Math Notes: Stable Location Parameters and Pairwise Stability Identification for Electrostatic Surface Potentials in Calreticulin Diversity

Jeff Cromwell, PhD

The Mathematical Learning Space Research Portfolio: Math Note 4

Abstract:

Chaperone proteins in the endoplasmic reticulum such as calreticulin, bind to the three glucose residues on the core N-linked glycan. Given that a protein repeatedly fails to properly fold, it is excreted from the endoplasmic reticulum and degraded by cytoplasmic proteases. Electrostatic surface potentials (ESPs) of aromatic rings in (a) tryptophan, (b) tyrosine, (c) phenylalanine, and (d) histidine offer the possibilities that electronic effects play a role in the binding to glycans. For a collection of protein crystals a collection of calreticulin with secondary structures from the protein crystals such as 1hhn 1k91 1k9c 2clr 3dow 3o0v 3o0w 3o0x 3pos 3pow 3rg0 5hcf 5lk5 5v90 6eny a sample of 50 sequences from 3000 proteins across species is the basis for a comparison of similarity with a feature matrix of Stability Index Binding Potential ALiphatic f.1 CpH5 CpH7 CpH9 scales. Maximum Likelihood estimations for sequences across species and with respect to secondary structures are provided and stable location parameter and a potential unstable scale parameter are examined for amino pairs.

Keywords: Protein Structure Function Protein-protein Interactions Small Molecules Peptides Chemical Biology glycobiology aromatic rings

1 Introduction

Glycobiology is the study of the (1) structure, (2) biosynthesis, and (3) biology of saccharides (sugar chains or glycans) where new anti-cancer drugs have both possibilities and potentials in glycobiology. Anti-cancer drugs with different algorithmic steps together with anti-inflammatory and anti-infection drugs have contemporary clinical trials. [1][2].

The comparison of electrostatic surface potentials (ESPs) of aromatic rings in (a) tryptophan, (b) tyrosine, (c) phenylalanine, and (d) histidine may have electronic effects also play a role in the binding to glycans [3]

N-linked glycans have an extremely important in proper protein folding in eukaryotic cells. Chaperone proteins in the endoplasmic reticulum, such as calnexin and calreticulin, bind to the three glucose residues present on the core N-linked glycan and fold the protein that the glycan is attached to and when proper folding is achieved the three glucose residues are removed, and the glycan moves on to further processing reactions. If their is folding failure, the three glucose residues reattach and the protein re-associate with the chaperones in the form of a recycle. This cycle may repeat n times until a protein reaches its proper conformation, however, if the protein repeatedly fails to properly fold, it is removed from the endoplasmic reticulum and degraded by cytoplasmic proteases. [4]

2 Results

Consider a collection of calreticulin with secondary structures from the protein crystals such as 1hhn 1k91 1k9c 2clr 3dow 3o0v 3o0w 3o0x 3pos 3pow 3rg0 5hcf 5lk5 5v90 6eny in Table 1.

	Accession	Name
1	1hhn	Calreticulin P-domain
2	1k91	Solution Structure of Calreticulin P-domain subdomain (residues 221-256)
3	1k9c	Solution Structure of Calreticulin P-domain subdomain (residues 189-261)
4	2clr	THREE DIMENSIONAL STRUCTURE OF A PEPTIDE EXTENDING OUT ONE END OF A CLASS I MHC BINDING SITE
5	3dow	Complex structure of GABA type A receptor associated protein and its binding epitope on calreticulin
6	3o0v	Crystal structure of the calreticulin lectin domain
7	3o0w	Structural basis of carbohydrate recognition by calreticulin
8	3o0x	Structural basis of carbohydrate recognition by calreticulin
9	3pos	Crystal structure of the globular domain of human calreticulin
10	3pow	Crystal structure of the globular domain of human calreticulin
11	3rg0	Structural and functional relationships between the lectin and arm domains of calreticulin
12	5hcf	T. cruzi calreticulin globular domain
13	5lk5	Crystal structure of the globular domain of human calreticulin mutant D71K
14	5v90	Crystal structure of ERp29 D-domain in complex with the P-domain of calreticulin
15	6eny	Structure of the human PLC editing module

