De novo Discovery of Short Linear Motifs

Haiyan Hu

University of Central Florida

Outline

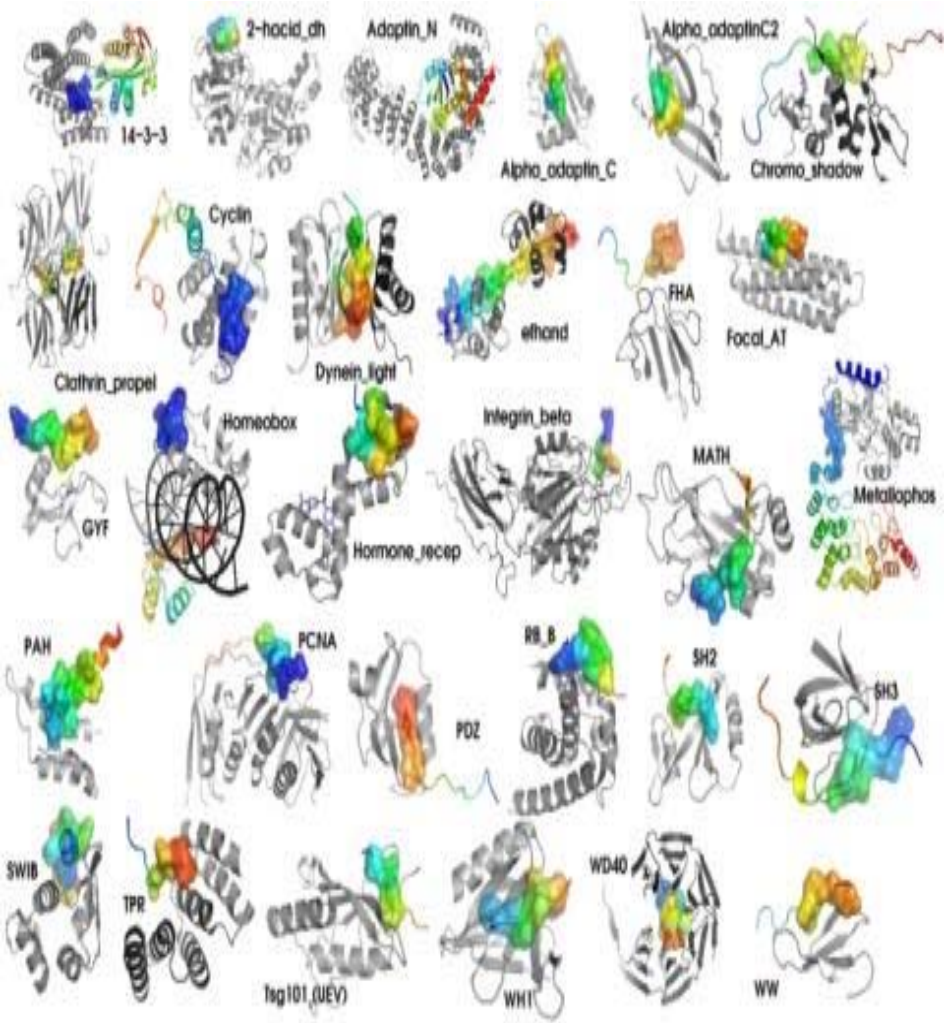
- What are short linear motifs (SLiMs)?
- The importance of identifying SLiMs.
- Current methods for SLiM discovery
- A new approach, DoubleFlex
- Results and discussion
- Acknowledgement

What are SLiMs?

