UNIL | Université de Lausanne

Unicentre
CH-1015 Lausanne
http://serval.unil.ch

Year : 2016

Understanding Molecular Evolution Of Vesicular Trafficking Proteins By Using Multiple Sequence Information

Srivastava Nicee

Srivastava Nicee, 2016, Understanding Molecular Evolution Of Vesicular Trafficking Proteins By Using Multiple Sequence Information

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_E4722FC4320D2

Droits d'auteur
L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

Copyright
The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Département des neurosciences fondamentales

Understanding Molecular Evolution Of Vesicular Trafficking Proteins By Using Multiple Sequence Information

Thèse de doctorat ès sciences de la vie (PhD)
Integrated Experimental and Computational Biology program

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Nicee Srivastava

Masters in Bioinformatics, Banaras Hindu University (BHU), India

Jury

Prof. Richard Benton, Président du comité
Prof. Dirk Fasshauer, Directeur de thèse
Prof. Nicolas Salamin, Expert du comité
Prof. Michael Hothorn, Expert du comité

Lausanne, UNIL, 2016.

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Ecole Doctorale

Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président · e	Monsieur	Prof.	Richard	**Benton**
Directeur · rice de thèse	Monsieur	Prof.	Dirk	**Fasshauer**
Experts · es	Monsieur	Prof.	Michael	**Hothorn**
	Monsieur	Prof.	Nicolas	**Salamin**

le Conseil de Faculté autorise l'impression de la thèse de

Madame Nicee Srivastava

Master in Bioinformatics Banaras Hindu University, India

intitulée

Understanding Molecular Evolution
Of Vesicular Trafficking Proteins
By Using Multiple Sequence Information

Lausanne, le 22 avril 2016

pour le Doyen
de la Faculté de biologie et de médecine

Richard Benton

Prof. Richard Benton

Acknowledgement

I am sincerely thankful to Prof. Dirk Fasshauer for giving me the opportunity to work on this project. I would like to thank him for his guidance, support, patience, and help throughout my project.

I am thankful to Dr. Tobias Kloepper for his advice and guidance during the first half of my project. My sincere thanks to Dr. Nickias Kienle for all his help, support and advice throughout my project.

I also wish to thank my thesis committee, Prof. Richard Benton, Prof. Nicolas Salamin and Prof. Michael Hothorn for their insightful comments, advice and encouragement. I am especially grateful to Prof. Nicolas Salamin for letting me attend and present at his group’s lab meetings that provided me good opportunity for useful discussions.

I am extremely thankful to Dr. Linda Dib for useful discussions, tips and advice during the course of the project.

I thank my colleagues Nickias Kienle, Janeta Iordanova, Czuee Morey, Emilio Iraheta, Xiong Chen, Yves Mingard, Frédérique Varoqueaux, Dany Khalifeh and Jules Duruz for all the discussions and for a fun filled and warm atmosphere.

I would like to thank all my friends in Switzerland for their support and all the fun time I had during all these years.

I want to thank my family, especially my parents and my sister for their love and support. I am indebted for their support and help specially while I was writing my thesis.

Most importantly, I want to thank my husband, Srini for lots of useful discussions and ideas. His love, support and encouragement kept me going through tough times.

Lausanne, 15th March 2016. Nicee Srivastava

Abstract

Vesicle trafficking is an essential process by which molecules (nutrients, protein, lipids, etc.) are transported from a donor compartment to an acceptor within the eukaryotic cell. Currently, it is becoming clear that the important proteins participating in vesicular trafficking are highly conserved, not only between different species but also between different vesicle trafficking steps. Previous phylogenetic analyses showed that the interacting proteins of the vesicle fusion apparatus arose by duplication and diversification of prototypic protein machinery. It is conceivable that these proteins might show common patterns of episodes of duplication and diversification. Exploring the sequence information of the key proteins of the vesicle fusion machinery can provide better insights about their evolution, function and interacting surfaces. The main aim of this thesis was to extract novel insights about the structure and function of the proteins of the vesicle fusion machineries, by exploiting the covariation pattern from the multiple sequence alignments of these proteins. To perform a comprehensive sequence analysis, a tool with different analysis options and visualizations was developed. As a first step towards exploring sequence information, the tool was used to identify intra-protein covariation for the SM protein family, which has been studied for two decades, but their molecular role is still debated. The covarying residues obtained were found to be lying in distant regions in the tertiary structure of the protein. The result appeared to hint at conformational changes and allosteric coupling as discussed in the recent literature. In order to understand the functional relevance of the covarying sites obtained, heatmap visualization and improved analysis pipeline with clustering of the covariation data was developed. The improved approach was applied on test cases from previous studies, on simulated dataset and on SNAP proteins. SNAP proteins are structurally and functionally different from SM proteins and participate in the dissociation of the SNARE complex together with NSF ATPase. In all the test cases, the clusters of covarying residues were found to be restricted to a particular region of the protein. Many residues with known structural and functional importance were also identified. The improved analysis pipeline was applied again on the SM proteins. Differences between the clusters of covarying residues were observed for SM proteins and proteins like enzymes and SNAPs. Most of the covarying residues within the clusters were lying close to each other, while some were lying distant to each other. The distant residues appeared to be forming a chain in the 3D structure of the protein. Comparable pattern of covarying residues within the 3D structure was observed for all the analyzed subfamilies of SM proteins, suggesting that the analysis uncovered a common feature within the protein family. An inter-protein analysis of Sly1 and Syntaxin 5 pair also revealed clusters of distantly lying residues that appeared to form networks connecting different regions of the protein. Overall, the results suggested a possible communication between the two binding sites involving a network of co-varying residues along the structure of SM/Syntaxin complexes. Thus, the approach of clustering the covariation data and detecting the groups of covarying residues helped to identify putative novel structural and functional features for the SM proteins and SNAP proteins. The future scope of this work would include performing mutational and biochemical studies to test the importance and allosteric nature of the identified covarying sites.

Résumé

Le trafic vésiculaire est un processus essentiel pour lequel des molécules (nutriments, protéines, lipides, etc.) sont transportés d'un compartiment donneur à un accepteur dans la cellule eucaryote. Actuellement, il est devenu clair que les protéines qui participent au trafic vésiculaire sont hautement conservées, non seulement entre les différentes espèces, mais aussi entre les différentes étapes du trafic vésiculaire. Les analyses phylogénétiques antérieures ont montrées que les protéines d'interaction venant de l'appareil de fusion vésiculaire sont nées de la duplication et de la diversification de la machinerie protéique prototypique. Il est concevable que ces protéines pourraient montrer des modèles communs d'épisodes de duplication et de diversification. Explorer l'information de séquence des clés protéines de la machine de fusion des vésicules peut offrir de meilleures compréhensions de l'évolution, la fonction et les surfaces en interaction. L'objectif principal de cette thèse était d'extraire des informations nouvelles sur la caractéristiques structurales et fonctionnelles des protéines des machineries vésicules de fusion, en exploitant le modèle de covariation des alignements de séquences multiples de ces protéines. Pour effectuer une analyse complète de la séquence, un outil comprenant différentes options d'analyse et de visualisations a été développé. La première étape de cette étude s'articule autour de l'exploration de l'information de séquence. L'outil a été utilisé pour identifier les covariations intra-protéiques de la famille des protéines SM, qui a été étudié pendant deux décennies, mais où leur rôle moléculaire est encore débattu. Les résidus covariants obtenus se sont avérés se trouvant dans des régions éloignées de la structure tertiaire de la protéine. Le résultat semblait faire allusion à des changements de conformation et à des couplages allostériques tels que discutés dans la récente littérature. Afin de comprendre la pertinence fonctionnelle des sites covariants obtenus, une visualisation de la carte de chaleur et une meilleure canalisation de l'analyse avec le regroupement des données de covariance ont été développées. L'approche améliorée a été appliquée à partir d'études antérieures, sur des données simulées et sur des protéines SNAP. Les SNAP protéines sont structurellement et fonctionnellement différentes de protéines SM et participent à la dissociation du complexe SNARE avec NSF ATPase. Tous les groupes de covariants se sont révélés être limité à une région particulière de la protéine. Beaucoup de résidus ayant une importance structurale et fonctionnelle connue ont également été identifiés. La canalisation d'analyse améliorée a été appliquée à nouveau sur les protéines SM. Les différences entre les groupes de résidus covariants ont été observées chez les protéines SM, chez les protéines tels que les enzymes et chez les SNAP. Les plus résidus covariants dans les groupes étaient allongés les uns près des autres, alors que certains étaient couchés et éloignés les uns des autres. Les résidus éloignés sembleraient être relié et sembleraient ainsi former la chaîne dans la structure 3D de la protéine. Le modèle de comparaison des résidus covariants se trouvant au sein de la structure 3D a été observée pour toutes les sous-familles de protéines SM analysées, suggérant ainsi que les régions similaires des différentes sous-familles covarient. Une analyse inter-protéines de Sly1 et Syntaxin 5 paire aussi délectait amas de résidus de loin couché qui est apparu pour former des réseaux reliant les différentes régions de la protéine. Dans l'ensemble, les résultats suggèrent une possible communication entre les deux sites de liaison impliquant un réseau de résidus de covariables le long de la structure des complexes SM / de Syntaxin. Ainsi, l'approche du regroupement des données de covariation et la détection des groupes des résidus covariants ont permit d'identifier des nouvelles caractéristiques structurales et fonctionnelles concernant les protéines SM et les protéines SNAP. La future portée de ce travail pourrait comprendre la réalisation d'études mutationnels et biochimiques pour tester l'importance et la nature allostérique des sites covarying identifiés.

Contents

Acknowledgement i

Abstract iii

Résumé v

List of Figures ix

List of Tables xi

List of abbreviations xiii

1 Introduction 1
- 1.1 Intracellular membrane trafficking 1
  - 1.1.1 The SNARE protein family 3
  - 1.1.2 Sec1/Munc18 (SM) protein family 5
    - 1.1.2.1 Interaction of SM proteins with SNAREs 7
    - 1.1.2.2 Interaction of SM proteins with other proteins 10
  - 1.1.3 SNARE complex disassembly 12
  - 1.1.4 A previous phylogenetic analysis showed comparable patterns of Qa SNAREs and SM proteins changes in metazoans 17
- 1.2 Sequence analysis 19
  - 1.2.1 Molecular Covariation – Current Theories and Hypothesis 21
  - 1.2.2 Complexity of covariation 22
  - 1.2.3 Molecular Covariation detection methods – Different strategies and controversies 23
    - 1.2.3.1 Mutual Information (MI) 25
    - 1.2.3.2 Corrected Mutual Information (MIp) 25
  - 1.2.4 Groups of covarying residues 25

2 Aims of the project 27

3 Materials & Methods 29
- 3.1 Sequence Alignments 29
- 3.2 Phylogenetic Tree Generation 29
- 3.3 Generation of Simulated Alignments 30
- 3.4 Covariation Analysis 34
- 3.5 Network Analysis 36
- 3.6 Scoring and ranking the clusters 37
- 3.7 Visualizations 38
- 3.8 Analysis Pipeline 40

4 Results 43
- 4.1 Initial Covariation analysis on the SM protein Sly1 44
- 4.2 Development of an improved analysis approach for detecting groups of covarying residues and its application on the test cases from prior publications 49
  - 4.2.1 Test case 1: The Methionine amino peptidase1 (MAP1) contains two types of covarying positions 49
  - 4.2.2 Test case 2: UDG (Uracil-DNA glycosylase) data set from Lui & Bahar in 2012 59
- 4.3 Application of the improved analysis pipeline on SNAP proteins 64
- 4.4 Application of the improved analysis pipeline for intra-protein analysis of SM proteins 75
- 4.5 Application of the improved analysis pipeline for inter-protein analysis of Sly1 and Syntaxin 5 83
- 4.6 Intra-protein analysis of other SM protein subfamilies 95

5 Discussion ... **99**
5.1 Covariation analysis by MIp and average linkage clustering ... 101
5.2 Application of MIp and average linkage clustering on α-SNAP proteins identified networks of functionally and structurally important residues in N- and C-terminal regions and in TPR region. ... 105
5.3 Application of MIp and average linkage clustering on SM protein identified networks of covarying residues. ... 108
5.4 Application of MIp and average linkage clustering for inter-protein analysis on Sly1 and Syntaxin 5 proteins identified complex network of functionally and structurally important inter- and intra-protein residues. ... 113
6 Conclusions & Perspective ... **117**
A Appendix : Application note ... **121**
Alifea– Alignment Feature analysis tool ... 121
A.1 Alignment Preprocessing ... 122
A.2 Sequence Analysis ... 123
A.3 Subtype Specific analysis ... 124
A.4 Covariation Analysis ... 124
A.5 Network Analysis Algorithm ... 129
A.6 Scoring and ranking the clusters ... 130
A.7 Visualizations ... 130
References ... **131**
Contributions ... **143**
Curriculum Vitae ... **145**

List of Figures

Figure 1.1: Different steps of intracellular vesicular transport. 2
Figure 1.2:Domain compositions of SNARE subfamilies showing the SNARE motif flanked by the N-terminal domain and the C-terminal transmembrane domain. 3
Figure 1.3: Model of SNARE cycle during vesicle docking and fusion. 5
Figure 1.4: Schematic summary of known mammalian SNARE complexes and SM proteins and their sites of action along the exocytic and endocytic pathways. 7
Figure 1.5: Topology of SM protein, Munc18a of *Rattus norvegiucus* 7
Figure 1.6: Interaction of SM and SNARE proteins with tethering factors. 11
Figure 1.7: Representation of SNARE complex disassembly. 13
Figure 1.8: Model of SNF-mediated SNARE complex disassembly derived from cryo-EM structures generated in the Zhao et. al. 2015. 13
Figure 1.9: Unrooted phylogenetic tree of SNAP protein from various eukaryotic lineages 14
Figure 1.10: Structure of *Saccharomyces cerevisiae* α–SNAP, Sec17 (PDB code 1QQE) 16
Figure 1.11: Structure of monomer A of the γ -SNAP of Brachydanio rerio (PDB code 2IFU) 16
Figure 1.12: Schematic depiction of the evolutionary changes of the SNARE and SM proteins in animals. 17
Figure 1.13: Information extracted from multiple sequence alignments (MSAs). 19
Figure 3.1 : Heat map of MIp scores from the simulated alignment. 32
Figure 3.2: Heatmap of MIp scores of simulated coevolving residues. 33
Figure 3.3: Schematic representation of the workflow of the improved analysis pipeline. 41
Figure 4.1: Covarying residues with Z-score > 4 marked on structure of *Saccharomyces cerevisiae* Sly1 in complex with the N-peptide of Sed5 (1MQS). 45
Figure 4.2: Paired residues with Z-score>4, marked on the structure of *Saccharomyces cerevisiae* Sly1 in complex with the N-peptide of Sed5 (1MQS). 46
Figure 4.3: An example of residues with Z-score>4 appearing as groups, marked on the structure of *Saccharomyces cerevisiae* Sly1 in complex with the N-peptide of Sed5 (1MQS). 47
Figure 4.4 : Some of the residues with high MIp score mapped on the structure of *Escherichia coli* MAP1 (1C24). 50
Figure 4.5: Heat map of MIp scores of MAP1 dataset from (Gloor et al. 2005). 52
Figure 4.6 : Residues identified in the current analysis and in the previous study obtained for MAP1 alignment. 55
Figure 4.7: Residue from the selected clusters marked on the structure of *Escherichia coli* MAP1 (1C24). 55
Figure 4.8: Heatmap of MIp scores arranged according to the ranking of clusters from the analysis on MAP1 dataset from Gloor et al. 2005. 57
Figure 4.9: Network of residues of cluster 1 from MAP dataset. 58
Figure 4.10: Residues from the top 90% clusters mapped onto the structure of human UDG (PDB id: 4SKN). 61
Figure 4.11: Network of residues of cluster 1 from the UDG dataset. 62
Figure 4.12 : Heatmap of MIp scores arranged according to the ranking of clusters from the analysis on the UDG dataset. 63
Figure 4.13:Sequence logo of the entire SNAP alignment. 65
Figure 4.14: Heat map of MIp scores from the entire SNAP alignment. 66
Figure 4.15: Some of the residue pairs from the top 50 covarying residue pairs of the α-SNAP analysis. 68
Figure 4.16: Some of the residue pairs from the top 50 covarying residue pairs of the α-SNAP analysis. 68
Figure 4.17: Heatmap of MIp scores from α-SNAP alignment arranged according to the ranking of clusters. 69
Figure 4.18: A magnified portion of heat map MIp scores from α-SNAP alignment arranged according to the ranking of clusters. 70
Figure 4.19: Residue from the selected clusters marked on the structure of Saccharomyces cerevisiae α-SNAP (Sec17, PDB code 1QQE). 70
Figure 4.20: Network of residues of cluster 1 from α-SNAP analysis. 72

Figure 4.21: Helical wheel view of residues of cluster 1 from α-SNAP analysis. 72
Figure 4.22: Network view of three different clusters from α-SNAP analysis. 73
Figure 4.23: Highest scoring pair 3<--> 4 from cluster 1 mapped on phylogenetic tree of α-SNAP. 74
Figure 4.24: Heat map of MIp scores of Sly1 data...77
Figure 4.25: Heatmap of MIp scores from the Sly1 dataset, arranged according to the ranking of clusters. 78
Figure 4.26: : Residue from the top 90% clusters marked on the structure of *Saccharomyces cerevisiae* Sly1 (1MQS). 79
Figure 4.27: Network view of cluster1 of the Sly1 dataset. 80
Figure 4.28: Network view of cluster1 of the Sly1 dataset after removing low scoring edges...............80
Figure 4.29: Network view of first group of high scoring residues from cluster1 of Sly1 dataset. 81
Figure 4.30: Network view of second group of high scoring residues from cluster1 of Sly1 dataset.....81
Figure 4.31: Residue from the two subclusters marked on the structure of *Saccharomyces cerevisiae* Sly1 (1MQS). 82
Figure 4.32: Heat map of the MIp scores of the Syntaxin 5 data. 84
Figure 4.33: Heatmap of the MIp scores of the Syntaxin 5 dataset, arranged according to the ranking of clusters. 85
Figure 4.34: Residue from the top 90% clusters marked on the homology model of the Sly1/Syx5 structure. 86
Figure 4.35: Heatmap of MIp scores of the concatenated alignment of Sly1 and Syntaxin5 proteins. ..87
Figure 4.36: Residue from the top 90% clusters marked on the structure obtained by homology model of the Sly1/Syx5 structure. 89
Figure 4.37: Residue from the inter-protein cluster2 marked on the structure obtained by homology model of the Sly1/Syx5 structure. 90
Figure 4.38: SNARE motif of Syntaxin 5 from rat and yeast. 90
Figure 4.39: Residue from the inter-protein cluster2 marked on the structure obtained by overlapping 1MQS (of Sly1) and homology modeled structure of Syntaxin 5 on SSO1. 91
Figure 4.40:Network view of inter-protein cluster2 after removing low scoring edges. 92
Figure 4.41: Residue from the inter-protein cluster2 marked on the structure obtained by overlapping 1MQS (of Sly1) and homology modeled structure of Syntaxin 5 on SSO1. 92
Figure 4.42: Weblogo of Syntaxin5 SNARE motif. 93
Figure 4.43: Weblogo of Sly1 specific loop region. 93
Figure 4.44: Residue 258 of Syntaxin5 from inter-protein cluster2, mapped on the phylogenetic tree of Syntaxin 5. 94
Figure 4.45: Residue 550 of Sly1 from inter-protein cluster2, mapped on the phylogenetic tree of Sly1. 95
Figure 4.46: Residue from the top 85% clusters marked on the structure of *Rattus norvegicus* Munc18-1 (3C98). 96
Figure 4.47: Residues from the top 90% clusters marked on the structure of *Rattus norvegicus* Vps45 homology modelled on Munc18-1 of *Rattus norvegicus*. 97
Figure 4.48: Residue from the top 85% clusters marked on the structure of *Chaetomium thermophilum* Vps33 (5BUZ). 97
Figure 5.1: Sequence logo of residues from Munc18/Sec1 alignment lying near or within the N-peptide binding region. 112

List of Tables

Table 4.1: Top 6 residue pairs with Z-score>4 from Sly1 MIp data. 46
Table 4.2: Table of residues appearing as groups, identified in the initial analysis on Sly1. 47
Table 4.3 : Top 50 covarying residues ranked according to their MIp scores..51
Table 4.4: Top 50 covarying residues ranked according to their MIp scores from UDG dataset. 60
Table 4.5: Top 50 covarying residues ranked by their MIp scores from the α-SNAP dataset. 67
Table 4.6: Distribution of highest scoring pair 3<--> 4 from cluster 1 of α-SNAP. 75
Table 5.1: Maximum and minimum MIp scores obtained for different protein families/subfamilies studied. 104
Table 5.2 : Distribution of covarying residues identified in the N-peptide binding region in different lineages..112

List of abbreviations

SNARE	Soluble N-ethylmaleimide-sensitive factor Attachment protein Receptors
SM	Sec1/Munc18
NSF	N-ethylmaleimide-Sensitive Fusion
SNAP	Soluble NSF Association Proteins
AAA+	ATPase Associated with diverse cellular Activities
COPI	Coat protein complex I
COPII	Coat protein complex II
CATCHR	Complex Associated with Tethering Containing Helical Rods
Munc18	Mammalian uncoordinated protein 18
GARP	Golgi-associated retrograde protein
COVERT	Core vacuole/endosome tethering
HOPS	HOmotypic fusion and vacuole Protein Sorting
MAP	Methionine amino peptidase
UDG	Uracil-DNA glycosylase
TGN	Trans Golgi network
TM	Trans-membrane
ER	Endoplasmic reticulum
MTC	Multisubunit tethering complexes
Vps	Vacuolar protein-sorting
TPR	Tetra-tricopeptide repeats
MSA	Multiple sequence alignment
McBASC	McLachlan based substitution correlation
SCA	Statistical coupling analysis
MI	Mutual information
MIp	Corrected mutual information
OMES	Observed versus expected frequencies of residue pairs
ELSC	Explicit Likelihood of Subset Variation
DCA	Direct coupling analysis
PSICOV	Protein sparse inverse covariation
HMM	Hidden Markov Model
CLAG	CLusters AGgregation
BIS	Blocks In Sequences
rmsd	Root mean square deviation
Cryo-EM	Cryo-electron microscopy

1 Introduction

1.1 Intracellular membrane trafficking

A salient feature of eukaryotic cells is the presence of a complex system of intracellular membrane delimited compartments, called organelles. These organelles are biochemically distinct and exchange material through vesicle trafficking. Vesicles are small membrane bound carriers that bud from a donor organelle and then fuse specifically with an acceptor organelle. This enables cells to take up nutrients and transport soluble molecules and membrane components. Thus, vesicle trafficking is an essential process of the eukaryotic cell.

The newly synthesized proteins and lipids are transported within the cell through the exocytic pathway. Exocytosis is a process in which the secretory vesicle fuses with the plasma membrane and releases its contents in the extracellular space or to put lipids and protein into the plasma membrane. Many essential processes in our body such as signaling, insulin secretion by pancreas, secretion of neurotransmitters, etc. are dependent on it. Constitutive exocytosis occurs in all cells and helps to release the newly synthesized membrane proteins that can be incorporated in the plasma membrane. Regulated exocytosis occurs in some specialized cells where the vesicles require a stimulus or signal for fusion with the plasma membrane. For example, in neurons, the vesicle filled with neurotransmitters fuses with the presynaptic plasma membrane upon the rise in intracellular Ca^{2+} concentration. The Ca^{2+} influx occurs because of the opening of the voltage-gated calcium channels upon the arrival of action potential in neurons. Conversely, the cells take up the nutrients through the endocytic pathway. Endocytosis is a process by which a cell retrieves the proteins and other molecules from the plasma membrane.

Each of the vesicular transport reactions can be divided into four major steps. These steps include budding, transport, tethering and fusion (Bonifacino & Glick 2004; Cai et al. 2007) (Figure 1.1). Over the last decades, a large amount of work has been performed to identify the molecular machines involved in these processes. Each step is mediated by conserved homologous sets of protein machineries. The vesicle budding step, during which a vesicle buds form the donor organelle is mediated by vesicle coat proteins, like the COPI, COPII and Clathrin complexes. The molecular motor proteins, such as microtubules or actin, transport the vesicle to their final destination along the cytoskeletal tracks. The vesicles are then specifically tethered to their target organelle with the help of tethering proteins belonging to the CATCHR (Complex Associated with Tethering Containing Helical Rods) family and Rab proteins (Stenmark 2009). These proteins play an important role in determining the specificity of the target membrane. The final step of vesicle fusion is catalyzed by a protein machinery that involves SNARE (Soluble N-ethylmaleimide-sensitive factor Attachment protein REceptors) proteins (Jahn & Scheller 2006) and SM (Sec1/Munc18) proteins (Toonen & Verhage 2003).

Currently, it is becoming clear that the molecular machineries involved in the major steps of vesicular trafficking are highly conserved, not only between different species but also between different vesicle trafficking steps within the cell. Most probably they arose by duplication and diversification of prototypic protein machineries during evolution. This suggests that the proto-eukaryotic cell was already equipped with the various compartments and the vesicle transport machinery found in contemporary cells.

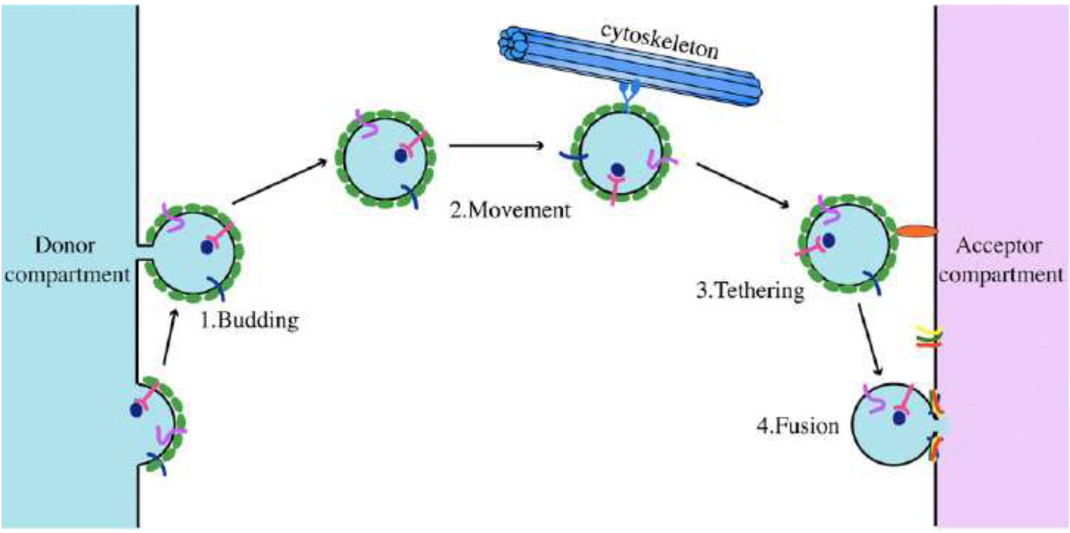


Figure 1.1: Different steps of intracellular vesicular transport.
The vesicle buds from the donor compartment (Budding), moves along tethers and finally fuses with the receptor compartment. Modified from (Cai et al. 2007).

1.1.1 The SNARE protein family

The members of the SNARE proteins family form the core machinery for each intracellular vesicle fusion process (Wickner & Schekman 2008; Sudhof & Rothman 2009; Jahn & Fasshauer 2012). These are relatively small, mostly tail-anchored, cytoplasmically oriented membrane proteins. Their defining feature is the presence of a so-called SNARE motif. It is conserved domain of 60-70 amino acids, arranged in heptad repeats and connected to the C-terminal transmembrane domain by a short linker (Figure 1.2).

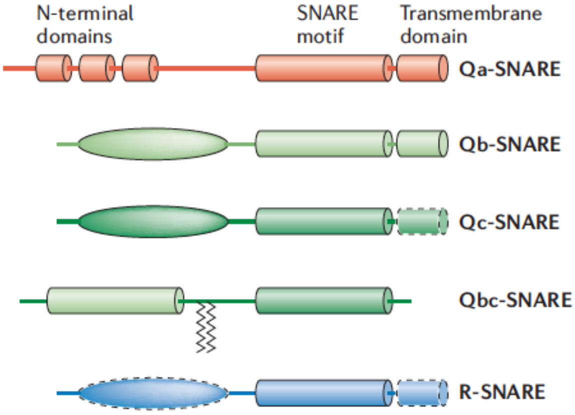


Figure 1.2:Domain compositions of SNARE subfamilies showing the SNARE motif flanked by the N-terminal domain and the C-terminal transmembrane domain.
Qa SNAREs have a short N-terminal peptide and a three-helical domain (Habc domain). The three-helical domain can also be found in members of Qb and Qc SNAREs. Qbc SNAREs represents a small subfamily of SNAREs-the SNAP-25 subfamily, that contains an N-terminal Qb and a C-terminal Qc SNARE motif that are interconnected by a linker region. The linker often carries a cysteine stretch that is known to be palmitoylated (zig-zag lines) and serves as a membrane anchor. Most R-SNAREs have an N-terminal profilin/longin domain and a transmembrane anchor. Ykt6 is an exception as it does not contain the longin domain and the transmembrane domain is replaced by farnesylated CAAX box. Adapted from (Jahn & Scheller 2006).

The SNARE proteins are present either on the vesicle membrane or on the target organelle membrane. They assemble into a tight four-helix bundle complex between opposing membranes. This complex assembly occurs in a zipper-like fashion from the N-terminus towards the C-terminus of the SNARE motif, thereby pulling the opposing membranes into close proximity (Figure 1.3), resulting into fusion. It is believed that the formation of a tight four-helix complex between SNAREs from the opposing membranes provides the energy to drive membrane fusion. Initially, the SNAREs were classified as v-SNARE and t-SNARE depending on their association with either the vesicle membrane or the target membrane. However, very similar crystal structure of three distantly related SNARE complexes (Sutton et al. 1998; Antonin et al. 2002; Zwilling et al. 2007) revealed high degree of structural similarity, suggesting that they all form complexes consisting of an elongated parallel four-helix bundle. The interior of the bundle contains 16 (−7 to +8) layers of highly conserved, mostly hydrophobic residues, except for a central ('0') hydrophilic layer that contains three conserved glutamine (Q) residues and one conserved arginine (R) residue. These highly conserved 0-layer residues lead to the classification of SNAREs into Q- and R- SNAREs (Weimbs et al. 1997; Fasshauer et al. 1998). SNARE complex formation requires three different type of Q-SNAREs (Qa, Qb, Qc) and one R-SNARE. For example, in the neuronal SNARE core complex, Syntaxin1a contributes the Qa SNARE domain, SNAP-25 contributes the Qb and Qc-SNARE domain and Synaptobrevin contributes the R-SNARE domain. Later, a detailed phylogenetic classification of different SNAREs reflected their participation in different trafficking steps and divided them into four subclasses: group I, involved in trafficking towards the endoplasmic reticulum (ER; I), group II, involved in trafficking towards the Golgi apparatus (II), group III, involved in trans-Golgi network vesicle trafficking (TGN; III.a) and digestive endosomal compartment trafficking (III.b), and group IV, involved in secretion (IV) (Kloepper et al. 2007). The study revealed that in the cell distinct sets of SNARE proteins are present that work together in different trafficking steps. In each set, four basic types of SNAREs occur, reflecting their position in the four-helix complex. A basic set of SNAREs has already been existed in the least common eukaryotic ancestor or the proto-eukaryotic cell. This basic SNARE set then evolved in different lineages giving rise to distinct types found in all eukaryotes.

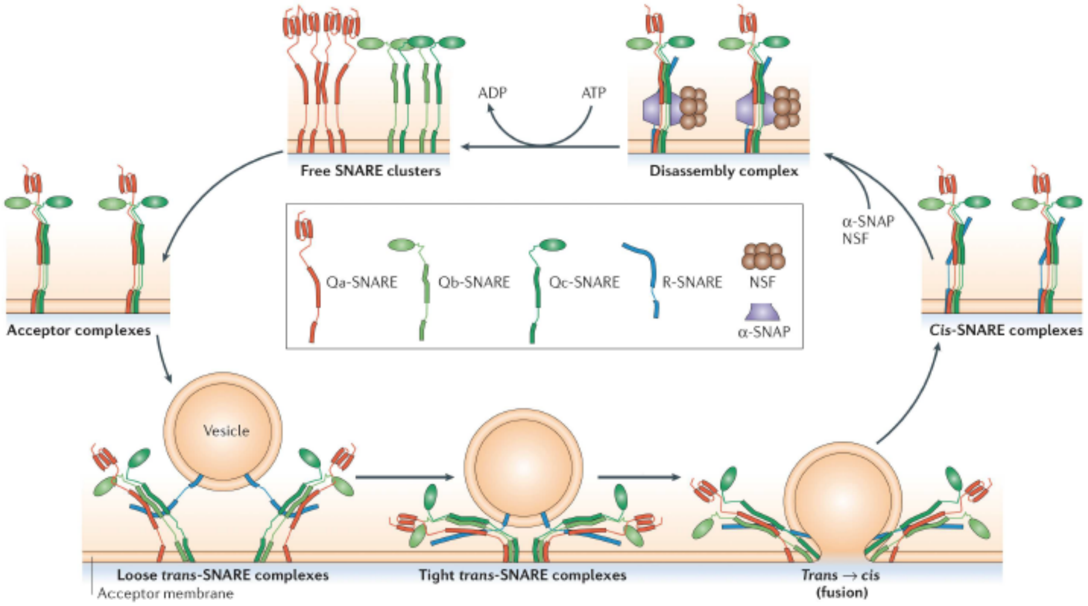


Figure 1.3: Model of SNARE cycle during vesicle docking and fusion.
The three Q-SNARE motifs from SNARE proteins present in clusters on the acceptor membrane, form the acceptor complex (Qabc). The vesicle R-SNARE interacts with the acceptor complex and forms a four-helical trans-complex (SNAREs anchored on opposing membranes). This zipper-like process starts at the N- terminus and assembles towards the transmembrane anchors at the C-terminus. This leads to opening of the fusion pore and fusion of vesicle with the plasma membrane, transferring the SNAREs into a cis-complex-configuration (SNAREs anchored on the same membrane). Disassembly of the cis-complex is managed by the AAA+ ATPase NSF together with its cofactor SNAP (soluble NSF attachment protein). Adapted from (Jahn & Scheller 2006).

1.1.2 Sec1/Munc18 (SM) protein family

Members of the SM protein family are essential regulatory factors in different vesicular trafficking steps (Jahn & Fasshauer 2012; Burkhardt et al. 2008) that genetically and biochemically interact with the core SNARE fusion machinery, specifically with the Syntaxin or Qa-SNARE, of that particular membrane fusion step. The SM proteins are a small family of cytosolic proteins of 60–70 kDa with moderate sequence homology (Halachmi & Z. Lev 1996). The functional importance of SM proteins in intracellular trafficking was shown by several independent genetic studies (Novick & Schekman 1979; Ossig et al. 1991; Harrison et al. 1994; Verhage et al. 2000). The first SM genes were independently identified in genetic screens of membrane trafficking mutants in *Caenorhabditis elegans* and yeast (Brenner 1974; Novick & Schekman 1979). Unc-18 in *Caenorhabditis elegans* and Sec1 in yeast, both were found to be critical for functioning. Loss-of-function mutations for nine SM genes in four species have been carried out which showed severe impairment of vesicle trafficking and fusion, most of which lead to a lethal phenotype (Toonen & Verhage 2003). These studies established SM proteins to be a central and indispensable factor in intracellular trafficking. However, the exact molecular mode of interaction of SM proteins is still not entirely understood.

According to the literature, the SM protein family contains four basic members: Sec1/Munc18, Sly1, Vps33, and Vps45 (Jahn & Südhof 1999; Chen & Scheller 2001) (Figure 1.4). Sec1/Munc18 regulate exocytosis by interacting with Qa SNAREs type IV. Sly1 functions in transport between the endoplasmic reticulum (ER) and Golgi apparatus and interacts wit Qa SNAREs of type II (Sed5 or Syntaxin5) and type I (Ufe1 or Syntaxin18). Vps33 is active in the endocytic/lysosomal sorting system and interacts with Qa SNARE of type III a, Vam3 or Syntaxin 7. Vps33 is also a part of the large HOPS protein complex. Vps45 acts in the trans-Golgi network and endosomal pathways and interacts with Qa SNAREs of type III b, Tlg2 or Syntaxin 16 (Chen & Scheller 2001). An unpublished phylogenetic analysis of the SM protein family in my laboratory has uncovered a fifth member of the SM protein family, for which no SNARE protein binding partner could be identified yet. Different rounds of duplication during vertebrate evolution have resulted in three homologs of Munc18 (Munc18a, Munc18b and Munc18c) and two homologs of Vps33 (Vps33a and Vps33b).

SM proteins from different sub-types show high sequence variation, for example, Sec1, Sly1, Vps33 of yeast have only 22-23 % sequence identity, however, they have a high structural similarity as can be seen by comparing the different solved X-ray structures of Sly1 from *Saccharomyces cerevisiae* (1MQS), Munc18-1 from *Rattus norvegicus* (3C98), Munc18-2 from *Homo sapiens* (4CCA), Munc18-3 from *Mus musculus* (2PJX), Unc18 from *Monosiga brevicolis* (2XHE), Sec1 from *Loligo pealei* (1FVF) and Vps33 from *Chaetomium thermophilum* (5BUZ). These crystal structures reveal that SM proteins have a highly conserved overall fold. SM proteins display an arch shaped structure with a large central cavity and 3 domains (domain 1-3). The third domain is a large insertion between the third and fourth parallel strands (b9 and b12) of domain 2 (Figure 1.5).

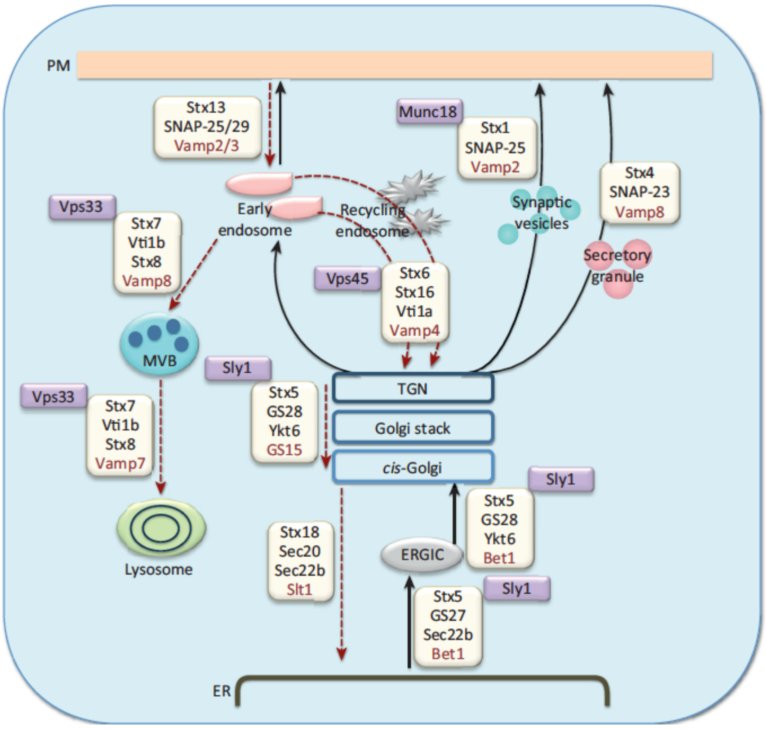
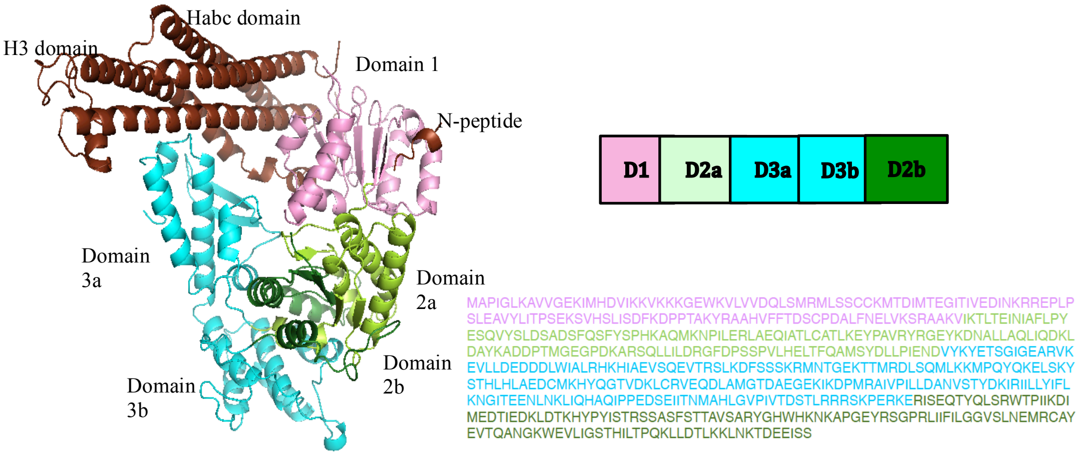


Figure 1.4: Schematic summary of known mammalian SNARE complexes and SM proteins and their sites of action along the exocytic and endocytic pathways.
Potential v-SNAREs are indicated in red; the SM proteins are indicated in purple. ER: endoplasmic reticulum; ERGIC: ER–Golgi intermediate compartment; MVB: multivesicular bodies; PM: plasma membrane; TGN: trans-Golgi network. Adapted from (Hong & S. Lev 2013).



MAPIGLKAVVGEKIMHDVIKKVKKKGEWKVLVVDQLSMRMLSSCCKMTDIMTEGITIVEDINKRREPLP
SLEAVYLITPSEKSVHSLISDFKDPPTAKYRAAHVFFTDSCPDALFNELVKSRAAKVIKTLTEINIAFLPY
ESQVYSLDSADSFQSFYSPHKAQMKNPILERLAEQIATLCATLKEYPAVRYRGEYKDNALLAQLIQDKL
DAYKADDPTMGEGPDKARSQLLILDRGFDPSSPVLHELTFQAMSYDLLPIENDVYKYETSGIGEARVK
EVLLDEDDDLWIALRHKHIAEVSQEVTRSLKDFSSSKRMNTGEKTTMRDLSQMLKKMPQYQKELSKY
STHLHLAEDCMKHYQGTVDKLCRVEQDLAMGTDAEGEKIKDPMRAIVPILLDANVSTYDKIRIILLYIFL
KNGITEENLNKLIQHAQIPPEDSEIITNMAHLGVPIVTDSTLRRRSKPERKERISEQTYQLSRWTPIIKDI
MEDTIEDKLDTKHYPYISTRSSASFSTTAVSARYGHWHKNKAPGEYRSGPRLIIFILGGVSLNEMRCAY
EVTQANGKWEVLIGSTHILTPQKLLDTLKKLNKTDEEISS

Figure 1.5: Topology of SM protein, Munc18a of *Rattus norvegiucus*.
Different domains of Munc18a are color-coded and are also shown in a separate bar. Binding partner Syntaxin1 is shown in brown. The displayed protein sequence of *Rattus norvegiucus* of Munc18a is also color-codded as per the domains.

1.1.2.1 Interaction of SM proteins with SNAREs

Specific SM proteins are thought to interact with Qa SNAREs of particular membrane fusion steps. The exact molecular role of SM proteins, in particular their binding mode with syntaxins, is still debated as different binding modes were proposed for the different SM proteins (Toonen and Verhage, 2003). More recently, biochemical and structural data suggest that SM proteins and syntaxins generally utilize two spatially separated binding sites. Generally, SM proteins seem to interact with their cognate Syntaxin via two different binding surfaces, the "closed" conformation in which the SNARE (H3) domain folds back onto the Habc domain, such that it is inaccessible for SNARE complex formation and the very N-terminal region of Syntaxin, called the 'N-peptide'.

The mammalian isoform of Sec1, Munc18a was initially shown to bind to syntaxin1 in the closed conformation and thus to block SNARE assembly (Misura & Weis 2000; Pevsner et al. 1994; Chen & Scheller 2001). In order to form a SNARE complex, Syntaxin 1 must dissociate from the tight grip of Munc18a and switch to open conformation (Burkhardt et al. 2008; Pevsner et al. 1994). By contrast, Sly1 was found to bind with nanomolar affinity to another binding site of Sed5, the N-peptide (Yamaguchi et al. 2002; Kosodo 2002; Peng & Gallwitz 2004). Consistent with the high affinity interaction, the crystal structure showed that the N-peptide binds via an extensive interface to the outer surface of Sly1 (Bracher & Weissenhorn 2002). A similar binding was also shown for Vps45 and the TGN Syntaxin Tlg2p (Dulubova 2002). These SM proteins were believed not to be interacting with the remaining of Syntaxin. It was thus suggested that this mode of binding somehow assists in the SNARE complex formation, rather than inhibiting it (Dulubova 2002; Peng 2002; Peng & Gallwitz 2004; Carpp et al. 2006; Furgasona et al. 2009). This mode of interaction is also supported by the notion that that Sly1 can stay bound during SNARE complex formation (Peng 2002). Such an idea was strengthened by the observation that yeast Sec1 interacts with the assembled secretory SNARE complex but not with its isolated Qa-SNARE (Carr et al. 1999; Scott 2004; Togneri et al. 2006). Unfortunately there is so far no structural confirmation for the interaction of Sec1 with its cognate SNARE.

Thus, a decade ago, the general consensus in the field regarding the interaction of SM protein with Syntaxin was that SM proteins function by binding to the N-terminal peptide region of their partner syntaxins. Munc18a binding involving only the close conformation of Syntaxin was considered to be a special adaptation for neuronal exocytosis.

Recent biochemical and structural studies have shed new light into this discrepancy by showing a possible common mode of interaction, involving two spatially separated binding sites for SM proteins and Syntaxins (Burkhardt et al. 2008; Demircioglu et al. 2014). A refined crystal structure of Munc18a-Syntaxin1a complex from rat revealed that Munc18a could also bind to the second spatially separate N-peptide of Syntaxin1 although with a lower affinity (Burkhardt et al. 2008) (PDB ID: 3C98). This observation was similar to the interaction of Sly1 with Sed5 N-peptide. Interestingly, a so-called open mutant of Syntaxin, containing a double mutation L165A/E166A in the linker region, binds tightly to Munc18a and allows the SNARE complex formation as well (Dulubova et al. 1999). Thus, the N-peptide of Syntaxin acts as a switch controlling Munc18a to either bind to a closed conformation of Syntaxin and thereby inhibiting the SNARE complex formation or allowing the bound Syntaxin to form the SNARE complex. Interestingly, both N-peptide and closed conformation binding modes were also observed in the crystal structure of Munc18-Syntaxin1 complex of the choanoflagellate *Monosiga brevicollis* (Burkhardt et al. 2011). Choanoflagellates are a group of single-celled eukaryotes that are thought to be the closest living relatives of metazoans. This suggests that the binding mode involving two different binding sites could be evolutionarily conserved. A comparable binding mode was shown for the trans-Golgi network trafficking SM protein Vps45. Vps45 interacts with Tlg2p lacking the N-peptide, possibly in closed conformation (Furgasona et al. 2009; Burkhardt et al. 2008). Similar results were found for Sly1p, which was shown to interact not only with the N-peptide but also with the remainder of Sed5, making use of both binding modes (Demircioglu et al. 2014).

Recent biochemical and structural studies on Munc18a/Syntaxin1 (Khvotchev et al. 2007; Burkhardt et al. 2008), Vps45/Syntaxin16 (Carpp et al. 2006), Vps45/Tlg2p (Furgasona et al. 2009; Burkhardt et al. 2008), Munc18b (Hackmann et al. 2013) Munc18c/Syntaxin 4 (Aran et al. 2009), Unc18 (Johnson et al. 2009) have shown that SM proteins generally bind to their respective Syntaxin using both modes of interaction, however with different relative binding affinities. For example, mammalian Munc18a binds to the closed conformation of Syntaxin 1 with a higher affinity but has a weak interaction with the N-peptide (Burkhardt et al. 2008). This suggests that the SM proteins share a common mode of interaction with Syntaxin involving the two modes of interaction, although with different binding affinities. Also, different trafficking steps may favor one over the other.

However, not all SM proteins interact with N-peptide of their cognate Syntaxin. For example, the crystal structure of Vps33 from *Chaetomium thermophilum* (Baker et al.

2013) and of the human homolog Vps33A (S. C. Graham et al. 2013), showed that the N-peptide binding pocket in Vps33 is blocked. Vps33 is also a part of the multisubunit-tethering complex, the homotypic fusion and protein-sorting (HOPS) complex. Despite having a blocked N-peptide binding pocket and a low sequence similarity, the overall structure of Vps33 is very similar to other SM proteins (Baker et al. 2013; S. C. Graham et al. 2013).

It is possible that the occurrence of high structural conservation in SM protein is important to preserve a similar or conserved molecular function. Both binding sites seem to be crucial for the function of the complex in vesicle fusion, but it is still unclear how the two binding sites are able to communicate. The existence of a possible conformational switch between the two binding (open and close) interactions that enables the SM protein to control the accessibility of the bound Syntaxin, is still not clear.

1.1.2.2 Interaction of SM proteins with other proteins

Many other recent biochemical studies also suggest that SM proteins interact with a number of accessory proteins also known as tethering factors. Tethering factors are multiple protein complexes that help the vesicles to link or tether to their target membranes. They can be divided into two major groups, homodimeric long coiled-coil proteins and multisubunit tethering complexes (MTCs). MTCs can further be divided into two major classes, complexes associated with tethering containing helical rods (CATCHR) and Class C vacuolar protein-sorting (Vps) complexes. Multisubunit tethering complexes such as the exocyst, the Conserved Oligomeric Golgi (COG), the Golgi-associated retrograde protein (GARP) and DSL1 complex belong to the CATCHR family. Homotypic fusion and protein sorting (HOPS) and class C core vacuole/endosome tethering (CORVET) belong to the Class C Vps complexes. They all are involved in different vesicular trafficking steps within the cells and interact with the SNARE and SM proteins of those trafficking steps (Hong & S. Lev 2013).

