Biophysical Journal, Volume *120*

Supplemental information

Intrachain interaction topology can identify functionally similar intrinsically disordered proteins

Jonathan Huihui and Kingshuk Ghosh

This supporting material contains principal component plot for PSC protein family, Euclidean distance matrices (used to classify proteins in a family), Sequence Charge Decoration matrices ($SCDM$) presented as color coded maps for PSC and RAM family. It also contains methods and results for three control studies: i) using charge composition, ii) shuffling the $bSCDM$ matrices to see the role of topology of the charge decoration matrices, and iii) charge product matrices. We also provide color-coded K_d values for RAM sequences, and the sequences that were used for Ste50, PSC, and RAM families.

First 2 Principal Component Plot for PSC

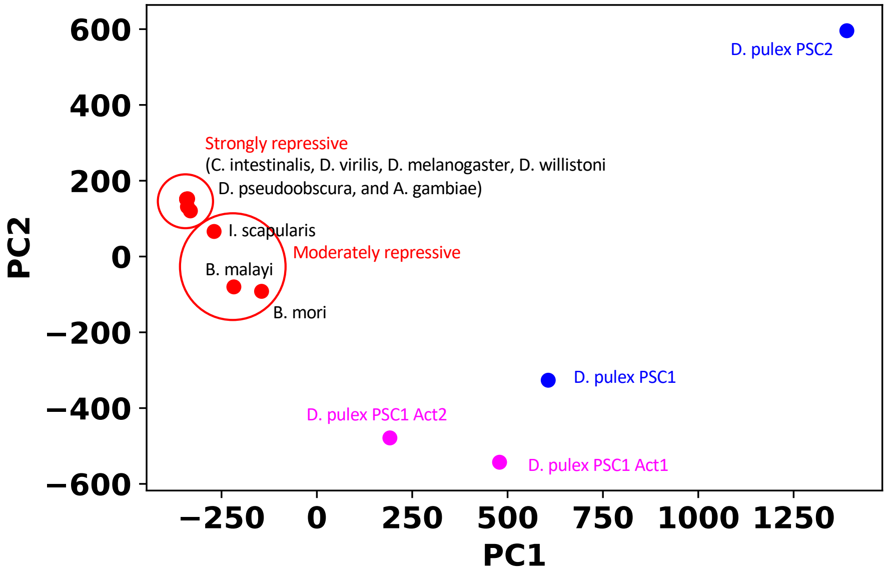

Figure S1: **Scatter Plot of the First 2 Principal Components of PSC.** PSC family proteins represented in terms of the fist two principle components. Collectively, PC1 and PC2 account for 74% of the variance (56% and 18% respectively). Classification using two PCs is in line with that shown in Figure 3 of the main manuscript.

Distance Matrices Used for Clustering

Here we include the distance matrices that represent the distances between the Principal Components of each protein within a family.

	RAD26	**SC5A**	**SCCharge**	**LKCharge**	**PEX5**
RAD26	0.0	131.2	127.0	138.7	143.0
SC5A	131.2	0.0	80.9	78.0	93.2
SCCharge	127.0	80.9	0.0	27.4	69.5
LKCharge	138.7	78.0	27.4	0.0	45.7
PEX5	143.0	93.2	69.5	45.7	0.0

Table S1: **Euclidean distances between Principal Components of Ste50 proteins.**

	D. w.	**Act2**	**B. mo.**	**A. g.**	**I. s.**	**D. ps.**	**D. p.2**	**D. p.1**	**D. v.**	**Act1**	**D. m.**	**C. i.**	**B. ma.**
D. w.	0	939	612	81	378	1.2	1803	1204	0.6	1162	56	11	525
Act2	939	0	839	895	852	939	1649	971	938	744	908	938	792
B. mo.	612	839	0	545	315	612	1728	1067	612	1023	582	603	155
A. g.	81	895	545	0	300	82	1793	1198	82	1134	39	74	449
I. s.	378	852	315	300	0	379	1749	1216	378	1084	334	368	195
D. ps.	1.2	939	612	82	379	0	1803	1204	1.2	1163	57	12	525
D. p.2	1803	1649	1728	1793	1749	1803	0	1412	1803	1516	1800	1799	1766
D. p.1	1204	971	1067	1198	1216	1204	1412	0	1204	986	1213	1202	1144
D. v.	0.6	938	612	82	378	1.2	1803	1204	0	1163	56	12	525
Act1	1162	744	1023	1134	1084	1163	1516	986	1163	0	1141	1160	1022
D. m.	56	908	582	39	334	57	1800	1213	56	1141	0	52	485
C. i.	11	938	603	74	368	12	1799	1202	12	1160	52	0	516
B. ma.	525	792	155	449	195	525	1766	1144	525	1022	485	516	0

Table S2: **Euclidean distances between Principal Components of PSC-CTR proteins.** Protein names have been shortened due to formatting constraints. **D. p.2**, **D. p.1**, **Act2** and **Act1** denote *D. pulex PSC2*, *D. pulex PSC1*, *D. pulex1 Act2* and *D. pulex1 Act1* respectively.

	9	**12**	**7**	**3**	**4**	**11**	**1**	**5**	**10**	**13**	**2**	**6**	**8**	**WT**
9	0.0	79.3	80.5	92.9	53.4	100.8	76.9	45.2	88.9	102.9	62.7	60.6	70.1	70.7
12	79.3	0.0	88.0	66.0	81.1	81.3	83.1	86.6	88.2	77.7	81.0	78.3	80.4	79.3
7	80.5	88.0	0.0	65.6	62.4	61.9	45.8	64.6	35.1	64.4	39.5	66.0	57.5	33.7
3	92.9	66.0	65.6	0.0	78.0	42.0	45.4	80.0	57.9	43.4	64.1	75.6	72.8	55.3
4	53.4	81.1	62.4	78.0	0.0	89.3	59.4	22.7	71.9	90.8	35.0	39.7	65.2	49.4
11	100.8	81.3	61.9	42.0	89.3	0.0	66.0	89.2	48.6	9.7	77.4	87.7	85.0	70.9
1	76.9	83.1	45.8	45.4	59.4	66.0	0.0	61.8	44.1	70.9	40.6	62.7	61.2	26.2
5	45.2	86.6	64.6	80.0	22.7	89.2	61.8	0.0	74.5	91.7	40.3	52.9	58.8	53.6
10	88.9	88.2	35.1	57.9	71.9	48.6	44.1	74.5	0.0	53.5	60.2	76.1	69.6	49.9
13	102.9	77.7	64.4	43.4	90.8	9.7	70.9	91.7	53.5	0.0	79.5	87.7	85.5	74.1
2	62.7	81.0	39.5	64.1	35.0	77.4	40.6	40.3	60.2	79.5	0.0	44.3	56.6	20.5
6	60.6	78.3	66.0	75.6	39.7	87.7	62.7	52.9	76.1	87.7	44.3	0.0	70.9	56.1
8	70.1	80.4	57.5	72.8	65.2	85.0	61.2	58.8	69.6	85.5	56.6	70.9	0.0	58.2
WT	70.7	79.3	33.7	55.3	49.4	70.9	26.2	53.6	49.9	74.1	20.5	56.1	58.2	0.0

Table S3: **Euclidean distances between Principal Components of RAM permutations.**

