Analysis of Motif Distributions in Regions of Endocytic Proteins

Chante Bethell[1] , Zuzana Kadlecova PhD[2], and Stephanie J. Spielman, PhD[3]
[1]Bioinformatics Program, Rowan University
[2] Cambridge Institute for Medical Research, Cambridge University, UK
[3]Department of Biological Sciences, Rowan University

Abstract

A short linear motif (SLiM) is a recurring pattern of approximately three to ten amino acids found in proteins. SLiMs are important for cellular signaling and the regulating of proteins, often times by acting as binding sites for protein-binding domains. While SLiMs exist both in ordered regions of proteins where there is a tertiary structure and in disordered regions where there is no structure, they are primarily functional in disordered regions. An important example of SLiM-mediated processes and the focus of this study is endocytosis. Endocytosis is the process by which cells engulf molecules from the extracellular environment. There are specific motifs that mediate and trigger endocytosis. However, the short length of motifs means that it is easy to overlook those that may be important to biological functions. The goal of this study is to identify previously unrecognized proteins that may be involved in endocytosis by analyzing the distribution of motifs in the ordered and disordered regions of the human proteome. Using a bioinformatics approach, we systematically searched the entire human proteome for motifs known to be involved in endocytosis. We hypothesize that the proteins we find to be enriched with motifs in disordered regions may be functionally important for endocytosis. These proteins will be targeted for experimental validation.

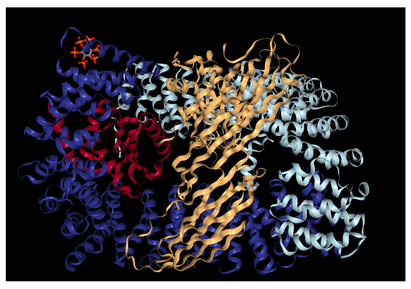


Figure 1: The AP2 Protein Adaptor Complex, colored by its subunits: AP2A, AP2B1, AP2S1, AP2M1.[1]

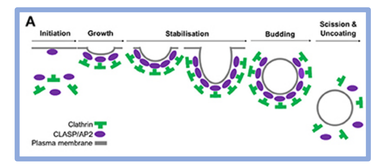
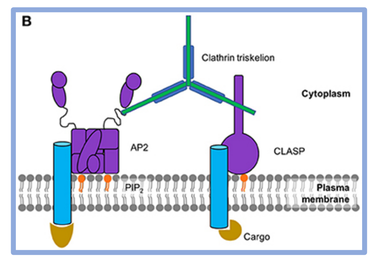


Figure 2A: Steps of the clathrin-coated pit formation.[2]
Figure 2B: The major players in the initiation of CME. The cargo binds to its respective transmembrane receptor, which AP2 then binds to in order to create a bridge to clathrin.[2]



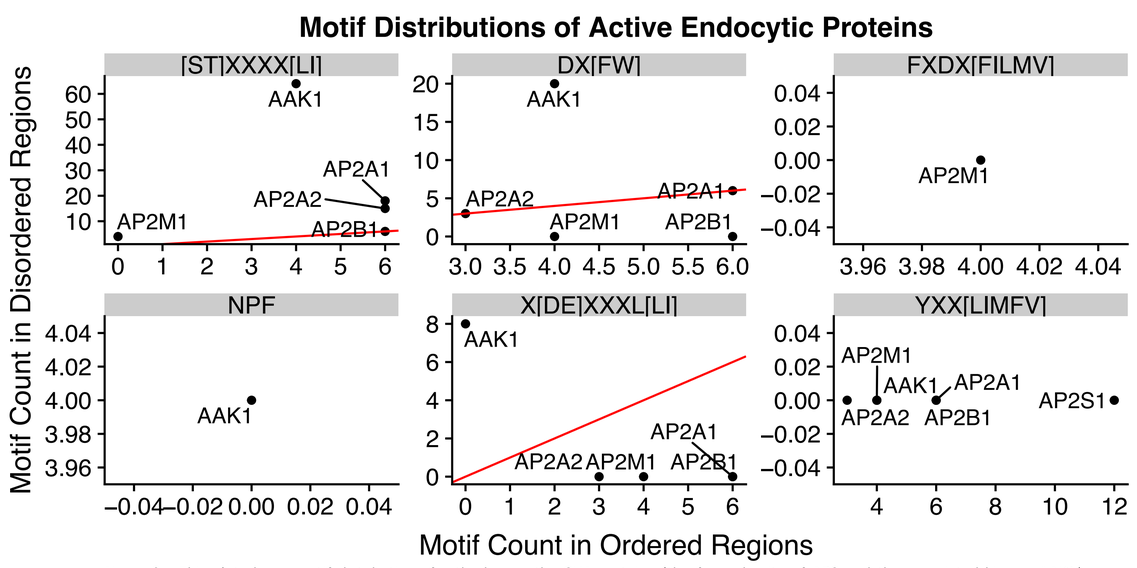
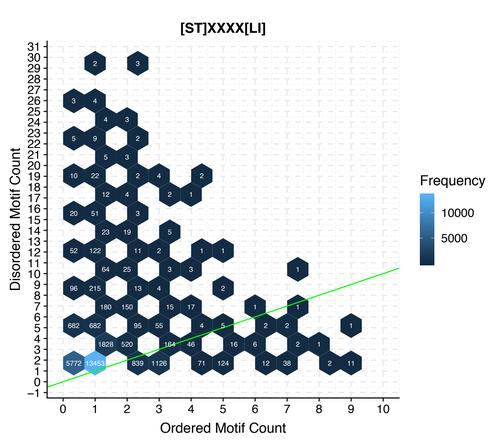
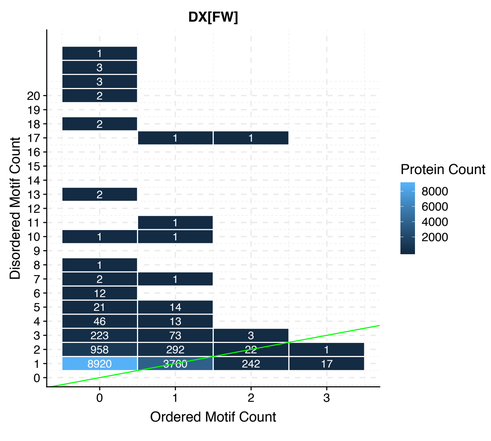
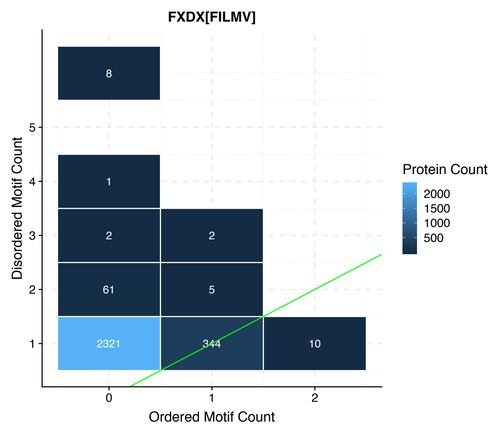


