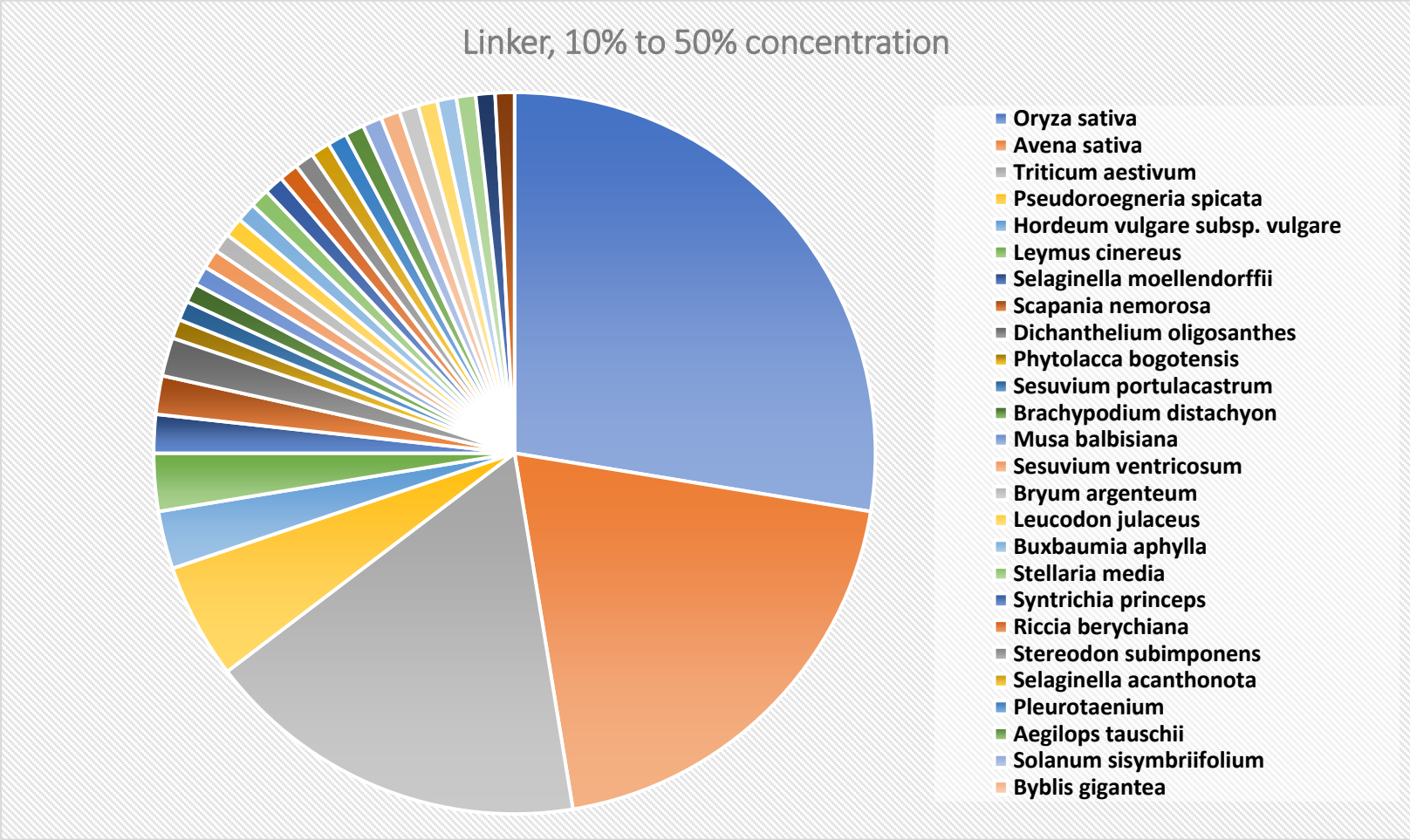
Statistics and developments

Hololectins and chimerolectins statistics



Statistics:
206 sequences
80 species

The longest linkers

Barbilophozia barbata	8	hololectin	SSNSGAEGLRTGSTTAGIALASAIATMLFLGSEKGWRAPGWRLIFALLCLTASLQLGLATE DGECANSNGCPEQFKYGE		
Morus alba var. multicaulis	8	chimerolectin	WSSPPPPSPPPPPPSPPPPSPPPPSPPPPSPPPPSPPPPSGPERPDH	WTSLLSALISNGNNAISK	
Selaginella acanthonota	8	hololectin	GLQSSSSAKPPATSTTASFNGSPTGSTLN	GNQNNHPTGPANSTTV	
Pandorina morum	8	hololectin	WDPPPPPSNIPPGGSAPPSPPPPARGFTW		
Pleurotaenium	8	hololectin	NSTVWAAAPPDSLLTPDSVCSAIV	KPTSRNCTAK	GGDSCKYGSCN
Pandorina morum	8	chimerolectin	ASLDTSDGALLRAVADRRKDTTRP	IQYPSPPPKPPAPPPA	
Gonium pectorale	8	hololectin	WWPPPPPALANTTSPPPPSPGTVL		
Prumnopitys andina	8	chimerolectin	YSNSVPSNSMIMIMLVYIAPAGN	YSRVSS	
Osmunda javanica	8	hololectin	TGSANSPQSGPASSPSKSDT		
Volvox globator	8	chimerolectin	LPNGALLRSVASRKTPPARP	TQYPPSPPRPSAPPPF	WNKFSPPSPPRSPPPQPPPPPPPPASYG
Lobochlamys segnis	mix	chimerolectin	TADCRLLRSVTSRVPPPPRP	GVTRPPPPTVGPMASPPPPSSVG	KRPPPPSQSHGSPPPPVSSPPPAPPPPRPAPSPP PLASPPPPPPPPPASPPPPPPASPPPPPPASPPL PAANSPPPPANSPPPPVSVSRPPPPKGKAKRRP PKVPPSPSPPPAPAKAQGF
Avena sativa	mix	hololectin	PRYGKRANVDGNVIPGNA	TGAKIDQDDVPGNA	TGAKLNEDVVPNA
Orthotrichum lyellii	8	hololectin	EGDTRTGPTSAPGATTPM		
Bambusina borreri	8	hololectin	SAPSPESGQSPPTPSSDG		
Buxbaumia aphylla	10	hololectin	GDGLKAIQEMRSDYNGK		
Scapania nemorosa	8	hololectin	YVSSPNSSPPDSGGLGE		
Scapania nemorosa	8	chimerolectin	YVSSPNSSPPDSGGLGE	YRSSHG	
Capsicum baccatum	6	hololectin	VLETRQSQWEEPAETDL		
Scapania nemorosa	8	hololectin	NVSRSSKTGSEPSHQGE		
Pteromonas	8	hololectin	NVEPPRGGHYGGPVAPL		
Avena sativa	8	hololectin	YGKRDNVDGNAVPGNA	TGAKFNENVVPDNA	TGAMLNEDGVPNA
Avena sativa	8	hololectin	YGKRANVDGNVIPGNA	TGAKINQDDVPGNA	TGAKLNEDVVPNA
Avena sativa	8	hololectin	YGKRANVDGNVIPGNA	TGAKINQDDVPGNA	TGAKLNEDVVPNA
Avena sativa	8	hololectin	YGKRANVDGNVIPGNA	TGAKINQDDVPGNA	TGAKLNEDVVPNA
Avena sativa	8	hololectin	YGKRANVDGNVIPGNA	TGAKINQDDVPGNA	TGAKLNEDVVPNA
Avena sativa	8	hololectin	YGKRANVDGNVIPGNA	TGAKINQDDVPGNA	TGAKLNEDVVPNA
Scapania nemorosa	8	hololectin	LLNSPSSSPSNSDDGS		
Avena sativa	8	hololectin	YGKRANVDGNIVPGNA	TRAELNENIVPGNA	TGGKLNEDVVSNA
Avena sativa	8	hololectin	YGKRANVDGNIVPGNA	TRAELNENIVPGNA	TGGKLNEDVVSNA
Avena sativa	8	hololectin	YGKRANVDGNIVPGNA	TRAELNENIVPGNA	TGGKLNEDVVSNA
Avena sativa	8	hololectin	YGKRANVDGNIVPGNA	TRAELNENIVPGNA	TGGKLNEDVVSNA
Avena sativa	8	hololectin	YGKRANVDGNIVPGNA	TRAELNENIVPGNA	TGGKLNEDVVSNA
Avena sativa	8	hololectin	YGKRANVDGNAIPGNA	TGAKLNQDVVPGNA	TGAKLNEDVVPNA
Avena sativa	8	hololectin	YGKRANVDGNIVPGNA	TRAELNENIVPGNA	TGGKLNEDVVSNA
Avena sativa	8	hololectin	YGKRANVDGNIVPGNA	TRAELNENIVPGNA	TGGKLNEDVVSNA
Avena sativa	8	hololectin	YGKRANVDGNIVPGNA	TRAELNENIVPGNA	TGGKLNEDVVSNA
Avena sativa	8	hololectin	YGKRANVDGNAIPGNA	TGAKLNQDVVPGNA	TGAKLNEDVVPNA
Avena sativa	8	hololectin	YGKRANVDGNAIPGNA	TGAKLNQDVVPGNA	TGAKLNEDVVPNA
Avena sativa	8	hololectin	YGKRANVDGNAIPGNA	TGAKLNQDVVPGNA	TGAKLNEDVVPNA
Avena sativa	8	hololectin	YGKRANVDGNAIPGNA	TGAKLNQDVVPGNA	TGAKLNEDVVPNA
Avena sativa	8	hololectin	YGKRANVDGNVIPGNA	TGAKINQDDVPGNA	TGAKLNEDVVPSA
Avena sativa	8	hololectin	YGKRANVDGNAIPGNA	TGAKLNQDVVPGNA	TGAKLNEDVVPNA
Leucodon julaceus	8	hololectin	AAPSDPYASTVPNIDV	STNRKKPADPTTPPKTANAGDT	
Neckera douglasii	6	hololectin	GATAVVYPAAVYATRD		
Schwetschkeopsis fabronia	6	chimerolectin	GAVTVVQPAPVYATRD	GIGAPAPAVGS	GAVTYGN
Avena sativa		8hololectin	YGKRANVDGNAIPGNA	TGAKLNQDVVPGNA	
Fontinalis antipyretica	8	chimerolectin	HGRGHGAYGVLAAAE	HGRGHGAYGILATAE	HGHGGPRGDTGRASYYTTYVPSACYGNDESQL PRNRHMAAVNDRH
Solanum tuberosum	8	hololectin	PSPPSPPSPFPPPPG	KNRTIDDYECGHQG	PPPPPPPPPPPYAIG
Stellaria media	mix	hololectin	HNTPLSEIEPTAAGQ		
Solanum ptychanthum	8	hololectin	PPPPLPPPPPPPAPG	RLPPEAPPPPPDAQ	

Hololectins and chimerolectins statistics

Protein ID	Species	n Cys	Lectin type	Linker 1	Linker 2	Linker 3	n ND	conc
MRKX-2002612	Phytolacca bogotensis	8	hololectin	DYD	NFW		3	0.5
HZTS-2101312	Sesuvium portulacastrum	8	hololectin	DSGN	DYGR	DYG	4	0.363636
I1IJ99	Brachypodium distachyon	mix	hololectin	FTNNR	RDDHS	YDQR	5	0.357143
A0A4S8J4P4	Musa balbisiana	mix	hololectin	NDSGY	DGGSGSGDSDA		5	0.3125
OPZX-2050165	Sesuvium ventricosum	8	hololectin	DYLE	DYG	DYG	3	0.3
DQ462308.2	Triticum aestivum	8	hololectin	YDNKI	WADL	STDKP	4	0.285714
M0YEA8	Hordeum vulgare subsp. vulgare	8	hololectin	YNNKM	WADL	STDKP	4	0.285714
BE403750.1	Triticum aestivum	8	hololectin	YDNKI	WADL	STDKP	4	0.285714
JZ883150.1	Triticum aestivum	8	hololectin	YDNKI	WADL	STDKP	4	0.285714
GO588722.1	Avena sativa	8	hololectin	YGKRDNVDGNAVPGNA	TGAKFNENVVPDNA	TGAMLNEDGVPNA	12	0.27907
GO583073.1	Avena sativa	8	hololectin	TGAKINQDDVPGNA	TGAKLNEDVVPNA		7	0.259259
GO585665.1	Avena sativa	8	hololectin	YGKRANVDGNVIPGNA	TGAKINQDDVPGNA	TGAKLNEDVVPNA	11	0.255814
GO585827.1	Avena sativa	8	hololectin	YGKRANVDGNVIPGNA	TGAKINQDDVPGNA	TGAKLNEDVVPNA	11	0.255814
GO581912.1	Avena sativa	8	hololectin	YGKRANVDGNVIPGNA	TGAKINQDDVPGNA	TGAKLNEDVVPNA	11	0.255814
GO584240.1	Avena sativa	8	hololectin	YGKRANVDGNVIPGNA	TGAKINQDDVPGNA	TGAKLNEDVVPNA	11	0.255814
GO582968.1	Avena sativa	8	hololectin	YGKRANVDGNVIPGNA	TGAKINQDDVPGNA	TGAKLNEDVVPNA	11	0.255814
D8RQT6	Selaginella moellendorffii	8	hololectin	NQEI			1	0.25
D8S008	Selaginella moellendorffii	8	hololectin	NQEI			1	0.25
JMXW_scaffold_2007524	Bryum argenteum	8	hololectin	NMLSTTPAVNDA	NVKVDSAA		5	0.25
IRBN-2159389	Scapania nemorosa	8	hololectin	LLNSPSSSPSNSDDGS			4	0.25
AL817495.1	Triticum aestivum	8	hololectin	RADF			1	0.25
CK169134.1	Triticum aestivum	mix	hololectin	ASYDNKI	WADL	STDKP	4	0.25
GO585787.1	Avena sativa	mix	hololectin	PRYGKRANVDGNVIPGNA	TGAKIDQDDVPGNA	TGAKLNEDVVPNA	11	0.244444
GO581539.1	Avena sativa	8	hololectin	YGKRANVDGNIVPGNA	TRAELNENIVPGNA	TGGKLNEDVVSNA	10	0.232558
GO582252.1	Avena sativa	8	hololectin	YGKRANVDGNIVPGNA	TRAELNENIVPGNA	TGGKLNEDVVSNA	10	0.232558
GO582291.1	Avena sativa	8	hololectin	YGKRANVDGNIVPGNA	TRAELNENIVPGNA	TGGKLNEDVVSNA	10	0.232558
GO585011.1	Avena sativa	8	hololectin	YGKRANVDGNIVPGNA	TRAELNENIVPGNA	TGGKLNEDVVSNA	10	0.232558
GO584693.1	Avena sativa	8	hololectin	YGKRANVDGNIVPGNA	TRAELNENIVPGNA	TGGKLNEDVVSNA	10	0.232558
GO583188.1	Avena sativa	8	hololectin	YGKRANVDGNAIPGNA	TGAKLNQDVVPGNA	TGAKLNEDVVPNA	10	0.232558
GO582673.1	Avena sativa	8	hololectin	YGKRANVDGNIVPGNA	TRAELNENIVPGNA	TGGKLNEDVVSNA	10	0.232558
GO582580.1	Avena sativa	8	hololectin	YGKRANVDGNIVPGNA	TRAELNENIVPGNA	TGGKLNEDVVSNA	10	0.232558
GO586735.1	Avena sativa	8	hololectin	YGKRANVDGNIVPGNA	TRAELNENIVPGNA	TGGKLNEDVVSNA	10	0.232558
GO581548.1	Avena sativa	8	hololectin	YGKRANVDGNAIPGNA	TGAKLNQDVVPGNA	TGAKLNEDVVPNA	10	0.232558
GO584697.1	Avena sativa	8	hololectin	YGKRANVDGNAIPGNA	TGAKLNQDVVPGNA	TGAKLNEDVVPNA	10	0.232558
GO585043.1	Avena sativa	8	hololectin	YGKRANVDGNAIPGNA	TGAKLNQDVVPGNA	TGAKLNEDVVPNA	10	0.232558
GO582770.1	Avena sativa	8	hololectin	YGKRANVDGNAIPGNA	TGAKLNQDVVPGNA	TGAKLNEDVVPNA	10	0.232558
GO582754.1	Avena sativa	8	hololectin	YGKRANVDGNVIPGNA	TGAKINQDDVPGNA	TGAKLNEDVVPSA	10	0.232558
GO584450.1	Avena sativa	8	hololectin	YGKRANVDGNAIPGNA	TGAKLNQDVVPGNA	TGAKLNEDVVPNA	10	0.232558
FF346516.1	Pseudoroegneria spicata	8	hololectin	RADI	STDKP		2	0.222222
AL818134.1	Triticum aestivum	8	hololectin	RADI	STDQP		2	0.222222
GH727783.1	Triticum aestivum	8	hololectin	STDKP			1	0.2
CD924836.1	Triticum aestivum	8	hololectin	STDKP			1	0.2