Sequence charge decoration matrices for PSC-CTR

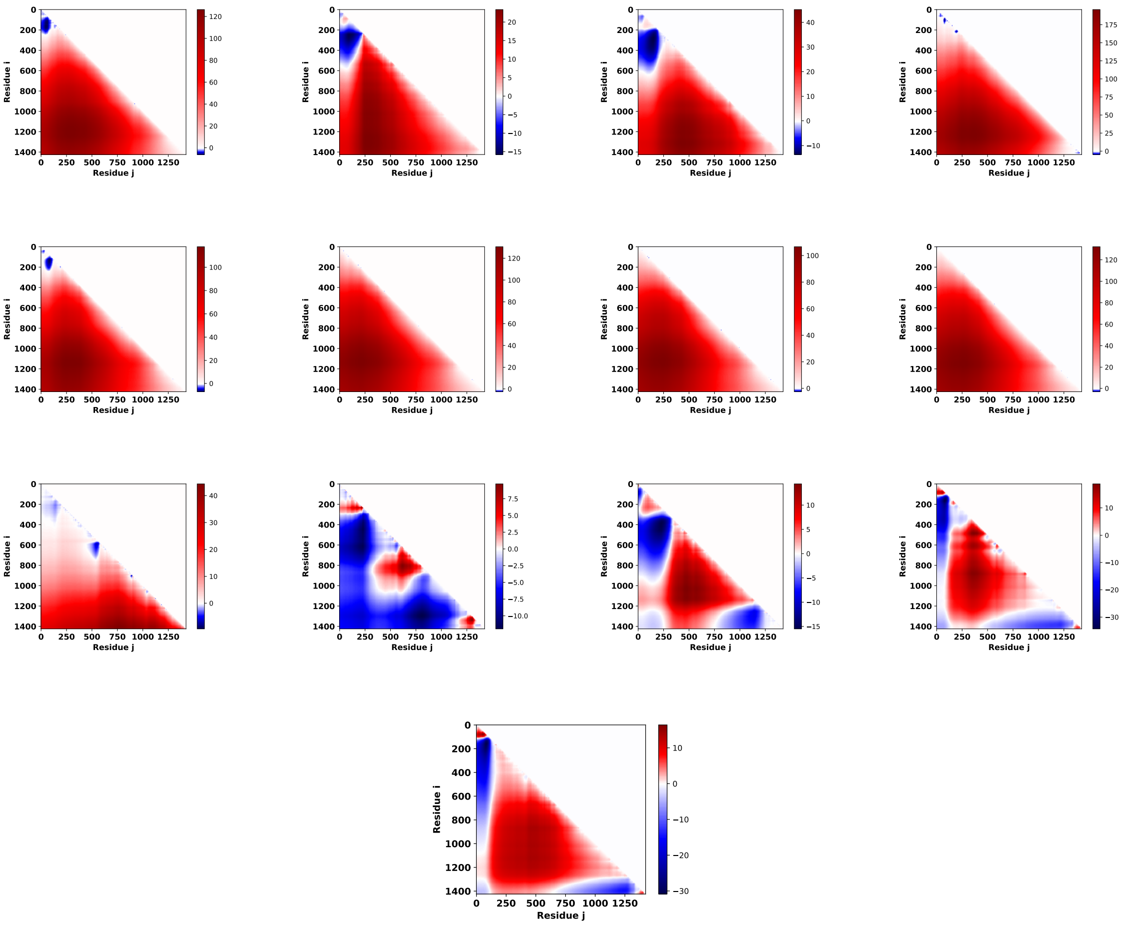

Figure S2: **Sequence Charge Decoration Matrices for PSC-CTR offer further visual evidence for links to function.** The color coding above depicts where electrostatics is predicted to promote expansion (red) or compaction (blue). From top left to bottom right, the rescaled $SCDM$s are included for *A. gambiae*, *B. malayi*, *B. mori*, *C. intestinalis*, *D. melanogaster*, *D. pseudoobscura*, *D. virilis*, *D. willistoni*, *I. scapularis*, *D. pulex2*, *D. pulex1*, *D. pulex1 Act1*, and *D. pulex1 Act2*. There is a clear visual trend in the matrices that distinguish inhibitory and non-inhibitory sequences, see main text for discussion.

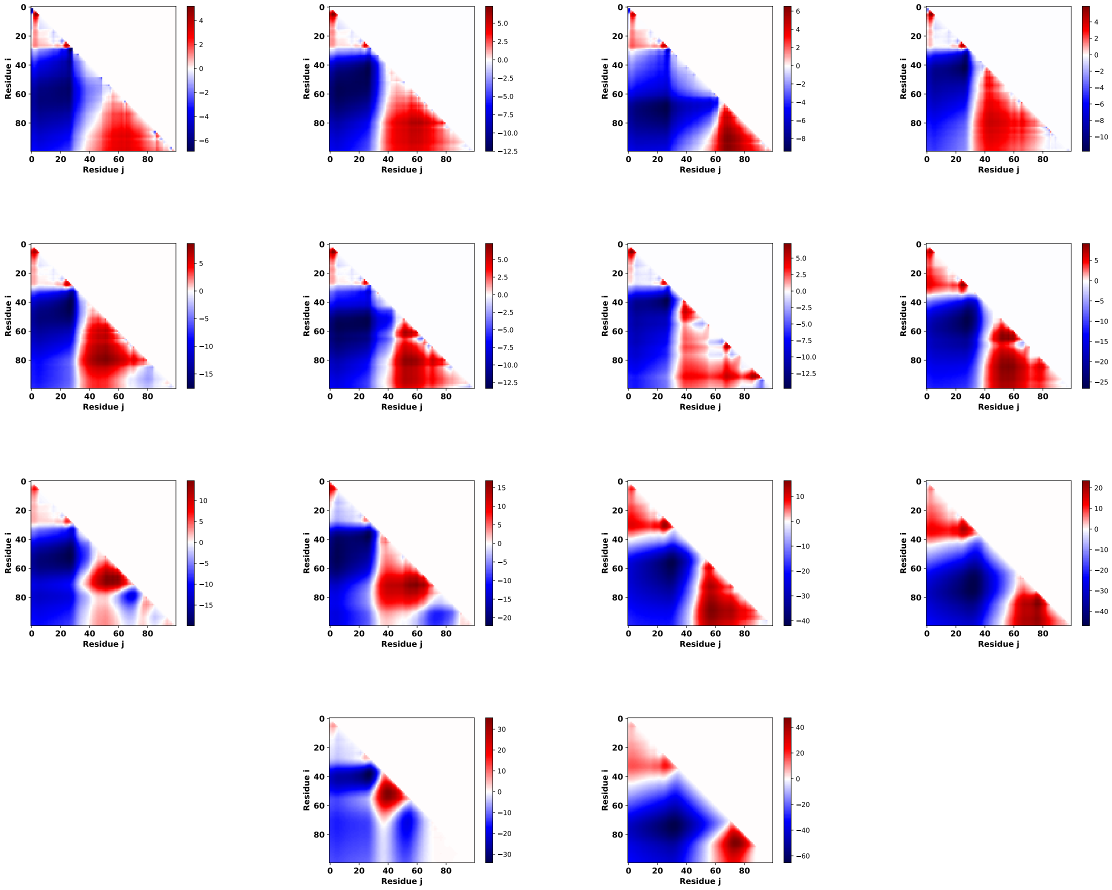

Figure S3: **Sequence Charge Decoration Matrices for RAM sequences provide visual evidence for overall trend.** The color coding above shows regions where electrostatics is predicted to promote expansion (red) or compaction (blue). $SCDM$s are included for (from top left to bottom right) RAM 1, 2, 3, 4, 5, WT, 6, 7, 8, 9, 10, 11, 12, and 13. RAM 12 is visually different from all the others and RAM 3, 11, and 13 look similar, agreeing with the dendrogram in the main text. See main text for discussion.

Composition Based Clustering

Methods

Clustering was performed with the fraction of positive and negative residues as independent coordinates. The same hierarchical agglomerative clustering algorithm as clustering with binary SCDMs was then employed to determine which proteins were most similar to each other by the Euclidean distance between these individual coordinates. This method was not used for the RAM proteins because all of the sequences were generated by shuffling the original sequence while maintaining the same composition.

Results

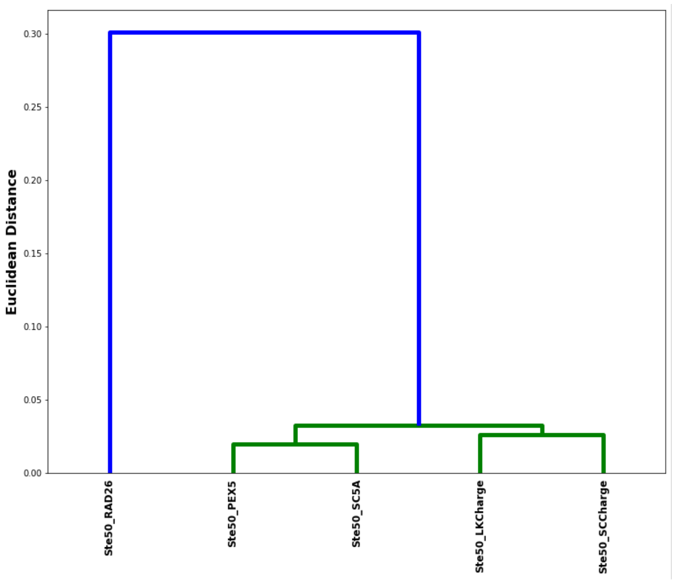

Figure S4: **Compositional clustering for Ste50.** The panel shows the resulting dendrogram based on the clustering by charge composition for the Ste50 proteins. This method clusters non-functional SC5A and functional PEX5 together demonstrating the inadequacy of the algorithm.

