PHYSICOCHEMICAL n-GRAMS TOOL (PnGT) TUTORIAL

Introduction

Physicochemical n-Grams Tool is an open source standalone software for calculating physicochemical descriptors of protein developed using the Python scripting language.

The algorithms used in the development of this tool are based on the n-Grams curation techniques which are mainly used for natural language processing. Apart from language processing, now a days n-Grams modelling is also making a room in the computational biology and genomics.

The software currently calculates 33 physicochemical descriptors and the sequence length for given protein primary sequence. To find more specific information about our tool, please contact us.



Figure.1 Physicochemical Amino Acids Property Groups

Utility: The tool is able to assess the **physicochemical properties** of the protein in an easier way with bi-gram, tri-Grams, tetra-Grams assessment of protein primary sequence.

Assessment of the physicochemical n-Grams properties of the protein: With the help of this PnGT tool, the analysis of proteomics data in the term of physicochemical n-Grams statistics is easy-to-do with the only primary sequence as feeding.

Implementation in machine learning model development as feeding: The outcomes in terms of physicochemical feature vectors from the Physicochemical n-Grams Tool can be directly implemented for the development of statistical (PCA, Regression Analysis) or machine learning model like SVM, *k*NN *etc.*

Keywords: Proteomics science, physicochemical parameter, n-Grams, machine learning, statistical modelling.

Physicochemical n-Grams Descriptors and Calculations: Conservation of multiple physicochemical groups along the sequence in a sliding window of length n residue. For example, physicochemical 2-grams: Aliphatic++; where n equals to 2

[[if aa]] _i ∈S^* AND [[aa]] _(i+1)∈S^*

where, aai = ith residue of the protein

Table 1. Physicochemical property groups of amino acids

Sr. no.	Name of amino acid property group	Amino acids in the specific group (S*)							
1.	Tiny amino acids group	Ala, Cys, Gly, Ser, Thr							
2.	Small amino acids group	Ala, Cys, Asp, Gly, Asn, Pro, Ser, Thr, Val							
3.	Aliphatic amino acids group	Ile, Leu and Val							
4.	Nonpolar amino acid groups	Ala, Cys, Phe, Gly, Ile, Leu, Met, Pro, Val, Trp, Tyr							
5.	Aromatic amino acid group	Phe, His, Trp, Tyr							
6.	Polar amino acid group	Asp, Glu, His, Lys, Asn, Gln, Arg, Ser, Thr							
7.	Charged amino acid group	Asp, Glu, His, Arg, Lys							
8.	Basic amino acid group	His, Lys, Arg							
9.	Acidic amino acid group	Asp, Glu							
10.	Hydrophobic acid group	Ala, Cys, Phe, Ile, Leu, Met, Val, Trp, Tyr							
11.	Hydrophilic acid group	Asp, Glu, Lys, Asn, Gln							

Installation Steps for User

1. PnGT-Installer: To download "PnGT-Installer" go to <u>Download</u> tab and download PnGT-Installer to install Physicochemical n-Grams Tool.

2. Extract the zip file and open it, then locate "PnGT" as shown in Figure 2 below.

	en Share with 🔻 New folder			= • 🔳 (
🔆 Favorites	Name	Date modified	Туре	Size
📓 Recent Places ■ Desktop 🗉	🍌 PnGT - Data_files	1/7/2019 11:00 PM	File folder	
	👰 PnGT	1/7/2019 10:58 PM	Application	32,464 KE
📕 Downloads 🖳				
Documents				
Documents Music				

Figure.2 PnGT Installer

Home Page

1. The PnGT will open as shown in Figure 3 below.

Protocol Convince		N=+DimA
Rout About Yes Help		
स्ट्रीय औषधीय 1	शिक्षा एवं अनसंधान संस्थान (नाईपर), एस.ए.एस. नगर	Allow the second s
National Institute of Pha	armaceutical Education and Research (NIPER) S.A.S. Nagar	158 (201)
Sector 67, S.A.S. Nagar (Mo	shali)-160062, Punjab (India), Phone: +91-172-2214682-87, Pax:+91-172-2214692	The second secon
S.A.S. Nagar		
		Han Tao
	Name of Amino Acid Property Smap	Amino Acido in the Specific Dimag
1	Tiny Aming Abds Design	Ala, Cyn, Oy, Ser, The
2.	Tenall Arrine Acits Group	Ala, Cyn, Aop, Bly, Ann, Pin, Ser, The, Val
3.	Alphatic Arrino Acids Group	Be, Low, Yol
a.	Nonpolar Amino Acuto Droxp	Also, Cyra, Phys. Die, Leon, Mett, Prot. Val. Trp., Tyr
5.	Asomatic Ammo Acids Broup	Ptie, Hits, Trip, Tyr
б.	Volas Amino Acids Gasap	Ann, Diu, Hin, Lyn, Ann, Siln, Airg, Ser, Thr
7.	Charged Arwing Acids Group	Avp. Sky, Hill, Avg. Lyt.
8.	Basic Amino Acide Group	His, Lyt, Aug
b.	Acidic Assists Acids General	dags, Gily
98.	Hydropholia: Arenne Acide Group	Alts, Cara Piter, Ber, Leu, Met, Val. Tru, Tur-
Physicochemical n-Grams Descriptor		
Carried report for output right right a separate Densitie Fern	PETER Sergin Protein Separtitie	
Physicschereical 2-televe Decoulou Physicschaescal 2-Gran	a Deconstrain (Physicacherenaul 4-Grams Deconstrain)	

