河豚鱼多药耐药基因柯氏质粒序列片段分析

Analysis of Fugu Cosmid sequence fragment with two Multi-Drug Resistance Genes

徐礼鸣

Xu, Liming

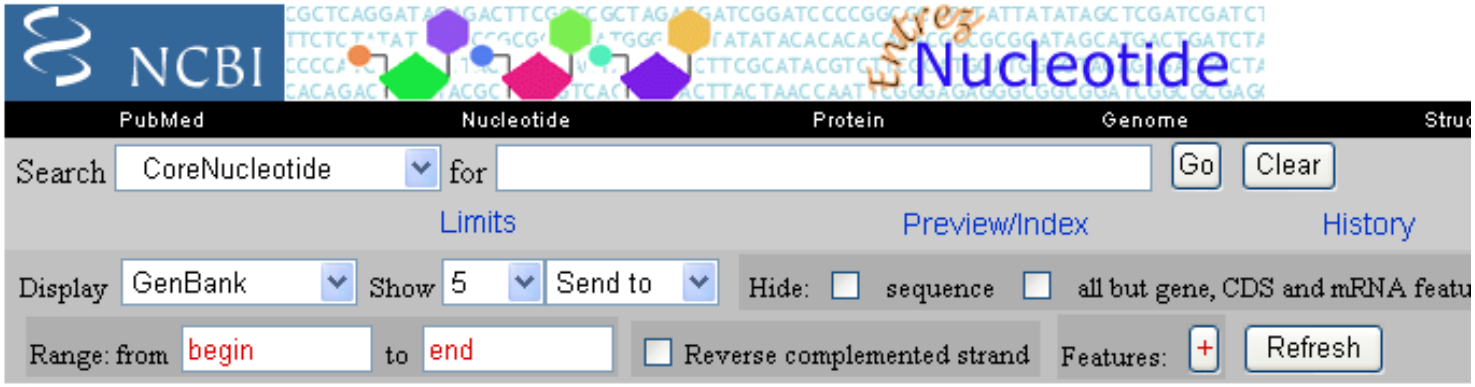
Mar 17, 2009

简介

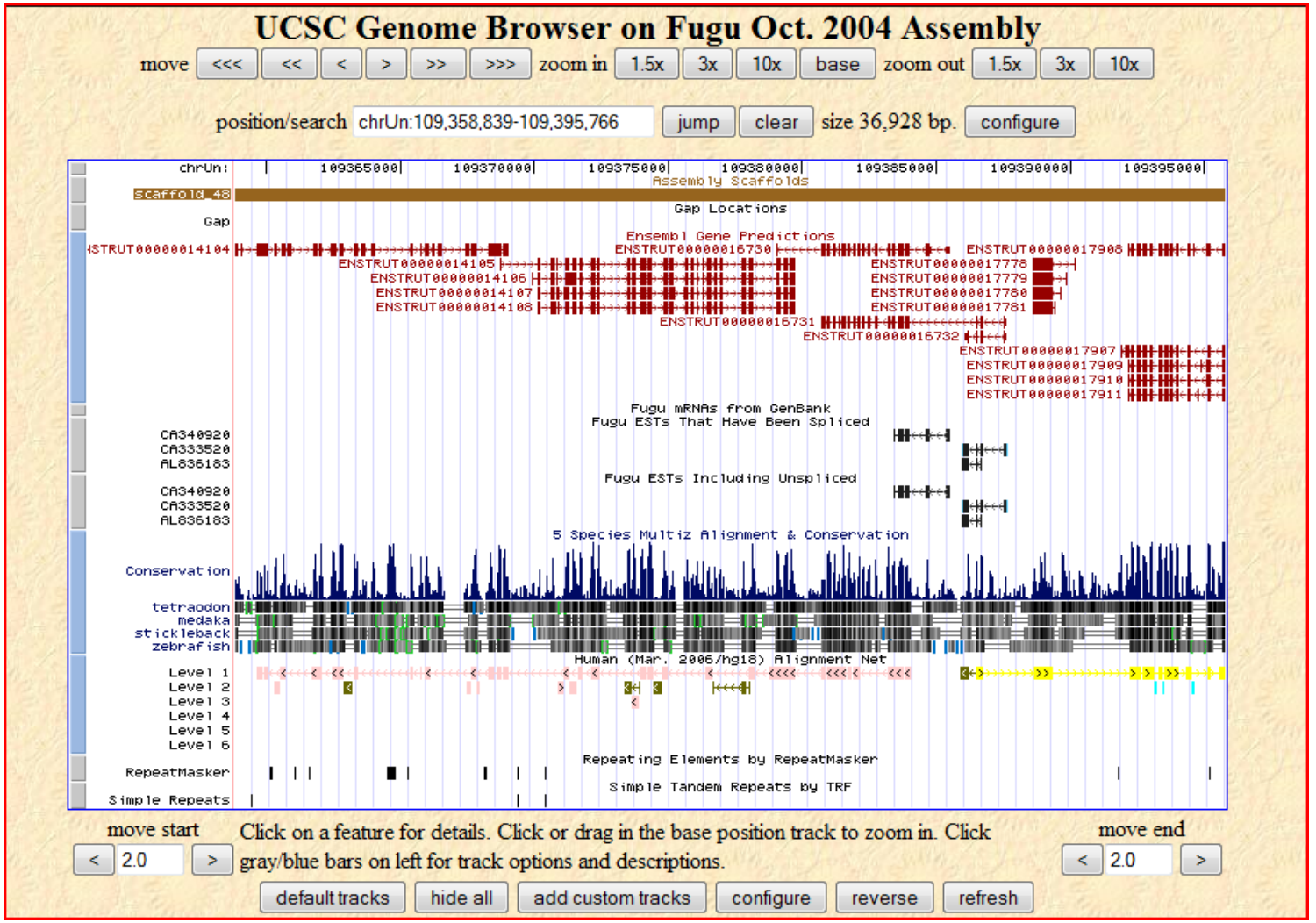
- AF164138 是一个长39757 bp的基因组序列片段,由协和医科大学刘勇等于1999年从河豚鱼中克隆到并递交到GenBank。
- 通过BioLand中整合的EMBOSS软件包中的needle，water，dottup等工具，以及基因预测程序GeneScan等，分析该序列片段，选取其中的功能片段进行blast，以推测其功能。
- 分析表明，该序列片段5’端含两个高度相似的多药耐药基（Multi-drug Resistance gene，简称MDR）。

方法和结果

1. 从GeneBank中找出MDR(AF164138)



