

棉纤维伸长促进基因KCS的 生物信息研究

The bioinformatic research of cotton fibre elongation-promoting

gene KCS

报告人:黄盖 组员:高晓雪、刘丹、屈颖 2014/1/10



3

Content

- Background and information collection
- Genome-wide analysis
 - Phylogeny construction and
 - chromosome location
- 3D structure prediction
- Summary



Why we choose cotton gene KCS?















About cotton evolution



(Andrew H. et al., 2012)

- Cotton is an important economic crop.
- Gossypium genus contains tetraploid and
 diploid species. The tetraploid cotton species
 AD-genome (such as G. hirsutum) are thougth
 to have formed by an allopolyploidzation of
 A and D-genome species.
- G. hirsutum (AD-genome) produce most widely used natural fiber, but G. raimondii (D-genome) is fibreless.
- The draft genome of G. raimondii is sequenced by our lab.(Kunbo Wang. et al.,2012)



Previous Research of Our Lab

Ethylene



(Shi *et al*., 2006)



Previous Research of Our Lab

Very Long Chain Fatty Acids (VLCFAs)





About VLCFAs



⁽Riezman, 2007, revised)

- VLCFAs (>18 carbons are essential components of plant lipids, suberins and cuticular waxes.
- VLCFAs is synthesized by four successive enzyme reactions, including KCS,KCR, HCD and ECR
- Previous results in our lab reveal that VCFAs may promote cotton fiber elongation by regulating ethylene biosynthesis.



Gene names	Alias	Entry name	Length of amino acids
KCS1	EL1 At1g01120 T25K16.11	<u>Q9MAM3_</u> ARATH	528
KCS2	At4g34510 T4L20.90	<u>065677</u> _ARATH	487
KCS3	At1g07720 F24B9.18	Q9LQP8_ARATH	478
KCS4	At1g19440 F18O14.21	Q9LN49_ARATH	516
KCS5	CER60 At1g25450 F2J7.9	Q9C6L5_ARATH	492
KCS6	CER6 EL6 At1g68530 T26J14.10	Q9XF43_ARATH	497
KCS7	At1g71160 F23N20.15	<u>Q9C992</u> ARATH	460
KCS8	At2g15090 T15J14.13	Q4V3C9_ARATH	481
KCS9	At2g16280 F16F14.22	Q9SIX1_ARATH	512
KCS10	FDH EL4 At2g26250 T1D16.11	Q570B4_ARATH	550
KCS11	At2g26640 F18A8.1	<u>048780</u> _ARATH	509
KCS12	At2g28630 T8O18.8	Q9SIB2_ARATH	476
KCS13	HIC At2g46720 T3A4.10	<u>Q9ZUZ0</u> ARATH	466
KCS14	At3g10280 F14P13.12	Q9SS39_ARATH	459
KCS15	At3g52160 F4F15.270	Q9SUY9_ARATH	451
KCS16	EL2 At4g34250 F10M10.20	Q9SYZ0_ARATH	493
KCS17	At1g04220 F20D22.1	Q5XEP9_ARATH	528
KCS18	FAE1 At4g34520 T4L20.100	Q38860_ARATH	506
KCS19	At5g43760 MQD19.11	Q9FG87_ARATH	529
KCS20	At5g49070 K20J1.4	Q9FH27_ARATH	464
KCS21	At5g04530 T32M21.130	Q9LZ72_ARATH	464

The database contains 21 reviewed sequences of Arabidopsis thaliana and they belong to FAE protein family. Besides, none of their protein has 3D structure .