- 3-11 amino acid long peptide patterns that mediate protein activities.
- Often occurs in protein disorder regions
- High abundant in proteins
- An example, the FFAT SLiM, [DE].{0,4}E[FY][FYK]D[AC].[ESTD]

The function of SLiMs

- activation or deactivation of proteins
- modifying proteins through post-translational modifications
- serving as target sites for cleavage by proteases
- targeting proteins to specific subcellular localization



PLoS One, 2008, 3(7), e2524

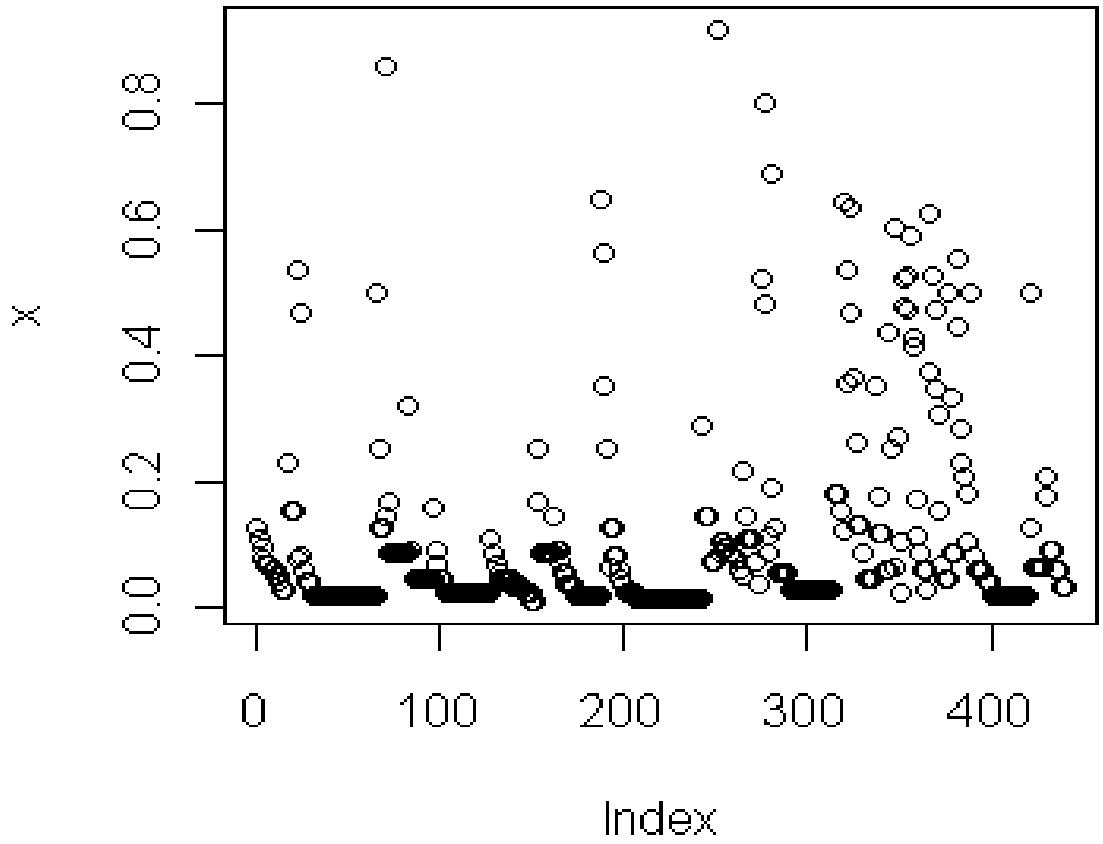
Experimental approaches to discover SLiMs

- Typical experiments involve raising an antibody to peptide that expresses a SLiM and then using this antibody to test the surface exposure or accessibility of the SLiM (*Nucleic Acids Res* 26*,* 5486-5491, 1998), followed by the mutation or deletion analysis.
- Experimental approach (Time consuming, cannot detect transient and subtle signals)