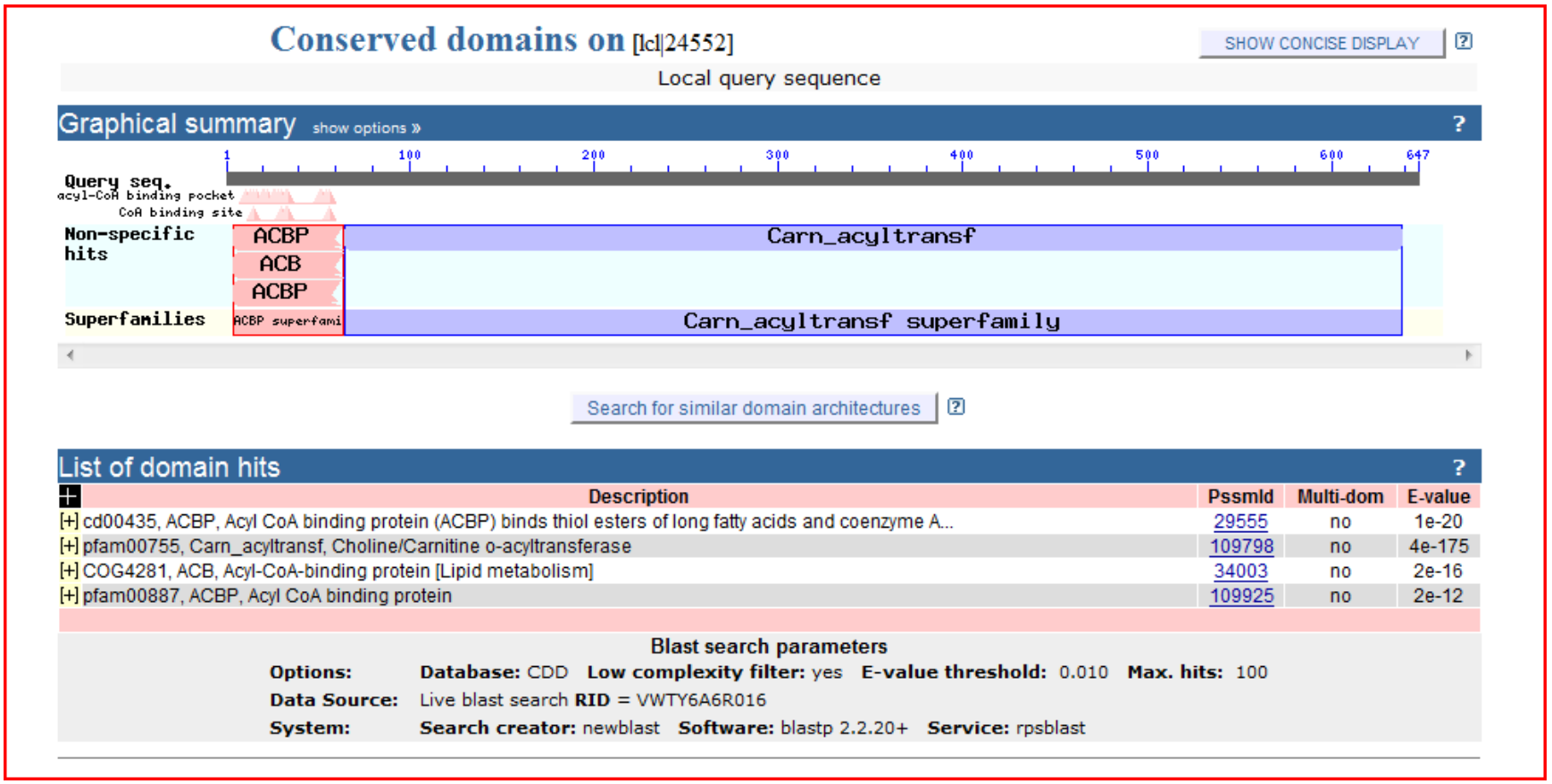

```
LOCUS       AF164138               39757 bp    DNA     linear   VRT 12-JAN-2003
DEFINITION  Takifugu rubripes cosmid 124A22 Mdr2, Mdr3, carnitine
            octanoyltransferase, RNase P38, and N-myristoyltransferase 2 genes,
            complete cds.
ACCESSION   AF164138
VERSION     AF164138.1  GI:27656756
KEYWORDS    .
SOURCE      Takifugu rubripes (Fugu rubripes)
```



Chromosome Name	Ensembl Gene ID	Gene Start (bp)	Gene End (bp)	Ensembl Transcript ID	Transcript Start (bp)	Transcript End (bp)	Associated Gene Name
scaffold_48	ENSTRUG00000005770	103925	124831	ENSTRUT00000014104	103925	114129	Q804Z6_FUGRU
scaffold_48	ENSTRUG00000005770	103925	124831	ENSTRUT00000014106	115010	124831	Q804Z6_FUGRU
scaffold_48	ENSTRUG00000006780	124127	132705	ENSTRUT00000016731	125809	132705	Q804Z5_FUGRU
scaffold_48	ENSTRUG00000007197	133649	135269	ENSTRUT00000017781	133655	134504	Q804Z4_FUGRU
scaffold_48	ENSTRUG00000007236	136977	140852	ENSTRUT00000017907	136977	140852	Q804Z3_FUGRU

Ensemble BioMart

有EST的部分序列分析



TrEMBL Results

Unreviewed, UniProtKB/TrEMBL **Q804Z5** (Q804Z5_FUGRU)

Last modified October 14, 2008. Version 25. History...

Clusters with 100%, 90%, 50% identity | Third-party data | Customize display

Names and origin · Protein attributes · General annotation (Comments) · Ontologies · Sequences · References · Cross-references · Entry inf

Names and origin

Protein names	*Submitted name:* **Carnitine octanoyltransferase** EMBL AAO20903.1
Organism	**Fugu rubripes (Japanese pufferfish) (Takifugu rubripes)** EMBL AAO20903.1
Taxonomic identifier	31033 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Actino Tetraodontiformes › Tetradontoidea › Tetraodontidae › Takifugu

Protein attributes

Sequence length	647 AA.
Sequence status	Complete.
Sequence processing	The displayed sequence is not processed.
Protein existence	Predicted.

General annotation (Comments)

Sequence similarities	Contains 1 ACB (acyl-CoA-binding) domain. Spearmint SPM000582

方法和结果

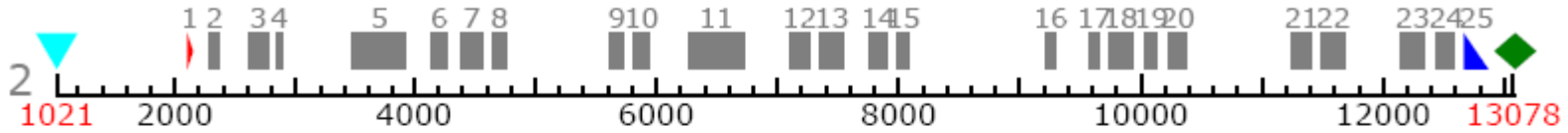
- 使用软件预测编码序列

Softberry - Comprehensive gene identification server (Commercial)
HMMgene - Prediction of vertebrate and C. elegans genes (CBS, Denmark)
GenScan - Identification of complete gene structures in genomic DNA (MIT)
DIGIT-consensus-based programs

Show picture of predicted genes in PDF file

FGENESH 2.6 Prediction of potential genes in Fish genomic DNA
Time : Sun Mar 16 22:25:17 2008
Seq name: AF164138 AF164138.1 Takifugu rubripes cosmid 124A22 Mdr2, Mdr3, carnitine octa
Length of sequence: 39757
Number of predicted genes 7: in +chain 4, in -chain 3.
Number of predicted exons 87: in +chain 54, in -chain 33.
Positions of predicted genes and exons: Variant 1 from 1, Score:924.954395

G	Str	Feature	Start		End	Score	ORF		Len
1	+	1 CDSi	53	-	199	13.65	53	- 199	147
1	+	2 CDSl	272	-	478	16.38	272	- 478	207
1	+	PolA	967			1.26			
2	+	TSS	1021			-8.03			
2	+	1 CDSf	2099	-	2139	4.91	2099	- 2137	39
2	+	2 CDSi	2283	-	2367	3.42	2284	- 2367	84
2	+	3 CDSi	2603	-	2768	3.74	2603	- 2767	165
2	+	4 CDSi	2845	-	2878	4.23	2847	- 2876	30
2	+	5 CDSi	3461	-	3911	23.78	3462	- 3911	450
2	+	6 CDSi	4120	-	4244	9.31	4120	- 4242	123
2	+	7 CDSi	4364	-	4535	23.62	4365	- 4535	171
2	+	8 CDSi	4621	-	4734	7.05	4621	- 4734	114
2	+	9 CDSi	5595	-	5705	6.89	5595	- 5705	111
2	+	10 CDSi	5792	-	5917	15.14	5792	- 5917	126
2	+	11 CDSi	6242	-	6700	42.22	6242	- 6700	459
2	+	12 CDSi	7078	-	7239	18.17	7078	- 7239	162
2	+	13 CDSi	7331	-	7525	17.70	7331	- 7525	195
2	+	14 CDSi	7740	-	7886	7.32	7740	- 7886	147
2	+	15 CDSi	7966	-	8070	5.48	7966	- 8070	105
2	+	16 CDSi	9190	-	9282	7.99	9190	- 9282	93
2	+	17 CDSi	9556	-	9639	8.50	9556	- 9639	84
2	+	18 CDSi	9713	-	9916	22.26	9713	- 9916	204
2	+	19 CDSi	10018	-	10118	8.79	10018	- 10116	99
2	+	20 CDSi	10216	-	10356	11.96	10217	- 10354	138
2	+	21 CDSi	11230	-	11386	7.60	11231	- 11386	156
2	+	22 CDSi	11474	-	11671	24.93	11474	- 11671	198
2	+	23 CDSi	12123	-	12329	23.24	12123	- 12329	207
2	+	24 CDSi	12432	-	12578	13.15	12432	- 12578	147
2	+	25 CDSl	12651	-	12857	17.03	12651	- 12857	207
2	+	PolA	13078			1.26			