COG subunit Cog4 interacts with Sly1 and Vps45 (Figure 1.6-A) (Hong & S. Lev 2013). Dsl1 complex that belong to the CATCHR protein family interacts with Sly1 on the endoplasmic reticulum (ER) membrane (Figure 1.6-B) (Hong & S. Lev 2013). The yeast SM protein Vps33 is a part of multisubunit tethering HOPS complex (Homotypic fusion and vacuole Protein Sorting) (Figure 1.6-C) (Hong & S. Lev

2013). The yeast exocyst complex subunits interact with Sec1 (Figure 1.6-D) (Hong & S. Lev 2013). The exocytotic SM protein Munc18 may work together with Munc13 to chaperone the assembly of neuronal SNARE complex (Ma et al. 2011; Ma et al. 2013; Rizo & Rosenmund 2008) (Figure 1.6-E). Possible interactions between Munc18a and accessory exocytic proteins like Mint1, granuphilin, Rab3, Doc2 and phospholipase D (Okamoto & Sudhof 1997; Coppola et al. 2002; Verhage et al. 1997; Dulubova et al. 1999; M. E. Graham et al. 2008) , have been put forward. However, the function and binding regions of many of these interactions remains poorly understood. Both, COG and Dsl1 complexes show structural homology to Munc13 (W. Li et al. 2011).

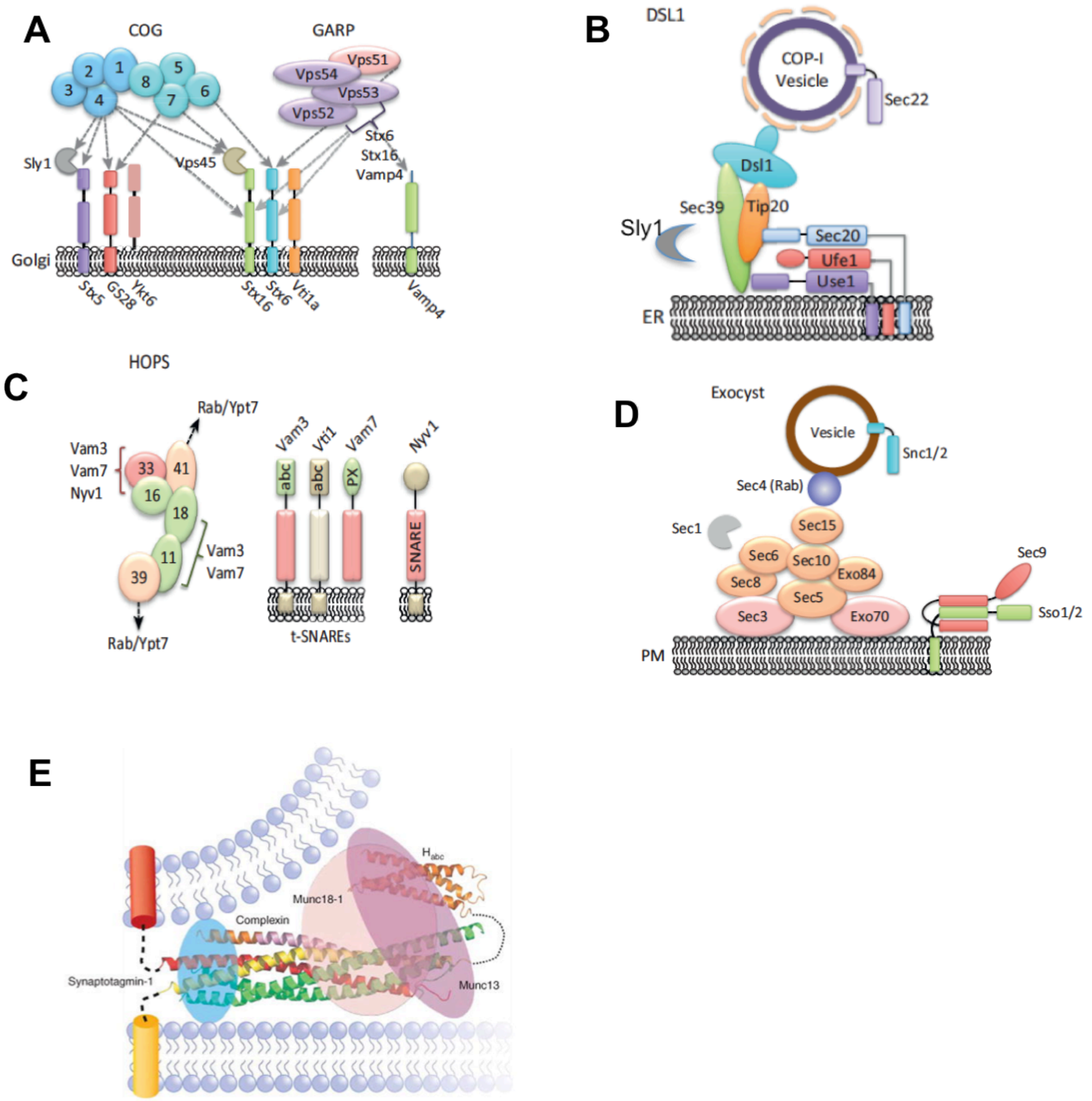


Figure 1.6: Interaction of SM and SNARE proteins with tethering factors.
A) A schematic representation of COG and GARP complex interaction with multiple SNAREs on the Golgi membranes. The light blue and dark blue circle represent the eight subunits of the COG complex organized into two structurally and functionally distinct lobes. The Cog4, Cog6, and Cog7 subunits bind the SNARE domains of the indicated SNAREs. Cog4 subunit also binds the Sec1/Munc18 (SM)

proteins Sly1 and Vps45. The Vps52, Vps53, and Vps54 subunits of GARP are shown in purple circles and Vps51 subunit in pink. The Vps51 subunit interacts with the N-terminal regulatory Habc domain of Syntaxin 6. The N-terminal regions of human Vps53 and Vps54 bind to the SNARE motifs of Stx6, Stx16, and Vamp4. B) A schematic representation of the yeast HOPS complex interaction with SNARE motifs (pink) of target membrane SNAREs (t- SNAREs) Vam3 and Vam7 and the transport vesicle SNARE (v-SNARE) Nyv1 via its Vps33 subunit (pink). The Vps11, Vps16, and Vps18 subunits (green) interact with the regulatory Habc and PX domains of the t-SNAREs Vam3 and Vam7 on the vacuole or lysosome membrane (green). Vps39 and Vps41 subunits (yellow) bind the small Rab GTPase Ypt7. C) A schematic representation of the yeast DSL1 complex showing its interaction with the t-SNAREs Sec20 and Use1 via its Tip20 and Sec39 subunits. The interaction induces t-SNARE gathering on the endoplasmic reticulum (ER) membrane. The Dsl1p subunit interacts with subunits of the COP-I coat and tethers COP-I vesicles to the ER membranes. D) A schematic representation of the yeast exocyst complex that tethers secretory vesicles to the plasma membrane (PM). The Sec3 and Exo70 (pink) subunits interact with the PM via their positively charged residues, whereas the Sec15 subunit interacts with the Rab GTPase Sec4, which localizes on a secretory vesicle. The interactions of the Sec6 subunit with the t-SNARE Sec9 and the SM protein Sec1 and of the Sec3 subunit with Sso1/2 regulate the assembly the Sec9–Sso1 t-SNARE complex that binds the v-SNARE Snc1/2. E) A schematic representation of interaction of Munc18-1, Mun13, complexin and synaptotagmin-1 with the SNARE complex. The ribbon diagram shows the structure of the complexin-I–SNARE complex. Figures A), B), C) and D) adapted from (Hong & S. Lev 2013) and E) adapted from (Rizo & Rosenmund 2008).

1.1.3 SNARE complex disassembly

After vesicle fusion, the SNARE complex needs to be disassembled to regenerate the SNAREs for consecutive rounds of fusion. The disassembly process is carried out by the AAA+ ATPase NSF (N-ethylmaleimide-Sensitive Fusion) protein together with its cofactor SNAP (Soluble NSF Attachment Protein) protein (Clary et al. 1990; Hanson et al. 1997; Sollner et al. 1993; Marz et al. 2003) (Figure 1.7). NSF is a member of AAA+ (ATPase Associated with diverse cellular Activities) superfamily. It is thought to form ring-shaped homohexamers that utilizes ATP hydrolysis for SNARE complex disassembly (Hanson et al. 2005). NSF consists of an N-terminal domain (N-domain) and two consecutive ATPase domains (D1 and D2) (Sollner et al. 1993; Zhao et al. 2015). Domain D1 is considered to be mostly responsible for ATPase activity, while domain D2 contributes mainly to the hexamerization. The N-domain of NSF is considered to be involved in binding with SNAP and possible binding with SNARE complex (Zhao et al. 2015; Hanson et al. 1997). SNAP acts as an adaptor protein between NSF and SNARE complexes as SNAREs do not have a direct binding site for NSF. SNAP binds to the SNARE complex and stimulates ATP hydrolysis of NSF, which causes conformational changes that induces the disassembly of the SNARE complex (Clary et al. 1990; Sollner et al. 1993; Hanson et al. 1997).

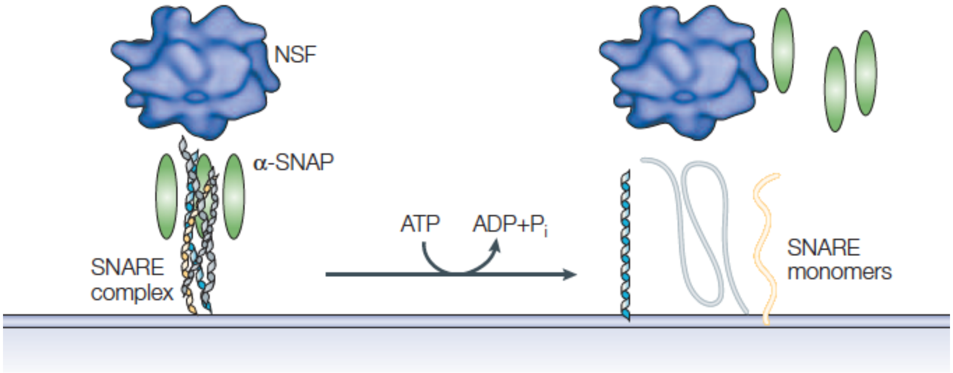


Figure 1.7: Representation of SNARE complex disassembly.
The AAA+ protein NSF, together with its adaptor protein SNAP, disassembles the SNARE complex formed during membrane fusion. The released individual SNAREs can then participate in new membrane-fusion events. The figure is modified from Hanson et al. 2005 and is based on cryo-electron-microscopy reconstruction of the structurally similar AAA+ protein.

Cryo-electron-microscopy (cryo-EM) structures from a recent study (Zhao et al. 2015) have suggested that depending on the SNARE complex composition, up to four molecules of SNAP can be present during the disassembly process. Based on the cryo-EM structure a working model for SNARE complex disassembly was proposed. The first step in the process of disassembly is the binding of SNAP to the SNARE complex (Figure 1.8-a). This involves association of SNAP with the membrane through a putative, conserved membrane attachment site (Figure 1.8-b), which could facilitate the binding to the SNARE complex (Winter et al. 2009). The N-domain of NSF then binds to SNAP, which stimulates the ATPase hydrolysis of NSF (Figure 1.8-c). This provides sufficient force to exert a torque that loosens the SNARE complex (Figure 1.8-d).

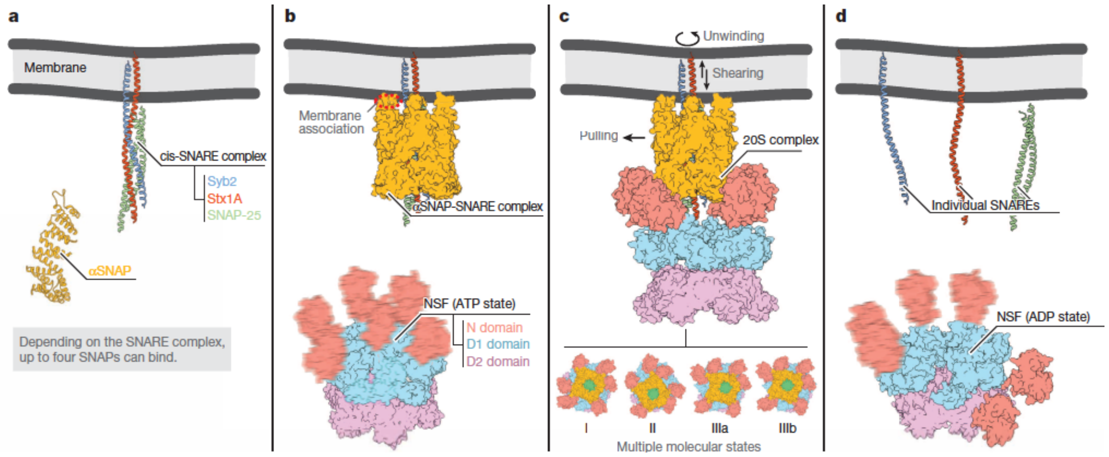


Figure 1.8: Model of SNF-mediated SNARE complex disassembly derived from cryo-EM structures generated in the Zhao et. al. 2015.
The model refers to the neuronal SNARE complex (consisting of synaptobrevin-2 (Syb2), syntaxin-1A (Stx1A), and SNAP-25) and α-SNAPs, but is applicable to other SNARE complexes as well. Figure taken from (Zhao et al. 2015).

Eukaryotes have three isoforms of SNAPs: α-, β-, and γ-SNAP (Clary et al. 1990), whereas fungi only posses the single α-SNAP homolog Sec17. α-SNAP, is a ubiquitous SNAP isoform, β-SNAP is a neuronal isoform, and γ-SNAP is strongly expressed in heart, kidney, liver and spleen (Bitto et al. 2007). Studies have shown that γ-SNAP can promote the intercisternal Golgi transport in-vitro, however, less efficiently than α-SNAP and β-SNAP (Clary et al. 1990).

Previous phylogenetic analysis carried out in my group (Kienle 2010) revealed two distinct groups with α-SNAP and β-SNAP in one group and γ-SNAP in another group. β-SNAP is a duplication of α-SNAP specific to vertebrates. γ-SNAP was found to be lost in Fungi except for some basal fungi species. The analysis shows that the duplication of SNAP into α-SNAP and γ-SNAP must be ancient, which suggests that these two SNAP proteins were already present in the assumed proto-eukaryotic ancestor (Figure 1.9). Multiple subtypes of SNARE proteins exists in eukaryotes, but there are only a few SNAPs and one NSF species (Zhao et al. 2015). It is still not known as how α-Snap recognizes the different SNAREs involved in different membrane fusion reactions.

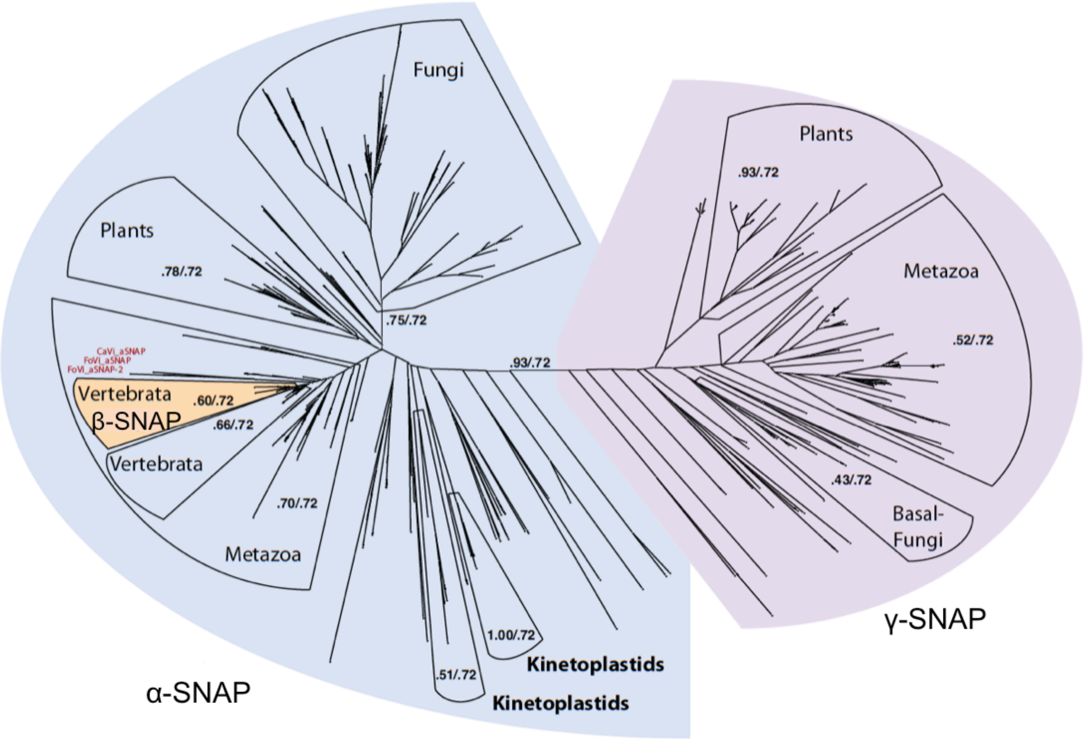


Figure 1.9: Unrooted phylogenetic tree of SNAP protein from various eukaryotic lineages.
The α-, β- and γ-SNAP branches are color-coded. The labels on the major branches represent the Likelihood Mapping (left) and AU support values (right). Figure modified from Dissertation of (Kienle 2010).

Several deletion mutagenesis and *in vitro* binding studies have been carried to understand the interaction between SNAPs and the SNAREs complex, as well as the interaction between SNAP and NSF. These studies showed 63 N-terminal and 37 C-terminal residues of α-SNAP to be essential for binding to the SNARE complex (Hayashi et al. 1995). Subsequent electron microscopy studies revealed that SNAP appears to coat the SNARE complex along its length (Hohl et al. 1998). The extreme C-terminal amino acids of α-SNAP have been shown to be required for its ability to stimulate the ATPase activity of NSF (Barnard et al. 1997). The N-terminal domain of NSF has been shown to be essential for the interaction between NSF, SNAP and the SNARE complex (Rice & Brunger 1999). The extreme C-terminal tail of γ-SNAP has also been shown to bind to NSF in absence of the SNARE bundle (Tani et al. 2003). The 23 N-terminal residues and 89 C-terminal residues of γ-SNAP also form a complex with Gaf-1/Rip11 (Tani et al. 2003). Evidences have also shown that alteration in expression levels of α-SNAP may be associated with certain pathological conditions. The ‘‘Hydrocephaly with Hop-gait’’ (HYH) missense mutation in the gene encoding α-SNAP, causes abnormalities in the apical protein localization and cell fate determination in neuroepithelial cells (Chae et al. 2004). The aforementioned membrane attachment site of α-SNAPs is located in the loop connecting the first two helices of the N-terminal region (Winter et al. 2009). Mutation of the two conserved residues (α-SNAPF-27S, F28S) in this loop did not support disassembly on the liposomes, suggesting that the loop accommodates the membrane attachment site (Winter et al. 2009). It was speculated that the membrane might be the first SNAP attachment site which could increase the local concentration of SNAP or induce a conformational change, thereby facilitating its binding to the SNARE complex (Winter et al. 2009).

The Sec17 (α-SNAP homolog in yeast) structure shows that it has 14 α-helices (Rice & Brunger 1999) (Figure 1.10). The N-terminal region has a twisted sheet of 9 α-helices (Rice & Brunger 1999). The C-terminal region has a globular bundle formed by 5 α-helices which are connected by loops of variable lengths (Rice & Brunger 1999). The N-terminal α-helices form tetra-tricopeptide repeats (TPRs) with the exception of α-helices α1and α2. TPR-containing proteins are considered to be rigid proteins that do not undergo large conformational changes upon ligand binding, but they do have some degree of flexibility (Cortajarena & Regan 2006). This typical arrangement creates a significant cleft on one face of the molecule, which provides a concave face whose curvature complements the convex surface of SNARE complex.

The concave face has more conserved residues than the convex one and is characterized by slightly basic charge distribution, whereas the convex face has a negative charge distribution (Rice & Brunger 1999). Mutation studies have also shown concave face to be the SNARE complex binding surface of α-SNAP (Marz et al. 2003).

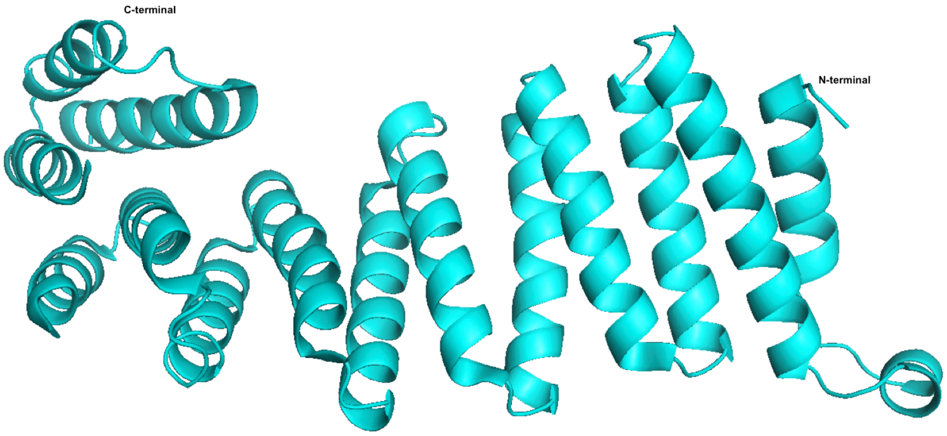


Figure 1.10: Structure of *Saccharomyces cerevisiae* α–SNAP, Sec17 (PDB code 1QQE)

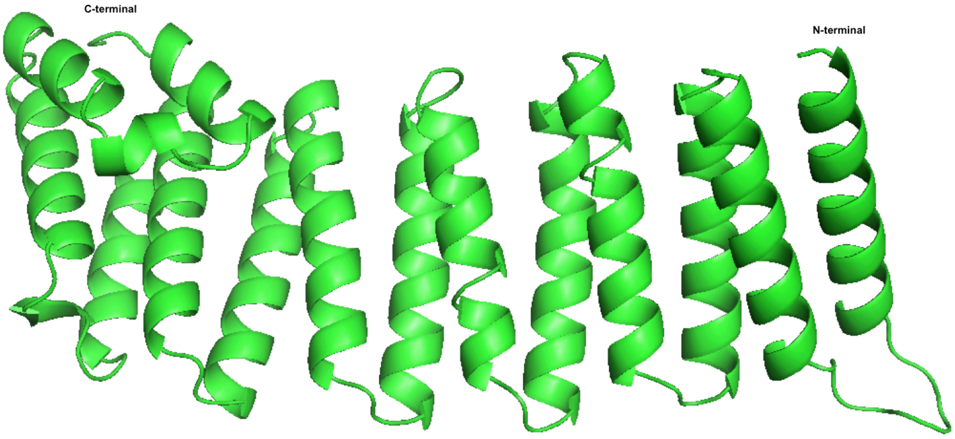


Figure 1.11: Structure of monomer A of the γ -SNAP of Brachydanio rerio (PDB code 2IFU)

The crystal structure of γ-SNAP from *Brachydanio rerio* (Bitto et al. 2007) revealed an elongated α-helical structure similar to Sec17, with 12 α-helices forming twisted antiparallel pairs at the N-terminal and 3 α-helices forming the globular C-terminal (Figure 1.11). The overall structural alignment of the Sec17 and γ-SNAP is poor with a root mean square deviation (rmsd) of 3.4Å (Bitto et al. 2007). This poor alignment is due to the differences in the twist from the middle portion of γ-SNAP (Bitto et al. 2007). However, the N-terminal portion and the C-terminal portion of γ-SNAP align well with the corresponding regions of the Sec17 (Bitto et al. 2007). Like Sec17, γ-SNAP possesses a similar charge distribution on the concave side (i.e., positively charged residues) (Bitto et al. 2007).

1.1.4 A previous phylogenetic analysis showed comparable patterns of Qa SNAREs and SM proteins changes in metazoans

A previous phylogenetic analysis established in my laboratory has shown that the current set of molecular machines involved in vesicle tethering and fusion arose by duplication and diversification of a prototypic protein machinery (Kloepper et al. 2007). It is therefore highly likely that the interacting proteins of the vesicle fusion apparatus may show common duplication and diversification patterns. A preliminary phylogenetic analysis of SNARE proteins and SM proteins already revealed comparable changes (Figure 1.12). For example, of the five different Qa-SNARE types, only the ones involved in endosomal/vacuolar trafficking and in secretion were duplicated and diversified in animals. Comparable patterns of change were observed for SM proteins, as only Vps33 (endosomal/vacuolar trafficking) and Sec1/Munc18 (secretion) proteins duplicated and diversified. However, it is unclear whether these similarities denote co-evolution between the interacting proteins.

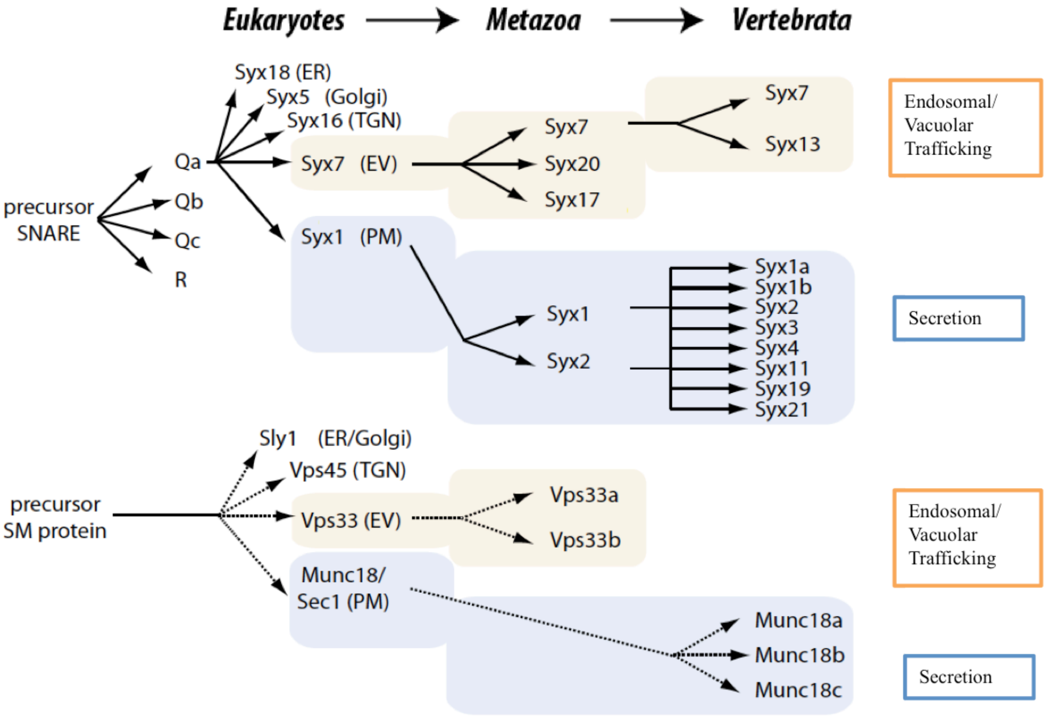


Figure 1.12: Schematic depiction of the evolutionary changes of the SNARE and SM proteins in animals.
The details of the evolution of Qa-SNAREs and SMs in metazoans are shown. Qa-SNARE and SM proteins that are involved in endosomal/vacuolar trafficking (yellow) and in secretion (blue) were duplicated and diversified in animals.

Many of the essential cellular processes (e.g. replication, transcription, translation, protein degradation, signal transduction and vesicular trafficking) are carried out by assemblies of 10 or more protein and ribonulceoproteins. These assemblies are often referred to as molecular machines. Often these machines have a catalytic core on which different layers of accessory factors bind. Some interactions in molecular machines are strong and long-lasting, whereas others are weak and highly transient. One can distinguish between very stable assemblies like the ribosome and dynamic machines that have to come together first to carry out their functions. The protein machinery involved in vesicle trafficking, for example, is different from other large protein machineries, such as ribosomes, as they interact transiently and assemble into intermediates.

Until now, the principal organization of the secretory and endocytic pathways has been established. However, a complete description of the vesicle pathway has not been achieved yet as there are clear differences between different eukaryotic lineages. Although a large amount of cell biological, genetic and biochemical work has been aimed to understand the underlying protein interaction network of vesicle trafficking; yet only fragmented insights are obtained. Often the biological processes are studied only in a few different model organisms, leaving it unclear if a certain feature is a special adaptation or a common principle. Bioinformatics investigations can fill in the gap as the entire spectrum of sequences and species can be integrated and investigated. Sequence analysis opens up new questions and new direction for further biochemical analysis. Deeper insights into the evolution of vesicle trafficking proteins and their changes within different eukaryotic lineages will be able to shed more insight on the molecular event.

Recent developments in sequencing technologies have led to an almost exponential growth of publicly available sequence data. From these, a wealth of information can be readily obtained by computational methods. Several sequence analysis methods are now available that can provide new insights about the function and interaction of various molecular machines. Sequence covariation analysis is one such approach that can help to identify the changes occurring within the protein in order to maintain its function and structure. This could help to recognize the structurally and functionally important sites within the protein that might also be involved in maintaining its interaction with different proteins. Novel insights about the vesicle fusion machineries and their changes within different eukaryotic lineages would provide explanations for different molecular events and would thus provide new directions for biochemical research.

1.2 Sequence analysis

Sequence analysis can be used to identify the residues that are under evolutionary constraints and have certain functional significance. Such residues can be conserved within the protein family or conserved within the subfamily or coevolve (Figure 1.12). Sequence analysis provides a novel insight into the evolution of protein family and the protein-protein interaction network. The experimental characterization of function and functionally important sites is usually expensive and time consuming. Sequence analysis can provide a starting point of a complete inspection for biochemical and structural analysis as the most interesting changes can later be tested. Several different methods have been developed to explore sequence space and thus predict functionally important regions.

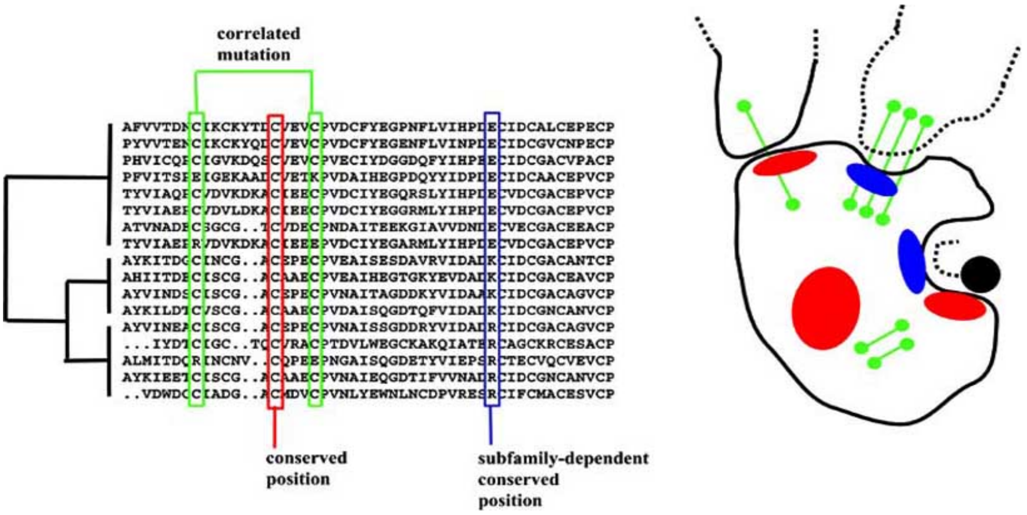


Figure 1.13: Information extracted from multiple sequence alignments (MSAs).
Left: fully conserved, family-dependent conserved positions and pairs of positions showing a correlated behavior are shown in a multiple sequence alignment. Right: An ideal model illustrating the relationships between these positions and functional and structural features. Conserved positions (red) are in the structural core of the protein and in the active sites. Family dependent conserved positions (blue) are also present in the active site conferring specificity. Coevolving positions (green) indicate structural or functional dependencies. In the case of inter-protein correlations, these pairs are many times pointing to the interaction surface but not directly on it. Modified from (Pazos & Bang 2006).

Conserved positions

Conserved positions represent functionally or structurally important residues. These positions are usually the first indicators of functionality and can be related to all types of functions (e.g.: active sites, ligand binding, protein-protein interaction sites, nucleic acid binding) as well as to structural requirements like forming the structural core of the protein (Pazos & Bang 2006). Many different approaches have been developed to locate conserved positions in an alignment (Pazos & Bang 2006). Some methods detect sequence conservation at a position by calculating the difference between the maximum possible entropy and the entropy of the observed amino acids distribution (e.g., sequence logo (Schneidery & Stephens 1990). Some methods make use of complex models with phylogenetic trees to avoid the artifacts caused by a potential uneven distribution of sequences in an alignment (e.g. highly similar sequences can result in more conserved positions (Pazos & Bang 2006). Methods based on sequence profiles are also used to identify sequence conservation (Pazos & Bang 2006). Sequence profiles are extracted from the multiple sequence alignment (MSA) and provide information about the conservation as well as the amino acid distribution within the position.

Subtype-specific positions

Information regarding positions that are conserved but vary in their amino acids type within different subgroups of proteins can also be extracted from the alignment. The subgroups can be defined based on phylogenetic, phenotypic or functional criteria. These positions are functionally important as they provide functional specificity to the subfamilies and can also be used to define subfamilies. Various methods have been developed to analyze and predict protein subtypes from alignments. Livingstone & Barton in 1993 developed a method to annotate MSAs to identify conserved positions within subtypes using the amino acid properties and sequence similarity. Principal component analysis was used by Casari et al. in 2004 to identify positions conserved across the protein family as well as subtype specific residues. Olivier Lichtarge et al. in 1996, proposed an evolutionary trace method to explore the phylogenetic tree of a protein family for sequence conservation at different branches and locate the protein binding surfaces. Sjolander in 1998 developed a method of phylogenetic inference, in which the nodes are represented by a sequence profile of sequence at that node. This method ensures that highly conserved sites have higher weights. Hannenhalli &

Russell in 2000 proposed an approach that compares the intra and inter-group residue entropy for every possible split in the tree.

Covarying positions

Most of the sequence analysis methods proceed with the common assumption that sites or positions within biological sequences (DNA, RNA or proteins) vary independently. However, several biochemical and biophysical studies have shown evidence that this assumption is not biologically realistic. Some sites are under strong evolutionary constraints to maintain the basic structure and function. Substitutions/mutations at such sites can partly destabilize the molecule's structure or function and are thus compensated by subsequent (Yanofsky et al. 1964)) or simultaneous (Fitch & Markowitz 1970) substitution/mutation at another site. Such positions thus imply an important conserved interaction that can be ultimately detected in MSAs and are commonly referred to as covarying or coevolving positions.

Covarying positions are the positions undergoing correlated or compensatory mutations (i.e., a mutation at one site is compensated by a mutation at another site). Such positions share a common selection constraint and undergo correlated evolution or coevolution. Therefore, they show great potential for the prediction of functional and structural characteristics of molecules. Such covarying mutations can indicate residues that interact within the protein to carry out a specific functions such as, catalysis, structure stabilization, protein-protein interactions, and allosteric regulation (Teppa et al. 2014). They also provide useful information to understand evolutionary processes, to predict protein structure and to predict the effects of site-directed substitutions. Different studies have shown that such compensatory mutations are frequent and are involved in functional and biophysical properties of proteins (Teppa et al. 2014).

1.2.1 Molecular Covariation – Current Theories and Hypothesis

The precise model of sequence changes leading to the observation of covariation is still under debate. However, two scenarios have been proposed for how sites can covary: the second-site suppressor model proposed by Yanofsky et al. in 1964 and the concomitantly variable codon model or covarion model proposed by Fitch & Markowitz in 1970. Under second-site suppressor model, the first substitution is a deleterious mutation and leads to a decrease in fitness. This is then followed by substitution at a second site that could suppress the mutation and restore the function (Talavera et. al. 2015). In the second scenario of covarion model or directional selection model, the first substitution is neutral or near neutral and the second substitution then provides a selective advantage (Talavera et. al. 2015). However, both the hypotheses imply that coevolution occurs in order to generate the compensatory change (Wollenberg & Atchley 2000; Gloor et al. 2010; Talavera et al. 2015; Ackerman & Gatti 2011). Many experimental studies have been conducted proving either of the theories. Poon in 2005 measured compensatory mutation in DNA Bacteriophage ΦX174 and found that in most cases they occurred at a second site. Merlo et al. in 2007 performed domain-swapping experiments on triosephosphate isomerase and detected the rare covarions. Despite several studies been carried out there is still a clear controversy regarding the mechanism of covariation.

1.2.2 Complexity of covariation

Covariation between amino acids at two sites/position can be detected by using an MSA. Wollenberg & Atchley in 2000, suggested that covariation observed between two sites i and j can be decomposed into:

$$C_{ij} = C_{phylogeny} + C_{structure} + C_{function} + C_{interactions} + C_{stochastic}$$

$C_{structure}$ and $C_{function}$ signify the covariation occurring due to the selective constraints acting on the sites to maintain the structure and function. These two sources of covariation are difficult to distinguish, as often they are not independent from each other. $C_{interaction}$ reflects the interactions between the amino acid sites and is also a structural or functional component of covariation. $C_{stochastic}$ represents the stochastic factors such as, random effects from uneven or incomplete sequencing. Identifying this covariation is the realm of statistics used for detection of covariation. The statistics based covariation detection methods usually distinguish the stochastic component of covariation. $C_{phylogeny}$ is the correlation due to the shared ancestry between homologous sequences. This component can result in a global correlation of patterns occurring throughout the data set and thus add to the random noise further complicating the detection of coevolving positions. Thus, distinguishing phylogenetic correlation from functional correlation is a crucial and challenging step in detection of coevolution (Dunn et al. 2008; Dutheil 2011).

1.2.3 Molecular Covariation detection methods – Different strategies and controversies

For the past 20 years, many studies have been dedicated to identifying covarying site pairs in MSAs. Several parametric and non-parametric methods have been implemented to identify important residues based on the theory of site covariation.

Initially the methods to detect covariation were solely based on multiple sequence alignments. A number of different scoring ways have been proposed to identify coevolving positions including methods based on substitution matrix (McLachlan based substitution correlation (McBASC) (Gobel et al. 1994), methods based on perturbation (statistical coupling analysis (SCA)) (Lockless & Ranganathan 1999) mutual information based methods (MI) (Martin et al. 2005), methods to detect difference between observed versus expected frequencies of residue pairs (OMES) (Larson et al. 2000; Marks et al. 2012). Some of the methods were later modified to either add information about the conserved sites, as in positional conservation based SCA (SCAnew) (Halabi et al. 2009) or to remove phylogenetic and background noise, as in corrected mutual information (MIp) (Dunn et al. 2008). Marks et al. in 2011 and 2012 divided these methods into methods based on local statistical models and those on global statistical models. MI, SCA, McBASC, OMES, MIp belong to local statistical model based methods. They assume that pairs or residue positions are statistically independent of the other pairs of residues. In contrast, global methods like direct coupling analysis (DCA) (Marks et al. 2012), Bayesian network model (Burger & van Nimwegen 2010) and protein sparse inverse covariation (PSICOV) (Gouveia-Oliveira & Pedersen 2007), consider correlated residue pairs to be dependent on each other and thereby remove the effect of transitivity. These methods have shown success in detecting covariations that relate to contact in 3D-strutcute and have been used for detection of 3D-structures by several studies (Marks et al. 2011; Marks et al. 2012; Hopf et al. 2012; Hopf et al. 2014; Hopf et al. 2015).

Several other methods were developed that attempt to model covariation of sites directly on a phylogenetic tree. These studies take the evolutionary history of the sequences into account (Pollock et al. in 1999, Tuffery & Darlu in 2000, Dutheil & Galtier in 2007 and Linda Dib et al. in 2014). These methods aim to account for phylogenetic correlation due to shared ancestry, by incorporating the phylogenetic correlation in observed statistics. They require a phylogenetic tree reconstruction of the sequence alignment and then compute the coevolutionary statistics from the phylogeny.

Phylogenetic noise is one of the major problems faced by methods based on sequence alignments such as MI. It is due to the fact that the proteins sequences are not independent and have an inherent signal due their phylogenetic relationship (Wollenberg & Atchley 2000; Gouveia-Oliveira & Pedersen 2007). A number of different approaches have been proposed to lower the background phylogenetic noise and thus enable more accurate identification of the coevolving positions (Dutheil 2011; de Juan et al. 2013). These include correction of the sequence based MI methods like MIp and the previously mentioned tree based methods.

Drawbacks have been identified for both approaches (Caporaso et al. 2008). There is an on-going debate on the validity of assumptions used on both the approaches. Such debates are further complicated by the fact that the field of molecular coevolution/covariation is at the crossroads of two distinct communities, structural/functional biologists and evolutionary biologists (Dutheil 2011). The covariation of two positions in a sequence alignment can refer to associate pattern from comparative sequence analysis point of view. However, from phylogenetic perspective it can refer to simultaneous substitution pattern of sites that they undergo during their evolutionary history. The tree-based methods are computationally expensive and have the disadvantage of being sensitive to the model that is used to generate a phylogenetic tree. The tree-ignorant or sequence based methods are easy to implement with low computational cost required, but are sensitive to the effect of shared ancestry of the sequences. However, as mentioned earlier, sequence based methods were further improved to explicitly account for the shared ancestry (e.g. MIp).

Comparative analysis studies have shown that tree-ignorant methods that account for shared ancestry out-performed the tree-based methods in identifying coevolving positions (Caporaso et al. 2008) and are thus more reliable than tree- based methods. Such methods, like MIp, are easy to implement, fast and minimizes the assumption on the data and therefore can be a reliable choice for a large data sets analysis (Dutheil 2011).

1.2.3.1 Mutual Information (MI)

Mutual information (MI) is the measure of reduction of uncertainty and is based on Shannon's entropy. The MI between two columns of a MSA reflects the degree to which the knowledge of the amino acid at one position helps to predict the identity of the amino acids at the other position. High MI values indicate correlation between the two positions. MI score ranges between 1 and 0, with higher MI value reflecting a higher interdependence between the two positions of an MSA.

Korber et al. 1993 first applied MI to MSAs of a short variable loop of an HIV envelope protein to identify correlated pairs. The initial formulations of MI were affected by high variability positions in MSAs and by the effect of phylogenetic background (Caporaso et al. 2008; del Sol Mesa et al. 2003; de Juan et al. 2013) and thus subsequent, improved versions of this approach had to be developed (Tillier & Lui 2003; Marino Buslje et al. 2010; Gouveia-Oliveira & Pedersen 2007; Dunn et al. 2008).

1.2.3.2 Corrected Mutual Information (MIp)

The mutual information approach was corrected to suppress the phylogenetic bias by normalizing the observed covariance of a pair of column by the background covariance of the columns. This correction on MI provided a substantial improvement compared to previously published methods for predicting covarying positions (Dunn et al. 2008).

Different comparative studies have also shown it to be better than tree-based methods for identifying coevolving positions (Caporaso et al. 2008). It is a fast and easy to implement method and has been shown to be a good choice for large data set analysis (Dutheil 2011).

1.2.4 Groups of covarying residues

Most of the covariation or coevolution detecting algorithms provide pairwise scores of residues. However, due to flexibility of the protein structure, it is possible that such compensatory mutations do not always occur only as pairs but can be compensated by other substitutions occurring at different positions. Groups or networks of covarying residues may reflect a series of mutational events within a local region of a molecule. Networks of such residues can also show a coordinated action of these residues in a functional or structural context. Such groups can thus provide information that is not related to direct contacts. They can provide information about long-range indirect interactions that are induced by chains of directly interacting residue pairs that run through the protein to connect distal pairs.

A few studies have demonstrated the existence of groups of covarying residues (Lockless & Ranganathan 1999; Süel et al. 2002; Halabi et al. 2009; Burger & van Nimwegen 2010). These studies have shown that such residues create physically connected networks that link the distant positions in the tertiary structure of proteins. They believed them to be thermodynamically coupled residues involved in allosteric communications within the protein. Allosetry represents the dynamics of proteins, where a perturbation at one site (substrate binding, covalent modifications or mutations) affects a spatially and sequentially distant site. Such distant sites within a protein communicate through local conformational changes that produce dynamics that leads to global changes. A recent review by Motlagh et al. in 2014, explains that allosetry can also be associated with changes in dynamics of proteins and large scale conformational disorders occurring within the protein. A recent study has also shown the involvement of coevolving fragments in the binding specificity and folding constraints that explains folding intermediates, peptide assembly and known mutations with roles in genetic diseases (Dib & Carbone 2012).

Thus, looking at just the pairs of residues as attempted so far may underestimate the covariation signal within a protein. However, performing an exhaustive search for groups of residues of arbitrary size is also difficult due to high number of possible combinations within a protein. It is also difficult to evaluate the roles of so many pairwise residue couplings. Clustering techniques are standard methods, which can be used to cope with such issues.

2 Aims of the project

Vesicle trafficking is a vital process of eukaryotic cells by which molecules (nutrients, protein, lipids, etc.) are transported from a donor compartment to an acceptor compartment via small membrane-bound carriers called vesicles. SNARE (Soluble N-ethylmaleimide-sensitive factor Attachment protein REceptors) and SM (Sec1/Munc18) proteins form the core machinery for vesicle fusion. Various other factors including Rab proteins, NSFs (N-ethylmaleimide-Sensitive Factor), SNAPs (Soluble NSF Attachments Proteins) and tethering proteins belonging to the CATCHR (Complex Associated with Tethering containing Helical Rods) family also co-ordinate the vesicle fusion process. All these proteins are highly conserved, not only between different species but also between different vesicle trafficking steps within the cell. Probably they arose by duplication and diversification of prototypic protein machineries during evolution. Although a considerable amount of research has been conducted on the vesicle fusion machinery, it has remained challenging to come to a comprehensive understanding the underlying protein interaction networks, possibly as most of the studies have been carried out only in few model organisms. Proteins from different eukaryotic lineages undergo different adaptations and losses or gains and thus it becomes difficult to clarify if a certain feature is a result of special adaptation of a particular protein or whether it is a shared trait. Sequence analysis can help in filling this gap as the whole set of sequences across all species can be investigated.

A large amount of work has previously been carried out in my group towards analyzing the sequences of the various factors involved in the vesicle fusion step. To store and analyze the sequences of factors involved in vesicle fusion, my group has implemented a database management system, which provides an opportunity to search for functionally and structurally important residues using computational approaches. Previously, my group has also analyzed the evolutionary history of the key proteins participating in vesicular fusion. The phylogenetic analysis of SNARE proteins and SM proteins had revealed some comparable patterns of duplications and diversifications between the closely interacting proteins. It would thus be interesting to explore the covariation patterns of these different protein families to gain better insight about the structurally and functionally important sites in these proteins.

The main aim of my project was therefore to identify functionally and structurally important residue networks by extracting their covariation relationship from multiple sequence alignments and thus explore their evolutionary changes across different eukaryotic lineages. This will provide novel insights into the structural and functional constraints as well about the complex mutational dynamics that took place during the evolution of proteins involved in the vesicular fusion process. The information obtained can later be used as a guide to structural or mutational studies on vesicle trafficking proteins so as to gain a better understanding of their function and interaction.

To achieve this aim, development of a comprehensive bio-informatics framework for sequence analysis was needed as the available software solutions had limited scope and provided insufficient visualizations for my purpose. The objective of the software was to provide a conglomerated set of information about a MSA. The basics are similar to software such as JalView, however, the focus of my software was on enriching data with statistical information, covariation detection analytics, advanced subtyping, visualizations and automatic analytic pipelining

SM proteins form the core of vesicle fusion machinery along with SNARE proteins. They interact with SNARE protein Syntaxins and thus control and guide the vesicle fusion process. Since more than a decade, the molecular role of SM proteins, in particular their binding mode with Syntaxins is debated. The two proteins generally make use of two spatially separated binding sites, but it is unclear how the two binding sites are able to communicate. To gain novel insights into their interplay I aimed at extracting novel information by investigating at their sequence covariation. As initial covariation results were rather complex, an improved analysis approach by combining the covariation detection methods and network analysis methods needed to be developed to be able to better understand the co-variation pattern of proteins of the vesicle fusion machinery.

3 Materials & Methods

3.1 Sequence Alignments

Previously, my group has developed and implemented a database management system and analyzed the evolutionary history of the SNARE protein (Kloepper et al. 2007; Kloepper et al. 2008; Kienle et al. 2009) and Rab protein family (Klöpper et al. 2012). The database system was consequently extended to incorporate additional protein families involved in intercellular traffic. Currently the database has sequences of SNARE proteins, SM proteins, Rab proteins, SNAP proteins and AAA proteins. Sequences of the SM protein subfamilies and SNAP proteins were collected from the in-house database. They were aligned by using a previously developed strategy (unpublished work). The well-conserved regions in domains were aligned with the HMM output (Krogh et al. 1994) and the inserts were realigned with MUSCLE (Edgar 2004). The alignments were further refined, by using an iterative block strategy. This strategy searches for conserved blocks and realigns and recursively refines the unconserved blocks in between them. Additionally, a conserved alignment filter was used to ensure that alignment contains only significant information. This filter removes the columns and rows with information content (either in terms of gap or entropy) below a certain threshold.

3.2 Phylogenetic Tree Generation

The phylogenetic trees used to generate the simulated alignments were obtained from a combination of three different programs IQ-TREE (Nguyen et al. 2014), RAxML (Stamatakis 2006), and PhyML (Guindon et al. 2009). IQTREE was first used with the test option to estimate the best fitting parameters. For all trees generated, the LG matrix was the most fitting substitution model together with gamma rate heterogeneity. All the three programs were used with 1000 bootstrap replicates. After the successful reconstruction of the trees, site-wise log likelihoods were calculated with RAxML. Consel (Shimodaria & Hasegawa 2001) was then used to rank the trees. The best tree was taken as a reference. TREE-PUZZLE (Schmidt et al. 2002) was then used as an additional independent confidence estimator to run likelihood mapping on the best tree.

3.3 Generation of Simulated Alignments

In molecular evolution, computer simulation of data has been widely used to test a hypothesis, compare or evaluate tools or methods, to access the fit of a model and study complex evolutionary processes.

At molecular level, the analysis of dependent or coordinated substitutions provides information about the potential structurally or functionally important positions along a DNA/RNA or protein sequences. A wide variety of algorithms have been developed to detect covarying positions from a MSA. Since there is no general analytical evolutionary model for covariation, simulated MSAs have been used as a tool by many studies to test their methods for filtering out the background coevolutionary signal (Dib et al. 2014; Fares & Travers 2006; Fodor & Aldrich 2004; Tillier & Lui 2003; Martin et al. 2005; Gloor et al. 2005; Ackerman et al. 2012; Dutheil & Galtier 2007; Dutheil 2011). The simulated alignments are generated based on certain properties of the real alignments, such as, conservation level, distribution of amino acid at each site, pairwise similarities between the sequences.