ololectins and chimerolectins statistics

IGUH_scaffold_2164028	Leucodon julaceus	8	hololectin	AAPSDPYASTVPNIDV	STNRKKPADPTTPPKTANAGDT		7	0.184211
HRWG_scaffold_2068669	Buxbaumia aphylla	10	hololectin	GDGLKAIQEMRSDYNGK			3	0.176471
A0A1E5W581	Dichanthelium oligosanthes	8	hololectin	CPNR	REDR	CEHG	2	0.166667
EG395772.1	Leymus cinereus	8	hololectin	YTSK	RANI	STDK	2	0.166667
EG378972.1	Leymus cinereus	8	hololectin	YTSK	RANI	STDK	2	0.166667
FG952083.1	Oryza sativa	8	hololectin	CSSQ	RADI	CPEN	2	0.166667
E1UYT9	Stellaria media	mix	hololectin	HNTPLSEIEPTDA			2	0.153846
GRKU-2016385	Syntrichia princeps	8	chimerolectin	PTSEGA	APTAPKGVFPGHLLFDYIGANGVTINFNDVPTALAGVDYA LGLSFAIDMNANGATQNGV		10	0.153846
WJLO-2036720	Riccia berychiana	8	chimerolectin	AAEND	SYNAPKGLKPGRMLFDYLGSNGVPITFNEIPITQ		6	0.153846
M25536.1	Triticum aestivum	8	hololectin	WTSK	RADI	STDKP	2	0.153846
AGI1_WHEAT Agglutinin isolectin 1	Triticum aestivum	8	hololectin	WTSK	RADI	STDKP	2	0.153846
AGI2_WHEAT Agglutinin isolectin 2	Triticum aestivum	8	hololectin	WTSK	RADI	STDKP	2	0.153846
AGI3_WHEAT Agglutinin isolectin 3	Triticum aestivum	8	hololectin	WTSK	RADI	STDKP	2	0.153846
FF356830.1	Pseudoroegneria spicata	8	hololectin	YTSK	RADI	STDKP	2	0.153846
FF351450.1	Pseudoroegneria spicata	8	hololectin	YTSK	RADI	STDKP	2	0.153846
BG365763.1	Hordeum vulgare subsp. Vulgare	8	hololectin	YTSK	RADI	STDKP	2	0.153846
FF351988.1	Pseudoroegneria spicata	8	hololectin	YTSK	RADI	STDKP	2	0.153846
FF367588.1	Pseudoroegneria spicata	8	hololectin	YTSK	RADI	STDKP	2	0.153846
FF343459.1	Pseudoroegneria spicata	8	hololectin	YTSK	RADI	STDKP	2	0.153846
CJ776108.1	Triticum aestivum	8	hololectin	YTSK	RADI	STDKP	2	0.153846
AL820037.1	Triticum aestivum	8	hololectin	WTSK	RADI	STDKP	2	0.153846
CD901987.1	Triticum aestivum	8	hololectin	WTSK	RADI	STDKP	2	0.153846
CD901264.1	Triticum aestivum	8	hololectin	WTSK	RADI	STDKP	2	0.153846
AL818572.1	Triticum aestivum	8	hololectin	WTSK	RADI	STDKP	2	0.153846
BY845808.1	Hordeum vulgare subsp. Vulgare	8	hololectin	YTSK	RADI	STDKP	2	0.153846

Hololectins and chimerolectins statistics

A0A1E5UY24	Dichanthelium oligosanthes	8	hololectin	CPNR	CEK		1	0.142857
LNSF_scaffold_2010351	Stereodon subimponens	8	hololectin	NVAEGTTVDAASAA			2	0.142857
EG396761.1	Leymus cinereus	8	hololectin	YTSK	RANIK	STDKP	2	0.142857
ZYCD_scaffold_2042177	Selaginella acanthonota	8	hololectin	GLQSSSSAKPPATSTTASFNGSPTGSTLN	GNQNNHPTGPANSTTV		6	0.133333
MOYY_scaffold_2007053	Pleurotaenium	8	hololectin	NSTVWAAAPPDSLLTPDSVCSAIV	KPTSRNCTAK	GGDSCKYGSCN	6	0.133333
M8CGU9	Aegilops tauschii	mix	hololectin	GARHGNKI	WADLKCGHLANGPEFCGARCQNGACSTDKP		5	0.131579
CD925154.1	Triticum aestivum	8	hololectin	WTSK	RADI		1	0.125
BQ246423.1	Triticum aestivum	8	hololectin	WTSK	RADI		1	0.125
FG969726.1	Oryza sativa	8	hololectin	RADI	CPEK		1	0.125
FG968377.1	Oryza sativa	8	hololectin	RADI	CPEK		1	0.125
FG955637.1	Oryza sativa	mix	hololectin	RADI	CPEK		1	0.125
FG967518.1	Oryza sativa	8	hololectin	RADI	CPEK		1	0.125
CI157196.1	Oryza sativa	8	hololectin	RADI	CPEK		1	0.125
FG949952.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG943860.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG966473.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG947977.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG967501.1	Oryza sativa	mix	hololectin	CSSQ	RADI		1	0.125
FG962101.1	Oryza sativa	mix	hololectin	CSSQ	RADI		1	0.125
FG945257.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG963243.1	Oryza sativa	mix	hololectin	CSSQ	RADI		1	0.125
FG958852.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG962209.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG969759.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG966676.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG965855.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG959219.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG960185.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG963939.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG945507.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG963217.1	Oryza sativa	mix	hololectin	CSSQ	RADI		1	0.125
FG969260.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG945838.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG957301.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG960988.1	Oryza sativa	mix	hololectin	RADI	CPEK		1	0.125
FG968778.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
CI157189.1	Oryza sativa	mix	hololectin	RADI	CPEK		1	0.125
FG945322.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125
FG969017.1	Oryza sativa	8	hololectin	CSSQ	RADI		1	0.125

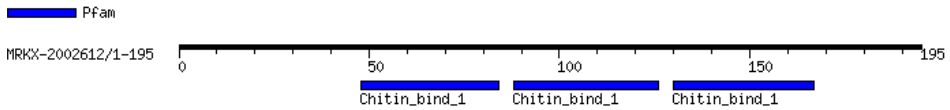
Hololectins and chimerolectins statistics

MG720758.1	Camellia fraterna	8	chimerolectin	SPTPG	SPTPG	SPTPPPPPPPPPTPPPPAAPSPPSSGDITD	2	0.05	without ND set 2	9
LQJY-2003271	Solanum xanthocarpum	8	chimerolectin	PEEPSPPPGA	PEAPSPNPPPTPSPPPSPSPPPSPGSPPGPAEDVSS		2	0.043478	without ND set 2	8
KUJU_scaffold_2050619	Gonium pectorale	8	hololectin	WWPPPPPALANTTSPPPPSPGTVL			1	0.041667	without ND set 2	8
AQFM-2001500	Pseudolarix amabilis	8	chimerolectin	GED	GGGSTPTPTTPTPTPTTPSGQGVAS		1	0.035714	with ND set 1	2
DLJZ_scaffold_2004921	Solanum ptychanthum	8	hololectin	PPPPLPPPPPPPAPG	RLPPEAPPPPPDAQ		1	0.034483	without ND set 2	4
ISPU_scaffold_2004539	Volvox globator	8	chimerolectin	LPNGALLRSVASRKTPPARP	TQYPPSPPRPSAPPPF	WNKFSPPSPPRSPPPQPPPPPPPPASYG	2	0.03125	without ND set 2	6
PYHZ-2072627	Isoetes sp	8	chimerolectin	GTSSNG	TPTSGTPPTSGTSGVSSIISSSTFSAFFPYM		1	0.027027	without ND set 2	11
KIIX-2004332	Pilularia globulifera	8	chimerolectin	P	TTTTSTNGQ	TTTPSPPPPPPTTTTPPPPTTTTGSVSS	1	0.026316	without ND set 2	7
OFUE_scaffold_2010323	Lobochlamys segnis	mix	chimerolectin	TADCRLLRSVTSRVPPPPRP	GVTRPPPPTVGPMASPPPPSSVG	KRPPPPSQSHGSPPPPVSSPPPAPPPPRPAPSPPPLASPPPPPPPPPASPPPPPPASPPPPPPASPPLPAANSPPPPANSPPPPVSVSRPPPPKGKAKRRPPKVPPSPSPPPAPAKAQGF	3	0.018405	with ND set 1	6
POIR_scaffold_2003815	Volvox aureus	8	hololectin	IAYPPPPPRPPSPPP	WYKPRSPPPRPRPPPPPSPRPPLWKG	LSSPPPPTPPGDTGPM	1	0.017544	without ND set 2	11
ERXG-2060416	Eschscholzia californica	8	hololectin	TKLIPEG	SI		0	0	with ND set 1	4
HPXA-2008397	Ptilidium pulcherrimum	6	hololectin	GAAAGT	GGTSYGS		0	0	with ND set 1	4
RKGT-2057060	Eschscholzia californica	8	hololectin	TKLIPEGR	SIR		0	0	with ND set 1	7
TMAJ-2134096	Neckera douglasii	6	hololectin	GAEAGT			0	0	with ND set 1	7
A0A1U8FVS6	Capsicum annuum	8	hololectin	PGPIRVR	SGPYPSG		0	0	without ND set 2	2
A0A1U8G3G4	Capsicum annuum	8	hololectin	PGPIPVG	SGPYPSG		0	0	without ND set 2	2
DN141224.1	Panicum virgatum	8	hololectin	SGFGTLSAE			0	0	without ND set 2	4
FE620753.1	Beta vulgaris	8	hololectin	SGFGTLSAE			0	0	without ND set 2	5
FE621365.1	Panicum virgatum	8	hololectin	SGFGTLSAE			0	0	without ND set 2	5
GOWD-2015396	Sphagnum lescurii	8	hololectin	R			0	0	without ND set 2	5
JPYU_scaffold_2004505	Marchantia	8	hololectin	GVSPESSPAQGS			0	0	without ND set 2	7
BG368849.1	Hordeum vulgare	mix	hololectin	YTSK			0	0	wga blast	13

Top few

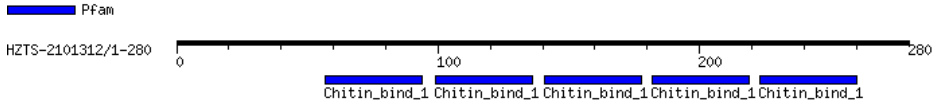
>MRKX-2002612/1-195 |Phytolacca bogotensis|Phytolaccaceae|Caryophyllales

```
MKTSNSIVVMLVLVLSSLVLLLPVEGQGHEGHGVGEILLMGKLGEPVCGVSASGRVCPNG
HCCSEWGYCGTTNEYCGKGCQSQCDYDRCGQEFGGKKCHHDLCCSQYGWCGYSQDAHCGE
GCQSQCNFWRCGKDFGGRICTGNLCCSKYGWCGYTEDHCRDGCQSQCIPSSLLPSPLHRT
IAIRKLKANLANMLS
```



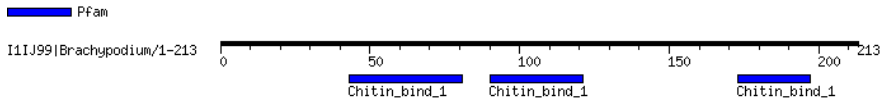
>HZTS-2101312/1-280 |Sesuvium portulacastrum|Aizoaceae|Caryophyllales

```
MAKTKKSSIGSAAMLAATLYMSTVVLHLAALPVLGQGDRVEQPMQLMPLWQKLGGPECGR
QGGGKVCPNNRCCSRWGYCGDTDAFCGNGCQSQCDSGNRCGKDFGGRVCPNGECCSRWGY
CGGDEMHCGYGCQSQCDYGRRCGKDFDDRLCPNDECCSEHGYCGVTRAHCEKGCQSQCDY
GRCGTEFDDKECEEGMCCSERGYCGVTDAHCGTGCQSQCEQQRCGKEFGGRYCPDWECCS
EEGYCGVTDDHCGKGCQSQCRCHLGAKALPATVARLLKLV
```