2	+		TSS	1021			-8.03			
2	+	1	CDSf	2099	-	2139	4.91	2099	-	2137	39
2	+	2	CDSi	2283	-	2367	3.42	2284	-	2367	84
2	+	3	CDSi	2603	-	2768	3.74	2603	-	2767	165
2	+	4	CDSi	2845	-	2878	4.23	2847	-	2876	30
2	+	5	CDSi	3461	-	3911	23.78	3462	-	3911	450
2	+	6	CDSi	4120	-	4244	9.31	4120	-	4242	123
2	+	7	CDSi	4364	-	4535	23.62	4365	-	4535	171
2	+	8	CDSi	4621	-	4734	7.05	4621	-	4734	114
2	+	9	CDSi	5595	-	5705	6.89	5595	-	5705	111
2	+	10	CDSi	5792	-	5917	15.14	5792	-	5917	126
2	+	11	CDSi	6242	-	6700	42.22	6242	-	6700	459
2	+	12	CDSi	7078	-	7239	18.17	7078	-	7239	162
2	+	13	CDSi	7331	-	7525	17.70	7331	-	7525	195
2	+	14	CDSi	7740	-	7886	7.32	7740	-	7886	147
2	+	15	CDSi	7966	-	8070	5.48	7966	-	8070	105
2	+	16	CDSi	9190	-	9282	7.99	9190	-	9282	93
2	+	17	CDSi	9556	-	9639	8.50	9556	-	9639	84
2	+	18	CDSi	9713	-	9916	22.26	9713	-	9916	204
2	+	19	CDSi	10018	-	10118	8.79	10018	-	10116	99
2	+	20	CDSi	10216	-	10356	11.96	10217	-	10354	138
2	+	21	CDSi	11230	-	11386	7.60	11231	-	11386	156
2	+	22	CDSi	11474	-	11671	24.93	11474	-	11671	198
2	+	23	CDSi	12123	-	12329	23.24	12123	-	12329	207
2	+	24	CDSi	12432	-	12578	13.15	12432	-	12578	147
2	+	25	CDSl	12651	-	12857	17.03	12651	-	12857	207
2	+		PolA	13078			1.26			

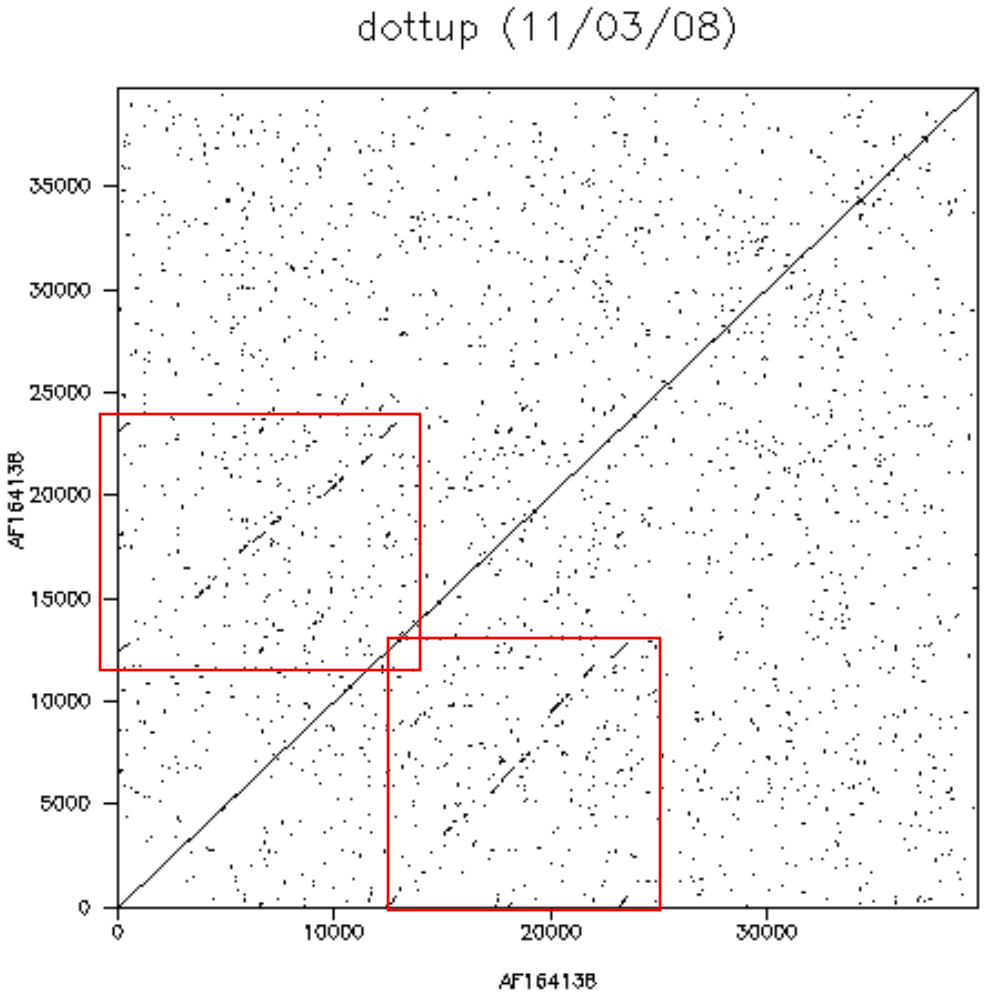
方法和结果

2. 从找到的GeneBank序列中找出重复片段

1）提取FASTA格式的全序列：seqret 文件名

2）dottup将fasta格式的数据进行自我比对，参数为wordsize=10

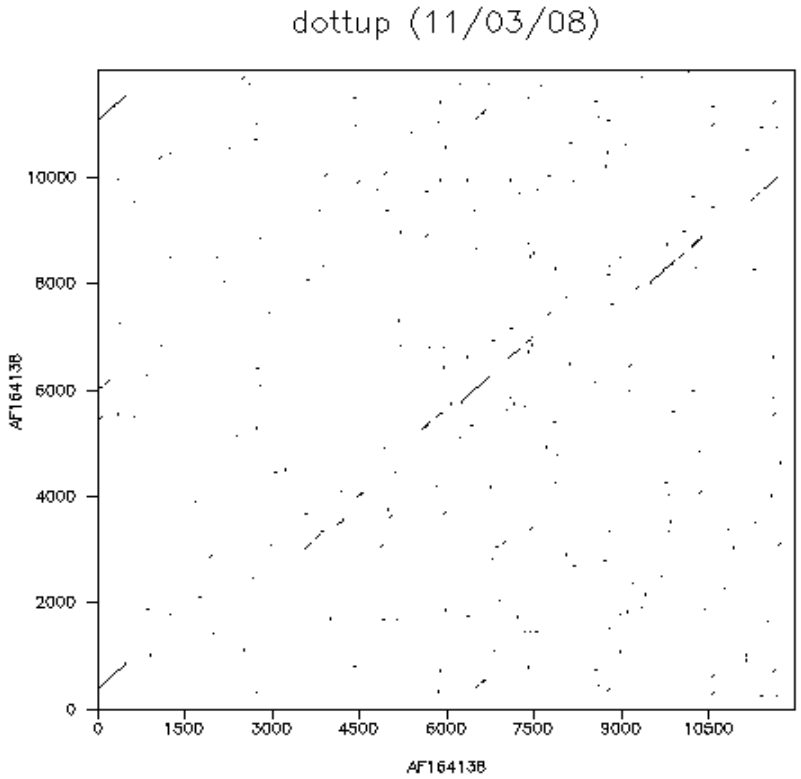
方法和结果


dottup (11/03/08)
35000
30000
25000
20000
15000
10000
5000
0
AF164138
0
10000
20000
30000
AF164138

