https://www.agr.nagoya-u.ac.jp/~bioinfo/

第6回　遺伝子の探索と遺伝子の機能予測

本日学ぶこと

データベースを使って遺伝子を探して、
かつその遺伝子の機能を予測する。

これまで大学の授業の中で、DNA、RNA、アミノ酸、
蛋白質について構造や特性などを学習してきた。

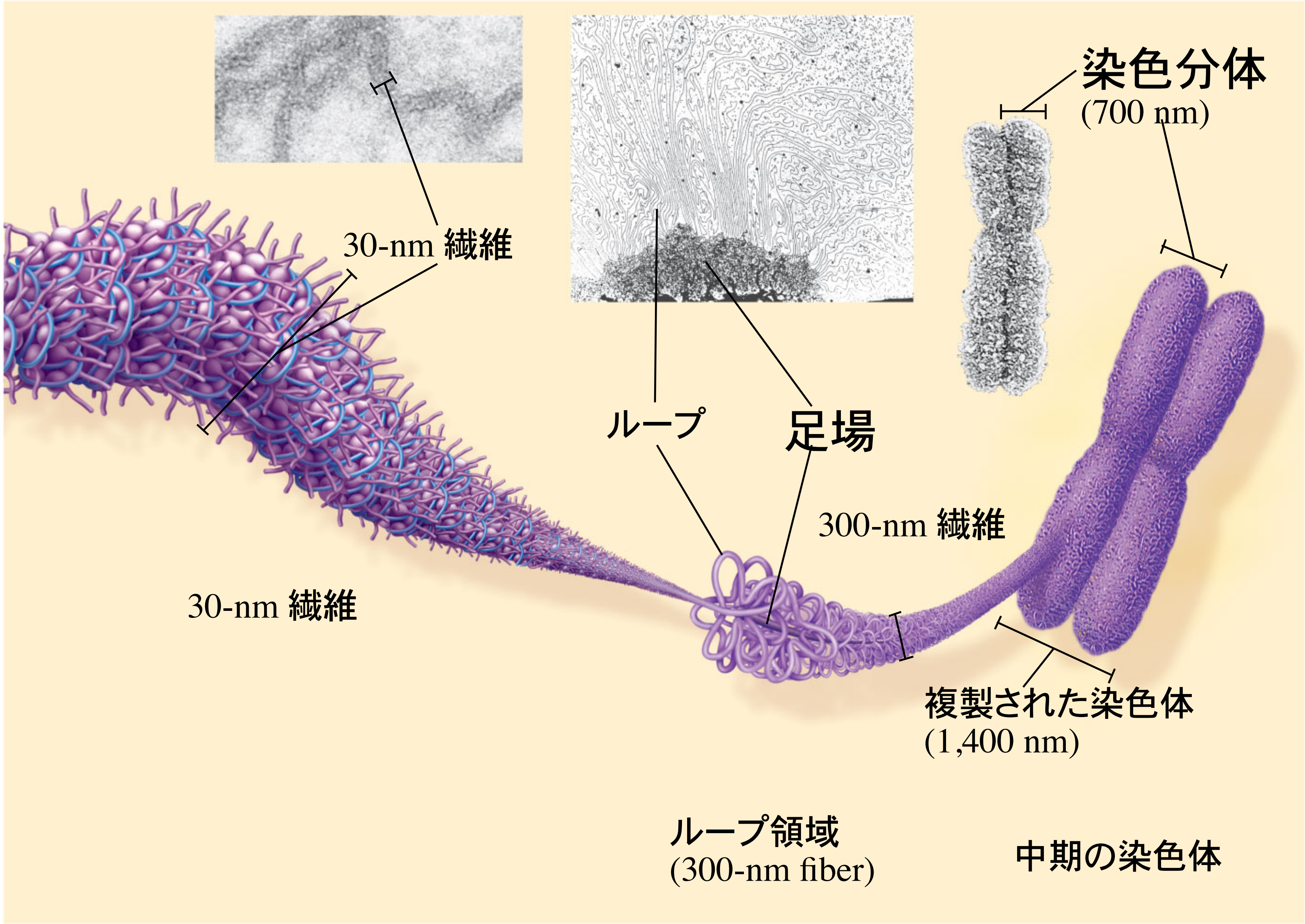
また、DNAの一部が転写され、スプラシングをうけた後、
蛋白質に翻訳されることを学習してきた。

しかし、実際長いDNA配列のどの部分が遺伝子なのだろうか?

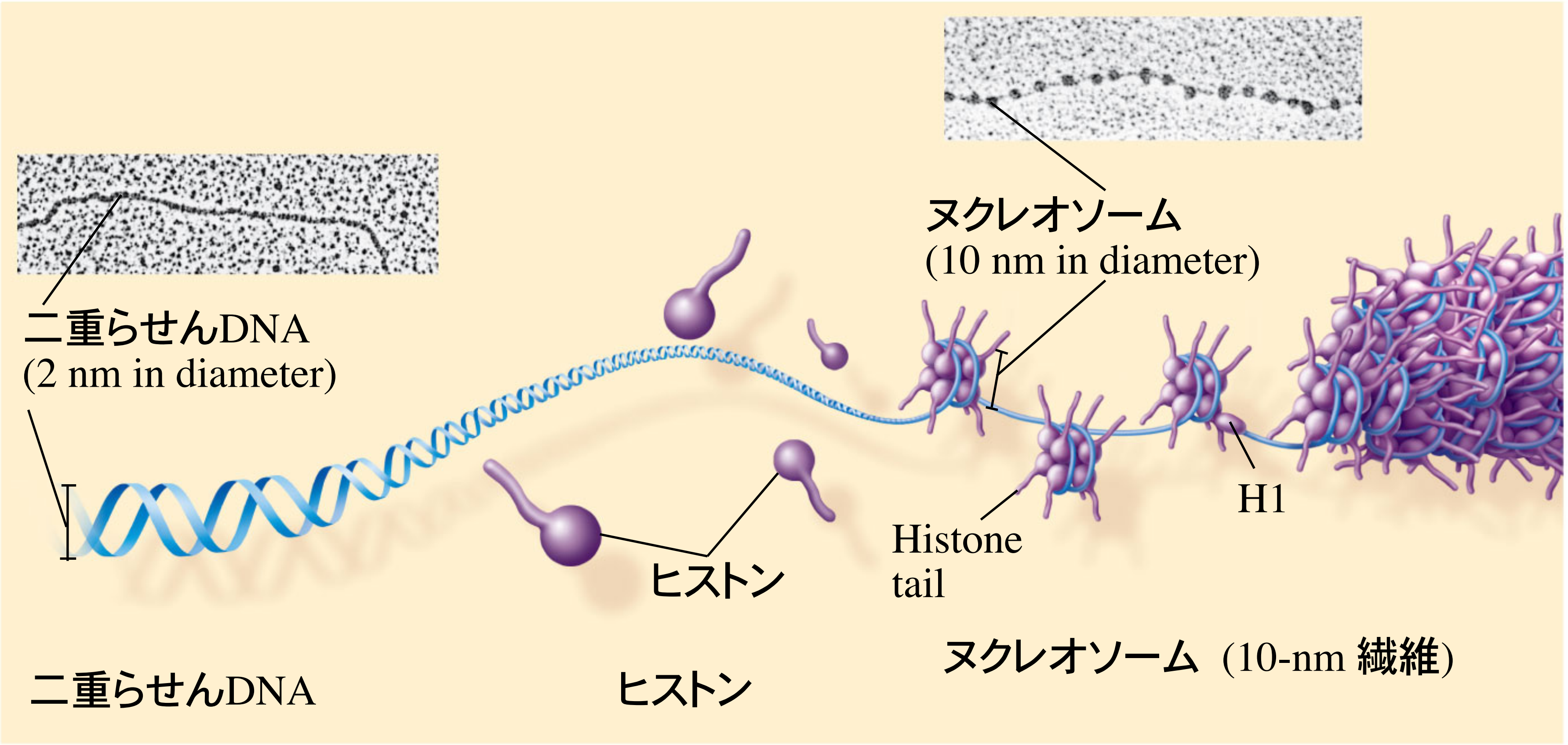
少し復習をしてみよう。

人間の染色体（核型）

キャンベル参照

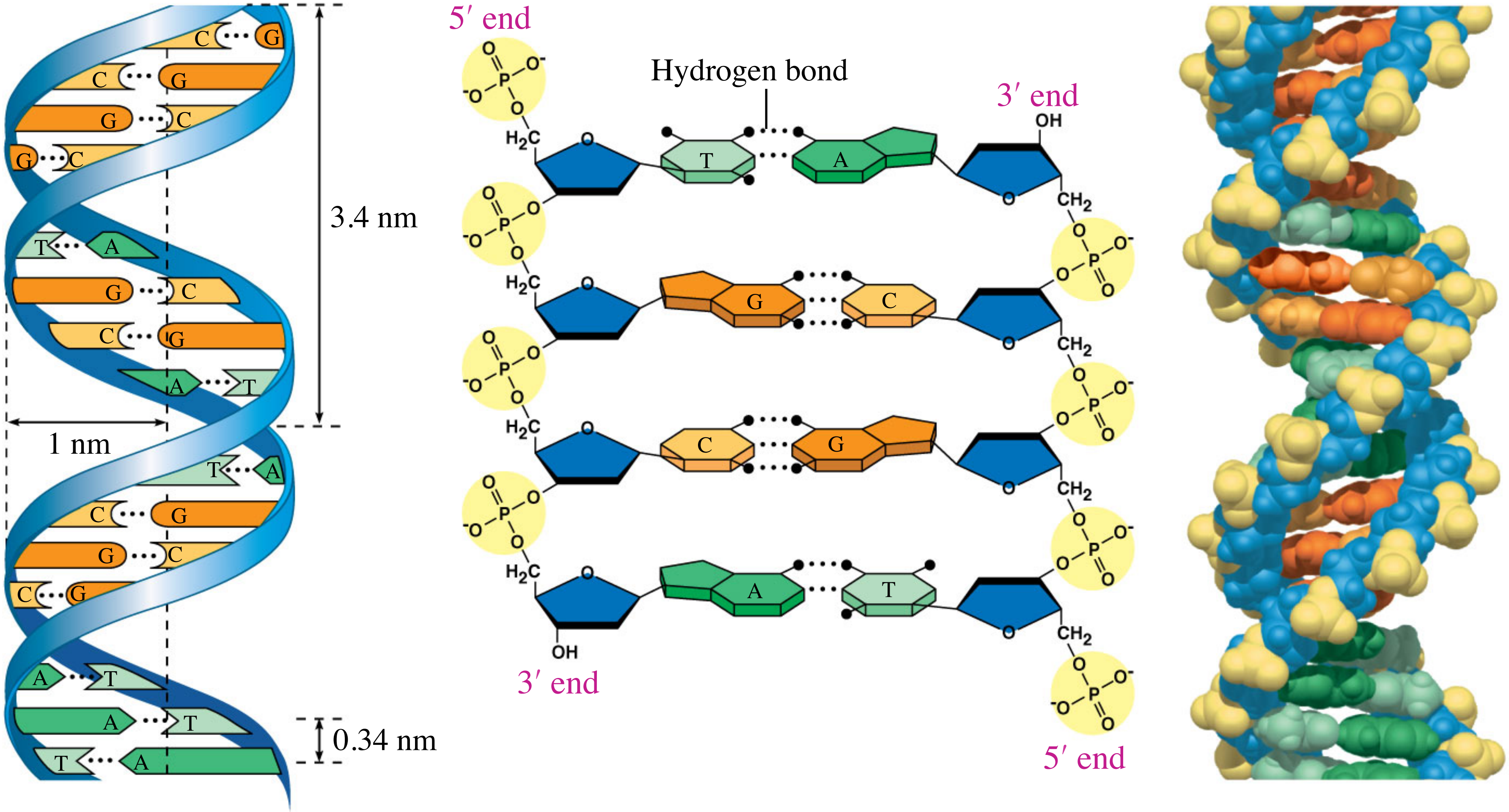

30-nm 繊維
染色分体
(700 nm)
ループ
足場
300-nm 繊維
30-nm 繊維
複製された染色体
(1,400 nm)
ループ領域
(300-nm fiber)
中期の染色体