Table 2 has A protein whose instability index is smaller than 40 is predicted as stable, a value above 40 predicts that the protein may be unstable. A protein have high binding potential if the index value is higher than 2.48. The relative volume occupied by aliphatic side chains (Alanine, Valine, Isoleucine, and Leucine) is considered to be a positive factor for the increase of thermostability of globular proteins. net charge of a protein sequence based on the Henderson-Hasselbalch equation for pH levels of 5 7 and 9. f.1 is the crosscovariance based on one of the sequence S at lag 1 is given by f.1 with property1 = Hydrophobicity based on the KyteDoolittle Scale and property2 =Hydrophobicity based on the Eisenberg scale. [1001] [500]

*Mathematical Learning Space Research Portfolio
Email address: http://mathlearningspace.weebly.com/ (Jeff Cromwell, PhD)

	Protein	Stability Index	Binding Potential	ALiphatic	f.1	CpH5	CpH7	CpH9
1	1hhn	34.38	3.30	38.71	-1.79	-12.25	-16.52	-18.12
2	1k91	42.89	2.84	31.62	-1.79	-5.84	-7.77	-8.85
3	1k9c	27.37	3.61	30.41	-1.98	-8.73	-11.76	-12.96
4	2clr	34.76	2.30	63.32	-0.74	25.38	-10.15	-31.25
5	3dow	45.97	1.93	78.10	-0.54	5.81	0.95	-1.48
6	3o0v	37.99	1.90	65.62	-0.63	-6.34	-13.86	-18.40
7	3o0w	38.33	1.88	67.98	-0.62	-5.45	-12.86	-17.40
8	3o0x	40.26	1.95	66.01	-0.66	-14.49	-29.68	-38.09
9	3pos	39.95	1.95	64.42	-0.65	-22.79	-45.60	-59.98
10	3pow	39.67	2.04	64.42	-0.68	-7.15	-14.88	-20.13
11	3rg0	38.85	2.09	65.45	-0.77	-7.37	-16.61	-21.64
12	5hcf	28.13	1.89	76.84	-0.57	4.40	-71.30	-102.91
13	5lk5	40.16	1.95	64.99	-0.66	-62.87	-138.63	-185.04
14	5v90	52.23	2.12	78.03	-0.79	-6.35	-11.74	-13.45
15	6eny	44.81	2.01	68.41	-0.66	-5.89	-60.35	-91.82

Table 1: A protein whose instability index is smaller than 40 is predicted as stable, a value above 40 predicts that the protein may be unstable. A protein have high binding potential if the index value is higher than 2.48. The relative volume occupied by aliphatic side chains (Alanine, Valine, Isoleucine, and Leucine) is a positive factor for the increase of thermostability of globular proteins. net charge of a protein sequence based on the Henderson-Hasselbalch equation for pH levels of 5 7 and 9. f.1 is the crosscovariance based on one of the sequence S at lag 1 is given by f.1 with property1 = Hydrophobicity based on the KyteDoolittle Scale and property2 =Hydrophobicity based on the Eisenberg scale.

Table 2A has the FASGAI vectors (Factor Analysis Scales of Generalized Amino Acid Information) for the 50 calreticulin similarities with differences in the alpha turns with respect to the sign change. The FASGAI vectors is a set of amino acid descriptors, that reflects hydrophobicity, alpha and turn propensities, bulky properties, compositional characteristics, local flexibility, and electronic properties to represent the sequence structural features of peptides or protein motifs. [1001]