方法和结果

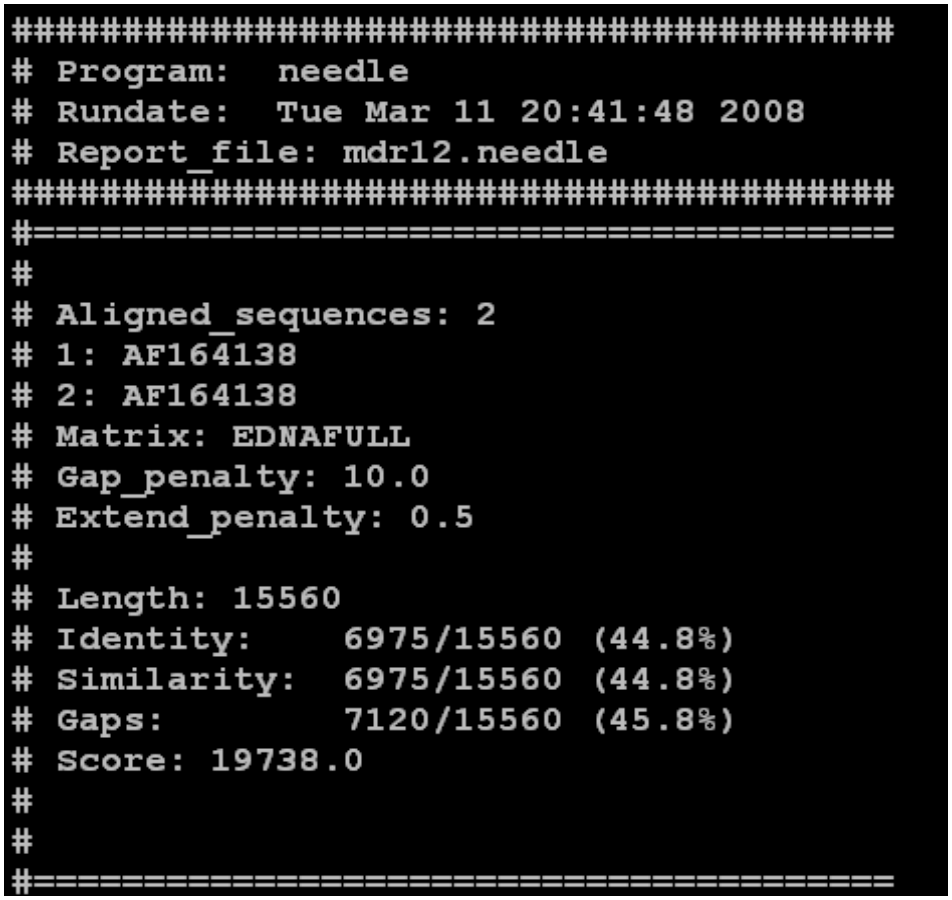
3）根据上图结果，用seqret命令将这两段序列分别取出来（1-12000bp和12001-24000bp），dottup比对

seqret 序列名 –sbegin 1 –send 12000>新序列名



方法和结果

4）用needle进行比



```
########################################
# Program:  needle
# Rundate:  Tue Mar 11 20:41:48 2008
# Report_file: mdr12.needle
########################################
#=======================================
#
# Aligned_sequences: 2
# 1: AF164138
# 2: AF164138
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 15560
# Identity:    6975/15560 (44.8%)
# Similarity:  6975/15560 (44.8%)
# Gaps:        7120/15560 (45.8%)
# Score: 19738.0
#
#
#=======================================
```

方法和结果

3. 对AF164138的编码序列的分析

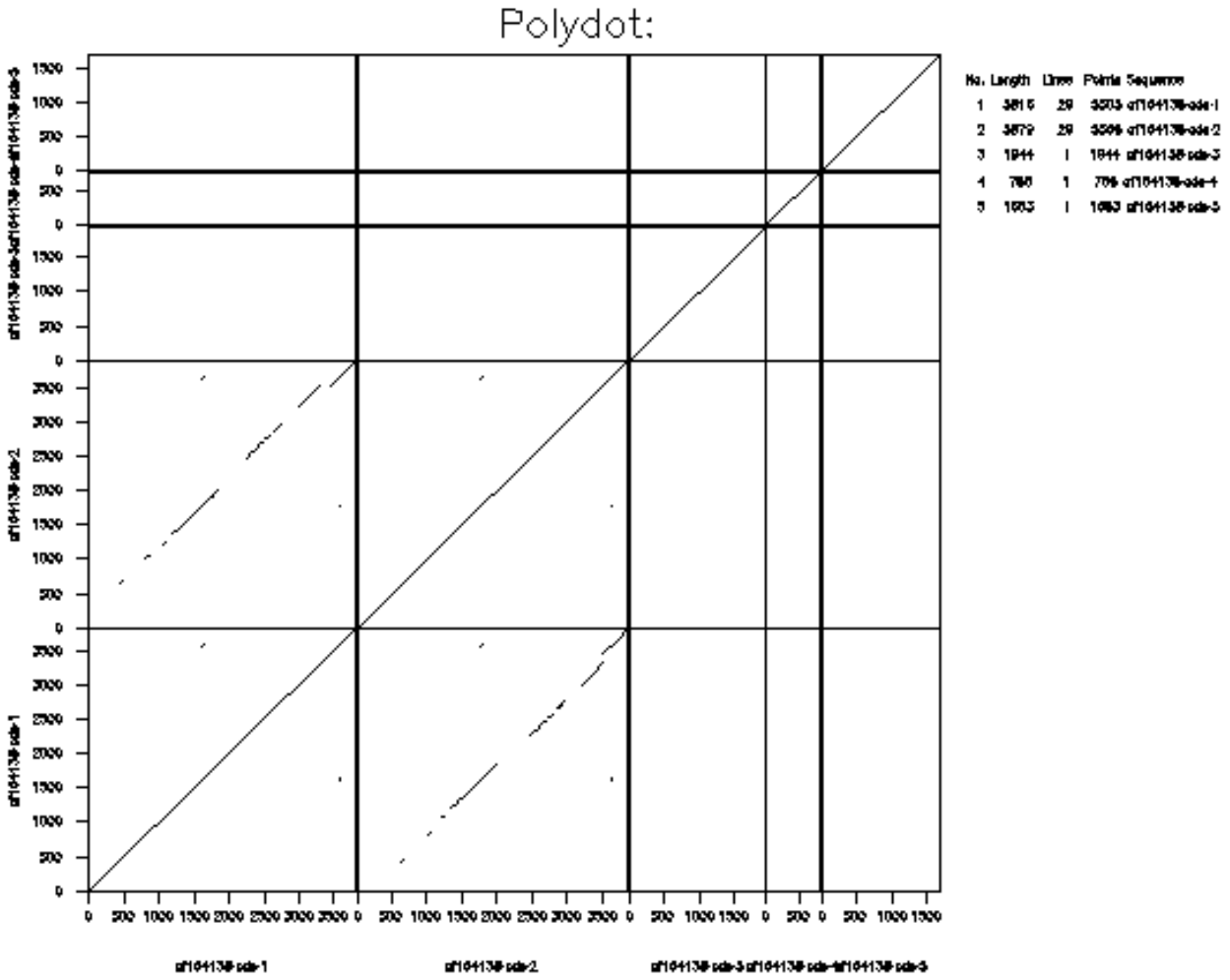
1）使用coderet命令从第一步下载的GeneBank格式的数据中提取所有CDS,mRNA和蛋白fasta格式的序列信息（codret 文件名），并用seqretsplit命令将序列分开，得到CDS、mRNA和蛋白序列各5条（seqretsplit 文件名）。（注意coderet 和seqret的区别）

2）比较编码区序列

使用cat命令将5个cds序列连成一个文件（cat file1 file2 file3 file4 file5>newfile）

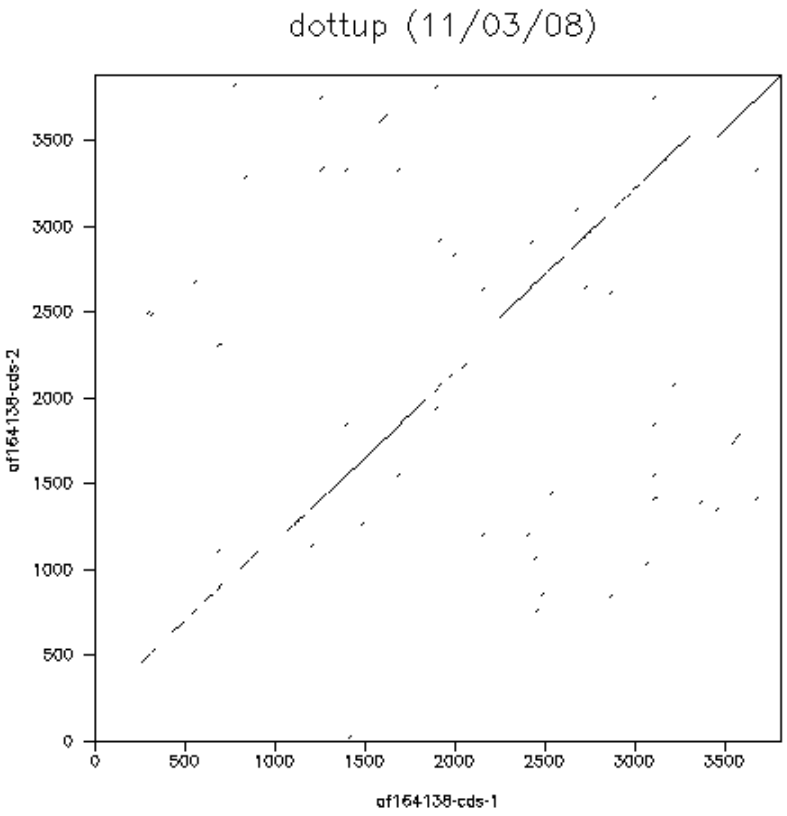
使用polydot分析相似性，wordsize=20.（polydot newfile）