キャンベル参照


ヌクレオソーム
(10 nm in diameter)
二重らせんDNA
(2 nm in diameter)
H1
Histone
tail
ヒストン
ヌクレオソーム (10-nm 繊維)
二重らせんDNA
ヒストン

キャンベル参照

Figure 16.7



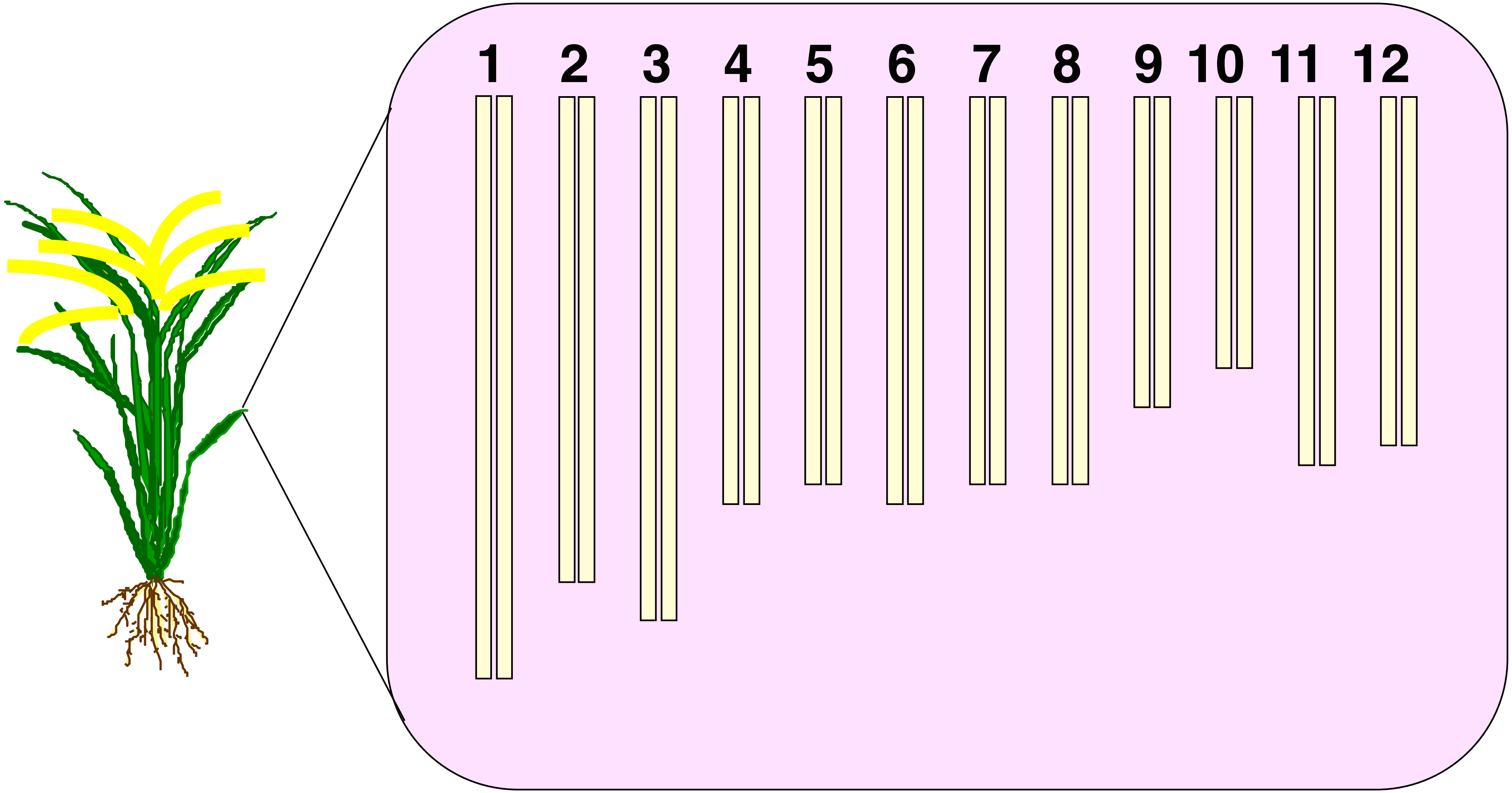
(a) Key features of DNA structure

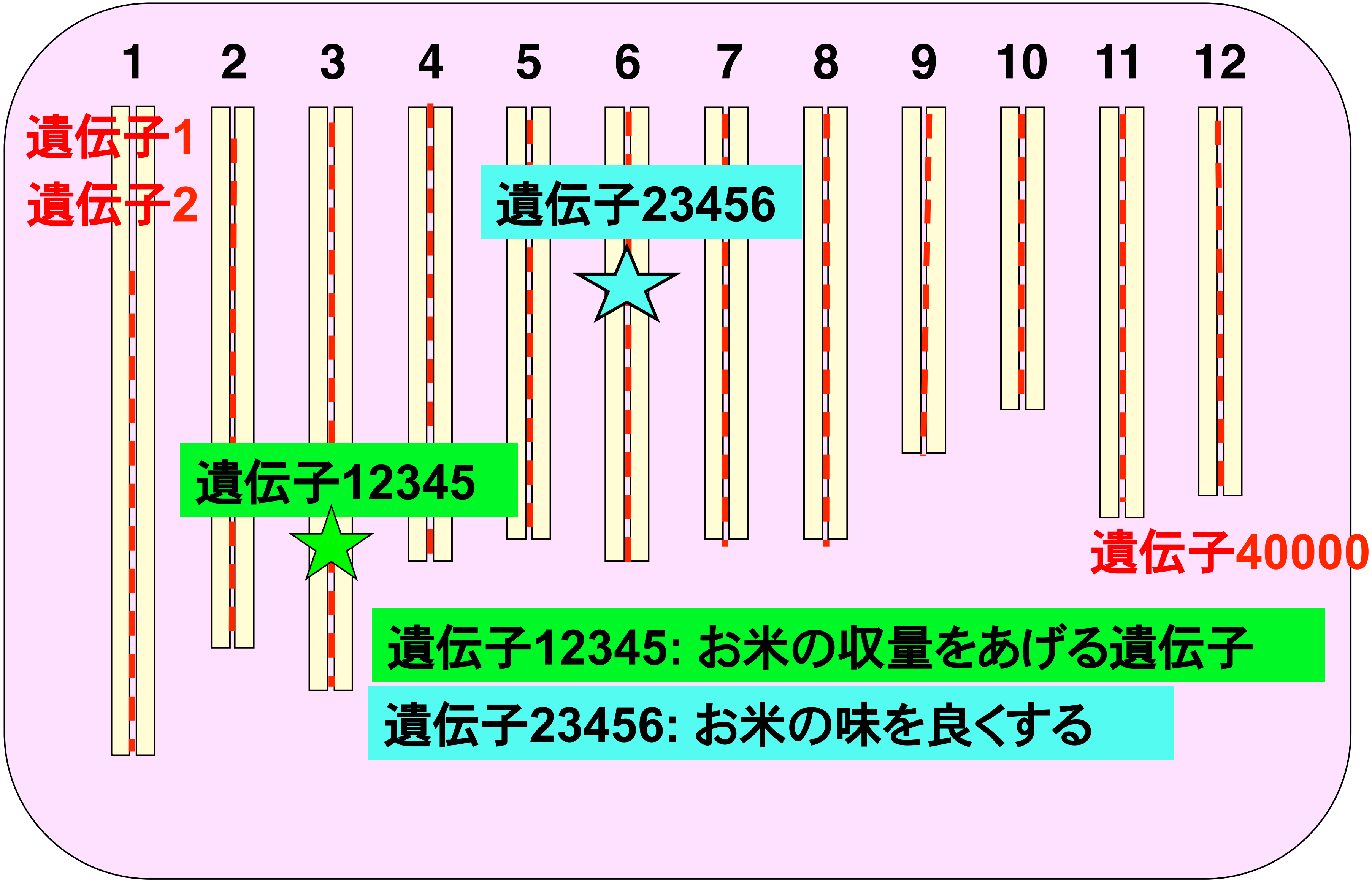
(b) Partial chemical structure

(c) Space-filling model

キャンベル参照

イネの細胞の中の染色体とDNA