Earlier available methods for simulation used Markov models to simulate each position independently and thus were not appropriate to evaluate coevolution of positions in nucleotide or amino acid sequences. In this study, simulator developed by (Dib et. al. 2015) was used, which was based on their evolutionary model, Coev Markov model (Dib et al. 2014). Given a rooted binary tree in Newick format and the values of the 4 continuous parameters of Coev model (Dib et al. 2014) the method simulates the nucleic and protein pairs of positions. The method randomly picks a coevolving profile and lets it evolve along the branches of the tree as per the Coev substitution matrix. Given a profile, the instantaneous rate matrix Q of the model, Coev, is modeled as follows:

$$Q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ by two nucleotide positions,} \\ r1 & \text{if } \{i,j\} \notin \phi \text{ and if } i \text{ differs from } j \text{ at position 1,} \\ r2 & \text{if } \{i,j\} \notin \phi \text{ and if } i \text{ differs from } j \text{ at position 2,} \\ s & \text{if } i \in \phi \text{ and } j \notin \phi, \\ d & \text{if } i \notin \phi \text{ and } j \in \phi \end{cases}$$

where, the parameter s is the rate of transition from a coevolving combination present in the profile to a non-coevolving combination, parameter d is the rate of transition from one non-coevolving to a coevolving combination; the additional parameters r1 and r2 are the rates of transitions between two non-coevolving combinations at positions 1 and 2, respectively. A combination of d>> s would simulate coevolving pairs, while a combination of s=d would simulate non-coevolving pairs.

To evaluate the efficiency of the analysis pipeline in detecting the coevolving residues, simulated MSA was generated using the CoEv Simulate.

Application of the developed improved analysis pipeline on simulated data

The developed approach for identifying groups of covarying residues was applied on a simulated dataset, where certain residue pairs were artificially forced to be coevolving with each other. This was done to verify the accuracy of the range of MIp scores obtained on the real dataset and also to check if the top scoring MIp residues indeed matched with those that are set to be coevolving in the simulation.

A rooted binary tree was generated by using a combination of three different softwares (IQ-TREE (Nguyen et al. 2014), RAxML (Stamatakis 2006), and PhyML (Guindon et al. 2009) from the alignment of an SM protein subfamily, Sly1, with a few sequences of another SM protein subfamily, Vps45 used as outgroup (for details see Methods section). Using this tree as an input two simulated alignments with only coevolving and only non-coevolving pairs were generated. A simulated alignment was generated with 50 coevolving pairs (100 residues), using s parameter equal to 10 and d parameter equal to 10000. Another simulated alignment was generated with 200 coevolving pairs (100 residues), using s parameter equal to 10 and d parameter equal to 10. The two alignments were then concatenated into one alignment of 500 amino acid positions. The number of coevolving pairs was chosen such that it is about L/5, where L is the total length of the concatenated alignment. Earlier studies like Gobel et. al. 1994 used a cut off of L/5 for coevolutionary residue prediction and so this was chosen as a reasonable number of coevolving sites.

MIp and then average linkage clustering with varying stopping criteria from mean, 1-σ, 2-σ, 3-σ and 4-σ was applied on the concatenated alignment. The MIp scores generated were visualized with the help of a heat map (Figure 3.1) and the selected clusters were also highlighted in a heat-map organized according to the rank of the clusters (Figure 3.2).

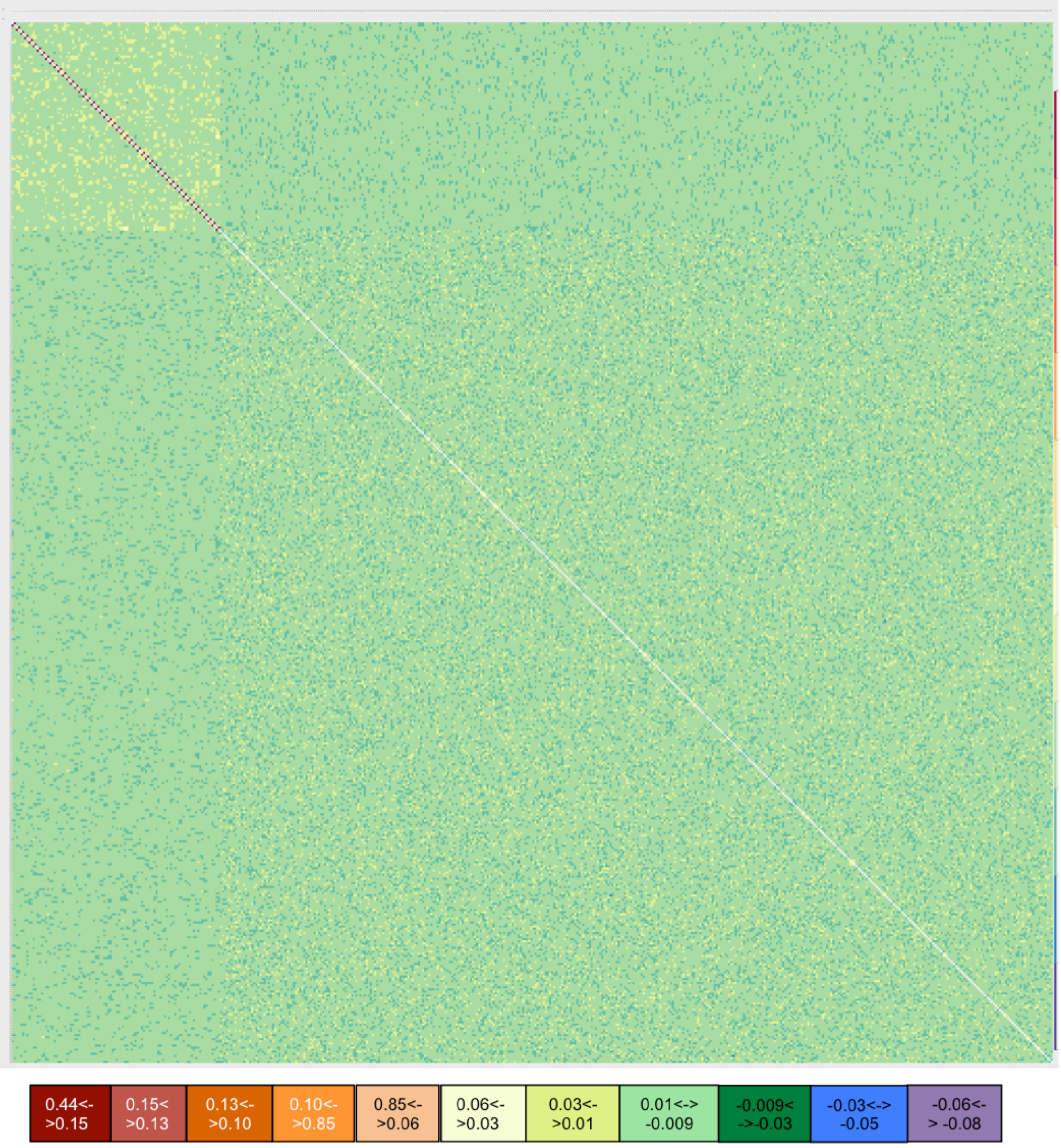


Figure 3.1 : Heat map of MIp scores from the simulated alignment.
The color code bar represents the color corresponding to the range of MIp score. The dark red, red, dark orange, orange, represent the high MIp scores and light green, dark green, blue, purple represent the low MIp scores. The original heat map can be found in the supplements that can be enlarged to have a clear view of the residue numbers.

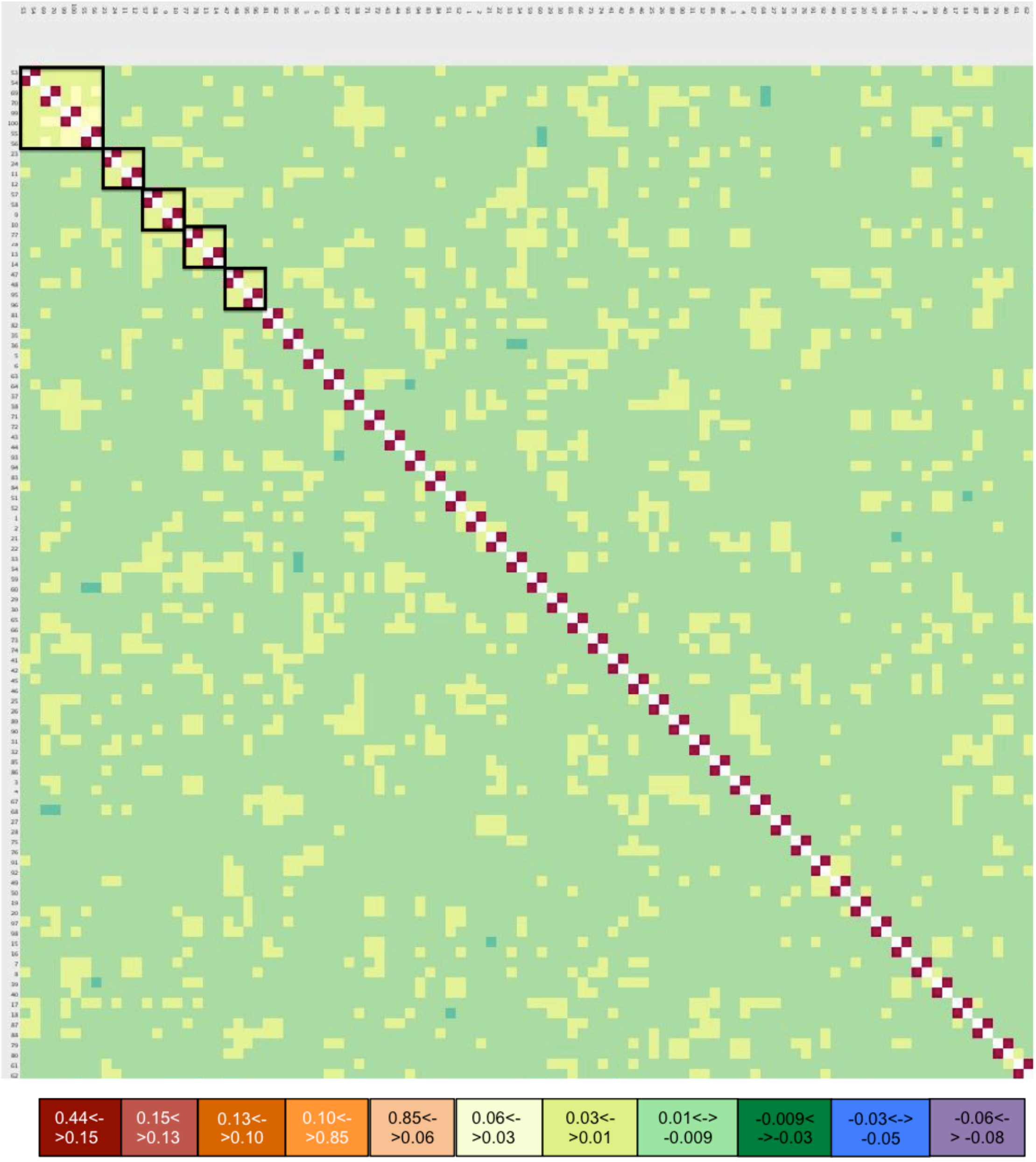


Figure 3.2: Heatmap of MIp scores of simulated coevolving residues.
The heatmap is arranged according to the ranking of clusters. Only the part containing the simulated pairs is shown. The residues appearing as clusters of more than 2 residues are marked in Black squares.

To identify covarying pairs, the clusters were scored and ranked as per average weighted degree of the cluster. Stopping criteria of mean and 1-σ resulted in clusters with both coevolving and non-coevolving residues being clustered together. Stopping criteria of 2-σ resulted in clusters of 8, 4 and 2 residues with coevolving and non-coevolving residues in separate clusters. Stopping criteria of 3-σ, resulted in one cluster of 4 residues and rest all with 2 residues with coevolving and non-coevolving residues in separate clusters. 4-σ stopping criteria resulted only in clusters with coevolving residues. Based on the result obtained at different cut off, 2-σ was chosen as stopping criteria. Cluster selection cut off of 95% or 90% of cluster scores was chosen as it resulted in all selected cluster with coevolving residues.

Standard performance methods were then used to evaluate the pipeline. The number of positions correctly predicted as coevolving (true positives, TP), the number of positions correctly predicted as non-coevolving (true negatives, TN), the number of non-coevolving positions predicted as coevolving (false positives, FP) and the number of coevolving residues predicted as non-coevolving (false negatives, FN), were estimated. Using these measures, the sensitivity TP/(TP + FP) and specificity TN/(TN + FP) of the approach was calculated. In my case, the true positives were 100, as at 95% cluster score cut off, all the cluster with total of 50 coevolving pairs were obtained and since no cluster with non-coevolving position was found the false positive and false negative were 0 and true negatives were 400. Thus, the sensitivity of the approach was 1 and specificity was also 1. Further, a true positive rate (TPR=1-specificty or TP/TP+FN) and a false positive rate (FPR= FP/TN+FP) can be calculated. The combinatorial approach resulted in a TPR of 1 and FPR of 0, thus indicating maximum sensitivity and specificity.

3.4 Covariation Analysis

Mutual Information (MI)

Mutual information (MI) is a method based on Shannon’s entropy that indicates the dependencies of the two columns. It is a measure of reduction of uncertainty. The MI between two columns of a MSA reflects the degree to which the knowledge of the amino acid at one position helps to predict the identity of the amino acids at the other position. A high MI values indicates correlation between the two positions. The implementation of MI is based on (Martin et al. 2005).

MI between two positions of MSA is calculated by:

$$MI(x,y) = H(x) + H(y) - H(x,y)$$

where, $H(x)$ and $H(y)$ are entropy of columns x and y, calculated as,

$$H(x) = -\sum_{i=1}^{K} p(x_i)\log_b p(x_i)$$

where $p(x_i)$ is probability of amino acid i in column x, $k=20$ (for 20 amino acids). The logarithm base b=20. The value of $H(x)$varies from 0, in case of complete conservation to 1, when all 20 amino acids are equally distributed. $H(x,y)$Is the joint entropy, which is defined as:

$$H(x,y) = -\sum_{i=1}^{k}\sum_{j=1}^{l} p(x_i,y_j)\log_b p(x_i,y_j)$$

where $p(x_i,y_j)$is joint probability of amino acid i in column x and amino acid j in column y, k=l=20 for amino acids, and b is logarithm base, here set to 20. The joint entropy can range from 0 to 2.

MI score ranges between 1 and 0 with high MI value reflecting a higher interdependence between the two positions of a MSA.

The initial formulations of MI were affected by high variability positions in MSAs and by the effect of phylogenetic background (de Juan et al. 2013) and thus many subsequent version of this approach were developed. One of them is described next.

Mutual Information Corrected (MIp)

The mutual information approach was corrected to suppress the phylogenetic bias by normalizing the observed covariance of a pair of column by the background covariance of the columns. The background covariance is the average covariance score of the column with all the other columns (Dunn et al. 2008). Thus,

$$MIp(a,b) = MI(a.b) - APC(a,b)$$

where $APC(a,b)$ is the average product correction of the background scores, calculated as:

$$APC(a,b) = \frac{MI(a,\bar{x})MI(b,\bar{x})}{\overline{MI}}$$

where $MI(a,\bar{x})$ is mean mutual information of column a, defined as:

$$MI(a,\bar{x}) = \frac{1}{m}\sum MI(a,x)$$

where n is the number of columns in MSA, and $m=n-1$, and summation is over $x=1$ to n, $x\neq a$, and $\overline{MI}$ denotes the overall mean mutual information,

$$\overline{MI} = \frac{2}{mn}\sum MI(x,y)$$

where indices run from $x=1$ to m, $y=x+1$ to n.

This correction on MI, provided a substantial improvement compared to other previously published methods for predicting covarying positions (Dunn et al. 2008; de Juan et al. 2013).

3.5 Network Analysis

Clustering is an approach to structure data by placing similar cases together in a group called cluster. Partition and hierarchical methods are the two major classes of clustering methods. However, there are wide ranges of different algorithms like model-, density- and grid-based methods.

Partitioning methods divide n cases into small discrete k classes, as described by p variable. They suffer from two major problems. Firstly, the value of k needs to be pre-defined and secondly, the solutions are not unique as the iterative algorithm that starts at random locations is used.

Hierarchical clustering does not require pre-specifying the number of clusters needed. It arranges data into a hierarchy based on the distance or similarity. *Alifea* provides three hierarchical clustering methods.

While using covariation data as a network, the covariation score is used as the distance between the two nodes or the residues. The higher the covariation score, the shorter is the distance between the residues and vice-versa.

Single Linkage Clustering

In single linkage clustering, also known as minimum or nearest neighbor methods, the distance between the clusters is the minimum distance between the members of the two clusters. This method produces long chains that can form loose clusters. The two clusters are fused together if the minimum distance (maximum covariation score) between the members of the two clusters is more than the average covariation score.

Complete Linkage Clustering

The complete linkage clustering is also known as maximum or furthest neighbor method. The distance between the clusters is the greatest distance between members of the two clusters. This method tends to produce very tight clusters of similar scores. The two clusters are fused together if the maximum distance (minimum covariation score) between the members of the two clusters is more than the average covariation score.

Average Linkage Clustering

The average linkage clustering uses the average values between the members of the two clusters as the distance between the clusters. The two clusters are merged if the average distance (or the average covariation score) between the members of the two clusters is more than the total average covariation score. It is a compromise between the sensitivity of complete-link clustering to outliers and the tendency of single-link clustering to form long chains that do not correspond to the intuitive notion of clusters as compact, spherical objects.

A stopping criterion was used for the average linkage clustering. This stopping criterion can be varied in different statistical steps of the distribution of the edge weights (such as 1-σ, 2- σ, 3- σ and 4- σ from the mean-μ of the weights. Only residue pairs that have the covariation score above the stopping criteria were considered for clustering. This step was used to remove weak connections between the residues, so that only those residues that have strong covariation were considered.

3.6 Scoring and ranking the clusters

Each cluster was then scored and ranked. Clusters can be scored either based on the average weighted degree of the cluster or based on the average weight of the cluster and then ranked in the decreasing order of the score.

Average weighted degree score of the cluster is calculates as:

$$\frac{1}{N}\sum W_i$$

where, W_i is the weight of the connection i and N is the total number of nodes. This score can be summarized as sum of edge scores/number of nodes. In general, this should result in large clusters being ranked higher and can be used to identify the groups of covarying residues.

Average weighted score of the cluster can be calculated as:

$$\frac{\sum W_i}{E}$$

where, W_i is the weight of connection i and E is the total number of connections or edges. This score can be summarized as sum of edge scores/number of scoring pairs. In general, this should result in small clusters being ranked higher and can be used to identify small groups of covarying residues.

A cut-off was then used to select the top clusters that have significant residue-to-residue interactions within the cluster. The cluster selection cut-off can be varied from 10% to 90% of the cluster scores.

Defining these cut-offs appropriately plays an important role in the quality of clusters that are identified. Different choices for selecting the values of cut offs (either the cluster selection cut-off or the stopping criteria) are given, as a certain cut-off might be more suitable for a particular MIp dataset. Thus, the cut-off values need to be adapted to the type of protein.

A simulation-based method was used to define the cut-offs for the particular protein family under analysis. Simulation was performed using the software CoEv Simulate to generate two simulated alignments, one with some coevolving residues and another with some non-coevolving residues. The co-evolving residues were then uniformly distributed with the non-coevolving residues, resulting in a concatenated alignment with known coevolving residues. The clustering process was then repeatedly performed with different values of stopping criteria and cluster selection cut off. The value of stopping criteria and cluster selection cut off that resulted in identifying the co-evolving residues of the simulated data set accurately and uniquely (with least false positives) was chosen.

3.7 Visualizations

Heat Map View

A heatmap view is a graphical representation of data, where individual pairwise values contained in a matrix generated by covariation analysis method, are represented by varying degree of colors. This view was developed and implemented so as to have a detailed look at the entire covariation (MIp) dataset rather than looking only at certain high scoring sites. It can help to provide a complete picture of the distribution of covarying residues within a protein. A heatmap visualization of the MIp scores illustrates the correlated mutation properties of residue pairs. The covariation matrix is mapped to an appropriate color map displayed as a symmetrical grid.

The heatmap can be arranged according to the alignment positions or to a representative sequence positions or according to the ranking of the clusters. The user has options to select different color schemes. The default color scheme maps higher covariation scores into red regions (dark red, red, dark orange, orange, light orange) and lower covariation scores blue region (lemon, light green, dark green, blue, purple). The user can also set up different values of maximum and minimum covariation score. There is also a choice to display only the positive covariation scores or all the covariation scores. The zero values of the covariation matrix can also be shown with a different color to differentiate them from the rest of the score. Information about the secondary structural elements can also be mapped on top of the heatmap. This information can be extracted from the PDB file if available or from the secondary structure prediction file provided by JPred.

Network View

The cluster obtained after applying clustering approach can be visualized with the help of the developed and implemented network view. In this view, nodes are shown as circles, representing the residue. The circle or nodes are connected by an edge representing the covariation score between the two residues. The color of the edge represents the strength of the covariation score, with red being highest and blue being the lowest. The color and size of the node represents the importance of the residue in terms of the average of the score of the edges connected to it.

Structure View

A Jmol viewer [75] has been integrated in tool, which can be used to map the results directly on a 3D structure, if available. The user has a choice to add one or more than one structure of different proteins. There is also possibility to map the identified positions along with the conserved positions (conservations degree as chosen by the user) position. The program can also automatically mark these positions on the MSA and show them as a sequence logo. It can also extract information like distance between the residues and secondary structure elements form the PDB file.

Hierarchy/Tree View

A tree view or the hierarchy view has been implemented that allows the user to visualize any phylogenetic tree in form of a hierarchy. This view can also be used to visualize the generated clusters.

3.8 Analysis Pipeline

Figure 3.1 illustrates the developed analysis pipeline.

The basic required input for the analysis was an MSA. The alignments would undergo a preprocessing step, where columns with 90% gaps would be removed. The resulting alignment was subjected to a covariation analysis using MIp method. The gaps were treated as signals or the 21^{st} amino acid type. A conservation check for 80% conserved columns was added so that MIp scores were calculated only if neither of the columns was 80% conserved. The resulting MIp dataset can be visualized with the help of the developed heat map visualization. Average linkage clustering with a certain stopping criteria was then performed on the complete MIp dataset. In order to identify groups of strongly covarying residues the obtained clusters were scored and ranked based on average weighted degree of the cluster. Top clusters above a certain cut-off were selected as the groups of covarying residues. The final selected clusters can be visualized as a network. The residues within the selected clusters can be mapped on a PDB structure, if available. In order to identify the evolutionarily dependent residues, a clade-specific analysis, where the covariation of the identified residues is mapped on the phylogenetic tree, was then performed.

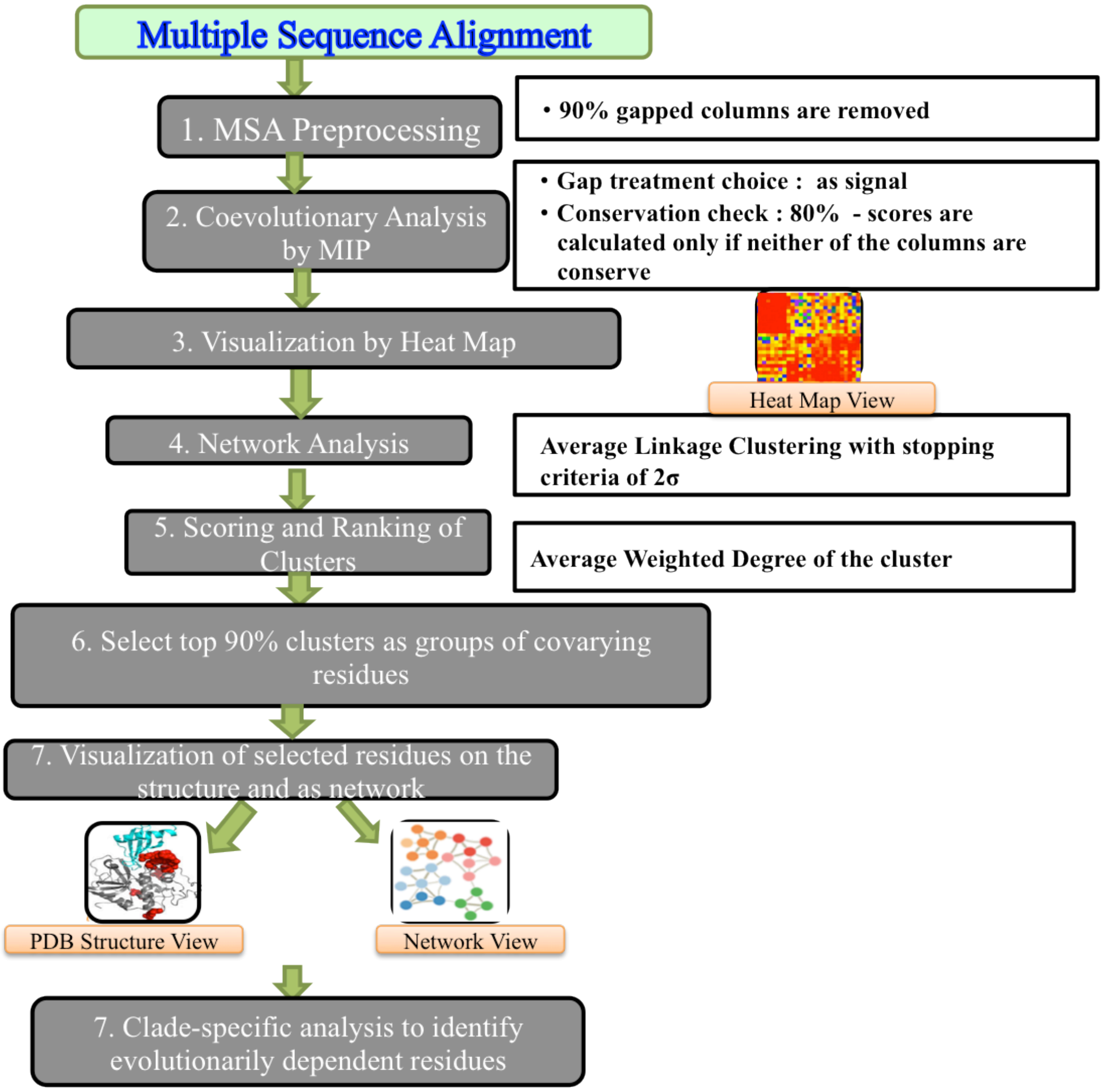

Multiple Sequence Alignment
1. MSA Preprocessing
• 90% gapped columns are removed
2. Coevolutionary Analysis by MIP
• Gap treatment choice : as signal
• Conservation check : 80% - scores are calculated only if neither of the columns are conserve
3. Visualization by Heat Map
Heat Map View
4. Network Analysis
Average Linkage Clustering with stopping criteria of 2σ
5. Scoring and Ranking of Clusters
Average Weighted Degree of the cluster
6. Select top 90% clusters as groups of covarying residues
7. Visualization of selected residues on the structure and as network
PDB Structure View
Network View
7. Clade-specific analysis to identify evolutionarily dependent residues

Figure 3.3: Schematic representation of the workflow of the improved analysis pipeline.

4 Results

As mentioned in the Introduction, my group had previously developed a database management system to store and analyze sequences from various factors involved in the vesicle fusion step. Among others, my group has collected sequences of SNARE proteins, Rab proteins, SNAP proteins, NSF family, and the SM protein family. With its well-curated sequence information this database provides a basis for the search for functionally and structurally important residues using computational approaches. My group has already analyzed the evolutionary history of several protein families participating in vesicular fusion. Systematic evolutionary analysis of SNAREs (Kloepper et al. 2007; Kloepper et al. 2008; Kienle et al. 2009) and Rabs (Klöpper et al. 2012) has been performed already. Preliminary phylogenetic analysis has also been carried out for the SM protein family. The phylogenetic analysis of SNAREs and SM proteins has brought to light some comparable patterns of duplications and diversifications between these closely interacting proteins. For example, of the five different Qa-SNARE types, only the one involved in endosomal/vacuolar trafficking and in secretion were duplicated and diversified during the rise of animals. Comparable patterns of change were seen for SM proteins, where the ones involved in endosomal/vacuolar trafficking and in secretion, duplicated and diversified in animals. These comparable patterns might denote co-evolution. To obtain more functional and structural information from the large sequence collection, I decided to explore the covariation relationship residing within the multiple sequence alignments. This, so the reasoning would help to explore the evolutionary changes occurring in the proteins of vesicle fusion machinery across different eukaryotic lineages.

To begin with I choose to investigate the SM protein family, as these proteins, compared to other proteins of the vesicle fusion machinery, have a manageable number of subtypes. By contrast, the previously studied SNARE and Rab protein families are composed of many different subfamilies, about 20 or more subclasses for SNAREs can be distinguished (Kloepper et al. 2007) and even more subclasses for Rab protein types were found (Klöpper et al. 2012). However, the SM protein family consists of only five subtypes that probably had been present in the last eukaryotic common ancestor.

As explained already in the Introduction section, SM proteins interact with Qa-SNAREs/Syntaxins and are essential factors in vesicular fusion machinery. The SM protein family comprises of five basic members: Sec1/Munc18, Sly1, Vps33, and Vps45 and scfd2 that participate in different trafficking steps within the cell. The exact the molecular role of SM proteins, in particular their binding mode with Syntaxins is still debated. Recent biochemical and structural studies on different SM/Syntaxin complexes suggest, however, that SM proteins and syntaxins generally make use of two spatially separated binding sites (Carpp et al. 2006; Furgasona et al. 2009; Aran et al. 2009; Martin et al. 2005; Johnson et al. 2009; Burkhardt et al. 2008; Demircioglu et al. 2014; Hackmann et al. 2013). Both binding sites appear to be important, although it is still unclear how the two binding sites are able to communicate in the complex. It is possible that there is a conformational switch that enables the SM protein to control the accessibility of the bound Syntaxin. So to gain deeper insights into these proteins and to extract additional functional and structural information, covariation analysis was initially applied on SM protein subfamilies.

4.1 Initial Covariation analysis on the SM protein Sly1

As a first step, intra-protein covariation analysis was performed on SM dataset using a widely used sequence covariation detection method, MIp. Following the approach used in earlier studies (Gloor et al. 2005; Martin et al. 2005; Dunn et al. 2008) only pairs that scored significantly (above Z-score=4.0 cutoff, corresponding to 99.99% cutoff) were selected. In the following, I briefly present the initial covariation analysis performed on the SM protein Sly1. Sly1 was used as it occurs in most eukaryotic genomes as singleton. A table of all residue pairs scoring higher than Zscore=4 is given in the Appendix section (Appendix A.1 (table can be provided on request)). The residues identified by this analysis were found to be spread all over the protein as can be seen by mapping the residues on the structure of Saccharomyces cerevisiae Sly1 (PDB-ID: 1MQS) (Figure 4.1). Some residue pairs were found to be lying close to each other, while some residues were found in groups or clusters of many residues. Often individual residue pairs were found lying next to each other. Figure 4.2 shows examples of pairs found in close proximity in the structure. An example of residues that appear to form a larger group is given in Figure 4.3. Often the residues that formed a group are lying distant from each other in tertiary structure. Some of the residues that formed the group, overlapped in such a way that they appeared to form a network.

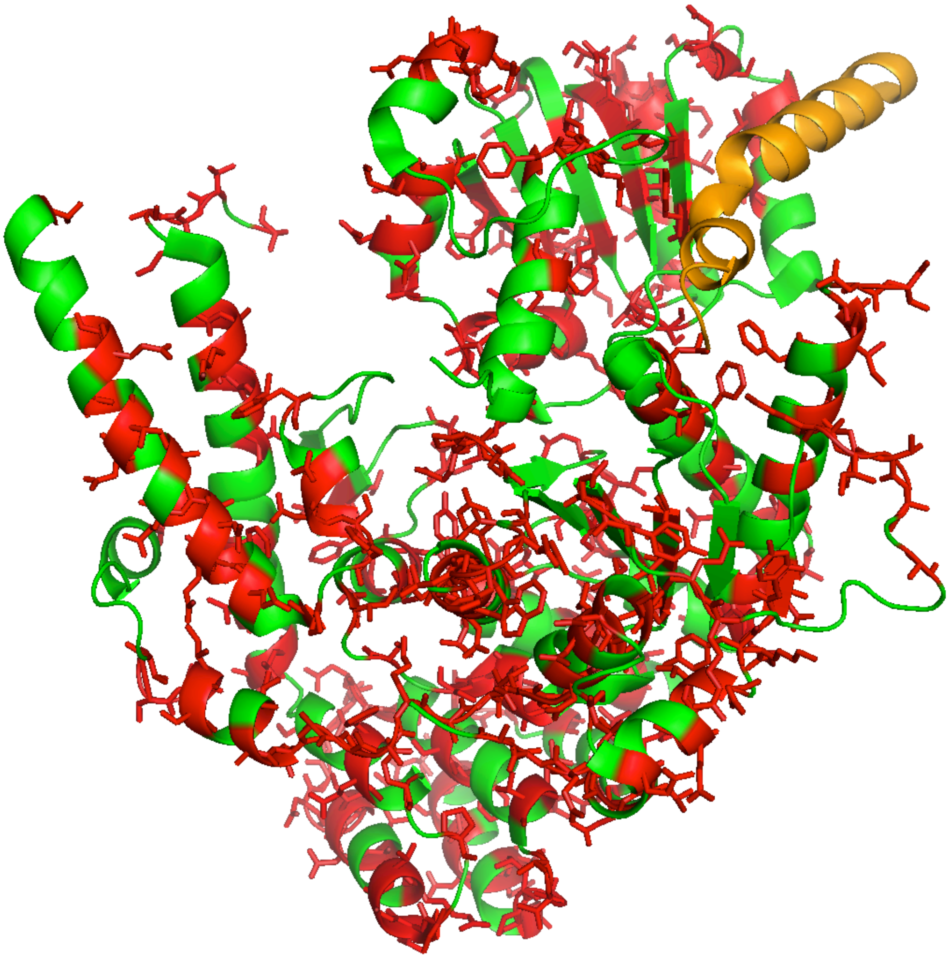


Figure 4.1: Covarying residues with Z-score > 4 marked on structure of *Saccharomyces cerevisiae* Sly1 in complex with the N-peptide of Sed5 (1MQS).
Sly1 structure is shown in green and the Sed 5-N-peptide structure is shown in orange. Residues identified in the intial analysis that have scores above a cut off of Z-score=4 are shown in red color.

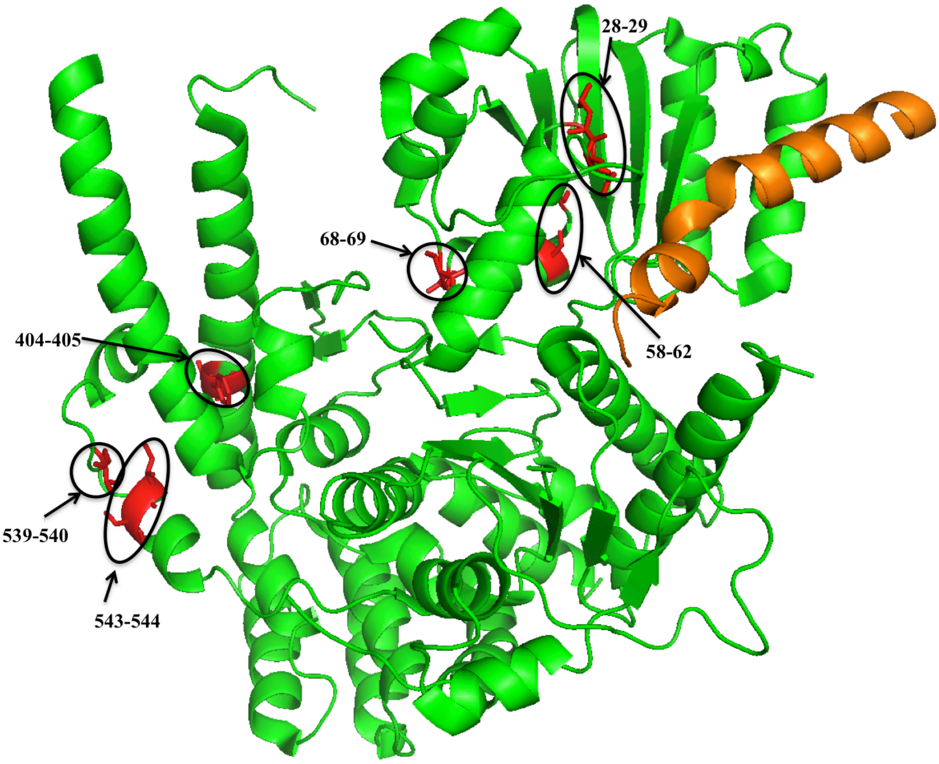


Figure 4.2: Paired residues with Z-score>4, marked on the structure of *Saccharomyces cerevisiae* Sly1 in complex with the N-peptide of Sed5 (1MQS).
Sly1 structure is shown in green and Sed5-N-peptide structure is shown in orange. The paired residues that are identified in the initial analysis and shown in Table 4.1, are circled and marked in red. Most of the paired residues were found next to each other.

Table 4.1: Top 6 residue pairs with Z-score>4 from Sly1 MIp data.

Residue1	Residue2	MIp score	Z-score
I543	S544	0.137	7.95
L28	N29	0.134	7.78
F404	A405	0.132	7.66
S68	V69	0.118	6.84
L58	S62	0.116	6.72
V539	G540	0.115	6.66

The numbers in the first two columns indicate the residue position as in *Saccharomyces cerevisiae* Sly1p and the letters indicate the amino acid at that position.

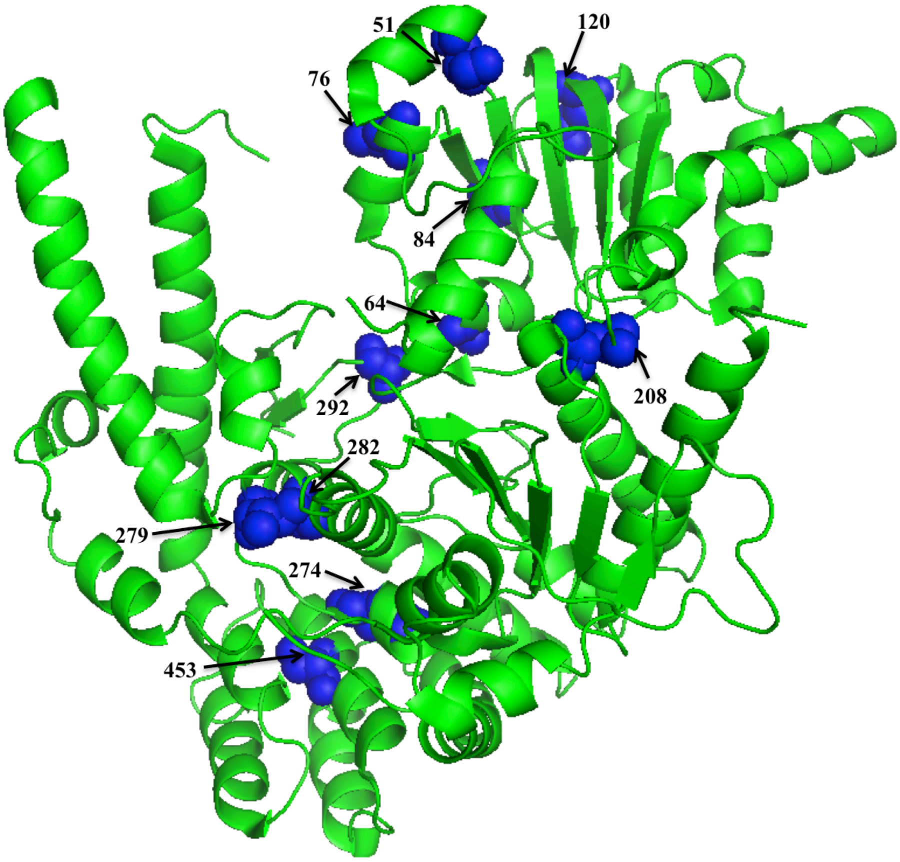


Figure 4.3: An example of residues with Z-score>4 appearing as groups, marked on the structure of *Saccharomyces cerevisiae* Sly1 in complex with the N-peptide of Sed5 (1MQS).
Sly1 structure is shown in green and Sed5-N-peptide structure is shown in orange. The distantly lying residues identified in the initial analysis and shown in Table 4.2, are shown in blue. The residues lie distant from each other and appear to form a connected network.

Table 4.2: Table of residues appearing as groups, identified in the initial analysis on Sly1.

Residue1	Residue2	MIp Score	Z-score
F274	I279	0.102	5.90
L76	S84	0.095	5.53
K120	S453	0.094	5.43
A64	L76	0.085	4.91
L76	K120	0.082	4.74
L76	R292	0.079	4.60
A64	R292	0.077	4.44
I51	L76	0.075	4.34
S84	K120	0.0743	4.29
I279	R292	0.07425	4.28
S62	L76	0.0731	4.23
F274	C282	0.0713	4.12
L76	C282	0.0710	4.08
S84	R292	0.0705	4.07
I279	C282	0.0704	4.07
L76	S453	0.0703	4.06
A64	I279	0.069	4.02

The numbers in the first two columns indicate the residue position as in *Saccharomyces cerevisiae* Sly1p and the letters indicate the amino acid at that position.

Covarying positions, as shown by previous studies on other proteins, are usually occupying adjacent positions (Gloor et al. 2005; Gobel et al. 1994; Wollenberg & Atchley 2000; Tillier & Lui 2003; Fodor & Aldrich 2004; Martin et al. 2005; Fares & Travers 2006; Gouveia-Oliveira & Pedersen 2007; Dunn et al. 2008; Marks et al. 2011; Hopf et al. 2014). In earlier reports often the top high scoring residue pairs were found to be lying close to each other and in close proximity to the enzyme's catalytic/substrate-binding site. The earlier studies generally used a cut-off, for example Z-score>4.0, and then considered the highest scoring pairs. In my case, using a preliminary cut-off seemed to limit the results to certain sites and to neglect some other sites, suggesting that it might be advisable to analyze the entire MIp data set first. The results obtained during this initial analysis may hint at conformational changes and allosteric coupling as discussed in the recent literature (Lockless & Ranganathan 1999; Süel et al. 2002; Halabi et al. 2009; Baussand & Carbone 2009; Burger & van Nimwegen 2010; Dib Linda 2012b).

When I started the co-variation analysis on SM proteins, I expected to obtain results similar to previously published studies. That is, I expected to find co-varying residues pairs in close proximity, possibly also lying close to the SM-Syntaxin binding sites. However, the result from my initial covariation analysis on Sly1 appeared to deviate from earlier reports and I often observed covarying residues that appeared to from a network within the tertiary structure. These deviating findings on the SM protein Sly1 made it challenging, at first glance, to understand the significance of the obtained MIp scores. Apart from few studies (Gloor et al. 2005; Halabi et al. 2009; Burger & van Nimwegen 2010; Dib Linda 2012b) occurrence of distantly lying residues was not shown or published for large-scale studies. Instead, most of studies were performed only on small proteins domains or enzymes. So possibly, networks of co-varying positions do occur in more complex proteins like Sly1. Maybe they even can be found in many other protein types as well, but the current strategy to analyze the co-variance data focuses the attention on co-varying residues in close proximity, which are easier to interpret. In fact, it has not been shown that larger networks of co-varying positions are indeed important for the structure and function of proteins. As my initial results did not appear in accord with published findings, it seemed possible that the current approaches to analyze the MIp scores by mostly selecting the most high-scoring pairs were not suitable and sufficient to analyze the co-varying positions in more complex proteins like Sly1. I therefore proceeded to device a new analysis approach, as I will outline in the following.

4.2 Development of an improved analysis approach for detecting groups of covarying residues and its application on the test cases from prior publications

To further investigate and understand the initial observations, an improved analysis pipeline was developed to visualize and analyse the networks of covarying residues. To test and improve the analysis pipeline, it was first applied on dataset of enzymes obtained from prior publications. This was done in order to see whether the developed approach helps to have a better look at the entire covariation data and to find out whether this approach can detect groups of covarying residues in a known dataset.

4.2.1 Test case 1: The Methionine amino peptidase1 (MAP1) contains two types of covarying positions

In the following, this approach will be explained along with the example of the Methionine amino peptidase (MAP), previously studied for positional coevolution by (Gloor et al. 2005). The basis for the analysis was a structure-based alignment of 174 protein sequences of the Methionine amino peptidase (MAP). The available crystal structure of methionine amino peptidase from *Escherichia coli* (PDB 1C24) was used for residue mapping.

Gloor et al. had calculated the mutual information (MI) scores between all ungapped positions in the alignment. To reduce the influence of entropy on MI, they normalized the raw MI scores. The raw MI scores were divided by the joint entropy of the position. Then they calculated a Z-score for each normalized ratio and used a Z-score cut-off of 4 to identify coevolving positions. They found two kinds of coevolving sets. One subsets contained sites that are not near the active sites or the subunit interface. Such sites were found to coevolve only with another site in close proximity. The other subset comprised of interconnected sites that had Z-score > 4 with more than one positions. These sites were found near the active site of the protein and belong to one large cluster.

They had speculated that the latter group of residues might have functional importance and might form a network along with some of the conserved residues in that region. By contrast the isolated pairs were found to be lying away from the active site and on the surface of the protein. They reasoned that these pairs have coevolved to maintain the local structure and that thus might be involved in protein folding and structural stability.

Application of MIp on MAP1 protein

For my analysis, I used MIp on the MAP1 dataset from Gloor et al. 2005. MIp is a statistical correction of MI for removing the phylogenetic noise. This resulted in residues that were identified in the earlier prediction (Gloor et al. 2005). Some additional residues were also found, probably because I did not use the same stringent cut-off of Z-score=4.

A detailed look at the 50 highest scoring residue pairs (Table 4.3) confirmed that some of the residue pairs were lying in close proximity, for example residue pairs 6 and 179, 154 and 158, 122 and 230, and 147 and 187, while some residue pairs were lying distant from each other in the tertiary structure, for example, residue pairs 78 and 202, and 95 and 177 (Figure 4.4). Some residues appeared to form connections with many other residues, thus forming groups, for example, residues, 21, 59, 67, 71, 78, 95, 96, 99,101, 109, 110,112, 177, 201 and 202.

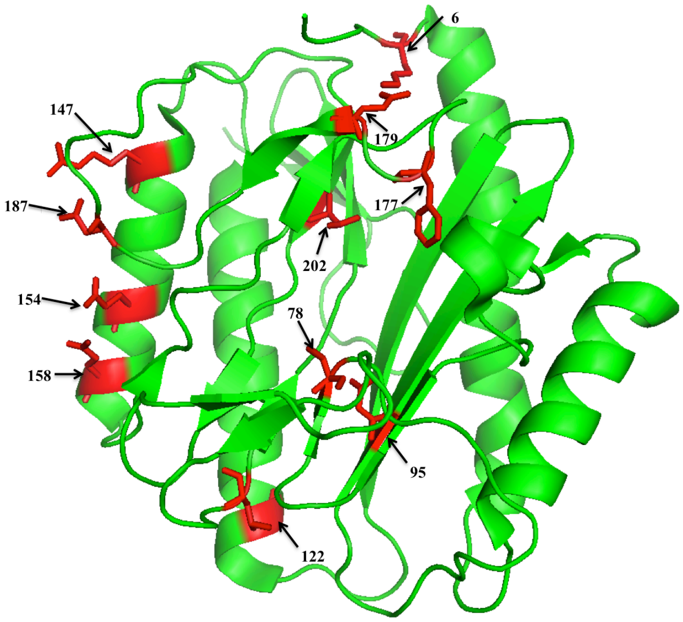


Figure 4.4 : Some of the residues with high MIp score mapped on the structure of *Escherichia coli* MAP1 (1C24).
E.coli MAP1 structure is shown in green and the residues with high MIp scores are shown in red. Some of these residues lie close to each other, while some lie distant from each other.

Table 4.3 : Top 50 covarying residues ranked according to their MIp scores.

Residue 1	Residue 2	MIp scores
C78	T202	0.1582
K6	E179	0.1560
Q154	E158	0.1530
G122	L230	0.1470
R147	D187	0.1424
N95	F177	0.1410
N95	S110	0.1392
D187	N192	0.1370
C59	F177	0.1332
M217	K224	0.1304
Y168	T216	0.1301
E190	N192	0.1294
A58	I101	0.1288
Y48	Q53	0.1263
C59	T99	0.1262
N95	T99	0.1230
C78	N95	0.1211
S110	M112	0.1192
N46	L60	0.1186
V140	V240	0.1140
A21	T99	0.1130
I101	S110	0.1129
C78	T99	0.1120
C78	F177	0.1115
N95	T202	0.1104
T99	F177	0.1095
T99	T202	0.1092
L60	F177	0.1088
N95	F201	0.1077
F177	F201	0.1063
T99	S110	0.1058
A21	T109	0.1049
D187	E190	0.1041
F177	T202	0.1037
E148	E190	0.1030
K67	N95	0.10292
S110	F177	0.10290
I144	E148	0.1010
C59	N95	0.1001
K67	F177	0.0995
V239	L248	0.0990
C59	I71	0.0989
C59	C78	0.0983
A121	N95	0.0963
K67	C78	0.09631
G150	T202	0.0960
K67	T202	0.0953
C59	L60	0.0935
E148	D187	0.0922
K67	P142	0.0910

The numbers in the first two columns indicate the residue position as in MAP1 of *Escherichia coli* and the letters indicate the amino acid at that position.

Development of a heatmap visualization

To better understand the initial observations, I wanted to have a detailed look at the entire covariation (MIp) dataset rather than looking only at certain high scoring sites. Heatmap visualization was thus developed to have a complete picture of the distribution of covarying residues within a protein. It is a graphical representation of data, where individual pairwise values contained in a matrix generated by covariation analysis method, are represented by varying degree of colors. It basically maps the covariation matrix values to an appropriate color map and displays it as a symmetrical grid. It allows easier comparative analysis and provides a compact view of the complete covariation dataset. It also offers to the user a selective choice of color scale, with default value option, where red regions (dark red, red, dark orange, orange, light orange) indicate a higher covariation scores and blue region (lemon, light green, dark green, blue, purple) indicate a lower covariation scores. It thus helps to highlight and detect interesting groups or clusters of residue with high covariation scores.