Figure.3 PnGT Home Page

Physicochemical n-Grams Descriptors Calculation

• The tool is leveraged with two input method: FASTA format and simple sequence pasting method, if the user doesn't have the primary sequence in FASTA format.



Figure.4A



Figure.4B

Population Page 1			ALC: UNK
Rout About Yes Help			
राष्ट्रीय औषधीय शिक्षा National Institute of Pharma SAS Nagar SAS Nagar (Nobul)-1	ा एवं अनुसंधान संस्थान (नाईपर), एस.ए.एस. नगर aceutical Education and Research (NIPER) S.A.S. Nagar 18082, Punjab (India), Phone: +91-172-2214882-87, Fax+91-172-2214882	* (<u>)</u>	
Gerregts Feature for Multiple Protein Sequence (Geraute Feature for Seq	yle Potose Jasueren		
Upload Sequence in Parts Format			(FF)
Uploaded Fasta File			I E I
		Calculate Physic	ochemicel 2-Grame Descriptor
Pile Rate (Neader): C.12	Save 243 Decknyr 01 93-2019 Pe07 - Date_Net 118 CDNIT_CDND FeSTil	Colculate Physic	ochemical 3- Gramo Descriptor
Disenter of Tequence 118		Calculate Physic	schenisal 4-Grame Descriptor
Provers Section 1 1000	WARMAN AND HERE ADDRESS AND ADDRESS AND ADDRESS AND ADDRESS AND ADDRESS ADDRES ADDRESS ADDRESS ADDR	THERE ADDRESS AND ADDRESS ADDR	Seventile
	REDY, DUBL, HANNAN, MARANAN, MARON, LAMANDA, JAHMITSA, CHARLER SHELTTAN, KAN KUTAKARANAN MANON, MANON HALL, COLL, LANNAN, MARON, KUTAN, KUTAN, MARON, LANNAN, KAN	Unit of each of the Could end Shareball up of the Could by Investor back of the Could by Unit of the Could by Could by Other Table Shareball and Unit of the Could by Could by Unit of the Could by Could by Unit of the Could by Could by Unit of the Could by Could by Could by Unit of the Could by Could by Could by Unit of the Could by Could by Could by Could by Unit of the Could by Could by Could by Could by Could by Could by Could by	Appard Fle
Physicochemical n-Grams Descriptor			
Generated Feature for Multiple Protein Sequence Generated Feature for S	Singly Protein Sequence		
Physicschemical 2-Grams Descriptor Physics here call 3-Grams Descriptor	nation Physicacherence of a Grants Description		