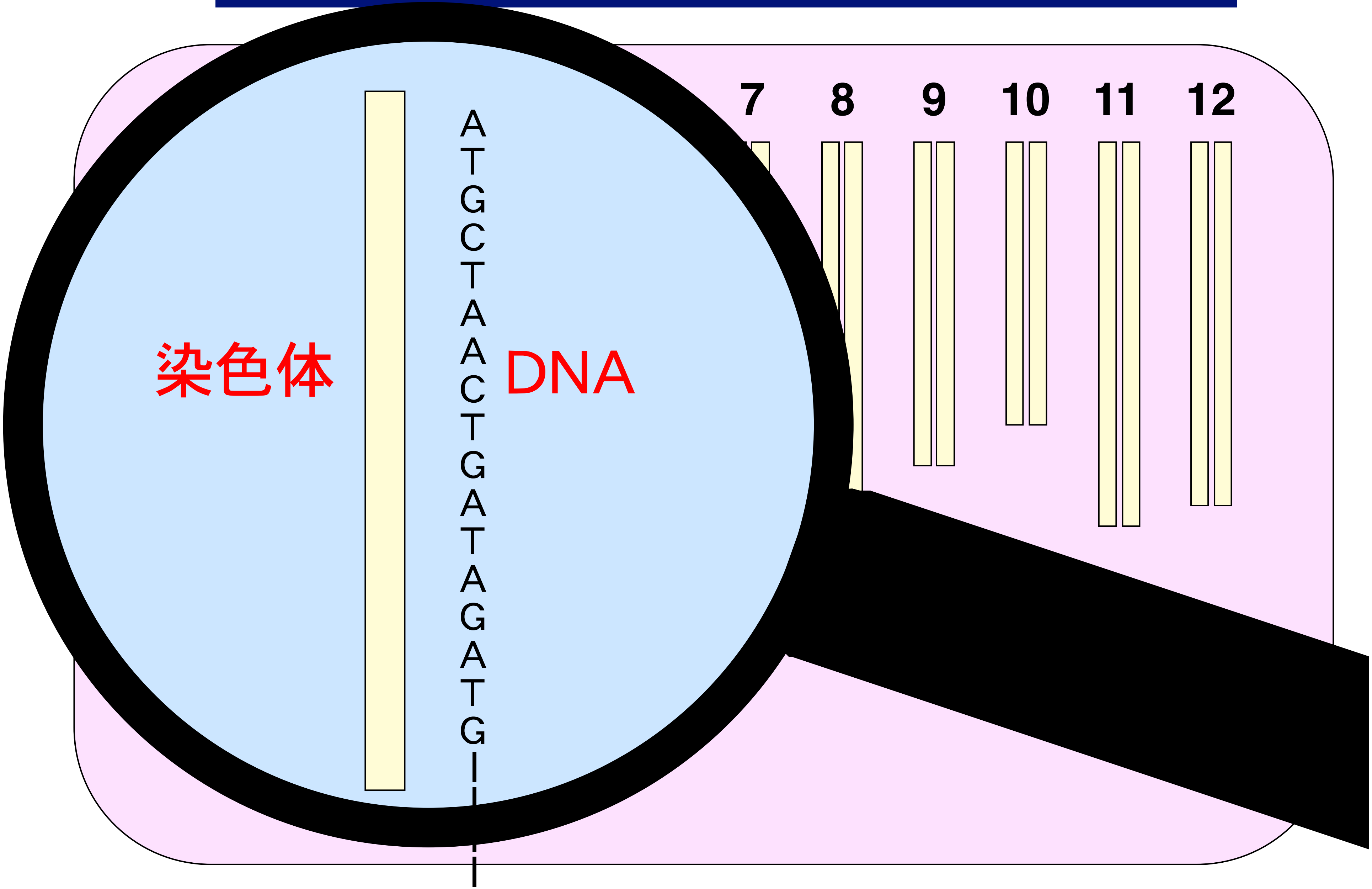


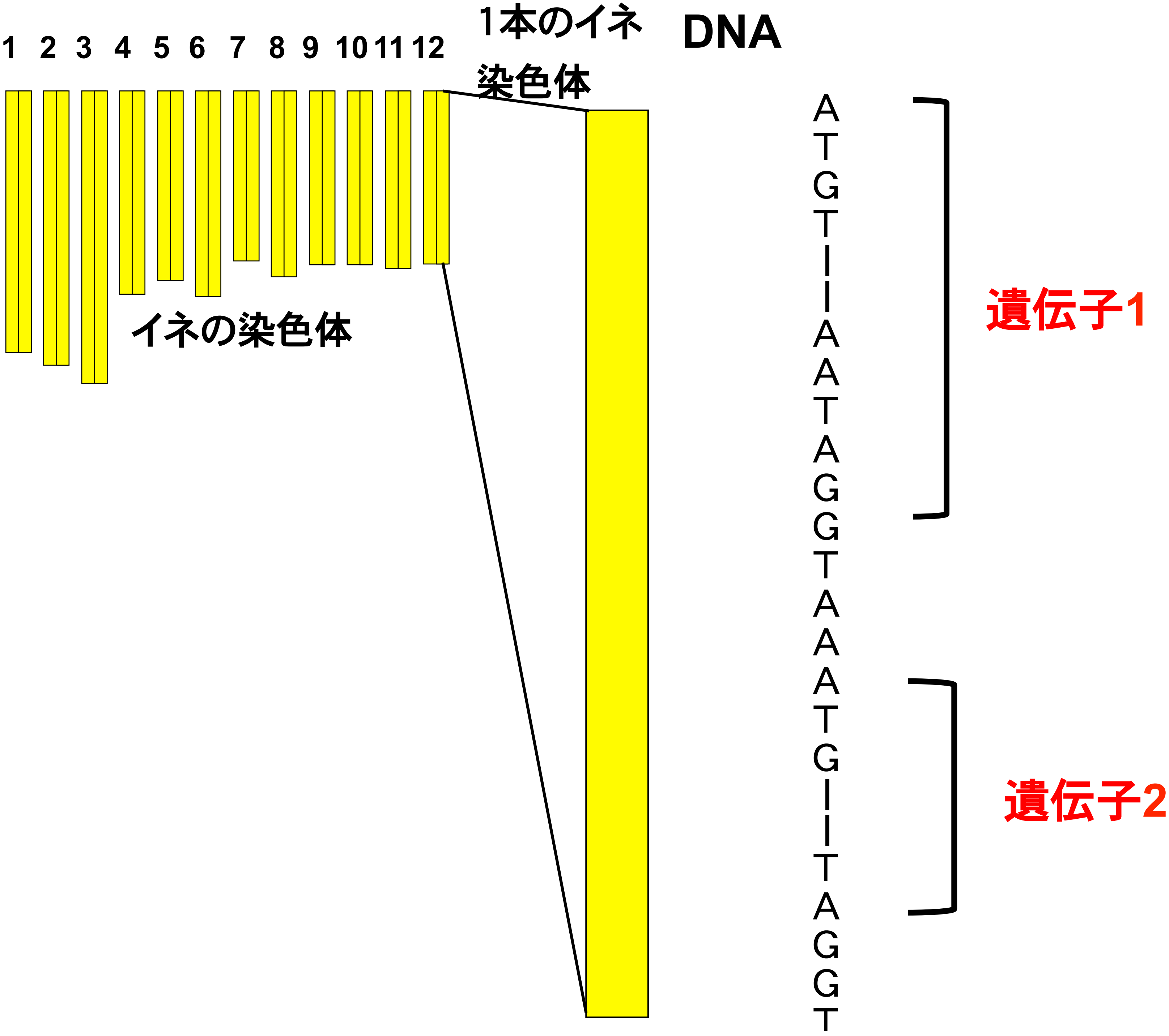


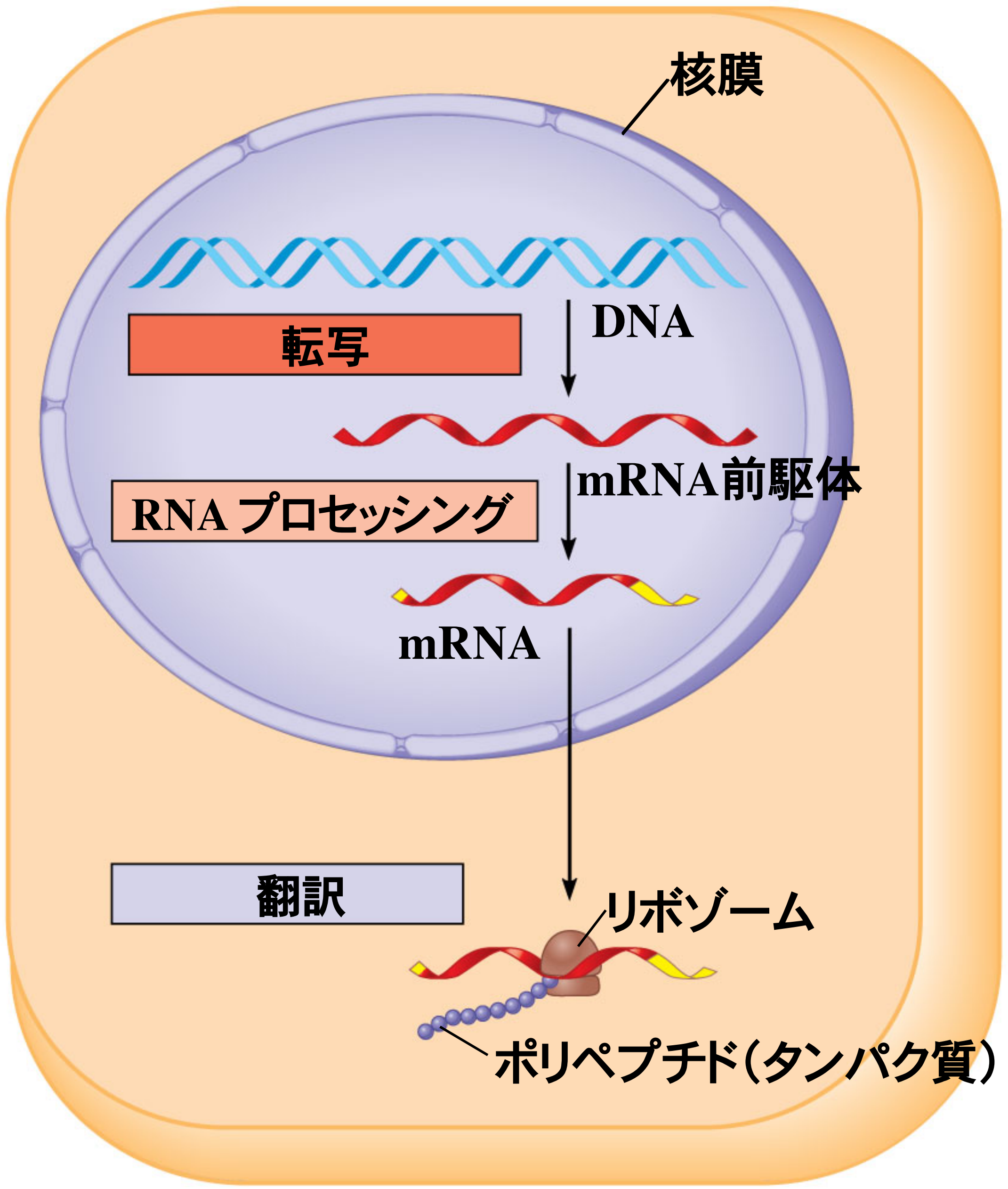
遺伝子12345: お米の収量をあげる遺伝子

遺伝子23456: お米の味を良くする

イネの細胞の中の染色体とDNA



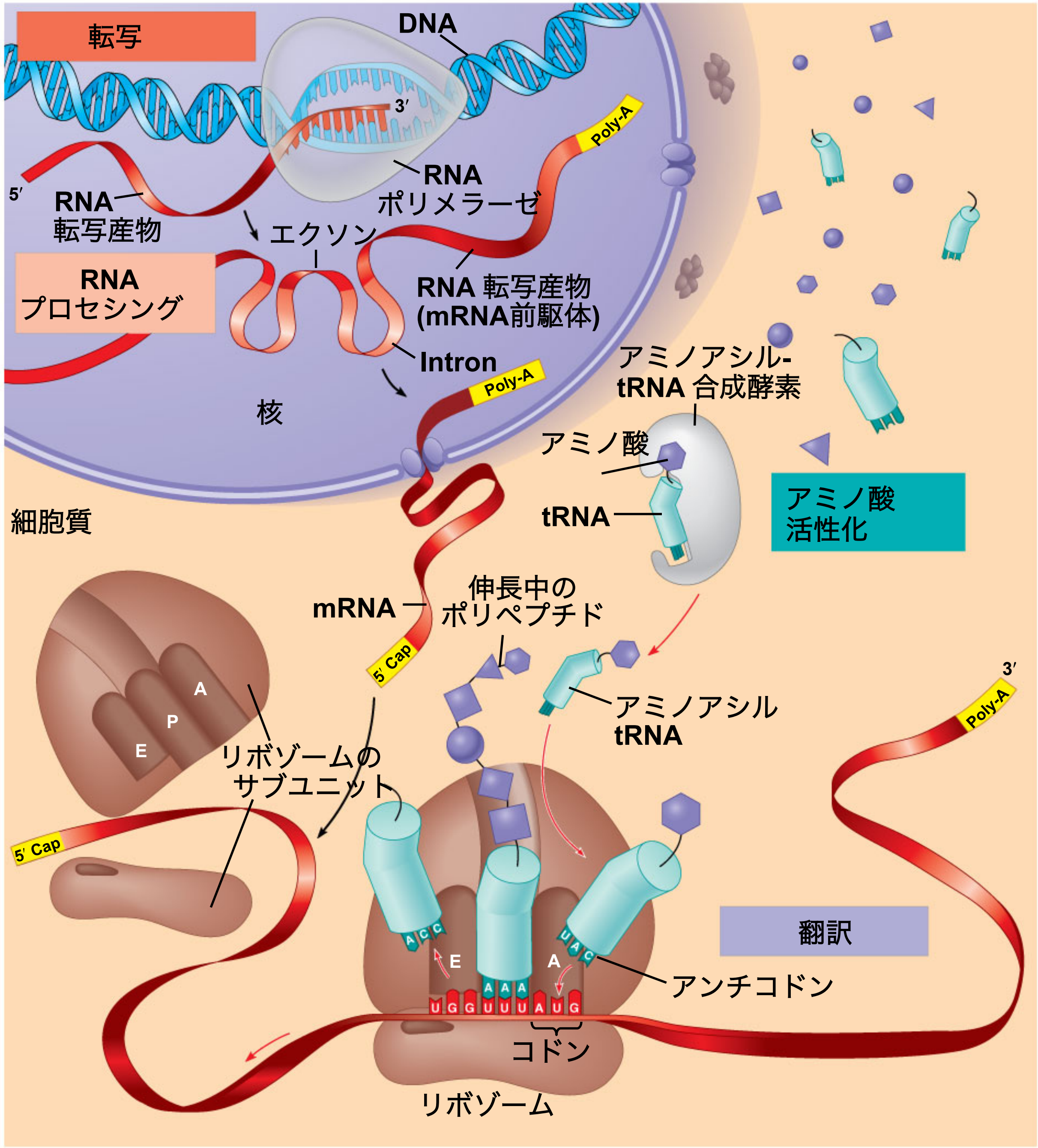

1 2 3 4 5 6 7 8 9 10 11 12
イネの染色体
1本のイネ染色体
DNA
A
T
G
T
I
I
A
A
T
A
G
G
T
A
A
A
T
G
I
I
T
A
G
G
T
遺伝子1
遺伝子2



