically made. It's possible to do manual annotations along with the automatic one, to ensure that all the results are actually included. Further research on the *Psychrobacter* spp. could perhaps include a deeper search in RAST, doing more studies on the subsystems and the genes in them, and seeing the genes in closer relations to Mauve would be interesting.

5.2 Mauve

In this thesis, the whole point of using Mauve was just to see a general view of evolutionary differences between the genomes, not going deeper into anything. As the figures in **section 3.2** showed, there are a lot of things that could be investigated further here, such as focusing on specific genes and following them in the genomes, or mapping the differences and similarities in much more detailed manners. If further studies on these *Psychrobacter* genomes are interesting, deeper searches in mauve could be very useful.

5.3 Laccase-like Multicopper Oxidases

In the section of pocket analysis using PyMOL, it was discovered that the pocket sequences were of different length. By comparing the parts of the pockets that had identical residues, and specially in the parts where some of the LMCO's had multiple residues while others had just one or two residues. Two cases of this was shown to appear in the areas of trinuclear sites, where a fully conserved and a semi-conserved Leu residue was the only residue found in some of the pockets, while others included whole stretches of residues, including the HXH pattern. It would be possible to see if this Leu residues actually had anything to do with holding the substrate in place, or if there was any bugs in Phyre2's Investigator/Fpocket leading to not finding the complete substrate pockets.

5.4 Transcriptome, proteome and promoters

This is the section that has the most potential for further research, as two of the topics here were barely touched. Trying to figure out why the *LMCO* in P11F6 was not on the top ten list of transcribed genes when growing on media containing added substrate, or trying to find a substrate that will upregulate the transcription of *LMCO*, are both possible studies that could be interesting. As the proteome part of this thesis was simply to see if the top ten list of upregulated genes were found in among the proteins, there is still plenty of work to do here, such as identifying the two missing upregulated genes in the proteome.

In the promoter section, the promoters found here are purely theoretical. As many RBS's were found, some of them even being most likely completely wrong, one interesting study here would be to figure out which ones of these were actually the correct ones, if any. In the case of DoxX, hypothetical membrane protein and conserved hypothetical secreted protein, it was possible to find sequences which resembled the RBS's that were searched for. These resembling sequences had perhaps a base wrong or had two of them in different places. These are, hypothetically, possible RBS's, although with a different strength, as the changes in bases could alter binding. Discovering if these are real could be a part of further studies, and both studies like this or general studies on the promoters could be done by knocking out various RBS's or introducing mutations to the found promoter sequences and see how this would affect the gene expression.

Chapter 6

Conclusion

The automatic annotation of four genomes of *Psychrobacter* spp. by RAST showed four genomes of high similarity, although with some differences, just as expected. The four genomes were all between 3.2 and 3.4 millions bp, had a GC-content of 41.9 to 42.9 % and contained between 2674 and 2914 genes. The distributions found by RAST showed a generally even number of genes in each of the 27 subsystems, with a few exceptions. The exceptions here might be due to a low rate of subsystem coverage by RAST. Using Mauve for evolutionary analysis showed how whole segments of genes were moved around in the genomes as a result of evolution, keeping some parts more or less conserved and at the same time having rearrangements of blocks. Both the RAST results and Mauve could be used in deeper studies of genome and evolution.

By analyzing the LMCO's in various ways and with various tools, they were found to have a length of 565 - 568 aa's of a generally similar distribution, having a Mw of 63.7 - 64 KDa and pI of 7.07 - 8.59, a predicted half-life of ">10 h" and being classified as "stable". Cu binding sites were found in all six LMCO's, both T1, T2 and T3. All of the sites were included in the signature sequences as expected, and the signature sequences for members of the multicopper oxidase family were confirmed. The laccase specific signature sequences were only found partly, which concludes that the LMCO's are definitely members of the multicopper oxidase family, although most likely not laccases.

By comparing 3D models and superpositioning the proteins, the differences and similarities were further investigated and showed some unexpected variations in the LMCO's that were suspected to be similar, while others showed opposite tendencies. 3D models of the substrate pockets were also made, showing some of the pockets being a lot smaller than the others. The pockets were in general a lot more different than expected, although there were some clearly conserved residues. Deciding if the sizes of the pockets are correct could be investigated further in the lab.

Signal sequences were all found to be destined for the periplasm, and an alignment of them showed that they where not as similar as expected. Aligning the whole proteins sequences did however show a lot of conserved residues, especially in the parts containing the Cu binding sites and signature sequences.

