

BLAST結果の処理法（2）可視化

Shuji Shigenobu / 重信秀治

Aim

- BLASTの結果を可視化し、知識発見に結びつける。

```
BLASTX 2.5.0+
out.b0.txt -- Edited ~

Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.

Database: mouse_proteins.pep.fasta
49,870 sequences; 21,467,191 total letters

Query= XM_002716705.3 PREDICTED: Oryctolagus cuniculus autophagy related 3
(ATG3), mRNA
Length=1516

Sequences producing significant alignments:
Score E
(bits) Value
Q9CPX6 Ubiquitin-like-conjugating enzyme ATG3 OS=Mus musculus GN=... 598 0.0
Q8R1P4 Ubiquitin-like-conjugating enzyme ATG10 OS=Mus musculus G... 35.0 0.13
A0A0R4J029 Autophagy-related 10 (Yeast), isoform CRA_a OS=Mus mu... 34.7 0.14
A0A0X1K6G2 Negative elongation factor B OS=Mus musculus GN=Nelfb... 30.4 5.0
Q8VF22 Coiled-coil domain-containing protein 138 OS=Mus musculus... 30.0 6.9
Q8C4Y3 Negative elongation factor B OS=Mus musculus GN=Nelfb PE... 30.0 7.1
Q9JK38 Glucosamine 6-phosphate N-acetyltransferase OS=Mus muscul... 29.3 8.3
Q8CSK5 Uncharacterized protein CXorf38 homolog OS=Mus musculus P... 29.6 8.8

>Q9CPX6 Ubiquitin-like-conjugating enzyme ATG3 OS=Mus musculus GN=Atg3
PE=1 SV=1
Length=314
Score = 598 bits (1542), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 309/314 (98%), Positives = 313/314 (99%), Gaps = 0/314 (0%)
Frame = +1
Query 418 MNQVINTVKGKALEVAEYLTPVLKESKFRETGVITPEEFVAAGDHLVHHCTWQWATGEE 597
Sbjct 1 MNQVINTVKGKALEVAEYLTPVLKESKFRETGVITPEEFVAAGDHLVHHCTWQWATGEE 600
Query 598 LKKVAYLPTGKQFLVTKNWPCYKRCOMEYSDELEAITFEEDGGGGWDTYHNTGITGIT 777
Sbjct 61 LKKVAYLPTGKQFLVTKNWPCYKRCOMEYSDELEAITFEEDGGGGWDTYHNTGITGIT 120
Query 778 EAVKEITLKNKSIKLQDCSALCEEEDEGEAADMEEYEESSGLLETDEATLDRKIVE 957
Sbjct 121 EAVKEITLKNKSIKLQDCSALCEEEDEGEAADMEEYEESSGLLETDEATLDRKIVE 180
Query 958 ACKAKADAGGEDAILQTRTYDLYITYDKYQTPRLWLFYGYDEORQPLTVEHMYEDISODH 1137
Sbjct 181 ACKAKADAGGEDAILQTRTYDLYITYDKYQTPRLWLFYGYDEORQPLTVEHMYEDISODH 240
Query 1138 VKKTVTIENHHPHLP PPPMCSVHPCRHAENVKKIETVAEGGGELGVHMYLLIFLKFVQAV 1317
Sbjct 241 VKKTVTIENHHPHLP PPPMCSVHPCRHAENVKKIETVAEGGGELGVHMYLLIFLKFVQAV 300
Query 1318 IPTIEYDYTRHFTM 1359
Sbjct 301 IPTIEYDYTRHFTM 314

>Q8R1P4 Ubiquitin-like-conjugating enzyme ATG10 OS=Mus musculus GN=Atg10
PE=1 SV=1
Length=215
Score = 35.0 bits (79), Expect = 0.13, Method: Compositional matrix adjust.
Identities = 25/116 (22%), Positives = 52/116 (45%), Gaps = 11/116 (9%)
Frame = +1
Query 994 AILQTRTYDLYITYDKYQTPRLWLFYGYDEORQPLTVEHMYEDISODHKK-----TV 1152
Sbjct 88 AVAEVTKHEYHVLVYCSYQVPLVYFRASFLDGRPLALFDINFGVHECYKPRLLLOGPWDTI 147
```