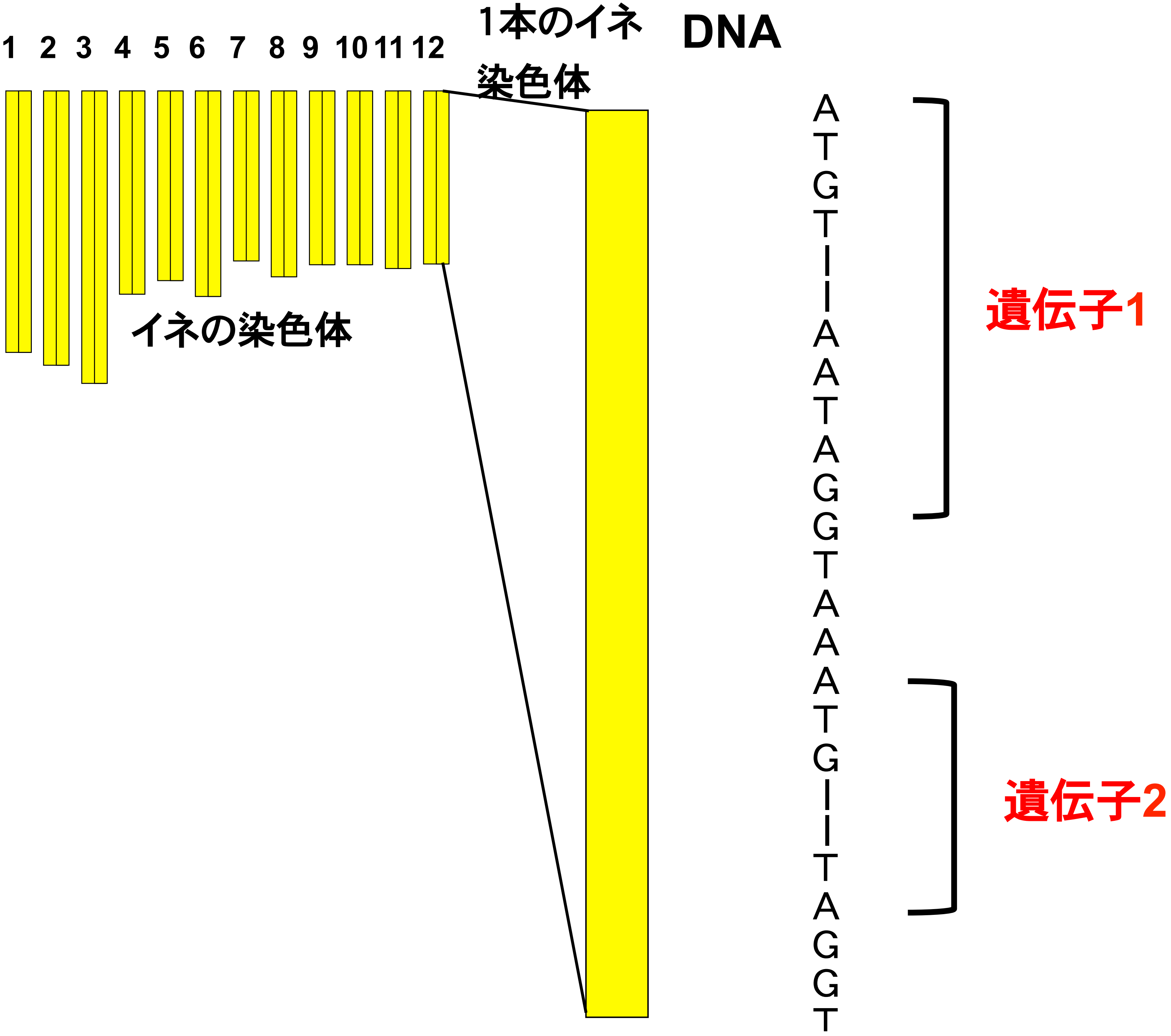
(b) 真核生物の細胞

キャンベル参照

Figure 17



キャンベル参照



1 2 3 4 5 6 7 8 9 10 11 12

イネの染色体

1本のイネ
染色体

DNA

ATGTIIAATAGGTAAATGIITAGGT

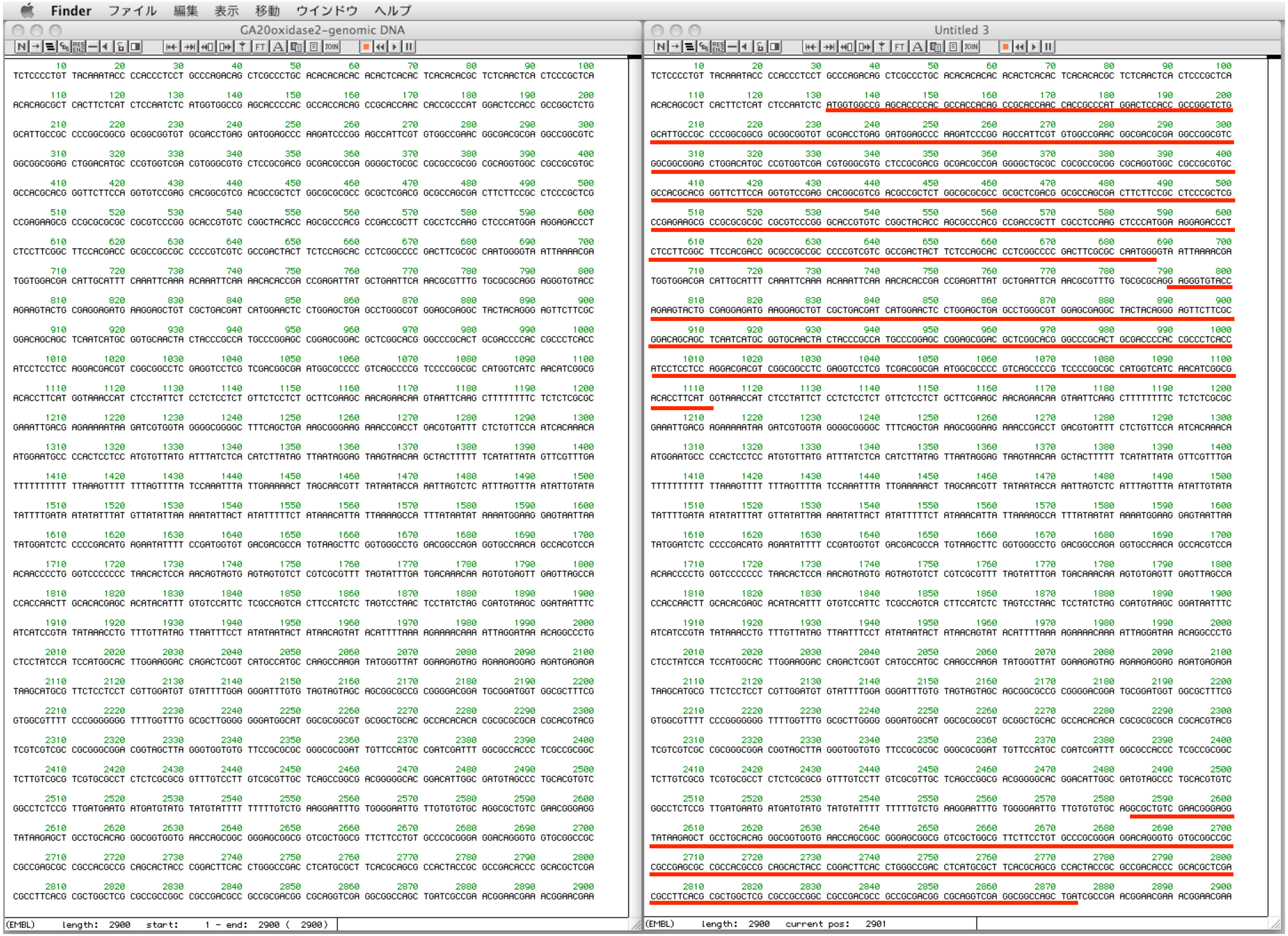
遺伝子1

遺伝子2

```
        10         20         30         40         50         60         70         80         90        100
TCTCCCCTGT TACAAATACC CCACCCTCCT GCCCAGACAG CTCGCCCTGC ACACACACAC ACACTCACAC TCACACACGC TCTCAACTCA CTCCCGCTCA

       110        120        130        140        150        160        170        180        190        200
ACACAGCGCT CACTTCTCAT CTCCAATCTC ATGGTGGCCG AGCACCCCAC GCCACCACAG CCGCACCAAC CACCGCCCAT GGACTCCACC GCCGGCTCTG

       210        220        230        240        250        260        270        280        290        300
GCATTGCCGC CCCGGCGGCG GCGGCGGTGT GCGACCTGAG GATGGAGCCC AAGATCCCGG AGCCATTCGT GTGGCCGAAC GGCGACGCGA GGCCGGCGTC

       310        320        330        340        350        360        370        380        390        400
GGCGGCGGAG CTGGACATGC CCGTGGTCGA CGTGGGCGTG CTCCGCGACG GCGACGCCGA GGGGCTGCGC CGCGCCGCGG CGCAGGTGGC CGCCGCGTGC

       410        420        430        440        450        460        470        480        490        500
GCCACGCACG GGTTCTTCCA GGTGTCCGAG CACGGCGTCG ACGCCGCTCT GGCGCGCGCC GCGCTCGACG GCGCCAGCGA CTTCTTCCGC CTCCCGCTCG

       510        520        530        540        550        560        570        580        590        600
CCGAGAAGCG CCGCGCGCGC CGCGTCCCGG GCACCGTGTC CGGCTACACC AGCGCCCACG CCGACCGCTT CGCCTCCAAG CTCCCATGGA AGGAGACCCT

       610        620        630        640        650        660        670        680        690        700
CTCCTTCGGC TTCCACGACC GCGCCGCCGC CCCCGTCGTC GCCGACTACT TCTCCAGCAC CCTCGGCCCC GACTTCGCGC CAATGGGGTA ATTAAAACGA

       710        720        730        740        750        760        770        780        790        800
TGGTGGACGA CATTGCATTT CAAATTCAAA ACAAATTCAA AACACACCGA CCGAGATTAT GCTGAATTCA AACGCGTTTG TGCGCGCAGG AGGGTGTACC

       810        820        830        840        850        860        870        880        890        900
AGAAGTACTG CGAGGAGATG AAGGAGCTGT CGCTGACGAT CATGGAACTC CTGGAGCTGA GCCTGGGCGT GGAGCGAGGC TACTACAGGG AGTTCTTCGC

       910        920        930        940        950        960        970        980        990       1000
GGACAGCAGC TCAATCATGC GGTGCAACTA CTACCCGCCA TGCCCGGAGC CGGAGCGGAC GCTCGGCACG GGCCCGCACT GCGACCCCAC CGCCCTCACC

      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
ATCCTCCTCC AGGACGACGT CGGCGGCCTC GAGGTCCTCG TCGACGGCGA ATGGCGCCCC GTCAGCCCCG TCCCCGGCGC CATGGTCATC AACATCGGCG

      1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
ACACCTTCAT GGTAAACCAT CTCCTATTCT CCTCTCCTCT GTTCTCCTCT GCTTCGAAGC AACAGAACAA GTAATTCAAG CTTTTTTTTC TCTCTCGCGC

      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300
GAAATTGACG AGAAAAATAA GATCGTGGTA GGGGCGGGGC TTTCAGCTGA AAGCGGGAAG AAACCGACCT GACGTGATTT CTCTGTTCCA ATCACAAACA

      1310       1320       1330       1340       1350       1360       1370       1380       1390       1400
ATGGAATGCC CCACTCCTCC ATGTGTTATG ATTTATCTCA CATCTTATAG TTAATAGGAG TAAGTAACAA GCTACTTTTT TCATATTATA GTTCGTTTGA

      1410       1420       1430       1440       1450       1460       1470       1480       1490       1500
TTTTTTTTTT TTAAAGTTTT TTTAGTTTTA TCCAAATTTA TTGAAAAACT TAGCAACGTT TATAATACCA AATTAGTCTC ATTTAGTTTA ATATTGTATA

      1510       1520       1530       1540       1550       1560       1570       1580       1590       1600
TATTTTGATA ATATATTTAT GTTATATTAA AAATATTACT ATATTTTTCT ATAAACATTA TTAAAAGCCA TTTATAATAT AAAATGGAAG GAGTAATTAA

      1610       1620       1630       1640       1650       1660       1670       1680       1690       1700
TATGGATCTC CCCCGACATG AGAATATTTT CCGATGGTGT GACGACGCCA TGTAAGCTTC GGTGGGCCTG GACGGCCAGA GGTGCCAACA GCCACGTCCA

      1710       1720       1730       1740       1750       1760       1770       1780       1790       1800
ACAACCCCTG GGTCCCCCCC TAACACTCCA AACAGTAGTG AGTAGTGTCT CGTCGCGTTT TAGTATTTGA TGACAAACAA AGTGTGAGTT GAGTTAGCCA

      1810       1820       1830       1840       1850       1860       1870       1880       1890       1900
CCACCAACTT GCACACGAGC ACATACATTT GTGTCCATTC TCGCCAGTCA CTTCCATCTC TAGTCCTAAC TCCTATCTAG CGATGTAAGC GGATAATTTC

      1910       1920       1930       1940       1950       1960       1970       1980       1990       2000
ATCATCCGTA TATAAACCTG TTTGTTATAG TTAATTTCCT ATATAATACT ATAACAGTAT ACATTTTAAA AGAAAACAAA ATTAGGATAA ACAGGCCCTG

      2010       2020       2030       2040       2050       2060       2070       2080       2090       2100
CTCCTATCCA TCCATGGCAC TTGGAAGGAC CAGACTCGGT CATGCCATGC CAAGCCAAGA TATGGGTTAT GGAAGAGTAG AGAAGAGGAG AGATGAGAGA

      2110       2120       2130       2140       2150       2160       2170       2180       2190       2200
TAAGCATGCG TTCTCCTCCT CGTTGGATGT GTATTTTGGA GGGATTTGTG TAGTAGTAGC AGCGGCGCCG CGGGGACGGA TGCGGATGGT GGCGCTTTCG

      2210       2220       2230       2240       2250       2260       2270       2280       2290       2300
GTGGCGTTTT CCCGGGGGGG TTTTGGTTTG GCGCTTGGGG GGGATGGCAT GGCGCGGCGT GCGGCTGCAC GCCACACACA CGCGCGCGCA CGCACGTACG

      2310       2320       2330       2340       2350       2360       2370       2380       2390       2400
TCGTCGTCGC CGCGGGCGGA CGGTAGCTTA GGGTGGTGTG TTCCGCGCGC GGGCGCGGAT TGTTCCATGC CGATCGATTT GGCGCCACCC TCGCCGCGGC

      2410       2420       2430       2440       2450       2460       2470       2480       2490       2500
TCTTGTCGCG TCGTGCGCCT CTCTCGCGCG GTTTGTCCTT GTCGCGTTGC TCAGCCGGCG ACGGGGGCAC GGACATTGGC GATGTAGCCC TGCACGTGTC

      2510       2520       2530       2540       2550       2560       2570       2580       2590       2600
GGCCTCTCCG TTGATGAATG ATGATGTATG TATGTATTTT TTTTTGTCTG AAGGAATTTG TGGGGAATTG TTGTGTGTGC AGGCGCTGTC GAACGGGAGG

      2610       2620       2630       2640       2650       2660       2670       2680       2690       2700
TATAAGAGCT GCCTGCACAG GGCGGTGGTG AACCAGCGGC GGGAGCGGCG GTCGCTGGCG TTCTTCCTGT GCCCGCGGGA GGACAGGGTG GTGCGGCCGC

      2710       2720       2730       2740       2750       2760       2770       2780       2790       2800
CGCCGAGCGC CGCCACGCCG CAGCACTACC CGGACTTCAC CTGGGCCGAC CTCATGCGCT TCACGCAGCG CCACTACCGC GCCGACACCC GCACGCTCGA

      2810       2820       2830       2840       2850       2860       2870       2880       2890       2900
CGCCTTCACG CGCTGGCTCG CGCCGCCGGC CGCCGACGCC GCCGCGACGG CGCAGGTCGA GGCGGCCAGC TGATCGCCGA ACGGAACGAA ACGGAACGAA