Na	mes	Alias	Entry name	Family	Length of AA
K	CS2		A9XUG6_GOSHI		529
	CS1	Gr10021139	M9Z5H0_GOSRA		533
K	CS2	Gr10031660	M9Z380_GOSRA		529
K	CS3	Gr10022901	M9ZC32_GOSRA		466
K	CS4	Gr10018297	M9Z385_GOSRA		504
K	CS5	Gr10014500	M9ZC12_GOSRA		496
K	CS6	Gr10032226	M9ZC26_GOSRA		467
K	CS7	Gr10016173	M9ZA63_GOSRA		510
K	CS8	Gr10026783	M9Z6N0_GOSRA	1	535
K	CS9	Gr10018147	M9Z5I3_GOSRA		504
KC	S10	Gr10015926	M9Z5H7_GOSRA		531
КС	CS11	Gr10017624	M9ZC02_GOSRA	1	510
К	CS	Gr10018148	M9Z5I6_GOSRA	1	418
KC	S13	Gr10031991	M9Z390_GOSRA		492
КС	S15	Gr10032475	M9ZC21_GOSRA		462
КС	S16	Gr10000033	M9Z6L3_GOSRA		390
KC	S17	Gr10028536	M9ZC07_GOSRA		537
KC	S18	Gr10034062	M9ZA66_GOSRA		515
КС	KCS19 Gr10010162	M9Z6K4_GOSRA	TALLCOIL APPA	512	
КС	S20	Gr10019136	M9Z3A9_GOSRA		437
КС	S21	Gr10017854	M9Z397_GOSRA		457
K	CS	Gr10019659	M9ZA81_GOSRA		439
К	CS	Gr10004055	M9Z6K7_GOSRA		533
K	CS	Gr10036653	M9ZC17_GOSRA		515
K	CS	Gr10040229	M9Z5K7_GOSRA		501
K	CS	Gr10018150	M9ZA87_GOSRA		493
K	CS	Gr10018149	M9ZA72_GOSRA		504
K	CS	Gr10040228	M9Z3B3_GOSRA		501
K	CS	Gr10019657	M9Z6M6_GOSRA		424
K	CS	Gr10019616	M9Z5J9_GOSRA		449
K	CS	Gr10033317	M9Z3A3_GOSRA		274
K	CS	Gr10018195	M9Z6L8_GOSRA		285
KC	S14	4 Gr10009484 M9Z5J2_GOSRA		<u>MLO</u>	282
KC	S12	Gr10018194	M9ZA77_GOSRA		286

The database contains 33 KCS sequences of G. raimondii and 1 KCS sequence of G. hirsutum. Besides, none of them are reviewed and their sequences were submitted by our lab.





enome-wide analysis





Cotton database construction

F:\blast\bin 的目录

2014/01/06	16:20	<dir></dir>	
2014/01/06	16:20	<dir></dir>	
2013/03/13	06:34	5,488,640	blastdbcheck.exe
2013/03/13	06:34	6,592,512	blastdbcmd.exe
2013/03/13	06:34	4,174,336	blastdb_aliastool.exe
2013/03/13	06:34	9,608,704	blastn.exe
2013/03/13	06:34	9,604,608	blastp.exe
2014/01/06	19:12	88,207	blastpout
2013/03/13	06:34	9,597,440	blastx.exe
2013/03/13	06:34	9,463,808	blast_formatter.exe
2014/01/06	00:02	1,279	CDS.fas
2014/01/06	00:07	26,794	cdsout
2013/03/13	06:34	5,219,840	convert2blastmask.exe
2014/01/05	12:38	775,246,172	Cotton_D.all.anchored.fa
2014/01/05	19:03	337,026	Cotton_D.all.anchored.fa.nhr
2014/01/05	19:03	53,448	Cotton_D.all.anchored.fa.nin
2014/01/05	19:03	194,138,605	Cotton_D.all.anchored.fa.nsq
2013/03/13	06:34	9,780,224	deltablast.exe
2013/03/13	06:34	5,517,824	dustmasker.exe
2014/01/04	13:38	49,567,883	Gr-cds.fa.txt
2014/01/04	16:49	4,975,290	Gr-cds.fa.txt.nhr
2014/01/04	16:49	491,788	Gr-cds.fa.txt.nin
2014/01/04	16:49	11,348,776	Gr-cds.fa.txt.nsq
2014/01/04	17:07	18,193,161	Gr-PEP.fa.fasta
2014/01/04	17:08	5,016,266	Gr-PEP.fa.fasta.phr
2014/01/04	17:08	327,888	Gr-PEP.fa.fasta.pin
2014/01/04	17:08	15,083,278	Gr-PEP.fa.fasta.psq
2013/03/13	06:34	51,345	logaoy_blact.pl
2013/03/13	06:34	6,111,232	makeblastdb.exe
2013/03/13	06:34	5,744,128	makembindex.exe
2013/03/13	06:34	5,085,184	makeprofiledb.exe
2014/01/06	19:12	554	PEP.fas
2013/03/13	06:34	9,755,136	psiblast.exe
2013/03/13	06:34	9,621,504	rpsblast.exe
2013/03/13	06:34	9,613,312	rpstblastn.exe
2013/03/13	06:34	5,727,744	segmasker.exe
2013/03/13	06:34	9,716,224	tblastn.exe