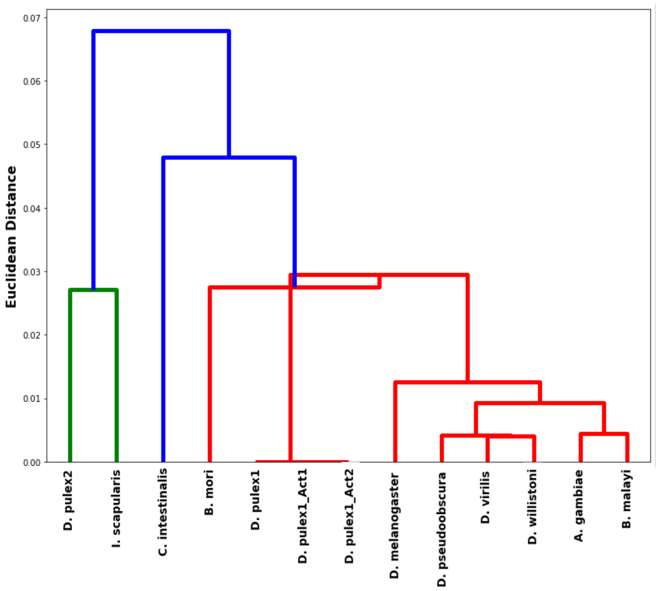

Figure S5: **Compositional clustering for PSC-CTR.** The panel shows the resulting dendrogram based on the clustering by charge composition for the PSC-CTR proteins. Non-repressive and repressive proteins are clustered together indicating inability of this metric to properly classify proteins.

Clustering using shuffled matrices Methods

Binarized sequence charge decoration matrices ($bSCDM$) were calculated for all of the proteins and were randomly shuffled, with the average $bSCDM$ tracked. The amount of times the $bSCDM$ matrices were shuffled depended on the cumulative change in the average $bSCDM$. A mathematical representation of this criteria would be $\delta = \sum_{i=2}^{N} \sum_{j=1}^{i} |\langle bSCDM_{i,j} \rangle_T - \langle bSCDM_{i,j} \rangle_{T+1}|$, where $\langle bSCDM_{i,j} \rangle_T$ is the average of the binary sequence charge decoration matrix after T iterations and δ is the difference between the average at the T and $T+1$ iteration. The average matrix was then subjected to the same PCA and clustering technique used to create the dendrogram. Multiple δ values were tested and resulted in 10^5 to 10^6 iterations performed for each individual matrices. The dendrograms were compared at each δ value and the appropriate δ value was chosen after visually determining the dendrogram did not significantly change (not shown).

Results

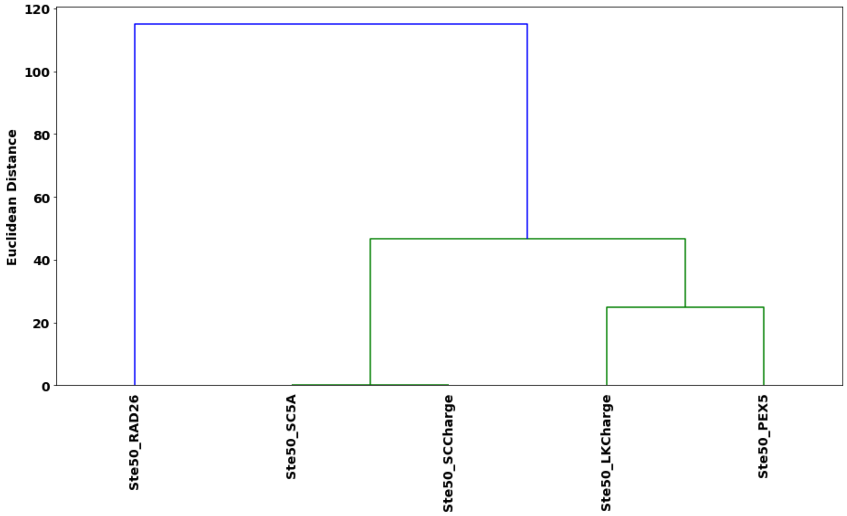

Figure S6: **Results of shuffling the topology of the $bSCDM$ matrix for Ste50.** The dendrogram based on the clustering of the Principal Components (capturing about 100% of the variance) of the average binary sequence charge decoration matrices does not agree with experimental data. For example, SCCharge (functional) and SC5A (non-functional) are clustered together.

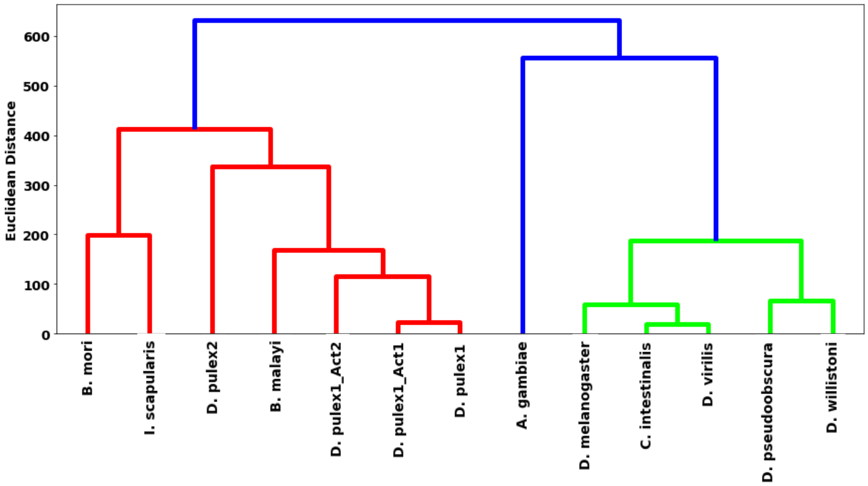

Figure S7: **Results of shuffling the topology of the** $bSCDM$ **matrix for PSC-CTR.** The dendrogram based on the clustering of the Principal Components (capturing about 97% of the variance) of the average binary sequence charge decoration matrices does not agree with experimental data. For example, non-repressive (*D. pulex PSC1*, *D. pulex PSC2*) and repressive proteins are clustered together. See main text for more.

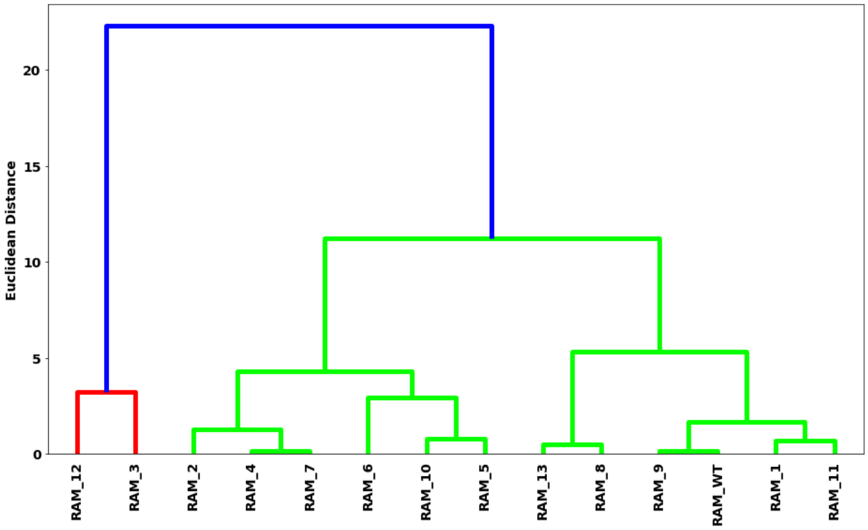

Figure S8: **Results of shuffling the topology of the** $bSCDM$ **matrix for RAM.** The dendrogram based on the clustering of the Principal Components (capturing about 99% of the variance) of the average binary sequence charge decoration matrices does not agree with classification using experimentally measured K_d data. See main text for more.