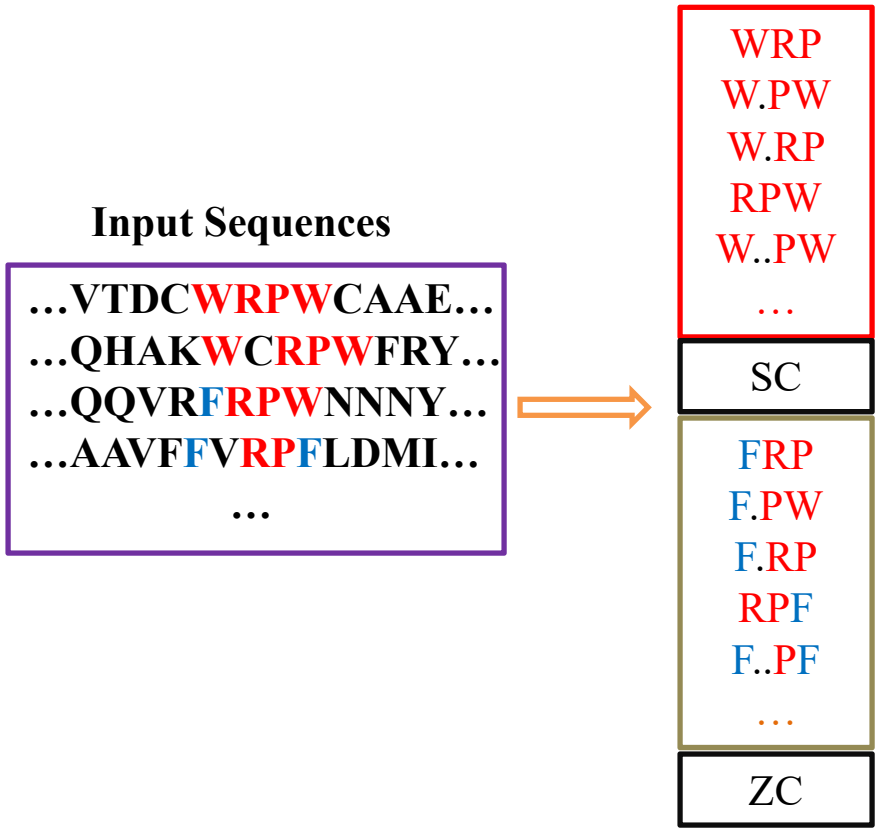
Computational approaches

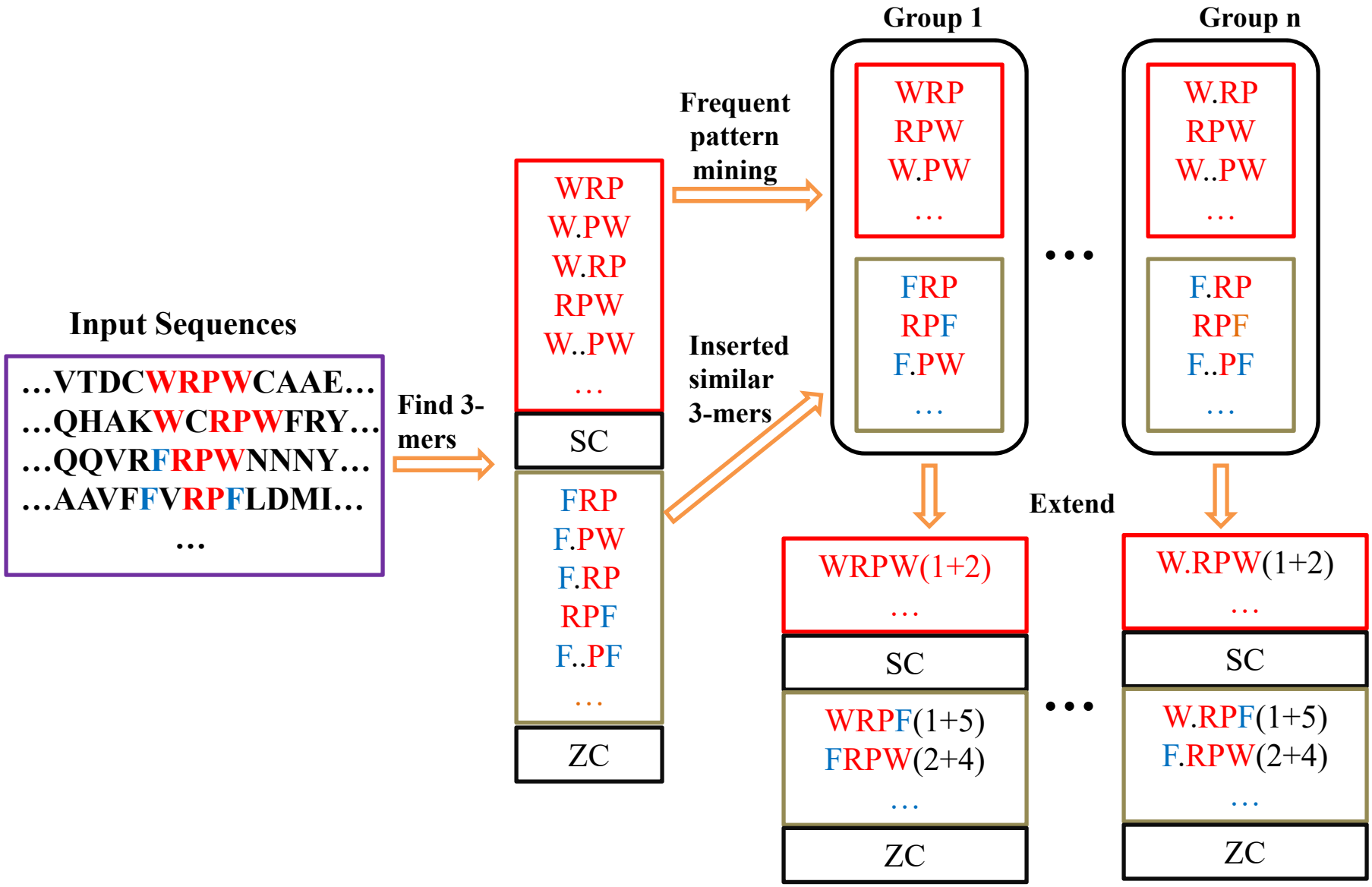
- Teiresias, 1998
- D-Motif, 2006
- Dilimot, 2006
- SLiMFinder, 2007
- SLiMDisc, 2010
- BayesMotif, 2010
- FirePro, 2010