1. Download local blast software from NCBI.

2.Download G. raimondii genome sequence from CGP.

3.To execute blast in DOS system



To acquire conserved domain



HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam 25.0 (March 2011, 12273 families)

The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). <u>More...</u>



Recent Pfam blog posts

ØHide this

No, seriously, we've made a release @ (posted 1 April 2011)

Wall it should have been out shout 6 menths and but finally the long qualited

pfam (<u>http://pfam.sanger.ac.uk/</u>)





Pfam search results



FAE1_CUT1-RppA HMM:

ylarrrrkvyLvdyacykpedelkvstetfleivkrvkkldeesleflrkilersGlgeetyvPrslleipee ktlaeareEaeevlfgavdellaktkvkpkdigilvvncslfsptPslsamvvnryklredvksynLsgm GCsaglisidlakdllqvhkntlalvvstEnitlnwYvGnersmllsnclFRvGgaavllsnksadrrr akykLvhvvRthkgaddkayrcvlqeeDeegkvGvslskdlvkvagealkknlttlgplvLPlsEkl rflaslvarkl

To query cotton genome database

	.1 000
000	+ 6='7000
	111-230

Sequences producing significant alignments:	Score (Bits)	e E Value
Cotton D gene 10017624 locus=scaffold163.642071	•643603•+ 497	4e-147
Cotton D gene 10018297 locus=scaffold71.210359.	711873.+ 425	2e-146
Cotton D gene 10016173 locus=scaffold172:162964	1.1631173 422	3e-145
Cotton D gene 10036653 locus=scaffold109:119017	2:1191719:+ 418	1e-143
Cotton D gene 10034062 locus=scaffold4:3957287:	3958834 - 417	2e-143
Cotton D gene 10010162 locus=scaffold130:399751	•401289•- 417	4e-143
Cotton D gene 10015926 locus=scaffold166:101637	7.1017972.+ 414	6e-142
Cotton D gene 10018147 locus=scaffold152:156765	7:1569171:+ 413	8e-142
Cotton D gene 10028536 locus=scaffold43:1655946	:1657559:+ 409	1e-139
Cotton D gene 10018150 locus=scaffold152:1584096	6:1585466:+ 405	1e-139
Cotton D gene 10031660 locus=scaffold84:3835910	:3838652:- 408	2e-139
Cotton D gene 10021139 locus=scaffold31:112479:	114389:- 408	2e-139
Cotton D gene 10018149 locus=scaffold152:157782	3:1579337:+ 404	4e-138
Cotton D gene 10014500 locus=scaffold72:1357197	:1358687:+ 402	2e-137
Cotton Desence 10031991 homesescentfold8:311495x3	1 23723	2e-136
Cotton D gena 1004 55 Ocust scart D3 Avat 10		5e-132
Cotton D gene 10000033 locus=scaffold1232:972:2	144:+ 380	3e-130
Cotton D. gene 10026783 _ locus_coaffold67:784594:'	786550:+ 376	6e-127
	376:- 318	1e-107
Cotton_D_gene_10018148 locus=scaffold152:1574866	ð:1576236:+ 320	1e-106
Cotton_D_gene_10017854 locus=scaffold167:1150904	4:1152277:- 297	2e-097
Cotton_D_gene_10040228 locus=scaffold1:7996116:	7997621:+ 291	2e-094
Cotton_D_gene_10040229 locus=scaffold1:8009851:8	3011356:+ 290	7e-094
Cotton_D_gene_10019136 locus=scaffold36:1496467	:1497780:- 284	3e-092
Cotton_D_gene_10032226 locus=scaffold8:2307409:2	2308812:+ 273	1e-087
Cotton_D_gene_10022901 locus=scaffold25:472748:4	474148:+ 270	1e-086
Cotton_D_gene_10019616 locus=scaffold207:1262504	4:1263853:- 266	2e-085
Cotton_D_gene_10032475_locus=scaffold75:3586793	<u>:3588181:- 256</u>	3e-081
	<u> 591087:- 248</u>	4e-081
Cotton_D_gene_10019659 locus=scaffold207:1692130	0:1693449:+ 254	1e-080
<u>Cotton D gene 10019657</u> locus=scaffold207:165727(0:1658544:+ 251	<u>1e-079</u>
Cotton_D_gene_10019658 locus=scaffold207:1664233	3:1664877:+ 187	8e-058
Cotton_D_gene_10038985 locus=scaffold30:686293:0	586952:- 72.4	4e-015
Cotton_D_gene_10000813 locus=scaffold471:148816	:149609:- 38.9	0.002
Cotton_D_gene_10026953 locus=scaffold260:135384	:136656:- 39.3	0.003
Cotton_D_gene_10023921 locus=scaffold97:1006873	:1009133:+ 38.5	0.004
Cotton_D_gene_10002256 locus=scaffold414:316833	:319857:- 37.4	0.009
Cotton_D_gene_10023920 locus=scaffold97:934662:3	936994:+ 37.0	0.013
Cotton_D_gene_10029217 locus=scaffold96:1906837	:1908121:- 37.0	0.014
Cotton_D_gene_10011223 locus=scaffold119:765522	:768233:+ 36.2	0.024
Cotton_D_gene_10007646 locus=scaffold236:85034:8	35.0	0.059