Control using charge-product Calculation

Methods

charge decoration matrices were calculated for all of the proteins within a family by using a charge product (CP) matrix defined as:

$$[CP]_{i,j} = q_i q_j \tag{S1}$$

where q is equal to +1 for positively charged amino acids (Lysine and Arginine), -1 for negatively charged amino acids (Glutamic and Aspartic acids), and 0 for all others. CP matrices are then rescaled to the largest protein as done previously. Principal Components were then calculated within a family of proteins and these components were then clustered in the same fashion as before.

Results

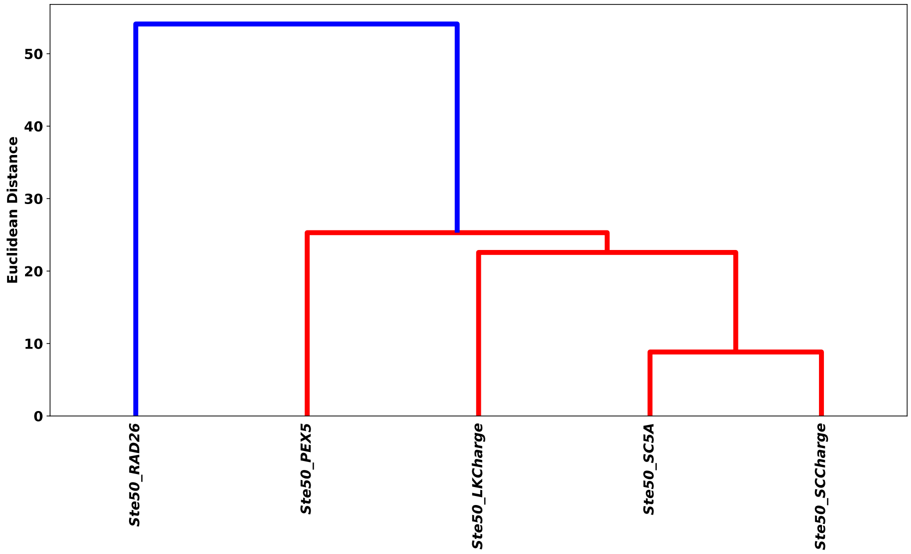

Figure S9: **Clustering based on charge product CP matrix used for Ste50.** The dendrogram based on the Principal Components (capturing about 100% of the variance) of the charge product matrix correctly classifies RAD26 outside of the functional proteins, however it incorrectly clusters SC5A within the functional protein group.

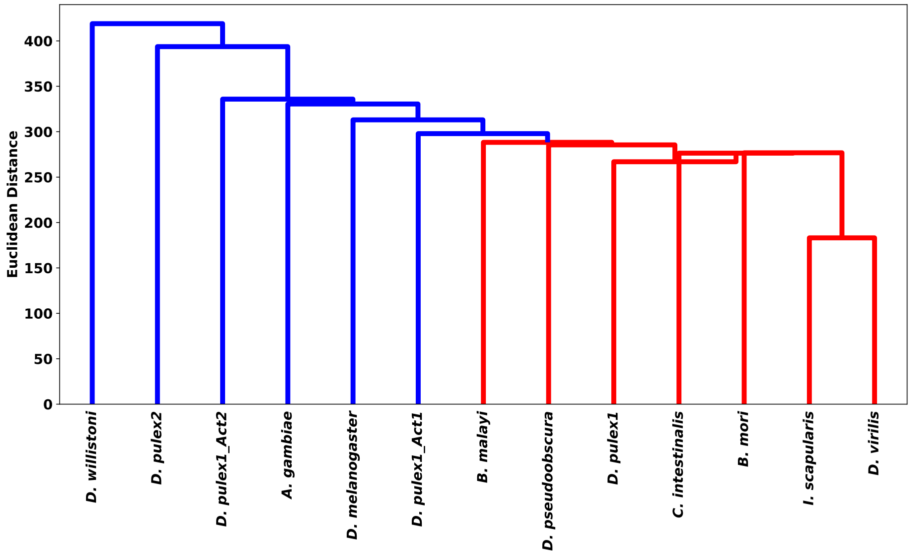

Figure S10: **Clustering based on charge product CP matrix used for PSC.** The dendrogram based on the Principal Components (capturing about 96% of the variance) of the charge product matrix incorrectly groups strongly repressive, moderately repressive, and non-repressive proteins together (red cluster).

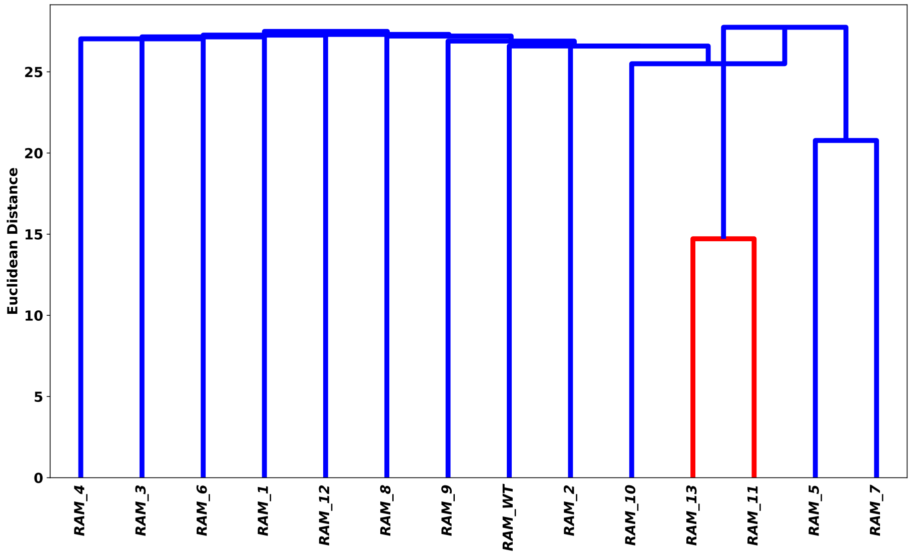

Figure S11: **Clustering based on charge product CP matrix used for RAM.** The dendrogram based on the Principal Components (capturing about 92% of the variance) of the charge product matrix reveals no trend between sequence patterning and K_d.

Color-coded K_d values for RAMANK sequences

Protein	K_d **(nM)**
RAMANK 1	10.1
RAMANK 2	11.3
RAMANK 3	11.8
RAMANK 4	9.7
RAMANK 5	16.2
RAMANK 6	22.3
RAMANK 7	17.2
RAMANK 8	11.8
RAMANK 9	18.9
RAMANK 10	32.2
RAMANK 11	29.1
RAMANK 12	99
RAMANK 13	38.5
RAMANK WT	9.2

Table S4: **Experimentally measured K_d values for RAMANK**. Color coding corresponding to clustering using our theoretical algorithm shown in the main text.