Occurrence number difference of different subpatterns



[WF].{0,1}RP[WF]

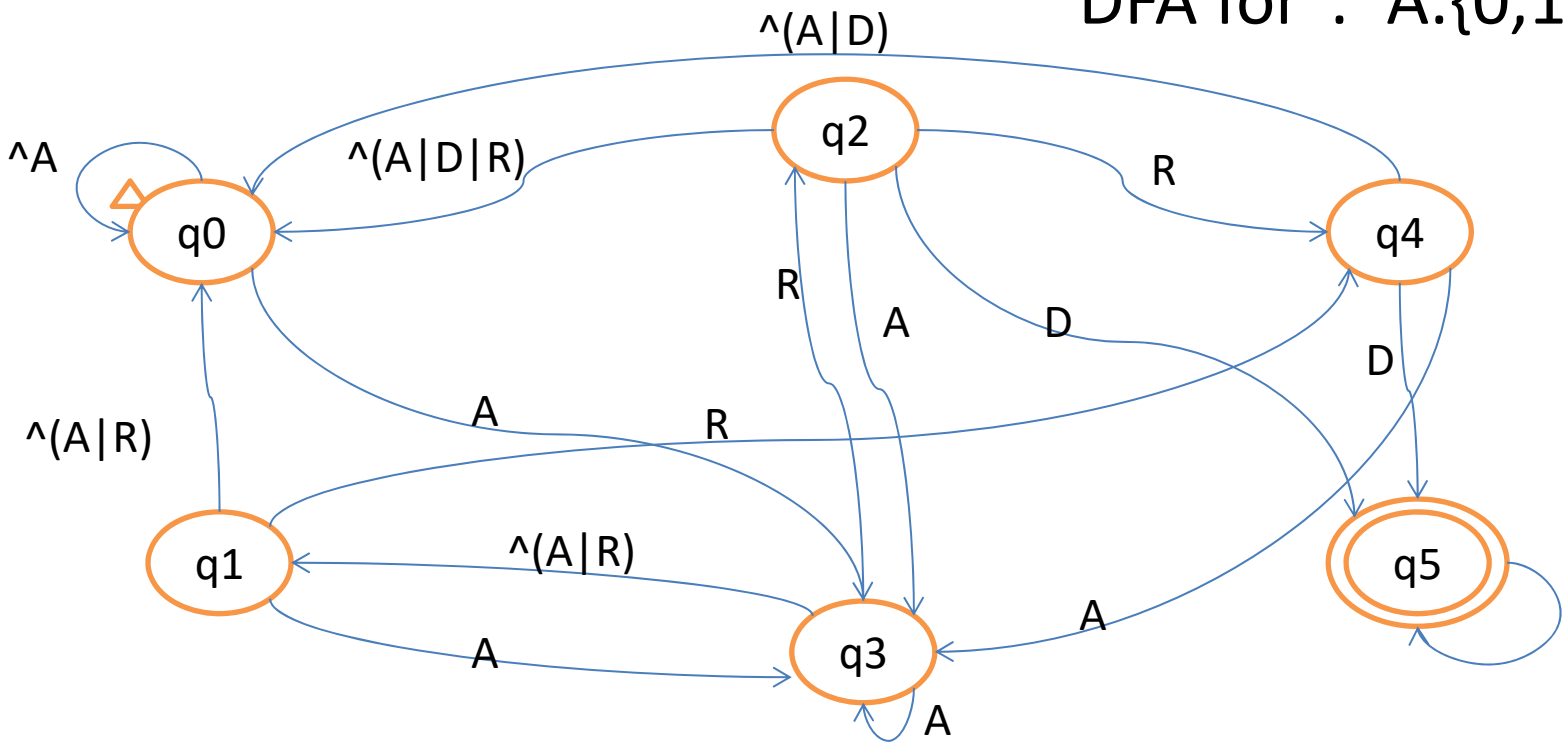

Input Sequences
…VTDCWRPWCAAE…
…QHAKWCRPWFRY…
…QQVRFRPWNNNY…
…AAVFFVRPFLDMI…
…
WRP
W.PW
W.RP
RPW
W..PW
…
SC
FRP
F.PW
F.RP
RPF
F..PF
…
ZC



IEEE transactions on knowledge and data engineering 17: 1347-1362 (2005).

Deterministic finite automaton

DFA for '.*A.{0,1}RD.*'



$$P = \begin{pmatrix} 0.95 & 0 & 0 & 0.05 & 0 & 0 \\ 0.90 & 0 & 0 & 0.05 & 0.05 & 0 \\ 0.85 & 0 & 0 & 0.05 & 0.05 & 0.05 \\ 0 & 0.90 & 0.05 & 0.05 & 0 & 0 \\ 0.90 & 0 & 0 & 0.05 & 0 & 0.05 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad prob(n) = V_{start} P^n e_{apt}$$

Test on simulated data

	DoubleFlex		SLimFinder	
	100	400	100	400
LIG_EH_1 .NPF.	NPF	NPF	NPF	NPF
LIG_HOMEOBOX [FY][DEP]WM	[FY][DEP]WM	[FY][DEP]WM	[FY][DE]WM	[FY][DE]WM
LIG_CtBP [PG][LVIPME][DENS]L[VASTRGE]	G.[DEKNS]L	K.{3,4}[NQT][G]	LKD..[AG]	N/A
	[EILMV][DENS]L(2)	[GP][IM].L(10)	G[LV][DE]L(7)	
LIG_CYCLIN_1 [RK].L.{0,1}[FYLIVMP]	K.L.{0,1}[FMNVY]	G.{2,3}P.{1,2}V	CQ..[LV]	S.VP
	[RKT].L[FIMPVY](2)	N/A	N/A	N/A
TRG_ER_FFAT_1 [DE].{0,4}E[FY][FYK]D[AC][ESTD]	E[FY][FYK]D[AC]	E[FY][FYK]D[AC]	E[FY][FY]D	E[FY][FY]D