Note that Liu & Bahar in 2012 also used heatmap to visualize and map the MIp scores in their study. This was developed in parallel to my study. Some other studies

(Gouveia-Oliveira et al. 2009; Dutheil 2011; Halabi et al. 2009; Ackerman & Gatti 2011) also used heatmaps like plots to represent the covariation score.

Heatmap of MIp results from MAP1 protein

The MIp scores generated for the Map 1 dataset from Gloor et. al. were visualized with the help of a heat map (Figure 4.5). This showed that certain hot spots for high covarying residue pairs.

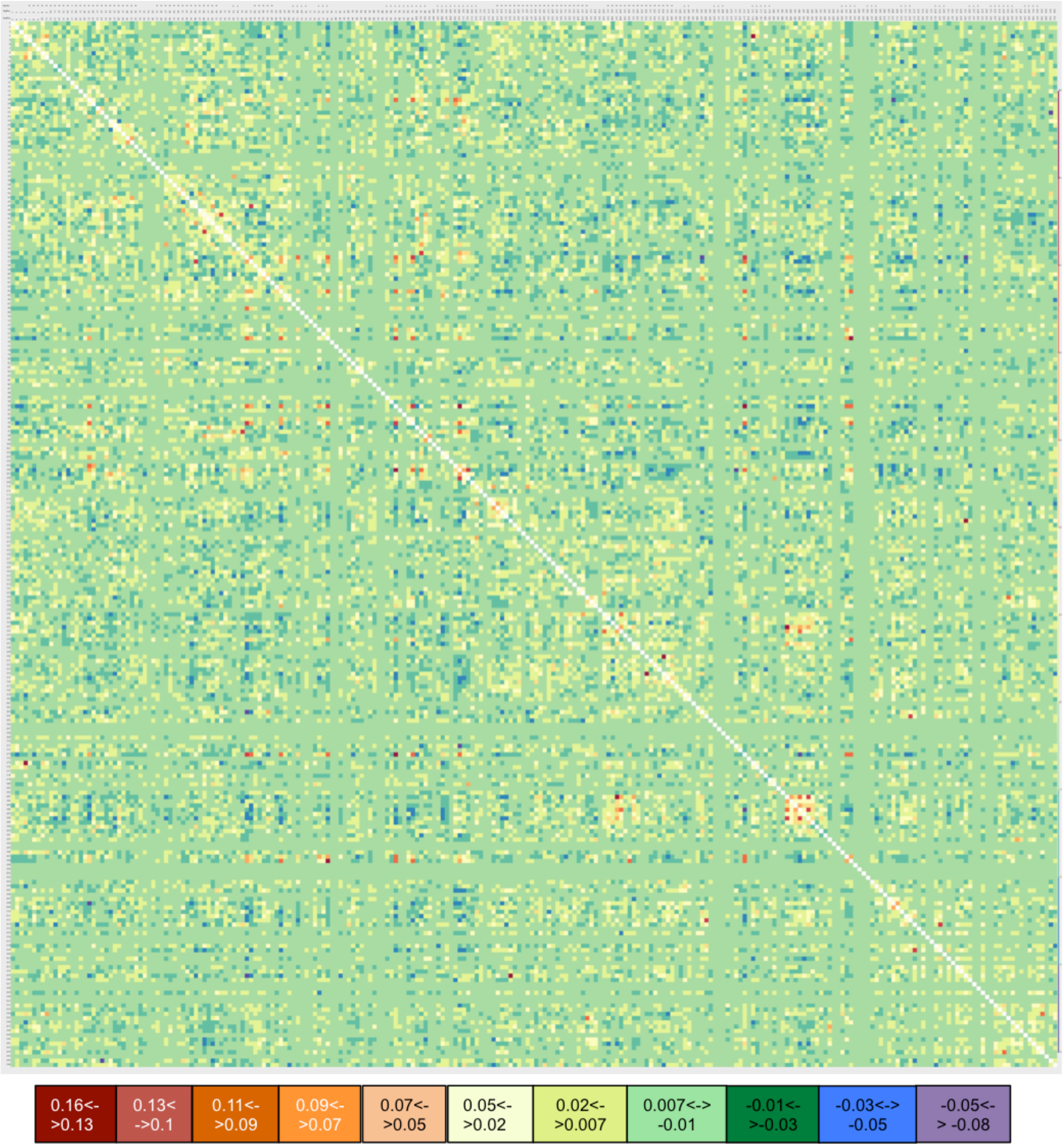


Figure 4.5: Heat map of MIp scores of MAP1 dataset from (Gloor et al. 2005).
The heatmap is arranged as per the residue numbers from the structure of *Escherichia coli* MAP1 (1C24). The color code bar represents the color corresponding to the range of MIp score. The dark red, red, dark orange, orange, represent the high MIp scores and light green, dark green, blue, purple represent the low MIp scores. The original heat map can be found in the supplementary data (provided on request) that can be enlarged to have a clear view of the residue numbers.

Development of a network theory based clustering method

To further investigate and understand initial observations, I needed to improve the analysis such that it would include the groups or networks of residues as well instead of only the pairs of residues. However, it is difficult to perform an exhaustive search for groups of residues on the entire dataset due to the existence of high number of possible combinations within a protein. It is also difficult to examine the role so many residues whether occurring in groups or in pairs. Clustering techniques from network theory are standard methods that can be used to handle such issues.

In order to identify groups or networks of strongly covarying residues, hierarchical clustering based average linkage methodology was applied on the covariation dataset considered as a residue-to-residue graph, with the residues as nodes and the covariation scores as edge weight. A stopping criterion was used for the average linkage clustering. This step was used to remove weak connections between the residues, so that only those residues that have strong covariation were considered. Each cluster was then scored and ranked. Clusters can be scored either based on the average weighted degree of the cluster or based on the average weight of the cluster and then ranked in the decreasing order of the score. In general, average weighted degree scoring of the cluster should result in large clusters being ranked higher and can be used to identify the groups of covarying residues. Average weight of the cluster should result in large clusters being ranked higher and can be used to identify the groups of covarying residues. A cut-off was then used to select the top clusters that have significant residue-to-residue interactions within the cluster (see Methods Section for details).

A few other studies were performed that also used clustering for detecting the groups of coevolving residues within a molecule, however they all used different methods. The study carried out by Dutheil & Galtier in 2007 used a tree-based method called CoMap to calculate the covariation score and complete linkage clustering to identify the groups of coevolving residues. The method they used for identifying the coevolving residues was not a sequence-based method as MIp, which was used in the current study. Also the complete linkage clustering always results in a complete graph, where all the nodes within a graph are connected to each other. Such a cluster would not give a clear indication about the strengths of the connection within the cluster. Liu et. al. in 2008 used mutual information (MI) and spectral graph partitioning to analyze the correlated mutations in HIV-1 protease. MI method does not account for the phylogenetic noise due to shared ancestry of the sequences. Studies have shown that it is crucial to distinguish correlation caused by the phylogenetic noise from the functional correlation (Dunn et al. 2008; Dutheil 2011). In my approach, I use the improved and wieldy used method, MIp, which corrects the phylogenetic bias and has been shown to out perform the tree-based methods for identifying coevolving positions (Caporaso et al. 2008; Dutheil 2011). Another study performed by Dib & Carbone in 2012, used a new combinatorial approach, Blocks In Sequences (BIS), to identify clusters of coevolving blocks using their previously developed automatic clustering algorithm CLAG (Dib & Carbone 2012) . CLAG (CLusters AGgregation) is an unsupervised non-hierarchical clustering algorithm designed to cluster a large variety of biological data and to provide a clustered matrix and numerical values indicating cluster strength. It allows for a position to belong to several clusters. In my approach, I use a hierarchical clustering method that does not allow a position to belong to several clusters and they are scored separately based on their the average weighted degree score.

Application of clustering on MAP1 protein

Average linkage clustering with stopping criteria of 2-σ as and cluster selection cut off of 90% was then applied on the MIp data obtained on MAP1 alignment from Gloor et.al. This resulted in three clusters. Most of grouped residues were identified among the top selected clusters (Figure 4.6, Figure 4.9) along with some new residues.

The selected top clusters were mapped onto the structure of MAP1 from *E.coli*. Each cluster was restricted to a certain region of the protein. Most of the residues within each cluster were in contact or in significant close proximity (Figure 4.7), similar to what was shown in the earlier study. However, cluster 1 (shown as red spheres) had residues that were lying in close contact as well as distant from each other. In contrast to the earlier result, the residues within this cluster appeared to be interconnected and form a network of residues. Again similar to what was already shown in Gloor et. al., I found some of the residues that were also located in close proximity to the active sites.

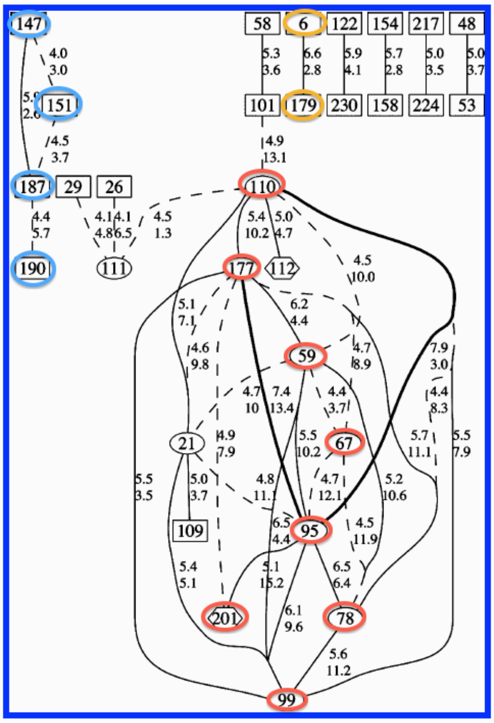


Figure 4.6 : Residues identified in the current analysis and in the previous study obtained for MAP1 alignment.

Linkages between positions in the MAP1 alignment with residue numbers from a representative structure *E.coli* MAP1(1C24). Residues that make contact with their partner are enclosed in rectangular boxes; residues at or near the active site are enclosed in ovals, and the hexagonal nodes represent the residues for which neither of the other associations holds. Bold lines represent Z-scores greater than 7; solid lines are Z-scores between 5 and 7, and dashed lines represent Z-scores between 4 and 5. The top number to the right of each line is the Z-score; the bottom number is the distance of closest approach between the two residues. The residues identified by the current clustering approach are indicated by circles. The residues circled in red are identified in the first cluster, blue in the second cluster and orange in third cluster. The figure is modified from (Gloor et al. 2005).

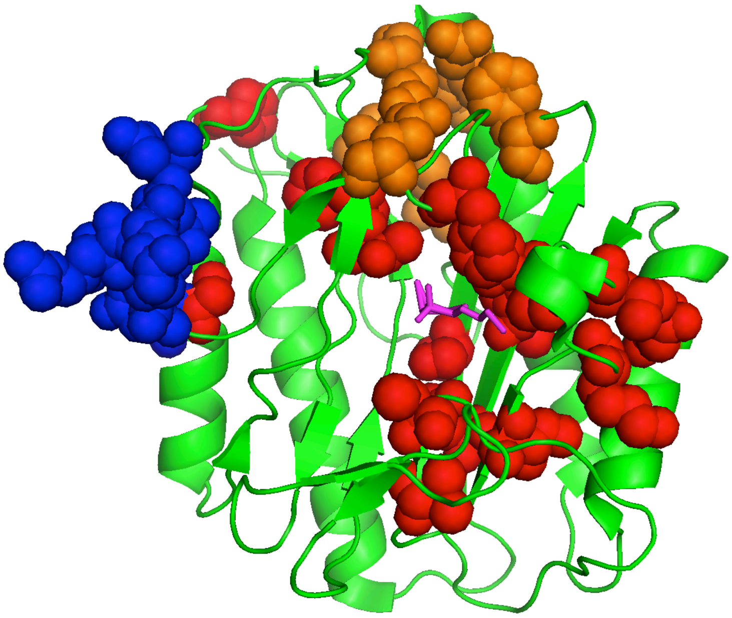


Figure 4.7: Residue from the selected clusters marked on the structure of *Escherichia coli* MAP1 (1C24).

E.coli MAP1 structure is shown in green. The ligands-MPJ & CO are in magenta. Residues in red are from cluster1, in blue from cluster2 and in orange from cluster3.

Clustering of the MAP1 MIp dataset of MAP1 corroborated the earlier reports by Gloor et al. of a network of co-varying positions. They identified these groups by simply selecting residue connections with Z-score>4.0. With the clustering approach, I found the residues that appeared as groups in the previous analysis as separate clusters. The first cluster contained the residues that had been identified as a large group, while the second cluster contained the residues identified before as small group. The third cluster contained one of the previously identified residue pairs. In contrast to the earlier observation (Gloor et. al.), this pair was part of a network with several other residues, although it clearly is the highest scoring pair within this cluster (Figure 4.6).

However, some high-scoring residue pairs were not included in the topmost clusters (Figure 4.8). These residues were not identified because the current cluster approach does not rank them at top-level, as it was designed to preferentially detect groups of co-varying positions. The approach was thus successful in identifying the groups of covarying residues lying near the active site and which are possibly involved in functional constrains of the protein (Gloor et al. 2005).

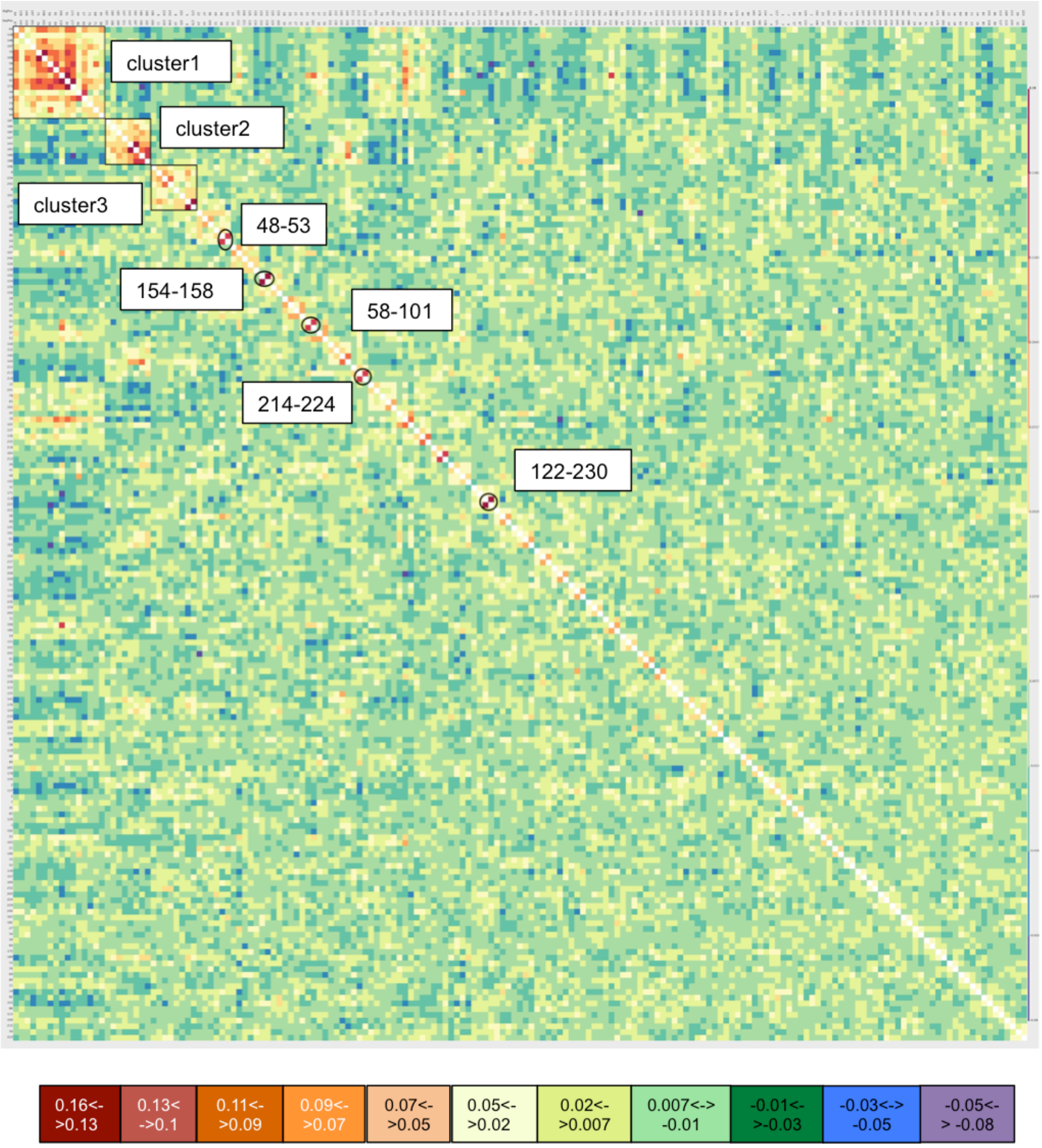


Figure 4.8: Heatmap of MIp scores arranged according to the ranking of clusters from the analysis on MAP1 dataset from Gloor et al. 2005.
Selected clusters are marked in Black squares. Residues previously shown as important but not identified in current study are shown by black circle. The color code bar represents the color corresponding to the range of MIp score. The dark red, red, dark orange, orange, represent the high MIp scores and light green, dark green, blue, purple represent the low MIp scores. The original heat map can be found in the supplements (provided on request) that can be enlarged to have a clear view of the residue numbers.

Development of a network visualization

To visualize and analyze the groups or networks of residues obtained from the improved analysis pipeline, a network view was also developed. After the clustering, each cluster can be represented as a network. In this view, nodes are shown as circles, representing the residue. The circle or nodes are connected by an edge representing the covariation score between the two residues. The color of the edge represents the strength of the covariation score, with red being highest and blue being the lowest.

The color and size of the node represents the importance of the residue in terms of the average of the score of the edges connected to it.

Network view of clusters from MAP1 protein

The clusters obtained on MAP1 can be visualized with the help of the developed network view. This cluster had many of the previously identified grouped residues. Many new connections were also found in this cluster. The cluster view showed that the strongest connection within the cluster is of residue 78 and 202. Although residue 78 was identified earlier as well, this particular connection was lost in the previous analysis. The figure also showed that residue 95 is the strongest node with most number of strong connections. This is similar to what was shown before (figure drawing) where 95 had many string connections with other residues.

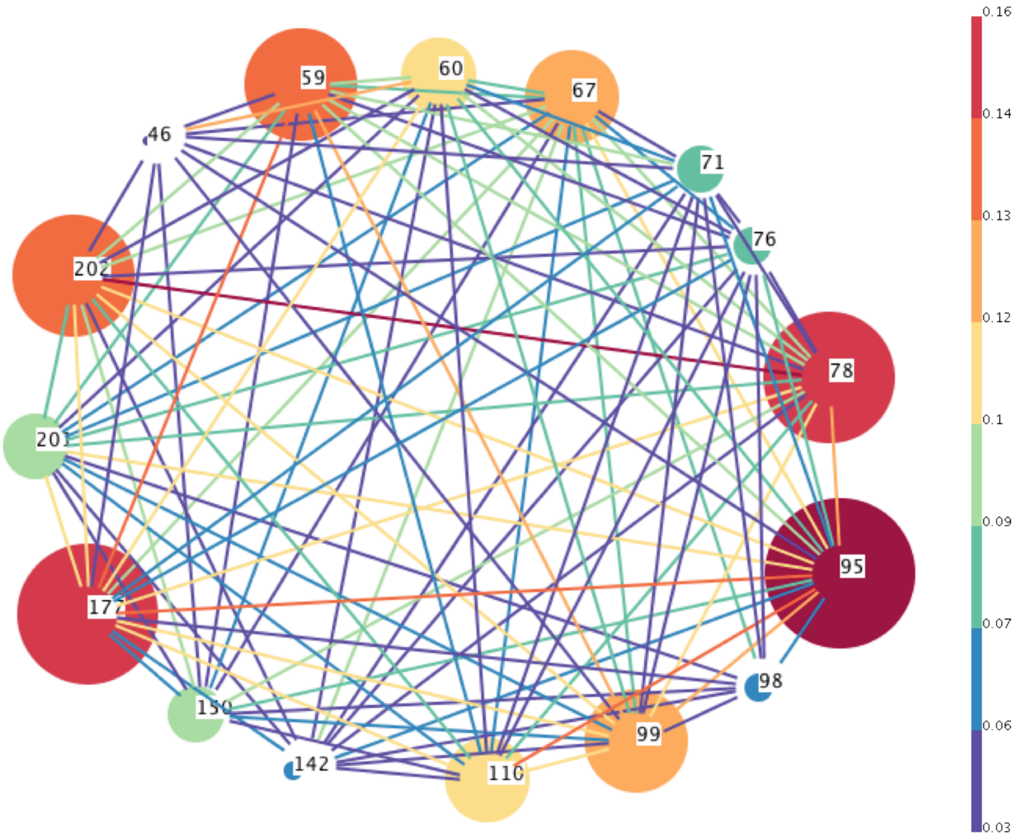


Figure 4.9: Network of residues of cluster 1 from MAP dataset.

The color of the edge represents the strength of the covariation score. Nodes are represented by circle with residues number from the representative structure 1C24. The color of the edge represents the strength of the covariation score, shown in the bar on the right, with red being highest and blue being the lowest. The color and size of the node represents the importance of the residue in terms of the average of the score of the edges connected to it.

Thus, the cluster approach detected large networks of covarying residues in MAP1. Note, however, that the cluster approach did not detect well individual high scoring paired residues. With the help of the example of previously published data, where only MI was used, this showed that my approach can identify additional information about groups of covarying residues.

4.2.2 Test case 2: UDG (Uracil-DNA glycosylase) data set from Lui & Bahar in 2012

The combinatorial analysis pipeline was applied on a dataset of a DNA repair enzyme, uracil-DNA glycosylase (UDG), previously studied for sequence coevolution (Liu & Bahar 2012). They showed that coevolving pairs or even groups of coevolving sites, distinguished by high MIp score, have uniquely high mobilities (obtained by Gaussian network model (GNM) analysis) and are involved in substrate recognition. The dataset had 6214 sequences and 163 sites and a PDB structure from UDG of human (PDB ID: 4SKN) was used for mapping the residues.

Application of MIp on UDG protein

MIp was initially applied on the dataset without any conservation check and gap column removal. A detailed look at the top 50 covarying residue pairs of the UDG dataset showed that most of the residues lying adjacent to each other (Table 4.4). Some residues had connections with more than one residue and formed small groups, such as, residues 131, 132, 133 and 134. Among the top 50 covarying residue pairs, none was found to be lying far apart in the structure.

The complete MIp dataset was visualized with the help of a heat map (Appendix figure A.10 (can be provided on request)). The heatmap of MIp scores revealed that the covarying residues were located at certain regions in the protein. They did not appear to be spread all across the protein but were restricted to certain hot spots within the protein. It also revealed that most of the high covarying residues were lying adjacent to each other in the sequence of the protein.

Table 4.4: Top 50 covarying residues ranked according to their MIp scores from UDG dataset.

Residue 1	Residue 2	MIp score
189	190	0.207
182	183	0.169
133	134	0.166
132	133	0.162
171	172	0.151
131	132	0.141
185	186	0.138
189	192	0.132
132	134	0.129
169	172	0.126
276	277	0.124
292	293	0.123
209	211	0.123
209	210	0.123
236	237	0.120
281	283	0.120
189	191	0.119
190	191	0.116
168	172	0.116
187	188	0.116
181	182	0.113
190	192	0.112
186	187	0.107
180	181	0.107
208	209	0.106
190	193	0.101
144	152	0.101
168	171	0.101
169	171	0.099
144	268	0.098
282	285	0.097
235	237	0.096
188	194	0.095
236	240	0.095
185	187	0.095
189	193	0.094
174	176	0.094
164	166	0.093
291	292	0.093
285	286	0.092
182	186	0.092
183	186	0.092
181	183	0.091
186	189	0.091
131	133	0.090
185	188	0.090
289	290	0.090
170	172	0.090
176	189	0.090
188	189	0.088

The numbers in the first two columns indicate the residue position as in USG of human. Residues appearing as a group and also identified in the current clustering approach are colored as red.

Application of clustering on UDG protein

To identify the clusters of covarying residues, average linkage clustering with the stopping criterion of 2-σ and a cluster selection cut off of 90% of the cluster scores was then applied on the MIp data. This resulted in five clusters of different sizes. Each cluster was found to restricted to a distinct region of the protein (Figure 4.10). The residues involved in substrate recognition and interaction with DNA that were previously identified were detected by my improved approach as well.

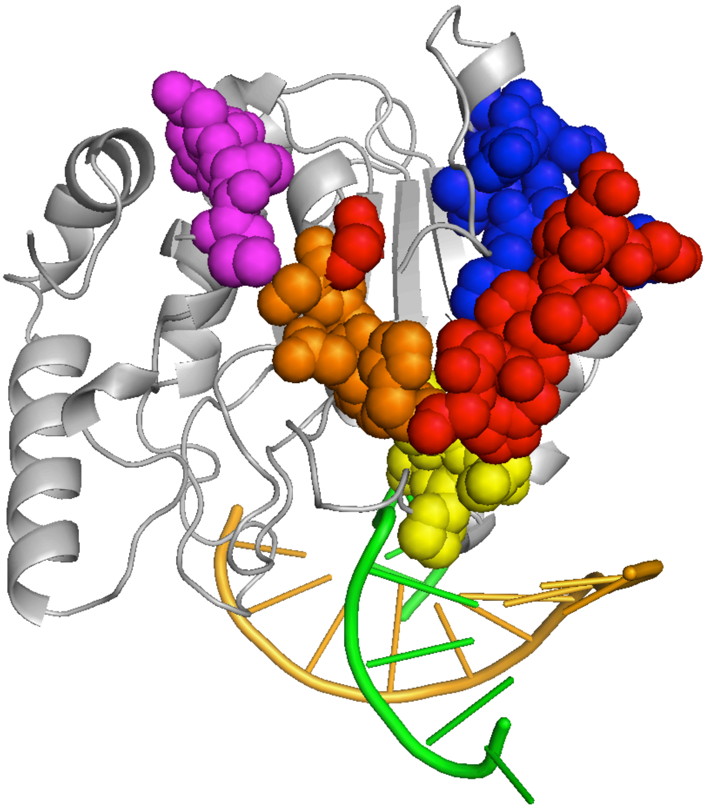


Figure 4.10: Residues from the top 90% clusters mapped onto the structure of human UDG (PDB id: 4SKN).
Human UNG structure is shown in green. Residues in red are from cluster1, in blue from cluster2 and in orange from cluster3, yellow from cluster4 and magenta from cluster5. The interacting DNA is shown in green and light orange.

Each cluster had residues that were lying next to each other as well as distant to each other. The near-by and distant residues within each cluster appeared to be connected by a chain of covarying residues (Figure 4.11). The high scoring pairs within the cluster were lying next to each other. For example, residues 182 and 183 from cluster 1 were lying adjacent to each other.

The authors considered only the list of high MIp scoring pairs in their analysis but they also found that some of the high scoring residues were covarying with many other residues (Liu & Bahar 2012), thus essentially forming clusters. For example, residues 180, 181,182, 183, 185, 186, 187, 188 and 194 (marked in red in the Table 4.4) have high covariation score with each other and appeared to form a group of inter-connected residues. They defined these cluster by observation based on the list of high scoring residue pairs. In my analysis, all these interconnected residues were identified in one cluster (cluster 1, Figure 4.11). Thus my cluster approach indeed specifically identifies co-varying networks within this protein.

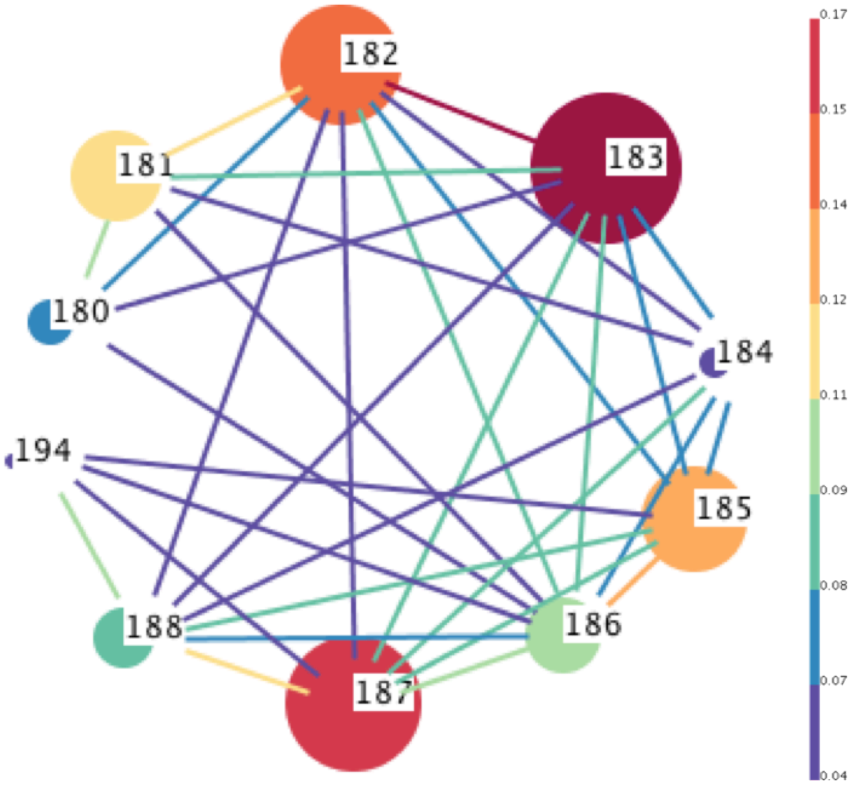


Figure 4.11: Network of residues of cluster 1 from the UDG dataset.
The color of the edge represents the strength of the covariation score. Nodes are represented by circle with residues number from the representative structure 4SKN. The color of the edge represents the strength of the covariation score, shown in the bar on the right, with red being highest and blue being the lowest. The color and size of the node represents the importance of the residue in terms of the average of the score of the edges connected to it.

The selected clusters are shown in a heatmap that is organized according to the rank of the clusters (Figure 4.12). The heatmap shows that the selected clusters had the top 50 high scoring residues embedded within them. Some previously identified residue pairs involved in DNA interaction and having high MIp score, such as 276 & 277, 278 & 279 (Liu & Bahar 2012), were not identified by the presented approach. The heat map (Figure 4.12) showed that even though these sites had a high MIp score, they did not form larger networks.

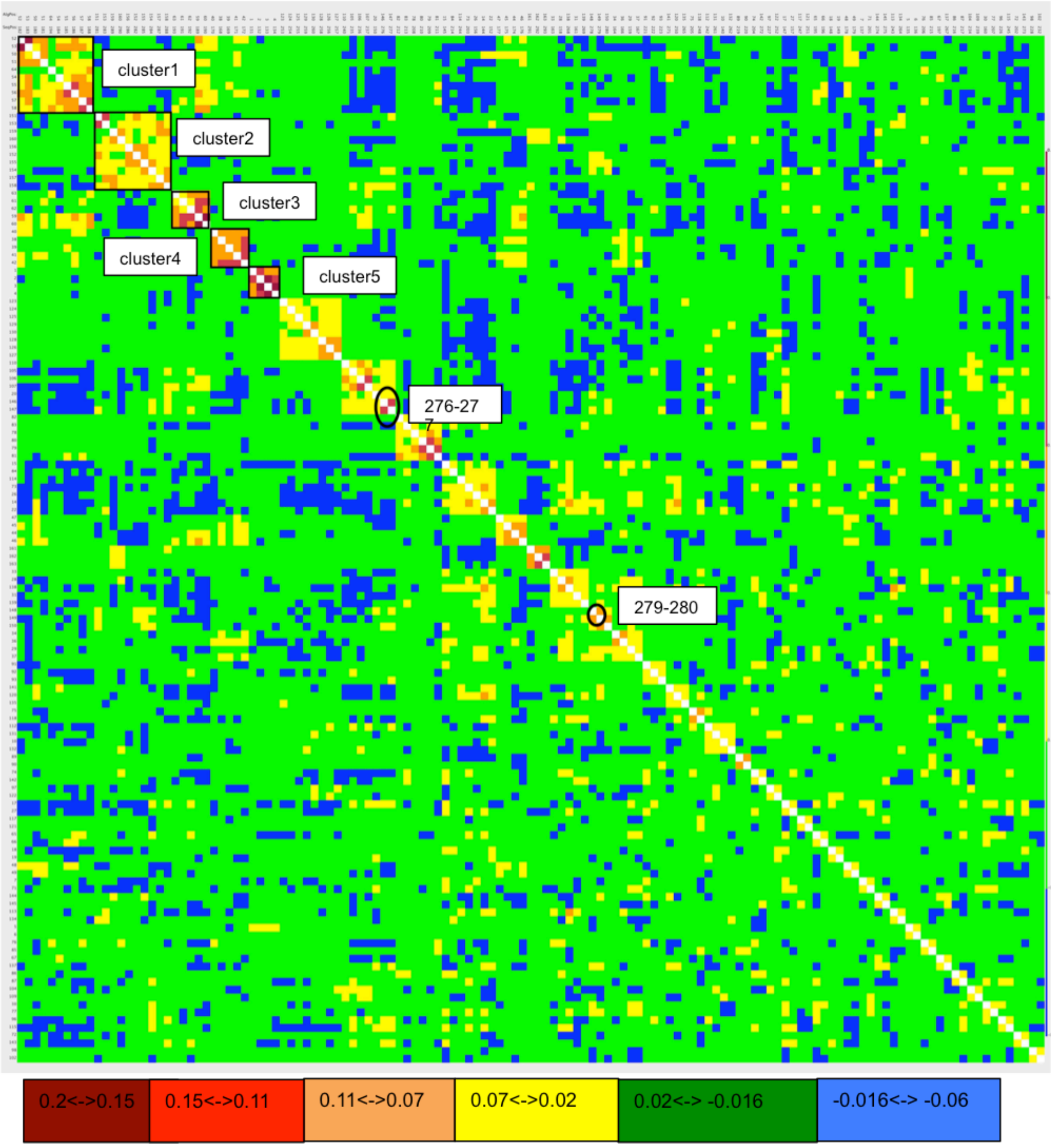


Figure 4.12 : Heatmap of MIp scores arranged according to the ranking of clusters from the analysis on the UDG dataset.

Selected clusters are marked in Black squares. Residue pairs previously shown as important but not identified in current study are shown by black circle.

Here using two test case examples from previously published data, I showed that my approach could identify clusters of covarying residues. The earlier studies did suggest the existence of such groups by inspecting the list of high-scoring residue pairs in a non-automated fashion. My approach of using a computational method is able to describe the entire cluster of the covarying residues.

4.3 Application of the improved analysis pipeline on SNAP proteins

I next applied the improved analysis pipeline on another protein of the vesicle fusion machinery. For this I selected the SNAP protein family as a test case, because this protein family is structurally and functionally different from SM proteins. SNAP proteins participate in the dissociation of the SNARE complex together with NSF ATPase. As already mentioned in Introduction, they act as an adaptor protein between NSF and SNARE complexes. Three molecules of SNAPs, along with one SNARE complex and one NSF hexamer form a complex that drives the disassembly of the SNARE complex. Sequences of SNAP proteins have also been collected and analyzed previously in my group. The phylogenetic analysis of SNAP has shown two distinct groups of SNAPs with α-SNAP and β-SNAP in one group and γ-SNAP in another group. γ-SNAP was found to be lost in Fungi except in some more basal Fungi and β-SNAP was shown to be a duplication of α-SNAP in Vertebrates (Kienle 2010). The structure of **γ**-SNAP from *Danio rerio* has a fold similar to the structure of *Saccharomyces cerevisiae* α-SNAP, Sec17. The monomer A of **γ**-SNAP aligns with Sec17 with a root mean square deviation (rmsd) of 3.4Å (Bitto et al. 2007). The N-terminal portion and the C-terminal portion of **γ**-SNAP align well with the corresponding regions of the Sec17, while the poor overall alignment of both structures is due to difference in the twist from the middle region of **γ**-SNAP (Bitto et al. 2007).

SNAPs have several tetra-tricopeptide repeats (TPR). TPR repeat containing proteins are thought to be rigid and not undergoing large conformational changes but they do have some flexibility for ligand binding (Cortajarena & Regan 2006). SM proteins, on the other hand, are globular and have two binding sites for one binding partner. Thus, SNAP proteins are different from enzymes and are a more static example than SM proteins.

The alignments were generated and were further refined, by using an iterative block strategy (see the Method Section for details). Any column that had more than 10% gaps was removed from the alignment. Columns with more than 80% conservation were also not considered in the analysis (Figure 4.13). The resulting data set for the entire SNAP family contained 740 sequences and 271 columns. The data set for α-

SNAP had 521 sequences and 278 columns. The crystal structure of Sec17 (homolog of α-SNAP) from *Saccharomyces cerevisiae* (PDB 1QQE) was used for residue mapping. The data set for the γ-SNAP had 219 sequences and 280 columns and the crystal structure of γ-SNAP of *Danio Rerio* (PDB 2IFU) was used for residue mapping.

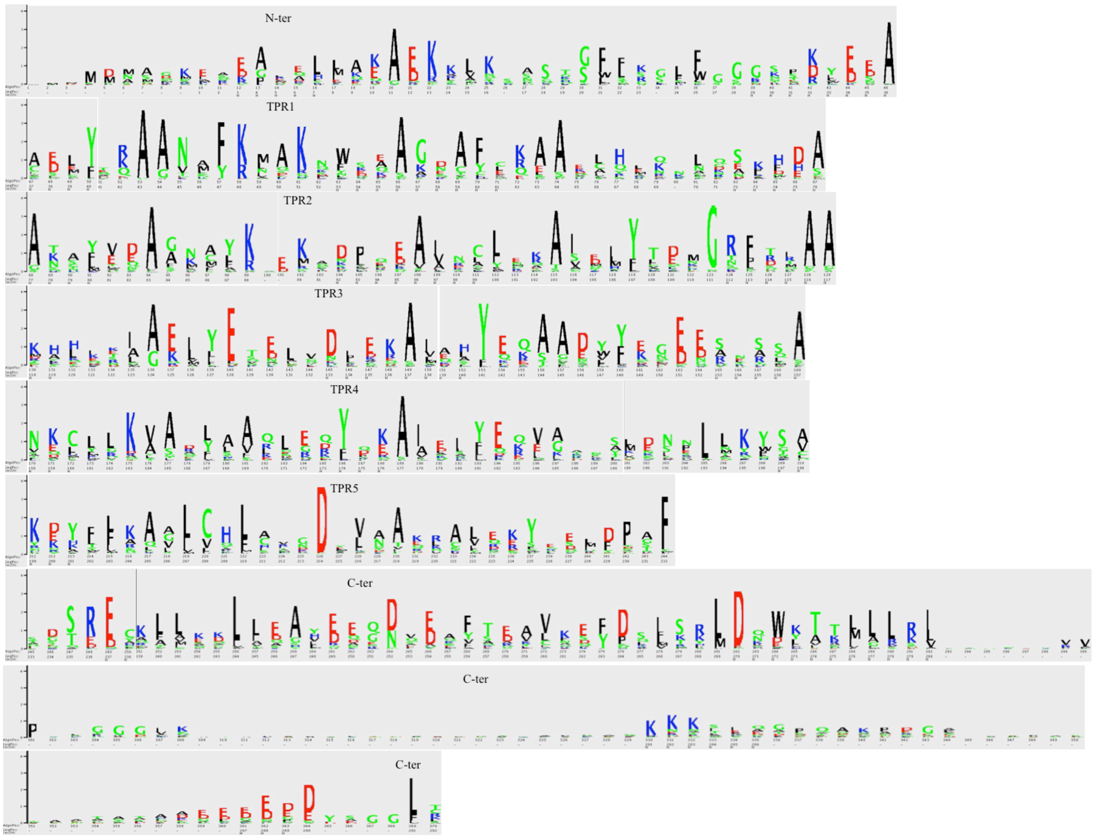


Figure 4.13:Sequence logo of the entire SNAP alignment.
The logo is arranged according to the sequence of Sec17, before removing 90% gapped columns or 80% conserved columns. The logo is arranged to show the structural elements, N-terminus, C-terminus, and different TPRs.

Correlated Mutation analysis of SNAP proteins

MIp scores were calculated for the alignment of the entire SNAP family, was well as for separate alignments of the two subfamilies α-SNAP and γ-SNAP. For the entire SNAP family most of the highly covarying residues were found in the N- and C-terminal regions. Only few were found in the central TPR region (Figure 4.14). Overall, most pairs with positive scores were lying near or around the diagonal in the heat map (Figure 4.14, Appendix Figure-A.25, Figure-A.46 (Appendix figures can be provided on request)). This indicates that most of the residues score highly with neighboring residues in the sequence.

Comparable results were obtained for the two subfamilies. The major difference between the two subfamilies was that for the γ-SNAP subfamily residues in the central TPR region did not score as high as for the α-SNAP subfamily.

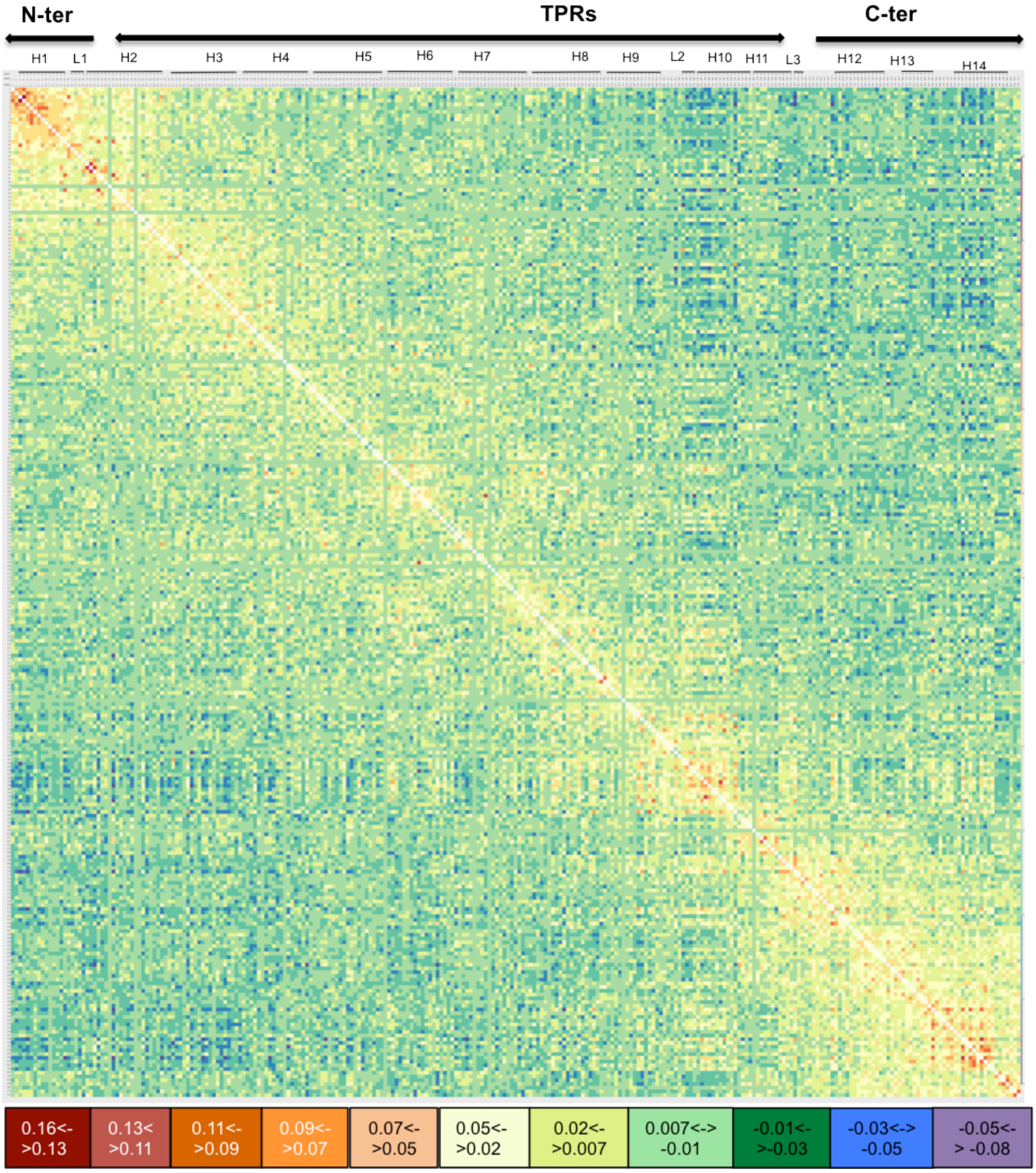


Figure 4.14: Heat map of MIp scores from the entire SNAP alignment.
The heatmap is arranged as per the residue numbers from the structure of *Saccharomyces cerevisiae* Sec17.The color code bar represents the color corresponding to the range of MIp score. The dark red, red, dark orange, orange, represent the high MIp scores and light green, dark green, blue, purple represent the low MIp scores. The original heat map can be found in the supplements that can be enlarged to have a clear view of the residue numbers. The structural element are marked on the top of the heatmap; H1, H2… : different helices, L1.L2…: loops, N-ter: N-terminal region, TPRs: tetra-tricopeptide repeats, C-ter: C-terminal region. The original heat map can be found in the supplements (provided on request) that can be enlarged to have a clear view of the residue numbers.

As the results for the entire family and the two subfamilies were very similar, only the (representative) results of α-SNAP analysis will be discussed further in detail.

Table 4.5: Top 50 covarying residues ranked by their MIp scores from the α-SNAP dataset.

Residue1	Residue2	MIp Score
D3	P4	0.1387
D200	K204	0.1291
D229	K273	0.1278
L24	F25	0.1266
S30	Y31	0.1197
T275	K279	0.1161
K120	C140	0.1149
E6	K9	0.1138
K261	N265	0.1125
E6	L8	0.1112
I274	N278	0.1104
D74	R110	0.1066
D171	G172	0.1057
L53	S194	0.1033
L291	L292	0.1020
M1	P4	0.0992
K14	Y31	0.0986
E6	L7	0.0981
F119	I162	0.0965
E6	R10	0.0959
Q209	T221	0.0954
R50	L53	0.0936
I162	K163	0.0926
M1	D3	0.0911
S2	K14	0.0909
L198	F266	0.0906
C260	L291	0.0904
K85	R114	0.0902
L53	N118	0.0894
Q112	1N18	0.0893
D3	L7	0.0884
Q285	Q286	0.0881
IO284	Q285	0.0875
T275	N278	0.0875
P4	R48	0.0869
N278	K279	0.0866
K9	R10	0.0862
Q150	C164	0.0859
R268	K271	0.0853
K68	N102	0.0851
V5	K14	0.0850
N78	R114	0.0849
I274	T275	0.0848
N54	L55	0.0847
R114	A155	0.0842
L222	L241	0.0842
D229	N231	0.0832
S188	A211	0.0832
N118	S194	0.0830
D65	K68	0.0830

The numbers in first two columns indicate the residue position in Sec17 and the letters indicate the amino acid at that position. The residue pairs lying in close contact to each other in the tertiary structure are colored orange and the residue pairs lying away from each other are colored in pink.

A closer inspection of the 50 highest scoring covarying residue pairs of the α-SNAP analysis revealed that some of the residue pairs are lying next to each other (Table 4.5). The neighboring residues 3 and 4 were found as the highest scoring pair. Similarly, the high-scoring pairs 200 & 204, 24 & 25, 30 & 31, 275 & 279, 120 & 140 are lying close to each other in the tertiary structure (Figure 4.15). Some high-scoring pairs were found to be further away from each other though, for example, the residue pairs 53 & 194, 119 & 162, 2 & 14, 198 & 266, 85 & 114, and 4 & 48 were lying distant from each other (Figure 4.16). Some residues appeared to be co-varying with many other residues thus forming groups of covarying residues, for example, residues 1, 3, 4, 6, 7, 8, 9 and 10.

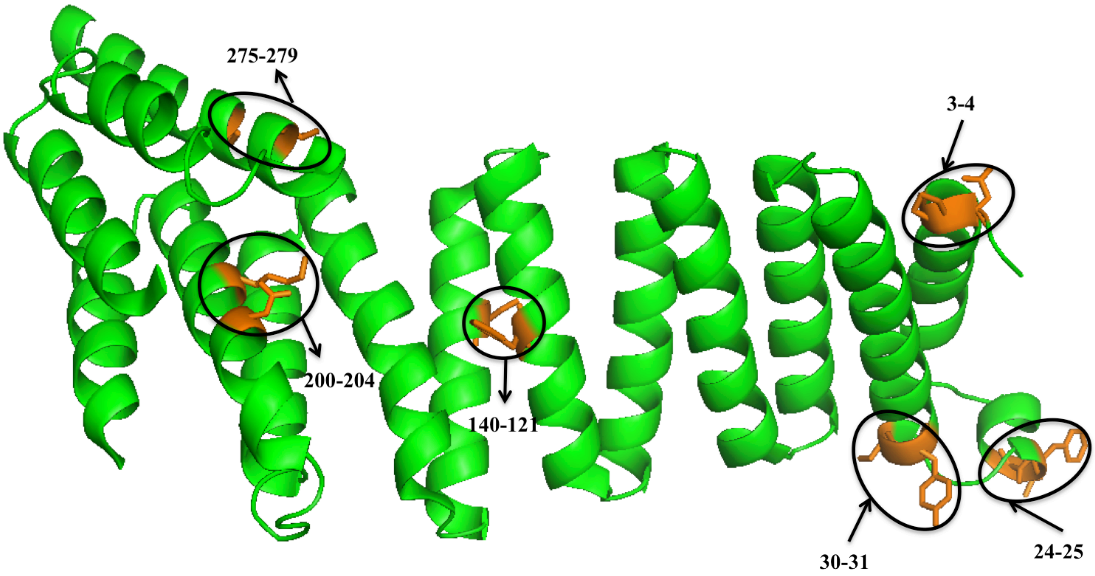


Figure 4.15: Some of the residue pairs from the top 50 covarying residue pairs of the α-SNAP analysis.
The residues are marked on the tertiary structure of Saccharomyces cerevisiae α-SNAP (Sec17, PDB code 1QQE). The Sec17 structure is shwon in gren and the residues that are lying close to each other in the structure are represented in orange.