(EMBL)     length:  2900    current pos:    697
```

人間の場合 – 32億塩基対
イネの場合 – 4億塩基対

どこからどこまでが遺伝子なのか？



```
        10         20         30         40         50         60         70         80         90        100
TCTCCCCTGT TACAAATACC CCACCCTCCT GCCCAGACAG CTCGCCCTGC ACACACACAC ACACTCACAC TCACACACGC TCTCAACTCA CTCCCGCTCA

       110        120        130        140        150        160        170        180        190        200
ACACAGCGCT CACTTCTCAT CTCCAATCTC ATGGTGGCCG AGCACCCCAC GCCACCACAG CCGCACCAAC CACCGCCCAT GGACTCCACC GCCGGCTCTG

       210        220        230        240        250        260        270        280        290        300
GCATTGCCGC CCCGGCGGCG GCGGCGGTGT GCGACCTGAG GATGGAGCCC AAGATCCCGG AGCCATTCGT GTGGCCGAAC GGCGACGCGA GGCCGGCGTC

       310        320        330        340        350        360        370        380        390        400
GGCGGCGGAG CTGGACATGC CCGTGGTCGA CGTGGGCGTG CTCCGCGACG GCGACGCCGA GGGGCTGCGC CGCGCCGCGG CGCAGGTGGC CGCCGCGTGC

       410        420        430        440        450        460        470        480        490        500
GCCACGCACG GGTTCTTCCA GGTGTCCGAG CACGGCGTCG ACGCCGCTCT GGCGCGCGCC GCGCTCGACG GCGCCAGCGA CTTCTTCCGC CTCCCGCTCG

       510        520        530        540        550        560        570        580        590        600
CCGAGAAGCG CCGCGCGCGC CGCGTCCCGG GCACCGTGTC CGGCTACACC AGCGCCCACG CCGACCGCTT CGCCTCCAAG CTCCCATGGA AGGAGACCCT

       610        620        630        640        650        660        670        680        690        700
CTCCTTCGGC TTCCACGACC GCGCCGCCGC CCCCGTCGTC GCCGACTACT TCTCCAGCAC CCTCGGCCCC GACTTCGCGC CAATGGGGTA ATTAAAACGA

       710        720        730        740        750        760        770        780        790        800
TGGTGGACGA CATTGCATTT CAAATTCAAA ACAAATTCAA AACACACCGA CCGAGATTAT GCTGAATTCA AACGCGTTTG TGCGCGCAGG AGGGTGTACC

       810        820        830        840        850        860        870        880        890        900
AGAAGTACTG CGAGGAGATG AAGGAGCTGT CGCTGACGAT CATGGAACTC CTGGAGCTGA GCCTGGGCGT GGAGCGAGGC TACTACAGGG AGTTCTTCGC

       910        920        930        940        950        960        970        980        990       1000
GGACAGCAGC TCAATCATGC GGTGCAACTA CTACCCGCCA TGCCCGGAGC CGGAGCGGAC GCTCGGCACG GGCCCGCACT GCGACCCCAC CGCCCTCACC

      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
ATCCTCCTCC AGGACGACGT CGGCGGCCTC GAGGTCCTCG TCGACGGCGA ATGGCGCCCC GTCAGCCCCG TCCCCGGCGC CATGGTCATC AACATCGGCG

      1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
ACACCTTCAT GGTAAACCAT CTCCTATTCT CCTCTCCTCT GTTCTCCTCT GCTTCGAAGC AACAGAACAA GTAATTCAAG CTTTTTTTTC TCTCTCGCGC

      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300
GAAATTGACG AGAAAAATAA GATCGTGGTA GGGGCGGGGC TTTCAGCTGA AAGCGGGAAG AAACCGACCT GACGTGATTT CTCTGTTCCA ATCACAAACA

      1310       1320       1330       1340       1350       1360       1370       1380       1390       1400
ATGGAATGCC CCACTCCTCC ATGTGTTATG ATTTATCTCA CATCTTATAG TTAATAGGAG TAAGTAACAA GCTACTTTTT TCATATTATA GTTCGTTTGA

      1410       1420       1430       1440       1450       1460       1470       1480       1490       1500
TTTTTTTTTT TTAAAGTTTT TTTAGTTTTA TCCAAATTTA TTGAAAAACT TAGCAACGTT TATAATACCA AATTAGTCTC ATTTAGTTTA ATATTGTATA

      1510       1520       1530       1540       1550       1560       1570       1580       1590       1600
TATTTTGATA ATATATTTAT GTTATATTAA AAATATTACT ATATTTTTCT ATAAACATTA TTAAAAGCCA TTTATAATAT AAAATGGAAG GAGTAATTAA

      1610       1620       1630       1640       1650       1660       1670       1680       1690       1700
TATGGATCTC CCCCGACATG AGAATATTTT CCGATGGTGT GACGACGCCA TGTAAGCTTC GGTGGGCCTG GACGGCCAGA GGTGCCAACA GCCACGTCCA

      1710       1720       1730       1740       1750       1760       1770       1780       1790       1800
ACAACCCCTG GGTCCCCCCC TAACACTCCA AACAGTAGTG AGTAGTGTCT CGTCGCGTTT TAGTATTTGA TGACAAACAA AGTGTGAGTT GAGTTAGCCA

      1810       1820       1830       1840       1850       1860       1870       1880       1890       1900
CCACCAACTT GCACACGAGC ACATACATTT GTGTCCATTC TCGCCAGTCA CTTCCATCTC TAGTCCTAAC TCCTATCTAG CGATGTAAGC GGATAATTTC

      1910       1920       1930       1940       1950       1960       1970       1980       1990       2000
ATCATCCGTA TATAAACCTG TTTGTTATAG TTAATTTCCT ATATAATACT ATAACAGTAT ACATTTTAAA AGAAAACAAA ATTAGGATAA ACAGGCCCTG

      2010       2020       2030       2040       2050       2060       2070       2080       2090       2100
CTCCTATCCA TCCATGGCAC TTGGAAGGAC CAGACTCGGT CATGCCATGC CAAGCCAAGA TATGGGTTAT GGAAGAGTAG AGAAGAGGAG AGATGAGAGA

      2110       2120       2130       2140       2150       2160       2170       2180       2190       2200
TAAGCATGCG TTCTCCTCCT CGTTGGATGT GTATTTTGGA GGGATTTGTG TAGTAGTAGC AGCGGCGCCG CGGGGACGGA TGCGGATGGT GGCGCTTTCG

      2210       2220       2230       2240       2250       2260       2270       2280       2290       2300
GTGGCGTTTT CCCGGGGGGG TTTTGGTTTG GCGCTTGGGG GGGATGGCAT GGCGCGGCGT GCGGCTGCAC GCCACACACA CGCGCGCGCA CGCACGTACG

      2310       2320       2330       2340       2350       2360       2370       2380       2390       2400
TCGTCGTCGC CGCGGGCGGA CGGTAGCTTA GGGTGGTGTG TTCCGCGCGC GGGCGCGGAT TGTTCCATGC CGATCGATTT GGCGCCACCC TCGCCGCGGC

      2410       2420       2430       2440       2450       2460       2470       2480       2490       2500
TCTTGTCGCG TCGTGCGCCT CTCTCGCGCG GTTTGTCCTT GTCGCGTTGC TCAGCCGGCG ACGGGGGCAC GGACATTGGC GATGTAGCCC TGCACGTGTC

      2510       2520       2530       2540       2550       2560       2570       2580       2590       2600
GGCCTCTCCG TTGATGAATG ATGATGTATG TATGTATTTT TTTTTGTCTG AAGGAATTTG TGGGGAATTG TTGTGTGTGC AGGCGCTGTC GAACGGGAGG

      2610       2620       2630       2640       2650       2660       2670       2680       2690       2700
TATAAGAGCT GCCTGCACAG GGCGGTGGTG AACCAGCGGC GGGAGCGGCG GTCGCTGGCG TTCTTCCTGT GCCCGCGGGA GGACAGGGTG GTGCGGCCGC

      2710       2720       2730       2740       2750       2760       2770       2780       2790       2800
CGCCGAGCGC CGCCACGCCG CAGCACTACC CGGACTTCAC CTGGGCCGAC CTCATGCGCT TCACGCAGCG CCACTACCGC GCCGACACCC GCACGCTCGA

      2810       2820       2830       2840       2850       2860       2870       2880       2890       2900
CGCCTTCACG CGCTGGCTCG CGCCGCCGGC CGCCGACGCC GCCGCGACGG CGCAGGTCGA GGCGGCCAGC TGATCGCCGA ACGGAACGAA ACGGAACGAA

(EMBL)      length:  2900    start:      1 - end:  2900 (  2900)
```