Figure.4C

Physiotherical 5 Gran Tax												OCD-HIM										
Rout Store the Help																						
राष्ट्रीय आवर्षीय विशेषा एव अनुसंधान सस्थान (नाइंपर), एस. ए.पस. नगर National Institute of Pharmaceutical Education and Research (NIPER) S.A.S. Nagar SAS. Nagar Sass to Volge Poten Separate Instational Page Instances																						
Centrato: Feature for Multiple Pro	tern Segurate Generate	nature for Sand	· Poter laneres																			
Upload Sequence in Parts Form	nat.																					
Uploaded Fasta Fi	le																					
	Pily Note (Note)	F) (55	er 241 Dickey (1-93-2079 Pu07 - De	in No. 111 COM	2.00074074				u u	icolan Physicochanica Icolan Physicochanica	- Dente Decrete - Dente Decrete - Dente Decrete - See File Append File - Hydrogekile *										
	Number of Yogan	ac) [226								6	éculiére Physicaichienis ca	oohamuud 2-50 area Deunistory cohamuud 2-50 area Deunistory cohamuud 2-50 area Deunistory cohamuud 2-50 area Deunistory Sever faie Append Tek T T 26 36 38 38 39 37 27										
Physicochemical n	Prenary Separate	1 1100 1100 1100 1100 1100 1100 1100 11	WHOMORE A LEATH 2017 COME - MINAN MILLIONE C, JAMAN VIGHT AND AND AND AND VIGHT AND		ERFLER TOLLED STOL	SPESION ANT ANTES CANAL PROVIDES COSTI COPULATION COSTI COPULATION COSTI COPULATION CONTRACTOR COSTI CONTRACTOR COSTI COSTI CONTRACTOR COSTI COSTI COSTI CONTRACTOR COSTI COSTI COS	DARSSALLATION THOLTHO-CHIMANE HILLIGHTLATIANCO HILLIGHTLA		CT 15550 (vol.4) PLOTATER OPENING PLOTATER OPE	SPORT = COLUM DATA NUME PARK PARK PARK PARK PARK PARK PARK PARK	F	See The Approx The										
Ceremeted Feature For Multiple P	tatan Sepience Grout	d Feature for S	ngin frotein Sequer	nt .	-																	
Ftyp:popersensel2-Grane Dep	oper Physiochemical	-Drama Desuma	tor Physicocharm	uel 4-Orama Deacharta	е.					1757	Charles with	1010										
Proton log-tifier	Sednesse reads	1997	Level .	manne	Nergelar	Lorde	Pater	Charged	and a	Appr	Hymanobe	Chronite III										
INTERPORT	214	÷	84	5 C		5		10		14	37	4										
THE REAL PROPERTY AND ADDRESS	116	- T	144	23	140	15	10	10	1													
CATHOLINET ARATH	444			23	72	14		47														
18 Percent and a second second	590	1.00									34	26										
or \$71268 BARL VEART	220	10	211	10	110	3 <u>4</u>	218	38	34	78	20.	26 38 81										
IEPZIZERPARI_YEAST	290 830 71.7	20 60	211. 80	39	177		238	18	25	13 13	20. 111 47	26 26 26 26										
IRF21288FARL_VEAST IRF29EF18FF81A_HUMAN IRF400008UMU SCHP0	200 838 112 238	20 18 21	41 211 80 84	39 11. 6	177 82 55	1	238 94	15 45 28	20 20	10 10 1 1	10. 111 47 23	26 38 81 38										
rpP21288/FAR1_YEAST rp[09EF18.0F911A_HUMAN rp[F40080[01M1_SCHP0 rpE4002228A8F.0AT	200 838 313 338 368	20 20 20 20	42 211 80 84 37	39 11 6	177 82 53 55	1 1 1 1	238 94 71 30	70 41 33	20 20 8	11 13 1 4 1	10. 111 47 23 25	26 38 81 38 30 37 2										
IEP21288/HARL_YEAST IEP098F188/PR1A_HUMAN IEP0987188/PR1A_HUMAN IEP098228(A88_JAAT IEP09228(A88_JAAT IEP0921187(CKE_CAETL	290 838 313 338 368 384	20 20 20 20 20	211 80 84 37 35	79 11. 6 2	177 62 53 55 30		73 238 94 71 38 61	20 20 20 20	20 20 8	11 13 1 1 1 1 1 1 1	20. 111 47 23 25 17	26 34 81 35 36 37 2 23										
representation of the second of the second s	290 838 313 338 368 368 368 364	20 50 18 20 28 18 29 28 28	211 80 84 37 35 45	39 11 6 6 2 2	177 82 55 55 30 44	1 8 1 1 1 1 2	71 228 94 71 95 61 45	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	11 29 20 8 8 8 20	11 13 1 4 3 7 3	20. 1111 47 223 25 17 17	//)) 第3日 第377 2 331 4										

Figure.4D

• The user can download the feature vector generated through physicochemical n-Grams assessment in CSV file format (by default). Figure 5 showing output descriptor in CSV file format.