Sequences Used For Classifications

Protein	**Sequence**
RAD26	DTANREYAKNDEQKDEDFEMATEQMVENLTDEDDNLSDQDYQMSGKESEDDEEEENDDKILKELEDLRFRGQPGEAK
SC5A	DVLDVMKTSSSSAPINTHGVSTTVPSSNNTIIPSSDGVSLSQTDYFDTVHNRQAPSRREAPVTVFRQPSLSHSKSLHKDSKNKVPQISTNQSHPSAVSTANAPGPAPNEALK
SCCharge	DVLDVMKTSSSSSPINTHGVSTTVPSSNNTIIPSSDGVSLSQTDYFDTVHNRQSPSRRESPVTVFRQPSLSHSKSLHKDSKNKVPQISTNQSHPSAVSTANEEGPEENEALK
LKCharge	DVLELIRRNNGNINTTEESFGTQPQPTGDYFDQQKHPLIINGSSGTTNNLGSNGSKSSVLRSGSSTASVPALASSNSFGGEEGGNSTNEPLK
PEX5	LIDDKRRMEIGPSSGRLPPFSNVHSLQTSANPTQIKGVNDISHWSQEFQGSNSIQNRNADTGNSEKAWQRGSTTASSRFQYPNTM

Table S5: **Sequences used for the Ste50 proteins.**

Protein	**Sequence**
RAM 1	DDRKRRRQHGQLWFPEGFKVSEASKKKRREDLEKTVVQELTWPALLANKESQTERNDLLLLGDFKDGEPNGMALDSMHVPAGPMFRDEQDARWDQHKDQD
RAM 2	MARKRRRQHGQLWFPEGFKVSEASKKKRRDPLGKESVGLDPLDNASDGALMDRNQNDWGDKDLETREFEFKDPVVLPELEDQTKHDQWTQQHLDAARLEM
RAM 3	EERKRRRQHGQLWFPEGFKVSEASKKKRRWEDVKDATQVWDTKLGELKSHLGMMNNRLGDRRQDLPDPENDQADLSEAHQQTALDPAMLDPFDLKFEVGD
RAM 4	MERKRRRQHGQLWFPEGFKVSEASKKKRRLFDMQDVVDRWQELEMDTLSENHAPDNASRQDWNRVEDLQLLTGLEPTGLDHQDKKDDLKFDAPGGAPKAE
RAM 5	MARKRRRQHGQLWFPEGFKVSEASKKKRRRPLGEDSVGLEPLDNASDGALMEENQNDWGDDKLDTERFRFDDPVVLPDLDEQTDHKQWTQQHLKAAKLEM
RAM 6	LFRKRRRQHGQLWFPEGFKVSEASKKKRRADPWWSSTVEEDPQDHEPDLLGDGALKRGFQGNTVKAQDEDDDALPLKLRMHLVMADQELEEDDMRNTNQK
RAM 7	MARKRRRQHGQLWFPEGFKVSEASKKKRRKPLGRKSVGLDPLENASDGALMEDNQNEWGEDDLDTEDFRFKKPVVLPDLEDQTEHDQWTQQHLDAARLDM
RAM 8	MARKRRRQHGQLWFPEGFKVSEASKKKRRKPLGDDSVGLKPLDNASEGALMEDNQNEWGDDDLETEEFDFEDPVVLPRLRKQTKHRQWTQQHLDAADLDM
RAM 9	RKRKRRRQHGQLWFPEGFKVSEASKKKRRAAQAQNEEHEDDLEQVAVNMGKFDVLDSLPDDLGLEDEETLDDDMPHQDAPLFGLDGLNWWRRQTPKMSKT
RAM 10	FHRKRRRQHGQLWFPEGFKVSEASKKKRRKKKRLLLQVVPRQLSTAPNMLDHWDTDDDDDDLLVAGFLNQDEEETQRPGAEMGPDAEQEEGAMDSDKLWN
RAM 11	DLRKRRRQHGQLWFPEGFKVSEASKKKRRLKKRKKQRRAPGMPELGWLQMHSLNVALNNSGADTDLDEPQMFHTAEDEEDDDDFDDLPVQQGLADVETEW
RAM 12	LMRKRRRQHGQLWFPEGFKVSEASKKKRRRATFALHDDDEEEEFDDDEDEDDDQDEDSLWLALNHRPWTQKGKANNKSVAQQRGPMVGGPMTLKLLLPVQ
RAM 13	LQRKRRRQHGQLWFPEGFKVSEASKKKRRRRKKKRKTVPAAAWLSQQPVMPTHTLSLQMQPNWLVNLGMFDDDDEDEEEDDEEDDDDDEANFGLGHALGL
RAM WT	MARKRRRQHGQLWFPEGFKVSEASKKKRREPLGEDSVGLKPLKNASDGALMDDNQNEWGDEDLETKKFRFEEPVVLPDLDDQTDHRQWTQQHLDAADLRM

Table S6: **Sequences used for the RAM proteins.**

Protein	**Sequence**
A. gambiae	RDAPMKYYYRIRTTESNPVELPEVALRRSPSLVTALPPAQRPSVDEEDDKE NRVRLDRIVSEAASNESDSSSSSSSNTIANTPRADASKPPTAAQVTPAPES PATPTQPRKNESIKLKIGLNKNTYVSILQSPQPDEPSTHSSSSSSSSSASS PGSEGAKSSSSSHKSEKSKRKRKDALATLQQMEENSRELKFKIEQMKDTGL VGSKSKSGKSSAKHHQHQQLALVPYKVELSGGLSQPSAVPDPERSDSKRLH SAKNGSNSSSSGSSPAYCKLKIKKSSPEDSKQPHHHHPIVLKIDQRSPEMA TATLKFGMPRKSEKSMTPSPPPLPPSPPTPPSPKQKFADEKSQFLNSFQLT PIKPAEQSSSPSKTSAGAATTTATTTTPPAAVESVAPAGKKSPTSTPSVPP VAAPTNSTSPPASNGTGTTTKRKAKDASSGGGVPRSGPKKPKLSNDEIKAI VEKTVAENIRSPSEHIVPPIFLKPKPPTTTAAAAASGQPSPPLPSSSAPTK AKDPSPKRDSPRPFVFKTPPPPPPPPIVSANNIPAVKPSQQVLPAPVPQKH APVPIRPALTTAPKPAVPQVPQTHIRKPTAPTKLPTSAAGTGGVKSASPPA QQPLSNGSQQQAAPSNATTHSVAAGANRQQKGLELKRAQSNPSINIPPPQS SAPPVTVPRDTEISKLRPEDLKKNQKVYGPQTVPEQQQQPPKPNTTTEATG ASFAVPGPKAAPKPSSSSAAAPVGNATKSSGSAQAGQGTKARPVNYLNYAL LNSKAAAAGSRTPIPSYSSSSPSYSPDSPQYSPNLNFSSKQFKYANPLAYN SHLQNMLNDRRTGSTSPPGSSTSTIPASSPSPPQDRPAATTPNASGNKRPA SALSPTAEDKKQQPPEKQPALLSAAAPNPADKFPSGIPDGLSVTLATDDDD AARIKNVNKQLKNNFIEIRALPEVPITEVKLPLPLPSSSTTTTASKPGRRT PPGKAVAAAAGSPATLSGSPMARKSSSPSVPAPTYSVAASAPPKTTVSSAA PAPANRPADALQRKIIDLIDKPSPSSSAKTSSKPPPTMPTVSRPSTPKGGT SGGFPPVNNGNKFKLPNATVNENGTLKLNNYREVDLIPKGAAASGAKSAPS PPSGASSRTMPPPTSSASSKSIMPIAPKPSSSQLQSFANGRMAMSPPIQRS PTATSGYQQPKSKTPPSQLPSMASMGPMFDIQMKSAIAAAAASGGGTTPSK KPPTSSTTLTSATVPRRKIVPTSNSTSLVPLKTSPVAASPSTTAGGAGSKL LSSNYSDYITLHPQGPVSSSAPPSRPLFGTPHQQQHAALTQILSENFARQC FNNLPFPYLLQQFAHHQPGAASMGSRGLGSDSVTITASPMGAARGAVSQNS LTVTAIPPGQQGGGGGGSGARGSGGGGLNGPRGGIGGGGPNPASRNSS
B. malayi	RLGPMKVLFTLQRHLEEEKPPVLDMEFMPELVAEEPLSQGSVSVAAAIETP VQLPALTVSLNTSMMEGGPNHQPIITTEVHPPPRKKRKSTAPTKKQVASPI PVQRMTGVSPLAKGPPPLMRLENTGLSKKSVSSGRIKSTEKTPAKTPPHED TPATKQAKLMPTSFDNKLQQIIDSSPSRSTKISKGSKTKTLAKAASGFAES SSKLVSNDKSMESSSKPGNKNGSLQAMKLISTDGIKTESSSSNVKTENVTN KEKTLSKLHTISKITENTTVITTISTPASTTAATATTISHPRPIQPRPLEM KTNYEALVKSYGLNGIGNKLSSFPLDGKHVGFSPPIHMQAPPLFMDPKIAA QPIKHILSGRGMPIVPEATPYLRNPALANFMHHLHMQPPPIPGLPGTSTPP LLSHSSSSTLSCSSSNQVTSTHSPPNSNVSSKNSQQQLKHPTAVPLPPATV SSNGSGNKLLNNSSSNSRASSPAAKLQQKIVSPIPIPTAHITPFMSHS