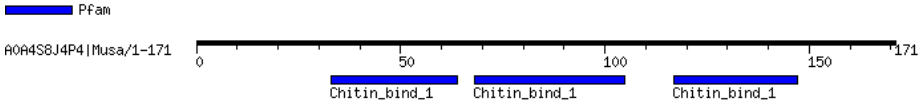
>I1IJ99|Brachypodium/1-213 distachyon|Poaceae|Poales

```
MAMTTTKLQQVVVLSTLLAAAALAPLTVHGAVDINLGVPCPKQCGREANNTVCGDNHCCS
GDGFCGLGGNYCGSGCQSGACFTNNRCSETSPCPNNQCCSVYGYCGFGQDYCGSGCRNGP
CRDDHSCEGGKLCPSNLCCRGKDKKCGLGGNYCSINGDQGCLSGACYDQRCSSAKPCSNG
YCCSVHGYCGVGRAYCGGDGSLTLLNGLVCVLS
```



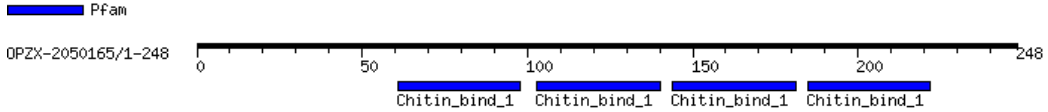
>A0A4S8J4P4|Musa/1-171 balbisiana|Musaceae|Zingiberales

```
MASLLFCLECRKMKLPFVVIFGLGSLLGSLAQQCGPAAGGKTCLDGLCCSKYGYCGSTFA
YCNDSGYQCGSQAAGGVCSAWQCCSQHGYCGNTSDYCGSGCQSQCDGGSGSGDSDAQCGS
QAAGALCSDGQCCSQYGYCGTTSDYCGVSRKSTTWTHKEFASISISVMLCG
```



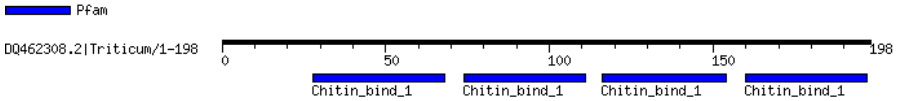
>OPZX-2050165/1-248 |Sesuvium ventricosum|Aizoaceae|Caryophyllales

```
MAKTKKSSIGSVAMLASILYYMSTVMLLQLAVVAVSGQGEVAEQVQVKPLSLWHKLTSFP
TCGWQVLGRVCPDGRCCSKWGYCGDTEDYCGSGCQSNCDYLERCGKDFGGRLCPNGECCS
QHGYCGVTNAHCDTGCQSQCDYGRCGLEFGDKVCDGGKCCSQWGYCGVTDAHCGNGCQSQ
CDYGRCGVEFGDGLCPNGECCSENGYCGVTKAHCKMGCQSQCEGLSKANDLPHNTNFTAT
AARLLKLI
```



>DQ462308.2|Triticum/1-198 aestivum|Poaceae|Poales

```
MKGLLLCALALAFAAVTTHAQLQSCPTRCGKQADGMECPNNLCCSKDGYCGLGVDYCSAG
AGCQSGAYYDNKICGAQANGTLCRNNHCCSSGGRCGYGREYCSNGCQGGPCWADLKCGHL
DNGKLCPNNLCCSQYGYCGLGPEFCGTGCQNGACSTDKPCGNKANGAPCTNNYCCSQYGS
CGLGKDYCGTGCQNGACN
```

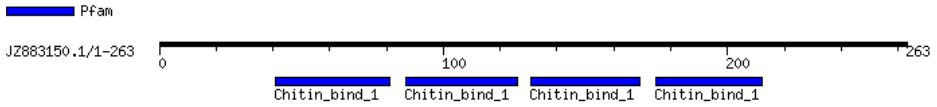


~25% (like avenatides)

>JZ883150.1/1-263 JN177_LCP131_D03 cDNA library of wheat cultivar Jinan 177 Triticum aestivum cDNA, mRNA sequence

```
STYPGRTLAMKGLFLCALALAFAMVTTHAQLQYCEKRCGKQADGMECPNNLCCSKD
GYCGLGVDYCSAAAGCQSGACYDNKICGAQAGGALCPNNHCCSSGGRCGYGSEYCSGSRG
CQSGPCWADLKCGHLANGKQCPNNLCCSQYGYCGLGPEFCGARCQNGACSTDKPCGNKAN
GARCTNNYCCSQYGSCGLGKDYCGTGCQSGACYTPSFLANILKCVP
```

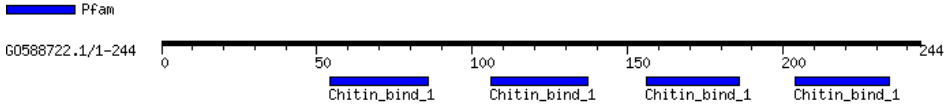
28.5% of N/D



>GO588722.1/1-244 ASAEN2NG-UP-037_F07_20DEC2007_053 ASAEN2NG Avena sativa cDNA, mRNA sequence

```
HHRKLSHPPN*LRKKTFPSSIHRELRASSMIMKALALGLLVLACAAATALGCSGGSPCPG
NQCCSKYGYCGLGGDYCGAGCQSGPCYGKRDNVDGNAVPGNACSSSSLCPSNQCCSKWGY
CGLGSDYCGSGCQSGPCTGAKFNENVVPDNACSSSSPCPSNQCCSKWGYCGLGGDYCGSG
CQSGPCTGAMLNEDGVPNACSSSSPCPGNQCCSKWGYCGLGGDYCGAGCQGGPCTGAALP
SDEM
```

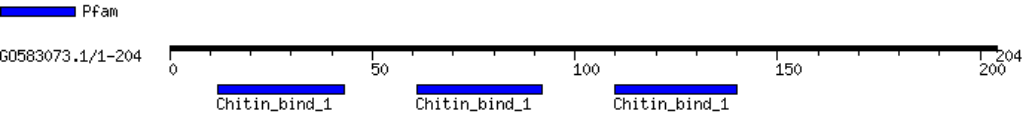
27.9% of N/D



>GO583073.1/1-204 03-ASAEN1NG-UP-023_B03_03JAN2007_029 ASAEN1NG Avena sativa cDNA, mRNA sequence

```
GNVIPGNACSSSSPCSGNQCCSKWGYCGLGGDYCGDGCQSGPCTGAKINQDDVPGNACSS
SSPCPGNQCCSKWGYCGLGGDYCGSGCQSGPCTGAKLNEDVVPNACSSSSPCSGNQCCSK
WGYCGLGGDYCGAGCQSGPCTGAALLSDEMW*TLSTIVHVVVLRVDLPPIK*SDPWMSQR
RMFVCVIIGTMSVPCNKDGVRFVV
```

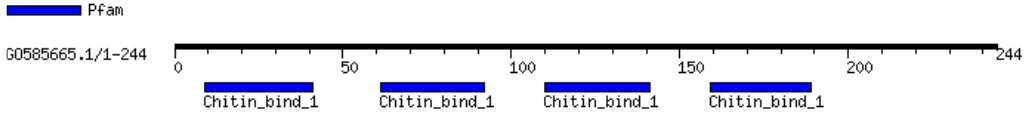
25.9% of N/D



>GO585665.1/1-244 03-ASAEN1NG-UP-052_B03_04JAN2007_029 ASAEN1NG Avena sativa cDNA, mRNA sequence

```
TTTALSCSGGSPCPGNQCCSKYGYCGLGGDYCGAGCQSGPCYGKRANVDGNVIPGNACSS
SSPCSGNQCCSKWGYCSLGGDYCGDGCQSGPCTGAKINQDDVPGNACSSSSPCPGNQCCS
KWGYCGLGGDYCGSGCQSGPCTGAKLNEDVVPNACSSSSPCSGNQCCSKWGYCGLGGDYC
GAGCQSGPCTGAALLSDEMW*TLSTIVHVVVLRVDLPPIK*SDPWMSQRRMFVCVIIGTM
SVPC
```

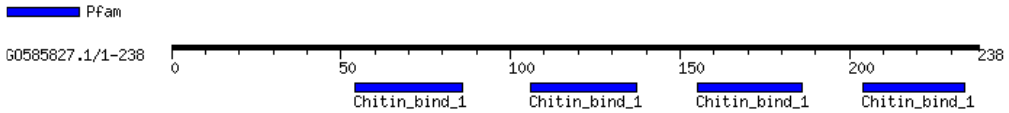
25.5% of N/D



>GO585827.1/1-238 ASAEN1NG_UP_002_D03_28NOV2006_025 ASAEN1NG Avena sativa cDNA, mRNA sequence

```
DRKLSHPPN*LS*KRKSSEIHPLQVRASNMAMKALASGLLVLACATTTALSCSGGSPCPG
NQCCSKYGYCGLGGDYCGAGCQSGPCYGKRANVDGNVIPGNACSSSSPCSGNQCCSKWGY
CGLGGDYCGDGCQSGPCTGAKINQDDVPGNACSSSSPCPGNQCCSKWGYCGLGGDYCGSG
CQSGPCTGAKLNEDVVPNACSSSSPCSGNQCCSKWGYCGLGGDYCGAGCQSGPCTGAA
```

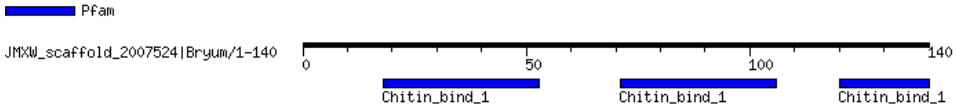
25.5% of N/D



>JMXW_scaffold_2007524|Bryum/1-140 argenteum|Bryaceae|Bryales

```
ILLVAMVVSVAAADEYAECGSQAHGAVCLFGMCCSKDGMCGRGENYCGNGCQEGAGRCNM
LSTTPAVNDAECGVQAHGAVCLFGACCSKDGMCGRGDKYCGNGCQEGAGRCNVKVDSAAE
CGSQAHGALCLFGACCSKDG
```

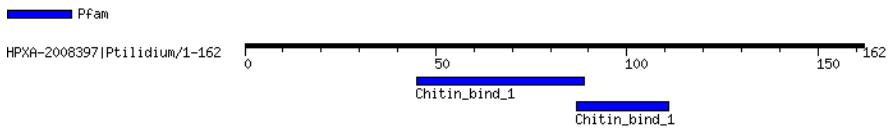
25.0% of N/D



Bottom few

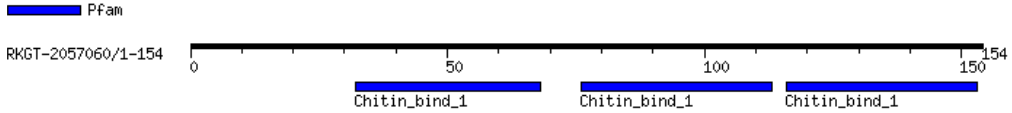
>HPXA-2008397|Ptilidium/1-162 pulcherrimum|Ptilidiaceae|Ptilidiales (inaccurate motif find)

```
MMIAATGVSLLLLSLVYNVAGNCLISGCGGGGCCSAYGYCGTGYCGAAAGTCLNYGCPAG
QCCSQYGYCGNTATYCGGTSYGSCYSTGCGGGLCCSQYGYCGSYGAYGSYCAVAKFLSRK
QPVSLEGEFQGQATYYNETMAGADYSTCGTSRARSLDENDEK
```



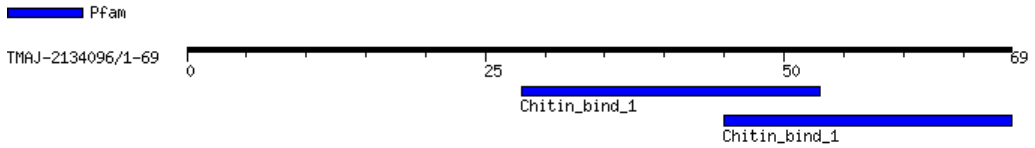
>RKGT-2057060/1-154 |Eschscholzia californica|Papaveraceae|Ranunculales

```
MRMSSSSTSFFLLFLFISSSSSIPFNEWTSDRCGTFGGSVCPEGSCCSIWGYCGNTNDYC
LYNCYSQCTKLIPEGRCGTEFSNALCPEGLCCSQWGYCGNTADHCGSGCQSQCSIRCGNV
FGDSRCPEGLCCSLWGYCGNTIEHCGADCQSQCS
```



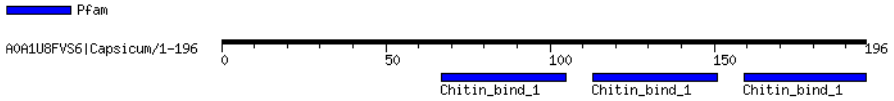
>TMAJ-2134096/1-69 |Neckera douglasii|Neckeraceae|Hypnales

```
MMIAATGMSLLLLSLVYNAAGNCLIGGCGAGACCSAYGYCGVGYCGAEAGTCRNYGCPAG
QCCSQYGYC
```



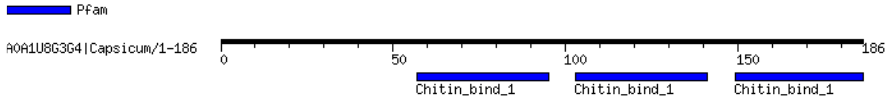
>A0A1U8FVS6|Capsicum/1-196 annuum |Solanaceae|Solanales

```
MIPRLITIHKMREIIISLLALALFLLKVSAKLSDVPFYLPANETFGLELIRERNTSDLAQ
SLLARSRCGWPAGGRWCPEGQCCNFDGWCGTTSAYCGENMCDFQCPGPIRVRRCGMQAGG
RPCPTGQCCRDTGWCGTTERYCNPAHCQSQCSGPYPSGRCGWQAGGRKCPTGLCCSLSGW
CGTTSIYCSREECQSQ
```



