UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Development of a generic, structural bioinformatics information management system and its application to variation in foot-and-mouth disease virus proteins

by

Tjaart Andries Petrus de Beer

Submitted in partial fulfilment of requirements for the degree Philosophiae Doctor
(Bioinformatics)
in the Faculty of Natural and Agricultural Sciences
Bioinformatics and Computational Biology Unit
Department of Biochemistry
University of Pretoria
Pretoria
November 2008

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Declaration

I, Tjaart Andries Petrus de Beer, declare that the thesis/dissertation, which I hereby submit for the degree Philosophiae Doctor at the University of Pretoria, is my own work and not previously been submitted by me for a degree at this or any other tertiary institution.

SIGNATURE .... DATE ....

Acknowledgments

I want to thank the following people:

- My parents who supported me through all my studies.
- My supervisors for all their guidance, support and help during the last few years.
- All the various funding agencies who made it possible for me to study.
- All my friends and special people for their valuable support.
- My fellow students at the BCBU over the years.
- Irene, ti ringrazio per essermi stata accanto, sostenendomi sempre. Te ne sono davvero grato.

Summary

Structural biology forms the basis of all functions in an organism from how enzymes work to how a cell is assembled. *In silico* structural biology has been a rather isolated domain due to the perceived difficulty of working with the tools. This work focused on constructing a web-based Functional Genomics Information Management System (FunGIMS) that will provide biologists access to the most commonly used structural biology tools without the need to learn program or operating specific syntax. The system was designed using a Model-View-Controller architecture which is easy to maintain and expand. It is Python-based with various other technologies incorporated. The specific focus of this work was the Structural module which allows a user to work with protein structures. The database behind the system is based on a modified version of the Macromolecular Structure Database from the EBI. The Structural module provides functionality to explore protein structures at each level of complexity through an easy-to-use interface. The module also provides some analysis tools which allows the user to identify features on a protein sequence as well as to identify unknown protein sequences. Another vital functionality allows the users to build protein models. The user can choose between building models online or downloading a generated script. Similar script generation utilities are provided for mutation modelling and molecular dynamics. A search functionality was also provided which allows the user to search for a keyword in the database. The system was used on three examples in Foot-and-Mouth Disease Virus (FMDV). In the first case, several FMDV proteomes were reannotated and compared to elucidate any functional differences between them. The second case involved the modelling of two FMDV proteins involved in replication, 3C and 3D. Variation between the several different strains were mapped to the structures to understand how variation affects enzymes structure. The last example involved capsid protein stability differences between two subtypes. Models were built and molecular dynamics simulations were run to determine at which protein structure level stability was influenced by the differences between the subtypes. This work provides an important introductory tool for biologists to structural biology.