B. mori	RNEPMRFFYQIIDYVAIRNRIFDINRKRSHFHDQKLSPVSTEDTSTSSPAPNLHDHASEASSGPSSPVPDDNNRNTPEVLTNDKMNVQNDESCNDKNDYSSTNKLDEDVEKSQFLNSFELTAKSSCIPVKSPQKFNTEKLSLAKEVVTKSITSKVKAEDPTPDNLKRKNHTSPPTPELKKLKVEISNCLPSFSVQPSSSSISTKTEENQRKHETVDCNKNNQSAIKNNAPSATPTTRDLKQPQTVKQQVGTSKQTLDNSGVKRTVVGPQNILSPKRKPPNESTAEQAMPQQQQQQKTLSPLKLQIPKLDAVSKTSEAPKPPLKKIPDLKPSMPMLQSAHSKSPAMNKVRMDLLANNSDPTIDRSKILSQVKSSMGVQSPAQNSGDPLKSLFDSCKINIPSSLSITLTDQKSDNRCPVDTLDPKKNFTNKNLAAASSSSISAHKVPSPPVHNYIEILKLPESDSNLKKIAKNEADSKTNSQCKPEISQTKPTTKGSETSTKGPVPNLKPIADTKLAKQAGNFSTPITFQQTFEQQLQSLQCDKKGKPKNKAQVPKLVPATPKSLSAVTKPIIPVNKPTNSSSTETKTGTALDLTTPHNIQSQLAVQQTFDKALETMHSIANLAKKQNLPSKGIPMSLTHSNIFPGITSRPLTAGINSVRLSSPNTINQVKLDKPNPNVPTVTGNNRQESSIKSSQVKQLGNMNLTLQSPAYQIPSAHPPSNAQPSPRSQTRSPSSSPKLVIAEEKQTSTTVMEHNVSQLNQVTSTHITNFGTPKGELSKTLPGPSKPSLKQVKNLNTNKVSGVWPSLTSTLKTTASSSMSSNLSQHIAKHMEVNAWIKAQRQYEFMKNMGHQNQNEYHKDKQ
C. intestinalis	HKTRPLLNIRSDQTLQDIVYKLVPGLRSDEMKRRRRFWGENPESKKDFAIWRELSPEELGDADQFDVATFSSKVTLVLENLKRRNKKADDLREWSEIASSLSKRYLRTSQDLTVNHLHKFLRAKLNEPISTEIVMLCGENVLPPTYTLADVRDTFSPVDHLLHLTYCIFVPRSLKRKPPPQRVAEKVEVEAKSRKTVAKKSSFRKKSATPHLKAFFNSQISPPTTEKQQRPFLKPISDYRKQDEIESLREAEEQKLIEWAAARDTRAKLPLFEKLQLTTVQRAAAIKRAALYKLANQKKAKEKQEYINSAASSSVQKLPQKKLDSQNEQTKLKSTKNEYKVTAQVPKGTNSPRRNIKQGSNEQFPSTGRWLNKQNNNRTRPTRVKCYSVLNIPVDEAVRSKPPTPVVDPLCPPVVLKRSSADPDNEAPPTKMKPFVQRTANNEVPMLNLSEHGNNKVAPKQQQFQRNRRKPIHPTHHASAGRTDSPGTVLLQKIDTSKTQFSTTPTRPISREPDRQQAGFDTIRIQSPNNGKFILLSTEGMERGHSQSHPAGLSMKLHSQMQQNRSSNDPRKLDNQGTMLNTANQDSQNKSQFQTRINSQAHRAALDAVRNTMGKVLLQAERPRQMPLKRPILPKGVSKPIHTGVGSIPIRLPTNQTRNIFQNEQVIYPMNKTVASSAASTSQSKAPIRTQPKPSPKSLSNNELEQLKKLREQQDFLNKLTEAAAINQLANRKKSSTDNSPQTSNQSPSTFRIKQHLSSQDNNRGRPPVLQADARVIPSPRFSQPSPAPRFQKPITQKPFERINSTSTRGRFQNSAPVSSPSLNRNSFPMRPTPQPNSNSHVNKQAQFTRLASGVQINSRPQQPSAKTLLLQSRAQDRPVGITPAEMQQRRQQYKTNPSTSIANNGRYNQQFGSRPPRFQQQQQQQHPLPVPRQFMLPKSNTNPRQQTFQLRSSPNASMNRHPIATNQRTRQVPSIIRRSFEKMNPRPKSVTPTNRGQIQARSNLHSRQAHVRVRSTSHEVLAPTATPPAGTKSPWSSRGYPLPAVPTAHPSEYATQHEIHKPPLAHQQPSSNNFARASTSIKTNALPLSDMQPLELTAKKNTNSTKQTIDDGAGQSNSDQPLCLVMKK

D. melanogaster	RDAPMRFYYRVYESPQPLVKPAPRRVLPLKLEKQERENQEQQLAVEVASSKVEPVSLAEDQKAEASIKVEGEESTREIVKEVIKDVAATPPTETLKLVINRNMLDKREKSHSPQLSSKSSSKSSPCTPVSSPSEPNIKLKIDLSKQNSVTIIDMSDPERREIVKPLKPEKESRSKKKDKDGSPKSSSSSSSSSSGERKRKSPSPLTVPPLTIRTERIMSPSGVSTLSPRVTSGAFSEDPKSEFLKSFALKPIKVKVESPERTLNNRAITPPSPSVQQSASPKSKGNNLDDSILMKPPSCMPPKSIASSKRKSKEPVKAVSKKQKLSPPLPTVDFKIRLPVTNGNSSGTASPKIEKPLMPPPAKPPMLAPRKLQPSAQFAPPPSPIHHHAGVQMSAPGNRTPIAKRYQPILPKASRPNPFANIPNDVNRLLKDAGTEIKSIGGGSVENNSNSAQKPHLYGPKGETKMGPPALPATTPSQGNKNVGKQAGNLPMSAPPNKGNSSNNYLNLALFNSNKCKGKEAPPGCRTPMYTPNSPIYSPSSPQYVPSYNIPTMPTYKYTPKPTPNSGSGNGGSGSYLQNMLGGGNGGSLGGLFPSPPTKSDQNTNPAQGGGGSSSATQSGGNNGIVNNNIYMPNEDAPEKQQVKVKSLLNSCNINIPSSLSITISRDNGDSSSPNNGQHPKHKSPVNNYIEIVKLPDQPQDQVQAAKEAQKRQSPPAAVPGHLAAKLPPPPPSKAIPSPQHLVSRMTPPQLPKVATPPPPSSPRVITPPKTSPPANAAKVTPLKPVLTPTQVDKKTPSPEKRTAAQMGSHSPTASENKSPKGGAAGVANSTGGTQNGDPAAKKFRPILPRQNGMPELAPKLPTLAPFVGFNPLQNPAAGKKVPPSKKSPNAGAAAHQSGQQKLVNGGQPQSAQQKTSPPAQKNQQQVKKVSKNPTPPPPSLPAVGKMMPHPVMHSQNAPLSIASSASAAAVASGQLDLSNFLKENLRRVHAAQAAQAAQVAAAANQSNMMYNLAQMGHMTPAMYNYQQAYFREQLSRMQRVGNEVFNDYLQKLKTAAATGGGGPVEGELKPMLPTVTLPSPGATPPAASPKTSPLPAGKLTAAATAPQTKGNSSSGAANARQQTAATGNNGATVPAASLPPATKSK