format 0

header

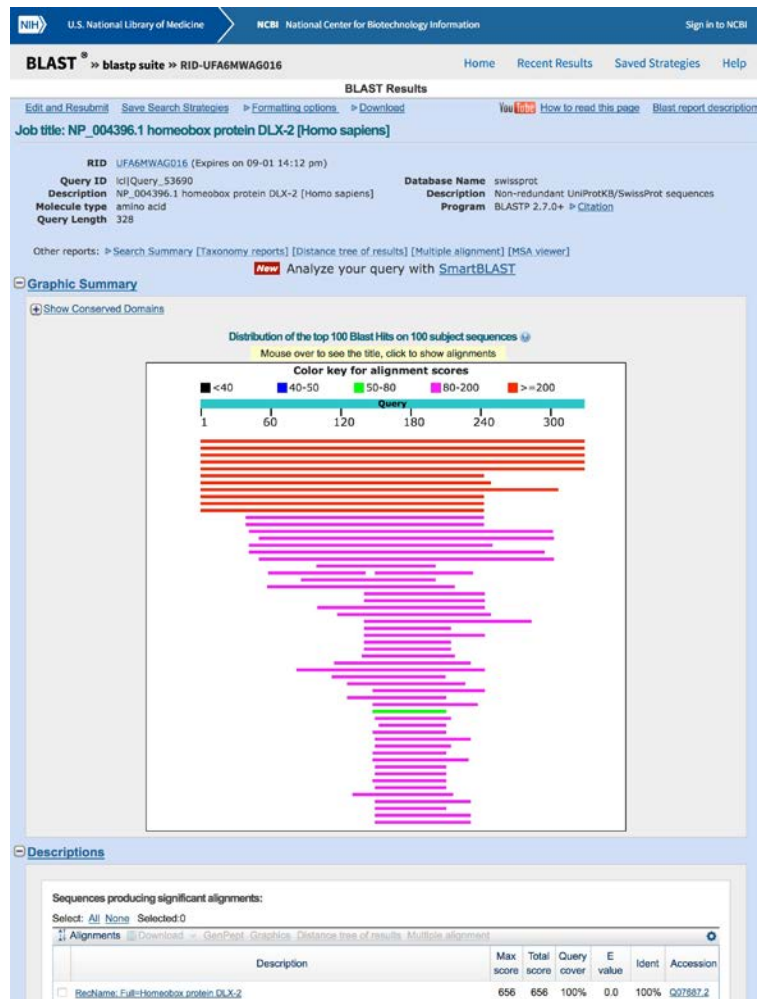
one-line summaries

alignments

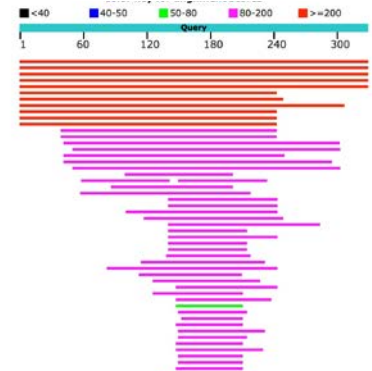
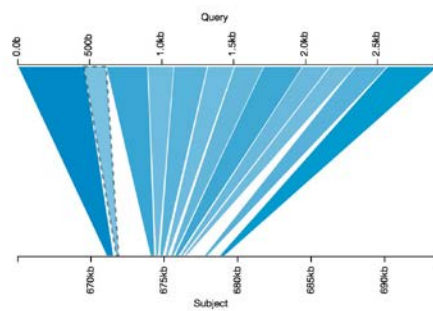
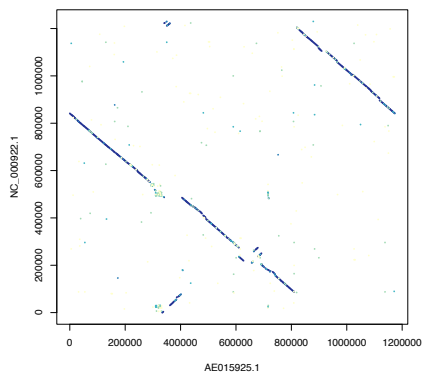
format 6