	Hydrophobic	Alpha Turns	Bulky	Compositional Characteristic	Local Flexibility	Electronic
1	-0.51	0.07	-0.13	0.23	0.13	-0.34
2	-0.39	0.03	-0.11	0.31	0.14	-0.31
3	-0.40	-0.04	-0.04	0.24	0.07	-0.08
4	-0.46	0.00	-0.10	0.23	0.09	-0.29
5	-0.41	0.06	-0.05	0.23	0.08	-0.16
6	-0.40	-0.10	-0.12	0.26	0.10	-0.28
7	-0.44	0.00	-0.12	0.23	0.13	-0.29
8	-0.40	-0.01	-0.04	0.23	0.07	-0.10
9	-0.31	-0.06	-0.09	0.27	0.08	-0.21
10	-0.46	0.07	-0.09	0.26	0.11	-0.28
11	-0.51	0.07	-0.13	0.24	0.13	-0.34
12	-0.38	-0.03	-0.01	0.24	0.12	-0.08
13	-0.42	-0.04	-0.11	0.26	0.11	-0.29
14	-0.39	-0.03	-0.12	0.29	0.07	-0.31
15	-0.39	-0.01	-0.01	0.24	0.13	-0.08
16	-0.39	-0.03	-0.04	0.24	0.05	-0.05
17	-0.51	0.04	-0.12	0.22	0.12	-0.32
18	-0.30	0.01	-0.05	0.26	-0.04	-0.09
19	-0.40	0.06	-0.14	0.29	0.07	-0.37
20	-0.39	-0.01	-0.00	0.24	0.15	-0.07
21	-0.49	0.07	-0.11	0.24	0.09	-0.29
22	-0.41	-0.03	-0.11	0.21	0.08	-0.31
23	-0.45	-0.02	-0.11	0.24	0.07	-0.31
24	-0.42	-0.00	-0.11	0.26	0.19	-0.31
25	-0.46	0.07	-0.07	0.20	0.08	-0.26
26	-0.46	-0.03	-0.11	0.23	0.11	-0.31
27	-0.43	0.03	-0.12	0.29	0.07	-0.22
28	-0.33	-0.09	-0.06	0.26	0.07	-0.01
29	-0.45	0.07	-0.16	0.27	0.14	-0.39
30	-0.38	-0.03	-0.01	0.24	0.10	-0.09
31	-0.38	-0.01	-0.02	0.23	0.08	-0.07
32	-0.41	0.06	-0.05	0.23	0.09	-0.15
33	-0.39	0.03	-0.11	0.27	0.10	-0.29
34	-0.43	0.02	-0.09	0.22	0.13	-0.30
35	-0.41	-0.02	-0.11	0.29	0.14	-0.32
36	-0.35	-0.03	-0.01	0.23	0.09	-0.10
37	-0.40	-0.04	-0.15	0.26	0.15	-0.28
38	-0.47	0.03	-0.15	0.28	0.08	-0.28
39	-0.42	0.00	-0.10	0.25	0.17	-0.29
40	-0.40	-0.01	-0.13	0.33	0.10	-0.28
41	-0.37	-0.03	-0.08	0.29	0.10	-0.24
42	-0.52	0.05	-0.11	0.24	0.10	-0.30
43	-0.44	0.04	-0.15	0.29	0.10	-0.30
44	-0.45	0.07	-0.09	0.22	0.10	-0.30
45	-0.39	-0.07	-0.07	0.27	0.13	-0.24
46	-0.29	-0.02	-0.03	0.20	0.13	-0.20
47	-0.40	0.00	-0.10	0.28	0.13	-0.30
48	-0.40	-0.00	-0.03	0.22	0.08	-0.09
49	-0.47	0.07	-0.12	0.27	0.08	-0.29
50	-0.44	0.12	-0.10	0.35	0.06	-0.22

Table 3 has the molecular properties of Table 2 for a sample of 50 of calreticulin from an N=3000.[1001] [500]