方法和结果


Polydot:
No. Length Lines Points Sequence
1 3815 29 3503 af164136-sds-1
2 3879 29 3566 af164136-sds-2
3 1944 1 1944 af164136-sds-3
4 786 1 786 af164136-sds-4
5 1663 1 1663 af164136-sds-5
af164136-sds-1
af164136-sds-2
af164136-sds-3 af164136-sds-4 af164136-sds-5

方法和结果

3）分别用dottup（wordsize=10)和needle对cds1和cds2进行比对

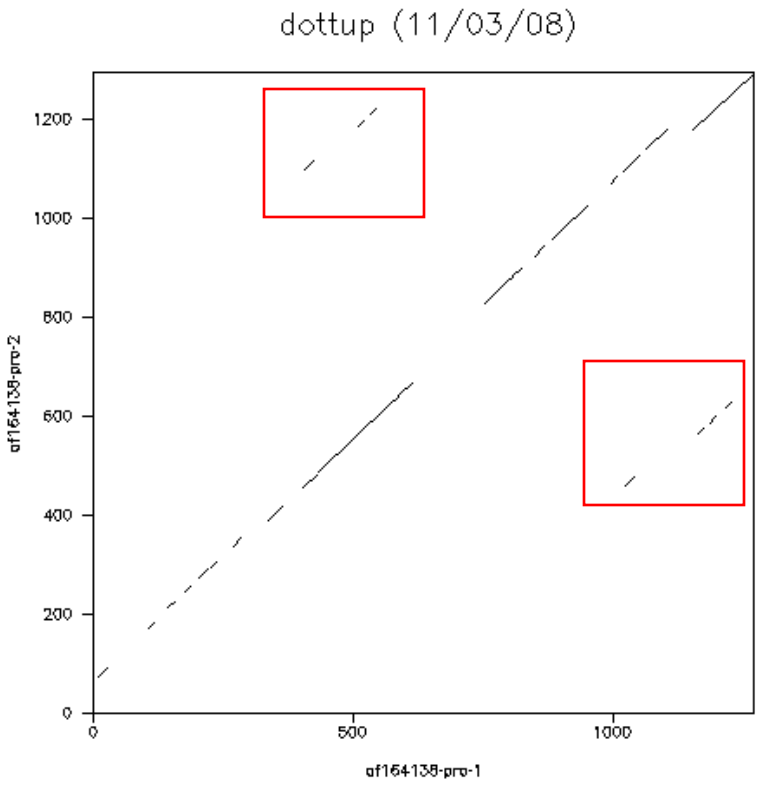



```
########################################
# Program:  needle
# Rundate:  Tue Mar 11 21:17:30 2008
# Report_file: cds12-.needle
########################################
#=======================================
#
# Aligned_sequences: 2
# 1: af164138_cds_1
# 2: af164138_cds_2
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 4159
# Identity:    3098/4159 (74.5%)
# Similarity:  3098/4159 (74.5%)
# Gaps:         623/4159 (15.0%)
# Score: 13047.5
#
#
#=======================================
```

方法和结果

4. 对AF164138的蛋白序列进行分析

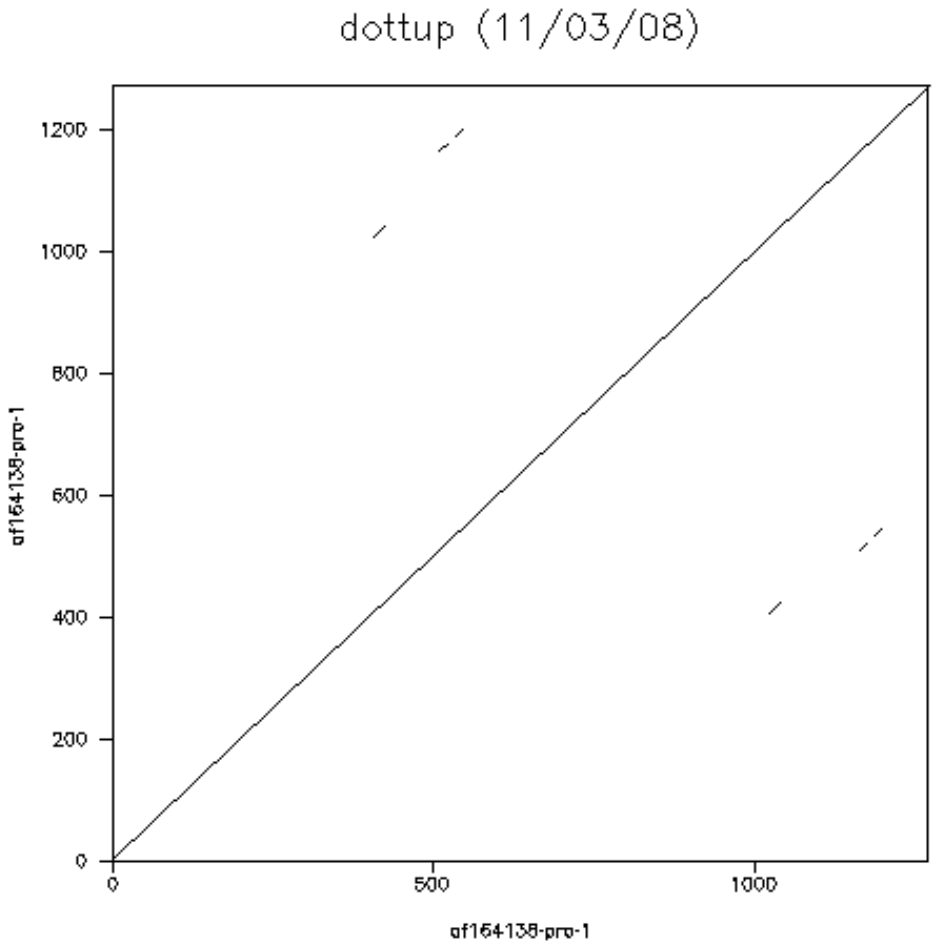
1）用dottup和needle对相应的PRO1和PRO2进行比对。




```
########################################
# Program:  needle
# Rundate:  Tue Mar 11 21:22:08 2008
# Report_file: pro12.needle
########################################
#=======================================
#
# Aligned_sequences: 2
# 1: af164138_pro_1
# 2: af164138_pro_2
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1372
# Identity:     976/1372 (71.1%)
# Similarity:  1077/1372 (78.5%)
# Gaps:         181/1372 (13.2%)
# Score: 4867.5
#
#
#=======================================
```