query	subject	%identity	align-len	mismatch	gap_open	q_start	q_end	s_start	s_end	bit-score	evalue
spo:SPAC212.11	sce:YMR190C	27.297	370	238	12	1169	1525	653	1004	7.08e-23	105
spo:SPAC212.11	sce:YDR021W	38.053	113	64	3	1417	1529	267	373	1.42e-14	75.9
spo:SPAC212.11	sce:YNL112W	25.076	331	225	12	1204	1525	148	464	1.06e-12	70.9
spo:SPAC212.11	sce:YBR237W	24.403	377	237	17	1190	1532	276	638	1.68e-12	70.5
spo:SPAC212.11	sce:YOR204W	28.505	214	128	8	1352	1559	339	533	2.80e-12	69.7
spo:SPAC212.11	sce:YOR046C	28.205	234	140	8	1310	1530	227	445	1.65e-11	66.6
spo:SPAC212.11	sce:YDR243C	23.512	336	227	10	1208	1524	216	540	7.40e-11	65.1
spo:SPAC212.11	sce:YGL078C	22.689	357	239	11	1188	1527	130	466	7.62e-11	64.7
spo:SPAC212.11	sce:YPL119C	27.619	210	129	8	1352	1556	351	542	2.84e-10	63.2
spo:SPAC212.11	sce:YGL064C	22.195	410	216	16	1215	1541	166	555	8.25e-10	61.6
spo:SPAC212.11	sce:YDL084W	24.424	217	156	5	1308	1524	201	409	7.84e-07	51.6
spo:SPAC212.11	sce:YHR169W	27.273	143	93	4	1418	1555	257	393	1.09e-06	51.2
spo:SPAC212.11	sce:YLR276C	26.667	210	120	10	1189	1377	38	234	2.85e-06	50.1
spo:SPAC212.11	sce:YLR276C	34.146	41	27	0	1484	1524	386	426	0.12	35.0
spo:SPAC212.11	sce:YLL008W	21.965	346	233	13	1195	1524	258	582	4.71e-06	49.3
spo:SPAC212.11	sce:YPL082C	24.074	108	81	1	1417	1524	1648	1754	0.002	40.8
spo:SPAC212.11	sce:YGL070C	37.838	37	23	0	1583	1619	52	88	2.3	29.3
spo:SPBPB10D8.05C	sce:YPL092W	30.730	397	224	9	10	356	11	406	3.04e-47	165
spo:SPBPB10D8.05C	sce:YGL195W	31.034	58	37	2	140	195	2014	2070	0.37	30.8
spo:SPBPB10D8.05C	sce:YDR283C	51.613	31	14	1	221	250	589	619	0.66	30.0
spo:SPBPB10D8.05C	sce:YBR028C	35.135	37	24	0	229	265	127	163	0.84	29.6
spo:SPBPB10D8.05C	sce:YPL027W	27.957	93	54	3	138	221	53	141	3.1	27.3
spo:SPBPB10D8.05C	sce:YDR161W	28.571	49	35	0	318	366	102	150	3.2	27.7
spo:SPBPB10D8.05C	sce:YER166W	27.451	51	35	1	50	98	1273	1323	7.0	26.9
spo:SPBPB10D8.05C	sce:YGR040W	23.256	86	55	3	214	299	3	77	7.8	26.6
spo:SPBPB10D8.05C	sce:YAR019C	44.444	18	10	0	235	252	30	47	9.5	26.2
spo:SPBC359.02	sce:YOR368W	29.885	87	56	2	53	138	7	89	0.88	29.3
spo:SPCC330.07C	sce:YHR197W	26.761	71	39	3	65	130	287	349	2.0	28.9
spo:SPCC330.07C	sce:YGR224W	23.881	67	50	1	181	247	152	217	5.1	27.7