When growing the *Psychrobacter* P11F6 on media containing 2-methoxy-phenol, it was expected to see an upregulation of *LMCO's*. This was not found, although a large amount of *phenol hydroxylases* were upregulated. By comparing the top ten list of upregulated genes to the proteome, only eight of them were found. All of the promoters for the top ten upregulated genes were found, and although being purely theoretical, the found sequences seems to fit the expected positions and contents of prokaryot promoters, apart from the RBS's of DoxX, hypthetical membrane protein and conserved hypothetical secreted protein. The given RBS's are probably wrong, and this section needs more work, such as trying to find alternative RBS's. This could be investigated further in the lab for confirmation.

The goal of this thesis was to get an overview of the genomes of the four *Psychrobacter* spp., investigate the six found LMCO's and finding the promoters to the top ten regulated genes as well as doing a brief comparison of transcriptome and proteome results. Even if there are still things that needs further work, the overall goals are reached, and hopefully the results from this work can be used for further research.

Bibliography

Ayala-del Río, H. L., Chain, P. S., Grzymski, J. J., Ponder, M. A., Ivanova, N., Bergholz, P. W., Di Bartolo, G., Hauser, L., Land, M., Bakermans, C., et al. (2010). The genome sequence of psychrobacter arcticus 273-4, a psychroactive siberian permafrost bacterium, reveals mechanisms for adaptation to low-temperature growth. *Applied and Environmental Microbiology*, 76(7):2304–2312.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., et al. (2008). The rast server: rapid annotations using subsystems technology. *BMC Genomics*, 9(1):75.

Baldrian, P., Merhautová, V., Gabriel, J., Nerud, F., Stopka, P., Hrubỳ, M., and Beneš, M. J. (2006). Decolorization of synthetic dyes by hydrogen peroxide with heterogeneous catalysis by mixed iron oxides. *Applied Catalysis B: Environmental*, 66(3):258–264.

Bertrand, G. (1984). Sur le latex de l'arbre à laque. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*, 118:1215–1218.

Bozal, N., Montes, M. J., Tudela, E., and Guinea, J. (2003). Characterization of several psychrobacter strains isolated from antarctic environments and description of psychrobacter luti sp. nov. and psychrobacter fozii sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, 53(4):1093–1100.

Claus, H. (2004). Laccases: structure, reactions, distribution. *Micron*, 35(1):93–96.

Colman, P. M., Freeman, H. C., Guss, J. M., Murata, M., Norris, V. A., Ramshaw, J. A. M., and Venkatappa, M. P. (1978). X-ray crystal structure analysis of plastocyanin at 2.7 Å resolution. *Nature*, 272:319–324.

Couto, S. R. and Herrera, J. L. T. (2006). Industrial and biotechnological applications of laccases: a review. *Biotechnology Advances*, 24(5):500–513.

Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14(7):1394–1403.

Diamantidis, G., Effosse, A., Potier, P., and Bally, R. (2000). Purification and characterization of the first bacterial laccase in the rhizospheric bacterium azospirillum lipoferum. *Soil Biology and Biochemistry*, 32(7):919–927.

Enguita, F. J., Martins, L. O., Henriques, A. O., and Carrondo, M. A. (2003). Crystal structure of a bacterial endospore coat component a laccase with enhanced thermostability properties. *Journal of Biological Chemistry*, 278(21):19416–19425.

Fee, J. A. (1975). *Copper proteins systems containing the "Blue" copper center*, pages 1–60. Springer.

Ferguson, A. D. and Deisenhofer, J. (2002). Tonb-dependent receptors—structural perspectives. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1565(2):318–332.

Gardy, J. L., Spencer, C., Wang, K., Ester, M., Tusnady, G. E., Simon, I., Hua, S., Lambert, C., Nakai, K., Brinkman, F. S., et al. (2003). Psort-b: Improving protein subcellular localization prediction for gram-negative bacteria. *Nucleic Acids Research*, 31(13):3613–3617.

Gasteiger, E., Hoogland, C., Gattiker, A., Wilkins, M. R., Appel, R. D., Bairoch, A., et al. (2005). *Protein Identification and Analysis Tools on the ExPASy Server*, pages 571–607. Springer.

Givaudan, A., Effosse, A., Faure, D., Potier, P., Bouillant, M.-L., and Bally, R. (1993). Polyphenol oxidase in azospirillum lipoferum isolated from rice rhizosphere: evidence for laccase activity in non-motile strains of azospirillum lipoferum. *FEMS Microbiology Letters*, 108(2):205–210.

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., and Lopez, R. (2010). A new bioinformatics analysis tools framework at embl–ebi. *Nucleic Acids Research*, 38(suppl 2):W695–W699.

Ihssen, J., Reiss, R., Luchsinger, R., Thöny-Meyer, L., and Richter, M. (2015). Biochemical properties and yields of diverse bacterial laccase-like multicopper oxidases expressed in escherichia coli. *Scientific Reports*, 5.