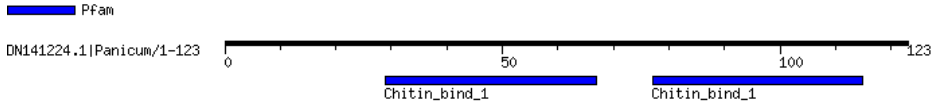
>A0A1U8G3G4|Capsicum/1-186 annuum |Solanaceae|Solanales

```
MREIIISLLSLALFLLKVSAKLSDVPFYLPANETLGLKLIRERNTSDLAQSLLAQSRCGW
PGGGRSCPTGQCCDFDGWCGTGDAYCDEDECDFQCPGPIPVGRCGMQAGGRPCPTGQCCR
DSGWCGTTEEYCNPLRCQSQCSGPYPSGRCGWQAGGRKCPTGLCCSLSGWCGTTSIYCSR
EECQSQ
```



>DN141224.1|Panicum/1-123 virgatum|Poaceae|Poales

```
MRSAVLAMKALVLSALLLTFAGVITHAQQCGSQAGGKKCPNNLCCSPWGYCGSGPDYCGN
GCQSGPCSGFGTLSAEQCGRQAGNKNCPNNLCCSQWGFCGLGGDYCGNGCQSGPCSGELG
AEQ
```