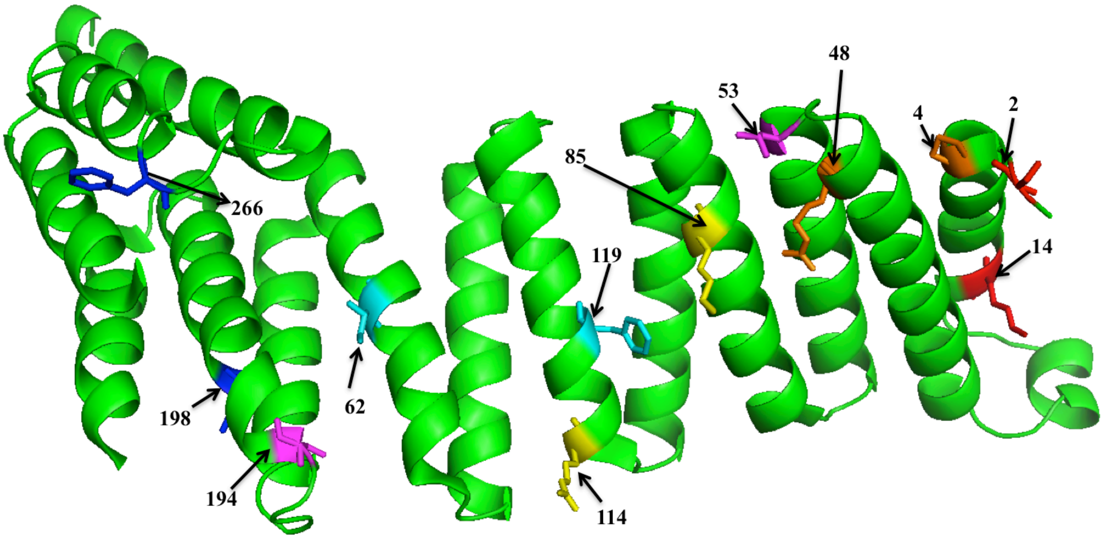


Figure 4.16: Some of the residue pairs from the top 50 covarying residue pairs of the α-SNAP analysis.
The residues are marked on the tertiary structure of Saccharomyces cerevisiae α-SNAP (Sec17, PDB code 1QQE). The residue pairs that are lying distant from each other in the structure are represented in same color.

Clusters of covarying residues in α-SNAP protein

In order to identify the groups of covarying residues, average linkage clustering of the MIP data was carried out. Average linkage clustering with stopping criteria of 2-σ of the MIp score, resulted in 49 clusters, of which, six clusters were above the cluster selection cut-off of 80%.

The selected clusters are shown in a Heat-map that is organized according to the rank of the clusters (Figure 4.17). The heat map showed that the selected clusters had most of the high-scoring residues embedded within them (Figure 4.17). Note that few high-scoring residue pairs were not included in the top 80% clusters.

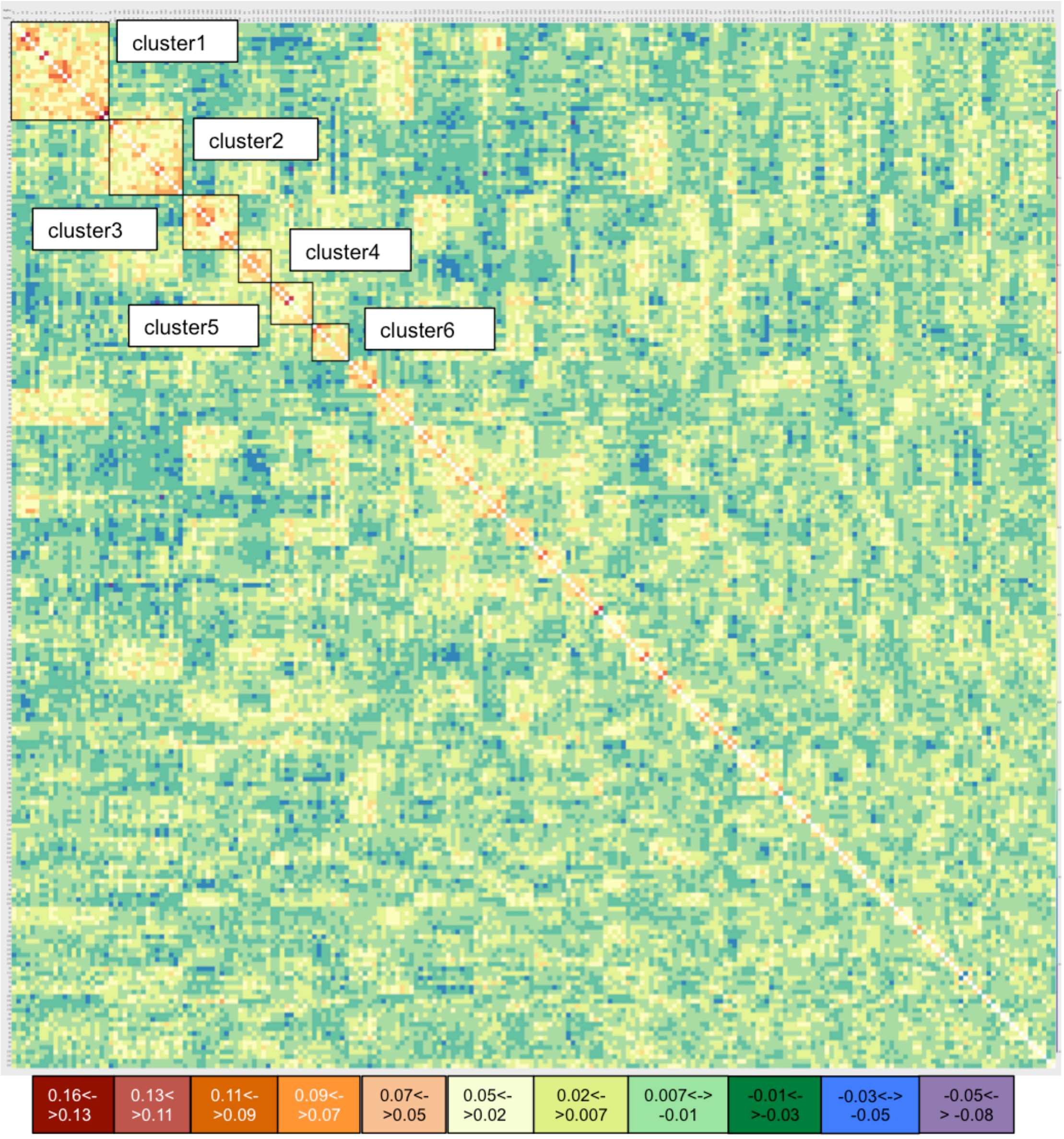


Figure 4.17: Heatmap of MIp scores from α-SNAP alignment arranged according to the ranking of clusters.

Selected clusters are marked in Black squares.

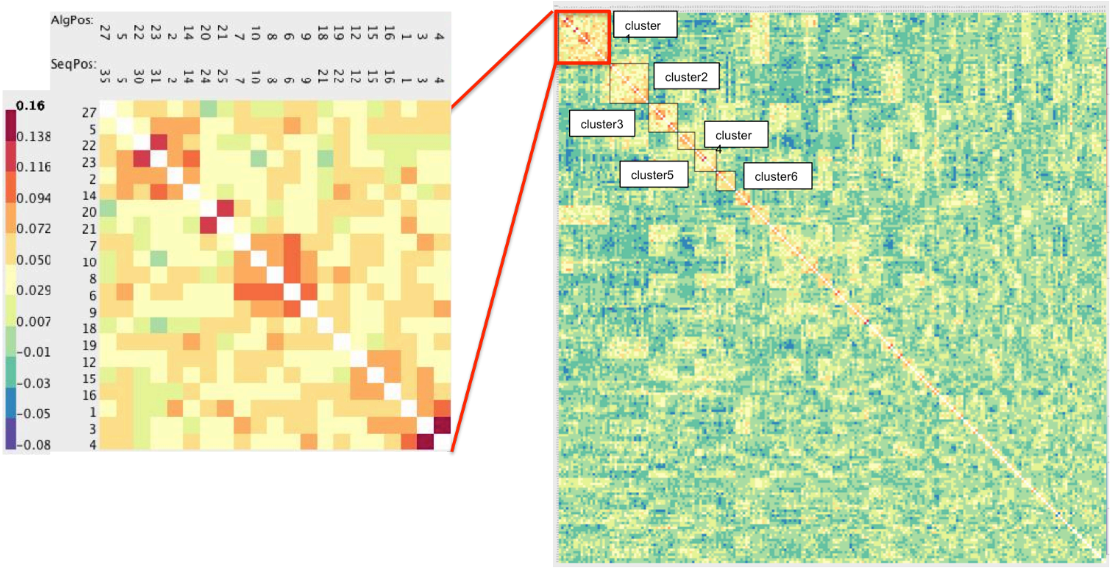


Figure 4.18: A magnified portion of heat map MIp scores from α-SNAP alignment arranged according to the ranking of clusters.
The magnified view is showing the first cluster identified after applying the average linkage clustering, scoring and ranking of the clusters.

To evaluate and illustrate the selected (top 80%) covarying residue networks, the residues from the top 80% clusters were mapped onto the structure of *Saccharomyces cerevisiae* α-SNAP, Sec17 (Figure 4.19). This revealed that each cluster is restricted to a certain region of the protein.

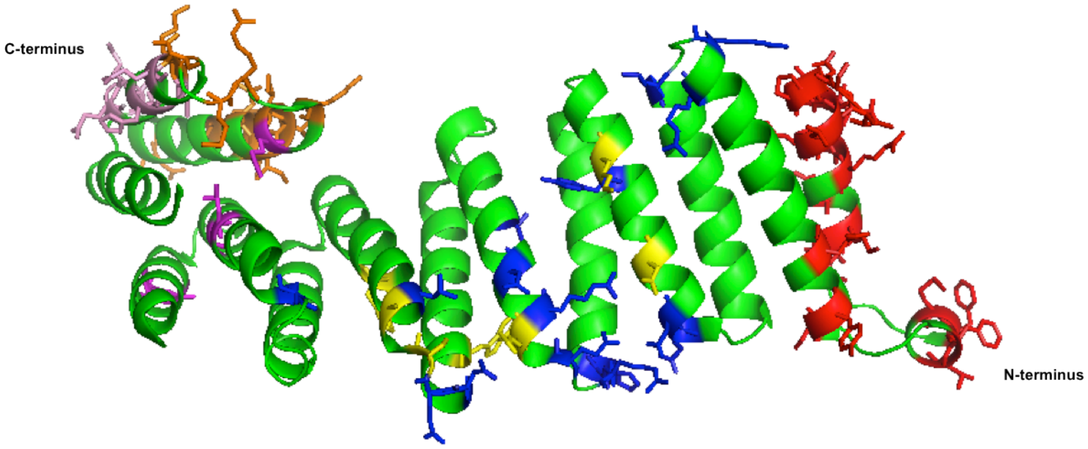


Figure 4.19: Residue from the selected clusters marked on the structure of Saccharomyces cerevisiae α-SNAP (Sec17, PDB code 1QQE).
The structure 1QQE is in green. Red residues represent the first cluster, blue residues represent the second cluster, orange residues represent the third cluster, yellow residues represent the fourth cluster, magenta residues represent the fifth cluster, light pink represents residues form sixth cluster.

All the residues of cluster 1 are localized to the N-terminal region that encompasses the first helix and a loop that is thought to be involved in membrane attachment. All the residues of cluster 2 and cluster 4 from the analysis of α-SNAP are in the middle TPR region. The residues in this cluster lie in TPR 1, 2, 3 and 4. The residues from cluster 3, 5 and 6 all are located at the C-terminal region and lie close to each other. The highest scoring pair of the entire MIp dataset residue 3 and 4 is a part of cluster 1. Some other residues with high MIp scores within this cluster are also lying close to each other. However, some other residue pairs, also with high scores, such as 8 & 22, 30 & 2, 31 & 2, 15 & 3, are lying further away in the tertiary structure. The cluster analysis thus helped to identify networks of residues.

Figure 4.20 shows a network view of this cluster. Many of the residues from the top 50 scoring pairs that were found to be lying as isolated pairs (Figure 4.15) and the residues that were found to be lying distant from each other (Figure 4.16) were clustered together into one cluster. For example, residues 3 & 4, 24 & 25 and 30 & 31, appeared as pairs (Figure 4.15) and residues 2, 5 and 14 were found to be lying distant from each other (Figure 4.16). All of these residues were clustered together into cluster 1 by the cluster analysis. Residues 3 and 4 have the strongest connection as they have the highest covariation score, however, they appeared to be less connected or connected by very low scores to most other residues. Residue 14 appeared to be a stronger node, with high number of strong connections. Distribution of amino acids from this cluster showed that most of the identified residues in this region are hydrophobic and polar or negatively charged (Figure 4.21). The hydrophobic and polar amino acids usually have an important role in protein folding and structural stability and also mediate binding to the substrates or other proteins. This suggests that the residues from this cluster might have some structural and functional importance.

Interestingly, when the stopping criterion for clustering was increased to 3σ the cluster 1 fell apart into three distinct clusters of 7, 5 and 3 residues (Figure 4.22). Thus, within this cluster there were small groups of highly covarying residues that had residues that appeared to be lying close or adjacent to each other. These groups were connected together by some low scoring connections thereby connecting the distantly lying covarying residues. This showed that selecting the cluster stopping criteria is very important for identifying the clusters of significant size. Higher cluster stopping criteria (3σ, 4σ) results in smaller clusters.

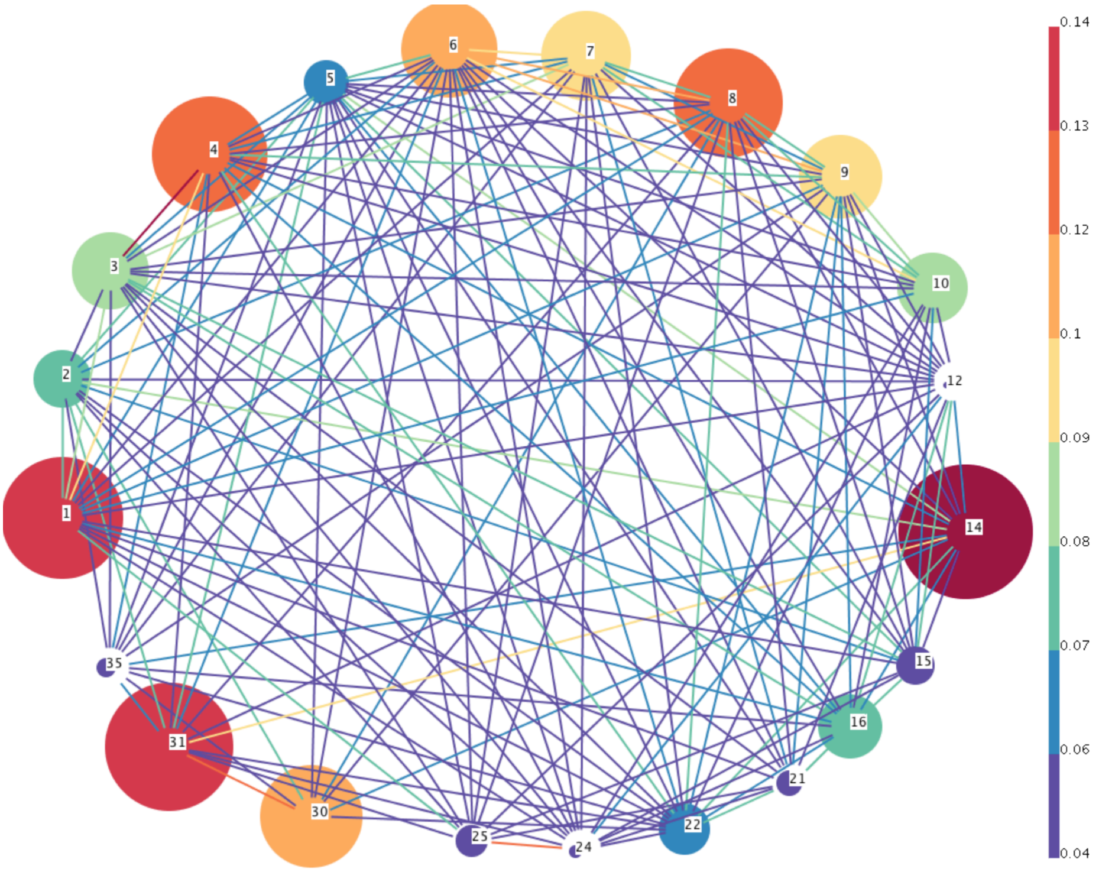


Figure 4.20: Network of residues of cluster 1 from α-SNAP analysis.
The color of the edge represents the strength of the covariation score. Nodes are represented by circle with residues number from the representative structure of Sec17. The color of the edge represents the strength of the covariation score, shown in the bar on the right, with red being highest and blue being the lowest. The color and size of the node represents the importance of the residue in terms of the average of the score of the edges connected to it.

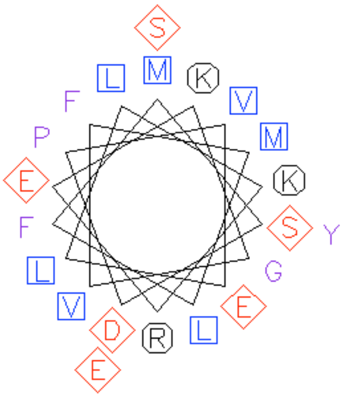


Figure 4.21: Helical wheel view of residues of cluster 1 from α-SNAP analysis.
The helical wheel represents the helix by a projection of the Cα backbone structure down the helix axis. Aliphatic/hydrophobic residues are shown as blue squares, polar or negatively charged or hydrophilic residues as red diamonds, positively charged residues as black octagons and special cases as purple letters.

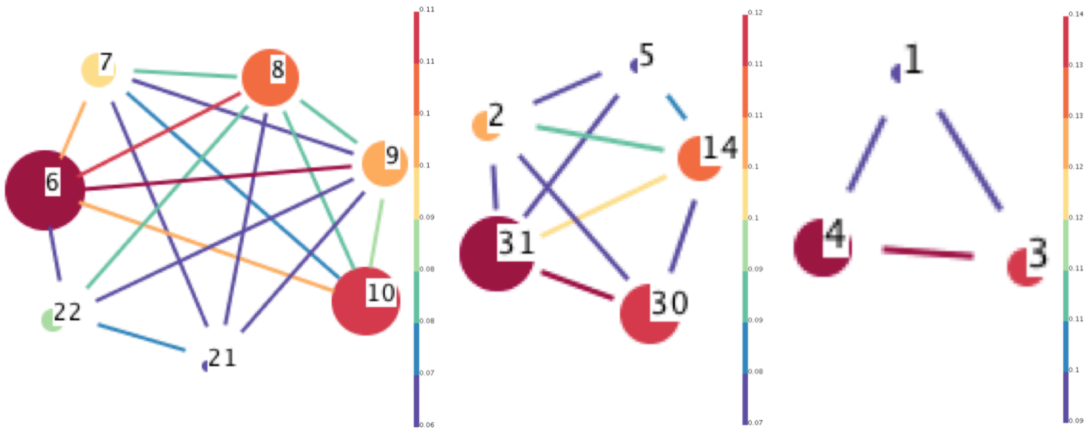


Figure 4.22: Network view of three different clusters from α-SNAP analysis.
Average linkage clustering with stopping criteria of 3σ on the MIp scores from α-SNAP, breaks the cluster1 into three different clusters. The three clusters have residues same as in the cluster 1 obtained at stopping criteria of 2σ. Color-codes are same as in Figure 4.20. The residues within the subclusters appear to be lying adjacent or close to each other.

Similar observations were noted for other clusters as well (Appendix Figure A.33-A.38 (can be provided on request)). The clustering thus helped to identify the networks of covarying residues occurring within the protein.

Clade-specific analysis of covarying residues of α-SNAP

The covariation of the highest scoring pair residue 3 and 4 occurring along the different lineages was mapped on the phylogenetic tree of the α-SNAP protein. These residues showed a change in pattern of substitution according to the major branches of the tree, i.e., Fungi, Viridiplantae, Metazoa and others (Figure 4.23).

A previous phylogenetic analysis of SNAP proteins carried out in my group showed that SNAP proteins underwent duplication to α- and β-SNAP in vertebrates (Kienle 2010). The distribution of the highest scoring pair revealed that in α- and β-SNAP of vertebrates, the residues 3 & 4 had EA, while in rest of the metazoans or invertebrates they had RA, RG, KA and KG. A strong change of one residue from being positively charged (K/R) to negatively charged (E) might relate to specialization specific to vertebrates. Examining this distribution further revealed that the branches closer to metazoans like Capsaspora also had KA at these two residue positions, which was a pattern similar to the metazoan invertebrates. Thus, Capsaspora is very similar and closer in evolution to the metazoans.

In other branches, the highest scoring pair residue 3 and 4, varied from Q-A in Basal Fungi, P-A and Q-G in Basidiomycota, D-P in Ascomycota and R-G, K-G, R-A, K-A in Viridiplantae. In most of the other eukaryotic lineages, this pair varied from K-A, R-A, E-A, R-G, Q-A and D-A, E-A, D-I, E-I, corresponding to further branching (Table 4.6). Some other eukaryotic lineages, like Kinetoplastida, Entamoebidae, Mycetozoa, Stramenophiles, Amoebozoa, Rhodophyta, Haptophyceae, Apusozoa, Rhizaria, Jakobida showed pattern similar to metazoans, i.e., K-A, R-A or K-G, R-G. Alveolata showed a different pattern as D-A, E-A, D-I, E-I, T-P, N-P or N-A. Similarly some other lineages like Diplomonadida, Euglenozoa, Malawimonadidae, Cryptophyta, Heterolobosea showed Q-A, S-A pattern at these residue positions.

Similar substitution pattern was observed for high scoring residues in the other clusters. Thus, the identified residues showed clade-specific covariation.

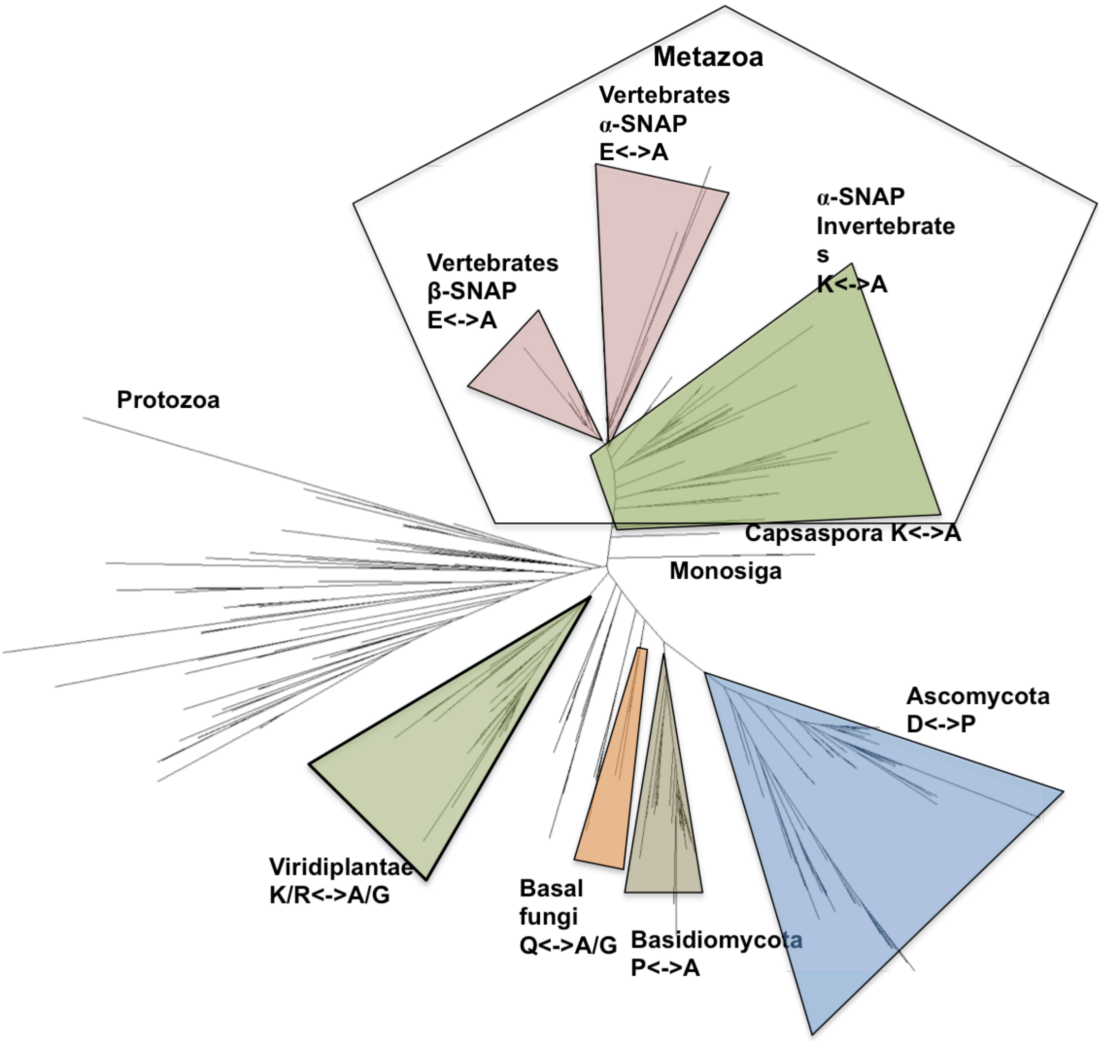


Figure 4.23: Highest scoring pair 3<--> 4 from cluster 1 mapped on phylogenetic tree of α-SNAP. The colors represent the changes in distribution of the covarying amino acid pairs. The branches that have same amino acid distribution are colored in same color and the ones where the amino acid distribution changes are colored differently. The detail of distribution of covarying amino acids in other branches for the highest scoring pair 3<->4 is shown separately in the Table 4.6.

Table 4.6: Distribution of highest scoring pair 3<--> 4 from cluster 1 of α-SNAP.

Branches	Residue 3	Residue 4
Kinetoplastida	K/R	G/A
Kinetoplastida duplication	S	A
Entamoebidae	K/R	A
Mycetozoa	K/R	A
Stramenophiles	K/R	G/A
Amoebozoa	K/R	I/G
Rhodophyta	R	I
Haptophyceae	K	A
Apusozoa	K	A
Rhizaria	K	A
Jakobida	K	A
Choanoflagellata	K	G/A
Diplomonadida	Q	A
Euglenozoa	Q	A
Malawimonadidae	S	A
Cryptophyta	Q	A
Heterolobosea	E/Q	V/A
Apicomplexa	D/E	A/I
Ciliophora	K	G
Chromera velia/Perkinsea	N	A

The rows are colored as per the changes in the other branches. The branches that have same amino acid distribution are colored in same color and the ones where the amino acid distribution changes are colored differently.

4.4 Application of the improved analysis pipeline for intra-protein analysis of SM proteins

As outlined earlier (in section 4.1), on performing the initial co-variation analysis on the SM protein, Sly1, several high-scoring residue pairs were found to be lying distant from each other and some covarying residues also appeared to form networks within the tertiary structure. To be able to further analyze the MIp data on Sly1, novel tools were developed as outlined in the previous sections.

Correlated Mutation analysis of Sly1

The heatmap of MIp scores of Sly1 (Figure 4.27) revealed several regions with several high-scoring residue pairs. Although several covarying residues were concentrated in the N-terminal region of subdomain d1, and in some parts in the subdomains d2 and d3, many covarying residues were found to be spread all over the sequence. This is consistent with initial observation that high-scoring co-varying residues were found to be located all over the structure of the protein (Figure 4.1).

Comparison with analysis on enzymes (MAP & UDG) and SNAP protein

When one compares the heatmaps of the MIp scores of Sly1 with that of the other proteins analyzed in this study (Map1- Figure 4.5, UDG-Figure in Appendix-A.10, α-SNAP – Appendix-A.23 (Appendix figures can be provided on request)) a different pattern can be noticed immediately. While many of the high-scoring pairs are well distributed in the heatmap of the Sly1 protein family, the other protein families often contain distinct hotspots of co-varying positions. As outlined before, their covarying residues are restricted to certain regions of the protein (MAP1- Figure 4.7 and UDG-Figure 4.10) (α-SNAP- Figure 4.19), while many of the covarying residues were found to be lying in different regions of Sly1.

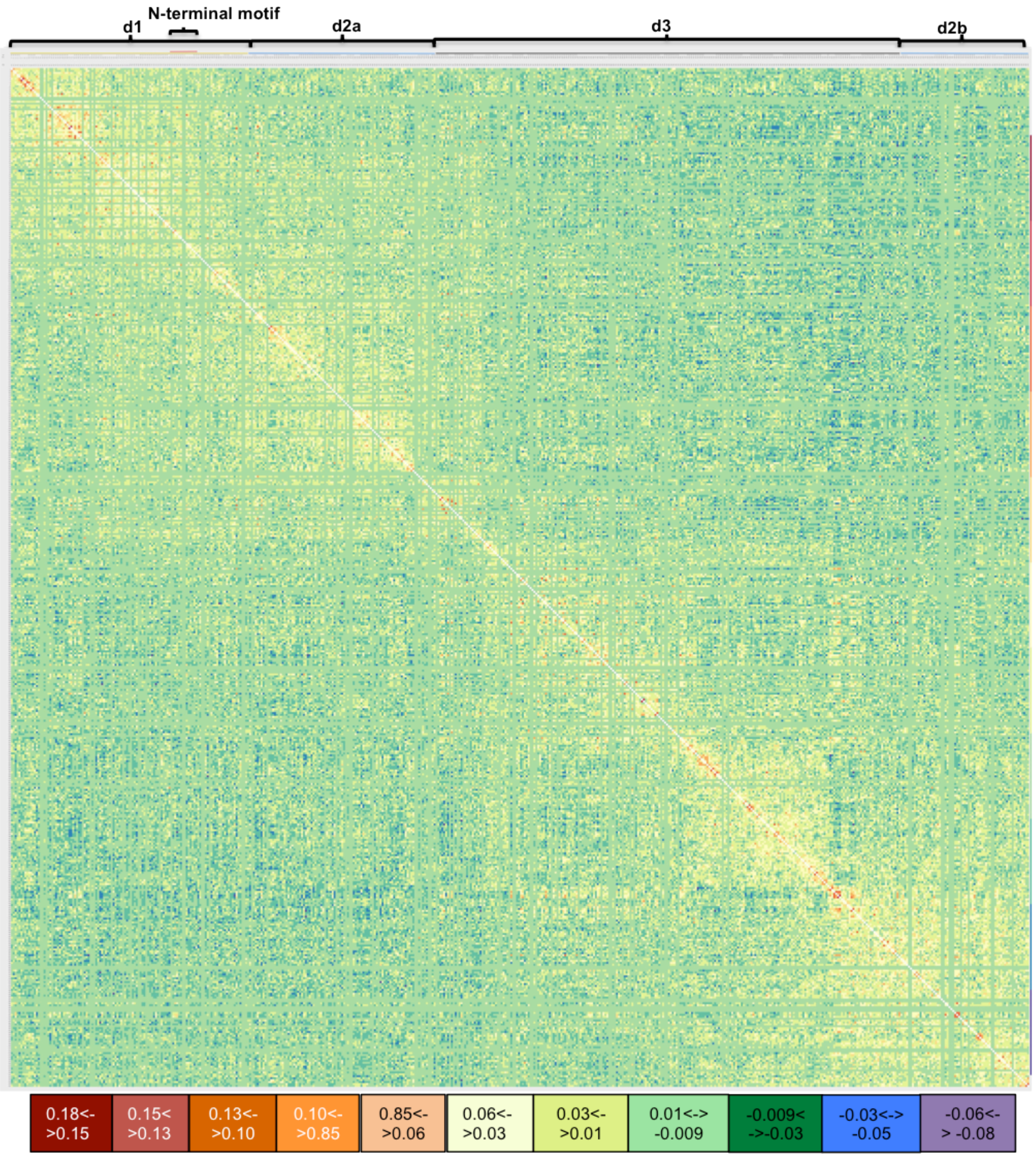


Figure 4.24: Heat map of MIp scores of Sly1 data.
The heatmap is arranged as per the residue numbers from the structure *Saccharomyces cerevisiae* Sly1 (1MQS). The color code bar represents the color corresponding to the range of MIp score. The dark red, red, dark orange, orange, represent the high MIp scores and light green, dark green, blue, purple represent the low MIp scores. The structural element are marked on the top of the heatmap; d1: domain d1, d2a: domain d2a, d3: domain d3, d2b: domain d2b. N-terminal motif: motif that binds to the N-peptide of the Syntaxin. The original heat map can be found in the supplements (provided on request) that can be enlarged to have a clear view of the residue numbers.

Clusters of covarying residues in Sly1 protein

Average linkage clustering was then applied on the MIp dataset. A stopping criterion of 2-σ and a cluster selection cut-off of 90% resulted in top four clusters being selected. As described in the Method section 3.6, the cut offs were selected based on the simulation of the Sly1 dataset. In Fig. 4.1.3-2, the selected clusters are shown in a heat-map that is organized according to the rank of the clusters. In the heat-map it can be seen that most of the high scoring residues (shown by red) were embedded within the selected top 90% clusters.

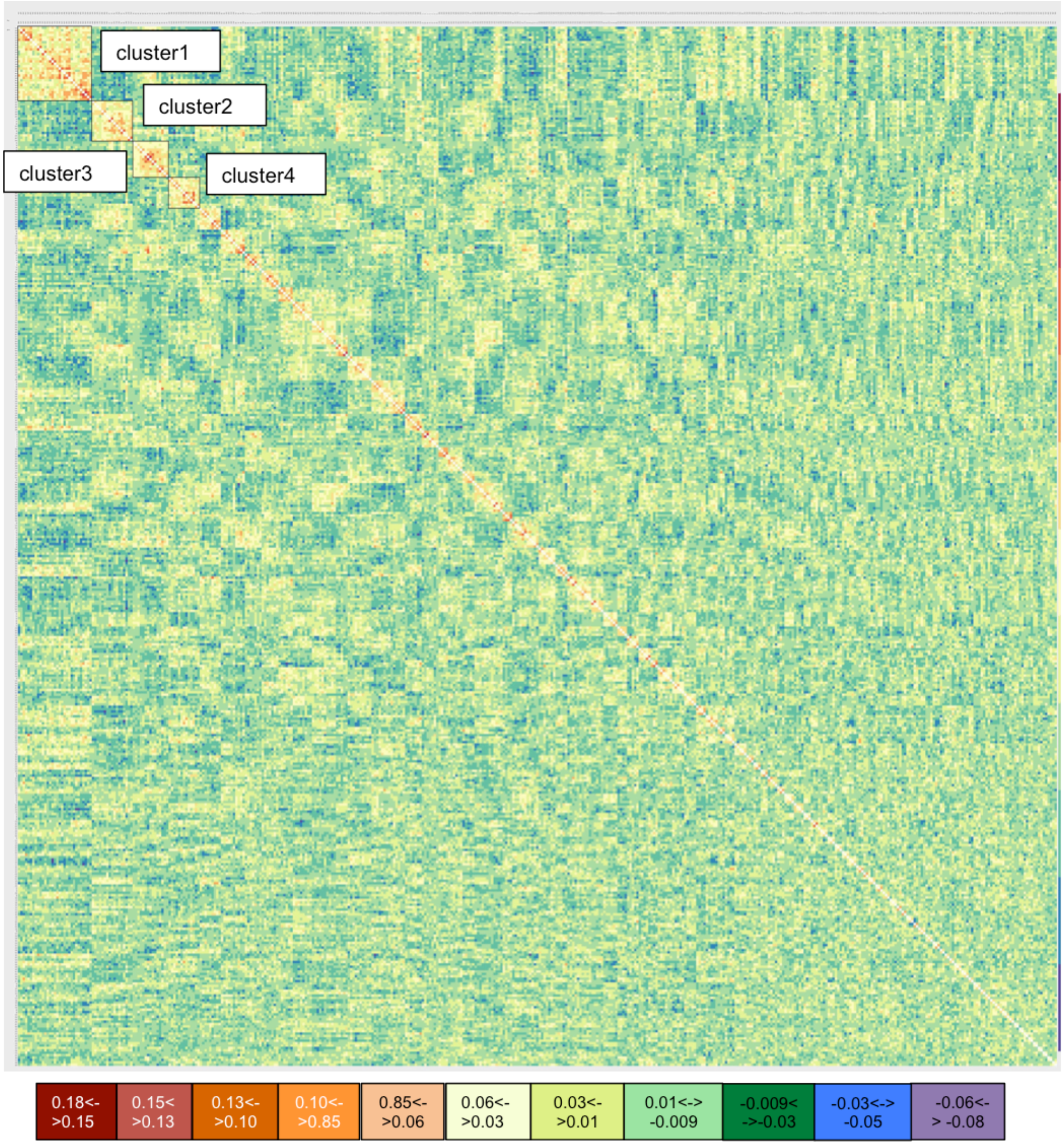


Figure 4.25: Heatmap of MIp scores from the Sly1 dataset, arranged according to the ranking of clusters.
Selected clusters are marked in Black squares.

Each cluster was then analyzed to check the distribution of residues within them.

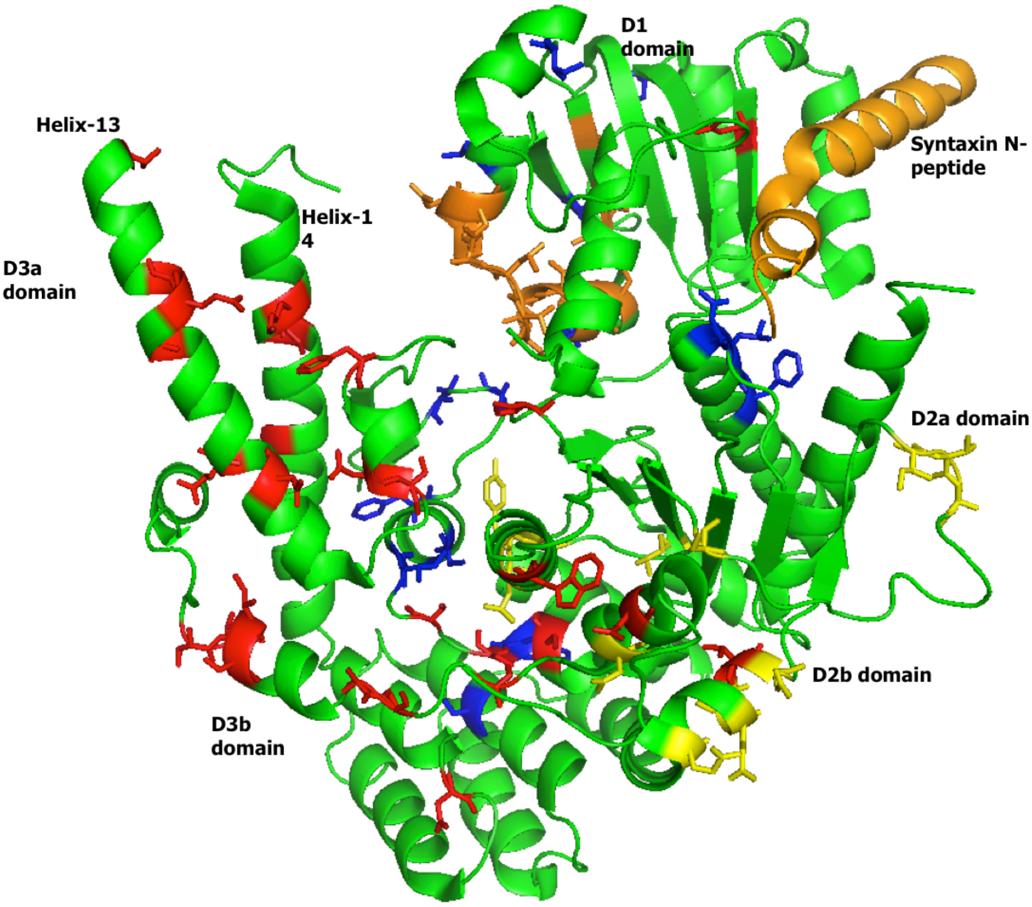


Figure 4.26: : Residue from the top 90% clusters marked on the structure of *Saccharomyces cerevisiae* Sly1 (1MQS).
Sly1 structure is shown in in green and Syntaxin5-N-peptide structure is shown in grey. Red residues represent the first cluster, blue residues represent the second cluster, orange residues represent the third cluster, yellow residues represent the fourth cluster.

The residues from cluster 1 (shown in red in Figure 4.26) were located in the domain d3, including helix-13 and helix-14, in the domain d2 and in the Sly1 specific loop region. The highest scoring pair, residues 543 & 544 were lying adjacent to each other (Figure 4.27), in the 3D structure. Some other pairs, such as 326 & 354 were 12Å and 354 & 579 were 34Å apart.

In order to have a clear detail view of the cluster, I removed the low-scoring edges from this cluster by putting a cut off of 0.09 and visualizing only the connections having MIp scores above this cut off (Figure 4.28). This resulted in two major sub clusters and many individual pairs. One subcluster contained all the connections with near-by lying residues, for example, 541, 542, 543, 544, 550 & 551 (Figure 4.29). The other subcluster had some distantly lying residues pairs, for example, 354 & 579 and 354 & 275 (Figure 4.30). This suggested that the distantly lying residues of cluster 1 were connected to each other by other low scoring pairs and thus appeared to form a chain in the 3D structure.

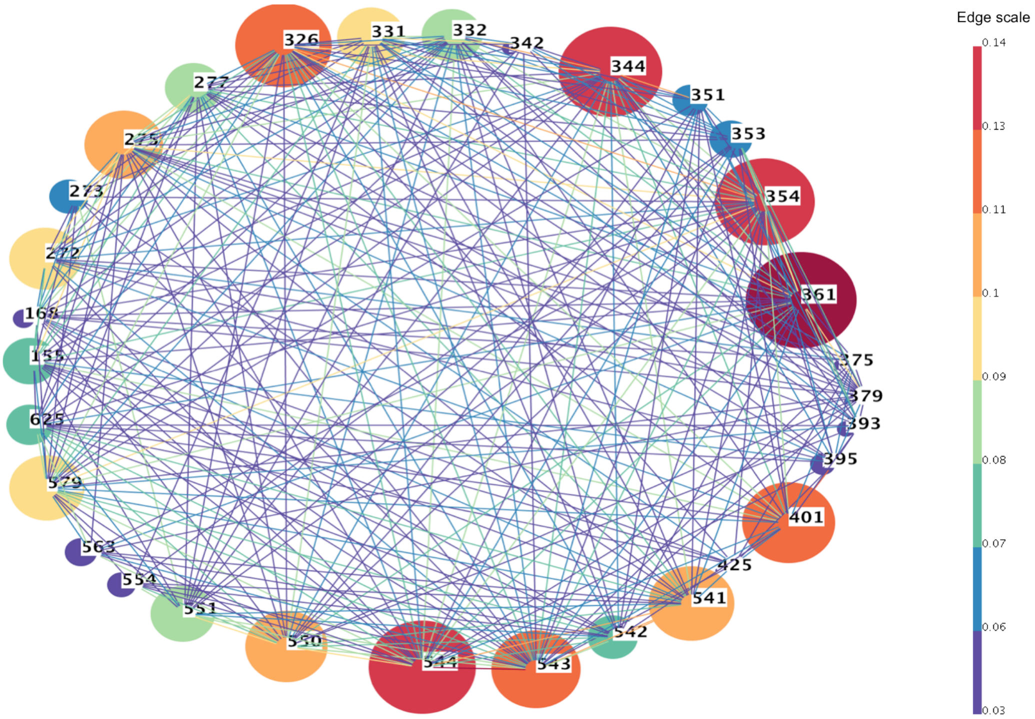


Figure 4.27: Network view of cluster1 of the Sly1 dataset.
The color of the edge represents the strength of the covariation score. Nodes are represented by circle with residues number from the representative structure 1MQS. The color of the edge represents the strength of the covariation score, shown in the bar on the right, with red being highest and blue being the lowest. The color and size of the node represents the importance of the residue in terms of the average of the score of the edges connected to it. Enlarged view can be found in the supplementary data (provided on request).

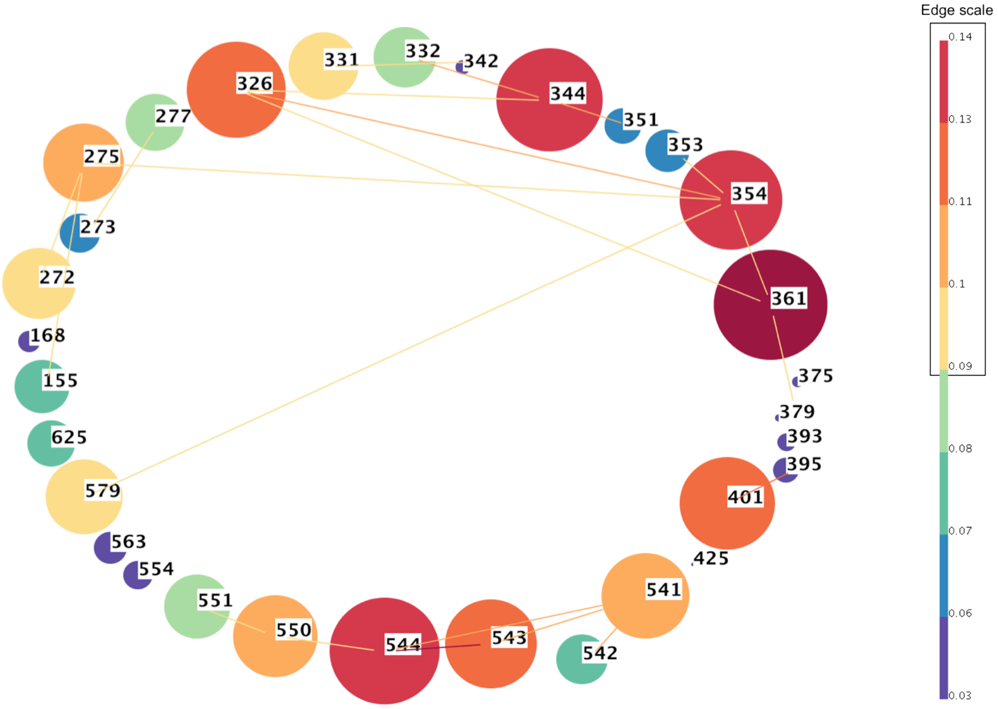


Figure 4.28: Network view of cluster1 of the Sly1 dataset after removing low scoring edges.
The color codes are same as in Figure 4.31. The black box around the edge scale indicates cutoff used to visualize the high scoring connections.

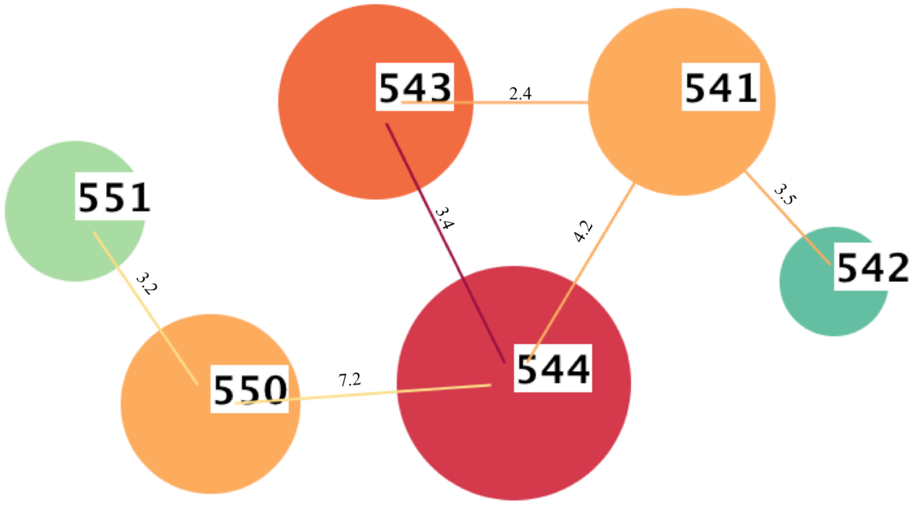


Figure 4.29: Network view of first group of high scoring residues from cluster1 of Sly1 dataset. Only residues connected by edge weight >0.090 are shown. The group shows residues lying close to each other. The numbers above the edges indicate the side chain distances between the residues. The color code is same as in Figure 4.27.

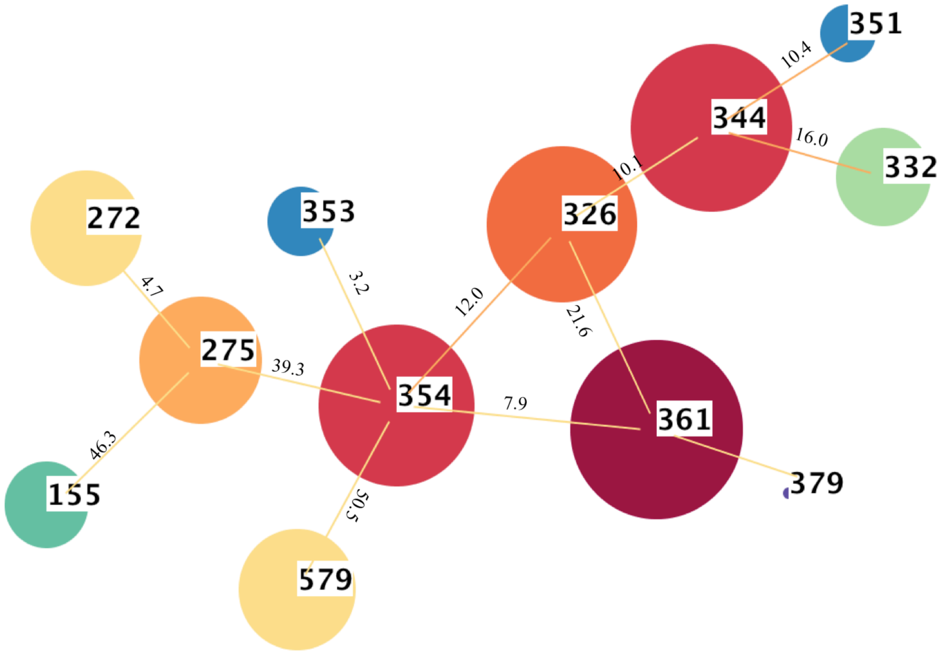


Figure 4.30: Network view of second group of high scoring residues from cluster1 of Sly1 dataset. Only residues connected by edge weight >0.090 are shown. The group shows residues lying close as well distant from each other. The numbers above the edges indicate the side chain distances between the residues. The residue 379 is missing is the structure and so distance value is not indicated for it. The color code is same as in Figure 4.27.