Figure 3: This plot shows motif distributions for the known the CME proteins (the five subunits of AP2 and the associated kinase, AAK1).

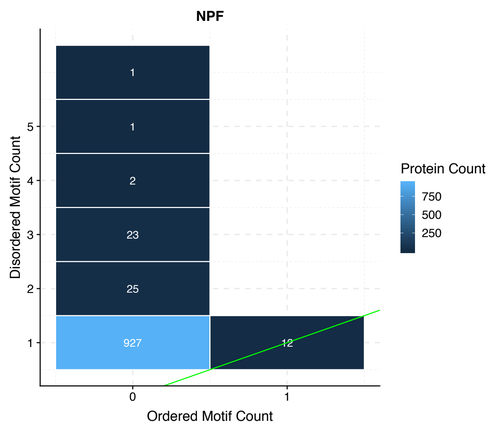
Distribution of select CME motifs across the human proteome

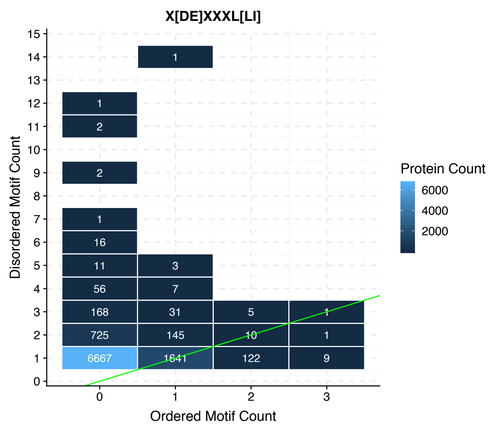


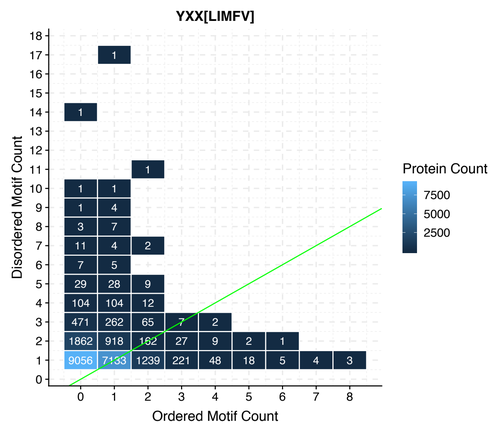


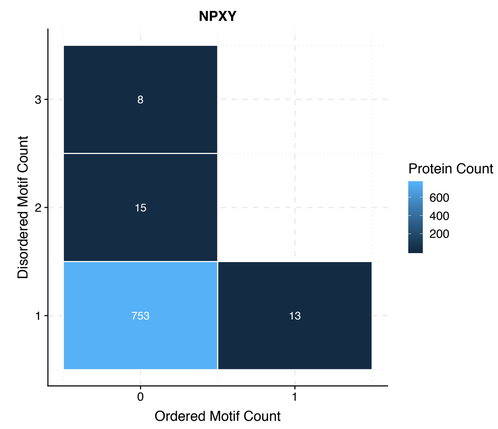


These plots represent heatmaps of proteins organized by their distribution in the disordered and ordered regions for the several endocytic motifs. Each rectangle (or hexagon) is colored based on the number of proteins that have the respective number of motif instances in the ordered and disordered regions. The rectangles found above the green *y=x* line on each plot represent the proteins that are enriched for the motif in the disordered regions.