HLP dataset statistics

Chitin-binding

2917 sequences, 1064 species, 329 families, 105 orders

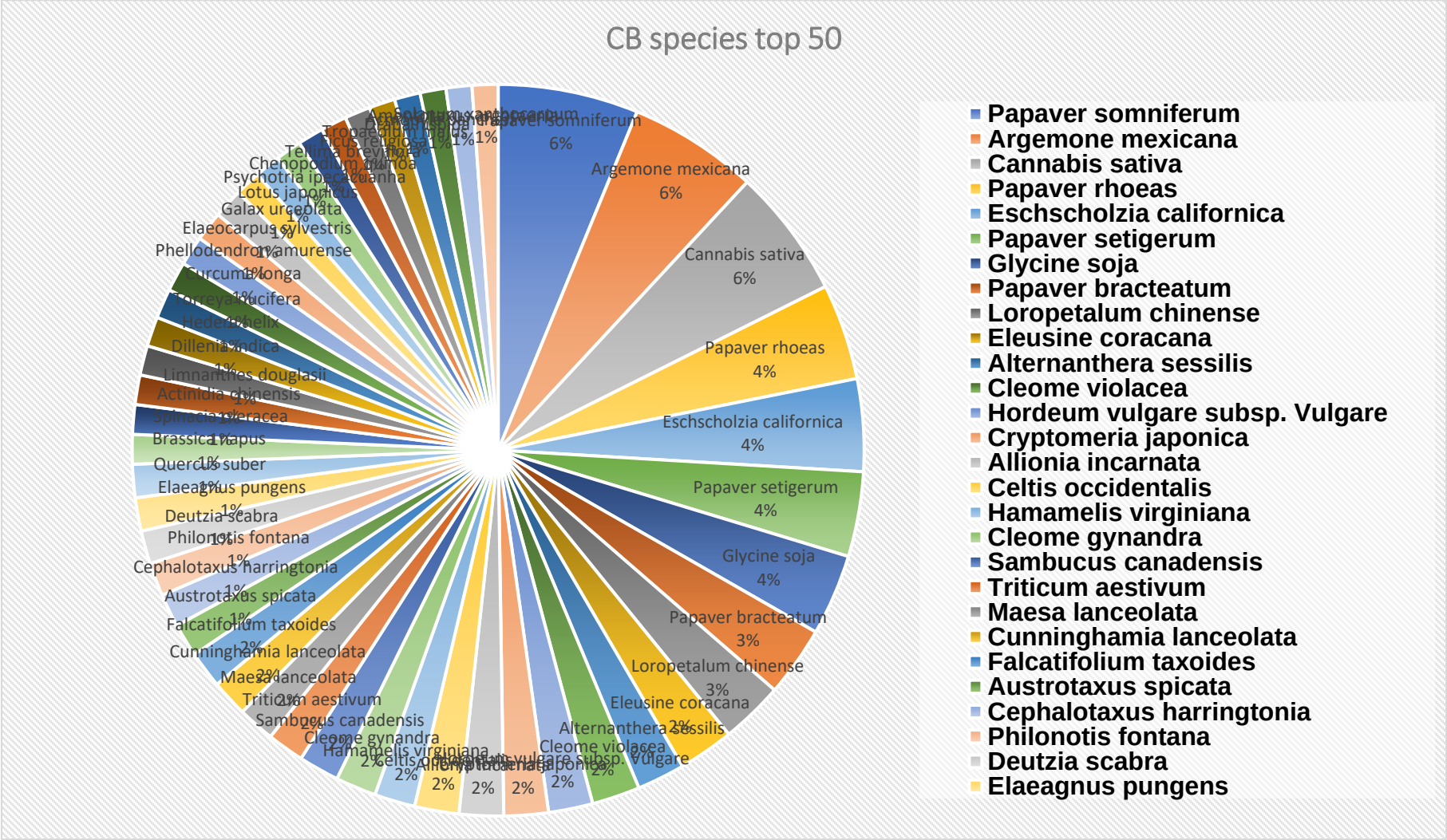
Non chitin-binding

6381 sequences, 1051 species, 411 families, 163 orders

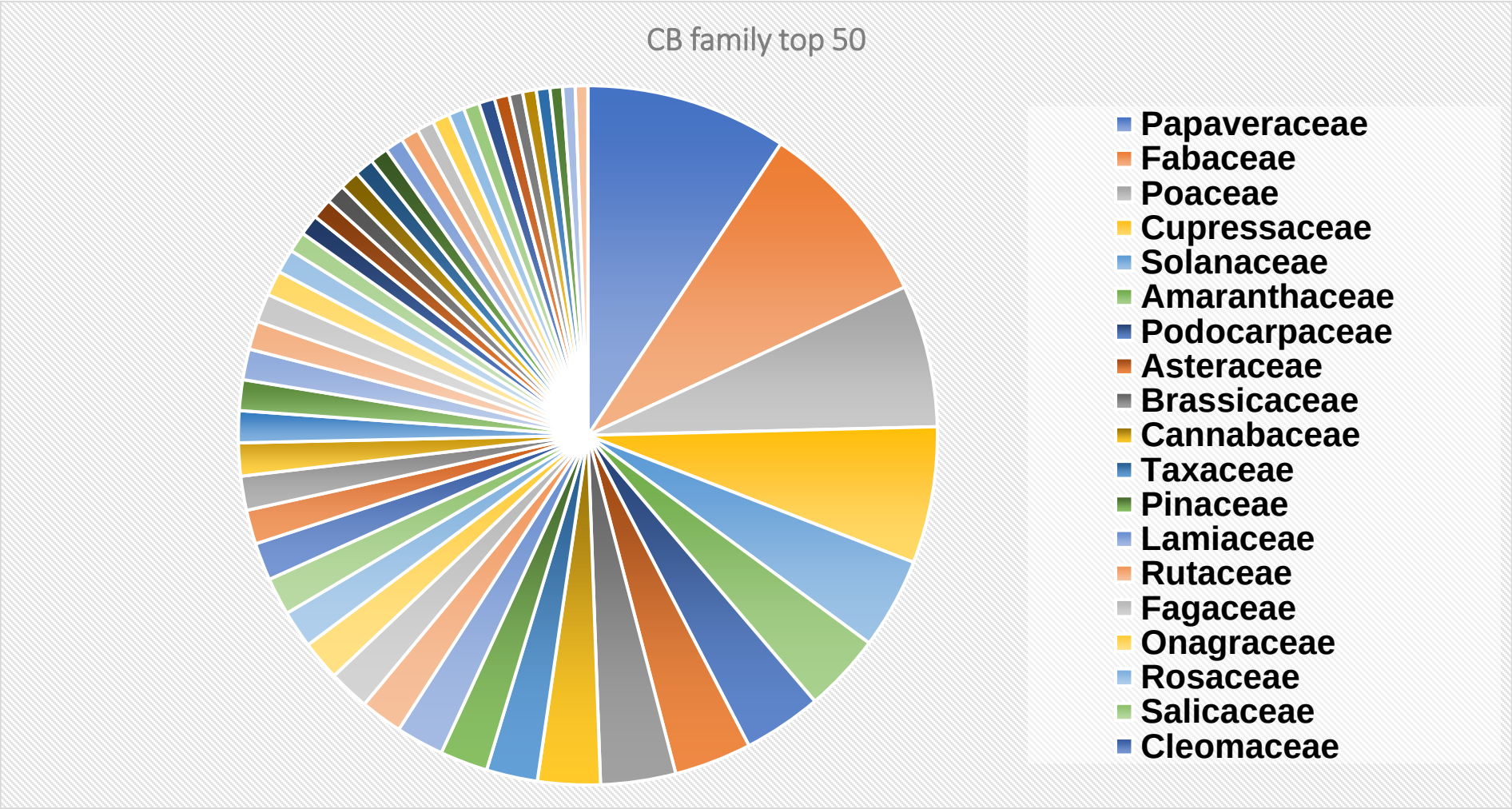
Combined

9298 sequences, 1377 species, 444 families, 166 orders

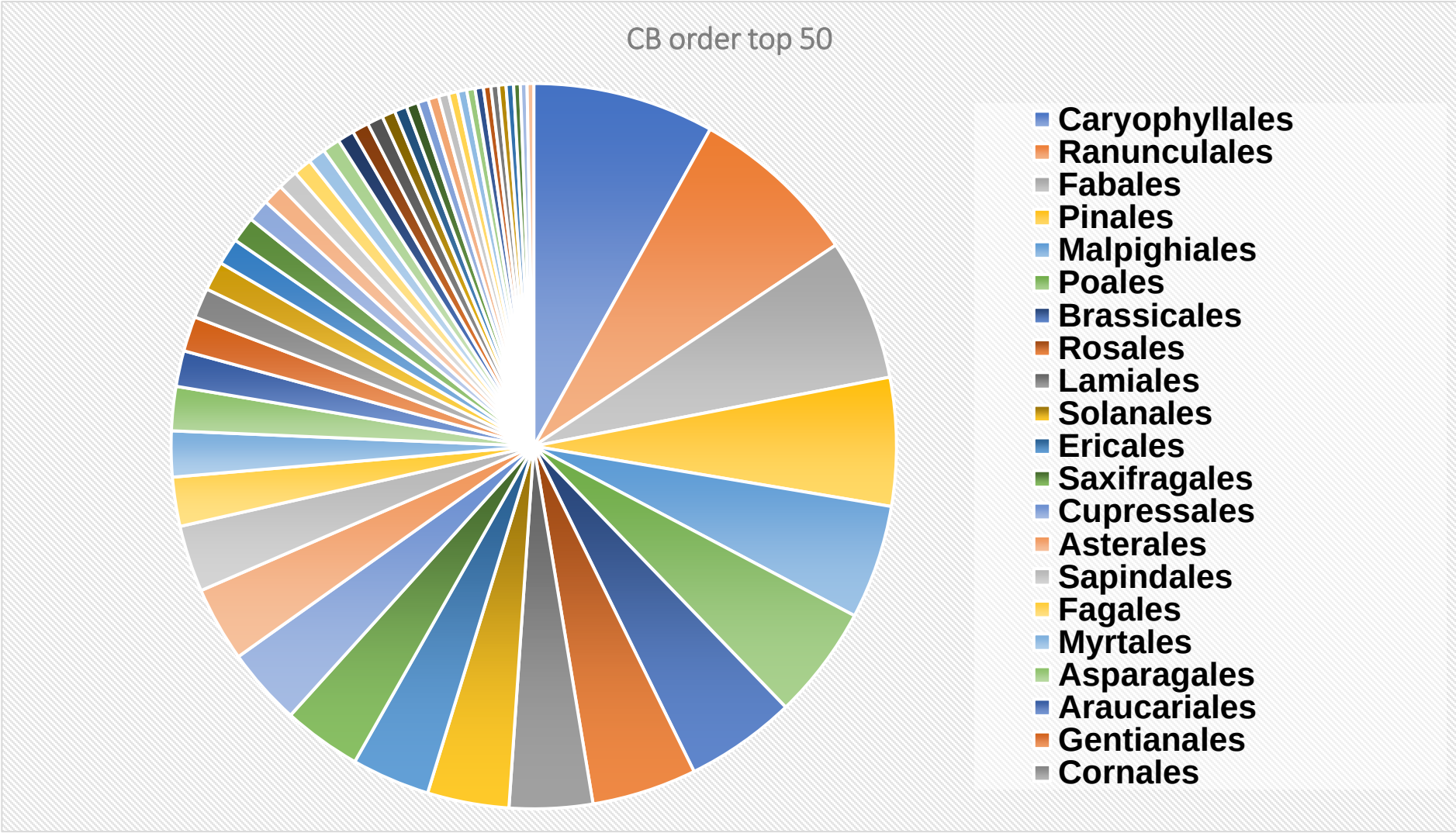
HLP distribution, CB species top 50



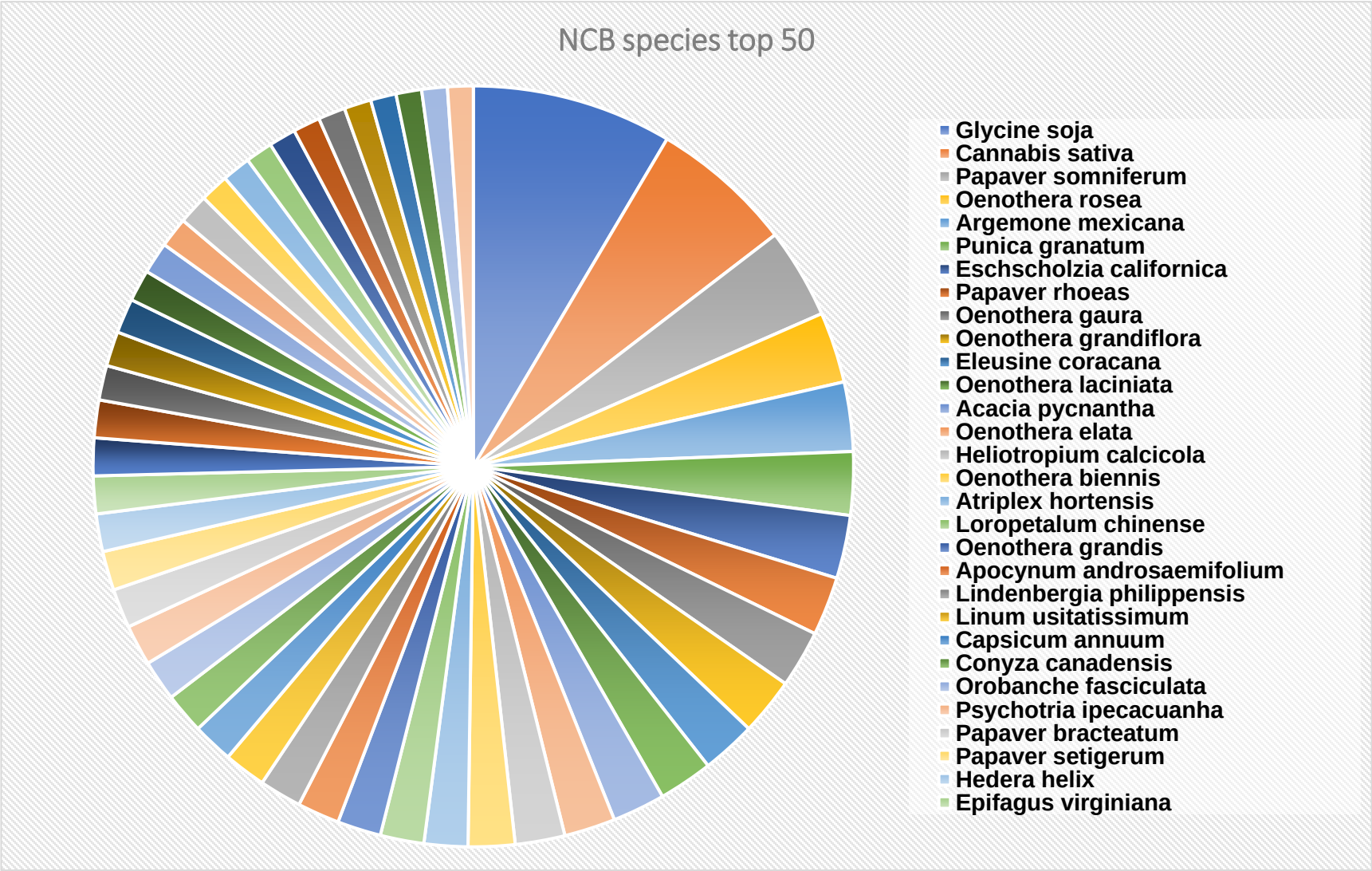
HLP distribution, CB family top 50



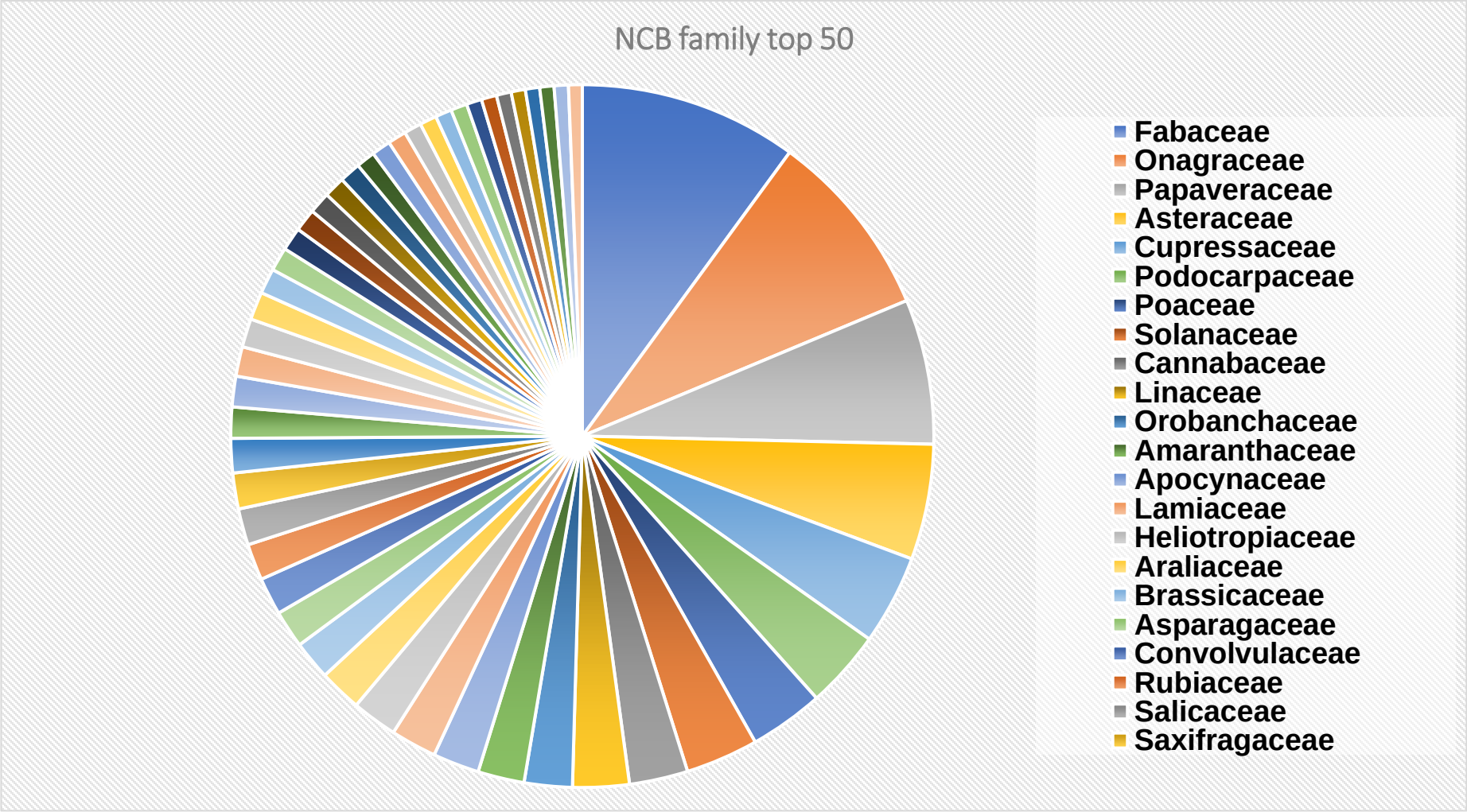
HLP distribution, CB order top 50



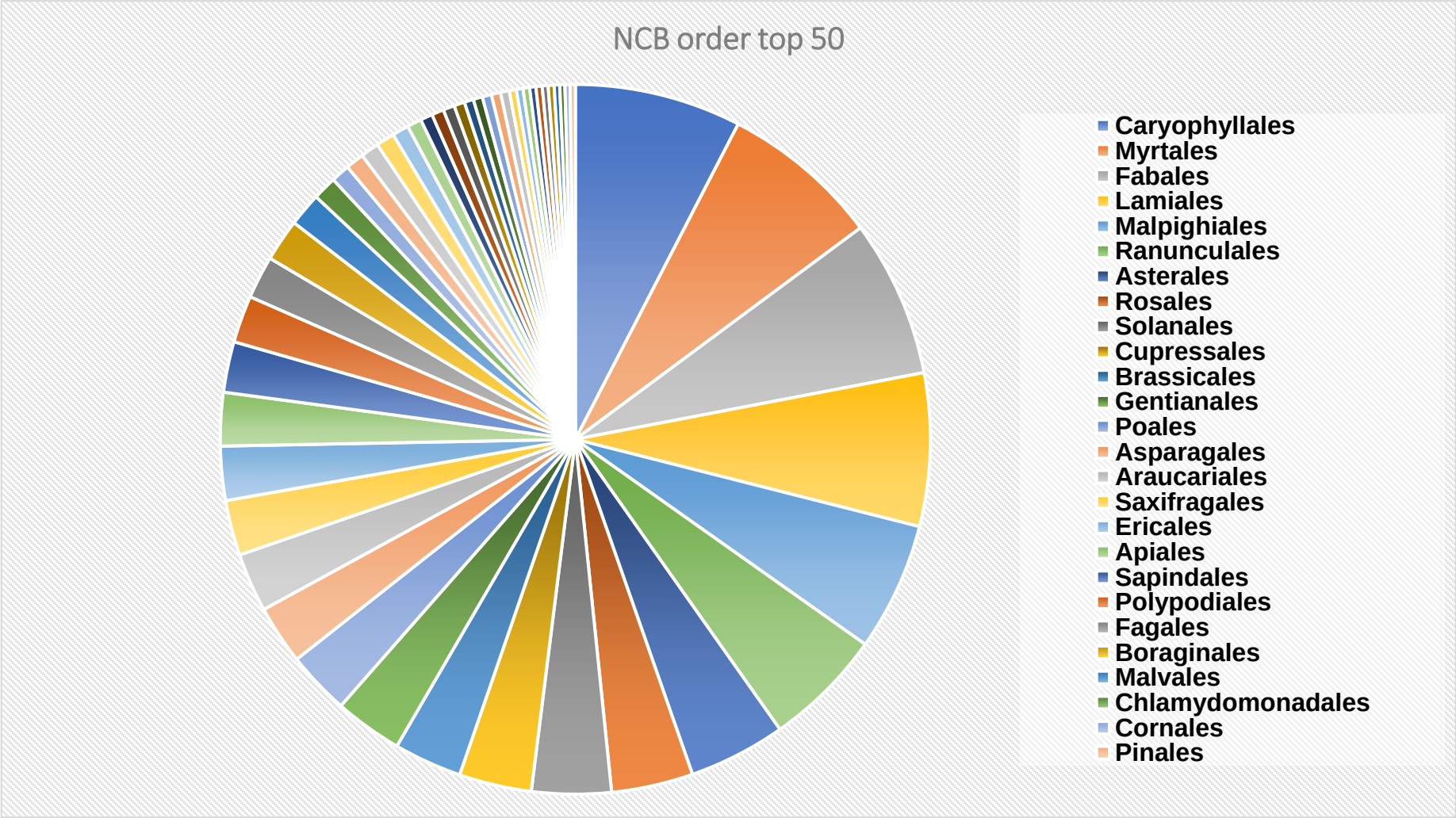
HLP distribution, NCB species top 50



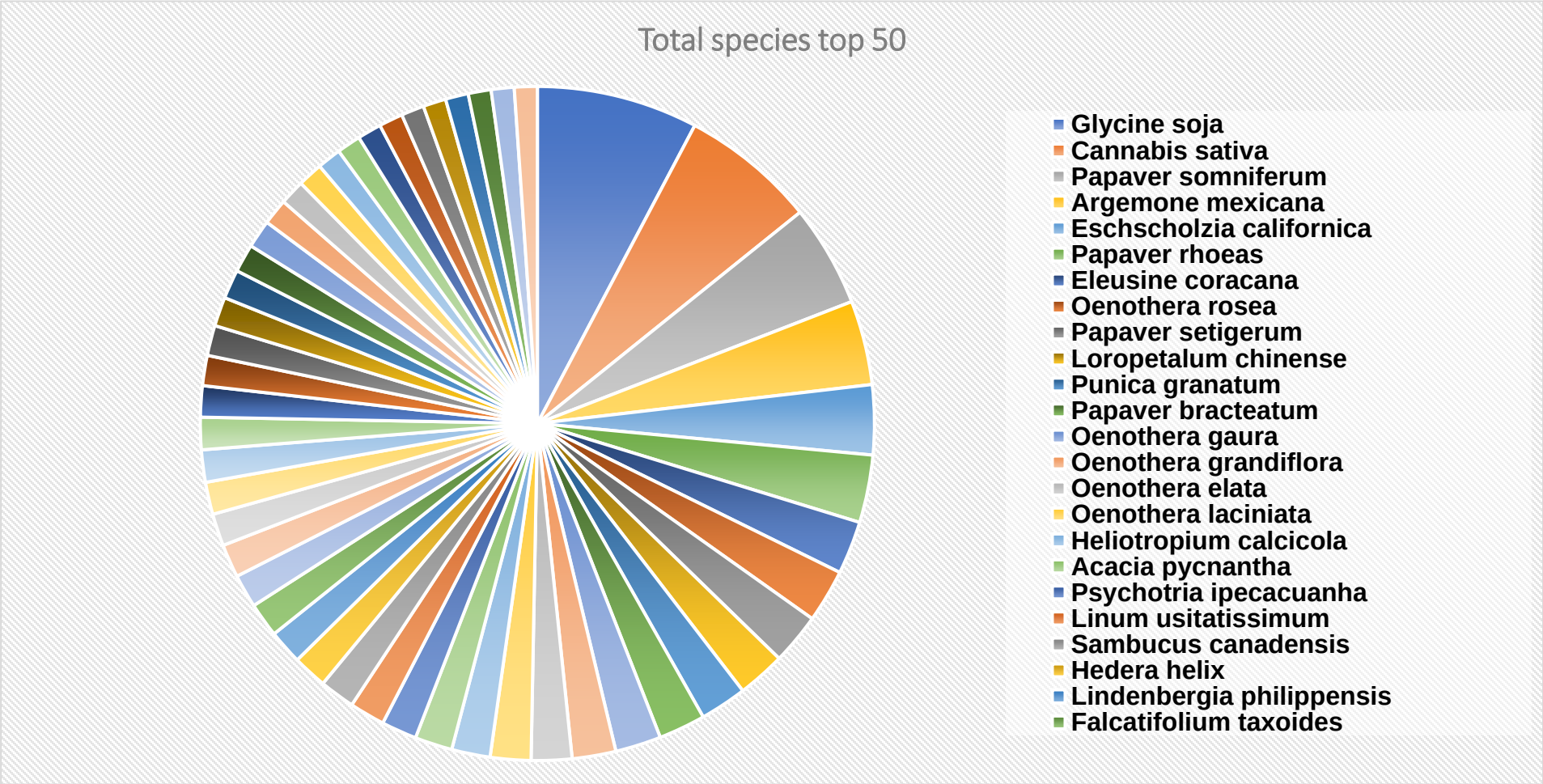
HLP distribution, NCB family top 50



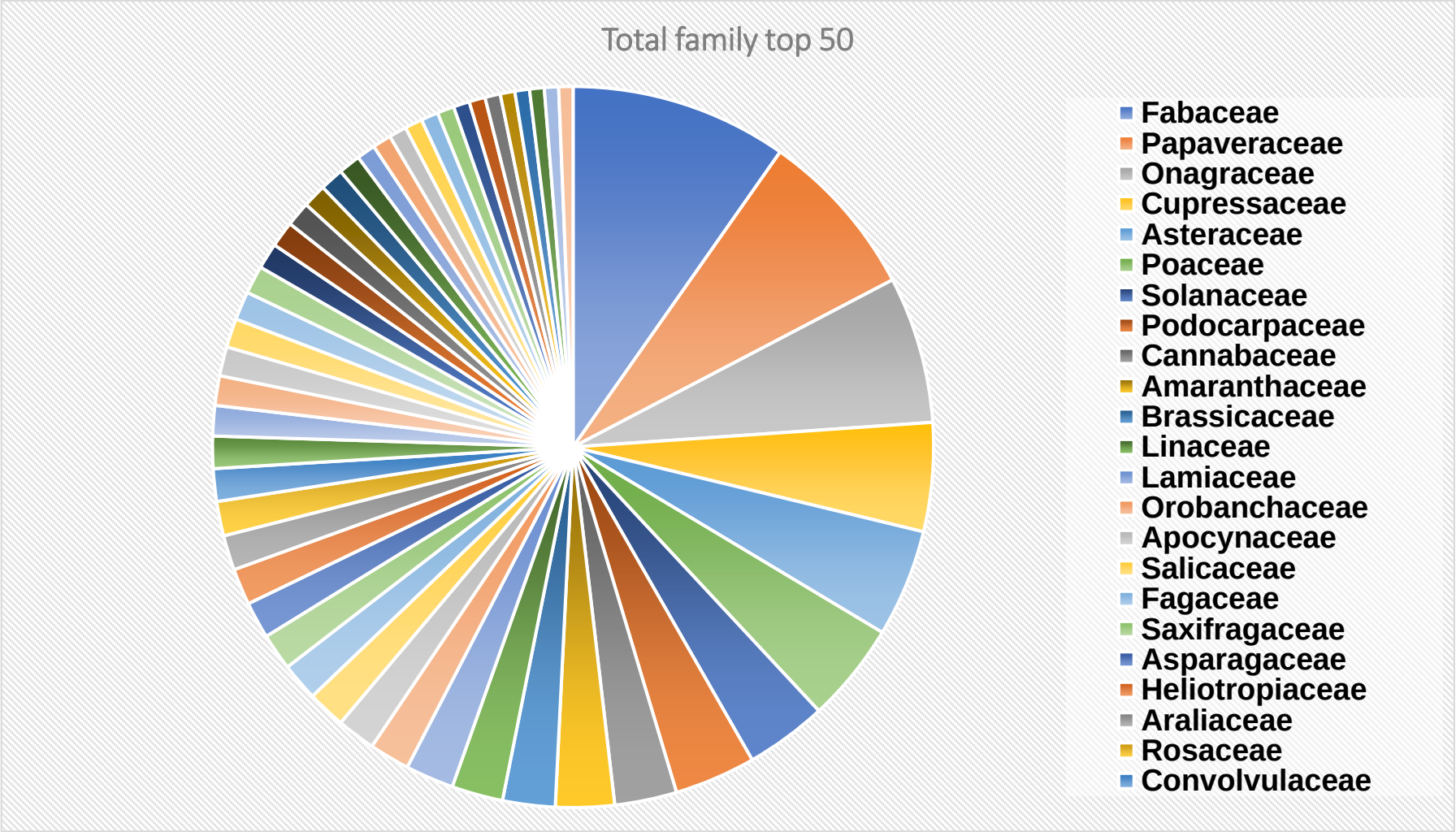
HLP distribution, NCB order top 50



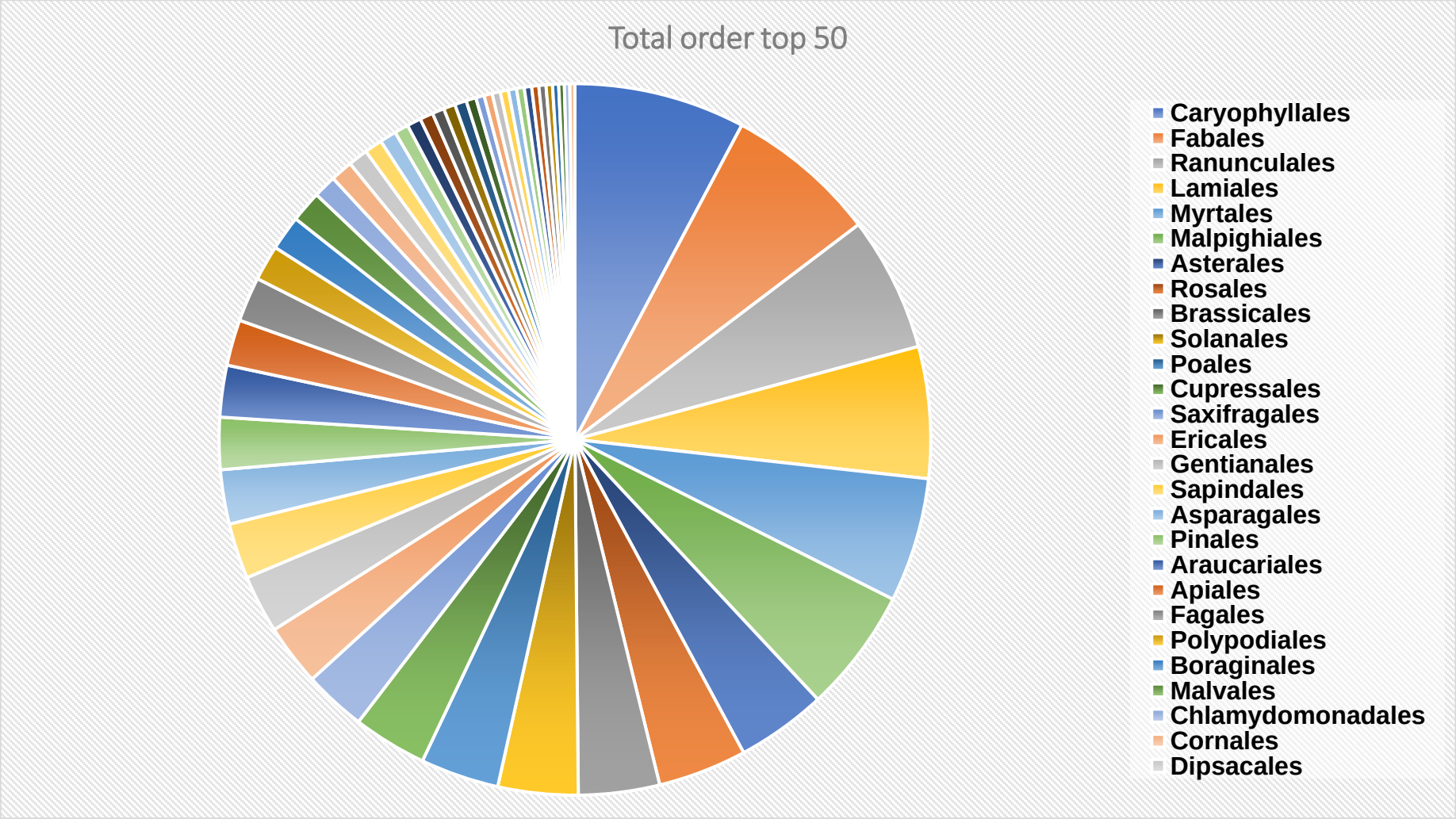
HLP distribution, total species top 50



HLP distribution, total family top 50



HLP distribution, total order top 50



Methodologies

- Motif searching
- Data formatting, management and storage
- DNA translation
- De novo assembly

Motif searching

1. Convert motif into PROSITE motif pattern

```
Caripe 1     GVIPCGESCVFIPCIST-VIGCSCKNKVCYRN

Caripe 2     G-IPCGESCVFIPCTITALLGCSCSKKVCYKN
Caripe 3      XIPCGESCVFIPCISAVVGSCSC-NKVCYNN
Caripe 4      -LICSSTCLRIPCLSPR---CTCRHHICYLN
Caripe 5        XCGESCVFIPCFTSV--GCSCKDKVCYRN
Caripe 6     GAI-CTGTCFRNPCLSRR---CTCRHYICYLN
Caripe 7     G-IPCGESCVFIPCTVTALLGCSCKNKVCYRN
Caripe 8     GVIPCGESCVFIPCITAAI-GCSCKKKVCYRN
Caripe 9           XCVFIPCTITALLGCSCSNNVCYKN

Caripe 10    GVIPCGESCVFIPCFSTVI-GCSCKNKVCYRN
Caripe 11    GVIPCGESCVFIPCISTVI-GCSCKKKVCYRN
Caripe 12    GVIPCGESCVFIPCFSSVI-GCSCKNKVCYRN
Caripe 13    G-IPCGESCVFIPCFTSVF-GCSCKDKVCYRN
Psyle E      GVIPCGESCVFIPCISSVL-GCSCKNKVCYRD
CYS              I    II    III      IV V    IV
```

X is anything but cysteine in this case

C X(3) C X(4) C X(4,7) CXC X(3,4) C

PROSITE motif pattern:
C-{C}(3)-C-{C}(4)-C-{C}(4,7)-C-{C}-C-{C}(3,4)-C.

Cyclotides isolated from an ipecac root extract antagonize the corticotropin releasing factor type 1 receptor
Fahradpour M., Keov P., Tognola C., Perez-Santamarina E., McCormick P.J., Ghassempour A., Gruber C.W.
(2017) *Frontiers in Pharmacology*, 8 (SEP) , art. no. 616