1. To query cotton genome

database based on HMM model

sequence

2. To retrieve the corresponding

gene sequences ,then confirm the

conserved domain in Pfam database

and smart database.

(http://smart.emblheidelberg.de/)

B. To further search the cotton

database in order to find all KCS

sequences.





Phylogeny construction and chromosome location







1. The phylogeny analysis suggests that all G. ranondii KCS genes are divided in two major branches. 2.In combination with Pfam prediction results, we know that the two major branches belong to FAEFAE1_CUT1_RppA (PF08392, PF08541) and Elo(PF01151, GNS1/SUR4 family) 3.FAE1_CUT1_RppA: described as 3ketoacyl-CoA synthases ; ACP_syn_III_C, ACP synthase III C terminal; ELO1 Members of this family are involved in LCFAs elongation systems that produce the 26-carbon precursors for ceramide and sphingolipid synthesis

Candidate gene in chromosome location



29 G.raimondii KCS genes except GrKCS11, GrKCS16, GrKCS2, Gr10004055 were located on 11 chromosomes, but chromosome 3 and 12 contain no KCS genes.



KCS may be required in fiber initiation and elongation



Several 3-ketoacyl-CoA synthase (KCS) genes, including **KCS2, KCS13 and KCS6**, were **only expressed in G. hirsutum**, whereas intermediate levels of KCS7 transcripts were observed in both G. hirsutum and G. raimondii, indicating that high-level expression of Sus and KCS family genes may indeed be required for fiber cell initiation and elongation.(Kunbo Wang, et al. *2012.Nature Genetics.)*



