INVESTIGATING THE RECOGNITION AND INTERACTIONS OF NON-POLAR α HELICES IN BIOLOGY

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Chemistry

2017

James Alexander Baker
orcid.org/0000-0003-0874-2298

Contents

Word count 22,000

List of Tables

List of Figures

Acronyms

SNARE Soluble N-Ethylmaleimide-Sensitive Factor Attachment Receptor. 11

The University of Manchester

James Alexander Baker
Doctor of Philosophy
Investigating the Recognition and Interactions of Non-Polar α Helices in Biology
November 17, 2017

Abstract

Non–polar helices figure prominently in structural biology, from the first protein structure (myoglobin) through trans-membrane segments, to current work on recognition of protein trafficking and quality control. Trans-membrane α helix containing proteins makep around a quarter of all proteins, as well as two-thirds of drug targets, and contain some of the most critical proteins required for life as we know it. Yet they are fundamentally difficult to study experimentally. This is in part due to the very features that make them so biologically influential: their non–polar trans-membrane helix regions. What is missing in the current literature is a nuanced understanding of the complexities of the helix composition beyond a hydrophobic region of around 20 residues. Currently, it is known that the properties of trans-membrane protein α helices underpin membrane protein insertion mechanisms.

By leveraging large datasets of trans-membrane proteins, this thesis is focused on characterising features of α helices en masse, particularly regarding their topology, membrane–protein interactions, and intramembrane protein interactions.

In this thesis, I make the argument that there are different classifications of trans-membrane α helices. These have markedly different evolutionary pressures, these different classes interact differently with the membrane, and each class serve the protein differently.

Lay Abstract

The survival of each of our cells relies on a a cellular barrier to separate themselves from the surrounding environment. This cellular skin can be thought of as the bag that contains all the important machinery required for normal cell function. The barrier works by being chemically very different to both the outside environment, and to the inside of the cell, which in both cases are mostly water. The membrane is fatty, and because of that repels water.

Proteins are the molecular machinery that form much of the cell structure and shape as well as carrying out many of the cell's routine tasks. Around a third of our genome codes for proteins that are permanently embedded in the membrane, but because these proteins are adapted for a life in the water repelling cell wall, they are very hard to study in laboratories which need to look at proteins in water.

In this thesis, we focus particularly on the parts of the protein that are embedded in the water repelling cellular skin. Traditionally, these regions are hard to study, because we must first remove them from the cellular wall, which causes problems since the embedded regions also repel water and this often causes them to stick to one another, making them hard to work with in a laboratory setting.

We analyse thousands of proteins to further our understanding of electrical charges in the embedded regions and find that negative charge on the outside of the cell has been evolutionarily selected across bacteria, animals, and plants. This is especially true for regions that specifically anchor the protein into the cellular wall. Where the embedded regions have additional function, for example ferrying something in or out of the cell, the negative charge "bias" can no longer be seen.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgements

I would like to thank all members of both the Eisenhaber research group, as well as the Curtis and Warwicker research group for discussion, but in particular Jim Warwicker, Frank Eisenhaber, Birgit Eisenhaber, and Wing-Cheong Wong for supervision and guidance during my research. I would also like to thank The University of Manchester and the A*STAR Singapore Bioinformatics Institute for funding the project. Furthermore, I would like to extend my gratitude to the research group of Professor Stephen High.

Chapter 1

Tail-Anchored Protein Datasets

1.1 Abstract

1.2 Introduction

This study aims to identify Soluble N-Ethylmaleimide-Sensitive Factor Attachment Receptor (SNARE) proteins in eukaryotic proteomes by filtering through large datasets using automatically predicted TrEMBL consensus, and manually annotated SWISS-PROT transmembrane regions. The pipeline generates a list of singlepass proteins with a transmembrane domain close to the C terminal, that are not splice isoforms. A previous study predicted 411 tail anchor proteins [1].

1.3 Methods

The original list UniProt protein database was queried for records containing "TRANSMEM" annotation on June 15, 2016, totaling 75826 records from swissprot, and 12322000 records from TrEMBL.

1.3.1 Filtering the Uniprot database

Steps carried out by Kalbfleisch *et al.* published in Traffic 2007 (8: 16871694) [1], were recreated using up to date tools. The nonredundant human dataset of 145,715 proteins from SwissProt and TrEMBL [2]. 2,478 singlepass proteins were programmatically

extracted according to the TRANSMEM count from that list. Then TMDs not within 15AA of the C terminal were removed, resulting in 455 proteins. No splice isoforms were detected according to searching for NON_TER annotation. 195 proteins of the 411 predicted proteins from the previous study were successfully mapped using the Uniprot mapping tools [2]. Duplicate IDs from the previously predicted tail anchored protein were removed from the set. The remaining dataset contained XXX proteins.

1.3.2 Calculating Hydrophobicity

1.3.3 Calculating Sequence Complexity

1.4 Results

1.4.1 An Up To Date Tail-Anchor Dataset

1.4.2 Potential Tail-Anchored SNARE Protein Discovery

1.4.3 Biology of Spontaneously Inserting Tail Anchored Proteins

Chapter 2

Protein Classification Based on Intra-membrane Complexity Arrangement

2.1 Abstract

2.2 Introduction

2.3 Methods

2.4 Results

Chapter 3

Conclusions

3.1 Outlook

3.1.1 The hydrophobicity–sequence complexity continuum

We hypothesize that the hydrophobicity–sequence complexity continuum contains nuanced codes for different functions and that such differentiation of sequence and structural properties will allow assignment to these varying functions. Additionally, we suggest probing functional classification of yet uncharacterized membrane proteins by similarities of combinations of complex TM sets to well studied membrane proteins and finding those classes of TM proteins where this principle is most directly applicable.

Bibliography

1. Kalbfleisch, T., Cambon, A. & Wattenberg, B. W. A bioinformatics approach to identifying tail-anchored proteins in the human genome. *Traffic* **8,** 1687–1694 (2007).

2. Bateman, A. *et al.* UniProt: A hub for protein information. *Nucleic Acids Research* **43,** D204–D212 (2015).