When I mapped the residues from the two subcluster onto the structure of Sly1 (Figure 4.34), the residues from the first subcluster were found to be located on the Sly1 specific loop region and the residues form the second cluster were located on the two hairpin helices of domain d3a and in domain 2. A recent crystal structure of Vps33 revealed it to be interacting with the SNARE motif of Qa-SNARE as well as the R-SNARE (Baker et. al. 2015). The two hairpin helices from the domain d3 region were found to involved in this interaction. This study suggested that SM proteins serve as the assembly platform for the SNARE complex formation. The residues from the identified second subcluster were found to be lying in the analogous R- and Qa-SNARE interacting region in Sly1. Different configurations of the hairpin helix region as observed from the different crystal structure of Munc18a (Burkhardt et al. 2008), Munc18b (Hackmann et. al. 2013) and Munc18c (Hu et. al. 2007) have also suggested that there is a conformational change in SM protein that could be involved in the opening of syntaxin (Baker et. al. 2015). The Sly1 specific loop consisting of α20 and α21 helices is known to be a Rab GTPase interaction site for Sly1. Point mutation (E532K) in α20 and deletion mutation of most of the loop has shown to suppress the requirement of Rab Ypt1 and Ypt6 (Dascher et al. 1991, Bracher & Weissenhorn 2002). This loop is also thought to be acting as a lid regulating the exposure of the R-SNARE binding site (Baker et. al. 2015). This suggested that the two important regions of Sly1, involved in (either simultaneous or subsequent) interactions with different proteins had high scoring covarying residues. However, low scoring and distantly lying covarying residues connected the two regions. The covarying residues thus appeared to form a chain within the 3D structure of Sly1 connecting different important regions of the protein.

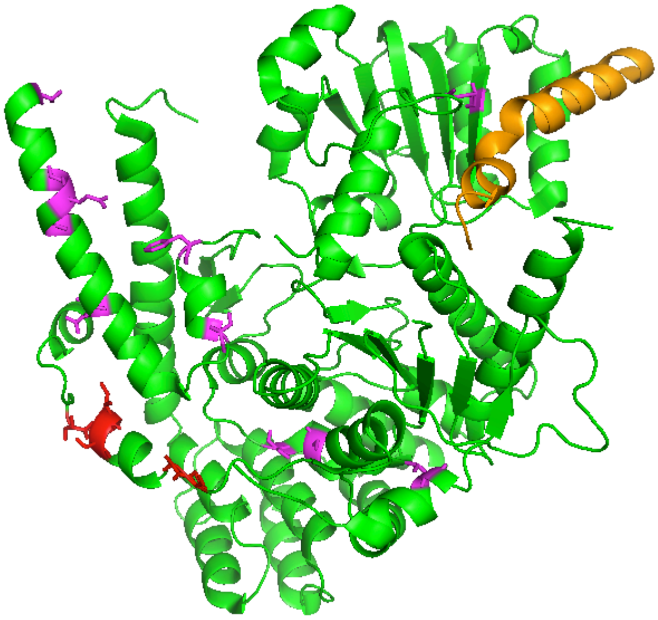


Figure 4.31: Residue from the two subclusters marked on the structure of *Saccharomyces cerevisiae* Sly1 (1MQS).
Sly1 structure is shown in in green and Syntaxin5-N-peptide structure is shown in grey. Red residues represent the first subcluster, pink residues represent the second subcluster.

Similar observations where many high-scoring residue pairs that appeared to be forming chains in the 3D structure and were spread within the protein were made for the residues from clusters 2 and cluster 4. The residues from cluster 3 were found to be restricted only in domain d1 and consisted of closely lying residues. Clustering approach thus helped to identify the groups of covarying residues occurring within the protein.

4.5 Application of the improved analysis pipeline for inter-protein analysis of Sly1 and Syntaxin 5

To identify the networks of functionally and structurally important residues involved in maintaining the interaction between the SM protein Sly1 and its partner Syntaxin 5, MIp followed by clustering was applied on the alignments of Sly1 and Syntaxin 5. Sly1 and Syntaxin 5 were chosen out of all other SM-Syntaxin pairs as they occur as singletons in most species. First, individual alignments of Sly1 and Syntaxin5 proteins were generated, refined, and processed as outlined before. The two alignments were then concatenated in order to have the same number and order of species and sequences.

Before performing the inter-protein analysis on the concatenated alignment of Sly1 and Syntaxin 5, intra-protein analysis was also performed on the Syntaxin 5 dataset,

Intra-protein analysis of Syntaxin 5

Correlated mutation analysis by MIp on Syntaxin 5 dataset alone showed high scoring residues to be lying at the N-terminal region, Habc domain, in the linker between Habc and SNARE domain, SNARE domain and transmembrane region (Figure 4.32). Again, the high scoring pairs were found to be spread all over the protein. The highest scoring pair was V339-N340 that is part of the transmembrane (TM) region of Syntaxin 5.

The average linkage clustering with stopping criteria of 2-σ of the MIp scores and cluster selection cut off of 90% resulted in five clusters with different numbers of residues for Syntaxin 5 (Figure 4.33). Clusters with residues lying close to each other as well as away from each other were found (Figure 4.34). The highest scoring pair residues 339 & 340 were lying adjacent to each other and were lying in the TM region of the protein. The cluster 1 had residues in Ha, Hb and Hc helices. Cluster 2 had residues in hc helix, SNARE motif, TM region and in the linker between Habc domain and the SNARE motif. Cluster 3 had residues in Hc helix, SNARE motif, in the linker between Habc domain and the SNARE motif and one residue in the N-terminal region. Cluster 4 had residues in SNARE motif and TM region. Cluster 5 had residues in Hb helix, SNARE motif and TM region. Thus the clusters had residues lying in different regions of proteins that appeared to form netwroks connecting these regions.

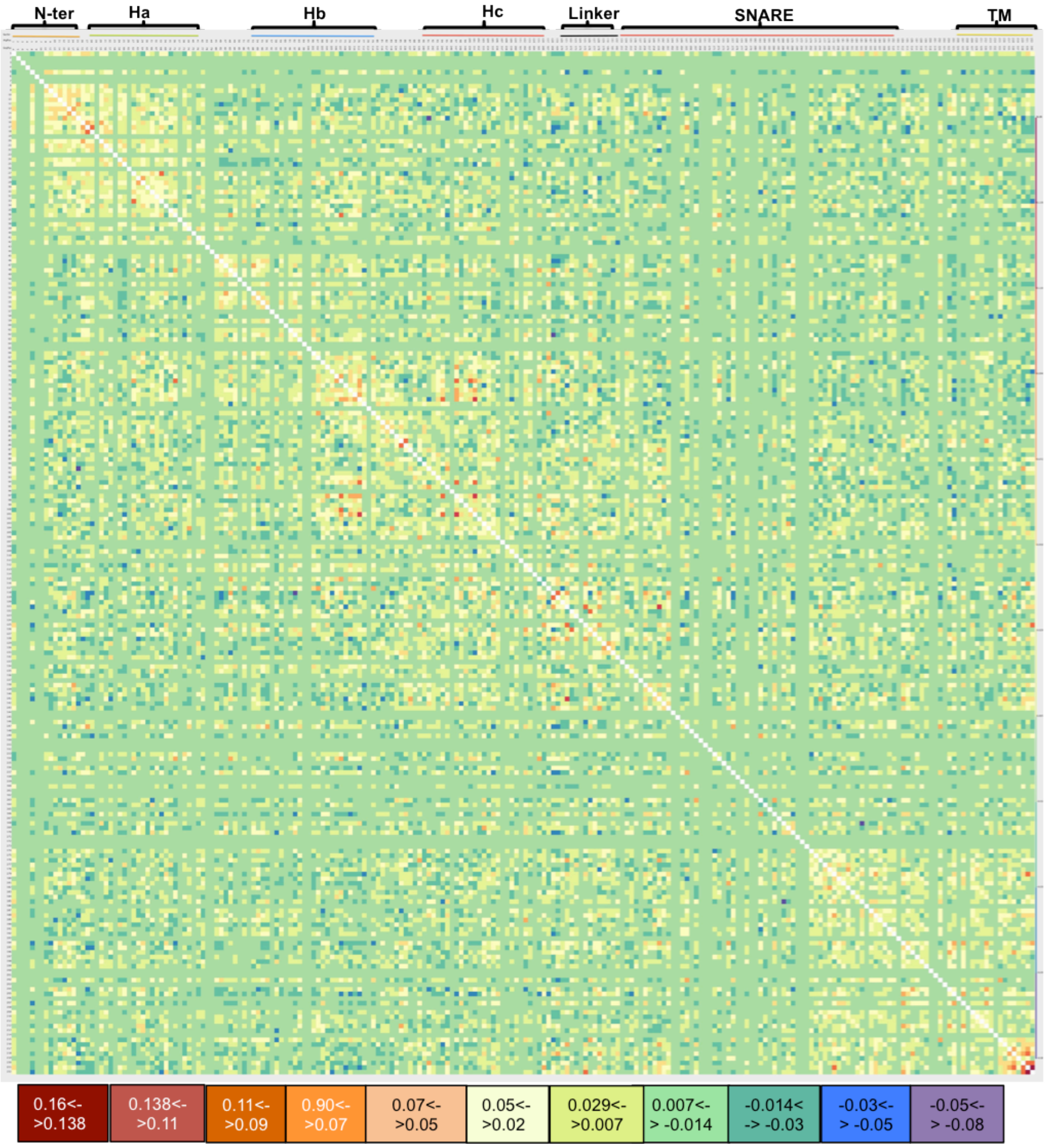


Figure 4.32: Heat map of the MIp scores of the Syntaxin 5 data.
The heatmap is arranged as per the residue numbers from the structure *Saccharomyces cerevisiae* Sed5. The color code bar represents the color corresponding to the range of MIp score. The dark red, red, dark orange, orange, represent the high MIp scores and light green, dark green, blue, purple represent the low MIp scores. The structural element are marked on the top of the heatmap; N-ter: N-peptide of the Syntaxin 5, Ha: helix Ha, Hb: helix Hb Hc: helix Hc, Linker: linker between Habc and NSARE

domain, SNARE: SNARE domain, Tm: transmembrane domain. The original heat map can be found in the supplements (provided on request) that can be enlarged to have a clear view of the residue numbers.

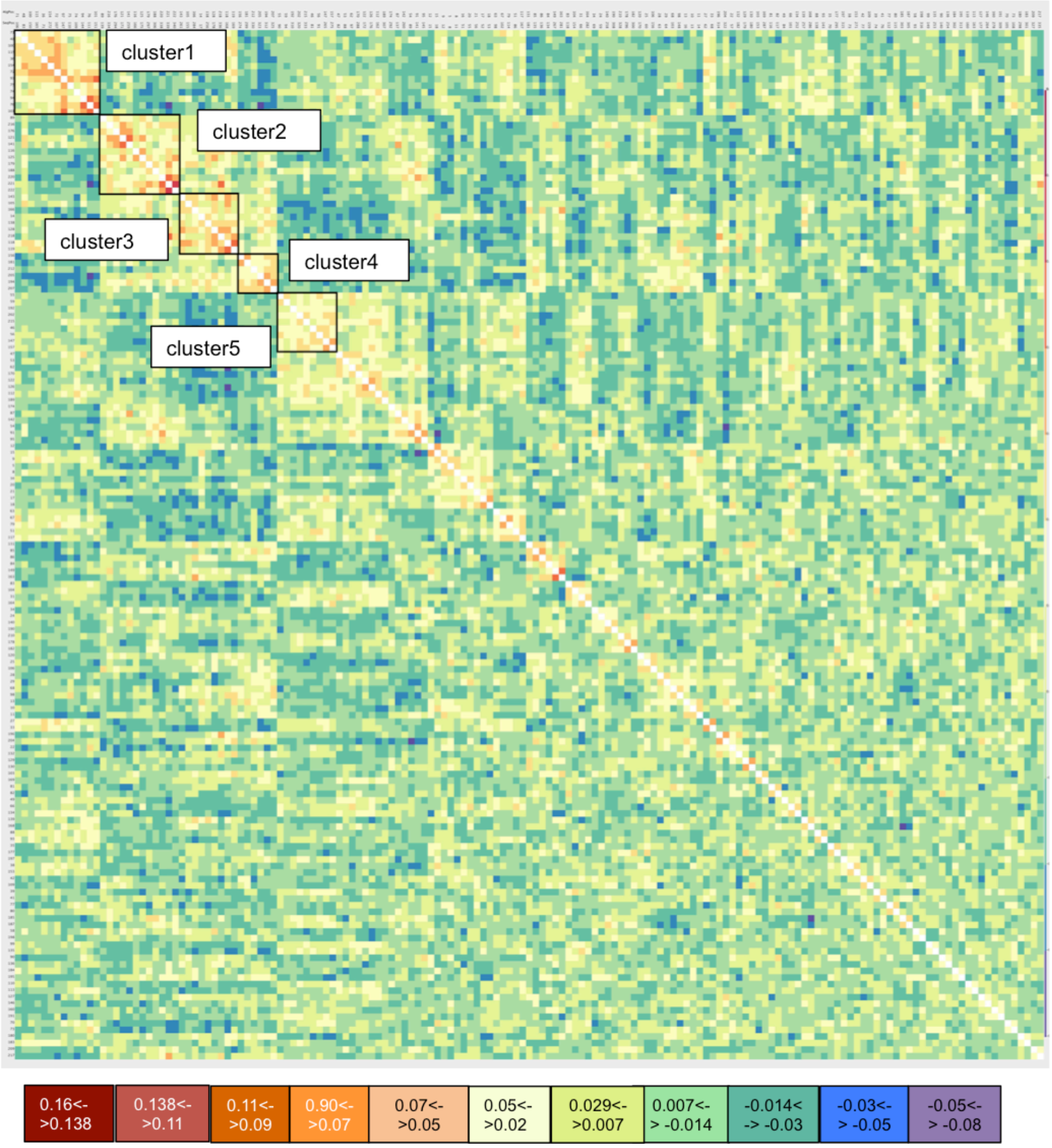


Figure 4.33: Heatmap of the MIp scores of the Syntaxin 5 dataset, arranged according to the ranking of clusters.
Selected clusters are marked in Black squares

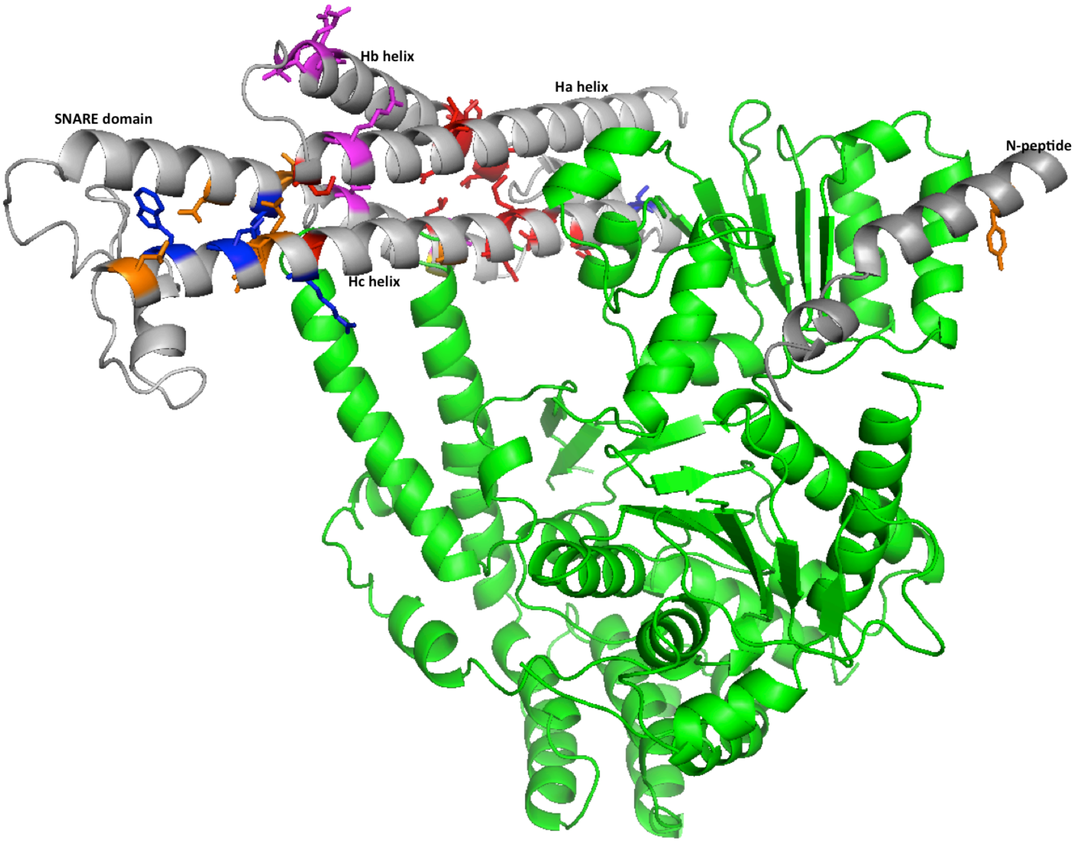


Figure 4.34: Residue from the top 90% clusters marked on the homology model of the Sly1/Syx5 structure.
The structure is obtained by overlapping 1MQS of Sly1, homology modeled structure of Syntaxin 5 on SSO1 and 3C98 of Munc18-1 in complex with Syntaxin1. Sly1 structure is shown in green and Syntaxin5 structure is shown in grey. Color code same as in Figure 4.49. As certain ergions like TMR, are missing in the homoly modelled structure, some fo the covarying residues occuring in the missing regions are not shown in the figure.

Correlated Mutation analysis of concatenated alignment of Sly1 and Syntaxin5

In Figure 4.35, the heat map of the MIp scores for the concatenated alignment of Sly1 and Syntaxin5 is shown. The highest scoring pair was an intra-protein residue pair, V339-N340 of Syntaxin5. In fact, most of the high-scoring pairs were intra-protein residue pairs either from Sly1 or Syntaxin5 protein. Overall, the covariation of residue pairs within a protein was stronger than between the two proteins. However, some high-scoring pairs were also found between the two proteins, for example, residue 69 of Sly1 & 175 of Sed5, 381 of Sly1 & 171 of Sed5, 616 of Sly1 & 325 of Sed5, etc.

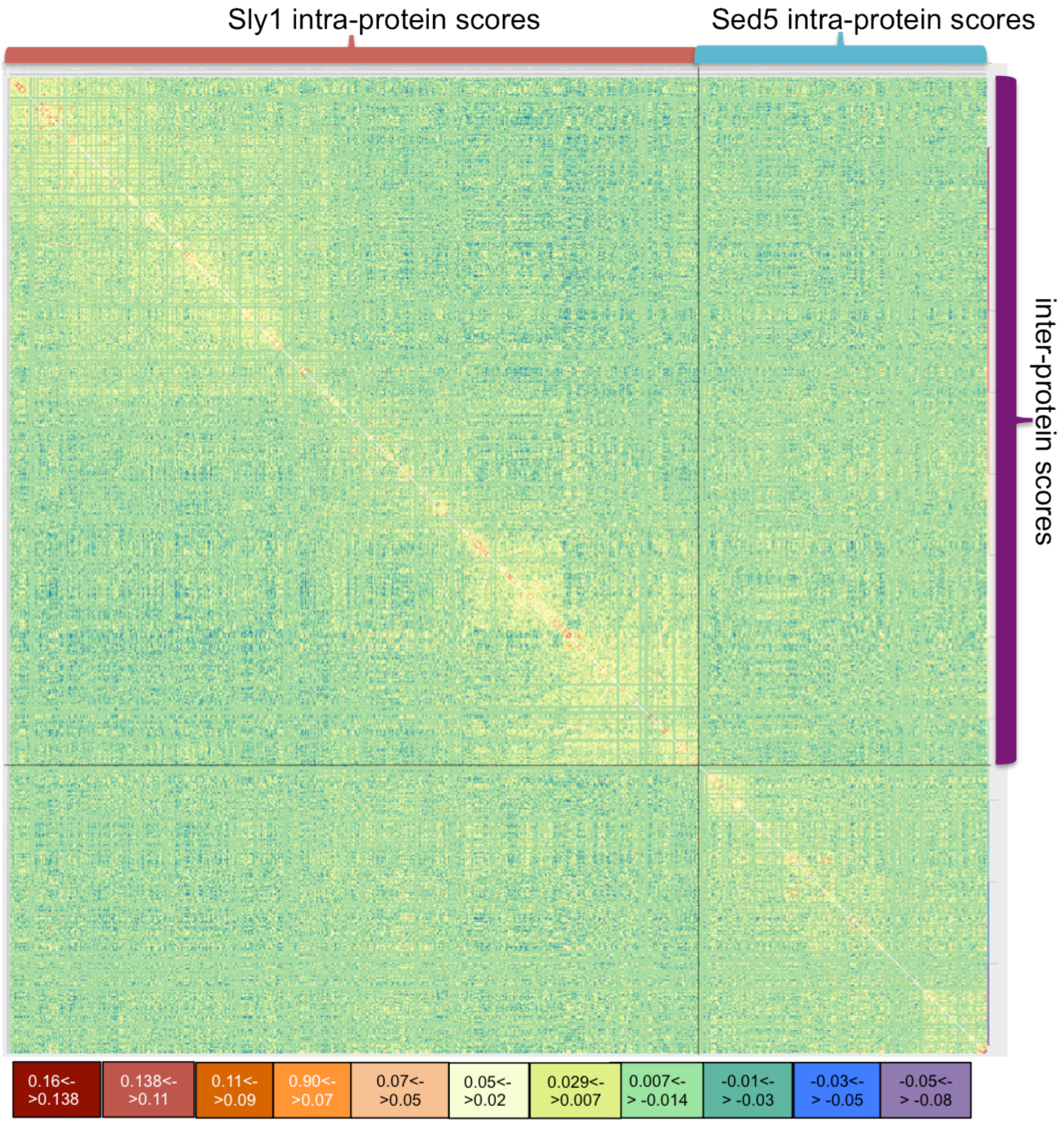


Figure 4.35: Heatmap of MIp scores of the concatenated alignment of Sly1 and Syntaxin5 proteins.
The heatmap is arranged as per sequence number of the concatenated alignment of Sly1 and Syntaxin5. The concatenated alignment had 301 sequences and 530 columns of Sly1 and 222 columns of Syntaxin 5.The color code bar represents the color corresponding to the range of MIp score. The dark red, red, dark orange, orange, represent the high MIp scores and light green, dark green, blue, purple represent the low MIp scores. The portions of the heatmap representing the intra-protein MIp scores of each protein and inter-protein MIp scores are demarcated. The original heat map can be found in the supplements (provided on request) that can be enlarged to have a clear view of the residue numbers.

Clusters of covarying residues within and between the Sly1 and Syntaxin5 proteins

To identify groups of strongly covarying residues, average linkage clustering was applied on the MIp dataset (figure of heatmap arranged by cluster ranking in Appendix A.180 (can be provided on request)). The average linkage clustering with stopping criteria of 2-σ of the MIp score and cluster selection cut off of 90% resulted in nine clusters with different numbers of residues from Sly1 and Syntaxin 5. The residues from the selected clusters were mapped on the combined structure (Figure 4.36). 54 residues of Syntaxin 5 were found to be co-varying with 69 residues of Sly1. . 2 residues in domain Ha, 10 in domain Hb, 9 in domain Hc, 15 in SNARE motif, 5 in Linker between Habc & SNARE motif, 1 in N-terminus and 8 in transmembrane (TM) domain of Sed5 were found covarying with Sly1. One residue was found in region between domain Ha and domain Hb, two residues were found in region between SNARE motif and TM domain and one residue was found in region between domain Hb and domain Hc. Similarly, 9 residues in domain d1, 9 residues in domain d2a, 3 in domain d2b, 48 in domain d3 domains of Sly1 were found covarying with residues in Sed5. However, 27 residues in domain d1 of Sly1 were found to be having intra-protein covariation, thus covarying with other residues in domain d1 of Sly1. Two-third of the identified inter-protein residues were located in regions known to interact with Sly1, i.e. the Habc domain and the SNARE motif of Syntaxin 5. The residues in the clusters were lying distant from each other and appeared to form a network of connected residues.

The inter-protein clusters had many residues that were also identified in the clusters of the intra-protein analysis of Sly1 and Syntaxin5. For example, cluster 1 from intra-protein analysis of Sly1 had many residues that were also identified in the cluster 2 of inter-protein analysis of Sly1 with Syntaxin5.

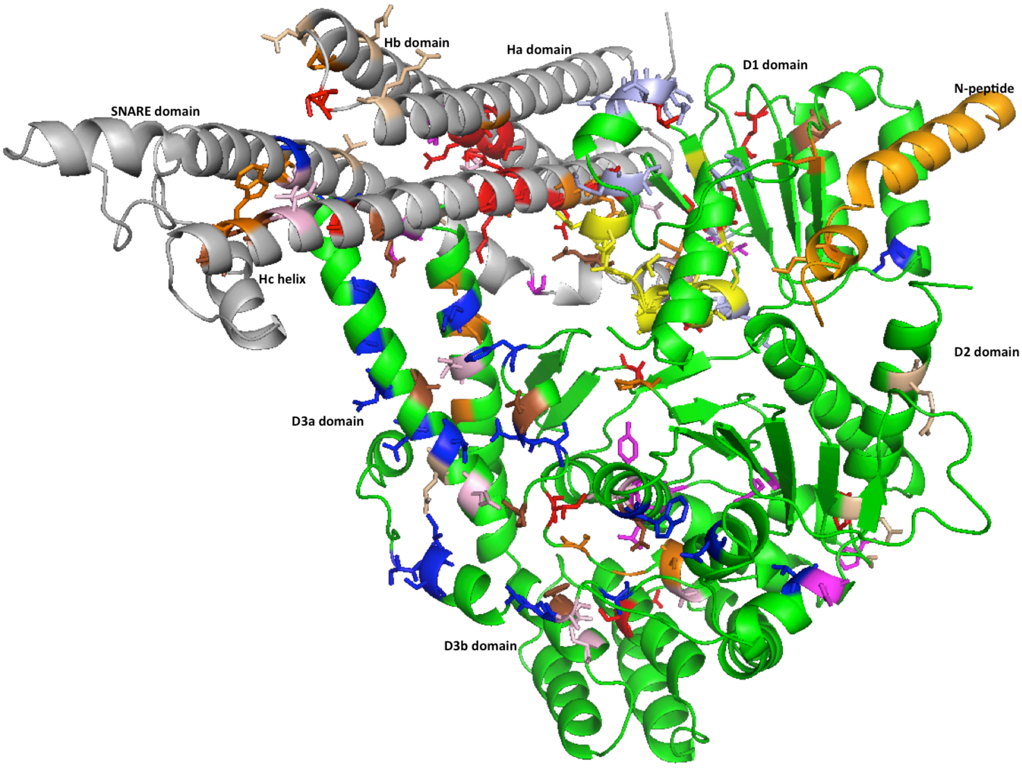


Figure 4.36: Residue from the top 90% clusters marked on the structure obtained by homology model of the Sly1/Syx5 structure.
Sly1 structure is shown in green, the structure of cytoplasmic domain of Syntaxin5 is shown in grey and the N-peptide of Syntaxin5 is shown in orange. The cytoplasmic domain of Syntaxin5 containing the Habc and SNARE domain was homology modeled on the crystal structure of the Syntaxin1 from rat by using Phyre2 server (Mezulis et al. 2015). The resulting structure was then overlapped with 1MQS of Sly in complex with N-peptide of Syntaxin5 and 3C98 of Munc18-1 in complex with Syntaxin1, using Pymol, so as to obtain an approximate position for the cytoplasmic domain of Syntaxin5. In the model of the Sly1/Syx5 complex, the hairpin helices of Sly1 and the Habc domain of Syx5 clash sterically. Red residues represent the first cluster, blue residues represent the second cluster, orange residues represent the third cluster, yellow residues represent the fourth cluster, magenta residues represent the fifth cluster, light pink represents residues form sixth cluster, light blue represent the sventh cluster, light brown represents the eigth cluster, dark brown represents the ninth cluster.2 residues in domain Ha, 10 in domain Hb, 9 in domain Hc, 15 in SNARE motif, 5 in Linker between Habc & SNARE motif, 1 in N-terminus and 8 in transmembrane (TM) domain of Syntaxin5 were found covarying with Sly1. 1 residue was found in region between domain Ha and domain Hb, 2 residues were found in region between SNARE motif and TM domain and 1 residue was found in region between domain Hb and domain Hc. Similarly, 9 residues in domain d1, 9 residues in domain d2a, 3 in domain d2b, 48 in domain d3 domains of Sly1 were found covarying with residues in Syntaxin5. However, 27 residues in domain d1 of Sly1 were found to be having intra-protein covariation, thus covarying with other residues in domain d1 of Sly1.

A detailed look on the cluster 2 from the inter-protein analysis of Sly1 and Syntaxin5 showed that many residues were lying in the two hairpin helices of domain d3a, in domain d2 and in the Sly1 specific loop regions of Sly1 and 4 residues were in the SNARE motif of Syntaxin5 (Figure 4.37). The inter-protein covarying residues in Syntaxin5 were lying at the N- and C-terminal region of SNARE motif (Figure 4.38). Most of them were lying adjacent to the conserved layered residues. Residue 258 was lying at the beginning of the SNARE motif.

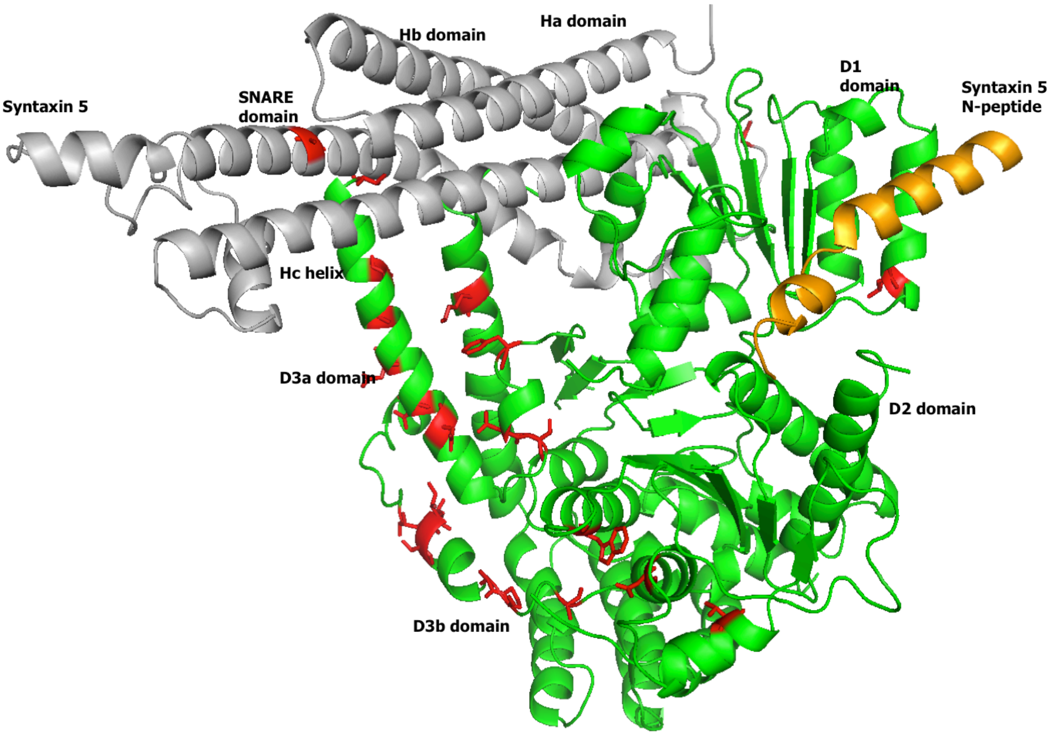


Figure 4.37: Residue from the inter-protein cluster2 marked on the structure obtained by homology model of the Sly1/Syx5 structure.
The structure is same as used in Figure 4.36. The covarying residues are indicated in red.

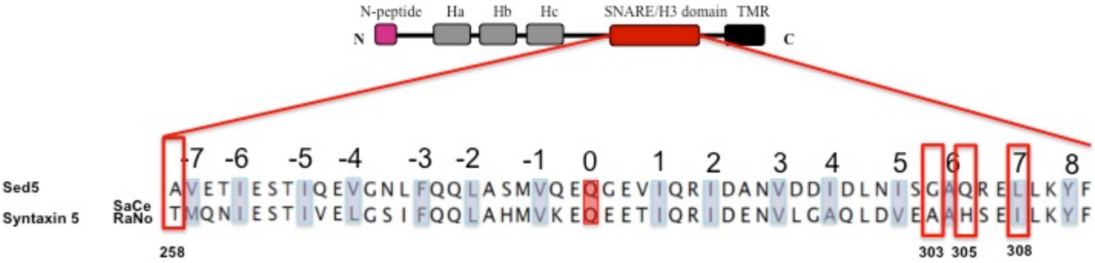

N-peptide
Ha
Hb
Hc
SNARE/H3 domain
TMR
N
C
-7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8
Sed5
AVETIESTIQEVGNLFQQLASMVQEQGEVIQRIDANVDDIDLNISGAQRELLKYF
Syntaxin 5
SaCe
RaNo
TMQNIESTIVELGSIFQQLAHMVKEQEETIQRIDENVLGAQLDVEAAHSEILKYF
258
303
305
308

Figure 4.38: SNARE motif of Syntaxin 5 from rat and yeast.
The inter-protein covarying residues are marked in red boxes

Mapping the 80% conserved residues on the structure showed that the covarying residues obtained were lying adjacent to the conserved residues (Figure 4.39).

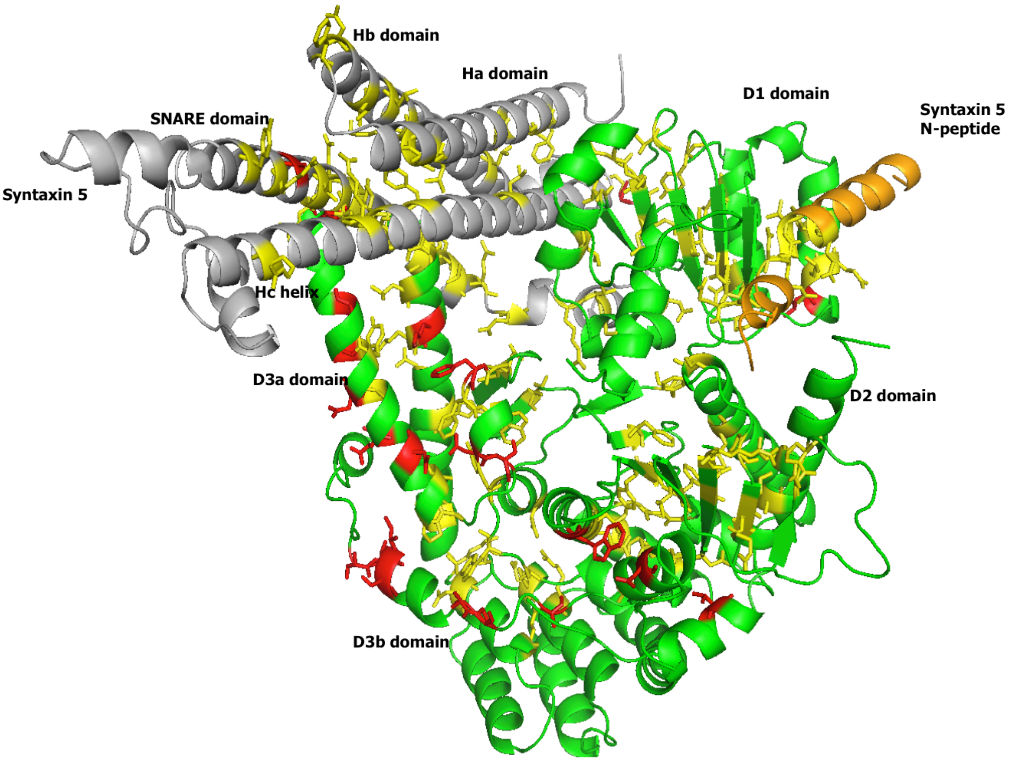


Figure 4.39: Residue from the inter-protein cluster2 marked on the structure obtained by overlapping 1MQS (of Sly1) and homology modeled structure of Syntaxin 5 on SSO1.
The structure is same as used in Figure 4.36. The covarying residues are indicated in red and the 80% conserved residues are indicated in yellow.

Having a detailed look at the network view of certain high scoring connections of inter-protein cluster 2, again showed the occurrence of two subclusters within the cluster (Figure 4.40). The two subclusters were same as observed in the intra-protein analysis of Sly1 and also occupied the same regions within the protein (Figure 4.28, Figure 4.29, Figure 4.30, Figure 4.31). This again suggested that the different important region of Sly1 were connected by chains of covarying residues within the 3D structure of the protein. The highest scoring inter-protein connection was between residues 258 of Syntaxin5 and 550 of Sly1. However, they appeared to be lying distant from each other. The residue 258 of Syntaxin 5 was in the beginning of the SNARE motif and the residue 550 of Sly1 was lying on the end of the α21 helix of Sly1 specific loop (Figure 4.41). The residues 258 of Syntaxin5 was also connected to residues 361 of Sly1 however with a low covariation score. These appeared to be lying close to each other as 361 of Sly1 was on the tip of hairpin helix of domain d3. This suggested that the two important regions of Sly1, responsible for interaction with R-SNARE, Qa-SNARE and Rab GTPase were coupled with the SNARE motif of Syntaxin5. This indicates a possible coevolution between the two interacting proteins.

It is thus possible that the SM protein prepares the bound syntaxin for SNARE complex assembly.

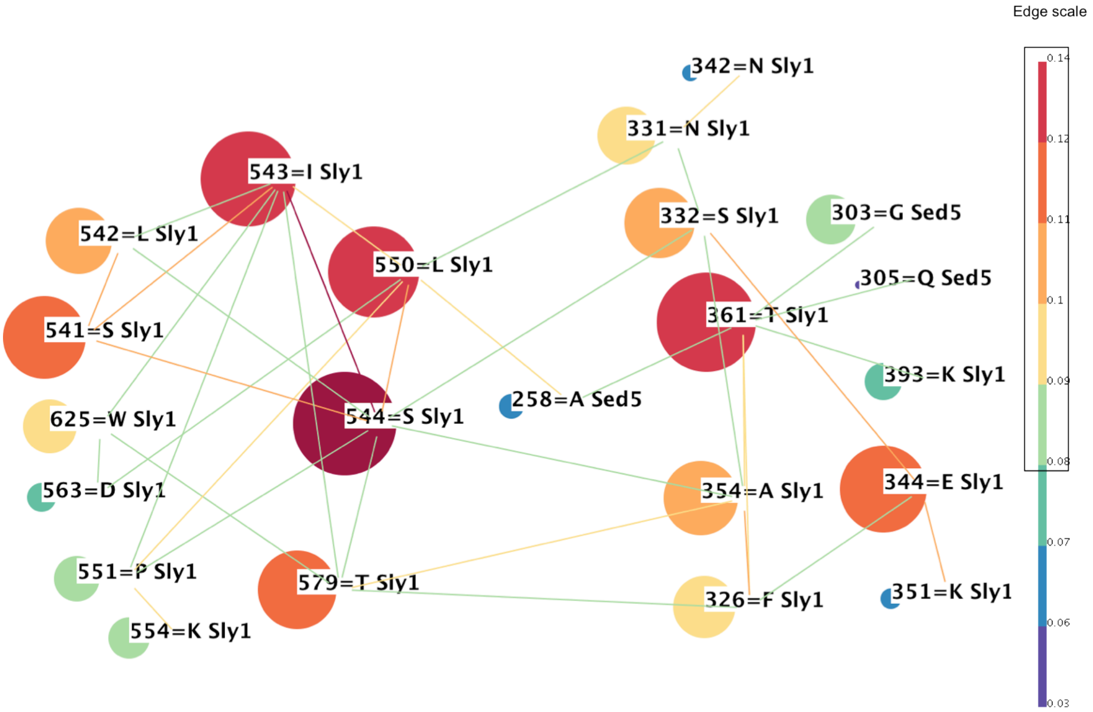


Figure 4.40:Network view of inter-protein cluster2 after removing low scoring edges.
The color codes are same as in Figure 4.31. The black box around the edge scale indicated the high scoring connections shown.

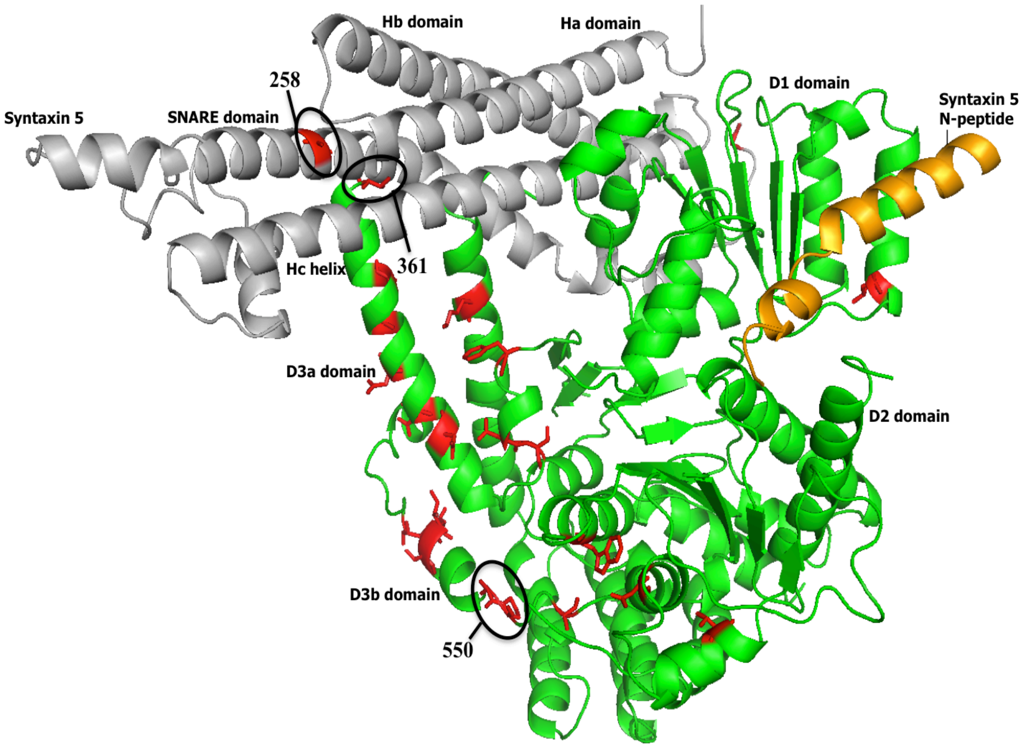


Figure 4.41: Residue from the inter-protein cluster2 marked on the structure obtained by overlapping 1MQS (of Sly1) and homology modeled structure of Syntaxin 5 on SSO1.
The structure is same as used in Figure 4.39. The covarying residues are indicated in red. The highest scoring inter-protein connections are shown in black circles.

Clade-specific analysis of inter-protein covarying residues

The aim of performing the clade specific analysis for the high scoring inter-protein residues within the cluster, was to see if the pattern of substitution of covarying site pair is congruent with the tree or not. The sites undergoing inter-protein covariation should show similar substitution pattern during evolutionary history of both the proteins.

The weblogo representation showed the distribution of amino acid at the highest scoring inter-protein residue pair of cluster2, 258 of Syntaxin5 and 550 of Sly1. The residue 258 of Sly1 had A, T, S and E (Figure 4.42). The residues 550 of Sly1 had L, V and I (Figure 4.43).

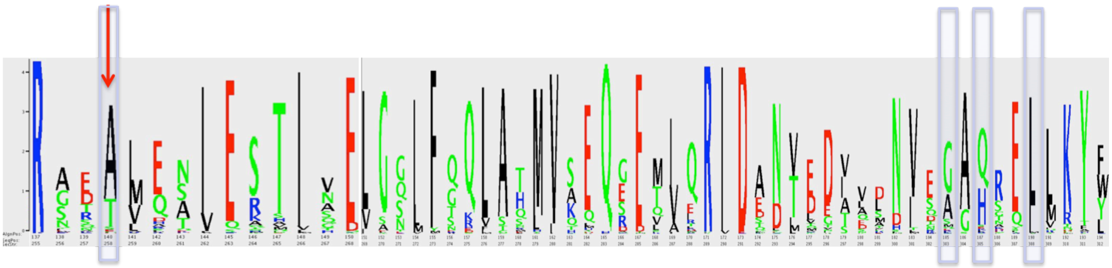


Figure 4.42: Weblogo of Syntaxin5 SNARE motif.
The covarying residues are indicated by purple boxes and the highest scoring inter-protein residue, 258 is indicated by a red arrow.

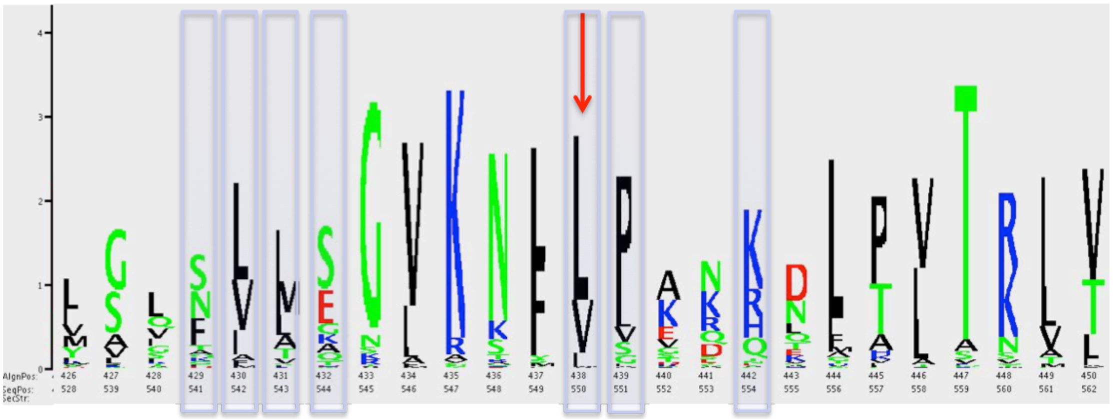


Figure 4.43: Weblogo of Sly1 specific loop region.
The covarying residues are indicated by purple boxes and the highest scoring inter-protein residue, 550 is indicated by a red arrow.

In order to see the amino acid distribution of the covarying sites in different lineages, the substitution pattern for the highest scoring inter-protein residue pair of cluster2 was then mapped on the tree of Sly1 and Syntaxin5, respectively. The residue 258 of Syntaxin5 varied from A in all Fungi species and plants, T in all Metazoa, S in Stramenophiles and E in Microsporidia (Figure 4.44). The residue 550 of Sly1 varied from L in all Fungi species and plants, V in all Metazoa, V/L in Stramenophiles and I in Microsporidia (Figure 4.45).

The substitution pattern showed that the inter-protein covarying residues changed at the same corresponding branches in the two proteins during the evolution. Similar substitution pattern was observed for other high scoring inter-protein covarying residues in other clusters. Thus, the identified inter-protein residues showed clade-specific covariation and might be involved in maintaining the interaction and thus evolution of the two proteins.