D. pseudoobscura	RDAPMRFYFRVYESPQRQVKPPPRRMLPAPLKVVKQEPTPAPEAPKVEQTSPTAAPVSPPASIKQELQEEIRVPSEQPLKLIINRNMLEKREKSHSPQSSKSSAKSNHHTPTTPSSSSSSSSCPSPSGELNIKLKIDLSKHNSVTIINMSDPERKEIVKPLKPEKESRSKNRSKDKDGSPKSSSSSSERKRKSPSPLTVPPLTIRTERILSPSGVSTLSPRCVASSSCHEDPKSEFLKSFALTPIKVKVESPERSPSSHRAPTPPKTTASGSGSGSHSHHSGRSKGTLEDRELMRPPAGMAPKSIASSKRKSKEPVKAVSKKPKLSPPLPREDFKIRLPATNSHSHPPPAPTTPPPFVGSLEKLMPPPPKPPMLASRKPQLAAQFAPPSPHHPGMQMAAPGNRTPIAKRYHPILPKAARPNPFANIPNDVNRLLKDAGTEIKSIGGSTSASSAKSHVYGPKADSKMGPPPPPAGAAAPHAARHTSGGQGKTGGNNQPQPHPAPSSNGSQNKAANNYLNLALFNASKSKGREAPPGCRTPMYTPNSPIYSPSSPQYVPNYNIPTMPTYKYTPKPSQATAGSYLQSMLGGGGGASGSGGGSLFPSPPTKADQNTNPAGAAPSSGHAFQRGASPSHEDAPEKQQVKVKSLLNSCNINIPSSLSITISRDNGDSSSASNGSHPKHKSPVNNYIEIVKLPDQPQDQGQKSAASVTEAQKRQSPPAPAPGRTPPPQLPAVAAPAPAAAAMRLTQPPPSKAIPSPQHLMSRMTPPQLPQTAPPPSSPSTATRGITPPKISPPASGKGTPLKTVLTPSQADSKKTPSPEKRSAAQMGSHSPTASENKSPKLAGQSAPGSATPNGDPAAKKFRPILPRQNAQIPDMAAKLPSLAPAFNFSQPQSQVQTGAKKVPTSKKSPNGGAAVFLPPPPKLPNGSHPAQKPSPPPKSQQTSGKKANKNPTPPPSSSAALGGGVQGNMGKLMPHPGLPGLNAPLSIASSAAAAAGQMDLNNFIKENLIRAQVAQAAQAAQAAQAAANQSNILYNFAQIGHMSPAMYNYQQAVFMEQLTRMQRAGNEAFNDYLQKLKNAANGQAGDGDHKPIMPMLPTVNLPSPSSATSAASPKTSALPNGKLTAAATAPSSHTPSSLAKAGSGASPRQQTAATPAAPLVAATKSK

D. virilis	RDAPMRFYFRVYESPQPLMKPALPMTLPAKPQVKQELATPVVTPTSSPPAAAALVKSPSPSPPAVAAAAATAQPLARIKLEQPQDEFRIAPKLPSPTEQSLKLFINRNQLEKQEKLPHERHHHHHHHHHSPKAAKSSPTTPTANSKFPPTGNYNKEEPNIKLKIDLSKQNSVTIINMSDPERKEIVKPLKPEKESRSKSKKDKDGSPKSSSSSSSSSSSSTSSSTSSERKRKSPSPLTVPPLTIRTERILSPNGVSTVLSPRVTSGACLEDPKSEFLKSFALTPIKVKLESPEKPASHAAPPAIAPPAAKSKTHLDDSLLMKPPSAMPPKSIASSKRKSKEPVKAVSKKPKLSPPLPREDFKIRLPAPNSCPSPPPPMLAAPVEKPLMPPPPAKPLPVPAARKAQLPHSPYPVHAPLPPHHQGMQMAAPGNRTPIAKRYQPILPKAARPNPFANIPSDVNRLLKDVGTEIKSIASQAKTHVYGPKMPEHKMGPPSAMHKPNNNSNNNHSNNNNNNNSNSNNNNKSNYLNLALFNASKSKGKEAPPGCRTPMYTPNSPIYSPSSPQYVSNYNIPTMPTYKYTPKPTTTNNNSNNSNNNNNSTTATTNASNYLQSMLNGTGAGGAGGGGLFPTPPTKTDQNTNPAAEDAPEKQQVKVKSLLNSCNINIPSSLSITISRDNGDASSPSSGGHAKHKSPVNNYIEIVKLPDQPAASAEQKEPTAAAKATPTPTPQPPVKLPAPPSKTIPSPQHLLARLTPPAAAATAAAVPAKTSPKATATAKPVLTPQQSDKKTPSPEKRAASQGSHSPNSSENKSPKSAQATSAAAGASGCATPNGGESAAKKFRPILPRQNATNGGATTEPKLLPQQPVGYNFAANLPNSKKVPASKKSPGAGGAIGGGGGGGSGTPAKLAHANGSSQALCKAGAKHKLATPTPPAALGSSLKFMGPPTGHAHPHLPNPNAPLSIASSANQLDLSNFLKDNLRAQAAAQVAQAAAAANQSNLLYNFAPAIYNYQQAYLMDQLSRMQRAGNEVFNDYLQKLKSAAIAGGEGAAGEHRQPVMPMLPTVTLPTAASQPIAASPKTSPHAAAHKLTPAATPTPTLAKSNSSSSSGGGGGSGSARPQAAATSNNALAKSK

D. willistoni	RDAPMRFYFRVYETQQQPALPPPPTSTSIISAPGATTPRRILPLKLEKREISPPAVVIKAPSPSPPSHPPASSPTPPTQKEHVPNAAVVTTPTVSPSHAPRIKQEKQEEFRIATKQLASPTEPLKLVINRTHYSPLSIASSASKMSSKSSHHHHNQPPTATTAAPSSPAAPQPPSSPKDEPNIKLKIDLSKQNSVTIINMNDPERKEIVKPLKPEKESRSKSKKDKDGSPKTSPSSSSSNGERKRKSPSPLTVPPLTIRTERILSPNGVSTLSPRITSGGLSEDPKSAFLKSFALTPIKVKVESPEKMLASTPSKLMKTNVDDSLLMKPPSSMPPKSIASSKRKSKEPVKAITKKPKLSPPLPREDFKIRLPGSPAAKSDDKPLMPPPMKPPMIAPRKQQQQQQQQQQQQLQQQSSGQFPVPSSPLFQGMQMAAPGNRTPIAKRYQPILPKAARPNPFANIPNDVNRLLKDAGTEIKSINNSSHANNKPHVYGPKTDAKMGPPPAPGRHVTNGGIAKPTNNHNNNQGSTSSSTSSSSSAAAGAAGLNSKSNNYLNLALFNASKSKGKEAPPGCRTPMYTPNSPIYSPSSPQYVPNYNIPTMPTYKYTPKPSTQASNYLQNILGSSSGAAAGNGGGLSAGLFPSPPTKADQNTNPAKSNTPPAAAAGASFNQRSASPNEDAPEKQQVKVKSLLNSCNINIPSSLSITISRDNGDSSSASNGAHPKHKSPVNNYIEIVKLPDQTPNAESQKRLSPPAPPISTASTGVTSSAPAPSVMKLPPAPPSKTIPSPQHLMSRLTPPQLPPVAAANPPRVITPPKTSPTNVKATPMKPVLTPTQGGDKKTPSPEKRSANHSPTASENKSPKSAGGSSSSSSSTSNGDPAAKKFRPILPRQNALPELAPKYSPQTNQQQQQQAHNVSAAVNNNNNSNNNNNNNVNKSKVQPSKKSPNTPNAAASGQKMSPPGQKQSPTLKKTAKNSTSTPPSQNKLMPHPGLAPLSIASSAAAAGQLDLSNFLKENLIRAQAAQVAAANQSNLLFNFAQIGQLPAMYNYQQACFMEHLSRMQRAGNEVFNDYLQKLKTAAGANGNGNGNVDVDYKPPVMPMLPTVTLPSLSNPGTAAASPKTSPLPTGKLTAAATPALALGAKGGNAASPRQQTAATSNGNRPSTPHSTTATPPPPPAAAAAKSK