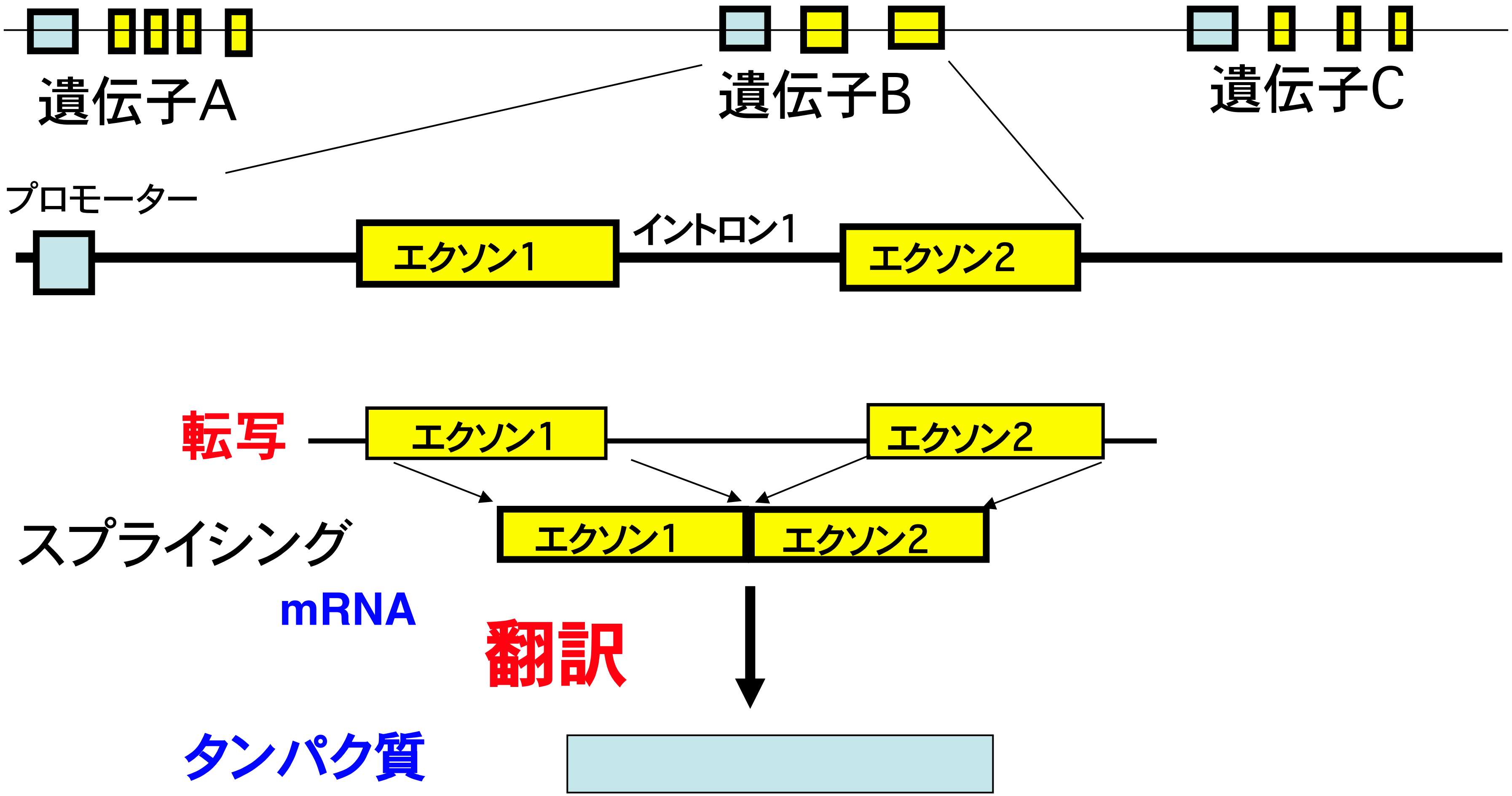
```
        10         20         30         40         50         60         70         80         90        100
TCTCCCCTGT TACAAATACC CCACCCTCCT GCCCAGACAG CTCGCCCTGC ACACACACAC ACACTCACAC TCACACACGC TCTCAACTCA CTCCCGCTCA

       110        120        130        140        150        160        170        180        190        200
ACACAGCGCT CACTTCTCAT CTCCAATCTC ATGGTGGCCG AGCACCCCAC GCCACCACAG CCGCACCAAC CACCGCCCAT GGACTCCACC GCCGGCTCTG

       210        220        230        240        250        260        270        280        290        300
GCATTGCCGC CCCGGCGGCG GCGGCGGTGT GCGACCTGAG GATGGAGCCC AAGATCCCGG AGCCATTCGT GTGGCCGAAC GGCGACGCGA GGCCGGCGTC

       310        320        330        340        350        360        370        380        390        400
GGCGGCGGAG CTGGACATGC CCGTGGTCGA CGTGGGCGTG CTCCGCGACG GCGACGCCGA GGGGCTGCGC CGCGCCGCGG CGCAGGTGGC CGCCGCGTGC

       410        420        430        440        450        460        470        480        490        500
GCCACGCACG GGTTCTTCCA GGTGTCCGAG CACGGCGTCG ACGCCGCTCT GGCGCGCGCC GCGCTCGACG GCGCCAGCGA CTTCTTCCGC CTCCCGCTCG

       510        520        530        540        550        560        570        580        590        600
CCGAGAAGCG CCGCGCGCGC CGCGTCCCGG GCACCGTGTC CGGCTACACC AGCGCCCACG CCGACCGCTT CGCCTCCAAG CTCCCATGGA AGGAGACCCT

       610        620        630        640        650        660        670        680        690        700
CTCCTTCGGC TTCCACGACC GCGCCGCCGC CCCCGTCGTC GCCGACTACT TCTCCAGCAC CCTCGGCCCC GACTTCGCGC CAATGGGGTA ATTAAAACGA

       710        720        730        740        750        760        770        780        790        800
TGGTGGACGA CATTGCATTT CAAATTCAAA ACAAATTCAA AACACACCGA CCGAGATTAT GCTGAATTCA AACGCGTTTG TGCGCGCAGG AGGGTGTACC

       810        820        830        840        850        860        870        880        890        900
AGAAGTACTG CGAGGAGATG AAGGAGCTGT CGCTGACGAT CATGGAACTC CTGGAGCTGA GCCTGGGCGT GGAGCGAGGC TACTACAGGG AGTTCTTCGC

       910        920        930        940        950        960        970        980        990       1000
GGACAGCAGC TCAATCATGC GGTGCAACTA CTACCCGCCA TGCCCGGAGC CGGAGCGGAC GCTCGGCACG GGCCCGCACT GCGACCCCAC CGCCCTCACC

      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
ATCCTCCTCC AGGACGACGT CGGCGGCCTC GAGGTCCTCG TCGACGGCGA ATGGCGCCCC GTCAGCCCCG TCCCCGGCGC CATGGTCATC AACATCGGCG

      1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
ACACCTTCAT GGTAAACCAT CTCCTATTCT CCTCTCCTCT GTTCTCCTCT GCTTCGAAGC AACAGAACAA GTAATTCAAG CTTTTTTTTC TCTCTCGCGC

      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300
GAAATTGACG AGAAAAATAA GATCGTGGTA GGGGCGGGGC TTTCAGCTGA AAGCGGGAAG AAACCGACCT GACGTGATTT CTCTGTTCCA ATCACAAACA

      1310       1320       1330       1340       1350       1360       1370       1380       1390       1400
ATGGAATGCC CCACTCCTCC ATGTGTTATG ATTTATCTCA CATCTTATAG TTAATAGGAG TAAGTAACAA GCTACTTTTT TCATATTATA GTTCGTTTGA

      1410       1420       1430       1440       1450       1460       1470       1480       1490       1500
TTTTTTTTTT TTAAAGTTTT TTTAGTTTTA TCCAAATTTA TTGAAAAACT TAGCAACGTT TATAATACCA AATTAGTCTC ATTTAGTTTA ATATTGTATA

      1510       1520       1530       1540       1550       1560       1570       1580       1590       1600
TATTTTGATA ATATATTTAT GTTATATTAA AAATATTACT ATATTTTTCT ATAAACATTA TTAAAAGCCA TTTATAATAT AAAATGGAAG GAGTAATTAA

      1610       1620       1630       1640       1650       1660       1670       1680       1690       1700
TATGGATCTC CCCCGACATG AGAATATTTT CCGATGGTGT GACGACGCCA TGTAAGCTTC GGTGGGCCTG GACGGCCAGA GGTGCCAACA GCCACGTCCA

      1710       1720       1730       1740       1750       1760       1770       1780       1790       1800
ACAACCCCTG GGTCCCCCCC TAACACTCCA AACAGTAGTG AGTAGTGTCT CGTCGCGTTT TAGTATTTGA TGACAAACAA AGTGTGAGTT GAGTTAGCCA

      1810       1820       1830       1840       1850       1860       1870       1880       1890       1900
CCACCAACTT GCACACGAGC ACATACATTT GTGTCCATTC TCGCCAGTCA CTTCCATCTC TAGTCCTAAC TCCTATCTAG CGATGTAAGC GGATAATTTC

      1910       1920       1930       1940       1950       1960       1970       1980       1990       2000
ATCATCCGTA TATAAACCTG TTTGTTATAG TTAATTTCCT ATATAATACT ATAACAGTAT ACATTTTAAA AGAAAACAAA ATTAGGATAA ACAGGCCCTG

      2010       2020       2030       2040       2050       2060       2070       2080       2090       2100
CTCCTATCCA TCCATGGCAC TTGGAAGGAC CAGACTCGGT CATGCCATGC CAAGCCAAGA TATGGGTTAT GGAAGAGTAG AGAAGAGGAG AGATGAGAGA

      2110       2120       2130       2140       2150       2160       2170       2180       2190       2200
TAAGCATGCG TTCTCCTCCT CGTTGGATGT GTATTTTGGA GGGATTTGTG TAGTAGTAGC AGCGGCGCCG CGGGGACGGA TGCGGATGGT GGCGCTTTCG

      2210       2220       2230       2240       2250       2260       2270       2280       2290       2300
GTGGCGTTTT CCCGGGGGGG TTTTGGTTTG GCGCTTGGGG GGGATGGCAT GGCGCGGCGT GCGGCTGCAC GCCACACACA CGCGCGCGCA CGCACGTACG

      2310       2320       2330       2340       2350       2360       2370       2380       2390       2400
TCGTCGTCGC CGCGGGCGGA CGGTAGCTTA GGGTGGTGTG TTCCGCGCGC GGGCGCGGAT TGTTCCATGC CGATCGATTT GGCGCCACCC TCGCCGCGGC

      2410       2420       2430       2440       2450       2460       2470       2480       2490       2500
TCTTGTCGCG TCGTGCGCCT CTCTCGCGCG GTTTGTCCTT GTCGCGTTGC TCAGCCGGCG ACGGGGGCAC GGACATTGGC GATGTAGCCC TGCACGTGTC

      2510       2520       2530       2540       2550       2560       2570       2580       2590       2600
GGCCTCTCCG TTGATGAATG ATGATGTATG TATGTATTTT TTTTTGTCTG AAGGAATTTG TGGGGAATTG TTGTGTGTGC AGGCGCTGTC GAACGGGAGG

      2610       2620       2630       2640       2650       2660       2670       2680       2690       2700
TATAAGAGCT GCCTGCACAG GGCGGTGGTG AACCAGCGGC GGGAGCGGCG GTCGCTGGCG TTCTTCCTGT GCCCGCGGGA GGACAGGGTG GTGCGGCCGC

      2710       2720       2730       2740       2750       2760       2770       2780       2790       2800
CGCCGAGCGC CGCCACGCCG CAGCACTACC CGGACTTCAC CTGGGCCGAC CTCATGCGCT TCACGCAGCG CCACTACCGC GCCGACACCC GCACGCTCGA

      2810       2820       2830       2840       2850       2860       2870       2880       2890       2900
CGCCTTCACG CGCTGGCTCG CGCCGCCGGC CGCCGACGCC GCCGCGACGG CGCAGGTCGA GGCGGCCAGC TGATCGCCGA ACGGAACGAA ACGGAACGAA

(EMBL)      length:  2900    current pos:    2901
```