Contents

Declaration i

Acknowledgments ii

List of Abbreviations vii

List of Figures x

List of Tables xiii

Chapter 1. Introduction 1

1.1. Biological Data Management 4

1.2. Data Storage 5

1.3. Data Models 6

1.4. Information Management Systems 11

1.5. Common Structural Analysis Needs of Biologists 13

1.6. The Functional Genomics Information Management System (FunGIMS) 16

1.7. Application to Foot-and-Mouth Disease Virus 23

Problem Statement 26

Specific Aims 28

Chapter 2. FunGIMS Design and Implementation 30

2.1. Overview 30

2.2. FunGIMS Design and Technologies 30

2.2.1. Technologies 31

2.2.1.1. Python 31

2.2.1.2. Web Development Framework . . . 32
2.2.1.3. Object-Relational Mapper . . . 32
2.2.1.4. Version Control . . . 34
2.2.1.5. Templating Language . . . 34
2.2.2. Development and Design . . . 35
2.2.2.1. The View . . . 39
2.2.2.2. The Controller . . . 39
2.2.2.3. The Model . . . 40
2.3. FunGIMS Core Functionalities . . . 41
2.3.1. User and Group Management . . . 41
2.3.2. Result Management . . . 42
2.3.3. Searching of Data and Results . . . 42
2.4. FunGIMS Data Model . . . 43
2.4.1. The Data Model . . . 43
2.5. Structural Module . . . 45
2.5.1. Overview . . . 45
2.5.2. Data Model . . . 45
2.5.2.1. Data Sources . . . 48
2.5.3. Functionalities . . . 48
2.5.3.1. Structural Data Representation . . . 49
2.5.3.2. Data Analysis . . . 53
2.5.3.3. Modelling and Molecular Dynamics . . . 57
2.5.3.4. Help Section . . . 60
2.5.3.5. Configuration . . . 61
2.6. Future Improvements . . . 62
2.6.1. FunGIMS . . . 62
2.6.2. Structural Module . . . 63
2.7. Conclusion . . . 65

Chapter 3. Reannotation of Foot-and-Mouth Disease Virus proteome 67

3.1. Introduction 67

3.2. Methods 71

3.3. Results and Discussion 72

3.3.1. Pfam Results 72

3.3.2. Prosite Results 73

3.3.3. Secondary Structure Results 79

3.3.4. Pepstat Hydrophobic Plot Results 81

3.4. Conclusion 82

Chapter 4. Modelling of Foot and Mouth Disease Virus 3C and 3D Non-structural Proteins 92

4.1. Introduction 92

4.2. Methods 95

4.2.1. 3C Protease 95

4.2.2. 3D RNA Polymerase 95

4.3. Results and Discussion 97

4.3.1. 3C Protease 97

4.3.2. 3D RNA Polymerase 106

4.4. Conclusion 108

Chapter 5. FMDV Capsid Stability and Variation Analysis 112

5.1. Introduction 112

5.2. Methods 114

5.2.1. Capsid Protomer 115

5.2.2. Capsid Pentamer 115

5.3. Results and Discussion 116

5.3.1. Capsid Modelling 116

5.3.2. Protomer Modelling and Variation Mapping 116

5.3.3. Pentamer Molecular Dynamics 120

5.4. Conclusion 131

Chapter 6. Concluding Discussion 135

Summary 140

Bibliography 142

Appendix 154

L 154

VP1 154

VP2 155

VP3 156

VP4 157

2A 158

2B 158

2C 159

3A 160

3B1 161

3B2 161

3B3 162

3C 162

3D 163

List of Abbreviations

Å	Angstrom
aa/AA	Amino Acid
A	Alanine
ANSI	American National Standards Institute
C	Cysteine
CHARMM	Chemistry at HARvard Macromolecular Mechanics
CG	Conjugate Gradient
D	Aspartic acid
DNA	Deoxyribonucleic Acid
E	Glutamic acid
EBI	European Bioinformatics Institute
EC	Enzyme Commission
Ec	*Escherichia coli*
EM	Electron Microscopy
EST	Expressed Sequence Tag
F	Phenylalanine
FMDV	Foot and Mouth Disease Virus
FuGE	Functional Genomics Experiment
FunGIMS	Functional Genomics Information Management System
G	Glycine
GB	Gigabytes
H	Histidine
HAV	Hepatitis A Virus
HRV	Human Rhino Virus

HS	Heparan Sulfate
HMM	Hidden Markov Model
I	Isoleucine
ISO	International Standards Organization
K	Kelvin
K	Lysine
kb	kilobases
kD	kilo Dalton
L	Leucine
M	Methionine
MB	Megabytes
MSD	Macromolecular Structure Database
MVC	Model View Controller
N	Asparagine
NCBI	National Center for Biotechnology Information
ns	nanosecond
P	Proline
PDB	Protein Data Bank
Pfam	Protein families database
ps	picosecond
Q	Glutamine
R	Arginine
S	Serine
SD	Steepest Descent
sid	System Identifier
SQL	Structure Query Language
T	Threonine
TMHMM	Trans-Membrane Hidden Markov Model
V	Valine
W	Tryptophane

