

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Beyond BLASTing: Ribonucleoprotein evolution via structural prediction and ancestral sequence reconstruction

A thesis presented in partial fulfilment
of the requirements for the degree of

Doctor of Philosophy in
Genetics

at Massey University, Manawatū Campus

Toni K. Daly
2016

Abstract

Primary homology in DNA and protein sequence has long been used to infer a relationship between similar sequences. However gene sequence, and thus protein sequence, can change over time. In evolutionary biology that time can be millions of years and related sequences may become unrecognisable via primary homology. This is demonstrated most effectively in chapter 4a (figure 10). Conversely the number of possible folds that proteins can adopt is limited by the attractions between residues and therefore the number of possible folds is not infinite. This means that folds may arise via convergence between evolutionarily unrelated DNA sequences.

This thesis aims to look at a process to will wring more information from the primary protein sequence than is usually used and finds other factors that can support or refute the placement of a protein sequence within the family in question. Two quite different proteins; the Major Vault Protein whose monomers make up the enigmatic vault particle and the argonaute family of proteins (AGO and PIWI) that appear to have a major hand in quelling parasitic nucleic acid and control of endogenous gene expression, are used to demonstrate the flexibility of the workflow.

Principally the method relies on prediction of three-dimensional structure. This requires at least a partially solved crystal structure but once one exists this method should be suitable for any protein. Whole genome sequencing is now a routine practice but annotation of the resultant sequence lags behind for lack of skilled personnel. Automated pipeline data does a good job in annotating close homologs but more effort is needed for correct annotation of the exponentially growing data bank of uncharacterised (and wrongly characterised) proteins. Lastly, in deference to budding biologists the world over, I have tried to find free stable software that can be used on an ordinary personal computer and by a researcher with minimal computer literacy to help with this task.

Acknowledgments

I have been a student since 1988 but now is the time to thank everyone and of course over the years there have been so many people that have inspired and encouraged me and I will miss some people that I shouldn't have.

Trevor Kitson made me realise that it was possible to be smart and funny, Mike Hardman for having the foresight to grab a box of tissues whenever I walked through his office door when I eventually became an internal student. Although Massey could bend the rules in those pre-studylink days a student loan required a minimum of seven papers.....but I only had one semester! So I physically couldn't attend all my classes, I just grabbed the notes, sat the exams and came back the following year to do the lab work. I want to thank Mark Patchett for just being the best tonic one could ever need when things looked glum, Rosie Bradshaw for her kindness, and Kathryn Stowell, a newly minted lecturer when I was an internal student, has encouraged me ever since.

Lesley Collins for allowing a complete unknown to contribute to her book, Austen Ganley from Albany for helping to prepare me for Vienna when Palmerston North was just too far away and lastly Andrew Sutherland-Smith and David Penny who have stoically listened to my troubles academic and personal once a week for six years. In retrospect I do not know how they coped and without them I surely would have quit. I also thank David for organising payment of my fees, and Massey for paying the bill.

It is an extraordinarily lonely thing to attempt a PhD without the camaraderie and the vicarious learning opportunities of watching fellow students give presentations etc. I want to especially thank Bruce White from the library for his help and patience and Tim White for helping me navigate computer-speak.

I set out to show the students of Northland New Zealand, that you can achieve your goals without the financial, geographical and educational advantages that you perceive everyone else to have. Protein annotation in particular needs help and people all over the world with access to a computer and an internet link can join in.

My biggest thanks though must go to my husband Dan Daly for his continual support despite landing him with three children as well as a full time job while I did my undergraduate stints, and I know that I have been a serious drain on finances and I know that we haven't been out in the kayaks (or out anywhere) for a long time and now I promise that our time will be spent together.

Preface

This thesis is written according to the regulations of the latest version of the Handbook for Doctoral Study (2016), published online by the Doctoral Research Committee. This thesis complies with the format of a thesis based on publication as described in the handbook and includes both published and unpublished chapters. The chapters do not follow the order of publication in order to better demonstrate the flow of the development of the work. Chapters 1 and 2 have been written by Toni Daly as an introduction and literature review and are not intended for publication.

Chapter 1. General Introduction.

Chapter 2. Literature review of the Major Vault Protein.

Chapter 3a. Toni K Daly *et al.* (2013) Beyond BLASTing: Tertiary and quaternary structure analysis helps identify Major Vault Proteins. *Genome Biology and Evolution* 5: 217-232.

Chapter 3b. Toni K Daly *et al.* (2013) In silico resurrection of the Major Vault Protein suggests it is ancestral in modern eukaryotes. *Genome Biology and Evolution* 5 (8): 1567-1583.

Chapter 4. Toni Daly, X. Sylvia Chen and David Penny (2011) How old are RNA networks? (L J Collins ed. *RNA Infrastructure and Networks*) Landes BioScience and Springer Science.