	Protein	Stability Index	Binding Potential	ALiphatic	f.1	CpH5	CpH7	CpH9
1	A0A026WU53	35.13	2.52	60.22	-1.05	-33.12	-48.96	-57.86
2	L8ICK2	47.05	2.52	59.88	-1.08	-38.34	-54.10	-60.66
3	A0A6J2W5S9	55.18	2.91	52.57	-1.28	-39.12	-55.33	-61.84
4	A0A670KG82	43.78	2.61	65.00	-1.03	13.72	2.32	-5.40
5	A0A0D9WX65	33.56	2.45	59.47	-1.05	-26.91	-42.15	-49.90
6	A0A2Y9MA80	48.57	2.55	59.88	-1.09	-39.97	-56.10	-62.66
7	A0A484L333	46.96	2.66	67.85	-1.02	9.34	-6.44	-15.06
8	A0A5A7TT24	40.51	2.73	59.10	-1.07	9.26	-6.44	-14.97
9	A0A5N5KZ91	38.91	2.58	55.45	-1.13	-38.90	-54.34	-60.83
10	A0A3Q2XB63	43.78	2.72	56.23	-1.16	-38.28	-54.14	-61.63
11	Q967Z2	32.58	2.34	60.23	-0.87	-28.06	-41.45	-49.98
12	F0JA25	37.48	2.46	60.66	-1.05	-27.74	-44.72	-54.04
13	A0A3Q7GQZ4	36.05	2.28	68.16	-0.88	2.60	-8.65	-15.89
14	A0A178W4E4	35.30	2.32	61.01	-0.97	-34.84	-48.62	-55.66
15	A0A5N3W2L9	45.49	2.55	59.88	-1.09	-42.13	-57.34	-63.66
16	A0A4S2KD39	34.64	2.32	63.16	-0.89	-19.29	-36.46	-45.40
17	A0A6P3E5I0	48.14	2.54	59.88	-1.08	-41.13	-56.34	-62.64
18	A0A5J5BA79	43.28	2.87	64.20	-1.06	6.64	-6.43	-14.79
19	Q96L12	47.66	2.27	66.56	-0.84	13.28	-2.74	-11.49
20	A0A1I8EYA7	33.18	2.73	55.75	-1.25	-16.91	-36.17	-44.85
21	A4I765	26.68	1.81	70.23	-0.62	-17.40	-32.41	-39.55
22	Q8WRU9	35.63	2.73	57.33	-1.19	-21.11	-38.68	-47.20
23	A0A2K2AA62	27.97	2.44	59.04	-1.07	-35.80	-49.59	-55.86
24	A1YB06	37.93	2.84	52.15	-1.22	-29.69	-45.37	-52.73
25	Q5MCL9	33.45	2.33	61.90	-0.97	-26.31	-41.91	-49.68
26	A0A673CIF2	26.70	1.91	70.05	-0.71	-10.96	-23.24	-31.99
27	A0A398APB3	31.28	2.30	61.25	-0.96	-30.02	-44.38	-51.69
28	A0A287SV63	37.70	2.15	64.76	-0.87	-17.66	-29.64	-36.63
29	M3VUU7	47.76	2.61	58.71	-1.12	-41.02	-56.34	-62.66
30	A0A673J4M0	42.50	2.70	56.03	-1.19	-35.78	-52.89	-60.67
31	A0A4S4ERW7	27.80	2.20	64.04	-0.86	-34.73	-50.14	-57.69
32	Q17MI1	29.86	2.40	62.51	-0.95	-32.74	-49.69	-58.14
33	A0A5J5DB49	37.99	2.50	57.79	-1.08	-38.83	-54.34	-60.83
34	I1CLI3	31.23	2.13	67.70	-0.82	-34.19	-46.05	-51.25
35	Q6R5P2	28.35	2.81	56.34	-1.29	-12.34	-28.64	-36.08
36	A0A3Q3X6R3	52.10	2.95	54.06	-1.29	-45.12	-61.36	-68.54
37	A0A6P4CKZ3	53.23	2.58	60.75	-1.03	7.10	-6.17	-14.07
38	I1LJN2	47.93	2.63	62.86	-1.03	9.77	-4.75	-14.50
39	A0A7K4S897	27.04	2.26	63.75	-0.88	-2.47	-14.41	-21.96
40	Q4DDX3	30.40	2.36	63.49	-0.93	-14.46	-32.15	-39.40
41	A0A3B4FJ43	58.06	2.93	55.77	-1.27	-47.05	-62.60	-69.33
42	A0A6A4RTT7	27.88	2.02	66.88	-0.73	-10.09	-24.02	-33.67
43	A0A671Y4D4	41.44	2.56	55.13	-1.11	-26.03	-42.63	-49.71
44	A0A4Z2HA26	25.30	1.89	68.34	-0.70	-14.21	-26.21	-34.21
45	C5Z0S1	39.97	2.95	58.51	-1.13	14.91	1.07	-6.97
46	A0A670XUY8	33.30	2.03	74.24	-0.76	25.61	14.81	7.09
47	A0A7K9BBV8	31.77	2.46	65.08	-0.93	8.17	-4.15	-11.07
48	A0A4U1ERY7	48.57	2.55	60.34	-1.09	-39.97	-56.10	-62.66
49	A0A3B6B861	33.23	2.21	63.76	-0.89	-18.57	-32.92	-40.46
50	A0A1D5R529	44.41	2.25	64.77	-0.84	11.35	-3.98	-12.47