Juni, E. and Heym, G. (1986). Psychrobacter immobilis gen. nov., sp. nov.: genospecies composed of gram-negative, aerobic, oxidase-positive coccobacilli. *International Journal of Systematic Bacteriology*, 36(3):388–391.

Kawai, S., Umezawa, T., Shimada, M., and Higuchi, T. (1988). Aromatic ring cleavage of 4, 6-di (tert-butyl) guaiacol, a phenolic lignin model compound, by laccase of coriolus versicolor. *FEBS Letters*, 236(2):309–311.

Kegg (2015). Phenol hydroxylase. `http://www.genome.jp/dbget-bin/www_bget?ec:1.14.13.7`, [Online; accessed 14-August-2015].

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. (2015). The phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, 10(6):845–858.

Kiiskinen, L.-L., Rättö, M., and Kruus, K. (2004). Screening for novel laccase-producing microbes. *Journal of Applied Microbiology*, 97(3):640–646.

Kumar, S., Phale, P. S., Durani, S., and Wangikar, P. P. (2003). Combined sequence and structure analysis of the fungal laccase family. *Biotechnology and Bioengineering*, 83(4):386–394.

Larkin, M. A., Blackshields, G., Brown, N., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., et al. (2007). Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–2948.

Larrondo, L. F., Salas, L., Melo, F., Vicuna, R., and Cullen, D. (2003). A novel extracellular multicopper oxidase from phanerochaete chrysosporium with ferroxidase activity. *Applied and Environmental Microbiology*, 69(10):6257–6263.

Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, 10(1):168.

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., Cowley, A. P., and Lopez, R. (2013). Analysis tool web services from the embl-ebi. *Nucleic Acids Research*, 41(W1):W597–W600.

MendelUniversityBrno (2015). Gene expression. `http://web2.mendelu.cz/af_291_projekty2/vseo/print.php?page=307&typ=html`, [Online; accessed 14-August-2015].

Moghadam, S. H., Cimmino, L., Albersmeier, A., Winkler, A., Wentzel, A., Hohmann-Marriott, M. F., Kalinowski, J., Rückert, C., Lale, R., and Rise, K. (2015). Genome sequencing of four laccase-like multicopper oxidase producing psychrobacter species and activity valdation of gene candidates found by genome mining. *Pending*.

Morozova, O., Shumakovich, G., Gorbacheva, M., Shleev, S., and Yaropolov, A. (2007). “blue” laccases. *Biochemistry (Moscow)*, 72(10):1136–1150.

Nakai, K. and Kanehisa, M. (1991). Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins: Structure, Function, and Bioinformatics*, 11(2):95–110.

O’Malley, D. M., Whetten, R., Bao, W., Chen, C.-L., and Sederoff, R. R. (1993). The role of of laccase in lignification. *The Plant Journal*, 4(5):751–757.

Ouzounis, C. and Sander, C. (1991). A structure-derived sequence pattern for the detection of type i copper binding domains in distantly related proteins. *FEBS Letters*, 279(1):73–78.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., et al. (2014). The seed and the rapid annotation of microbial genomes using subsystems technology (rast). *Nucleic Acids Research*, 42(D1):D206–D214.

Parrello, B. (2015). Figfams. `http://www.nmpdr.org/FIG/wiki/view.cgi/FIG/FigFam`, [Online; accessed 14-August-2015].

Reddy, G. S., Ara, S., Singh, A., Pinnaka, A. K., and Shivaji, S. (2013). Draft genome sequence of psychrobacter aquaticus strain cms 56t, isolated from a cyanobacterial mat sample collected from water bodies in the mcmurdo dry valley region of antarctica. *Genome Announcements*, 1(6):e00918–13.

Reiss, R., Ihssen, J., Richter, M., Eichhorn, E., Schilling, B., and Thöny-Meyer, L. (2013). Laccase versus laccase-like multi-copper oxidase: a comparative study of similar enzymes with diverse substrate spectra. *PLoS ONE*, 8(6).

Reiss, R., Ihssen, J., and Thöny-Meyer, L. (2011). Bacillus pumilus laccase: a heat stable enzyme with a wide substrate spectrum. *BMC Biotechnology*, 11(1):9.

Schrödinger, L. (2010). The PyMOL molecular graphics system, version 1.7.4.

Sharma, P., Goel, R., and Capalash, N. (2007). Bacterial laccases. *World Journal of Microbiology and Biotechnology*, 23(6):823–832.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, 7(1):539.

Solomon, E. I., Sundaram, U. M., and Machonkin, T. E. (1996). Multicopper oxidases and oxygenases. *Chemical Reviews*, 96(7):2563–2606.