Extra info on PROSITE motif pattern

[ALT] stands for Ala or Leu or Thr.
{AM} stands for all any amino acid except Ala and Met.
x(3) corresponds to x-x-x
x(2,4) corresponds to x-x or x-x-x or x-x-x-x
A(3) corresponds to A-A-A
<, > corresponds to N-, or C-terminal

Pattern	Explanation
[AC]-x-V-x(4)-{ED}	[Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}
<A-x-[ST](2)-x(0,1)-V	Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val at the N-terminal of the sequence
<{C}*>	No Cys from the N-terminal to the C-terminal i.e. All sequences that do not contain any Cys.
IIRIFHLRNI	Ile-Ile-Arg-Ils-Phe-His-Leu-Arg-Asn-Ile

ScanProsite - user manual: https://prosite.expasy.org/scanprosite/scanprosite_doc.html

Motif searching

2. Places to search using that PROSITE motif pattern

ExPasy ScanProsite

Searches UniprotKB, PDB and your own data (16 MB max.)

Supports taxonomic filter i.e. Viridiplantae, Poaceae, Homo sapiens, etc

GenomeNet Japan MOTIF Search

Searches RefSeq, in addition to SwissProt and PDB

No taxonomic filter. Use Python or do this manually.

ps_scan

Perl program to search motif from your own data in FASTA format. No size limitations.

Motif searching

ExPasy ScanProsite

This form requires to have JavaScript enabled to work correctly.

This form allows you to scan proteins for matches against the PROSITE collection of motifs as well as against your own patterns.

- ○ Option 1 - Submit PROTEIN sequences to scan them against the PROSITE collection of motifs.
- ◉ **Option 2 - Submit MOTIFS to scan them against a PROTEIN sequence database.**
- ○ Option 3 - Submit PROTEIN sequences and MOTIFS to scan them against each other.

Reset

STEP 1 - Enter a MOTIF or a combination of MOTIFS Examples [help]

```
C-{C}(3)-C-{C}(4)-C-{C}(4,7)-C-{C}-C-{C}(3,4)-C
```

Supported input:

- A PROSITE accession e.g. PS50240 or identifier e.g. TRYPSIN_DOM
- Your own pattern e.g. P-x(2)-G-E-S-G(2)-[AS]

» More

» Options [help]

STEP2 - Select a PROTEIN sequence database [help]

- ◉ UniProtKB
  - ☑ Swiss-Prot ☐ Include isoforms
  - ☑ TrEMBL
- ○ PDB
- ○ Your protein database
- ○ Randomized UniProtKB/Swiss-Prot

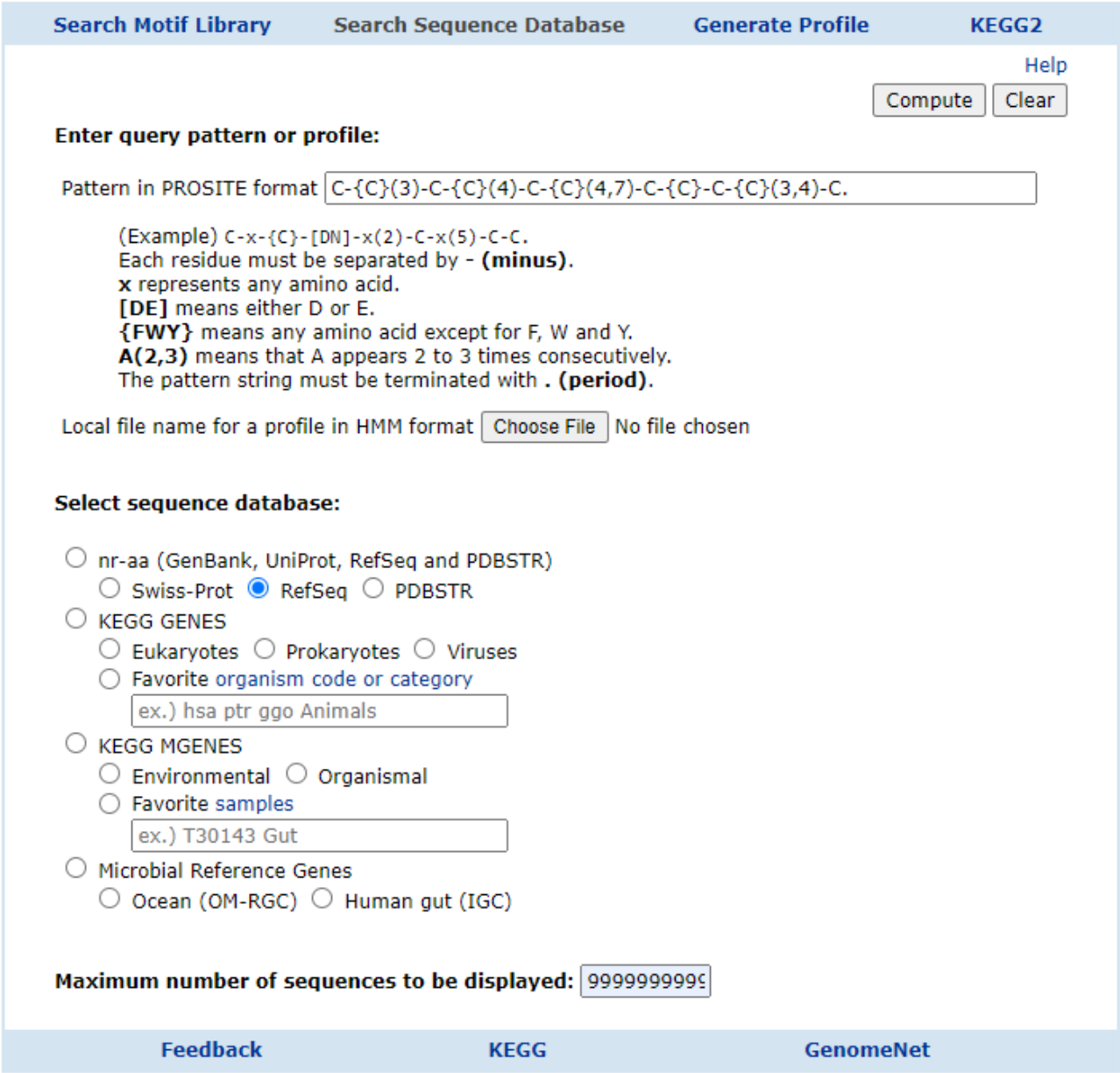
☐ Exclude fragments (concerns UniProtKB only)