For Table 3 the maximum likelihood estimation of stability provided a type 1 a 0.9816426 b 1.554865 location 25.29809 scale 33.56755 while for Table 2 the estimate was type 2 a 0.7458515 location 27.36622 scale 24.96957. Since a value above 40 predicts that the protein may be unstable both distribution have a stable location parameter and the potential of instability in the scale. Here the scale frequency of amino pairs for stablity classification is given by Figure 1 as

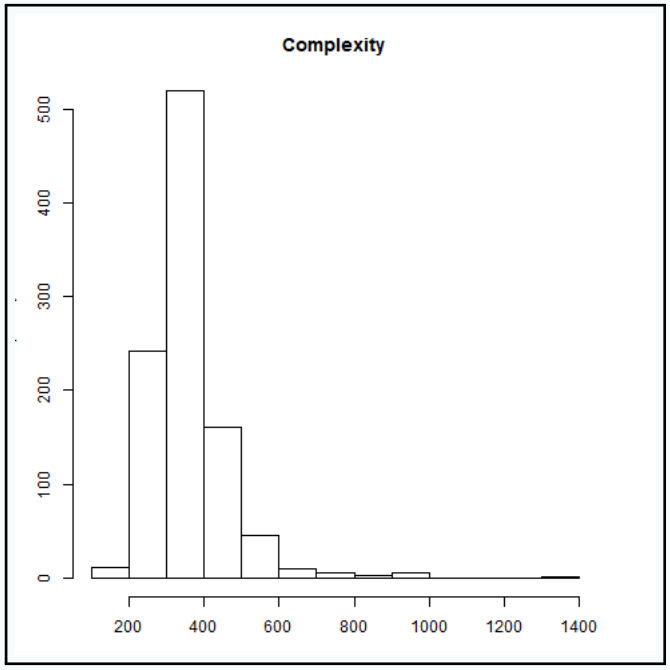

Figure 1: scale frequency of amino pairs for stablity classification [1001]

Table 4 has the different values for pairs of Amino Acids in the stability scale. Here pairwise combinations of (a) tryptophan, (b) tyrosine, (c)

phenylalanine, and (d) histidine are presented with AA1 descreasing and AA2 increasing values in the scale. [1001]

AA1	Scale	AA2	Scale
MH	58.28	YR	-15.91
RW	58.28	WT	-14.03
RR	58.28	WA	-14.03
MP	44.94	FK	-14.03
MS	44.94	NF	-14.03
HY	44.94	NG	-14.03
HI	44.94	DT	-14.03
YM	44.94	TW	-14.03
QS	44.94	TN	-14.03
NI	44.94	EW	-14.03
IE	44.94	VD	-14.03
RS	44.94	WG	-9.37
EC	44.94	HF	-9.37
SP	44.94	HG	-9.37
AC	44.94	YW	-9.37
CM	33.6	NW	-9.37
CH	33.6	WV	-7.49
CT	33.6	YT	-7.49
FY	33.6	YG	-7.49
KM	33.6	NT	-7.49
KR	33.6	IK	-7.49
EE	33.6	IV	-7.49
SC	33.6	RG	-7.49
LQ	33.6	DK	-7.49
WM	24.68	TG	-7.49

Figure 2 has Amino Pair (AP) Value Frequencies for (a) tryptophan, (b) tyrosine, (c) phenylalanine, and (d) histidine based on stability scale [1001] for electrostatic potentials.