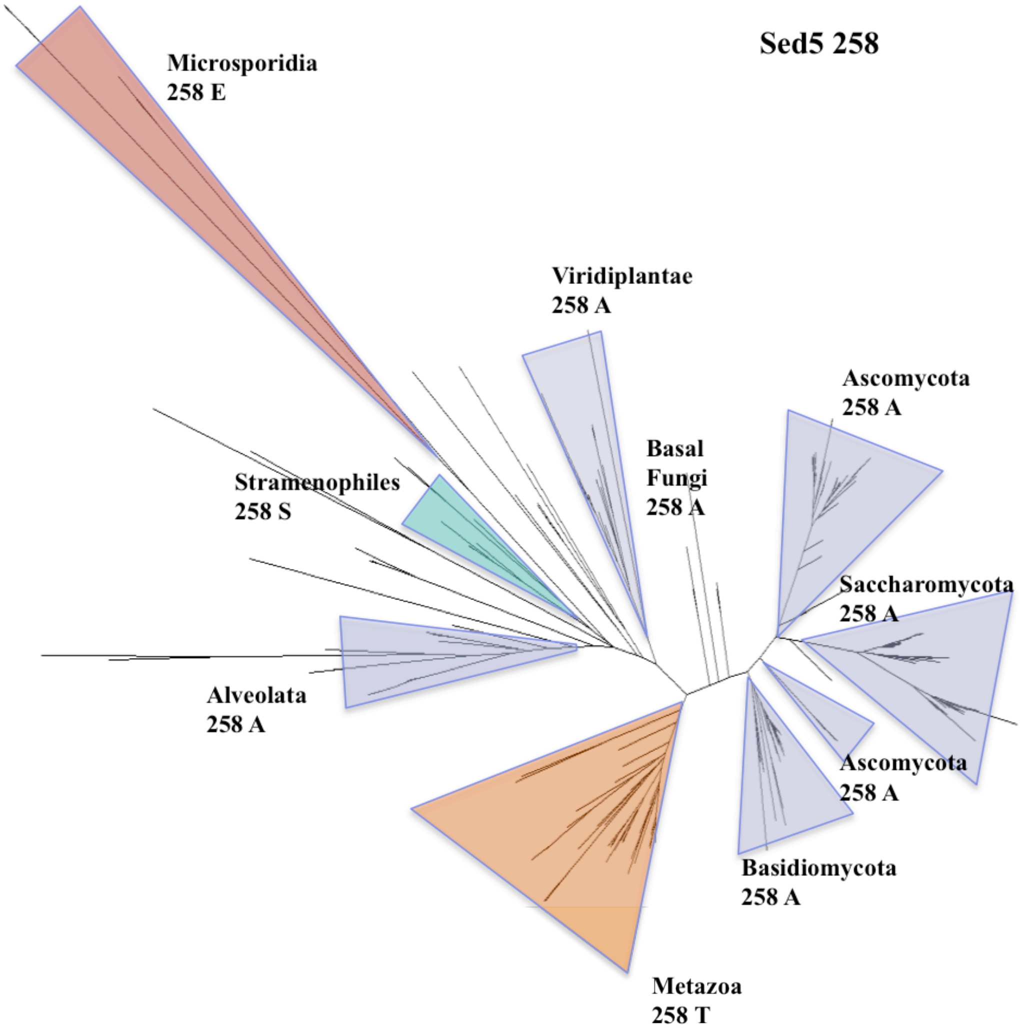


Figure 4.44: Residue 258 of Syntaxin5 from inter-protein cluster2, mapped on the phylogenetic tree of Syntaxin 5.
The colors represent the changes in distribution of the covarying inter-protein amino acids. The branches that have same amino acid distribution are colored in same color and the ones where the amino acid distribution changes are colored differently

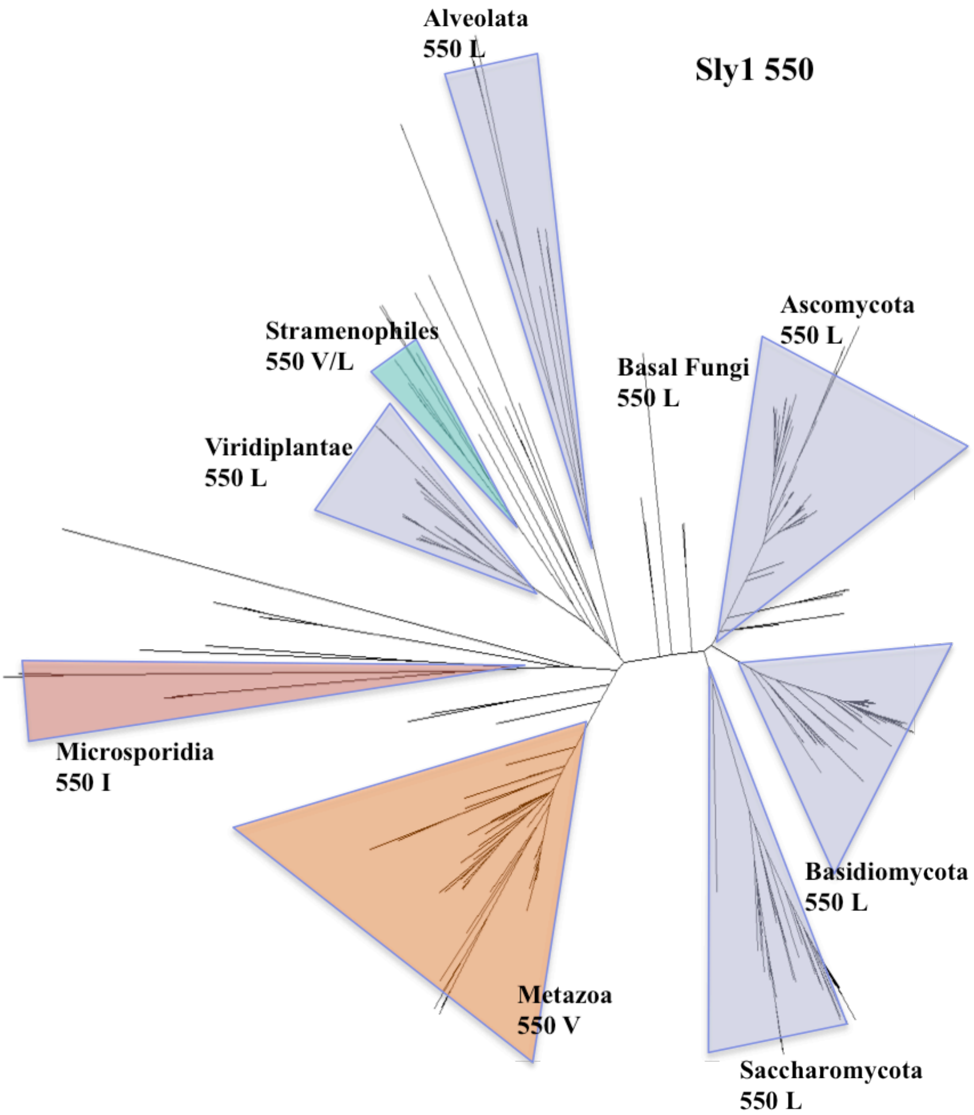


Figure 4.45: Residue 550 of Sly1 from inter-protein cluster2, mapped on the phylogenetic tree of Sly1.
The colors represent the changes in distribution of the covarying inter-protein amino acids. The branches that have same amino acid distribution are colored in same color and the ones where the amino acid distribution changes are colored differently.

4.6 Intra-protein analysis of other SM protein subfamilies

In order to analyze the other SM protein family members, alignments of each subfamily were generated and preprocessed as described before.

Correlated Mutation analysis on SM Proteins

The heatmap of the MIp scores for the Munc18/Sec1, Vps45, and Vps33 alignments (figure can be provided on request) revealed that most of the covarying residues lie in the beginning of domain d1, domain d2a, end of domain d3 and end of domain d2b.

Clusters of covarying residues in SM proteins

Similar to the result of Sly1, for Munc18/Sec1 subfamily, some of the topmost selected clusters (above the cluster selection cut off) had residues that were spread over different regions of the protein and thus forming networks connecting them, while some clusters had residues that were confined to a certain region of the protein. In case of Vps33 and Vps45 subfamilies, most of the topmost selected clusters had residues that appeared to be confined to a certain region of the protein. However, in all the SM subfamilies, the covarying residues occupied similar corresponding regions.

In all the three SM subfamilies, residues were identified in the beginning of domain d1, in the large cleft between domain d1 and domain d3 and in domain d3, including helix-13 and helix-14. The clusters identified on all three SM protein subfamilies again showed the occurrence of residues lying in close contact in the tertiary structure as well as residues that were lying further away from each other. Thus, for the different SM protein subfamilies a comparable pattern of highly scoring pairs within the 3D structure were found. This suggests that similar regions of the different subfamilies co-vary.

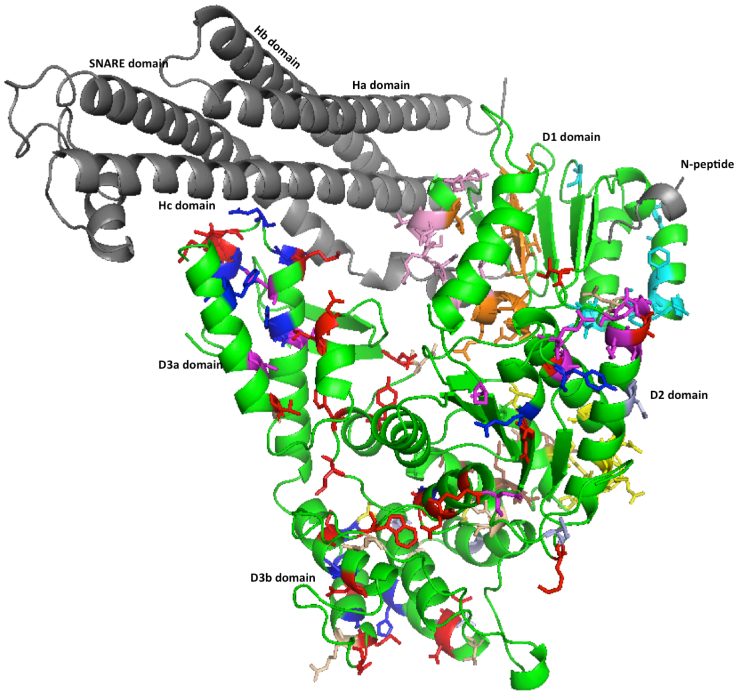


Figure 4.46: Residue from the top 85% clusters marked on the structure of *Rattus norvegicus* Munc18-1 (3C98).
Munc18-1 structure is shown in green and Syntaxin1 structure is shown in grey. Red residues represent the first cluster, blue residues represent the second cluster, orange residues represent the third cluster, yellow residues represent the fourth cluster, magenta residues represent the fifth cluster, light pink represents residues from sixth cluster, light blue represent the sventh cluster, light brown represents the eigth cluster, dark brown represents the ninth cluster, cyan represents the tenth cluster.

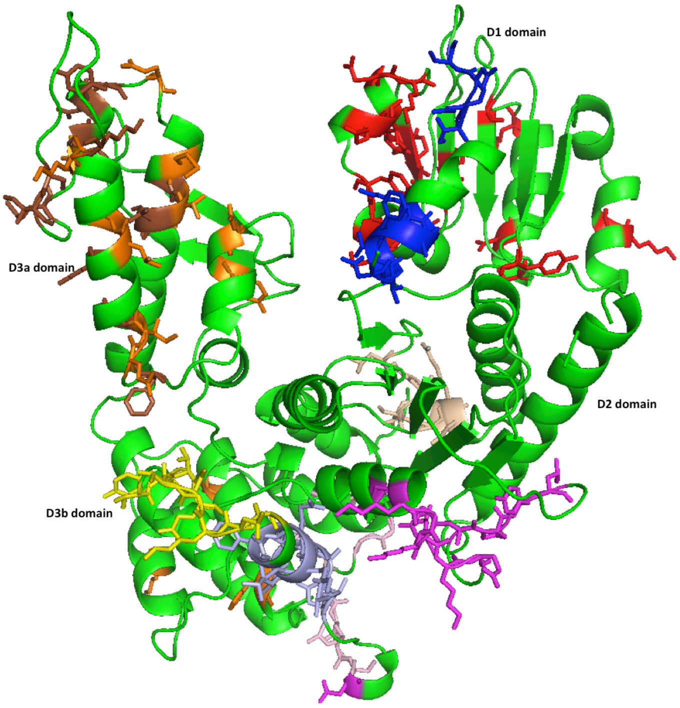


Figure 4.47: Residues from the top 90% clusters marked on the structure of *Rattus norvegicus* Vps45 homology modelled on Munc18-1 of *Rattus norvegicus*.
For Vps45 no crystal structure is available so far. Instead, a structure was obtained by homology modeling of *Rattus norvegicus* Vps45 on Munc18-1 of *Rattus norvegicus* by using the Phyre2 server (Mezulis et al. 2015). Vps45 structure is shown in green. Rest of the color code are same as in Figure 4.46.

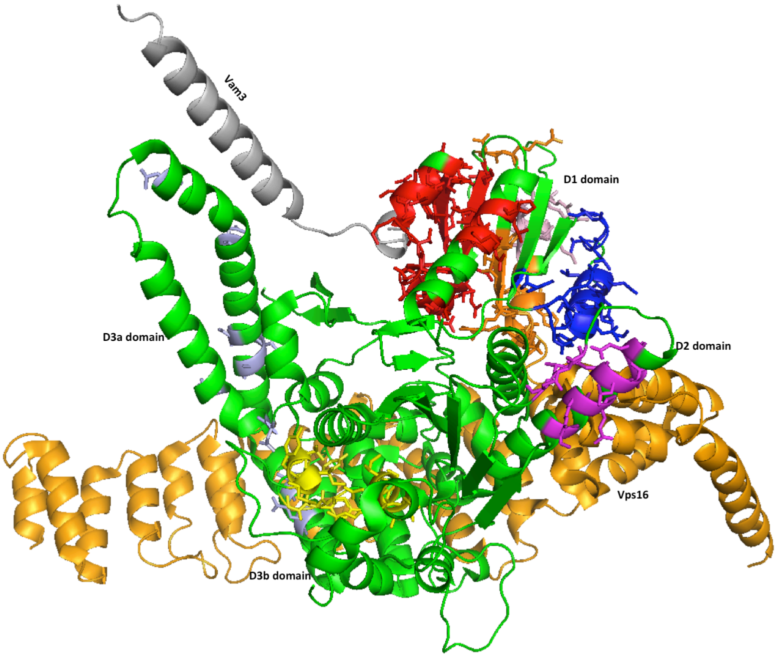


Figure 4.48: Residue from the top 85% clusters marked on the structure of *Chaetomium thermophilum* Vps33 (5BUZ).
The structure of *Chaetomium thermophilum* Vps33 is shown in green, the Vam3 structure is shown in grey and Vps16 structure is shown in golden. Rest of the color code are same as in Figure 4.46.

5 Discussion

Vesicle fusion is an essential process of the eukaryotic cells by which transport vesicles that bud from a donor compartment fuse specifically with an acceptor compartment. During past decades, conserved homologous sets of protein machineries involved in this process have been identified. The SNARE (Soluble N-ethylmaleimide-sensitive factor Attachment protein REceptors) proteins and the SM (Sec1/Munc18) proteins are considered to be the core of the vesicle fusion machinery. Several other conserved factors such as, Rab proteins, NSFs (N-ethylmaleimide-Sensitive Factor), SNAPs (Soluble NSF Attachments Proteins) and tethering proteins belonging to the CATCHR (Complex Associated with Tethering containing Helical Rods) family (Jahn & Scheller 2006; Jahn & Fasshauer 2012; Sudhof & Rothman 2009) also participate in the vesicle fusion process. Recently, it is becoming clear that these molecular machineries arose by duplication and diversification of a prototypic machine during evolution (Cai et al. 2007;Mast et al. 2014).

A large number of cell biological, genetic, and biochemical studies have been carried in last decades that helped to understand the protein interaction networks in vesicular trafficking. However, the exact order of molecular events is still unclear and there is still a gap in understanding the molecular features of the key proteins. As most of the studies are carried out only in few organisms, it is unclear whether a particular characteristic represents a special adaptation of a protein or is it a common phenomenon. Sequence analysis can help in filling this gap as the whole set of sequences across all species can be investigated at a time. It can help to understand how the mechanism has adapted in different eukaryotic lineages and in different vesicle trafficking steps within the cell. Novel insights about the structural and functional features of the vesicle fusion machineries as well as the co-evolutionary patterns between interacting proteins can help to explain the molecular events and provide new directions for the biochemical research.

Previously, a lot of work has been carried out in my group in analyzing the sequences of the various factors involved in the vesicle fusion step. A database management system was developed to store and analyze the sequences of various protein families involved in vesicle fusion, such as SNARE proteins, Rab proteins, SNAP proteins, NSF family, and the SM protein family. Phylogenetic analysis was carried out for SNARE proteins (Kloepper et al. 2007; Kloepper et al. 2008; FKienle et al. 2009) and Rabs proteins (Klöpper et al. 2012) as well as for SM protein family (unpublished). The phylogenetic analysis of SNARE proteins and SM proteins had revealed some comparable patterns of duplications and diversifications between the closely interacting proteins. This previous work provided an opportunity to explore the evolutionary changes occurring in the proteins of vesicle fusion machinery across different eukaryotic lineages using the large sequence collection.

The primary aim of this work was thus to extract functional and structural information and to explore the covariation pattern from multiple sequence alignments. Analysis of covarying positions in a protein family is thought to provide important information about the sites that have functional importance and are involved in the structural stability of the protein (Misura & Weis 2000; Wollenberg & Atchley 2000; Gloor et al. 2005; Tillier & Lui 2003; Travers & Fares 2007). The information thus gained would help to provide new directions for the future structural or biochemical research and thus provide a better understanding about the function and interactions of the vesicle fusion proteins.

With this aim, I started the intra-protein covariation analysis, initially on the SM proteins. Generally, SM proteins interact with Qa SNAREs (also known as Syntaxins) and thus control and guide the vesicle fusion process. Compared to the other protein families of the vesicle fusion repertoire, SM proteins had only few different, but highly conserved subtypes. The SM protein family comprises five subtypes: Sec1/Munc18, Sly1, Vps33, Vps45, and scfd2 that play key roles in different trafficking steps within the cell. The binding mode of SM protein with their partner Syntaxins is still not clear and is highly debated. Various recent structural and biochemical studies have suggested that, in general, the SM protein and Syntaxins interact tightly via two spatially separated binding sites (Bracher & Weissenhorn 2002; Khvotchev et al. 2007; Carpp et al. 2006; Furgasona et al. 2009; Aran et al. 2009; Johnson et al. 2009; Burkhardt et al. 2008; Demircioglu et al. 2014), (Hackmann et al. 2013). How these two binding interactions enable the SM proteins to control the accessibility of the bound Syntaxin is still not clear yet. The initial structure of secretory SM protein, Munc18-1 revealed it to be tightly bound to major portion of Syntaxin 1 in a closed conformation and thus to block SNARE assembly (Misura & Weis 2000). By contrast, Sly1 structure was solved where it was bound only to the N-peptide of Sed5 in an open conformation and thus assist in the SNARE complex formation (Bracher & Weissenhorn 2002). Later, it was shown that Munc18-1 also binds to the N-peptide of Syntaxin 1 although with a lower affinity (Burkhardt et al. 2008). Recently, Sly1 was also shown to interact not only with N-peptide but also with the remainder of Sed5 confirming the requirement of a conformational switch between a closed and open conformation of Syntaxin for its function (Demircioglu et al. 2014). Thus, both the binding sites allow the SM proteins to interact with their partner Syntaxin and control the conformational switch of the bound Syntaxin. However, the reason for the occurrence of two different binding sites has yet not been completely understood. It has been speculated by some researchers in the field that probably the two binding sites are allosterically coupled to control the accessibility of the bound Syntaxin (Dawidowski & Cafiso 2013; Colbert et al. 2013; Demircioglu et al. 2014). To understand how the two binding sites communicate and if there is a possible conformational switch between the two binding interactions, I aimed at extracting novel information by looking at their sequence covariation.

5.1 Covariation analysis by MIp and average linkage clustering

With the aim of exploring the sequence covariation, I initially implemented different sequence-based statistical methods into one bioinformatics framework. The methods implemented for the covariation analysis included McLachlan based Substitution Correlation (McBASC) (Gobel et al. 1994), Statistical Coupling Analysis (SCA) (Lockless & Ranganathan 1999), Positional Conservation based SCA (SCAnew) (Halabi et al. 2009), Mutual Information (MI) (Martin et al. 2005), Mutual Information Corrected (MIp) (Dunn et al. 2008), Explicit Likelihood of Subset Variation (ELSC) (Dekker et al. 2004), Observed Minus Expected Squared (OMES) (Larson et al. 2000). Of these, a rapid and widely used sequence covariation detection method, MIp (Dutheil 2011; Dunn et al. 2008; Liu & Bahar 2012; Buslje et al. 2009; Chakrabarti & Panchenko 2010; Dickson et al. 2010) was selected for an initial analysis in the current work, because it has been shown to identify a higher number of contacting residues compared to other coevolution detection methods (Tillier & Lui

2003; Dunn et al. 2008; Caporaso et al. 2008; Dutheil 2011). Nevertheless, the other methods need to be explored and compared to MIp in the future.

MI is a Shannon entropy based method that requires calculation of individual and joint amino acid frequencies between the columns. It suffers from problems like noise from phylogenetic background and many authors have pointed out that pairs with high MI scores are not always the true coevolving pairs (Tillier & Lui 2003; Wollenberg & Atchley 2000; Buslje et al. 2009; Gouveia-Oliveira & Pedersen 2007; Dunn et al. 2008; Dutheil 2011; Buslje et al. 2009), . Several modifications of MI were subsequently developed to correct for the phylogenetic biasness Tillier & Lui 2003; Dutheil 2011; Buslje et al. 2009; Gouveia-Oliveira & Pedersen 2007; Dunn et al. 2008). (Dunn et al. 2008) corrected this bias of MI by a simple multiplicative correction. This improved method is referred to as MIp (Dunn et al. 2008; 2012). Comparative studies have been carried out that showed that MIp is able to predict covarying positions better than other sequence based or even tree-based methods (Caporaso et al. 2008; Dunn et al. 2008; Dutheil 2011). To identify only closely contacting pairs, another method called Direct Coupling analysis (DCA) was proposed (Marks et al. 2011; Marks et al. 2012). This approach and its modifications have been used for the prediction of 3D-structure of protein complexes (Marks et al. 2011; Liu & Bahar 2012; Marks et al. 2012; Hopf et al. 2012; Hopf et al. 2014; Hopf et al. 2015). Since our aim was not to predict the 3D-contact and not to find only the closely contacting correlated residue pairs but to identify functionally and structurally important covarying residues, we decided to use a widely used sequence covariation analysis method, MIp.

As outlined above, several different methods to extract information about the covariant sites from MSAs are currently available, but have not attracted much attention among biologists. By contrast, methods to predict secondary structure, TMR regions, coiled coil regions, signal sequences, TPR repeats to name only a few, are regularly used by many researchers as they bring to light easy interpretable information about a protein. Until now, not many biologists have used the covariation detection methods, as it is still unclear whether co-varying positions are indeed important for the structure and function of a protein. Although this has been claimed in several studies on covariation, most of such studies were carried out only on enzymes or small protein families. A large landscape of proteins with different functions and structures still remains unexplored for identifying the covariant sites. When I started the analysis on SM proteins by using the approach introduced by Gloor et al. in 2005; Martin et al. in 2005; Dunn et al. in 2008, I found covarying residues that were not always in close proximity in the 3D structure but were spread over different regions of the proteins. Some residues were found in groups or clusters of many residues that were lying distant from each other in tertiary structure. This result was different to what had been shown by previous studies on other proteins and was not easy to understand. The earlier studies mostly used enzymes and results were easier to interpret (e.g:(Gloor et al. 2005; Martin et al. 2005; Dunn et al. 2008; Lee et al. 2012; Chakrabarti & Panchenko 2010; Ackerman & Gatti 2011). They often found the high scoring pairs to be lying close to each other and in close proximity to the enzyme’s catalytic/substrate-binding site. However, in some cases (e.g.: (Liu & Bahar 2012; Gloor et al. 2005)) they also found some networks on co-varying sites. These studies (Gloor et al. 2005; Liu & Bahar 2012) considered only the list of highly covarying residue pairs. They identified some sites as pairs and some that were coevolving with many other sites by inspecting the list of high-scoring residue pairs in a non-automated fashion. Their method wasn´t designed for analyzing covariant sites that form networks as they looked only at certain subset of high scoring pairs. Thus, I realized that although current approaches of looking just at certain pairwise sites were sufficient to extract meaningful information about some example proteins, but for proteins like SM proteins, the covariation data is more complex. So, I needed to modify the analysis approach in order to understand the complexity of the covariation happening in such proteins. I developed an automated approach based on clustering of the MIp dataset so as to detect residue pairs that have high scores not only with one partner but with several others as well. I also developed heat maps to visualize the entire MIp data set, as they would help to easily spot the high scoring covariant sites.

The approach was then applied on two test examples. These test examples included MAP (Gloor et al. 2005) and UDG (Liu & Bahar 2012) datasets from previous studies. In both the test cases, I was able to confirm their analysis with some minor additions by using my cluster approach. My approach revealed regional patterns of covariance that were relatively unexplored in the earlier work carried out by focusing just on pairs of columns. The clusters in MAP and UDG datasets were concentrated around the catalytic core. I also tried the approach on a different type of protein of the vesicle fusion machinery, SNAP, and again found the clusters to be confined to a particular region of the protein. The clusters had residues that were lying close to each other as well as some of the distantly lying residues.

The MIp scores obtained on the SM protein family, SNAP protein family and on the enzymes (MAP, UDG) were within a comparable range (Table 5.1).

Table 5.1: Maximum and minimum MIp scores obtained for different protein families/subfamilies studied.

Protein families/Subfamilies	Maximum MIp score	Minimum MIp Score
α-SNAP	0.138	-0.06
γ-SNAP	0.129	-0.06
Combined SNAP alignment	0.14	-0.07
Sly1	0.14	-0.08
Munc18/Sec1	0.17	-0.08
Vps45	0.14	-0.07
Vps33	0.13	-0.06
S1A protease	0.19	-0.03
MAP1	0.16	-0.06
UDG	0.20	-0.05

The value of MIp scores obtained for all studies was similar.

The covarying residues within the clusters identified in Sly1 were not restricted to a particular region, but were spread over different regions of the protein. Some of the residues within the clusters were lying close to each other, while some were lying distant from each other in the tertiary structure. The distantly lying residues appeared to be connected by chain or networks of other closely lying residues within the tertiary structure of the protein. The improved approach with clustering of MIp dataset thus helped to identify the networks of covarying residues occurring within the protein. However, it was noticed that the approach was not able to identify some of the paired residues that had high MIp scores. These residues mostly formed very small clusters, comprising of just 2, 3 or 4 residues and did not have covariation-based connection with many other residues. This could be an effect of the clustering approach because of which they were listed into small clusters. These clusters had very low average weighted degree score of the cluster and ranked very low and were thus neglected. In order to include such small clusters, one need to adapt the clustering criteria accordingly.

5.2 Application of MIp and average linkage clustering on α-SNAP proteins identified networks of functionally and structurally important residues in N- and C-terminal regions and in TPR region.

Application of the combinatorial analysis pipeline of MIp and average linkage clustering on the entire SNAP alignment and on the individual α-SNAP and γ-SNAP alignments resulted in defined groups of covarying residues that were localized and restricted to a particular region of the protein. This could be because the α-SNAP does not undergo large conformational changes during its interaction with SNARE complexes and NSF. This suggests that distinct regions of the protein carrying out certain activities do not communicate with each other, as the α-SNAP protein is extended and rigid these regions.

Important residues in N-terminal region

Of all the detected residues in the N-terminal region, residues 1-10, 12, 14-15 were located in the first α-helix, residue 16 is in the membrane attachment loop, while residues 21, 22, 24, 25 were in the short helix of the same loop and the residues 30, 31, 35 were at beginning of the second α-helix. Residues I1, S2, D3, P4 were lying on the very short loop at the beginning of the first α-helix. The residue 4 has a proline in case of Ascomycota or glycine in case of plants and lies just at beginning of the helix. Proline is known to cause kinks or breaks in the helix, however, it is often seen as the first residue of the helix due to its structural rigidity. PDBePISA (Krissinel & Henrick 2007) online service provided by EMBL was used to identify the solvent accessible residues in the structure of Sec17. Residues P4, L7, L8, E12, K14, G15, V16, S30, E35 were identified as the solvent accessible residues. Some of the solvent exposed residues also participate in polar interactions that might be important in determining the shape of the protein. Solvent exposed residue K14, forms a salt bridge with another solvent exposed residue E35 (Rice & Brunger 1999). This interaction is between the first and second α-helix and thus might be involved in stabilizing the end points of the extended loop. Residues P4 and V5, E6 and R10 were also involved in side chain polar interactions. The two previously mutated residues, F21 and M22 in Sec17 (F27 and F28 in α-SNAP), occurring in the membrane-binding loop, were also identified with the current approach. These residues have been shown to be required for binding to the membrane, which helps in efficient binding to the membrane-anchored SNARE complex (Winter et al. 2009).

The N-terminal region is known for the interaction of α-SNAP with SNARE complex. Deletion of 28 N-terminal residues carried out by Hayashi et al. in 1995, showed a reduction in α-SNAP-SNARE complex binding by 75%, but did not affect NSF binding. Griff et al. in 1992, showed that the first 17 residues are not essential for the proper folding or the production of the protein. They identified a SEC17 fragment that complemented the sec17-1 temperature sensitive mutation. This allele had a deletion in the N-terminal region, including the first exon, intron and 8 codons of second exon of SEC17. The expressed Sec17p from this allele was truncated with first 17 amino acids from the first helix deleted. However, the cells carrying this plasmid overproduced Sec17p of a lower molecular weight than the native Sec17p. Thus, these first 17 residues are not required for the production and proper folding of the protein. DeBello in 1995 showed that injection of peptides corresponding to 1-24 and 19-31 of squid SNAP (1-21 and 19-31 of Sec 17, including the first helix and the loop) and 144-163 of mammalian α-SNAP (140-159 of Sec 17) had an inhibitory effect on neurotransmitter release. These peptides might have competitively interfered with SNAP-SNARE binding and SNAP-NSF binding.

These observations from literature and available biochemical and structural data (Griff et al. 1992; Hanson et al. 1995; Hayashi et al. 1995; Barnard et al. 1997) suggests that some of the detected N-terminal residues from the first helix and the membrane attachment loop might be involved in interaction with SNARE complex or in SNAP-SNAP interaction, needed for SANRE binding. These residues might not be required for proper folding of the SNAP protein. As was shown by Griff et al. in 1992, that the allele with deleted N-terminal region was still able to express Sec17p truncated with first 17 amino acids and of lower molecular weight. Thus, the identified N-terminal residues might be essential for the formation of the first α-helix and the membrane attachment loop. Mutation of any of the identified residues could result in distortion of the first α-helix. These residues thus might participate in the correct placement of SNAP on the membrane, so that the membrane attachment loop can be anchored, and the binding to the SNARE complex can occur correctly.

Important residues in C-terminal region

Residue 234 and 235 of cluster 5 were lying in the beginning of the C-terminal region but were missing in the structure. Residues 253-255, 259-262 were lying on α12-helix, residues 265, 267, 268 were lying on the short connecting loop between α12 and α13-helix and residues 271, 273-276, 278-279, 282, 290-292 were lying on α13-helix. Residue K271 and T275, I274 and N278 have side chain polar interaction. Analysis by PDBePISA (Krissinel & Henrick 2007) identified only 255, 259, 262, 265 as the solvent accessible residues, while rest were inaccessible to the solvent. Residue D290 of cluster 3 forms a part of the negatively charged group (in mammals) involved in interaction with N-terminal domain of NSF (Zhao et al. 2015). Residue D229, N231, D234 and S235 of cluster 5 were missing in the structure and were a part of the loop end or the turn position of TPR5. Residue L291 of cluster 6, also missing in structure, corresponds to L294 of α-SNAP, mutation of which (L294A) was responsible for the inactivity of α-SNAP to simulate NSF ATPase (Barnard et al. 1997). The penultimate C-terminal residue, L304 (L311 in human γ-SNAP) of γ-SNAP was also identified. The penultimate leucine and 89 C-terminal residues have been shown to be required for the interaction with NSF and Gaf-1/Rip11 (Tani et al. 2003).

The C-terminal residues of SNAP are also known to be important for NSF interaction and stimulation of NSF ATPase activity. Barnard et al. in 1997 performed deletion of 10 residues from the extreme C-terminal that resulted in marked decrease in the ability of the mutant to simulate the ATPase activity of NSF. However, all mutants were able to bind to NSF. They demonstrated that the NSF binding sites are present in both N- and C-terminal regions of α-SNAP but only extreme C-terminal region interaction leads to NSF ATPase activation. Many residues with known functional importance were identified in this region. Cryo-EM structure of the 20S complex, developed by Zhao et al. in 2015 showed two distinct N-domain binding sites on the surface of the C-terminal region of α-SNAP, mutations of which impaired the kinetics of SNARE complex disassembly. Thus the residues identified in the C-terminal region of SNAP protein might have functional importance and might be participating in the interaction of SNAP with NSF, which effects the ATPase activation.

Important residues in TPR region

Most of the residues identified in this region were located at the beginning or end of the helices of the TPRs. The residues at the turn position between the two helices of a single TPR and between two TPRs have more structural importance (D'Andrea & Regan, 2003). Between adjacent TPRs, residues have roles with both structural and functional implications (D'Andrea & Regan, 2003). Residues 48 and 50 were lying at the end of the first helix of TPR1 and 53 was at the beginning of the second helix of TPR1. Residues E73, D74 were lying at the turn position between TPR1 and TPR2, H109, R110, Q112, R114, R115 were lying at the turn end of TPR2, D151 and Q152 at the turn end of TPR3, L156 at the turn position between TPR 3 and TPR4, S194 at the turn end of TPR4.

Residue D74 has a side chain polar interaction with residues Q112 and R110. Identified residues R48, Q112, N118 were shown by previous mutation studies to have reduced binding to the SNARE complex (Marz et al. 2003). PDBePISA (Krissinel & Henrick 2007) analysis identified residues 48, 53, 74, 84, 85, 102, 118, 123, 147, 151, 153, 155, 156, 161, 194 as the solvent accessible residues. Of these, residues 48, 50, 53, 84, 85, 123, were lying on the concave surface of Sec17, which is assumed to be region of interaction with SNARE complex. As up to four molecules of α-SNAPs are assumed to be present in the minimal 20S disassembly super complex, the other solvent accessible residues on the convex surface of the protein might be involved in SNAP: SNAP interaction. The residues identified in the TPR region, thus, appeared to be structurally important residues and might also be involved in proper folding of the protein.

5.3 Application of MIp and average linkage clustering on SM protein identified networks of covarying residues.

As explained already, SM proteins are one of the essential factors of the vesicle fusion machinery that genetically and biochemically interact with the core SNARE fusion machinery, specifically with the Syntaxins of that particular membrane fusion step (Jahn & Fasshauer 2012). Generally, SM proteins seem to interact with their cognate Syntaxin via two different binding surfaces, the "closed" conformation in which the SNARE (H3) domain folds back onto the Habc domain, such that it is inaccessible for SNARE complex formation and the very N-terminal region of syntaxin, called the 'N-peptide'. Recent biochemical and structural studies on different SM/Syntaxin pairs have shown that SM proteins generally bind to their respective Syntaxin using both modes of interaction, however with different relative binding affinities. Previously my groups had collected the sequences of SM proteins along with several other proteins of the vesicle fusion repertoire and had also analyzed the evolutionary history for these proteins. Comparable patterns of duplications and diversifications between the closely interacting proteins were revealed by the previous phylogenetic analysis. With the aim to extract novel structural and functional features of SM proteins from the large sequence collection, I performed intra-protein covariation analysis, initially on the SM proteins. SM proteins were chosen as compared to other proteins of the vesicle fusion machinery, they had a manageable number of subtypes (only five).

Initially, intra-protein covariation analysis was performed on the Sly1 protein using MIp and selecting the pairs that scored above the Z-score of 4.0. This resulted in residues that were spread all over the structure of the protein with some lying close to each other and others lying distant to each other. They formed groups that appeared to form networks within the protein. To investigate this result in more detail, I developed an improved analysis pipeline, where average linkage clustering was carried out on the MIp dataset.

Before re-analyzing the SM protein with the improved analysis approach, I first applied it on some different test examples. These test examples included MAP (Gloor et al. 2005) and UDG (Liu & Bahar 2012) datasets from previous studies. I also tried the approach on a different type of protein of the vesicle trafficking factor SNAP, so as to appraise better the occurrence of chains of more distantly lying covarying residues. In all the test examples, I detected clusters of covarying residues that were restricted to a particular region of the protein. However, applying the analysis on Sly1 showed a different result. The residues within the clusters appeared to form a network hinting at the possible conformational changes and allosteric coupling within the SM proteins. The improved approach with clustering of MIp dataset thus helped to identify the networks of covarying residues occurring within the protein, which might be involved in the communication between the two interacting sites of Sly1 with Syntaxin.

When the analysis was applied on other SM protein subfamilies, a comparable pattern of highly scoring pairs within the 3D structure was observed. In all SM proteins, some residues were found to be lying close to each other, while many were found lying distant from each other in the tertiary structure. They appeared to form a network of residues connecting the two binding sites. Some of the residues were lying in the beginning of domain d1 at the region that is interacting with domain Hc of Syntaxin. Some residues were found to be located on the flexible helix-13 and helix-14 that have been shown to be important by various mutagenesis, biochemical and structural studies (Boyd et al. 2008; Hashizume et al. 2009; Bracher & Weissenhorn 2001; Misura & Weis 2000; Burkhardt et al. 2008; Baker et al. 2013). Some residues were found lying at the direct interacting region with Syntaxin, for example, at the loop between helix-13 and helix-14. Many residues were lying at the domain d2, which appeared to have structural importance in maintaining the arch shape of the protein. This suggests that similar regions of the different subfamilies co-vary.

Many residues with known functional importance were also identified. In Munc18-1, residue R39 was identified. R39 is a surface residue that forms salt bridge with E234 of H3 domain of Syntaxin1, was previously mutated and shown to reduce the Syntaxin1-binding capabilities of Munc18-1 (Jorgacevski et al. 2011). Mutation of this residue has also been shown to be a disease related point mutation in Munc18-2 (Hackmann et al. 2013). This mutation was also believed to partially lose binding to the closed conformation of Syntaxin-1 (Johnson et al. 2009). Another residue that caused a disease related point mutation in Munc18-2, T345 was identified. T345M mutation resulted in loss of hydrogen bond to residue 341 (Hackmann et al. 2013). Residue P242 of Munc18-1, which is known for interaction with Mint protein, was also identified. Its mutation P242S was shown to have increased the fusion pore dwell-time (Jorgacevski et al. 2011). Residue K46 of Munc18-1 was identified that specifically contacts residues D231 and R232 in the H3 helix of synaxin-1. Mutation of K46 to 46E was shown to reduce the binding to Syntaxin1 (Han et al. 2009). A double mutant of another identified residues M38 of Munc18-1 with D34 was shown to bind poorly to Syntain-1 (Han et al. 2009). Residues M330, L331, K332, K333, M334, Q336, Q338, K339 of Munc18-1 that lay on the loop between helix-13 and helix-14 which interacts with H3 domain of Syntaxin-1, were also identified.

Many residues with known functional importance were also identified in Vps33. Many residue having intermolecular contact with VPS16 as identified by Baker et al. in 2013, were also identified in this study. These include R85, R114, T116, L117, F118, and A212. The tip of domain d3 has been shown to be interacting with Qa

SNARE (Baker et al. 2013). Many residues were identified in this region of d3 domain, which includes 316, 321, 347, 358, 368, 370 and 371. Recently the SNARE motif of R-SNARE Nyv1 has also been shown to be binding to the antiparallel α-helices of the domain 3a helical hairpin (Baker et al. 2015). Residues G321 and S368, identified in the current study lie in the Nyv1 binding groove. Single residues substitution mutation of these residues has been shown to reduce the binding to Nyv1 (Baker et al. 2015). Some of the missense mutations that have been identified as causing phenotypic defects in Vps33 or other SM proteins have also been identified in this study. These include E82 (D88 in yeast Vps33), mutation of which have shown defects in content mixing and fusion and S525 (T553 in yeast Vps33) corresponding to T531 of yeast Sly1, mutation of which was a suppressor of lethality of Ypt1 (Lobingier & Merz 2012; Pieren et al. 2010; Y. Li et al. 2007).

In Sly1 subfamily, only one residue was identified in the N-peptide binding pocket. In Vps45, three residues were identified in the same region. While in Vps33, many residues were identified in the N-peptide binding region. Vps33 is an outlier in terms of position and orientation of domain 1 (Baker et al. 2013) and also does not bind to closed Qa SNAREs. In Munc18/Sec1 subfamily also many residues were identified in in the N-peptide binding region. A detailed look (Figure 5.1 and Table 4.1)at these residues showed that these residues did not covary in Unc18, Munc18 proteins, nor in the vertebrate duplications, Munc18-1, Munc18-2, Mucn18-3 proteins . They did not covary within the plants as well. However, these residues did covary in Sec-1 proteins from fungi, which do not bind to the N-peptide of its partner Syntaxin Sso1 as Sso1 lacks the N-peptide sequence equivalent to the one found in Syntaxin1. As mentioned in Baker et al. in 2013 and shown by Lobingier & Merz in 2012, the SM proteins can be divided into two groups; proteins that bind to the Qa SNARE N-peptide, which includes Munc18, Sly1 and Vps45 and the other group of proteins that do not bind to Qa SNARE N-peptide, which includes Vps33 and Sec1. The results from the current study also appear to show the difference in the proteins that bind to the Qa SNARE N-peptide and those that do not for the residues from the N-peptide binding region.

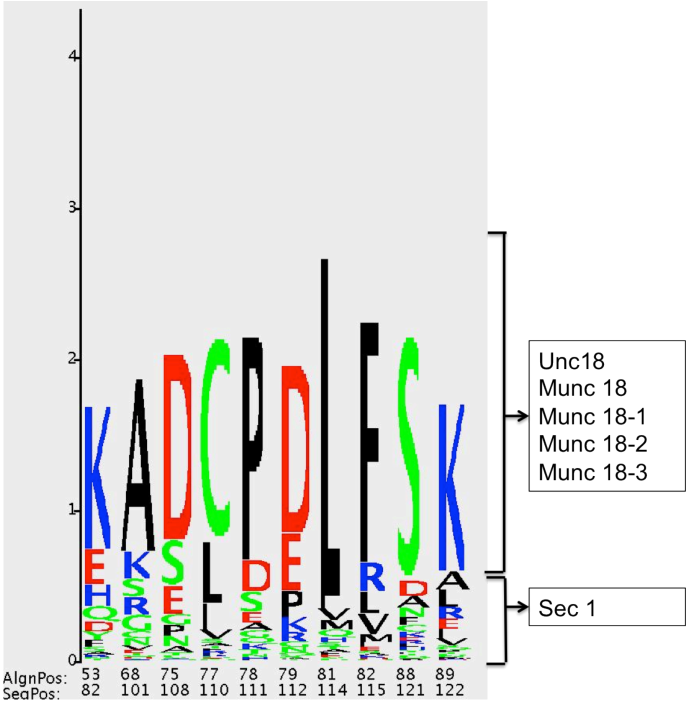


Figure 5.1: Sequence logo of residues from Munc18/Sec1 alignment lying near or within the N-peptide binding region.

The AlignPos are the position from the alignment of Munc18/Sec1 and the SeqPos are the positions from the Rat Munc18-1. These residues do not covary in Mucn18-1, Mucn18-2 and Mucn18-3 (the duplications in Vertebrates) but do so in Sec-1 proteins.

Table 5.2 : Distribution of covarying residues identified in the N-peptide binding region in different lineages

				Residue in the N-peptide binding region						
Residues	**82**	**101**	**108**	**110**	**111**	**112**	**114**	**115**	**121**	**122**
Metazoa	K	A	D/E	C	P	D/E	L	F	S	K
Viridiplantae	E	K	S	I/V/L	S	K/R	L	V/L/I	D	L
Chlorophyta	R/Q/N	R/K/S/T	S/N	L/A/P	S/Q/P/A	R/P	V/L	L	S/Q/N/C/A	V/L
Ascomycota	H	K/R	S/T	L	D/E	P	L/M/I.V	R	S/N	A/V/L/I
	Y/F	G	G	S/Q	P	S/T/Q/N	Q/S/T	S/Q	I/L/A	T/Q
	N	N/S	D	I/A/V/M	H	D/E	R	A/I/V/M	F	K
Basidiomycota	Q	S	D/E	L	D/E	D/E	L	F	S/N	D/E
Basal fungi	E/D	A	S/N	L	S/N	D	L	F	S/N	S
Protozoa	E/D	S	S/N	L/V/I	P/G	D/E	L/I/V/A	L/M/V	S/T/N/Q	L/V
	R	R/K		S/T	S/Q	G	H	F	C	K
	S			R/K	E/D					G

The number represents the residue position and the letters represent the amino acids. The residues were conserved in Metazoa and Viridiplantae but were covarying in Fungi.

5.4 Application of MIp and average linkage clustering for inter-protein analysis on Sly1 and Syntaxin 5 proteins identified complex network of functionally and structurally important inter- and intra-protein residues.

Combinatorial analysis pipeline of MIp and average linkage clustering on concatenated alignment of Sly1 and Syntaxin 5 resulted in clusters with inter and intra-protein residues from both the proteins as well as with cluster of only intra-protein residues from Sly1. The intra-protein clusters of Sly1 appeared to be localized to a certain region of the protein. The intra-protein residues 30, 31, 48, 49, 55, 57, 61 65, 68, 69, 74 and 78 of Sly1 were solvent accessible residues as per PDBePISA (Krissinel & Henrick 2007) analysis. They appeared to be involved in the interaction with Hc or SNARE domain of Sed5.

Examining residues from the clusters on the structure revealed that some of the identified residues in both Sly1 and Sed5 were lying in the interacting regions while some were found in regions away from the interaction sites. Some of the identified residues were lying close to each other while many were lying distant from each other. Many of the identified inter- & intra-protein residues in Sly1 were detected in the region corresponding to the closed conformation of Munc18a-Syntaxin1 interaction. This is line with the recent biochemical results shown for Sly1 that it interacts not only with N-peptide but also with the remainder of Sed5, making use of both the binding modes (Demircioglu et al. 2014). Some residues were detected on helix-13 and helix-14 and in the region between domain d3 and domain d2. Some of the detected residues were of known functional importance such as; Q155 in Sly1 was identified which is in known to stabilize the binding site by forming H-bonds. S541, L542, I543, S544 of Sly1 flank α20 and α21 helices in Sly1. α20 contains the Sly1-20 mutation that resulted in GTPase suppressor activity. Another identified residue K4 of Sed5 is known to have side chain contact with D158 of Sly1. K274 in Sly1 is known to form H-bond with S275. Most of the covarying residues were found to be lying in distant regions in the tertiary structure and thus, might be potentially involved in allosteric communication within the proteins.

Most of the detected inter- & intra-protein covarying residues in Sed5 were found in the domain Habc, SNARE motif and the N-terminal regions, which were known to be involved in interaction with SM proteins. Some residues from TM region were found to be covarying with residues from TM, SNARE and the Linker region. The residues from TM domain might play role in proper localization of Sed5 for its interaction with Sly1 and thus might be functionally important.

All of the detected inter and intra protein residues appeared to form a chain of interconnected residues. It is possible that these residues are participating in some sort of communication, possibly allosteric communication within the protein and between the two sites of interaction with Syntaxin. It has long been debated and questioned as to why the SM proteins have two spatially separated binding sites for interaction with Syntaxin. The reason for the occurrence of two different binding sites has yet not been completely understood. It has been speculated by some researchers in the field that probably the two binding sites are allosterically coupled to control the accessibility of the bound Syntaxin (Dawidowski & Cafiso 2013; Colbert et al. 2013; Demircioglu et al. 2014).

In my results, I identify inter- and intra- protein residues that were lying close as well as distant to each other. These residues appeared to be connected and form networks across the different regions of the protein. Burger & van Nimwegen in 2010 have also shown the existence of chains of statistically dependent residues. They explained that indirect interactions occur through the chains of directly interacting residue pairs that run through the protein and thus connect the distal pairs. Some other studies (e.g: (Lockless & Ranganathan 1999; Süel et al. 2002; Halabi et al. 2009; Baussand & Carbone 2009)) have also shown the existences of such thermodynamically coupled residues. These residues create physically connected networks linking the distant positions in the tertiary structure of protein and are thus involved in allosteric communication. This seems to be the case in the results of the current study as well.

The current findings suggest a possible communication between the two binding sites involving a network of covarying amino acids along the structure of the SM and Syntaxin complex. The residues identified in the current study appeared to form chain of coupled residues, however it remains a speculation if they are allosterically coupled or indicate a large conformational change. It is possible that the interaction with the N-peptide region of Syntaxin acts as an anchor point to push the SM proteins towards the membrane fusion site and establish the right conformation and location of Syntaxin to assemble into the complex. At this stage, the SM-Syntaxin complex can undergo conformational changes, which can also be induced by some docking factors. Then the SM protein can release the SNARE motif of Syntaxin making it available for the SNARE-complex formation. The residues identified by the current approach lie across the structure of protein, appearing to connect the two binding sites. Even though some of the individual pairs lie distant from each other, when the complete ensemble of identified residues is considered they appear to form a network of dependent residues thus establishing a connection between the two spatially distinct binding sites. However, these possibilities need to be confirmed by mutational and biochemical studies to be carried out in future.

6 Conclusions & Perspective

Proteins of the SNARE (Soluble N-ethylmaleimide-sensitive factor Attachment protein REceptors) and SM (Sec1/Munc18) families are essential components of the vesicle fusion machinery. Different sets of SNARE proteins drive the fusion of transport vesicles with an acceptor compartment by zippering into a tight four-helix bundle complex. In each trafficking step, SM proteins interact with the particular SNARE set, specifically with the Qa-SNARE (also called Syntaxins). SM proteins are thus believed to control and regulate the fusion process (Jahn & Fasshauer 2012) . Various other conserved factors such as Rab proteins, NSF (N-ethylmaleimide-Sensitive Factor), SNAP proteins (Soluble NSF Attachments Proteins), and tethering proteins belonging to the CATCHR (Complex Associated with Tethering containing Helical Rods) family are also part of the vesicle fusion machinery and are thought to facilitate the vesicle docking and fusion process. Although several structures of key factors of the vesicle fusion machinery have been elucidated during the last decades, the order of events and their exact interaction is not fully understood yet. In particular, it is unclear whether a particular molecular feature represents a special adaptation of the protein machinery or whether it is a shared trait. Previously, my groups had developed a database management system to collect and analyze the sequences of various protein families of the vesicle fusion machinery. The analysis of the evolutionary history of several protein families, e.g. SNARE proteins (Kloepper et al. 2007; Kloepper et al. 2008; Kienle et al. 2009) Rab proteins (Klöpper et al. 2012) (Klöpper et. al. 2012), and SM proteins had shown that these factors arose by duplication and diversification of a prototypic fusion machinery during evolution. Moreover, the analysis of SNAREs and SM proteins revealed some comparable patterns of duplication and diversification between these closely interacting proteins. This previous work lay the foundation for exploring the evolutionary changes within the vesicle fusion machinery across different eukaryotic lineages. Here, I explored their covariation pattern using multiple sequence alignments in order to extract novel insights into structural and functional features.

Initially, I performed intra-protein covariation analysis on the SM protein family, which consists of five distinct subtypes Sec1/Munc18, Sly1, Vps33, Vps45, and Scfd2.

It has been shown that they interact in general via two spatially separate binding sites with their cognate syntaxin binding partner. So far, it is not understood how this binding mode enables SM proteins to control the accessibility of the bound Syntaxin. Some recent studies indicate that the two binding sites might be allosterically coupled (Dawidowski & Cafiso 2013; Shu-Hong Hu et al. 2011; Colbert et al. 2013; Demircioglu et al. 2014).

When I applied the co-variation method, MIp on the SM protein Sly1, I found a significant portion of highly scoring co-variant residues to be scattered across the 3D-structure of the protein, while it had been reported in earlier studies that such residues were often in close vicinity and located in close proximity to the enzyme's catalytic site. I also realized that many of the co-variant residues in Sly1 scored highly with more than one other residue as if they were forming a co-variation network. Similar findings had been reported before (e.g : (Gloor et al. 2005; Liu & Bahar 2012)), but usually these covarying networks were confined to a certain area of the inspected protein. To further explore this initial observation, I developed an automated analysis pipeline that allowed me to perform, among other things, average linkage clustering on the MIp dataset. This analysis pipeline was first tested on test cases, such as MAP (Gloor et al. 2005) and UDG (Liu & Bahar 2012). This corroborated and extended earlier observations on their co-variation pattern. Using the improved approach, I investigated next another vesicular trafficking protein, the SNAP protein family. The SNAP family was chosen, because they are structurally and functionally very different from SM proteins. For SNAP proteins, I found several groups of co-varying residues that are restricted to a particular region. Most of the residues within the clusters were lying close to each other, while some more distantly lying residues appeared to be connected to each other by some near-by residues. Several residues with known structural and functional importance were identified using MIp

When I applied the improved analysis pipeline on the SM protein Sly1, I found that the networks of co-varying residues were not restricted to a particular region as observed for the other examples, but were much more spread, possibly connecting different regions of the protein. This suggests that certain regions that are not in direct contact within the 3D structure are somehow coupled in Sly1. Whether this occurs via a conformational change or "allosteric coupling" remains speculative, however. Exploring each cluster further revealed that many known structurally and functionally important residues were identified within these clusters. Similar pattern of covarying residues was observed for all the subfamilies of the SM proteins, suggesting that my analysis uncovered a common feature within this essential protein family. Possibly, this indicates that SM proteins share a common mode of interaction with Qa SNAREs involving two interaction surfaces. Further clade specific analysis of these residues revealed that the covariation pattern for the paired sites was changing in different branches of the tree. They were changing within the clades as well as between the clades. The detected residues thus might be co-evolving in different eukaryotic lineages.