Filters « [help]

- On length >= than:
- On length <= than:
- On taxonomy: Viridiplantae any taxonomical term e.g. Homo sapiens, e.g. Fungi; Arthropoda or corresponding TaxID e.g. 9606, e.g. 4751; 6656

Motif searching

GenomeNet Japan
MOTIF Search



Motif searching with ps_scan

```
ps_scan.pl -d cyclotide_syntax.dat translated_frame_1.fa -o fasta >result_frame_1.fa
```

cyclotide_syntax.dat - Notepad

File Edit Format View Help

```
//
ID   CYCLOTIDE; PATTERN.
AC   CYCLO1;
PA   C-{C}(3)-C-{C}(3,5)-C-{C}(4,7)-C-{C}-C-{C}(4)-C.
//
```

c_o6_cyclo_frame2.fa - Notepad

File Edit Format View Help

```
>NODE_65173_length_486_cov_13.228395/124-145 : CYCLO1 CYCLOTIDE
CWWFCSFRDCEGFNCLCRKFSC
```

contigs_output6_aa_frame1.fa - Notepad

File Edit Format View Help

```
>NODE_1_length_1147_cov_64.254578
FRSFI*YQLLFISINAECQINNLTDIILKQPQQIFEVIWVNSFNVMICHFNPQNMLVK*P
SKMNIKKLPIK*SLPNNAPNKFEV*KMFRLIN*VRIGLEC*FITNCCKK*CAFIKNYPRQ
QLIPFPC*PTSINSFLTLEVNSQPALQFLRCPHS*LII*IFKNVSSSDRNS*TVMMIMLV
LNLRSKMRTLIVKVNKARNLQHERERLLK*TGILPEESIQKFSVLLRELHKSYLLFITFN
*RS*SFLLRRRCSSLLNIRN*SKIKIMSCT*LFRMQLLTHHEK*FNIWQDVLTACRKWRQ
NTRWFRC*KH**VHQIRC*IRVALHLAYASAQL*PQWLQRCIDILYKSCRTLLF*QNLTI
LLI*GSNKRT*SLQNP*NCSPIRGSCTKKGFF
>NODE_2_length_119_cov_4.109244
PAGFEKAQRKISISDS*NSI*FVTELPSTQVLFCGPLLNTWKSATCPLL
>NODE_3_length_42_cov_73.523811
*LPFPSRPELQNDC*F*ILSASQG
>NODE_4_length_45_cov_49.400002
*F*ILSASQGMILHSIQKNISPLQK
>NODE_5_length_1450_cov_85.700691
NLPITGLHLLNNHLPFNSSICSYQGCRRHQCLLHNISTYTLISIFKFLLYCWNTSRQMNQ
GSPTTRDNTFLNSGKCSILCIFNPEFAVLKFSFCCSTNLYNCNTTGQFSNPFSKFLRIVY
RICLSQLLL*LCNPSL*LIL*CSISDDGCRVLSDFHLSGNTQILY*CILNQQTQVGCLVL
SSS*NSYILKKGLSSFTKSWSFDGSNIQNASQLVYNKSSQSLSSDIFSNNEQRSSYLCSF
LQQGDYVFYSRDLLICHQYTCILKFNNELLRVGNKLWRDVSSINLHSFLYIYCCIKRRGR
LNGQDPIRSNLVNCISNHGSNSIIISSRDRSNGFDVFTTFNRPGFLF*FINQSLYSFIYA
FLH*NWIGTRCNTK*TKSNDFASKN*RSGRAITRRIISLACNLSNQCSASIFHRICQLNC
PGNGHTIINHLGSPKFIQHNIPSPGTKSETNSISKFINPRLQSRP*LLIEGNFLSRSPHS
KLALCGFIIKALS
>NODE_41_length_148_cov_394.614868
```

Directory

Documents › ps_scan

Search ps_scan

Name	Date modified	Type	Size
usr	5/6/2020 8:10 pm	File folder	
cyclotide_syntax.dat	14/6/2021 3:36 pm	DAT File	1 KB
LICENSE	23/11/2015 10:08 pm	File	18 KB
pfscan	9/12/2003 11:36 pm	File	397 KB
pfscan	28/3/2007 5:18 pm	Application	184 KB
pfsearch	9/12/2003 11:36 pm	File	397 KB
pfsearch	28/3/2007 5:18 pm	Application	183 KB
ps_scan	5/6/2020 8:10 pm	Perl program file	81 KB
psa2msa	9/12/2003 11:36 pm	File	302 KB
psa2msa	28/3/2007 5:18 pm	Application	85 KB
README	23/11/2015 10:05 pm	File	10 KB
README.pfs	23/11/2015 10:19 pm	PFS File	7 KB
README	26/10/2004 4:12 pm	PICS Rules File	7 KB

gical University

gical University (1)

Data formatting, management and storage

We often store data and do analysis in Excel, but many bioinformatics programs handle data in FASTA format. Therefore we need tools to convert formatting, and we want them to work fast.

fasta_maker.py

XP_023740439	Lactuca sativa	Asteracea	Asterales	CAGGADGGKGCPCAGGVCCGKGCPCAGGAEGGKGCPCADGSEGGKGC
XP_023747635	Lactuca sativa	Asteracea	Asterales	CPANDAGCSAKDSGCPGGCCPAKDSGCPGGGCPDGGCPAKDGGCSGGSC
XP_010474430	Camelina sativa	Brassicace	Brassicale	CRIGLGNCGESCNDQCCDAKCAQSYNSGHGICDTHNEISLCQCKYPC
XP_010478289	Camelina sativa	Brassicace	Brassicale	CSGCEGACGRGERGPPPPCCKKDDDCKMHCPEGGYCSNNCEC
XP_010482005	Camelina sativa	Brassicace	Brassicale	CVGAIDMCTDTCPLSCCDRLCAIKYKNGRGGCVDYLGYRMCTCEYSC
XP_023637975	Capsella rubella	Brassicace	Brassicale	CVAHGGGKRCVVAGCTKSARGRTDCCVKHGGGKRCKSDGCEKSAQGSTDF
XP_021893127	Carica papaya	Caricaceae	Brassicale	CNVGIDRCTAACNEKCCDENCMSKFPEHLNCHGSCSDLLPPQFSVCIC
XP_028202525	Glycine soja	Fabaceae	Fabales	CNNGLGVCTLQCGDACCNANCARKYNQGTGMCSTIGNNNLCTCQYRC
XP_017421515	Vigna angularis	Fabaceae	Fabales	CNGSQGLCNDNCDEGCCNSKCAAKYKDGVGTCKLYVEGFNFCICKYAC

From Excel

A0A270R5T5 Panicum hallii Poaceae Poales CGQHAGGMLCPHNLCCSRSGLCGLGADYCGAGCQSGACCPSLRC
A0A270R5T5 Panicum hallii Poaceae Poales CGQHAGGMLCPHNLCCSRSGLCGLGADYCGAGCQSGACCPSLRC
A0A270R5T5 Panicum hallii Poaceae Poales CGQHAGGMLCPHNLCCSRSGLCGLGADYCGAGCQSGACCPSLRC
A0A270R5T5 Panicum hallii Poaceae Poales CGQHAGGMLCPHNLCCSRSGLCGLGADYCGAGCQSGACCPSLRC

Save as text

fasta_maker.py

>A0A270R5T5|Panicum hallii|Poaceae|Poales
CGQHAGGMLCPHNLCCSRSGLCGLGADYCGAGCQSGACCPSLRC
>A0A270R5T5|Panicum hallii|Poaceae|Poales
CGQHAGGMLCPHNLCCSRSGLCGLGADYCGAGCQSGACCPSLRC
>A0A270R5T5|Panicum hallii|Poaceae|Poales
CGQHAGGMLCPHNLCCSRSGLCGLGADYCGAGCQSGACCPSLRC
>A0A270R5T5|Panicum hallii|Poaceae|Poales
CGQHAGGMLCPHNLCCSRSGLCGLGADYCGAGCQSGACCPSLRC

Convert to FASTA

fasta_to_tab_text.py

From FASTA

```
>A0A270R5T5|Panicum hallii|Poaceae|Poales
CGQHAGGMLCPHNLCCSRSGLCGLGADYCGAGCQSGACCPSLRC
>A0A270R5T5|Panicum hallii|Poaceae|Poales
CGQHAGGMLCPHNLCCSRSGLCGLGADYCGAGCQSGACCPSLRC
>A0A270R5T5|Panicum hallii|Poaceae|Poales
CGQHAGGMLCPHNLCCSRSGLCGLGADYCGAGCQSGACCPSLRC
>A0A270R5T5|Panicum hallii|Poaceae|Poales
CGQHAGGMLCPHNLCCSRSGLCGLGADYCGAGCQSGACCPSLRC
```

fasta_to_tab_text.py

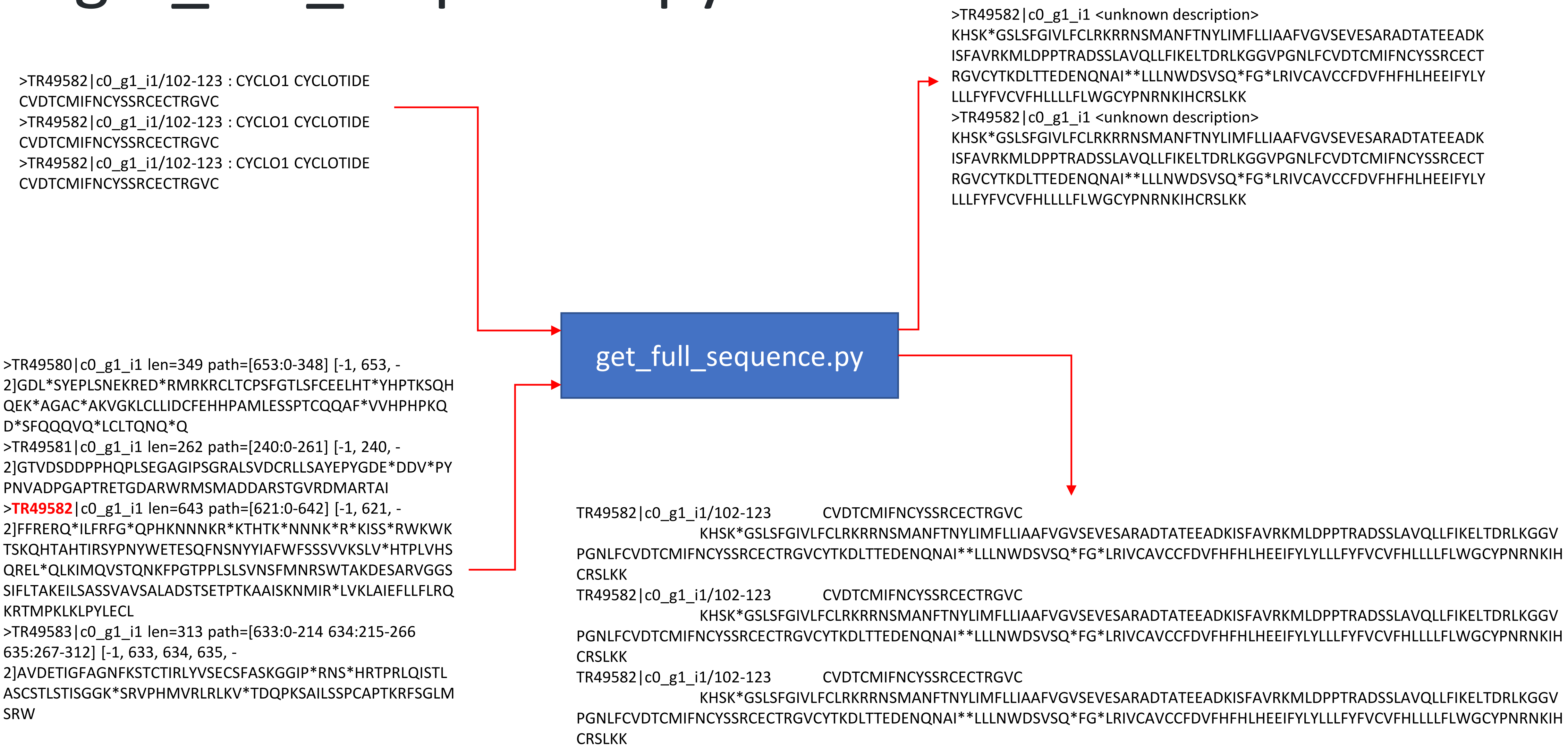
```
A0A270R5T5 Panicum hallii Poaceae Poales CGQHAGGMLCPHNLCCSRSGLCGLGADYCGAGCQSGACCPSLRC
A0A270R5T5 Panicum hallii Poaceae Poales CGQHAGGMLCPHNLCCSRSGLCGLGADYCGAGCQSGACCPSLRC
A0A270R5T5 Panicum hallii Poaceae Poales CGQHAGGMLCPHNLCCSRSGLCGLGADYCGAGCQSGACCPSLRC
A0A270R5T5 Panicum hallii Poaceae Poales CGQHAGGMLCPHNLCCSRSGLCGLGADYCGAGCQSGACCPSLRC
```