Multiple sequence alignment

	380	390	400 4	10 420
Gr10040229	KEGEIYMPSF	TAIQHECLPTSG	RALIGEIAKGLN	DGRDVEASLMTHRE
Gr10040228	KEGEIYMPSF	TAIQHECLPTSG	RALIGEIAKGLN	LDGRDVEASLMTLHRF
Gr10017854	ESAEIYTPRF	TVVQHECLPSSGI	KPLIREVAKGLN	LNGRNIEPALMTLHRF
Gr10015926	AKVSPYIPDF	LAFDHFCIHAGGE	RAVLDELOKNLO	LTDWHMEPSRMTLYRF
Gr10028536	AKVKPYIPDF	LAFERFCIRAGG	RAVLDELOKNLO	LTDWHMEPSRMTLHRE
Gr10004055	ARVKPYIPDF	LAVENECIHAGGE	RAVLDEIQKNLE	LTDWHMEPSRMTLHRF
Gr10016173	MKVKPYIPDF	LAFERFCIHAGGE	RAVLDELEKNLO	LSEWHMEPSRMTLYRF
Gr10017624	MKIKPYIPDF	LAFERECIHAGGE	RAVLDELEKNLO	LSEWHNEPSRMTLFRF
Gr10010162	MKIKPYIPDF	LAFERFCIHAGGE	RAVLDELEKNLO	LSEWHMEPSRMTLYRF
Gr10031660	MKIKPYIPDF	LAFERECIHAGGE	RAVLDELEKNLD	LTDWHMEPSRMTLYRF
Gr10021139	MKIRPYIPDF	LAFERFCIHAGGI	RAVLDELEKNLE	LSDWHMEPSRMTLYRF
Gr10031991	PKWKPYIPDF	QAFEH FC IHAGGI	RAVIDELOKNLO	LSAEHVEASRMTLHRF
Gr10014500	PKWKPYIPDF	LAFEHECIHAGGI	RAVIDELOKNLO	LSAEHVEASRMTLHRF
Gr10036653	ASVKPYIPDF	LAFDHECIHAGGI	RAVIDELEKNLO	LLPIHVEASRMTLHRF
Gr10034062	AGIKPYIPDF	LAFDHFCIHAGGI	RAVIDELEKNLO	LIPIHVEASRMTLHRF
Gr10000033	AGIKPYIPDF	LAFDHFCIHAGGI	RAVIDELEKNLO	LLPIHVEASRMTLHRF
Gr10018297	AKIKPYIPDF	LAFDHECIHAGGI	RAVIDELEKNLO	LIPVHAEASRMTLHRF
Gr10018148	AKIKPYIPDF	LAFERFCIHAGGI	RAVIDELEKNLH	LIPVHVEASRMTLHRF
Gr10018150	AKIKPYIPDF	LAFEHECIHAGGI	RAVIGELEKNLY	LLOVHVEASEMTLHEF
Gr10018147	AKIKPYVPDF	LAFEHECIHAGGI	RGVIDELEKNNO	LSPLHVEASRMTLHRF
Gr10018149	AKIKPYVPDL	LAFERFCIRAGG	RGVIDELEKNNO	LSPLHVEASRMTLHRF
Gr10026783	KSKTSLSPSSKPYIPDY	LAFERFCVHAASI	KTVLDELQKNLE	LSENNMEASRMTLHRF
Gr10019616	KTSSPES.GLNL	SGIDYFCIHPGGI	RAVIDAMGRSLG	LNEYDLEPTRMALHRF
Gr10032475	NNNKGKTPSLNM	SGFQHFCIHPGGI	RAVIDANGRSLG	LNEYDLEPTRMALHRF
Gr10019657	KGQNLSF.NLNL	SGVDHECLHPGGI	RAVIDGLGKSLG	LSEYDLEPTRMALYRF
Gr10019659	KGQNLSF.NLNL	SGVDHFCLHPGGI	RAVIDGLGKSLG	LSEYDLEPTRMALYRF
Gr10019136	KTAKPSLNL	TGIQHECIHPGGI	RAVIDGLGKSLG	LNEYDNEPARMALHRF
Gr10032226	HGSSHKGASQGPIKAGVNF	SGVDHECIHTGGI	KAVIDGIGISLD	LTEYDLEPARMTLHRF
G=10022901	UCCTVCTTTOCDIVACUNT	CIGATIN DISCUT DISCOU	VAUT DOT OPOT D	T TE VINITE DIA DIVITTURE