To further explore the interaction of SM proteins and Syntaxin, I performed an inter-protein analysis for a concatenated alignment of Sly1 and its partner Syntaxin 5. Sly1 and Syntaxin 5 were chosen out of all other SM-Syntaxin pairs as they occur as singletons in most species. Most of the co-varying residues were already detected when the individual proteins were investigated, but some residues were found to be co-varying between both the proteins. Most of the inter-protein covarying residues were not in direct contact with each other in the tertiary structure. They appeared to form a chain of interconnected residues along with the intra-protein covarying residues. Further clade specific analysis of the high scoring inter-protein residues revealed that the covariation pattern for the paired sites was changing at the same corresponding branches in the two proteins during their evolution. This suggested that the detected inter-protein covarying sites might be involved in maintaining the interaction and the evolution of the two proteins.

Together, my results suggest a possible communication between the two binding sites involving a network of co-varying residues along the structure of SM/Syntaxin complexes. However, it is difficult to arrive at a conclusion regarding the importance and allosteric nature of the identified covarying sites unless mutational and biochemical evidences are being carried out. Thus the next steps would be to selectively mutate the detected covarying sites, best by starting with the highest scoring pairs in order to establish whether the covarying sites are indeed functionally important. When designing such experiments one also needs to take into account the co-variation pattern of the larger networks detected. The mutations would need to be tested in the interaction studies using purified proteins, but also using in-vivo approaches. This could provide interesting insights for one of the long-standing questions in the field regarding the controversial function and interaction of the SM proteins with Syntaxins. Similar, studies can be carried out for the detected covarying residues of SNAP proteins to see if they affect the function and interaction of the protein with either the membrane, SNARE complexes or NSF. For example, the residues identified in the N-terminal region of the SNAP protein might be important for the formation or orientation of the first α-helix and the adjacent membrane attachment loop.

Very probably, the study identified novel structural and functional information for SM and SNAP proteins, which can be used for functional studies. Interestingly, co-variation analyses of these two structurally very different factors of the vesicle fusion machinery yielded in very different co-variation patterns. It should thus be explored whether co-variation analysis of even other protein types leads to even other co-variation patterns. As a first step, the novel analysis pipeline should be used for the sequences of other conserved factors of the vesicle fusion machinery that have been collected already in the laboratory.

A Appendix: Application note

Alifea– Alignment Feature analysis tool

An integrated tool for alignment feature extraction and visualization was developed which was called *alifea*. It is a standalone application written in JavaScript. The GUI is developed using the Netbeans platform. This tool can be used to predict structurally and functionally important residues in a protein, which can be putative candidates for mutational studies. The tool would hopefully facilitate and ease the sequence analysis. The tool can be made available on request as a standalone application.

The basic input for *alifea* is a MSA. The input alignment, if required, can be processed to remove sequences and sites containing too many gaps by specifying gap threshold or using the default parameters provided. The alignment can be visualized with the help of integrated Jalview alignment viewer (Waterhouse et al. 2009). Jalview functionality can be used for editing and visualizing the MSAs.

The tool allows the user to extract information such as fully conserved positions, conserved positions within a subtype, and coevolving positions from sequence alignments. Sequence Logo method has been implemented to visualize and identify the fully conserved positions. To detect subtype specific positions, Hannenhalli & Russell's in 2000 methods is implemented. This method identifies these positions, by comparing the distribution of intra and inter-group residue entropy for every possible branch in the phylogenetic tree of the protein family. The methods implemented for the coevolutionary analysis include McLachlan based Substitution Correlation, Statistical Coupling Analysis (SCA) (Lockless & Ranganathan 1999), Positional Conservation based SCA (SCAnew) (Halabi et al. 2009), Mutual Information (MI) (Martin et al. 2005), Mutual Information Corrected (MIp) (Dunn et al. 2008), Explicit Likelihood of Subset Variation (ELSC) (Dekker et al. 2004), Observed Minus Expected Squared (OMES) (Larson et al. 2000). Details of implementations of the algorithms are explained further in this section.

Another major advantage of this tool is the possibility to combine the results of the coevolutionary analysis with a network analysis to be able to highlight the important residues within a protein family. The results of the co-evolutionary analysis can be interpreted as a network with the positions as nodes and the co-evolutionary scores as edges. The tool includes the visualization of this network. The user can modify the network by interactively choosing the cut off for the coevolutionary results based on Z-score. The nodes and edges can be moved around, renamed, selected and searched.

Clustering methods can be utilized to extract positions with high scores. The tool contains average, single, and complete linkage hierarchical clustering methods to analyze the network. The clusters can be ranked based on the average coevolutionary score among them to identify the important residues. The clusters can be visualized as hierarchy or tree and as heat maps.

The results can be mapped directly on the 3D structure, if available, using the integrated Jmol (Jmol 2001) viewer. There is possibility to add more than one structure for different proteins in the MSA. The user can choose to map these positions on the structure along with conserved (conservations degree as chosen by the user) position. The program can also automatically mark these positions on the MSA and show them as a sequence logo.

This the first time that different methods for detecting conserved, subtype specific and coevolving residues have been integrated along with easy to comprehend visualizations. Previously, these operations were done manually or by using different softwares, but there was no single package to integrate them. Thus, this tool can be useful and practical for studies involving extraction of information from MSA.

The details of the methods and algorithms implemented in the tool are given below.

A.1 Alignment Preprocessing

The user has a choice to filter columns or sequences containing too many gaps by specifying the gap threshold or using the default parameters provided. The user also has a choice to treat the gaps as noise or signals during the covariation analysis. If gaps are treated as noise, the gapped positions will not be considered during any calculations. If gaps are treated as signal, the gaps are counted as the 21st amino acid during all calculations. A conservation filter is also implemented in the tool. This allows users to choose different values of conservation of columns. Any column having the chosen conservation will not be considered during the covariation analysis.

A.2 Sequence Analysis

Sequence Logo for conserved positions

The implementation of sequence logo for protein sequences is based on Schneidery & Stephens from 1990. It is a presentation of alignments, where characters representing the sequences are stacked over each other for each position in the alignment. It gives information about relative frequency of residues at each position, conserved residue at each position and conserved sites. The most conserved positions can be easily spotted from this representation; however, even variation at each site and distribution of amino acids at those sites can be inferred. The importance of a particular position is given by the information measured in bits. This information is calculated using the Shannon entropy as

$$H(l) = \sum_{b=a}^{t} f(b,l) \log_2 f(b,l)$$

, where H(l) is the uncertainty at the position *l, b* is one of the 20 amino acids, and $f(b,l)$ is the frequency of amino acid *b* at position *l*. The total information at the position is given by the decrease in the uncertainty

$$R_{sequence}(l) = log_2(20) - (H(l) + e(n))$$

in bits per position, where $R_{sequence}(l)$, is the amount of information present at position *l*, 20 is the maximum uncertainty for amino acids at any given position, *n* is the number of sequences in the alignment and $e(n)$ is the error correction factor for small *n* and is approximated as in Schneider et al. in 1986,

$$e(n) = (s - 1) / (2 * ln2 * n)$$

The set of $R_{sequence}(l)$ for the full alignment forms a curve representing the importance of each position. The height of each amino acids at each position gives the frequency of that amino acids at that position and is given as:

$$height\ of\ amino\ acid\ at\ position\ l = f(b,l) * R_{sequence}(l)$$

A.3 Subtype Specific analysis

The subtype specific analysis method implemented in the tool is the one given by Hannenhalli & Russell in 2000. The relative entropy of a position i for subtype s with respect to the entropy of that position for the subtype $\bar{s}$ (union of all the subtype excluding s), is calculated as:

$$RE_i^s = \sum_{for\ all\ x} P_{i,x}^s \log \frac{P_{i,x}^s}{P_{i,x}^{\bar{s}}}$$

where $P_{i,x}^s$ is the profile value for amino acid x at position i of subtype s, and $P_{i,x}^{\bar{s}}$ is the profile value for amino acid x at position i of subtype $\bar{s}$. $RE \geq 0$ and is exactly 0 when two distributions are identical. The profiles were build by the hmmbuild program of HMMER (Eddy 1999) in the original implementation. Cumulative relative entropies are then calculated to estimate the role of a position i in determining the subtypes as:

$$CRE_i = \sum_{for\ all\ sub-types\ s} RE_i^s$$

The cumulative relative entropy is converted to Z-score :

$$Z_i = \frac{CRE_i - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation.

In *alifea,* the HMMER package is not used. The probability profile score is generated directly using the implemented Dirichlet priors such that $\sum_{for\ all\ x} P_{i,x}^s = 1$, for each alignment position i. There are two choices of the mixture, Sjolander_1996 and Sjolander_2011.

A.4 Covariation Analysis

McLachlan based Substitution Correlation

The implementation of this method is based on Gobel et al. in 1994. An NxN matrix is generated, for each column i in the alignment, where N is the number of sequences in the alignment. The values in the matrix are filled from a substitution matrix that assigns a high score if there is an conserved substitution for a pair of residues in sequence k, l at column i and a low score if there is non-conserved substitution (s_{ikl}). Following Olmea et al. in 1999, McLachlan substitution matrix was used (available at GenomeNet database (MCLA710101) (McLachlan1971). The correlation score between two columns i and j is calculated as:

$$r_{ij} = \frac{1}{N^2} \sum_{kl} \frac{w_{kl} \left(s_{ikl} - \langle s_i \rangle\right) \left(s_{jkl} - \langle s_j \rangle\right)}{\sigma_i \, \sigma_j}$$

where, $\langle s_i \rangle$ is the mean of s_{ikl} and σi is the standard deviation of s_{ikl}.

The tool provides computation of Pearson correlation from the raw values as well as by Spearman correlation by rank values.

Statistical Coupling Analysis (SCA)

In this method, statistical coupling between the two sites of the MSA is calculated by computing the energy changes on perturbation. The perturbation is caused by making sub-alignments with a certain residue. The statistical coupling energy is calculated as:

$$\Delta\Delta G_{i,j}^{stat} = kT^* \sqrt{\sum_x \left(\ln \frac{P_{i|\delta j}^x}{P_{MSA|\delta j}^x} - \ln \frac{P_i^x}{P_{MSA}^x} \right)^2}$$

where, kT^* is an arbitrary energy unit, P_i^x is binomial probability of an amino acid x at site i in the alignment and $P_{i|\delta j}^x$ is the binomial probability of an amino acid x at site i in the sub-alignment, representing the perturbation of amino acid frequencies at site j.

The SCA implementation in the *alifea* tool, considers the most conserved residue in each column as the perturbation for creating the sub alignments.

The implementation is based SCA version 1.5 (Lockless & Ranganathan 1999). The value of kT constant, the non-normalization of P^x_{MSA}, and the acceptance criteria for sequence alignments are taken from the Gerstein coevolution tool (as described in their SCA note). The original SCA algorithm has been reworked (de Juan et al. 2013) and changed to SCAnew described below.

SCAnew / Positional Conservation based SCA

The implementation of Positional Conservation based SCA is based on the most recent version of SCA (Halabi et al. 2009). The correlations between the positions are weighted according to their conservation. The SCA correlation matrix is reduced by spectral decomposition. The principal components obtained represent the groups of coevolving positions. These coevolving groups are referred to as “protein sectors” that are nearly statistically independent. They are physically connected in the tertiary structure and have distinct functional role and constitute an independent mode of divergence in the protein family.

In alifea tool, only the calculation of SCA correlation matrix is implemented (based on the supplementary material of Halabi et al. in 2009 and the web supplement of SCA at http://systems.swmed.edu/rr_lab/Note109_files/Note109_v3.html). The structurally distinct protein sectors are detected by using clustering from network theory on the coevolutionary matrix.

Mutual Information (MI)

Mutual information (MI) is a method based on Shannon’s entropy that indicates the dependencies of the two columns. It is a measure of reduction of uncertainty. The MI between two columns of a MSA reflects the degree to which the knowledge of the amino acid at one position helps to predict the identity of the amino acids at the other position. A high MI values indicates correlation between the two positions. The implementation of MI is based on Martin et al. in 2005.

MI between two positions of MSA is calculated by:

$$MI(x,y) = H(x) + H(y) - H(x,y)$$

where, $H(x)$ and $H(y)$ are entropy of columns x and y, calculated as,

$$H(x) = -\sum_{i=1}^{K} p(x_i) \log_b p(x_i)$$

where $p(x_i)$ is probability of amino acid i in column x, $k=20$ (for 20 amino acids). The logarithm base b=20. The value of $H(x)$varies from 0, in case of complete conservation to 1, when all 20 amino acids are equally distributed. $H(x,y)$Is the joint entropy, which is defined as:

$$H(x,y) = -\sum_{i=1}^{k}\sum_{j=1}^{l} p(x_i,y_j)\log_b p(x_i,y_j)$$

where $p(x_i,y_j)$is joint probability of amino acid i in column x and amino acid j in column y, k=l=20 for amino acids, and b is logarithm base, here set to 20. The joint entropy can range from 0 to 2.

MI score ranges between 1 and 0 with high MI value reflecting a higher interdependence between the two positions of a MSA.

The initial formulations of MI were affected by high variability positions in MSAs and by the effect of phylogenetic background (de Juan et al. 2013) and thus many subsequent version of this approach were developed. One of them is described next.

Mutual Information Corrected (MIp)

The mutual information approach was corrected to suppress the phylogenetic bias by normalizing the observed covariance of a pair of column by the background covariance of the columns. The background covariance is the average covariance score of the column with all the other columns (Dunn et al. 2008). Thus,

$$MIp(a,b) = MI(a.b) - APC(a,b)$$

where *APC(a,b)* is the average product correction of the background scores, calculated as:

$$APC(a,b) = \frac{MI(a,\bar{x})MI(b,\bar{x})}{\overline{MI}}$$

where $MI(a,\bar{x})$ is mean mutual information of column a, defined as:

$$MI(a,\bar{x}) = \frac{1}{m}\sum MI(a,x)$$

where *n* is the number of columns in MSA, and *m=n-1*, and summation is over *x=1* to *n, x≠a*, and $\overline{MI}$ denotes the overall mean mutual information,

$$\overline{MI} = \frac{2}{mn}\sum MI(x,y)$$

where indices run from *x=1* to *m, y=x+1* to *n*.

This correction on MI, provided a substantial improvement compared to other previously published methods for predicting covarying positions (Dunn et al. 2008; de Juan et al. 2013).

Explicit Likelihood of Subset Variation (ELSC)

Like SCA, ELSC is a perturbation-based algorithm. It compares the characteristics of a full MSA with that of a subset of the MSA (Dekker et al. 2004). The implementation follows Dekker et al. in 2004. In this method, a subset of the MSA is created by constraining the identity of the amino acid in a column *i* of the MSA. Then, $N_{r,j}$ and $n_{r,j}$ is calculated. $N_{r,j}$ is the number of residues of type *r* at position *j* in the full MSA and $n_{r,j}$ is the number of residue type *r* at position *j* in the constrained subset.

The probability of drawing a random subset of size n_{total} from the MSA gives the observed amino acid composition at position *j* in the constrained subset. This is given by

$$L_j^{\langle i\rangle} = \frac{\prod_r \binom{N_{r,j}}{n_{r,j}}}{\binom{N_{total}}{n_{total}}}$$

N_{total} is the number of sequence in full MSA and n_{total} is the number of sequences in the subset MSA. The $L_j^{\langle i\rangle}$ is normalized as

$$L_{j,max}^{\langle i\rangle} = \frac{\prod_r \binom{N_{r,j}}{m_{r,j}}}{\binom{N_{total}}{m_{total}}}$$

For the normalization, an ideal representative subset $m_{r,j}$ is constructed such that

$$m_{r,j} \approx \left(N_{r,j}/N_{total}\right).n_{total}$$

The covariance score between the two positions is calculated by computing the normalized ratio,

$$^{<i>}_{j}\Lambda = \frac{^{<i>}_{j}L}{^{<i>}_{j,max}L} = \prod_{r} \frac{\binom{N_{r,j}}{n_{r,j}}}{\binom{N_{r,j}}{m_{r,j}}}$$

Observed Minus Expected Squared (OMES)

This method is based on chi-square statistics. The correlation between the positions is calculated by comparing the expected probability of the two residues occurring together in any one sequence to the frequency with which they actually do appear together (Larson et al. 2000). The correlation between residues *r* and *s*, appearing at positions *i* and *j*, is defined as:

$$\chi^2 = \sum_{\substack{i=r,\bar{r} \\ j=s,\bar{s}}} \frac{\left(p_{(i,j)} - p_{(i)}p_{(j)}\right)^2}{p_{(i)}p_{(j)}}$$

where $p_i(r)$ is the frequency of amino acid *r* occurring at position i and

$p_i(\bar{r}) = 1 - p_i(r),$ the total frequency of all other residues, and $p_{i,j}(r,s)$, is the frequency of *r* at position *i* and *s* at position *j* in the same sequence.

A.5 Network Analysis Algorithm

Hierarchical clustering methods arrange data into a hierarchy based on the distance or similarity. While using covariation data as a network, the covariation score is used as the distance between the two nodes or the residues. The higher the covariation score, the shorter is the distance between the residues and vice-versa. *Alifea* provides three hierarchical clustering methods.

Single Linkage Clustering

In single linkage clustering, the two clusters are fused together if the minimum distance (maximum covariation score) between the members of the two clusters is more than the average covariation score.

Complete Linkage Clustering

In complete linkage clustering the two clusters are fused together if the maximum distance (minimum covariation score) between the members of the two clusters is more than the average covariation score.

Average Linkage Clustering

In average linkage clustering the two clusters are merged if the average distance (or the average covariation score) between the members of the two clusters is more than the total average covariation score. A stopping criterion was used for the average linkage clustering such that only residue pairs that have the covariation score above the stopping criteria were considered for clustering. This stopping criterion can be varied in different statistical steps of the distribution of the edge weights (such as 1-σ, 2- σ, 3- σ and 4- σ from the mean-μ of the weights.

A.6 Scoring and ranking the clusters

Clusters can be scored either based on the average weighted degree of the cluster or based on the average weight of the cluster and then ranked in the decreasing order of the score. The details are described in the Method section.

A.7 Visualizations

Different visualizations including alignment/sequence view, heat map view of covariation data, network view of the clusters identified on the covariation data, hierarchy/tree view for visualizing the phylogenetic tree as well as the generated clusters and a structure view to map the results directly on a 3D structure, if available, were also implemented in the tool. The detail of all the different visualizations is provided in the Method section.

References

Ackerman, S.H. & Gatti, D.L., 2011. The contribution of coevolving residues to the stability of KDO8P synthase. *PloS one*, 6(3), p.e17459.

Ackerman, S.H., Tillier, E.R. & Gatti, D.L., 2012. Accurate Simulation and Detection of Coevolution Signals in Multiple Sequence Alignments A. Tramontano, ed. *PloS one*, 7(10), p.e47108.

Antonin, W. et al., 2002. Crystal structure of the endosomal SNARE complex reveals common structural principles of all SNAREs. *Nature Structural Biology*, 9(2), pp.107–111.

Aran, V. et al., 2009. Characterization of two distinct binding modes between syntaxin 4 and Munc18c. *The Biochemical journal*, 419(3), p.655.

Baker, R.W. et al., 2015. A direct role for the Sec1/Munc18-family protein Vps33 as a template for SNARE assembly. *Science (New York, NY)*, pp.1–4.

Baker, R.W., Jeffrey, P.D. & Hughson, F.M., 2013. Crystal Structures of the Sec1/Munc18 (SM) Protein Vps33, Alone and Bound to the Homotypic Fusion and Vacuolar Protein Sorting (HOPS) Subunit Vps16 H. Wanjin, ed. *PloS one*, 8(6), p.e67409.

Barnard, R.J.O., Morgan, A. & Burgoyne, R.D., 1997. Stimulation of NSF ATPase Activity by alpha-SNAP is Required for SNARE Complex Disassembly and Exocytosis. *The Journal of Cell Biology*, pp.1–9.

Baussand, J. & Carbone, A., 2009. A Combinatorial Approach to Detect Coevolved Amino Acid Networks in Protein Families of Variable Divergence R. Dunbrack, ed. *PLoS computational biology*, 5(9), p.e1000488.

Bitto, E. et al., 2007. Structure and dynamics of γ-SNAP: Insight into flexibility of proteins from the SNAP family. *Proteins*, 70(1), pp.93–104.

Bonifacino, J.S. & Glick, B.S., 2004. The Mechanisms of Vesicle Budding and Fusion. *Cell*, pp.1–14.

Boyd, A. et al., 2008. A Random Mutagenesis Approach to Isolate Dominant-Negative Yeast sec1 Mutants Reveals a Functional Role for Domain 3a in Yeast and Mammalian Sec1/Munc18 Proteins. *Genetics*, 180(1), pp.165–178.

Bracher, A. & Weissenhorn, W., 2001. Crystal structures of neuronal squid Sec1 implicate inter-domain hinge movement in the release of t-SNAREs. *Journal of molecular biology*, 306(1), pp.7–13.

Bracher, A. & Weissenhorn, W., 2002. Structural basis for the Golgi membrane recruitmentof Sly1p by Sed5p. *The EMBO Journal*, pp.1–11.

Brenner, S., 1974. The Genetics of Cenorhabditis elegans. *Genetics*, pp.1–24.

Burger, L. & van Nimwegen, E., 2010. Disentangling Direct from Indirect Co-Evolution of Residues in Protein Alignments P. E. Bourne, ed. *PLoS computational biology*, 6(1), p.e1000633.

Burkhardt, P. et al., 2008. Munc18a controls SNARE assembly through its interaction with the syntaxin N-peptide. *The EMBO Journal*, 27(7), pp.923–933.

Burkhardt, P. et al., 2011. Primordial neurosecretory apparatus identified in the choanoflagellate Monosiga brevicollis. *Proceedings of the National Academy of Sciences of the United States of America*, 108(37), pp.15264–15269.

Buslje, C.M. et al., 2009. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics (Oxford, England)*, 25(9), pp.1125–1131.

Cai, H., Reinisch, K. & Ferro-Novick, S., 2007. Coats, Tethers, Rabs, and SNAREs Work Together to Mediate the Intracellular Destination of a Transport Vesicle. *Developmental Cell*, 12(5), pp.671–682.

Caporaso, J.G. et al., 2008. Detecting coevolution without phylogenetic trees? Tree-ignorant metrics of coevolution perform as well as tree-aware metrics. *BMC evolutionary biology*, 8(1), p.327.

Carpp, L.N. et al., 2006. The Sec1p/Munc18 Protein Vps45p Binds Its Cognate SNARE Proteins via Two Distinct Modes. *The Journal of Cell Biology*, 173(6), pp.927–936.

Carr, C.M. et al., 1999. Sec1p Binds to SNARE Complexes and Concentrates at Sites of Secretion. *The Journal of Cell Biology*, pp.1–12.

Casari, G., Sander, C. & Valencia, A., 2004. A method to predict functional residues in proteins. pp.1–8.

Chae, T.H. et al., 2004. The hyh mutation uncovers roles for αSnap in apical protein localization and control of neural cell fate. *Nature Genetics*, 36(3), pp.264–270.

Chakrabarti, S. & Panchenko, A.R., 2010. Structural and functional roles of coevolved sites in proteins. *PloS one*, 5(1), p.e8591.

Chen, Y.A. & Scheller, R.H., 2001. Snare-Mediated Membrane Fusion. *Nature*, pp.1–9.

Clary, D.O., Griff, I.C. & Rothman, J.E., 1990. SNAPs, a Family of NSF Attachment Proteins Involved in Intracellular Membrane Fusionin Animals and Yeast. pp.1–13.

Colbert, K.N. et al., 2013. Syntaxin1a variants lacking an N-peptide or bearing the LE mutation bind to Munc18a in a closed conformation. *Proceedings of the National Academy of Sciences*, 110(31), pp.12637–12642.

Coppola, T. et al., 2002. Pancreatic *β*-Cell Protein Granuphilin Binds Rab3 and Munc-18 and Controls Exocytosis. pp.1–10.

Cortajarena, A.L. & Regan, L., 2006. Ligand binding by TPR domains. *Protein Science*, 15(5), pp.1193–1198.

D'Andrea, L. & Regan, L., 2003. TPR proteins: the versatile helix. *Trends in biochemical sciences*, 28(12), pp.655–662.

Dawidowski, D. & Cafiso, D.S., 2013. Allosteric Control of Syntaxin 1a by Munc18-1: Characterization of the Open and Closed Conformations of Syntaxin. *Biophysj*, 104(7), pp.1585–1594.

de Juan, D., Pazos, F. & Alfonsoo, V., 2013. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4), pp.249–261.

DeBello, 1995. DeBello_Augustine_Nature95. pp.1–5.

Dekker, J.P. et al., 2004. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics (Oxford, England)*, 20(10), pp.1565–1572.

del Sol Mesa, A., Pazos, F. & Alfonsoo, V., 2003. Automatic Methods for Predicting Functionally Important Residues. *Journal of molecular biology*, 326(4), pp.1289–1302.

Demircioglu, F.E., Burkhardt, P. & Fasshauer, D., 2014. The SM protein Sly1 accelerates assembly of the ER-Golgi SNARE complex. *Proceedings of the National Academy of Sciences*.

Dib Linda, C.A., 2012a. CLAG: an unsupervised non hierarchical clustering algorithm handling biological data. pp.1–14.

Dib Linda, C.A., 2012b. Protein Fragments: Functional and Structural Roles of Their Coevolution Networks. pp.1–21.

Dib, L., 2015. Coev-web: a web platform designed tosimulate and evaluate coevolving positionsalong a phylogenetic tree. *BMC bioinformatics*, pp.1–7.

Dickson, R.J. et al., 2010. Identifying and seeing beyond multiple sequence alignment errors using intra-molecular protein covariation. *PloS one*, 5(6), p.e11082.

Dulubova, I., 2002. How Tlg2p/syntaxin 16 "snares" Vps45. *The EMBO Journal*, 21(14), pp.3620–3631.

Dulubova, I. et al., 1999. A conformational switch in syntaxin during exocytosis: role of munc18. *EMBO Journal*, pp.1–11.

Dunn, S.D., Wahl, L.M. & Gloor, G.B., 2008. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics (Oxford, England)*, 24(3), pp.333–340.

Dutheil, J. & Galtier, N., 2007. Detecting groups of co-evolving positions in a molecule: a clustering approach. *BMC evolutionary biology*, 7(1), p.242.

Dutheil, J.Y., 2011. Detecting coevolving positions in a molecule: why and how to account for phylogeny. *Briefings in Bioinformatics*, 13(2), pp.228–243.

Eddy, S.R., 1999. Profile Hidden Markov Models. *Bioinformatics (Oxford, England)*, pp.1–9.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5), pp.1792–1797.

Fares, M.A.M. & Travers, S.A.A.S., 2006. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics*, 173(1), pp.9–23. Available at: http://www.genetics.org/cgi/doi/10.1534/genetics.105.053249.

Faruck Morcosa et al., 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *PNAS*, pp.1–9.

Fasshauer, D. et al., 1998. Conserved structural features of the synaptic fusion complex: SNARE proteins reclassified as Q- and R-SNAREs. *Proceedings of the National Academy of Sciences of the United States of America*, 95(26), pp.15781–15786.

Fitch, W.M. & Markowitz, E., 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics*, 4(5), pp.579–593.

Fodor, A.A. & Aldrich, R.W., 2004. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins*, 56(2), pp.211–221.

Furgasona, M.L.M. et al., 2009. The N-terminal peptide of the syntaxin Tlg2p modulates binding of its closed conformation to Vps45p. *Proceedings of the National Academy of Sciences*, 106(34), pp.14303–14308.

Gloor, G.B. et al., 2010. Functionally Compensating Coevolving Positions Are Neither Homoplasic Nor Conserved in Clades. *Molecular biology and evolution*, 27(5), pp.1181–1191.

Gloor, G.B. et al., 2005. Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions †. *Biochemistry*, 44(19), pp.7156–7165.

Gobel, U. et al., 1994. Correlated Mutations and Residue Contacts in Proteins. pp.1–9.

Gouveia-Oliveira, R. & Pedersen, A.G., 2007. Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Algorithms for Molecular Biology*, 2(1), p.12.

Gouveia-Oliveira, R. et al., 2009. InterMap3D: predicting and visualizing co-evolving protein residues. *Bioinformatics (Oxford, England)*, 25(15), pp.1963–1965.

Graham, M.E. et al., 2008. A gain-of-function mutant of Munc18-1 stimulates secretory granule recruitment and exocytosis and reveals a direct interaction of Munc18-1 with Rab3. *The Biochemical journal*, 409(2), p.407.

Graham, S.C. et al., 2013. Structural basis of Vps33A recruitment to the human HOPS complex by Vps16. *Proceedings of the National Academy of Sciences*.

Griff, I.C. et al., 1992. The Yeast SEC17 Gene Product Is Functionally Equivalent to Mammalian alpa-SNAP Protein*. *The Journal of biological chemistry*, pp.1–10.

Hackmann, Y., Graham, S.C. & Ehl, S., 2013. Syntaxin binding mechanism and disease-causing mutations in Munc18-2. In PNAS.

Halabi, N. et al., 2009. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell*, 138(4), pp.774–786.

Halachmi, N. & Lev, Z., 1996. The Sec1 Family: A Novel Family of Proteins Involved in Synaptic Transmission and General Secretion. pp.1–9.

Han, L. et al., 2009. Rescue of Munc18-1 and -2 Double Knockdown Reveals the Essential Functions of Interaction between Munc18 and Closed Syntaxin in PC12 Cells. pp.1–14.

Hannenhalli, S.S. & Russell, R.B., 2000. Analysis and prediction of functional subtypes from protein sequence alignments. *Journal of molecular biology*, 303(1), pp.61–76.

Hanson, P.I. & Whiteheart, S.W., 2005. AAA+ proteins: have engine, will work. *Nature Reviews Molecular Cell Biology*, 6(7), pp.519–529.

Hanson, P.I. et al., 1997. Structure and Conformational Changes in NSF and Its Membrane Receptor Complexes Visualized by Quick-Freeze/Deep-Etch Electron Microscopy. pp.1–13.

Hanson, P.I. et al., 1995. The N-Ethylmaleimeide-sensitive Fusion Protein and alpha-SNAP Induce a Conformational Change in Syntaxin. *The Journal of biological chemistry*, 270, pp.16955–16961.

Harrison, S.D. et al., 1994. Mutations in the Drosophila Rop Gene Suggest a Function in General Secretion and Synaptic Transmission. *Neuron*, pp.1–12.

Hashizume, K. et al., 2009. Yeast Sec1p Functions before and after Vesicle Docking. pp.1–13.

Hayashi, T. et al., 1995. Disassembly of the reconstituted synaptic vesicle membrane fusion complex in vitro. *The EMBO Journal*, 14(10), pp.2317–2325.

Hohl, T.M. et al., 1998. Arrangement of Subunits in 20 S Particles Consisting of NSF, SNAPs, and SNARE Complexes. pp.1–10.

Hong, W. & Lev, S., 2013. Tethering the assembly of SNARE complexes. *Trends in Cell Biology*, pp.1–9.

Hopf, T.A. et al., 2015. Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. *Nature Communications*, 6, p.6077.

Hopf, T.A. et al., 2014. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*, 3.

Hopf, T.A. et al., 2012. Three-Dimensional Structures of Membrane Proteinsfrom Genomic Sequencing. *Cell*, 149(7), pp.1607–1621.

Hu, S.-H. et al., 2007. Structure of the Munc18c/Syntaxin4 N-peptide complex defines universal features of the N-peptide binding mode of Sec1/Munc18 proteins. *PNAS*, pp.1–6.

Jahn, R. & Fasshauer, D., 2012. Molecular machines governing exocytosis of synaptic vesicles. *Nature*, 490(7419), pp.201–207.

Jahn, R. & Scheller, R.H., 2006. SNAREs — engines for membrane fusion. *Nature Publishing Group*, 7(9), pp.631–643.

Jahn, R. & Südhof, T.C., 1999. MEMBRANE FUSIONAND EXOCYTOSIS. *Annu. Rev. Biochem.*, pp.1–49.

Johnson, J.R. et al., 2009. Binding of UNC-18 to the N-terminus of syntaxin is essential for neurotransmission in Caenorhabditis elegans. *The Biochemical journal*, 418(1), p.73.

Jones, D.T. et al., 2012. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics (Oxford, England)*, 28(2), pp.184–190.

Jorgacevski, J. et al., 2011. Munc18-1 Tuning of Vesicle Merger and Fusion Pore Properties. *Journal of Neuroscience*, 31(24), pp.9055–9066.

Karpenahalli, M.R., Lupas, A.N. & Söding, J., 2007. TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences. *BMC bioinformatics*, 8(1), p.2.

Khvotchev, M. et al., 2007. Dual Modes of Munc18-1/SNARE Interactions Are Coupled by Functionally Critical Binding to Syntaxin-1 N Terminus. *Journal of Neuroscience*, 27(45), pp.12147–12155.

Kienle, N., 2010. *Phylogenetic studies of the vesicular fusion machinery*,

Kienle, N., Kloepper, T.H. & Fasshauer, D., 2009. Phylogeny of the SNARE vesicle fusion machinery yields insights into the conservation of the secretory pathway in fungi. *BMC evolutionary biology*, 9(1), p.19.

Kim, K. et al., 2012. Munc18b is an essential gene in mice whose expression is limiting for secretion by airway epithelial and mast cells. *The Biochemical journal*, 446(3), pp.383–394.

Kloepper, T.H., Kienle, C.N. & Fasshauer, D., 2007. An elaborate classification of SNARE proteins sheds light on the conservation of the eukaryotic endomembrane system. *Molecular biology of the cell*, 18(9), pp.3463–3471.

Kloepper, T.H., Kienle, C.N. & Fasshauer, D., 2008. SNAREing the basis of multicellularity: Consequences of protein family expansion during evolution. *Molecular biology and evolution*, 25(9), pp.2055–2068. Available at: http://mbe.oxfordjournals.org/cgi/doi/10.1093/molbev/msn151.

Klöpper, T.H. et al., 2012. Untangling the evolution of Rab G proteins: implications of a comprehensive genomic analysis. *BMC Biology*, 10(1), p.71.

Korber, B.T.M. et al., 1993. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. *Proceedings of the National Academy of Sciences of the United States of America*, pp.1–5.

Kosodo, Y., 2002. Binding of Sly1 to Sed5 enhances formation of the yeast early Golgi SNARE complex. *Journal of Cell Science*, 115(18), pp.3683–3691.

Krogh, A. et al., 1994. Hidden Markov Models. *Journal of molecular biology*, pp.1–32.

Larson, S.M., Di Nardo, A.A. & Davidson, A.R., 2000. Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *Journal of molecular biology*, 303(3), pp.433–446.

Lee, Y. et al., 2012. A Coevolutionary Residue Network at the Site of a Functionally Important Conformational Change in a Phosphohexomutase Enzyme Family A. M. Buckle, ed. *PloS one*, 7(6), p.e38114.

Li, W. et al., 2011. The Crystal Structure of a Munc13 C-terminal Module Exhibits a Remarkable Similarityto Vesicle Tethering Factors. *Structure/Folding and Design*, 19(10), pp.1443–1455.

Li, Y. et al., 2007. Mutations of the SM protein Sly1 resulting in bypass of GTPase requirement in vesicular transport are confined to a short helical region. *FEBS letters*, 581(29), pp.5698–5702.

Linda Dib, Daniele Silvestro & Salamin, N., 2014. Evolutionary footprint of coevolving positions in genes. *Bioinformatics (Oxford, England)*, pp.1–9.

Liu, Y. & Bahar, I., 2012. Sequence Evolution Correlates with Structural Dynamics. *Molecular biology and evolution*, 29(9), pp.2253–2263.

Livingstone, C.D. & Barton, G.J., 1993. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. pp.1–12.

Lobingier, B.T. & Merz, A.J., 2012. Sec1/Munc18 protein Vps33 binds to SNARE domains and the quaternary SNARE complex. *Molecular biology of the cell*, 23(23), pp.4611–4622.

Lockless, S.W. & Ranganathan, R., 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science (New York, NY)*, 286(5438), pp.295–299.

Ma, C. et al., 2011. Munc13 mediates the transition from the closed syntaxin–Munc18 complex to the SNARE complex. *Nature Publishing Group*, 18(5), pp.542–549. Available at: http://www.nature.com/doifinder/10.1038/nsmb.2047.

Ma, C. et al., 2013. Reconstitution of the Vital Functions of Munc18 and Munc13 in Neurotransmitter Release. *Science (New York, NY)*, 339(6118), pp.421–425.

Marino Buslje, C. et al., 2010. Networks of High Mutual Information Define the Structural Proximity of Catalytic Sites: Implications for Catalytic Residue Identification B. Rost, ed. *PLoS computational biology*, 6(11), p.e1000978.

Marks, D.S. et al., 2011. Protein 3D Structure Computed from Evolutionary Sequence Variation A. Sali, ed. *PloS one*, 6(12), p.e28766.

Marks, D.S., Hopf, T.A. & chriss, S., 2012. Protein structure prediction from sequence variation. *Nature Publishing Group*, 30(11), pp.1072–1080.

Martin, L.C. et al., 2005. Using information theory to search for co-evolving residues in proteins. *Bioinformatics (Oxford, England)*, 21(22), pp.4116–4124.

Marz, K.E., Lauer, J.M. & Hanson, P.I., 2003. Defining the SNARE Complex Binding Surface of -SNAP: IMPLICATIONS FOR SNARE COMPLEX DISASSEMBLY. *The Journal of biological chemistry*, 278(29), pp.27000–27008.

Mast, F.D. et al., 2014. Evolutionary mechanisms forestablishing eukaryotic cellularcomplexity. *Trends in Cell Biology*, 24(7), pp.435–442.

Merlo, L.M.F., Lunzer, M. & Dean, A.M., 2007. An empirical test of the concomitantly variable codon hypothesis. *PNAS*, pp.1–6.

Mezulis, S. et al., 2015. The Phyre2 web portal for protein modeling,prediction and analysis. *Nature Protocols*, 10(6), pp.845–858.

Misura, K.M.S. & Weis, R.H.S.W.I., 2000. Three-dimensional structure of the neuronal-Sec1-syntaxin 1a complex. *Nature*, pp.1–8.

Motlagh, H.N. et al., 2014. The ensemble nature of allostery. *Nature*, 508(7496), pp.331–339.

Nguyen, L.T. et al., 2014. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular biology and evolution*, 32(1), pp.268–274.

Novick, P. & Schekman, R., 1979. Secretion and cell-surface growth are blocked in a temperature sensitive mutant of *Saccharomyces cerevisae*. *Proceedings of the National Academy of Sciences of the United States of America*, pp.1–5.

Okamoto, M. & dhof, T.C.S., 1997. Mints, Munc18-interacting Proteins in Synaptic Vesicle Exocytosis*. pp.1–6.

Olivier Lichtarge, Bourne, H.R. & Cohen, F.E., 1996. An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families. *Journal of molecular biology*, pp.1–17.

Olmea, O., Rost, B. & Valencia, A., 1999. Effective use of sequence correlation and conservation in fold recognition. *Journal of molecular biology*, 293(5), pp.1221–1239.

Ossig, R. et al., 1991. The Yeast *SLY* Gene Products, Suppressors of Defects in the Essential GTP-Binding Ypt1 Protein, May Act in Endoplasmic Reticulum-to-Golgi Transport. pp.1–14.

Pazos, F. & Bang, J.-W., 2006. Computational Prediction of Functionally Important Regions in Proteins. *Current Bioinformatics*, pp.1–9.

Peng, R., 2002. Sly1 protein bound to Golgi syntaxin Sed5p allows assembly and contributes to specificity of SNARE fusion complexes. *The Journal of Cell Biology*, 157(4), pp.645–655.

Peng, R. & Gallwitz, D., 2004. Multiple SNARE interactions of an SM protein: Sed5p/Sly1p binding is dispensable for transport. *The EMBO Journal*, pp.1–11.

Pevsner, J. et al., 1994. Specificity and Regulationof a Synaptic Vesicle Docking Complex. *Neuron*, pp.1–9.

Pieren, M., Schmidt, A. & Mayer, A., 2010. The SM protein Vps33 and the t-SNARE H. *Nature Publishing Group*, 17(6), pp.710–717.

Pollock, D.D., Taylor, W.R. & Goldman, N., 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. *Journal of molecular biology*, 287(1), pp.187–198.

Poon, A., 2005. The Rate of Compensatory Mutation in the DNA Bacteriophage X174. *Genetics*, 170(3), pp.989–999.

Rice, L.M. & Brunger, A.T., 1999. Crystal Structure of the Vesicular Transport Protein Sec17: Implications for SNAP Function in SNARE Complex Disassembly. *Journal of Molecular Cell Biology*, pp.1–11.

Rizo, J. & Rosenmund, C., 2008. Synaptic vesicle fusion. pp.1–20.

Rizo, J. & Xu, J., 2015. The Synaptic Vesicle Release Machinery. *Annual Review of Biophysics*, 44(1), pp.339–367.

Schmidt, H.A. et al., 2002. TREE-PUZZLE: maximum likelihood phylognetic analysis using quartets and parallel computing. *Bioinformatics (Oxford, England)*, pp.1–3.

Schneider, T.D. et al., 1986. The Information Content of Binding Site on Nucleotide Sequences. *Academic Press Inc. (London)*, pp.1–34.

Schneidery, T.D. & Stephens, R.M., 1990. Sequence Logo: A New Way to Display Consensus Sequences. *Nucleic acids research*, pp.1–12.

Scott, B.L., 2004. Sec1p directly stimulates SNARE-mediated membrane fusion in vitro. *The Journal of Cell Biology*, 167(1), pp.75–85.

Shimodaria, H. & Hasegawa, M., 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics (Oxford, England)*, pp.1–2.

Shu-Hong Hu et al., 2011. Possible roles for Munc18-1 domain 3aand Syntaxin1 N-peptide and C-terminalanchor in SNARE complex formation. *PNAS*, pp.1–6.

Sjolander, K., 1998. Phylogenetic inference in protein superfamilies: Analysis of SH2 domains. pp.1–10.

Sollner, T. et al., 1993. SNAP receptors implicated in vesicle trageting and fusion. *Nature*, pp.1–7.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)*, 22(21), pp.2688–2690.

Stenmark, H., 2009. Rab GTPases as coordinators of vesicle traffic. *Nature Publishing Group*, 10(8), pp.513–525.

Stephane Guindon et al., 2009. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. pp.1–37.

Studer, R.A. & Robinson-Rechavi, M., 2010. Large-Scale Analysis of Orthologs and Paralogs under Covarion-Like and Constant-but-Different Models of Amino Acid Evolution. *Molecular biology and evolution*, 27(11), pp.2618–2627.

Sudhof, T.C. & Rothman, J.E., 2009. Membrane Fusion: Grappling withSNARE and SM Proteins. *Science (New York, NY)*, pp.1–4.

Sutton, R.B., Fasshauer, D. & Brunger, R.J.A.T., 1998. Crystal structure of a SNARE complex involved in ° synaptic exocytosis at 2.4 A resolution. pp.1–7.

Süel, G.M. et al., 2002. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology*, 10(1), pp.59–69.

Talavera, D., Lovell, S.C. & Whelan, S., 2015. Covariation Is a Poor Measure of Molecular Coevolution. *Molecular biology and evolution*.

Tani, K. et al., 2003. Mapping of Functional Domains of gamma-SNAP. *The Journal of biological chemistry*, 278(15), pp.13531–13538.

Teppa, E., Zea, D.J. & Marino Buslje, C., 2014. Identification of Coevolving Amino Acids using Mutual Information. pp.1–24.

Tillier, E.R.M. & Lui, T.W.H., 2003. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics (Oxford, England)*, 19(6), pp.750–755.

Togneri, J. et al., 2006. Specific SNARE complex binding mode of the Sec1Munc-18 protein, Sec1p. *PNAS*, pp.1–6.

Toonen, R.F.G. & Verhage, M., 2003. Vesicle trafficking: pleasure and pain from SM genes. *Trends in Cell Biology*, 13(4), pp.177–186.

Travers, S.A.A.S. & Fares, M.A.M., 2007. Functional coevolutionary networks of the Hsp70-Hop-Hsp90 system revealed through computational analyses. *Molecular biology and evolution*, 24(4), pp.1032–1044. Available at: http://mbe.oxfordjournals.org/cgi/doi/10.1093/molbev/msm022.

Tuffery, P. & Darlu, P., 2000. Exploring a Phylogenetic Approach for the Detection of Correlated Substitutions in Proteins. *Molecular biology and evolution*, pp.1–7.

Verhage, M. et al., 1997. DOC2 Proteins in Rat Brain: Complementary Distribution and Proposed Function as Vesicular Adapter Proteins in Early Stages of Secretion. pp.1–9.

Verhage, M. et al., 2000. Synaptic Assembly of the Brain in the Absence of Neurotransmitter Secretion. *Science (New York, NY)*, pp.1–7.

Waterhouse, A.M. et al., 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics (Oxford, England)*, 25(9), pp.1189–1191.

Weimbs, T. et al., 1997. A conserved domain is present in different families of vesicular fusion proteins: A new superfamily. *Proceedings of the National Academy of Sciences of the United States of America*, pp.1–6.

Wickner, W. & Schekman, R., 2008. Membrane fusion. *Nat Struct Mol Biol*, pp.1–17.

Winter, U., Chen, X. & Fasshauer, D., 2009. A Conserved Membrane Attachment Site in -SNAP Facilitates N-Ethylmaleimide-sensitive Factor (NSF)-driven SNARE Complex Disassembly. *The Journal of biological chemistry*, 284(46), pp.31817–31826.

Wollenberg, K.R. & Atchley, W.R., 2000. Separation of phylogenetic and functionalassociations in biological sequences byusing the parametric bootstrap. pp.1–4.

Yamaguchi, T. et al., 2002. Sly1 Binds to Golgi and ER Syntaxins via a Conserved N-Terminal Peptide Motif. *Developmental Cell*, 2(3), pp.295–305.

Yanofsky, C., Horn, V. & Thorpe, D., 1964. Protein Structure Relationships Revelaed by Mutational Analysis. *Science (New York, NY)*, pp.1–2.

Zhao, M. et al., 2015. Mechanistic insights into the recyclingmachine of the SNARE complex. *Nature*, 518(7537), pp.61–67.

Zwilling, D. et al., 2007. Early endosomal SNAREs form a structurally conserved SNARE complex and fuse liposomes with multiple topologies. *EMBO Journal*, pp.1–10.

Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/

'Protein interfaces, surfaces and assemblies' service PISA at the European Bioinformatics Institute. (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html)

Contributions

Sequence Alignments and Tracey

Initial alignments before refinement and preprocessing for SM proteins, Sly1, Vps33 and Vps45 were provided by Tobias H. Kloepper, while for SM protein, Munc18/Sec1, SNAP proteins and Syntaxin 5 were provided by Nickias Kienle.

Phylogenetic Trees

The actual calculations of all phylogenetic trees were conducted by Nickias Kienle.

Ms. Nicee Srivastava

Ph.D. student, UNIL

Chemin du Devin, 31 B
1012 Lausanne
Nicee.Srivastava@unil.ch
nicee.srivastava@gmail.com
Phone: +41-786167997
DOB: 4th May, 1985

EDUCATION

2007-2009: Masters in Bioinformatics
Banaras Hindu University (BHU), India. CGPA: 8.72/10 (Rank: 1)

2005-2007: Bachelors in Biology
I.T.College, Lucknow University, Lucknow.
Subjects: Botany, Zoology, Chemistry.

EXPERIENCE

03/2011- present: Understanding molecular evolution of vesicular trafficking proteins by using multiple sequence information

Ph.D. Studies, Supervisor: Dr. Dirk Fasshauer, UNIL, Switzerland.

Identifying structurally and functionally important sites in vesicular trafficking protein and development of a comprehensive protein sequence data analysis tool in Java.

11/2009-05/2010: Project Assistant
IMTECH, Chandigarh, India.
Project: Protein engineering and design involving molecular modeling and dynamics studies
-Molecular dynamics and simulation studies using amber tool and molecular modeling studies using modeller tool.

0507/2008: Summer Research Training
Dr.Lalitha Guruprasad's lab, School of Chemistry, University of Hyderabad, under UGC-Networking Resource Center
Project : Comparative analysis and modeling of PfkB-Kinase proteins in *Homo sapiens*

2007-2009: Masters in Bioinformatics
Banaras Hindu University (BHU), India. CGPA: 8.72/10 (Rank: 1)

Master Project: A survey of computational tools for coevolutionary analysis and *in silico* study of intramolecular coevolution in two tandem domains of EngA and SRP-SR GTP binding proteins

AWARDS AND CONFERENCES

Presented poster at Structural Dynamics in Cellular Communication, VIB Conference Series on February 9 and 10, 2015 in Brussels.
Presented posters at 1st and 2nd DNF Symposiums on 2014 & 2015.
Attended the ICTS Workshop and Conference on Evolutionary Origins of Compartmentalized Cells at Bangalore, India, on February 20 - March 2, 2012.
Received prestigious BHU gold Medal for securing first rank in Masters of Bioinformatics 2009.
Awarded INSPIRE Fellowship by DST, Govt. of India for 2010-2011
Qualified CSIR-NET Lectureship, Dec 2008.

TEACHING

Taught a practical course on introduction to bioinformatics tools to undergraduate students each year (2011-2015).

RELEVANT COURSES

- An Introduction to R
- Whole genome sequencing workshop
- Markov Process
- Introduction to network analysis
- Scientific writing and publishing
- Probability theory and network inference