I. scapularis	RDVPLQLFYRISENVARAPGPLPTGVAVMTAPPGLAGDGATREGAPQQQAQQQGPPRGDSSKGPAFLKDSVNFPGSSRSCKDAGTTPEGTDSTAKPLTTEEASDTKPACSDGTPGRAEPKVQSKADVAPTSTPPLDKSVPDLKAPTTALKAATKAKTPAPVQAKDVTEPVPENCRQLAKAKPYCDQTCTVPKSPVHGAEPGSAERSIPTGAAKCEKDSPCRPTAMPTEAEKSKTGVPPIRLKVPAECLELAKAHVHDVPAAELPRTAAKASSRADKEKALLQVGCASTPSGEDGSTARVEGASVTTIDRPPPVTLPNGAAKDLLKDLSEKLKVKGIVLELDPSSKRASLGSATAVESGVDPKLVNHTATPTIEDVVEAVSAIPEPVVQVSVAAETVLESAKLQTCFSRAADKARAKINALKAAASLKETSKAAVAEKEANEVVGLSVTLRANRQGAKGQLDCPVDKPAQKVASVAVSPPPVSPCEKKDNVLPSELPAATTSGATANRPVATAVPTYMTLSKSHPSLFHSSPRKRGRPRLATVNSLNEEIERAHMMAKRQQATAEKPKPAIPVITSLRIKPIPPPPPETPPPVDRAATGAEVPESERLQRRGSQSEVSEEKSDAEDSSGRRKSRRRRGPMELRNVVTQLKDMTLEKEQQAATQEPLRNLPGGPPSPAPAAAIPEKITLRVTRDEKSNLKVEKQLRPAAAAVVAETLHDSGFCEDVVAEGSRSPASEVKPKIEAATKTVRPTAPREPPLPSPCRKPDVAAAAHHGSNKKDMRKSKRRSVEDWVNEQSKWVRAHKAAAAVGGGDATPSPKAAKEHEDERPKRRPSLEEPPPAKGQQRRGRKRTNPVKITKPDPVVDAQQEKAASGSPPLTGGAPEKKGTTTATPPLVEKAEPSEAPARCPAGGPREPGKSRRELESPPKKLSELVIPRYIPNPATSIPLTITHARNKRLRETDKAPESASTCRRRIPSTPKQTGGERTRRRASSSRRR
D. pulex2	EREPLRLWYRISPTIKEEEPIQRKTTPPAEEVKVKGNQRASLVDITSKRSWQCNGKEERRTKRRKRSSAEKVAEKIQRMTPEASSVDSLVVPVAKTIFDEARTAEFSCSTPIPSAGSTAPPLETDSGESNKIHNWLPKAVFDLDDRALFAKRLQRITNPSEFRPEEVKEEEPQVDKTDKEVVVPTTKEESIESPDPLAPLRICVTPDPTGATSGGPDDEIHDSIGALDLSGSKGDSSDVSSPLSAGSCRSSASPVGSTSKMGPHPYFMTPSAVYHHVQQQDPAQSMAVLEAAAQSSTCSANNTKNLSSDDVKKPPIWDLFHQHIRPSTAHSSQQAQSLLDLLKSNPPLIFRGNNNKKKSKKTPAGKKFPSSSPNSAKGEFDPAFLYKVVT

D. pulex1	RKGPMRLKYRIYQRLQSSSPLTNGTNGSEEAPQKIKEEAVAEDKKMTNEVQ LEISECGVMSVPNVDAKNGIPEETQPTIPNTSSEKPEEVKAPSPKPEASSE APLVCVTVSTDISADLNGETSVNCSPSVDVKTGDSVEVKTTITVNENPTIT TGTTSSDTASNPPASHLTSSARNKIPSSTGHKTLKPPSSSWNQNVNRVGTK RPSSSVACNDGGSLNLANAEQTLPTPAKRATPSSPLKTPRFFKVRNASQPT DTNGGTCKATGTSIASVTESPVQEVAVNLTKASSSPKKPTDKERSSREGKE GKAPSPRPDIGSNPIRPYSVPVPSQSKRQPSPNLAAEDAAARLRHLLNPIS SAASTTTSPSSGDSSVQQLRFPAAAWLNLARGVPNRPVPLALGPGSPFNNR PPLSPAHFISSFIASHPHYPYLSPMGLPPAPDSKKSLPSSTASSSSSSIPS PQMHKSISSRTSNSFPTPTFNLNTLQQCTYPSSLSPLLPNLPRELVGSFYH SSYVPRPFMSARGGPLSVSPKSVSSSSSVGSNSSGGGFHPSMPPTVTTTTS NSSSSAAGRKSSPGLRPTVPRRNVAPPPPLVPIGTPSSVRSPPTLLPIKDI VEKESSCAKSTSPAASNCTSVVESCVPQMEEEIVKSGPVSKSTDGKPVDST SENQHSSAKENGKVIGDADSGKANTPAASPLEGSINKTTTDNANVVLENKS ESKVEIAAPAPS
D. pulex1 Act1	RKGPMDLEYDIYQDLQSSSPLTNGTNGSEEAPQEIEEEAVAEDEEMTNEVQ LEISECGVMSVPNVRAKNGIPKRTQPTIPNTSSKKPKKVKAPSPKPKASSK APLVCVTVSTDISAELNGETSVNCSPSVNVETGKSVKVDTTITVNENPTIT TGTTSSKTASNPPASHLTSSARNKIPSSTGHKTLKPPSSSWNQNVNRVGTK RPSSSVACNRGGSLNLANARQTLPTPAKRATPSSPLKTPRFFKVKRASQPT RTNGGTCDATGTSIASVTESPVQDVAVNLTEASSSPDEPTSSNRSSRKGKK GKAPSPRPKIGSNPIRPYSVPVPSQSEEQPSPNLAAEDAAAKLEHLLNPIS SAASTTTSPSSGRSSVQQLDFPAAAWLNLARGVPNRPVPLALGPGSPFNNE PPLSPAHFISSFIASHPHYPYLSPMGLPPAPRSKKSLPKKTARRSSSSIPS PQMHESISSRTSNSFPTPTFNLNTLQQCTYPSSLSPLLPNLPSSLVGSFYH SSYVPRPFMSARGGPLSVSPESVSSSSSVGSNSSGGGFHPSMPPTVTTTTS NSSSSAAGRKSSPGLRPTVPDEEVAPPPPLVPIGTPSSVRSPPTLLPIKDI VKKKSSCAKSTSPAASNCTSVVKSCVPQMSEEIVRSGPVSKSTDGKPVDST SENQHSSAKKNGKVIGRADSGEANTPAASPLDGSINDTTTDNANVVLENDS EEEVEIAAPAPS

D. pulex1 Act2	RKGPMDLEYDIYQDLQSSSPLTNGTNGSEEAPQEIEEEAVAEDEEMTNEVQLEISECGVMSVPNVKAKNGIPKRTQPTIPNTSSKKPKKVEAPSPKPEASSKAPLVCVTVSTKISAKLNGKTSVNCSPSVDVRTGKSVEVKTTITVNKNPTITTGTTSSDTASNPPASHLTSSARNKIPSSTGHETLKPPSSSWNQNVNRVGTEEPSSSVACNDGGSLNLANARQTLPTPAKRATPSSPLETPRFFKVENASQPTRTNGGTCKATGTSIASVTESPVQRVAVNLTEASSSPRKPTSSNDSSRKGEDGKAPSPRPKIGSNPIDPYSVPVPSQSRKQPSPNLAAKKAAAKLEHLLNPISSAASTTTSPSSGRSSVQQLDFPAAAWLNLARGVPNRPVPLALGPGSPFNNRPPLSPAHFISSFIASHPHYPYLSPMGLPPAPRSKKSLPEETARRSSSSIPSPQMHESISSRTSNSFPTPTFNLNTLQQCTYPSSLSPLLPNLPSSLVGSFYHSSYVPRPFMSARGGPLSVSPESVSSSSSVGSNSSGGGFHPSMPPTVTTTTSNSSSSAAGKKSSPGLRPTVPKEEVAPPPPLVPIGTPSSVRSPPTLLPIKRIVDKRSSCADSTSPAASNCTSVVKSCVPQMSKKIVRSGPVSKSTDGKPVDSTSENQHSSAKKNGKVIGDADSGEANTPAASPLRGSINDTTTDNANVVLENDSEEEVEIAAPAPS

Table S7: **Sequences used for the PSC proteins.**