A subset of identified candidate CME proteins

Protein name[4]	UNIPROT accession ID	Disordered Region Count	Ordered Region Count	Motif	Select Gene Ontology (GO) terms[4]
Insulin receptor substrate 2	**Q9Y4H2**	**17**	**1**	**YXX[LIMFV]**	**Insulin receptor binding, protein kinase binding, brain development, lipid homeostasis, signal transduction**
Insulin receptor substrate 1	**P35568**	**14**	**0**	**YXX[LIMFV]**	**Insulin receptor binding, transmembrane receptor protein tyrosine kinase adaptor activity, glucose homeostasis**
Transport protein Sec24B	O95487	11	2	YXX[LIMFV]	Coat protein complex II (COPII) vesicle cargo loading, COPII vesicle coating, intracellular protein transport
Protein PRRC2A	P48634	10	0	YXX[LIMFV]	RNA binding
Mucin-16	**Q8WXI7**	**10**	**1**	**YXX[LIMFV]**	**Cell adhesion, stimulatory C-type lectin receptor signaling pathway, transmembrane domain**
Trichohyalin	Q07283	14	1	X[DE]XXXL[LI]	Calcium ion binding, transition metal ion binding, intermediate filament organization
Uncharacterized protein FLJ40521	Q8N7P7	12	0	X[DE]XXXL[LI]	-
Golgin subfamily A member 6-like protein 4	A6NEF3	11	0	X[DE]XXXL[LI]	Cellular component of the Golgi apparatus
Coiled-coil domain-containing protein 136	**Q96JN2**	**11**	**0**	**X[DE]XXXL[LI]**	**Acrosome assembly, single fertilization, spermatogenesis, transmembrane domain**
Mucin-17	E7EPM4	9	0	X[DE]XXXL[LI]	Cellular homeostasis, O-glycan processing, stimulatory C-type lectin receptor signaling pathway
Receptor protein-tyrosine kinase	**E9PFD7**	**3**	**0**	**NPXY**	**ATP binding, transmembrane receptor protein tyrosine kinase signaling pathway**
Transmembrane channel-like protein	**F5GYU8**	**3**	**0**	**NPXY**	**Integral component of plasma membrane**
Calreticulin	K7EJB9	3	0	NPXY	Calcium ion binding, unfolded protein binding, protein folding
Calmegin	**O14967**	**3**	**0**	**NPXY**	**Calcium ion binding, protein folding chaperone, unfolded protein binding, binding of sperm to zona pellucida, transmembrane domain**
Receptor protein-tyrosine kinase	**B4DTR1**	**2**	**0**	**NPXY**	**ATP binding, transmembrane receptor protein tyrosine kinase activity**
Mucin-19	Q7Z5P9	119	1	[ST]XXXX[LI]	O-glycan processing, stimulatory C-type lectin receptor signaling pathway
Mucin-17	E7EPM4	114	1	[ST]XXXX[LI]	Cellular homeostasis, O-glycan processing, stimulatory C-type lectin receptor signaling pathway
Mucin-4	A0A0G2JR46	72	0	[ST]XXXX[LI]	Epithelial structure maintenance, regulation of signaling receptor activity
Kruppel-like factor 18 protein	A0A0U1RQI7	51	1	[ST]XXXX[LI]	Regulation of transcription by RNA polymerase II, nucleic acid binding
Adenomatous polyposis coli protein	P25054	42	7	[ST]XXXX[LI]	Protein kinase regulator activity, cell adhesion, cell migration, cell cycle arrest

Conclusions

- The distribution of motifs in the disordered versus ordered regions are not always enriched in endocytic proteins.
- We identify dozens of proteins that are enriched for CME motifs in their disordered regions. These can be targeted for further experimental testing for involvement in the CME pathway.
- Many of these identified proteins are not experimentally characterized with functional motifs according to the Eukaryotic Linear Motif database (http://elm.eu.org/).

Bibliography

1. RCSB Protein Data Bank. (n.d.). 4UQI. Retrieved April 10, 2019, from https://www.rcsb.org/3d-view/4UQI/1
2. Smith SM, Baker M, Halebian M and Smith CJ (2017) Weak Molecular Interactions in Clathrin-Mediated Endocytosis. Front. Mol. Biosci. 4:72. doi: 10.3389/fmolb.2017.00072
3. Tidyverse. (n.d.). Retrieved from https://www.tidyverse.org/
4. UniProt: A worldwide hub of protein knowledge. (2018). *Nucleic Acids Research, 47*(D1). doi:10.1093/nar/gky1049

All code and data is available from https://github.com/cbethell/motifs