Chapter 4a. Toni Daly, *et al.* (2016) Long Long AGO: The evolutionary history of Argonaute and PIWI in metazoa by ancestral protein inference and structure prediction. (submitted).

Chapter 4b. Toni K Daly *et al.* (2016) Argonaute gain and loss during fungal evolution. (submitted).

Chapter 4c. Toni Daly *et al.* Argonautes origins in eukaryotes. (in preparation).

Chapter 6. Conclusion.

David Penny is co-author on all of the published and prepared papers and Andrew Sutherland-Smith is co-author on five of them. Contributions to each paper are described in Appendix IV.

Contents

Chapter one: Introduction

1. Overview	1
1.1. Ribonucleoproteins	1
1.2. Protein Evolution	2
1.2.1. Evolutionary Aspects of the Chosen Proteins	3
1.3. Pipeline.....	4
1.3.1. Basic Local Alignment Search Tool (BLAST)	5
1.3.2. Protein Annotation and Prediction	6
1.3.3. Tree Calculations	7
1.3.4. Ancestral Sequence Reconstruction (ASR).....	8
1.4. Major Vault Protein (MVP)	8
1.4.1. MVP Form and Function.....	9
1.5. Vault Function.....	11
1.5.1. Cellular Location	12
1.5.2. Vault Cargo.....	13
1.5.3. Developmental / Scavenging Roles	13
1.5.4. Association with lipid rafts.....	14
1.5.5. Detoxification roles	15
1.5.6. Multi Drug Resistance (MDR)	15
1.5.7. Cell signalling.....	17
1.5.8. Possible future biotechnological use of the vault particle	18
1.6. VTRNA	19
1.6.1. Vault RNA nomenclature	20
1.6.2. VTRNA function	22
1.7. Summary	24
2. a: Beyond BLASTing	26
2.1. Sequence similarity identifiers	27
Abstract.....	29
Introduction.....	29
Materials and Methods.....	31
Results.....	34
Discussion.....	41
2. b: <i>In silico</i> Resurrection.....	45
Abstract.....	45

Introduction.....	45
Materials and Methods.....	57
Results.....	50
Discussion.....	56
Chapter Three: Introduction to the Argonaute Family	
3. The defence of the Dark Arts	62
How Old Are the RNA Networks?.....	67
Abstract.....	67
Introduction.....	67
Regulatory networks of small RNAs.....	68
RNA regulation and defence against the Dark Arts.....	71
Other regulatory RNAs.....	78
How old are the different interactions of RNA?.....	79
Conclusion.....	81
Chapter Four: The Evolution of the Argonautes	
4. An investigation into Argonaute evolution using 3-D structural prediction	83
4.1. Abstract	87
4.2. Introduction	87
4.2.1. Argonaute proteins.....	88
4.2.2. <i>In silico</i> analysis	92
4.3. Methods	93
4.4. Results	96
4.5. Ancestral reconstruction.....	103
4.6. Evolution	108
4.7. Discussion	111
4.7.1. Annotation issues.....	111
4.7.2. General.....	112
4. b: Argonaute gain and loss during fungal evolution.....	116
4b.1 Abstract.....	116
4b.2 Introduction.....	116
4b.3 Method.....	121
4b.4 Results.....	123
4b.4.1 Yeast and fungi.....	123
4b.4.2 The <i>R. irregularis</i> AGO expansion.....	127
4b.4.3 Microsporidia.....	134

4b.5 Discussion.....	139
4. c: Argonautes in eukaryotes.....	142
4c.1 Abstract.....	142
4c.2 Introduction.....	142
4c.3 Method.....	146
4c.4 Results.....	149
4c.4.1 SAR (Stramenopile, Alveolate and Rhizaria).....	151
4c.4.2 Amoebozoa.....	156
4c.4.3 Excavates.....	157
4c.4.4 Red and green algae.....	160
4c.4.5 Land plants.....	161
4c.5 Ancestral trees.....	163
4c.6 Discussion.....	167
Chapter Five: Conclusion	
5. Conclusion.....	170
5.1. The chosen proteins.....	170
5.2. Links with the past: Major Vault Protein.....	171
5.2.1. Bacteria.....	172
5.2.2. Archaea.....	174
5.3. Links with the past: Argonaute Family Proteins.....	176
5.4. Challenges of the method.....	178
5.4.1. BLASTp.....	178
5.4.2. MSA.....	179
5.4.3. Structural prediction.....	179
5.4.4. RosettaDock (ROSIE).....	180
5.4.5. Tree calculation.....	181
5.4.6. Ancestral Sequence Reconstruction (ASR).....	181
5.4.7. FATCAT.....	182
5.4.8. Philanthropy.....	183
5.5. The last word.....	184
Glossary.....	185
References.....	189

List of Figures

Chapter one:

Fig. 1.1 Pipeline evolution	4
Fig. 1.2 Vault ribonucleoprotein structure	10
Fig. 1.2 Refinement of the vault structure (2013).....	11

Chapter Two (published papers)