	430	440	450	460	470	480
Gr10040229	GNOSSSSMWYEL	AMMEARER	VKKGDKVLMLGM	GTOPKOGSC	VWECVRPIAG	DSNKNNPMR
Gr10040228	GNOSSSSMWYEL	AYNBAKER	VKKGDKVLMLGM	GTGPKCCSC	VWERVRPIAG	DSNKNNPMR
Gr10017854	GNOSSSSWWYEL	GYNEGKWE	VKKGDKIWVLGL	GTGIKCCSL	VLECLRPIVE	D.DKKSPMS
Gr10015926	GNTSSSSLWYEL	ANTEAKCE	VSDGDRVWQIAF	GSGFKCNSA	VWRALRSTPM	AESRGNPHK
Gr10028536	GNTSSSSLWYEL	ANTEAKGE	VSSGDRVWQIAF	GSGFKCNSA	VWRALRSPN	NELRGNPKK
Gr10004055	GNTSSSSLWYEL	AYTEAKCE	ISGGDRVWQIAF	GSGFKCNSA	VWRALRSTPM	SESRCNPKK
Gr10016173	GNTSSSSLWYEL	AYSBAKGE	IRRGDRTWOIAF	GSGFKCNSA	VWKALRTINF	VKEK,NPM
Gr10017624	GNTSSSSLWYEL	AYSEAKGE	IRKGDRTWOIAF	GSGFKCNSA	VWKALRTVNF	AKEK, NPMM
Gr10010162	GNTSSSSLWYEL	AYSEAKGE	IKKGDRTWQIAF	GSGFKCNSA	VWKALGTINF	AKEK, NPMM
Gr10031660	GNTSSSSLWYEL	AYSEAKGE	IRRGDRTWOIAF	GSGFKCNSA	VWKALKTINF	AKEK.SPMI
Gr10021139	GNTSSSSLWYEL	AYSEAKGE	IRKGDRTWQIAF	GSGFKCNSA	VWKALKTINF	AKEK.SPMI
Gr10031991	GNTSSSSLWYEN	SYIEAKGE	MKKGDRVWQIAF	GSGFKCNSA	VWKCNRTIKT	PKDGPMA
Gr10014500	GNTSSSSLWYEN	SYLEAKGE	MKKGDRIWQIAF	GSOFKCNSA	VWKCNRTIKK	PTDGPME
Gr10036653	GNTSSSSIWYEL	AYIEAKGE	IRNRNRIWQIAF	GSCFKCNSA	VWQALRNVKP	SSNGPME
Gr10034062	GNTSSSSIWYEL	AVIDANCE	MRKRNRVWQIAF	GSOFKCNSA	VWEALRNVKS	SSNGPME
Gr1000033	GNTSSSSIWYEL	AYIEAKGE	MRKRNRVWQIAF	GSGFKCNSA	VWEALRNVKS	SSNGPME
Gr10018297	GNTSSSSIWYEL	AYTEAKGE	MRKGNRVWQIAF	GSCFKCNST	VWVALRNVKP	SPNNP
Gr10018148	GNTSSSSIWYEL	AYMPAKGE	MRKGDRVWQIAF	SOFKENSA	VWLALKNVKP	SYNSPME
Gr10018150	CNTSSSSIWYEL	ANIDARGE	MRKGDRIWQIAF	SOFKENSA	VWLALKNVKP	SCNSP
Gr10018147	GNTSSSSIWYEL	AYMEAKGE	MRKGDRIWQVAF	SOFKENSA	VWLALKNVKP	SCGNPME
Gr10018149	GNTSSSSIWYEL	AYMDAKSE	MRKGDRIWQVAF	GSOFKCNSA	VWLALKNVKP	SCGNP
Gr10026783	GNTSSSSIWYEL	ANLDAKER	VERGORIWOIAF	SCFKENSV	VWRSMRRVRR	PSRD.NP
Gr10019616	GNTSAAGLWIVL	SENDARKS	LKKGDRILMISL	AGFRONNC	VWEVMEN, LL	DVNV
Gr10032475	GNTSAAGIWTVL	SIMPARKE	LKKGDRILMISL	AFFKONNC	VWEVNKDAMD	DVNV
Gr10019657	GNTSAGGLWIVL	SINDAKKE	LKKGDKIFMVSL	OACE MONNC	VWEVMKDGLD	DTRV
Gr10019659	GRTSAGGLWYVL	SEMPARKE	LKKGDKIFMVSL	DAGEMENNC	VWEVNKDGLE	DTRV
Gr10019136	GNTSAGGLWIVL	GENERAKKE	LEKGNEILMISL	GAGE MONNC	VWEVNED.LG	NGNVE
GF10032226	GRISASSLWIVL	AFROAKKE	LKKGDKVLMISF	AGFKONSC	LWEVVRD.LG	DGNVK
GF10022901	UNTRASSINTVL	ATNEAKK	LKKGDKVLMISF	DAUFKONSC	LWEIVRD.LE	DGNV

1. Align all KCS genes using ClustalX.

2.Resulting from poor alignment quality, we excluded all short sequences including Elo type protein and alignment again using ClustalX.