B 5- (* - +				Physical Revision - Gaussi Feature - Gaust												1	Saxe BD			a i	×.	
the frees later	of a Pinje Lepende Colo Rome	no i citere	feries	· Veri	telp:	Q Tel	-														A 164	-
A New	Cales - 12	200		49-1-1	E.Wa	Tatt	20	aniana"			TRU:	100	100	XP=	The I		Same	ker - e	5- 6) I		
Ph Casy -				1000	2307						10 Million	1.00	120	붠	思	##	100		5. 3	5		
Paste N Louis at Faister	8 1 1	2-4-3	医贫穷	10.45	in the	pt.R.Ceriter	* 13	5 - 96 :	1.2.	-11	Carvoloured	forest a	1.00	POPL	Detete P		# Own	6 8	OT IX. THE			
				1000			100				Parenter of	-1828	- Selver	100	dial 1	-	1000	1.00	eter i pere	42. I.		
000044	Pitera			whe			- 21	10,000				14941			See.			1995				
41 ° >	- A Pener	Feature litters	other .																			
41	a contraction of the			10	10	10		4		E)			6	M		0		0	10	121	1 I.	N.
Proton Pessive Identifier			TRAY	Smell	Alighatio	Norpola Ar	restal.	felar (Vergent B	tept.	AHE	Type costs in	wheely free	en fest	hey 3	inerii .	#lipherel	Porticitie -	Aperiet ic P	iai i	Prorport Bay	-
 President cal 2-Grame F 	eature Vector-up (COMME)	(ABN_AGATH			- 4	19		- 44	25		§	11	21 Phy	i loachei	- 18	- 44	£	41		- 80.		
3 Presidentical 2 Grams Feature Vector or (N46517) (094)8, HUMAR			20			34	2		- 22		9.3	38	17.8%	skoother.		.48	- R	18		.20	. 5.	
 Physicochemical 2-Grains Fe 	leature vector-sp(P49918)	CONSC. (HUMAN)	36	386	1.14	348	- 3	- 51	15		S 2	- 44	16-Phu	s icoche	- 10	1.15	2 40	\$43		- 23		
Provide the second state of the second secon			71	- 47		79	1	71	11	- 1	11 11	10	34 171	HONOR.	11	57	0	7		47	18	
 Missi confriences 2: Grama Pessaie: Vector sai GRYYS: OFFL, MATH Wyst conframetal 3: Grama Feature Vector op, P21268 (14Pc), ISAST 			65	111	39	177		118	241	1	41. It	211	81.9hu	i icadhili	- 31	110		18	1	101	24	
The contract of 2-Grane P	extore Verice 141 (298738)	WEDA, HUMAN	- 18		- 73	82	1		43.	- 1	ю. т	47	14 PTv	No.			(<u>1</u>)	- 15		- 47	78	
Physicochemical 3-Grams F	watshe Vector-spi P405801	R(MIL_SCHP0	28	- 84	1.4	53	1	71	181		8		17/Phi	alcoche.		- 19		32	2	-41		
Physicochemical 2-thems h	realiane Vector-spir (280229)	ANP_RAT	28	. \$7		55	- E	100	28		# 7.	-28	2.86	\$100714	18	10	× X			53	7	
Phee-coonemical 2-Grama N	Newforker Verotor-Apr (0221187)	001_0481	18	- 55	1	30.	- 1	- 61	18.		# T.	17	15 Pro	alcochel;	- 11	- 11	0.	13	- E	- 27		
Produced among at 3-Gramis P	reasone beccorrigi (218/2021	CONTA_HELLA	21	- 19	+	44	2	- 66	23	- 1	12	12	4.9%	100004			i #	38	- 1	- 28		
Physicochemical 2-Gramu fr	Vestaine Vector-sp(F51400.)	SEVEN, NUMBE	荷	- 31	1.1	46	- 6	- 30	- T		6 1	17	3.Phi	elozotve)	- 11	36	(ž		1.0	18	0	
Prysiciation call 2-Grant R	leafure vactor-up (QR023)	CONDA_RAT	15	- 44	1.18	48	4	- 24	18		£ 1	28.	3.Phy	Licadhe.		- 84	i 81	31		181	- A	
14 Physicochemical 2-Grams Peature Vector up (018687 (DRNS, Hukeki		21	- 31	6	541	2		LT.		4 8	34.	18 Phy	skotke.	- E	- 25	< 15	- 39	3	28			
18 Wysticschemistel 2-Grame Feature Pector-sp./FilleRej.0265A_MOULE		14			43	2	- 47	21	4	10 I	- 23	6.476	sicodv#	÷	28	E 18	- 24	3	21			
Physicochemister 3 Gramo P	Vestigate Vector ap (QS4364)	WAY INCOME	28		10 mil	28	- 1	51	18		9 2	- M.	3.Phg	10000		2	e (c			SE		
Physic cochemistrat 3-General Fi	Feature Vector-sp(GHCL8)	LEFT_ADUTH	28	- 47		23	1		25.		* *	11	21 PT-0	Excenter.	14	34	(#	- 1	1	-44	.11	