Solovyev, V. Salamov, A. (2011). Automatic annotation of microbial genomes and metagenomic sequences. *In Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies (Ed. R.W. Li)*, pages 61–78.

Xiao, Q., Zhang, F., Nacev, B. A., Liu, J. O., and Pei, D. (2010). Protein n-terminal processing: substrate specificity of escherichia coli and human methionine aminopeptidases. *Biochemistry*, 49(26):5588–5599.

Yu, Nancy, Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J., et al. (2010). Psortb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 26(13):1608–1615.

Appendices

Appendix A

Appendix A shows the full list of pocket residues for all LMCO's.

LMCO1	LMCO2	LMCO3	LMCO4	LMCO5	LMCO6
Ser41	His52	Gln40	Gln40	Val44	Asn40
Ala44	Lys53	Ala43	Ala43	Ile45	Val43
Thr45	Glu96	Val44	Val44	Asn46	Asn44
Ile46	Leu119	Asn45	Asn45	Pro55	Ser45
Ser48	Leu120	Asp47	Asp47	Glu93	Arg47
Gln50	Val121	Ala49	Ser49	Leu120	Ala48
Asn51	Pro122	Asp50	Asp50	Leu148	Asp49
Glu97	Phe123	Ile52	His51	Lys149	His50
Leu121	Glu124	Val53	Ile52	Gln150	Ile51
Lys150	Asp126	Pro54	Val53	Ser151	Glu94
Gln151	Thr146	Glu95	Pro54	Lys177	His113
Ser152	Lys149	His114	Phe61	Arg179	Trp114
Lys178	Gln150	Trp115	Leu91	Gly314	His115
Gly315	Ser151	His116	Met93	Ile315	Gly116
Asn316	Gly314	Gly117	Glu95	Asp316	Leu117
Asp317	Ile315	Leu118	Thr98	Met416	Leu118
Pro415	Asp316	Leu119	Val99	Ile484	Val119
Arg416	Pro414	Asp125	Ile101	Pro486	Met123
Met417	Met416	Leu147	His114	Arg489	Asp124
Asn418	Asn424	Val148	Trp115	Val490	Phe144
Leu419	Arg426	Gln149	His116	Ile492	Ile147
Asp420	Glu465	Ser150	Gly117	Gly508	Gln148
Arg490	Arg466	Lys176	Leu118	Met509	Ser149
Gly509	Val482	Gly313	Leu119	Trp510	Lys175
Met510	Ile484	Asn314	Val120	Arg523	Gly312
Trp511	Lys485	Asp315	Pro121	Val539	Ile313
Ser512	Pro486	Ala443	Phe122	Thr540	Asp314
Asp513	Asn487	Ile444	Glu123	Gly541	Pro412

Phe521	Arg489	Ile483	Asp125	Glu542	Met414
Gln522	Val490	Pro485	Lys144	Trp546	Asn415
Arg524	Ile492	Arg488	Phe145	Val564	Leu416
Asp539	Thr493	Val489	Lys146	Val566	Asn422
Val540	Met504	Ile491	Leu147		Arg424
Thr541	Leu506	His504	Lys148		Val480
Gly542	Gly508	Leu505	Gln149		Ile482
Glu543	Met509	His506	Ser150		Lys483
	Trp510	Gly507	Gly151		Pro484
	Ser511	Met508	Thr152		Gly485
	Leu513	Trp509	Tyr153		Arg487
	Val522	Arg522	Tyr171		Aal488
	Arg523	His524	Val172		Ile490
	Lys524	Val538	Lys176		Thr491
	His525	Thr539	Gly313		Met502
	Ile527	Gly540	Ile314		His503
	Phe537	Glu541	Asp315		Leu504
	Asp538	Trp545	Pro413		His505
	Val539	Val563	Asn423		Gly506
	Thr540	Val565	Arg425		Met507
	Gly541		Glu464		Trp508
	Glu542		Arg465		Ser509
	Ala543		Ile483		Asp510
	Trp546		Pro485		Leu511
	Trp548		Arg488		Gln519
	Arg562		Val489		Val520
	Glu563		Ile491		Arg521
	Val564		Thr492		Lys522
	Val566		Met503		His523

	His504	Thr524
	Leu505	Ile525
	His506	Phe535
	Gly507	Asp536
	Met508	Val537
	Trp509	Thr538
	Ser510	Gly539
	Leu512	Glu540
	Val521	Ala541
	Lys523	Trp544
	His524	Trp546
	Thr525	Arg560
	Ile526	Glu561
	Phe536	Val562
	Asp537	Val564
	Val538	
	Thr539	
	Gly540	
	Glu541	
	Trp545	
	Trp547	
	Arg561	
	Glu562	
	Val563	
	Val565	