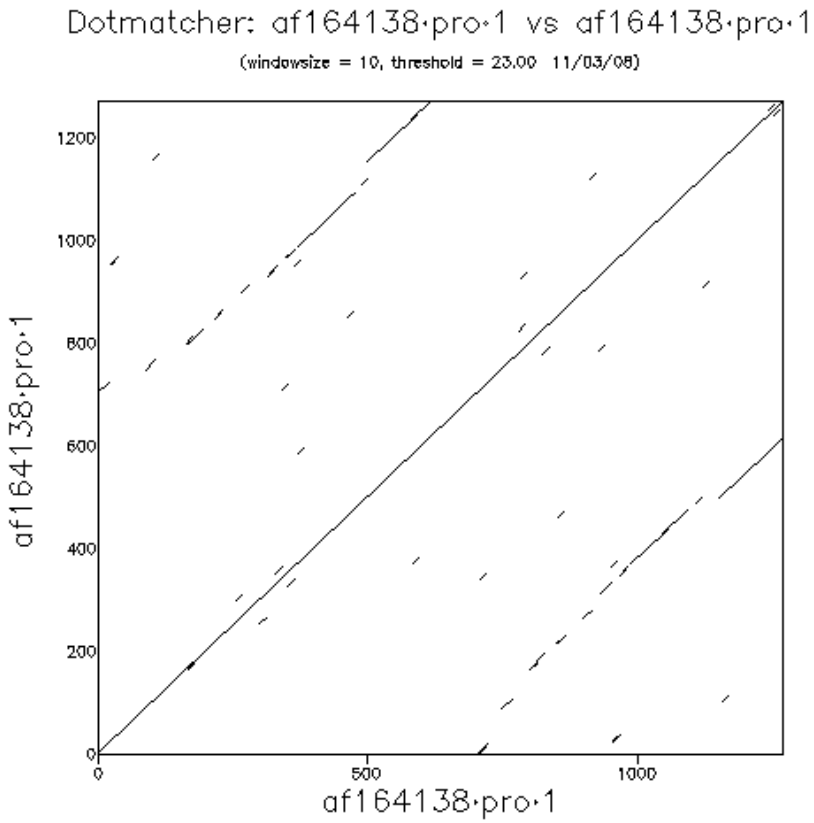
方法和结果

2）用dottup将pro1自我比对，参数为wordsize=10



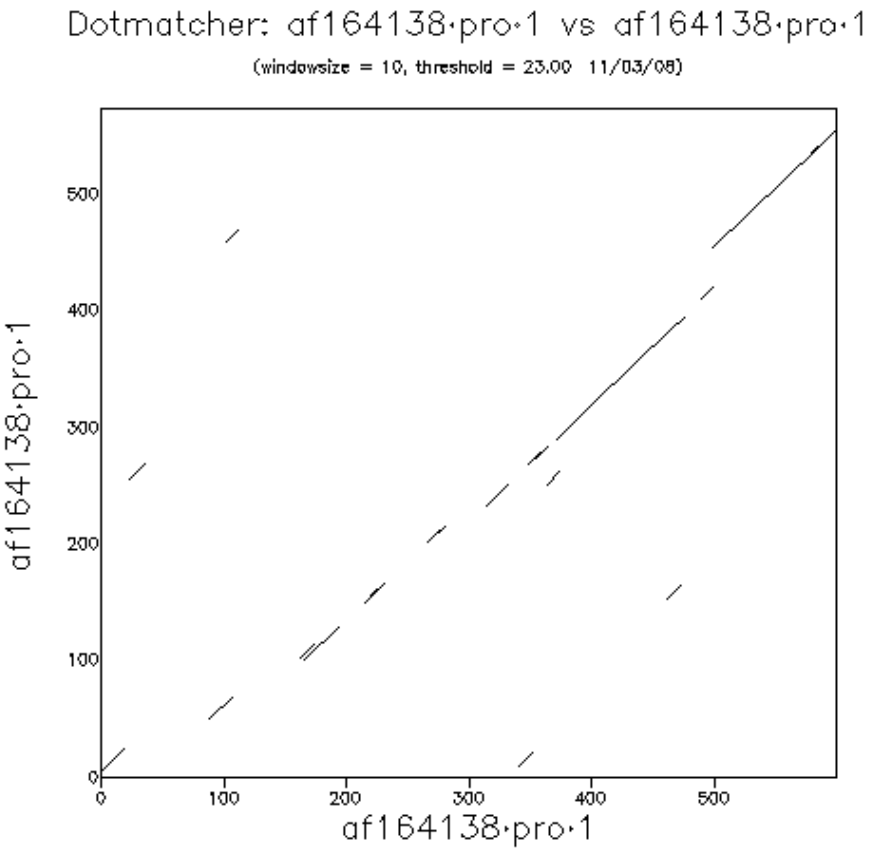
方法和结果

3）用dotmatcher进行进一步比较，设置参数为wordsize=10，threshold=23



方法和结果

4）根据上图用seqret命令截取两段重复序列1-600bp和700-1300bp，再将两者进行dotmatcher



方法和结果

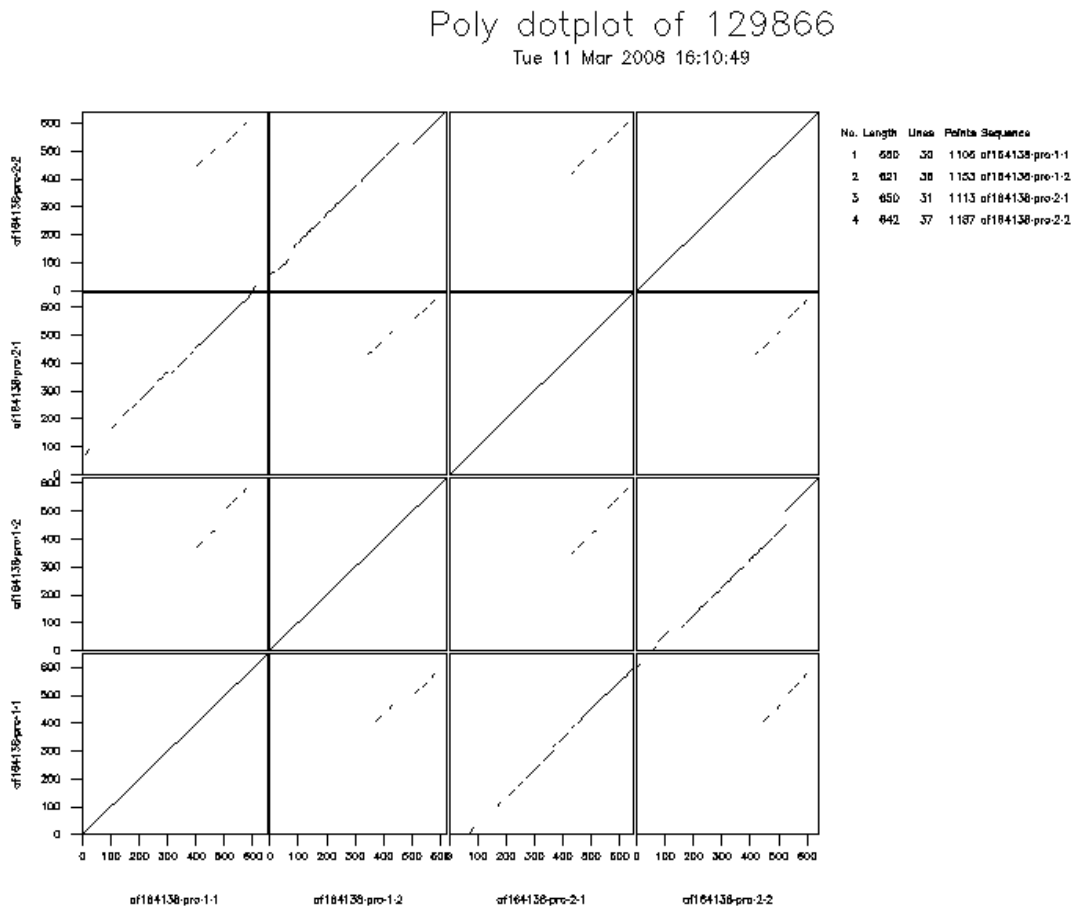
5）将两者进行needle和water比对

```
########################################
# Program:  needle
# Rundate:  Tue Mar 11 21:59:10 2008
# Report_file: pro1.1.2.needle
########################################
#=======================================
#
# Aligned_sequences: 2
# 1: af164138_pro_1
# 2: af164138_pro_1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 663
# Identity:     229/663 (34.5%)
# Similarity:   331/663 (49.9%)
# Gaps:         154/663 (23.2%)
# Score: 1013.5
#
#
#=======================================
```

```
########################################
# Program:  water
# Rundate:  Tue Mar 11 22:02:08 2008
# Report_file: pro1.1.2.water
########################################
#=======================================
#
# Aligned_sequences: 2
# 1: af164138_pro_1
# 2: af164138_pro_1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 638
# Identity:     229/638 (35.9%)
# Similarity:   331/638 (51.9%)
# Gaps:         132/638 (20.7%)
# Score: 1018.5
#
#
#=======================================
```