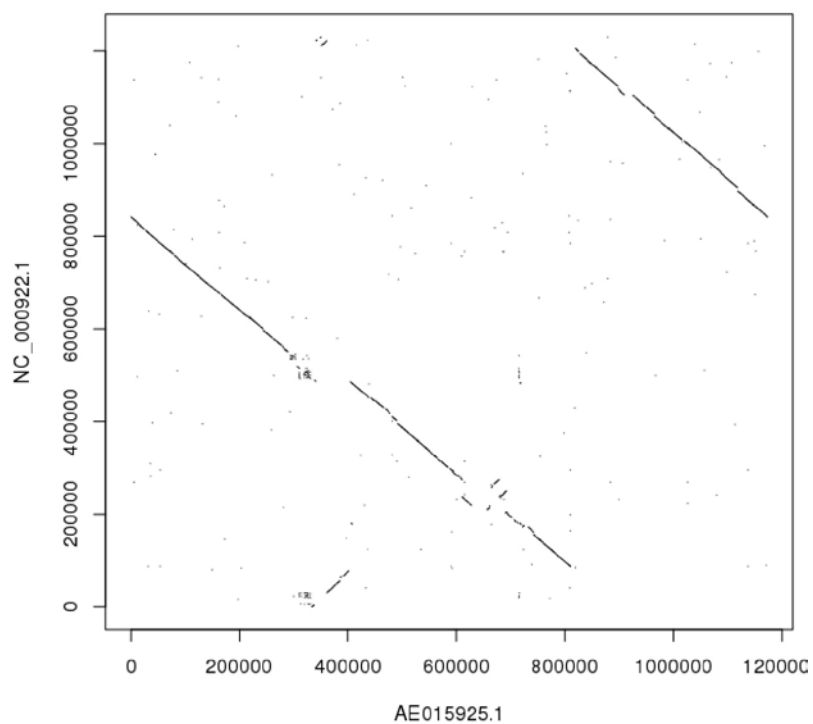
Web BLAST at NCBI



Visualize BLAST outputs



Draw dot plot



Draw dotplot from BLAST output by yourself

query	subject	%identity				q_start	q_end	s_start	s_end	evalue	bit-score
AE015925.1	NC_000922.1	79.515	14972	2951	49	793170	808087	104228	89319	0.0	13057
AE015925.1	NC_000922.1	77.062	14439	3149	68	107763	122130	730578	716232	0.0	10947
AE015925.1	NC_000922.1	70.626	23187	6113	286	874380	897222	1146890	1124058	0.0	10448
AE015925.1	NC_000922.1	76.018	12205	2796	72	1057699	1069843	968388	956255	0.0	8608
AE015925.1	NC_000922.1	94.369	5203	244	27	1020573	1025748	1005510	1000330	0.0	7984
AE015925.1	NC_000922.1	68.306	17729	5158	216	965915	983437	1057574	1040101	0.0	6082
AE015925.1	NC_000922.1	69.808	13550	3748	180	1029907	1043269	996470	983077	0.0	5490
...											



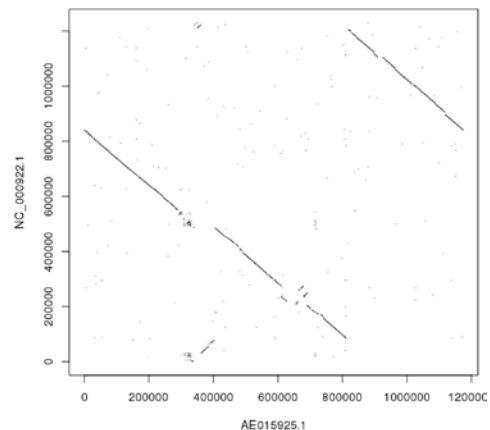
a short script

Table for R

AE015925.1	NC_000922.1
793170	104228
808087	89319
NA	NA
107763	730578
122130	716232
NA	NA
874380	1146890
897222	1124058
NA	NA
...	



```
[R]
> dat <- read.delim("table.dat")
> plot(x, type="line")
```



ex5-1

Let's try!