真核生物ではゲノムDNAの一部が遺伝子をコードしている

DNA

------TGGCATATGGCATGCAGACTGGCATTGGCATGCGTCTGCATGCCATTGGTTTCTGAGGTCGTGTGAAACTTATTTGGGCACAGACATGGCATGCAGACTGGCAC-----

------ACCGTATACCGTACGTCTGACCGTAACCGTACGCAGACGTACGGTAACCAAAGAGTCCAGCACACTTTGAATAAACCCGTGTCTGTACCGTACGTCTGACCGTG-----



遺伝子A

遺伝子B

遺伝子C

プロモーター

エクソン1

イントロン1

エクソン2

転写

エクソン1

エクソン2

スプライシング

エクソン1

エクソン2

mRNA

翻訳

タンパク質

エキソンとイントロンの境界

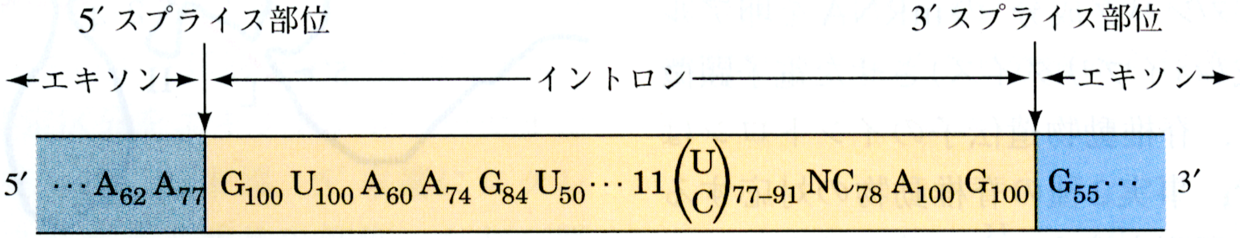
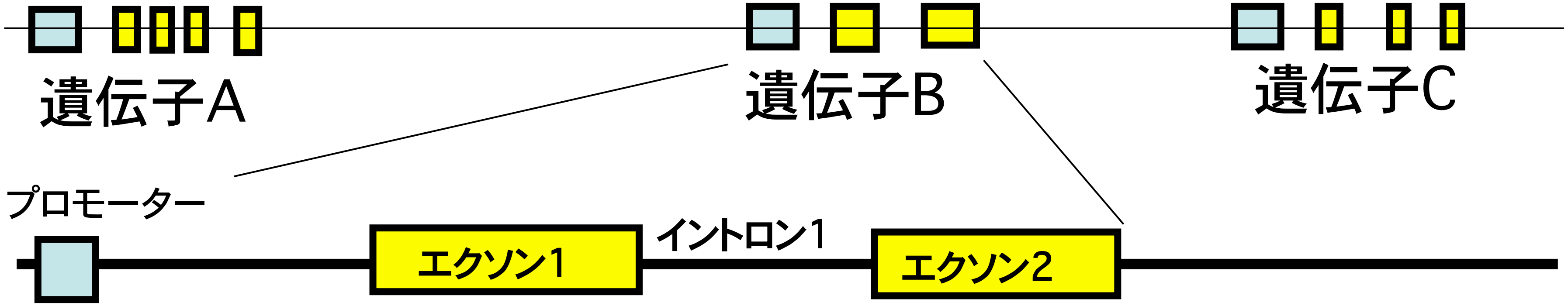


図 25・19　真核細胞 mRNA 前駆体のエキソン・イントロン接合部にあるコンセンサス配列　各塩基の添え数字はその塩基の出現頻度(%)．3′ スプライス部位のすぐ上流にピリミジンに富む 11 nt 区域がある［R.A. Padgett, P.J. Grabowski, M.M. Konarska, S.S. Seiler, P.A. Sharp, *Annu. Rev. Biochem.* **55**, 1123 (1986) のデータに基づく］

DNAからRNAへそしてタンパク質へ



コード鎖

DNA 5′--GTCTGC**ATGCCATTGGTTTCTGAGGTCGTGTGA**AACAATAAAGGGCA

DNA 3′--CAGACG**TACGGTAACCAAAGACTCCAGCACACT**TTGTTATTTCCCGT

鋳型鎖

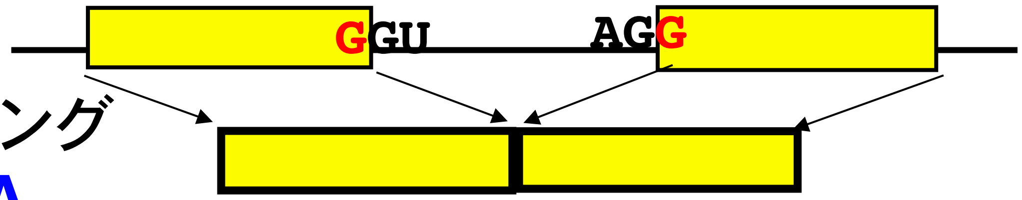
PolyA認識サイト

RNAポリメラーゼ

mRNA前駆体

CAP GTCUGC**AUGCCAUUGGUUUCUGAGGUCGUGUGA**AACAAUAAAGGGCAAAAAA

転写



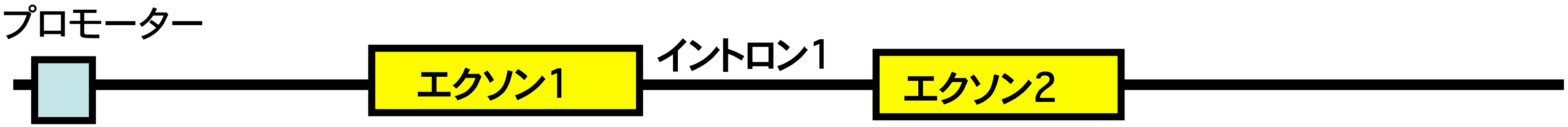
スプライシング

成熟mRNA

成熟mRNA CAPGTCUGC|**AUG**|**CCA**|**UUG**|**GUC**|**GUG**|**UGA**|AACUUAUUUGGGCAAAAAAAA

Met Pro Leu Val Val Stop

DNAとcDNA（complementary DNA）



コード鎖

DNA 5′--GTCTGC**ATGCCATTGGTTTCTGAGGTCGTGTGA**AACAATAAAGGGCA
DNA 3′--CAGACG**TACGGTAACCAAAGACTCCAGCACACT**TTGTTATTTCCCGT

鋳型鎖

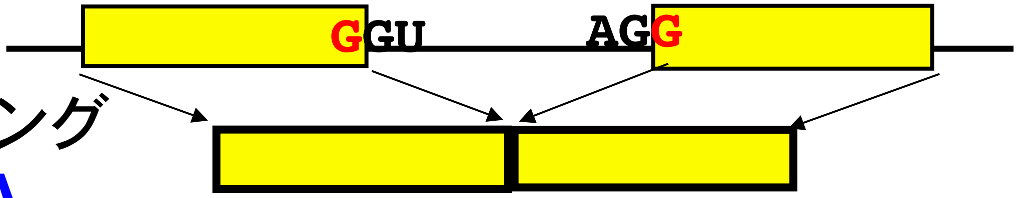
PolyA認識サイト

RNAポリメラーゼ

mRNA前駆体

CAP GTCUGC**AUGCCAUUGGUUUCUGAGGUCGUGUGA**AACAAUAAAGGGCAAAAAA

転写



スプライシング

成熟mRNA

成熟mRNA CAPGTCUGC|**AUG**|**CCA**|**UUG**|**GUC**|**GUG**|**UGA**|AACUUAUUUGGGCAAAAAAAA

Met Pro Leu Val Val Stop

逆転写酵素

cDNA

GTCTGC**ATGCCATTGGTCGTGTGA**AACAATAAAGGGCAAAAAAAA
CAGACG**TACGGTAACCAGCACACT**TTGTTATTTCCCGTTTTTTTT

DNA配列のどこが遺伝子なのか？ORFを検出してみる

(1)　CDS or cDNA（EST）情報が有るとき

EST (expressed sequence tag)とは、ある組織において発現しているタンパク質のmRNA 配列に由来するcDNA配列を部分的に決定したもののことを指す