Tested on benchmark data

Motif	DoubleFlex		SLiMFinder		DiLiMOT	
	close result	Z-score	close result	p-value	close result	p-value
LIG_WRPW_1, 96, 156	WRPW(1)	466.89	None*	None*	FSDPR(1)	0
[WFY]RP[WFY].{0,7}$	WRP[WFY](2)	247.1			V.PW(6)	0
LIG_EH_1, 32, 342	S.NPF(1)	38.08	NPF(1)	9.21E-08	NPFQ(1)	0
.NPF.	NPFL(2)	37.14	[GS].NPF(2)	1.84E-07	NPF(10)	0
LIG_Rb_LxCxE_1, 26, 326	L.C.E(1)	7.92	L.C.E(1)	2.15E-04	L.C.E(1)	0
[LI].C.[DE]	[LI].C.E(2)	7.42	A..G..T[IL](2)	0.11		
LIG_CtBP, 29, 487	D.PLDL(1)	39.93	P[ILM]DL(1)	2.00E-03	P.DL(1)	0
[PG][LVIPME][DENS]L[VASTRGE]	PLDLS(5)	24.71	D.P[IL]DL(2)	0.007	P.DLS(6)	1.13E-47
MOD_SUMO, 30, 256	VK.E(1)	11.41	[VIF]K.E(1)	2.71E-06	VK.E(1)	3.02E-24
[VILMAFP]K.E	[VIMF]K.E(2)	10.88	VK.E(2)	1.84E-06		
LIG_CYCLIN_1, 21, 280	KR.[IL].L(1)	15.88	[KR][KR]..F(1)	1.60E-02	YISP(1)	1.38E-35
[RK].L.{0,1}[FYLIVMP]	KR.L.[FLV](2)	15.6	[KR][KR]R[IL](2)	0.052		
LIG_PTAP_UEV_1, 20, 127	P[TS]AP(1)	12.12	P[TS]AP(1)	3.56E-09	PSAP(1)	7.18E-34
.P[TS]AP.	PSAP(2)	8.59	P[TS][AS]P	1.85E-05		
MOD_PKA_1, 20, 193	RR.S(1)	4.82	RR.S(1)	4.90E-02	RR.S(1)	3.66E-15
[RK][RK].([ST])[^P]..	VE.E(2)	4.60	[KR][HR].S	0.085		
TRG_ER_FFAT_1, 20, 375	EFFDA(1)	285.67	E[FY].DA.[DE](1)	1.29E-10	E.FDA(1)	0
[DE].{0,4}E[FY][FYK]D[AC].[ESTD]	F[FY]DA.E(2)	268.99	D.E[FY].DA(2)	1.69E-09		
LIG_HOMEOBOX, 16, 190	F[DP]WM(1)	218.18	None*	None*	PWM(1)	0
[FY][DEP]WM	[FY][DP]WM(2)	200.78			FPW(2)	4.79E-69

Test on chloroplast transit peptides

Motifs	#cTPs containing the SLiM (total 408)	#NECPs containing the SLiM (total 1924)	#proteins containing the SLiM (total 35386)	Enrichment p-value
[PS].S[FY]	119	495	5784	7.40E-28
[KN]P.{1,2}[FS]	101	601	7218	1.22E-31
SSS	175	776	9183	1.09E-46
S[PS]..[LY]	160	773	9091	1.57E-47
[HL]…[IS][S]	149	711	9549	1.13E-23
[LW]..SS	121	533	6065	2.20E-33
L..SS	117	503	5618	3.13E-33
[LS]….[KS]…F	102	560	6897	3.02E-26
[IS]S….[PQ]	124	622	7678	1.47E-29

Discussions

- Speed
- Dipeptide SLiMs
- Statistical significance

Acknowledgement

- Xiaoman Li, Burnett School of Biomedical Science, University of Central Florida
- Ping Ge, Department of EECS, University of Central Florida

BLOSUM62 Substitution Matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W