Presicochemical 2 Greme F	Seatsive Vector spl Q050821	ENT, ABATH				55	- 2		- 49		17	- 28	27 Pho	100001	11		8	:18			- 39	
Physicochemical 2-Grains F	Feature Vector-op/ GRAVIII	KRIHL_NRATH	- 43	- 48	1.18	58	- 2	209	28	- 4	11 7	.24	28 Pro	a iconte	- 24	- 64	(¥	. 30	- 4	76.	- 51	
Physicochemical 2 Grama P	restaire vector agi Q9,895	KRPS_ARATH	14		A	41	- 2	57	52		8、	- 22	20 Pro	ricoldre,	- 1 B	25	C	- 29		- 45	1	
Physicochemister 3-Grams F	essence vector-sp(0280.3)	121760_0445i		114		42	- 4	68	28		12	-41	17 Pro	# icach#	- 24	- 96		- 11	- R.	- 42	- 11	
The contrary of Commit	Restore Sector 14 (CPH055)	BURN, ARATH	- 28			49					8.1.1.4		21 Pe	1000		41	E	- 14		- 40	1	
Physicochamical 3-Grams P	feature Vector-ep(Q2R380.)	KRP3_06751	53	- 14		65 :	- 1	- 47	25		10 0	46	7:84	e logichei			X	- (W.		30	- 10	
Plusionhemical 2-Drams P	restare Vector sal GTRDHS	1894,0475	- 29					- 68	18		T 1		7.894	stoother,	- 14		()(
Physicochemical 3-Grams F	weture Vector-ear(057:15)	ERFE_DEX1	15		- 3	26	- 2	15			4 2	38	S Phy	E ICOC'NE		25	6. 8.	15		3	1	
Physical and Streets P	restore vector-spi A249W8	1101_001	10	114		87.			18		# · 7	-48	3.774	anong/re-	- 48	84		48	÷			
President and a Grame F	Anterest Anton-station (1997)	KBP5_08421	37	- 78		11.		- 63	10		£ 1	- 18	12-Pu	# Icoshe	10	11		- 11				
Fruction terrenal 2-Deams Fr	www.execurestication.com	APRIA, CHICK	10	- 18		88	E	- N.			F 8	47	27 Mu	+cutve	- 1	28	(<u> </u>	- 74		- 17	- 28	
19hezicochemical 3-Grame fi	feature Vector-sel PS4859.	SPI2_YEAST	12	- 28	C	34	2		17		5	28.		a koochie			())		R.	- 29		
Physicochemical 2-talans h	reative vector-sp(GR30PS)	CORDER_BACILIES	- 23	- 69	- 8	10	- 2	- 54	£1.		6 B	34	17.9%	sicane		19		- 84		- 26	- 4	
Physicochamilitar 2-Grania P	Newfork Vector-sp(G77817)	CDIV2A_MONDO	0 17	- 40		48.			17		4 2	29	8. Pro	a coohei	- 1	20		- 12		15-		
Physicsochemical 2-Grama F	reature becconspictments	BRE MONDO	- 28	- 18		42		- 28	11		* 1	11	6.0%	Excedite:				- 14	1			
Presidential 2 Grana	enture vector ap QR/H4	OWAT THEFT	10			17		- 45	30		B	14	12.7%	100070		31	- E	- 34		- 28	11	
Physicochemical 3 Grains R	enature vertor-cpi QR2781	SIM_ARATH	24			28	- 1	-41	28		r	13	8.9%	Excedite.	- 14		- ¥	- 18		20		
The codemical 2 Grand P	reative vector op OVICE2	WEZ APATH			- 7	17	- 1	57	22		2 1	1	17-Pha	scodys.						- 54	32	
Physicscherkstal 3 Grants F	earne vector-op (0120/2)	10,0900	15	+2		34	- 1	- 20			2 2	- 13	E ftie	sicodve:	10	30	Q.	12			0	
Pagaoohemaal 3 Grams N	Assence Assence all (CT1664)	0464_46479-	1			26	- 5		00		s	14	1.84	110009		24	- N				- 2	
The conternal of Grants P	Heatman Aeros-chi (06516)	INTEL ABATH		- 14	1	28	1	- 14	20		F	4	L1 Phu	erophe.	- 1	20	. R			- #.	T	
The contractor 2 Grand F	restore Vector NPI CRUNKE	IMRS_AMATH	1	14	- I		- 1	- 27	14		5		1.17	1000 m	- 4		- 2		- 8	18	1	
Physicoch	nemical = Grama Postur	 (8): 										+										
																	1.000	100	1001			-

Figure.5 Figure showing output descriptor in CSV file format; Only one click for CSV file generation (Save File).

• The PnGT tool consists of a menu bar, a graphical control element that contains few selectable drop-downs for user convenient such as Physicochemical Properties, References *etc.*



Figure.6 The Snapshot showing widget of About Menu: Amino Acid and widgets of Help Menu: Contact Us & References.