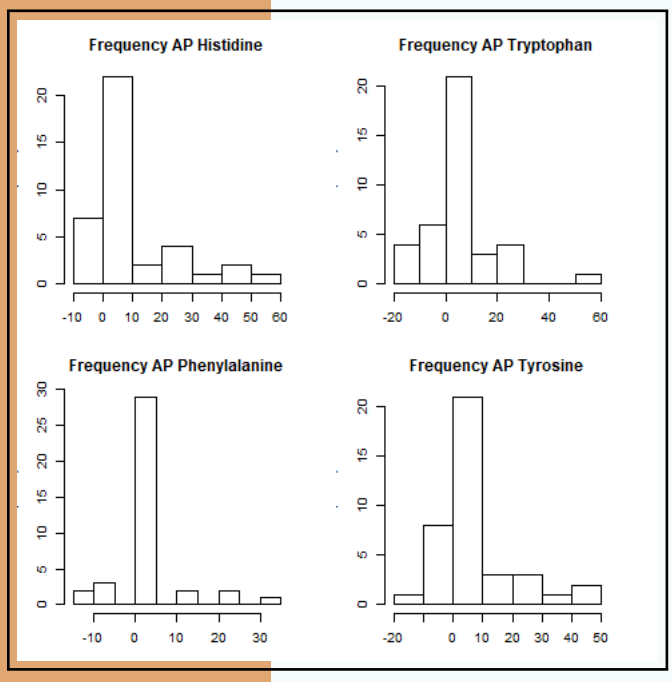

Figure 2: Amino Pair Value Frequencies for (a) tryptophan, (b) tyrosine, (c) phenylalanine, and (d) histidine [1001]

3 Conclusions

For a collection of protein crystals a collection of calreticulin with secondary structures from the protein crystals such as 1hhn 1k91 1k9c 2clr 3dow 3o0v 3o0w 3o0x 3pos 3pow 3rg0 5hcf 5lk5 5v90 6eny and a sample of 50 sequences from 3000 proteins across species a comparision was made by similarity with Stability Index Binding Potential ALiphatic f.1 CpH5 CpH7 CpH9 scales. Maximum Likelihood estimations for sequences across species and with respect to secondary structures as provided in Table 3 and Table 2 respectively provided stable location parameter and a potential unstable scale parameter. More research is needed to examine these differences in both the sample and population values of the sequence species library.

4 References

[1] Wikipedia contributors. Chemical biology. Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 2 Jan. 2021. Web. 18 Aug. 2021.

[2] Wikipedia contributors. Glycobiology. Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 24 May. 2021. Web. 18 Aug. 2021.

[3] Wikipedia contributors. Glycan-protein interactions. Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 13 Jul. 2021. Web. 18 Aug. 2021.

[4] Wikipedia contributors. Glycan. Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 7 May. 2021. Web. 18 Aug. 2021.

[5] Wikipedia contributors. Chaperone (protein). Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 8 May. 2021. Web. 18 Aug. 2021. MHRA style

[400] Kanehisa, Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K.; KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353-D361 (2017).

[401] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M.; KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457-D462 (2016).

[402] Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000).

[420] GeneCards Version 3: the human gene integrator

[430] DrugBank 5.0: a major update to the DrugBank database for 2018.

[440] COSMIC: the Catalogue Of Somatic Mutations In Cancer.

[450] Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders

[460] The ClinicalTrials.gov Results Database — Update and Key Issues

[470] PubChem Substance and Compound databases

[480] The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible

[490] MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search

[500] Blum M, Chang H, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, RichardsonL, Salazar GA, Williams L, Bork P, Bridge A, Gough J, Haft DH, Letunic I, Marchler-Bauer A, Mi H, Natale DA, Necci M, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A and Finn RD The InterPro protein families and domains database: 20 years on. Nucleic Acids Research, Nov 2020, (doi: 10.1093/nar/gkaa977)

[803] Cromwell, J. Mathematical Learning Space Research Portfolio Mathematical Learning Space Research Portfolio, http://mathlearningspace.weebly.com/ 8 3 2021. Web. 3 Aug. 2021.

[1000] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[1001] Osorio, D., Rondon-Villarreal, P. and Torres, R. Peptides: A package for data mining of antimicrobial peptides. The R Journal. 7(1), 4-14 (2015).