方法和结果

- 如果将这两个切下来的半段序列再进行polydot比较可以发现，这两个半段中前后也比较相似

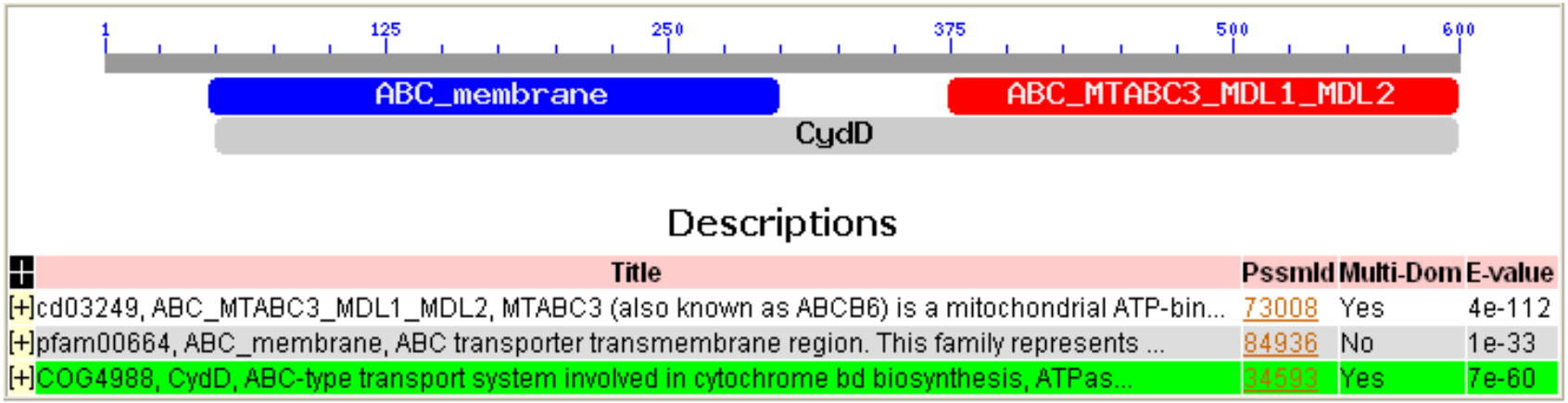


Polydot by Wei Liang

方法和结果

5.利用重复结构域序列BLAST查找其功能

取出PRO1的第一段序列（1-600bp）在NCBI的网站上进行blast，参数为：程序：BLASTP 2.2.18 、数据库：Non-redundant protein sequences (nr)、矩阵：BLOSUM 62 、Gap Costs：Existence 11，Extension 1。



方法和结果

- 最高比对结果ABC_MTABC3_MDL1_MDL2（ABCB6）结构域，它是一个线粒体ATP结合盒式蛋白，在铁稳态中起作用。MDL1是一个ATP依赖的渗透酶，且是ATM1的高拷贝抑制子，同时在氧压耐受性中起作用。
- CydD，ABC转运系统（ATP-binding cassette transporter）和细胞色素bd生物合成和相关，是渗透酶的组成成分。
- ABC_membrane是一个ABC转运体跨膜区域，这个家族每个六个跨膜螺旋为一个单位，许多ABC转运体家族成员有两个这样的区域。

讨论

1. 比较基因组重复序列(表格中的A)、编码区序列CDS1和CDS2（表格中的B）以及蛋白序列PRO1和PRO2（表格中的C）的序列相似性

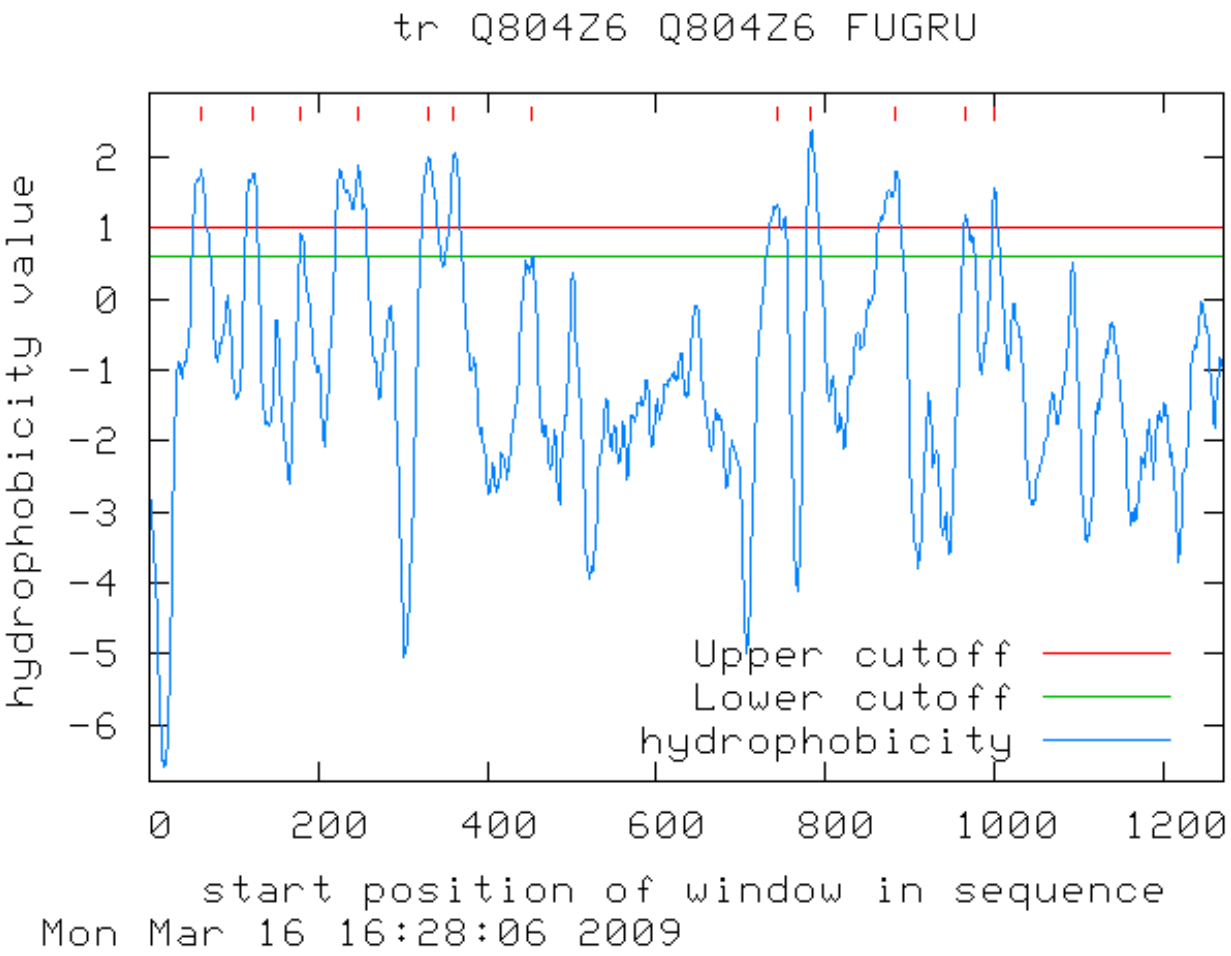
	A	B	C
Indentity	44.8%	74.5%	71.1%
Similarity	44.8%	74.5%	78.5%

A、B、C的相似性逐渐增加，推测是因为在A中存在了大量内含子，所以相似度最低，B中是编码区序列，C是蛋白序列,C去除了密码子简并性的影响，因而得到了最好的相似性。