Convert to TDT form

Paste into Excel

XP_023740439	Lactuca sativa	Asteracea	Asterales	CAGGADGGKGCPCAGGVCCGKGCPCAGGAEGGKGCPCADGSEGGKGC
XP_023747635	Lactuca sativa	Asteracea	Asterales	CPANDAGCSAKDSGCPGGCCPAKDSGCPGGGCPDGGCPAKDGGCSGGSC
XP_010474430	Camelina sativa	Brassicace	Brassicale	CRIGLGNCGESCNDQCCDAKCAQSYNSGHGICDTHNEISLCQCKYPC
XP_010478289	Camelina sativa	Brassicace	Brassicale	CSGCEGACGRGERGPPPPCCKKDDDCKMHCPEGGYCSNNCEC
XP_010482005	Camelina sativa	Brassicace	Brassicale	CVGAIDMCTDTCPLSCCDRLCAIKYKNGRGGCVDYLGYRMCTCEYSC
XP_023637975	Capsella rubella	Brassicace	Brassicale	CVAHGGGKRCVVAGCTKSARGRTDCCVKHGGGKRCKSDGCEKSAQGSTDF
XP_021893127	Carica papaya	Caricacea	Brassicale	CNVGIDRCTAACNEKCCDENCMSKFPEHLNCHGSCSDLLPPQFSVCIC
XP_028202525	Glycine soja	Fabaceae	Fabales	CNNGLGVCTLQCGDACCNANCARKYNQGTGMCSTIGNNNLCTCQYRC
XP_017421515	Vigna angularis	Fabaceae	Fabales	CNGSQGLCNDNCDEGCCNSKCAAKYKDGVGTCKLYVEGFNFCICKYAC

get_full_sequence.py



```
>TR49582|c0_g1_i1/102-123 : CYCLO1 CYCLOTIDE
CVDTCMIFNCYSSRCECTRGVC
>TR49582|c0_g1_i1/102-123 : CYCLO1 CYCLOTIDE
CVDTCMIFNCYSSRCECTRGVC
>TR49582|c0_g1_i1/102-123 : CYCLO1 CYCLOTIDE
CVDTCMIFNCYSSRCECTRGVC
```

```
>TR49580|c0_g1_i1 len=349 path=[653:0-348] [-1, 653, -
2]GDL*SYEPLSNEKRED*RMRKRCLTCPSFGTLSFCEELHT*YHPTKSQH
QEK*AGAC*AKVGKLCLLIDCFEHHPAMLESSPTCQQAF*VVHPHPKQ
D*SFQQQVQ*LCLTQNQ*Q
>TR49581|c0_g1_i1 len=262 path=[240:0-261] [-1, 240, -
2]GTVDSDDPPHQPLSEGAGIPSGRALSVDCRLLSAYEPYGDE*DDV*PY
PNVADPGAPTRETGDARWRMSMADDARSTGVRDMARTAI
>TR49582|c0_g1_i1 len=643 path=[621:0-642] [-1, 621, -
2]FFRERQ*ILFRFG*QPHKNNNKR*KTHTK*NNNK*R*KISS*RWKWK
TSKQHTAHTIRSYPNYWETESQFNSNYYIAFWFSSSVVKSLV*HTPLVHS
QREL*QLKIMQVSTQNKFPGTPPLSLSVNSFMNRSWTAKDESARVGGS
SIFLTAKEILSASSVAVSALADSTSETPTKAAISKNMIR*LVKLAIEFLLFLRQ
KRTMPKLKLPYLECL
>TR49583|c0_g1_i1 len=313 path=[633:0-214 634:215-266
635:267-312] [-1, 633, 634, 635, -
2]AVDETIGFAGNFKSTCTIRLYVSECSFASKGGIP*RNS*HRTPRLQISTL
ASCSTLSTISGGK*SRVPHMVRLRLKV*TDQPKSAILSSPCAPTKRFSGLM
SRW
```

get_full_sequence.py

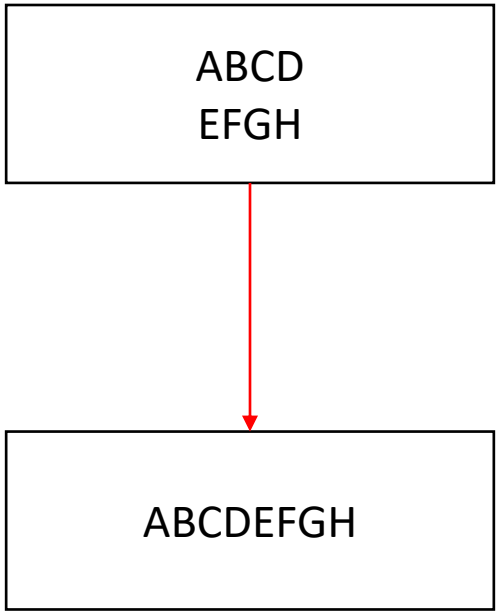
```
>TR49582|c0_g1_i1 <unknown description>
KHSK*GSLSFGIVLFCLRKRRNSMANFTNYLIMFLLIAAFVGVSEVESARADTATEEADK
ISFAVRKMLDPPTRADSSLAVQLLFIKELTDRLKGGVPGNLFCVDTCMIFNCYSSRCECT
RGVCYTKDLTTEDENQNAI**LLLNWDSVSQ*FG*LRIVCAVCCFDVFHFHLHEEIFYLY
LLLFYFVCVFHLLLLFLWGCYPNRNKIHCRSLKK
>TR49582|c0_g1_i1 <unknown description>
KHSK*GSLSFGIVLFCLRKRRNSMANFTNYLIMFLLIAAFVGVSEVESARADTATEEADK
ISFAVRKMLDPPTRADSSLAVQLLFIKELTDRLKGGVPGNLFCVDTCMIFNCYSSRCECT
RGVCYTKDLTTEDENQNAI**LLLNWDSVSQ*FG*LRIVCAVCCFDVFHFHLHEEIFYLY
LLLFYFVCVFHLLLLFLWGCYPNRNKIHCRSLKK
```

```
TR49582|c0_g1_i1/102-123          CVDTCMIFNCYSSRCECTRGVC
                    KHSK*GSLSFGIVLFCLRKRRNSMANFTNYLIMFLLIAAFVGVSEVESARADTATEEADKISFAVRKMLDPPTRADSSLAVQLLFIKELTDRLKGGV
PGNLFCVDTCMIFNCYSSRCECTRGVCYTKDLTTEDENQNAI**LLLNWDSVSQ*FG*LRIVCAVCCFDVFHFHLHEEIFYLYLLLFYFVCVFHLLLLFLWGCYPNRNKIH
CRSLKK
TR49582|c0_g1_i1/102-123          CVDTCMIFNCYSSRCECTRGVC
                    KHSK*GSLSFGIVLFCLRKRRNSMANFTNYLIMFLLIAAFVGVSEVESARADTATEEADKISFAVRKMLDPPTRADSSLAVQLLFIKELTDRLKGGV
PGNLFCVDTCMIFNCYSSRCECTRGVCYTKDLTTEDENQNAI**LLLNWDSVSQ*FG*LRIVCAVCCFDVFHFHLHEEIFYLYLLLFYFVCVFHLLLLFLWGCYPNRNKIH
CRSLKK
TR49582|c0_g1_i1/102-123          CVDTCMIFNCYSSRCECTRGVC
                    KHSK*GSLSFGIVLFCLRKRRNSMANFTNYLIMFLLIAAFVGVSEVESARADTATEEADKISFAVRKMLDPPTRADSSLAVQLLFIKELTDRLKGGV
PGNLFCVDTCMIFNCYSSRCECTRGVCYTKDLTTEDENQNAI**LLLNWDSVSQ*FG*LRIVCAVCCFDVFHFHLHEEIFYLYLLLFYFVCVFHLLLLFLWGCYPNRNKIH
CRSLKK
```

Some nifty macros for Word and Excel

Excel macro for removing newlines:

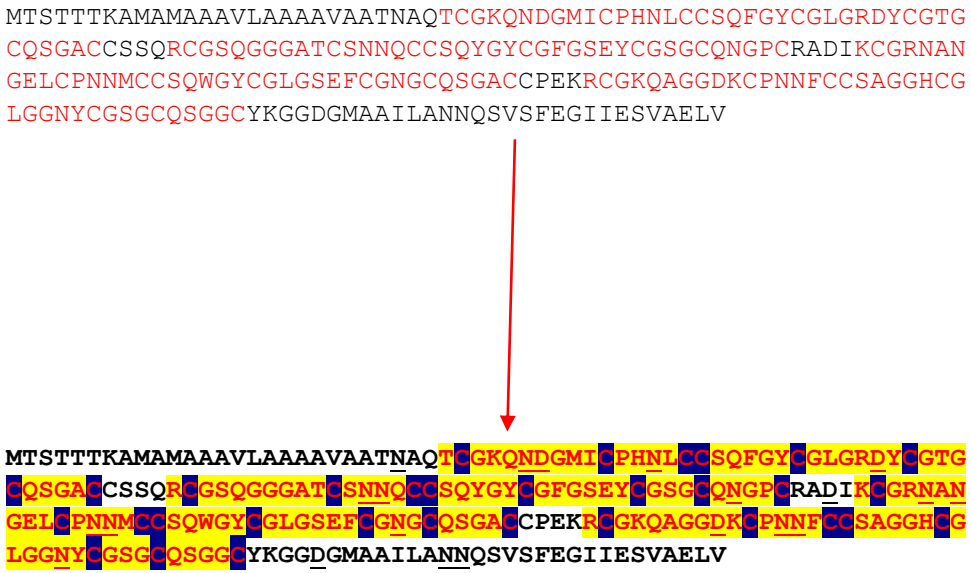
```
Sub remove_newlines()
'
' remove_newlines Macro
' allahu
'
' Keyboard Shortcut: Ctrl+r
'
    Dim MyRange As Range
    Application.ScreenUpdating = False
        Application.Calculation = xlCalculationManual

        For Each MyRange In ActiveSheet.UsedRange
            If 0 < InStr(MyRange, Chr(10)) Then
                MyRange = Replace(MyRange, Chr(10), "")
            End If
        Next

    Application.ScreenUpdating = True
    Application.Calculation = xlCalculationAutomatic
End Sub
```



Part 2 macro

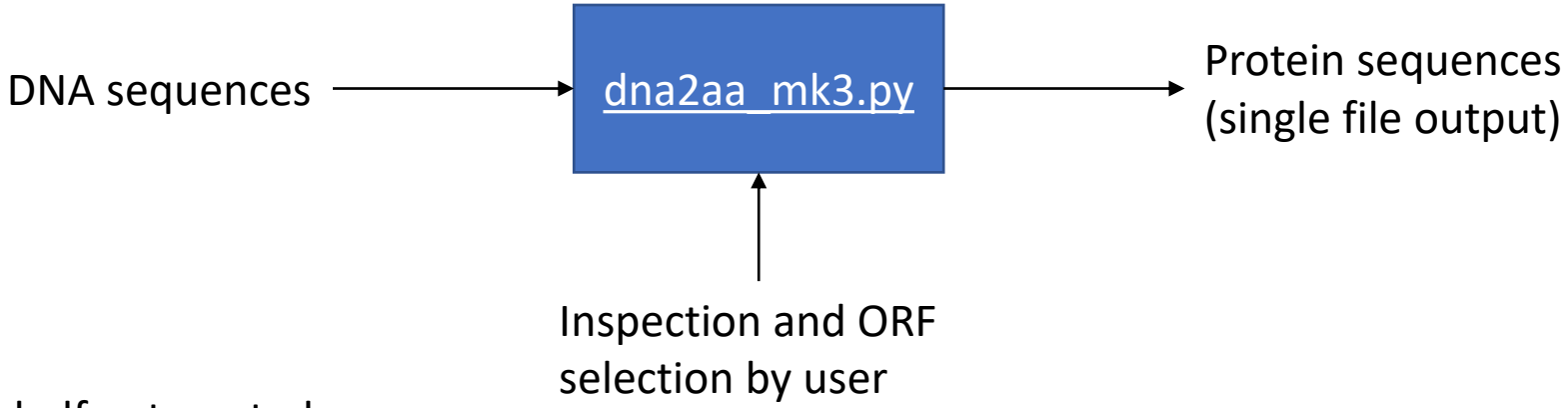
Word macro for highlighting peptide regions and cysteines, and underline N/D everywhere:

```
Sub ND_underliner()
' ND_underliner Macro
   Dim counter As Integer
   For counter = 1 To Len(Selection)
      Selection.Characters(counter).Font.Bold = True

      If Selection.Characters(counter).Font.ColorIndex = "6" Then
         Selection.Characters(counter).HighlightColorIndex = "7"
         If Selection.Characters(counter).Text = "C" Then
            Selection.Characters(counter).HighlightColorIndex = "9"
         End If
      End If

      If Selection.Characters(counter).Text = "N" Or Selection.Characters(counter).Text = "D" Then
         Selection.Characters(counter).Font.Underline = True
      End If
   Next
End Sub
```



DNA translation

- For use on transcriptome results received from BGI or similar
- DNA translation is easy until you have 56045 sequences to translate
- Python + BioPython library can help automate the process



Type 1: half automated

DNA translation

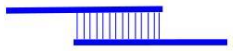
Type 2: fully automated

De novo assembly (not so important)

- If you download transcriptomes as sequencing data, in other words, an **Illumina Genome Analyzer II paired end sequencing data as compared to you sending sample to BGI, you need to do de novo assembly**

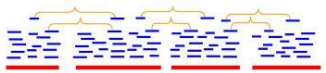
Conceptual steps in *de novo* assembly

1. Find reads that overlap by a specified number of bases (the k-mer size)



2. Merge overlapping, "good" reads into longer contigs

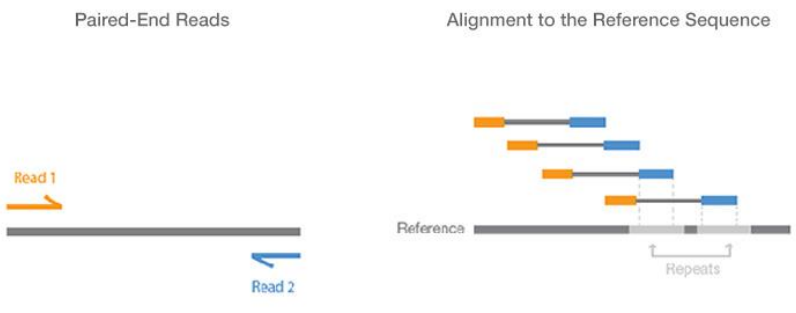
3. Link contigs to form scaffolds using paired-end information



Diagrams from Serafim Batzoglou, Stanford

velvetoptimizer

System requirements: Linux OS and about >16GB RAM



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Paired end reads

```
VelvetOptimiser -f "-shortPaired -fastq -separate reads_R1.fq reads_R2.fq" -s 15 -e 75 -d output_directory
```

Start k-mer, end k-mer.
I use 15 to 75 to ensure
good coverage