肺炎クラミジア *Chlamydomonas pneumoniae* のCWL029株と、GPIC株のゲノム全長をBLASTNで比較し、dotplotを描画することによって大規模な構造多型を視覚化しなさい。

- ▶ Seq1: CpneCWL029.NC_000922.uc.genome.fa
- ▶ Seq2: CpneGPIC.AE015925.uc.genome.fa

```
blastn -task blastn -subject CpneCWL029.NC_000922.uc.genome.fa \
  -query CpneGPIC.AE015925.uc.genome.fa -outfmt 6 \
  > CpneCWL029.vs.CpneGPIC.blast.out.fmt6

ruby blast6_to_rplot1.rb CpneCWL029.vs.CpneGPIC.blast.out.fmt6 \
  > CpneCWL029.vs.CpneGPIC.blast.out.fmt6.mat
```

```
(R)
> dat <- read.delim("CpneCWL029.vs.CpneGPIC.blast.out.fmt6.mat")
> plot(dat, type="l", lwd=2)
```

(参考: ex 5-2)

ex 5-1のdot plot描画を発展させ、blast bitscore によって色を変えてプロットするRスクリプトを自動生成する ruby scriptを書いてみました。

▶ Script: blast6_to_rplot2.rb

```
blastn -task blastn -subject CpneCWL029.NC_000922.uc.genome.fa \  
-query CpneGPIC.AE015925.uc.genome.fa -outfmt 6 \  
> CpneCWL029.vs.CpneGPIC.blast.out.fmt6
```

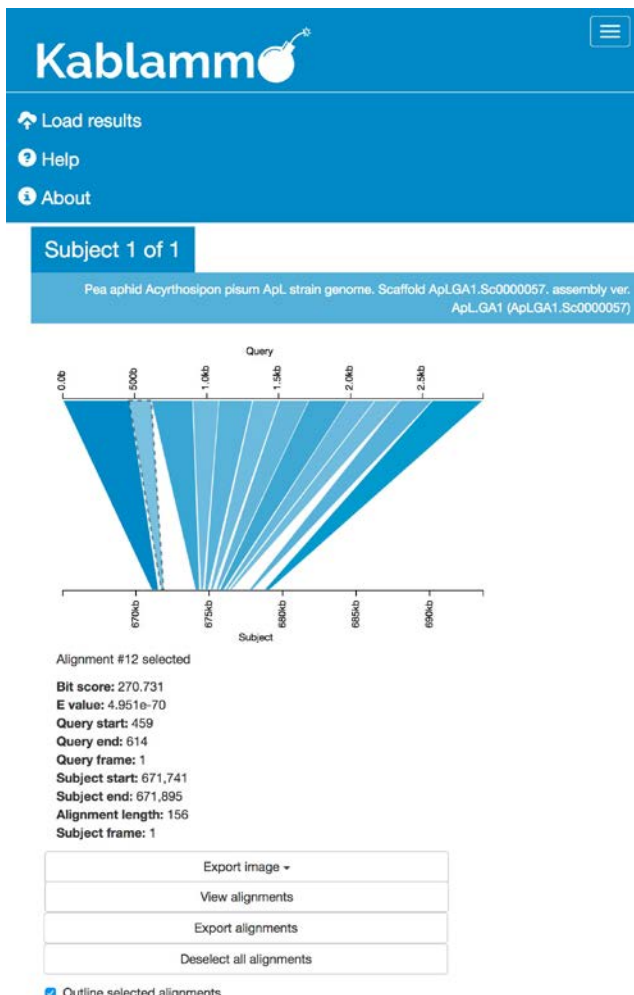
```
ruby blast6_to_rplot2.rb CpneCWL029.vs.CpneGPIC.blast.out.fmt6 \  
> CpneCWL029.vs.CpneGPIC.blast.out.fmt6.plot.R
```

(R)

```
> source("CpneCWL029.vs.CpneGPIC.blast.out.fmt6.plot.R")
```

Practice

▶ ex5-3: make a dotplot using the data set of ex1-8



<http://kablamm.wasmuthlab.org/>

- ▶ Kablammo is a web-based BLAST visualizer
- ▶ Easy setup. No need to be installed by users.
- ▶ Just upload BLAST results in XML format