XML	eXtensible Markup Language
Y	Tyrosine

List of Figures

1.1. The exponential growth of data deposits 3

1.2. A high level overview of the MSD data model 7

1.3. The different entities belonging to the Structure entity in MSD 8

1.4. The classes of `FuGE.Common` 10

2.1. TurboGears schematic diagram 33

2.2. The Model-View-Controller architecture 36

2.3. The overall design of the FunGIMS system 37

2.4. The result of a search for "dihydropteroate synthase" 43

2.5. The PDB file format 46

2.6. The relationship between the `Structure` object and the FuGE data model 47

2.7. The relationship between different secondary structures in a chain and the residues in a protein 52

2.8. The data model for the high level structure class 53

2.9. The primary view when a user views a protein 54

2.10. The chain summary view for a specific chain in a protein 55

2.11. The strand summary for a protein chain 56

2.12. The summary view for all helices in a protein chain 57

2.13. The summary view for all the turns in a protein chain 58

2.14. The tertiary structure view of a protein complex 59

2.15. The different tools available in the Structural module 60

2.16. The results from a transmembrane helix prediction 61

2.17. The results from a Hmmer search across Pfam 62

2.18. The results from a PROCHECK analysis 63

2.19. The Help page for the Structural Module 64

2.20. The modelling and dynamics interfaces 66

3.1. The distribution of FMDV outbreaks from 2000-2006 68

3.2. The genome organization of FMDV 69

3.3. The annotation of the L protein 75

3.4. The annotation of the VP1 protein 76

3.5. The annotation of the VP2 protein 77

3.6. The annotation of the VP3 protein 78

3.7. The annotation of the VP4 and 2A proteins 83

3.8. The annotation of the 2B protein 84

3.9. The annotation of the 2C protein 85

3.10. The annotation of the 3A protein 86

3.11. The annotation of the 3B protein 87

3.12. The annotation of the 3C protein 88

3.13. The annotation of the 3D protein 89

4.1. The structure of 3C^{pro} from FMDV 93

4.2. The structure of Polio virus 3D^{pol} 94

4.3. The alignments used in the modelling of 3C^{pro} 98

4.4. The alignment used in the modelling of 3D 99

4.5. SAT 3C^{pro} variation mapped onto a SAT 3C^{pro}model 101

4.6. The variation seen in the 3C^{pro} 102

4.7. The variation seen in the 3D RNA polymerase 102

4.8. The invariant region on a model of 3C^{pro} 103

4.9. SAT 3D variation mapped onto a SAT 3D model 104

4.10. The hypervariable and highly conserved regions in 3D 105

5.1. A schematic representation of the FMDV capsid 113

5.2. The capsid as generated using symmetry operations 117

5.3. The two complexes used during analysis 118

5.4. The acid inactivation of FMDV in BHK-21 cells 119

5.5. The alignments used to model the capsid proteins 122

5.6. The variation as mapped to a pentamer model 125

5.7. pH inactivation of ZIM/7/83 vs ZIM/5/83 126

5.8. The RMSD variation of the two strains during the simulation 127

5.9. The interaction interface between two pentamers 130

5.10. The hydrogen bond network of the pentamer interface 131

5.11. The location of Phe 493 in relation to His 575 132

List of Tables

1.1. Comparison between Flat-file data storage and Relational database data storage . . 6

1.2. The two categories of FuGE . . . 9

1.3. Comparison between the technologies in currently available LIMS. . . . 12

2.1. The technical specifications of FunGIMS. . . . 40

3.1. The Pfam pattern matches found in each proteome . . . 74

4.1. The SAT 3C and 3D sequences used in analysis . . . 96

4.2. The changes in the invariant section in 3C . . . 100

5.1. The results from the comparison of the VP1 chain of the 6 strains . . . 121

5.2. The results from the comparison of the VP2 chain of the 6 strains . . . 123

5.3. The results from the comparison of the VP3 chain of the 6 strains . . . 124

5.4. The differences between the P1 of ZIM/5/83/2 and ZIM/7/83/2 . . . 126

5.5. The pKa changes predicted by PROPKA . . . 126

5.6. Explanations for the pKa changes predicted by PROPKA . . . 129