3.The alignment results suggest that FAE type KCS protein share some strict conserved amiono acids.

4.Based on Netural Evolution of Kimura, most amino acids mutants are neutral, few mutants are harmful

and these sites are functional.





In order to find KCS proteins' active sites and study functional mechanism, we used popular methods to predict our target proteins' structures.





Obtain Squence

We extracted each one sequence from FAE type and Elo type , such as Gr10018148 and Gr10000033.

Prediction

We used Phyre and I-TASSER methods to predict KCS protein 3D sturcuture.



Analysis

To analysis some details about protein structure



Analysis of Gr10018148 Protein structure



Gr10018148-phyre



The secondary structure of Gr10018148 protein



Acc: accessibility, white is buried; cyan is intermediate; blue is accessible; blue with red borders is highly exposed. Hyd: hydropath, pink is hydrophobic; grey is intermediate; cyan is hydrophilic

Conserved domain in all KCS proteins

		α4	β3	η2	α5	
Gr10018148	160	170	180	➡ 200000 190	200	210
C=10019149	ENDER		ZMDCNIKCEN			
Gr10010140	FCPSPS	LSSMIINKI	SMRGNIKSEN			NNNCVATVISTET
Gr10040229	FCPSPS	LSSTTVNKVS	SMRSDVKSFN	LSGMGCSSGA	IGIDLAONLLKT	NNNCVATVISTET
Gr10017854	FCPSPS	LSSTTTNK	SMKSDIKSVN	LSCMCCSACT	CVDLAONLLKT	HENKTATVISTET
Gr10015926	FNPTPS	LSAMTVNHY	KLRTNINSYN	LGGMGCSAGL	ISTDLAKNLLOS	NPNTYALVVSTEN
Gr10028536	FNPTPS	LSAMTVKHY	KLRTDIKSYN	LGGMGCSAGL	ISVELAKNLLŐA	NPNTYAVVVSTEN
Gr10004055	FNPTPS	LSAMVINHY	KLRTDIKSYN	LGGMGCSAGL	ISVELAKNLLRA	NPNTYAVVVSTEN
Gr10016173	FNPTPS	LSAMVINHY	KLRGNIOSYN	LGGMGCSAGL	ISIDLAKHLLOV	HPNSYALVISMEN
Gr10017624	FNPTPS	LSAMVINHY	KLRGNIÕSYN	LGGMGCSAGL	LSIDLAKNLLŐV	HPNSYALVISMEN
Gr10010162	FNPTPS	LSAMVINHY	KLRGNIÕSYN	LGGMGCSAGL	ISIDLAKNLLÕV	H P N T Y A L V I S M E N
Gr10031660	FNPTPS	LSAVVVNR <mark>Y</mark> F	KFRGNILSYN	LGGMGCSAGL	ISIDLAKQLLRV	HPNSYALVVSMEN
Gr10021139	FNPTPS	LSAMIVNR <mark>Y</mark> F	KLRGNILSYN	LG <mark>GMGCS</mark> AGL:	ISIDLAKQMLQV	HP <mark>N</mark> SY <mark>A</mark> LVVSMEN
Gr10031991	FSPTPS	LSAMVINK <mark>Y</mark> F	KLR <mark>SN</mark> IKSFN	LS <mark>GMGCS</mark> AGL:	ISIDLARDLLQV	HP <mark>N</mark> SN <mark>A</mark> VVVSTEI
Gr10014500	FSPT <mark>P</mark> S	LSAMVINK <mark>Y</mark> F	KLR <mark>SN</mark> IKSFN	LS <mark>GMGCS</mark> AGL	ISIDLARDLLQV	HP <mark>N</mark> SN <mark>A</mark> VVVSTEI