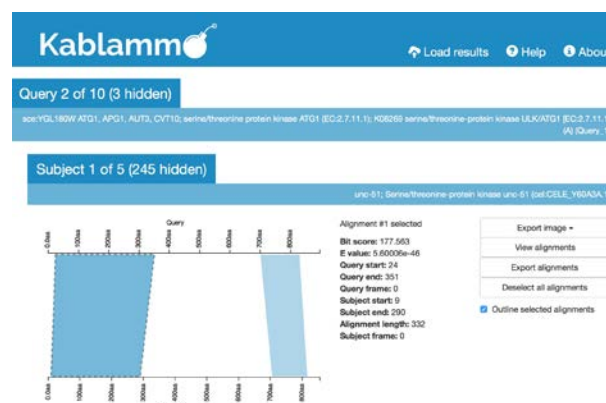
ex5-4

Practice: ex5-4

ex1-2で解析した、BLASTP の結果をKablammo で可視化しよう。
XML format で出力し、それをKablammoにアップロード

URL: <http://kablamm.wasmuthlab.org/>

```
blastp -query ScATGgenes.aa.fasta -db Cele.T00019.pep -outfmt 5 \
> out.xml
```



SequenceServer

<http://www.sequenceserver.com/>

Query= SI2.2.0.02551

BLAST 3 queries, 1 database

Query= SI2.2.0.02551

Query= SI2.2.0.02551

Query= SI2.2.0.02551

Download FASTA, XML, TSV

FASTA of all hits

FASTA of 1 selected hits

Number	Sequences producing significant alignments	Total score	E value	Length
1.	spIQ5SPL2 PHF12_MOUSE	372.06	1.07 x 10 ⁻¹⁰⁰	1305
2.	spIQ6QT6 PHF12_HUMAN	372.06	1.17 x 10 ⁻¹⁰⁰	1304
3.	spIQ6Q86 A27_SCHPO	105.53	2.15 x 10 ⁻²¹	907
4.	spIQ4779 YDC1_YEAST	82.03	6.69 x 10 ⁻¹⁶	984
5.	spIQ15154 TF1A_HUMAN	77.03	2.71 x 10 ⁻¹³	1050
6.	spIQ14839 CHD4_HUMAN	75.87	7.90 x 10 ⁻¹³	1312
7.	spIQFQ23 CHD4_MOUSE	75.87	8.69 x 10 ⁻¹³	1315
8.	spIQ4127 TF1A_MOUSE	75.10	1.29 x 10 ⁻¹²	1051
9.	spIQ24W6 TF1B_MOUSE	74.33	1.94 x 10 ⁻¹²	1242
10.	spIQ24W6 TF1B_MOUSE	73.94	2.75 x 10 ⁻¹²	1346

spIQ5SPL2|PHF12_MOUSE

spIQ6QT6|PHF12_HUMAN

HT length: 1004

1. Score	E value	Identities	Gaps	Positives
163.31 (410)	1.17 x 10 ⁻⁴⁰	84/102 (83.7%)	9/102 (4.6%)	120/102 (92.5%)

Query 142 EKRRSALVLTAAALVSPREFLPPELGLPINFPGSEK--ADYVGRVCKE-SGLD 197

Subject 203 QQLRPPPELIIAANKRPPTQLPPECCTALPDSERARKETTORVYRTQSLD 242

Query 198 SGLVPLPALKQYRGRGRKALACRQCYFFQDCLRPPTPTIQRVRRPPEKRP 207

Subject 207 N-VVPSL++CF C RSCR APLE CDYCL FR SCL+PPIDA P+GRVRRPSE E

▶ SequenceServer is a web-based interface of BLAST+.

▶ A visualization function like Web version of NCBI BLAST was implemented recently.

▶ Easy to set up.

ex5-5

Let's try

Sequenceserver をインストールし、線虫 *C. elegans* のタンパク質BLAST database を構築する。
C. elegans タンパク質のアミノ酸配列は、ex1-2 で使った, Cele.T00019.pep

1. Install sequenceserver

#今回使うMacにはインストール済みなので以下のコマンドは実行する必要はない
\$sudo gem install sequenceserver

2. Set up blastdb

makeblastdb -in Cele.T00019.pep -dbtype prot -parse_seqids

3. Start server

\$sequencesever

4. Access sequencesever with web browser

(access <http://localhost:4567>)

5. BLAST search and automatic visualization