2. BLAST结果分析

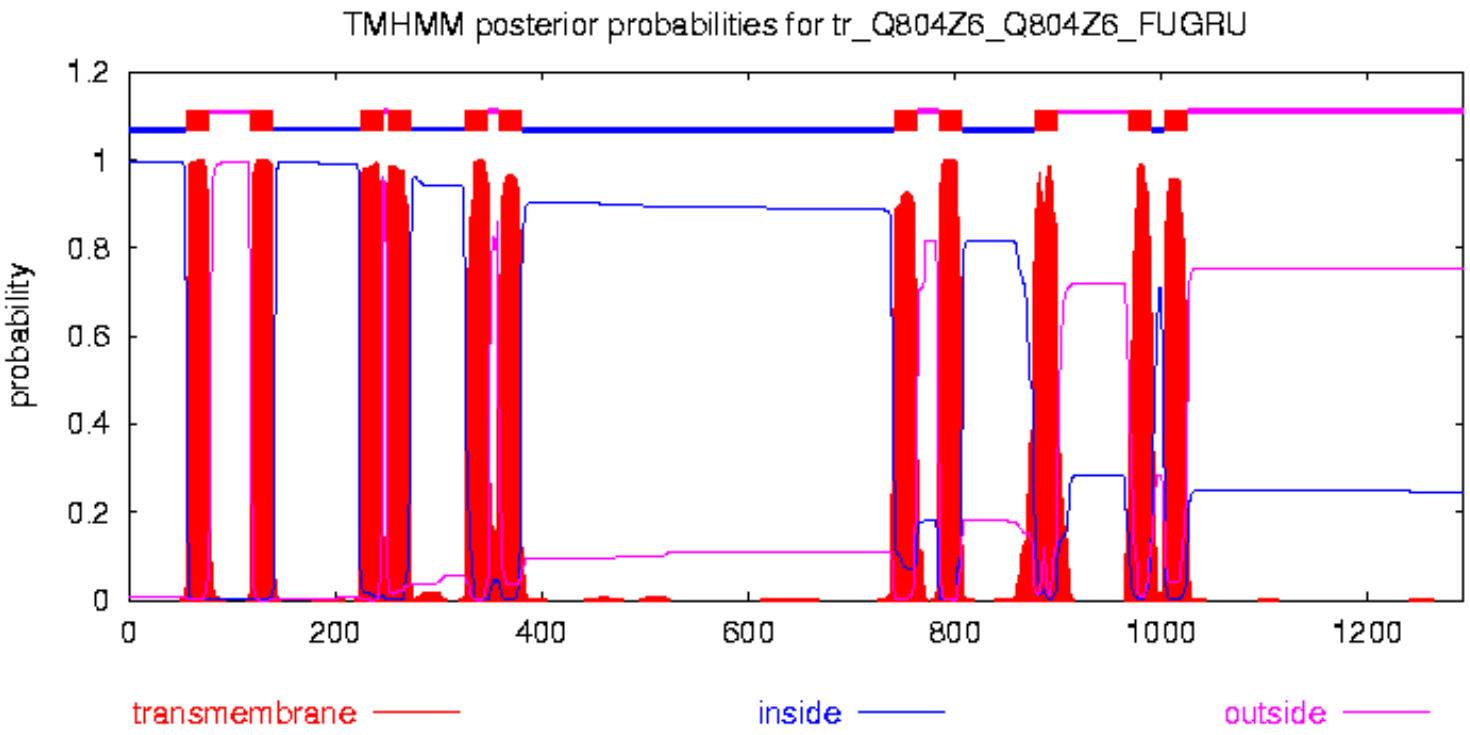
- ABC转运体是一个可以帮助代谢物进和出细胞的一个复合体，主要由四个结构域组成，两个细胞膜相关结构域，和两个在细胞膜内面的ATP结合结构域组成，这些结构域共同组成一个跨膜的中心孔。这四个核心的结构域可能作为分开的多肽编码，也可能以多种方式形成一个多结构域的多肽（GO: 0043190）。
- 根据上一段GO注释，我推测PRO1很有可能是一个同时含有四个核心结构域的ABC转运体。ABC转运体蛋白家族是细菌多药物抗性有关的蛋白[1]，由此可以推断原来的大段基因组序列确实是多药物抗性基因。

跨膜结构预测



TEMPRED 预测跨膜结构

TMHMM 预测


TMHMM posterior probabilities for tr_Q804Z6_Q804Z6_FUGRU
probability
1.2
1
0.8
0.6
0.4
0.2
0
0
200
400
600
800
1000
1200
transmembrane
inside
outside

TrEMBL Results

Unreviewed, UniProtKB/TrEMBL **Q804Z6** (Q804Z6_FUGRU)

Last modified February 10, 2009. Version 43. History...

Clusters with 100%, 90%, 50% identity | Third-party data | Customize display

Names and origin · Protein attributes · General annotation (Comments) · Ontologies · Sequences · References · Cross-references · Entry information

Names and origin

Protein names	*Submitted name:* **Mdr3** EMBL AAO20901.1
Organism	**Fugu rubripes (Japanese pufferfish) (Takifugu rubripes)** EMBL AAO20901.1
Taxonomic identifier	31033 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Actinopterygii › Ne Tetraodontiformes › Tetradontoidea › Tetraodontidae › Takifugu

Protein attributes

Sequence length	1292 AA.
Sequence status	Complete.
Sequence processing	The displayed sequence is not processed.
Protein existence	Inferred from homology.

General annotation (Comments)

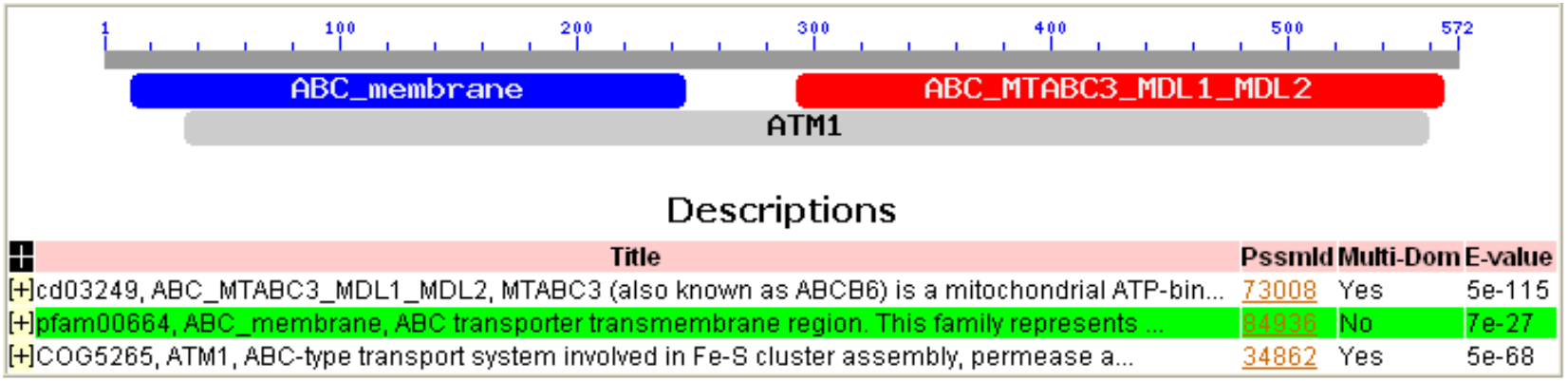
Sequence similarities	Belongs to the ABC transporter family. RuleBase RU000684V4

TrEMBL Results

Sequence databases	
EMBL	AF164138 Genomic DNA. Translation: AAO20901.1.
3D structure databases	
HSSP	HSSP built from PDB template 1MT0 based on UniProtKB P08716.
ModBase	Search...
Phylogenomic databases	
HOVERGEN	Q804Z6.
Family and domain databases	
InterPro	IPR001140. ABC_TM_transpt. IPR003439. ABC_transporter-like. IPR017871. ABC_transporter_CS. IPR017940. ABC_transporter_type1. IPR003593. ATPase_AAA+_core. [Graphical view]
Pfam	PF00664. ABC_membrane. 2 hits. PF00005. ABC_tran. 2 hits. [Graphical view]
ProDom	PD000006. ABC_transporter. 2 hits. [Graphical view] [Entries sharing at least one domain]
SMART	SM00382. AAA. 2 hits. [Graphical view]
PROSITE	PS50929. ABC_TM1F. 2 hits. PS00211. ABC_TRANSPORTER_1. 2 hits. PS50893. ABC_TRANSPORTER_2. 2 hits. [Graphical view]
ProtoNet	Search...

补充材料

- PRO1的第二段序列的比对结果



Descriptions

Title	PssmId	Multi-Dom	E-value
[+]cd03249, ABC_MTABC3_MDL1_MDL2, MTABC3 (also known as ABCB6) is a mitochondrial ATP-bin...	73008	Yes	5e-115
[+]pfam00664, ABC_membrane, ABC transporter transmembrane region. This family represents ...	84936	No	7e-27
[+]COG5265, ATM1, ABC-type transport system involved in Fe-S cluster assembly, permease a...	34862	Yes	5e-68

ATM也是ABC转运体系统中的一员，与Fe-S簇组装相关，是渗透酶和ATPase的组成成分。

参考资料

- [1] ATP-binding cassette transporter，From Wikipedia, the free encyclopedia，2008-3-11，**http://en.wikipedia.org/wiki/ATP-binding_cassette_transporter**
- NCBI
- GO

致谢

- 梁巍

  参考了梁巍同学的答案中polydot部分
- 罗老师

Thank you !