Preface

Fig. 2.1 Geneious alignment shading	28
---	----

Chapter 2a:

Fig. 1 Vault ribonucleoprotein structure.....	30
Fig. 2 MVP monomer comparison.....	32
Fig. 3 Structural effect of the 2ZUO*b constraint.....	35
Fig. 4 RosettaDock results from the crystal structure cap-helix.....	36
Fig. 5 RosettaDock results from the rat MVP shoulder region.....	37
Fig. 6 I-TASSER modelling results for the negative control sequences.....	38
Fig. 7 <i>Naegleria gruberi</i> MVP I-TASSER modelling.....	40

Chapter 2b:

Fig. 1 Problems with FastML and PAML.....	48
Fig. 2 Vault ribonucleoprotein structure.....	50
Fig. 3 MrBayes tree of unlikely placements.....	51
Fig. 4 Heatmap showing identity between the bacterial homolog pairs.....	53
Fig. 5 How inserts can affect the I-TASSER score.....	54
Fig. 6 A comparison of MVP structures with the stramenopile ancestor.....	55
Fig. 7 Heatmap of ancestors.....	56
Fig. 8 Structural diagrams of I-TASSER predictions.....	57
Fig. 9 ConSurf diagram showing conserved and non-conserved residues from multiple sequences.....	57

Chapter Three:

Fig. 1 Transcription of endogenous DNA that gives rise to dsRNA.....	70
Fig. 2 A comparison of RNAi networks involved with the defence of the Dark Arts...71	71
Fig. 3 Gemini viruses.....	72
Fig. 4 The CRISPR system.....	73
Fig. 5 Working backwards through four stages of the origin of life.....	76

Chapter Four (submitted papers)

Preface

Fig. 4.1 Workflow for chapter four.84

Chapter 4a:

Fig. 4a.1 The human AGO2 crystal structure PDB:4W5N.....91

Fig. 4a.2 Similarities and differences in predicted structure between the difficult-to-resolve flatworm sequences.98

Fig. 4a.3 The human AGO2 crystal structure aligned with the low scoring predicted structures identified in table 1.101

Fig. 4a.4 The difference between alignment and predicted structure in *X. tropicalis* PIWI3.103

Fig. 4a.5. The C terminal signature in PIWI-like ASR.104

Fig. 4a.6 Metazoan predicted structure for PIWI ancestors.105

Fig. 4a.7. The C terminal signature in AGO-like ASR.107

Fig. 4a.8 Metazoan predicted structure for AGO ancestors.108

Fig. 4a.9 Predicted structure for the putative sequences identified by BLAST in *S. rosetta*.110

Fig. 4a.10 An example where BLAST searches are ambiguous.113

Fig. 4b.1 Solved structures of the argonaute family119

Fig. 4b.2 A simplified Unikont tree showing the proposed relationship between the various phyla that contributed to the work.120

Fig. 4b.3 *A. gossypii* (UniProtKB:M9MXJ8) putative AGO sequence identified by BLAST.124

Fig. 4b.4. Structural predictions of the reconstructed ancestors from the tree of representative fungi species.126

Fig. 4b.5 Fates of the *R. irregularis* expansion.130

Fig. 4b.6 A comparison of UniProtKB:U9SQW1 with the solved structure of the human argonaute.131

Fig. 4b.7. A comparison of the microsporidian ancestor and *M. daphniae* with solved structures.135

Fig. 4b.8. An unrooted tree of fungi and metazoan sequences re-created by ASR.....137

Fig. 4c.1 I-TASSER structural prediction for *Trypanosoma brucei*.145

Fig. 4c.2 I-TASSER 3-D structural prediction of BLAST results with high number of residue changes per site.150

Fig. 4c.3 The divergent argonaute Twi12 from *T. thermophila*.154

Fig. 4c.4 A comparison between the canonical argonaute and PIWI-tryp within *Trypanosomes*.158

Fig. 4c.5 Mr Bayes tree of annotated argonaute proteins in trypanosomid protozoans.159

Fig. 4c.6 A rooted tree from the calculated ancestral sequences.	165
Fig. 4c.7 Tree of all eukaryote ancestral reconstructions.....	166
Fig. 5.1 <i>Escherichia coli</i> TolA (UniProtKB:P19934).....	172
Fig. 5.2 Bacterial MVP monomer.	173
Fig. 5.3 Full size bacterial MVP monomers.....	174
Fig. 5.4 Putative archaea homolog sequences.....	175
Fig. 5.5 The highly conserved C terminal from MVP.....	176
Fig. 5.6 I-TASSER and Phyre2 comparison.	180
Fig. 5.7 An I-TASSER comparison between raw and trimmed ASR node 1 sequences.	182
Fig. 5.8 FATCAT structural alignment of the truncated ancestor with HsAGO2.	182
Appendix I.....	204
Appendix II.....	209
Appendix III.....	213
Permission and contributions.....	217