Gr10036653	FNPT <mark>P</mark> S	LSAMIVNK <mark>Y</mark> F	KLRGNIRSFN	LG <mark>GMGCS</mark> AGV	IAVDLAKDLLQV	HR <mark>N</mark> TY <mark>A</mark> VVVSTEN
Gr10034062	FNPT <mark>P</mark> S	LSAMIVNK <mark>Y</mark> F	KLRGNIRSFN	L G <mark>G M G C S</mark> A G V I	IAVDLAKDMLQV	HR <mark>N</mark> TY <mark>A</mark> VVVSTEN
Gr1000033	FNPTPS	LSAMIVNK <mark>Y</mark> F	KLRGNIRSFN	LG <mark>GMGCS</mark> AGV:	IAVDLAKDMLQV	HR <mark>N</mark> TY <mark>A</mark> VVVSTEN
Gr10018297	FNPTPS	LSAMIINK <mark>Y</mark> F	KLRGNIRSFN	L G <mark>G M G C S</mark> A G V :	IAIDLAKDMLQV	HR <mark>N</mark> SY <mark>A</mark> VVVSTEN
Gr10018150	FNPT <mark>P</mark> S	LS <mark>S</mark> MIINK <mark>Y</mark> F	KMRGNIKSFN	LGGMGCSASV	ISVDLAKDMLQV	HR <mark>N</mark> NY <mark>A</mark> IVVSTEN
Gr10018147	FNPTPS	LTAMIINK <mark>Y</mark> F	KMRGNIKSFN	L S G M G C S A G V I	IAIDLAKDMLQV	HR <mark>N</mark> NY <mark>A</mark> VVFSTEN
Gr10018149	FNPTPS	LTAMIINK <mark>Y</mark> F	K M R G N N K S F N	LSGMGCSASV	IAIDLAKDMLQV	YRNNYAVVFSTEN
Gr10026783	FNPTPS	LSAMIINHYF	KMRGNILSYN	LGGMGCSAGI	IAVDLARDMLQA	NPNNYAVVVSSEM
Gr10019616	ISSVPS	IPARVINRYF	KMREDVKVFN	LSGMGCSASL	IAVDLVNHLFQT	YKNQFAIVVSSES
Gr10032475	ITSVPS	LPARVINRY	KMRDDVKVFN	LSGMGCSASV	LAVDLVHHLFKT	YMNSFAVIVSSES
Gr10019657	FSPAPS	LTSRIINRYF	KMRHNIKSFS	LSGMGCSASM:	LAIDLVQQLFKT	YKNQFAIVVSTES
Gr10019659	FSPAPS	LTSRIINRY	KMRDNIKSFS	LSGMGCSASM	ALDMVQQLFKT	YKNOFAIVVSTES
Gr10019136	FSPSPS	LTARIVNRY	KMRDNIKSFS	LSGMGCSASM	ALDLVQNLFKS	YKNAFAVVVSSET
Gr10032226	LSTVPS	LCSRIINHY	KMRPDIKSFN	LTGMGCSASL.	LSLDIVKNVFKS	YKNKFALLVTSES
GETUUZZ901	TTAPEC	T 2 2 K T T N H	MRQDIKCPN	CMCCSASL.		INNAIGLEVISES
consensus>/0	r.p.ps	15!!n.YK	cmr.#1.5%n.	L.GMGCSa	L.I#IA.GII	

Why these domains are conserved?



3D structure prediction suggest thatStrict domains locate on two helix ,five sheets and some loops.



Summary

- ★ Local cotton database construction is helpful to query target sequence and conserved domain prediction contributes to research loss of function in protein family.
- ★ Phylogeny construction and chromosome location suggest that there are some KCS genes that are required for fiber initiation and elongation.
- ★ Protein structure prediction can provide useful information to design point mutants.
- \star Bioinformatics analysis is useful, but we should be careful.





Prof. Luo Prof. Zhu All members in ABC class, especially G01 group. Our lab members