DNA配列とCDS or cDNA配列を比較する

(2) CDS or cDNA（EST）情報が無いとき

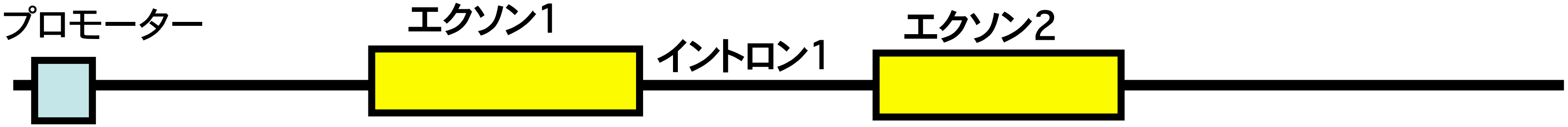
遺伝子予測ソフトで遺伝子領域を推定する

実際に遺伝子を探す

・マニュアルで遺伝子がどこにあるのか調べてみる。

cDNA情報が有るとき.1

DNA配列とcDNA配列を比較する



DNA ---GTCTGCATGCCATTGGTTTCTGAGGTCGTGTGAAACAATAAAGGGCA

DNA ---CAGACGTACGGTAACCAAAGACTCCAGCACACTTTGTTATTTCCCGT

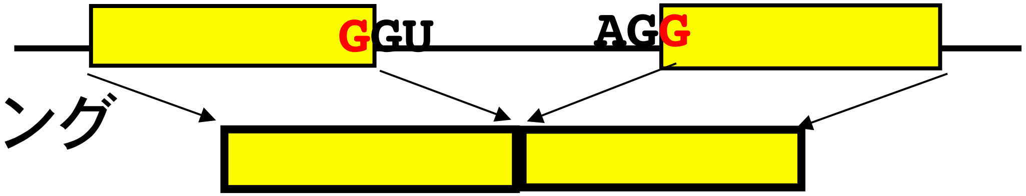
PolyA認識サイト

RNA

CAP GTCUGCAUGCCAUUGGUUUCUGAGGUCGUGUGAAACAAUAAAGGGCAAAAA

mRNA転写

スプライシング

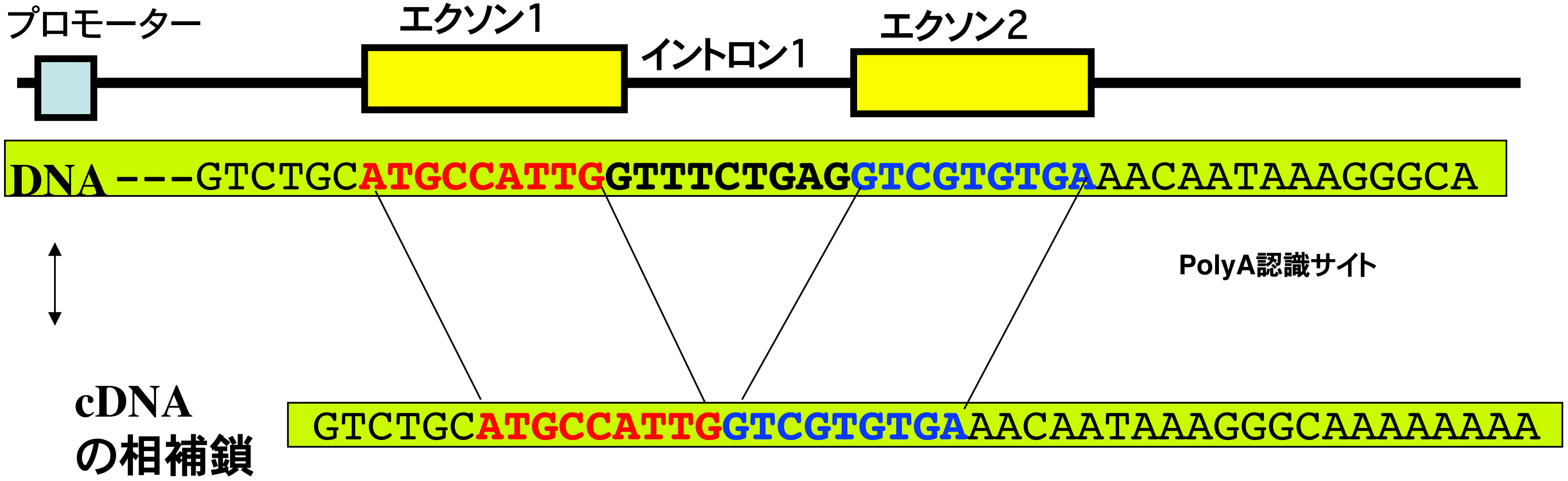


成熟mRNA CAPGTCUGCAUGCCAUUGGUCGUGUGAAACAAUAAAGGGCAAAAAAAA

cDNA CAGACGTACGGTAACCAGCACACTTTGTTATTTCCCGTTTTTTTTT

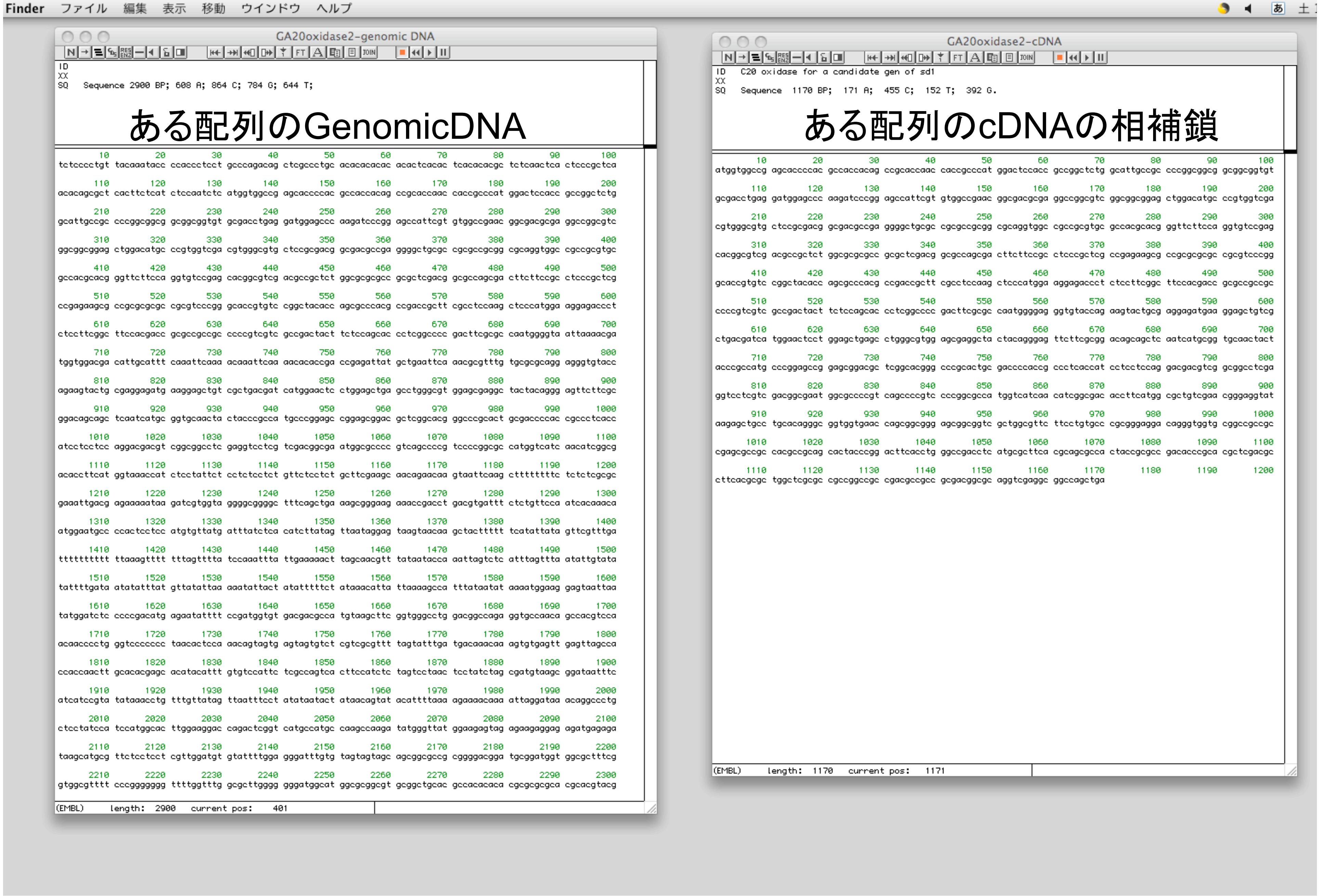
cDNAの相補鎖 GTCTGCATGCCATTGGTCGTGTGAAACAATAAAGGGCAAAAAAAA

cDNA情報が有るとき.2
DNA配列とcDNA配列（の相補鎖）を比較する



DNA ---GTCTGC**ATGCCATTGGTTTCTGAGGTCGTGTGA**AACAATAAAGGGCA

cDNA の相補鎖 GTCTGC**ATGCCATTGGTCGTGTGA**AACAATAAAGGGCAAAAAAAA



GA20oxidase2-genomic DNA

```
ID
XX
SQ   Sequence 2900 BP; 608 A; 864 C; 784 G; 644 T;
```

ある配列のGenomicDNA

```
        10         20         30         40         50         60         70         80         90        100
tctcccctgt tacaaatacc ccaccctcct gcccagacag ctcgccctgc acacacacac acactcacac tcacacacgc tctcaactca ctcccgctca

       110        120        130        140        150        160        170        180        190        200
acacagcgct cacttctcat ctccaatctc atggtggccg agcaccccac gccaccacag ccgcaccaac caccgcccat ggactccacc gccggctctg

       210        220        230        240        250        260        270        280        290        300
gcattgccgc cccggcggcg gcggcggtgt gcgacctgag gatggagccc aagatcccgg agccattcgt gtggccgaac ggcgacgcga ggccggcgtc

       310        320        330        340        350        360        370        380        390        400
ggcggcggag ctggacatgc ccgtggtcga cgtgggcgtg ctccgcgacg gcgacgccga ggggctgcgc cgcgccgcgg cgcaggtggc cgccgcgtgc

       410        420        430        440        450        460        470        480        490        500
gccacgcacg ggttcttcca ggtgtccgag cacggcgtcg acgccgctct ggcgcgcgcc gcgctcgacg gcgccagcga cttcttccgc ctcccgctcg

       510        520        530        540        550        560        570        580        590        600
ccgagaagcg ccgcgcgcgc cgcgtcccgg gcaccgtgtc cggctacacc agcgcccacg ccgaccgctt cgcctccaag ctcccatgga aggagaccct

       610        620        630        640        650        660        670        680        690        700
ctccttcggc ttccacgacc gcgccgccgc ccccgtcgtc gccgactact tctccagcac cctcggcccc gacttcgcgc caatggggta attaaaacga

       710        720        730        740        750        760        770        780        790        800
tggtggacga cattgcattt caaattcaaa acaaattcaa aacacaccga ccgagattat gctgaattca aacgcgtttg tgcgcgcagg agggtgtacc

       810        820        830        840        850        860        870        880        890        900
agaagtactg cgaggagatg aaggagctgt cgctgacgat catggaactc ctggagctga gcctgggcgt ggagcgaggc tactacaggg agttcttcgc

       910        920        930        940        950        960        970        980        990       1000
ggacagcagc tcaatcatgc ggtgcaacta ctacccgcca tgcccggagc cggagcggac gctcggcacg ggcccgcact gcgaccccac cgccctcacc

      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
atcctcctcc aggacgacgt cggcggcctc gaggtcctcg tcgacggcga atggcgcccc gtcagccccg tccccggcgc catggtcatc aacatcggcg

      1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
acaccttcat ggtaaaccat ctcctattct cctctcctct gttctcctct gcttcgaagc aacagaacaa gtaattcaag cttttttttc tctctcgcgc

      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300
gaaattgacg agaaaaataa gatcgtggta ggggcggggc tttcagctga aagcgggaag aaaccgacct gacgtgattt ctctgttcca atcacaaaca

      1310       1320       1330       1340       1350       1360       1370       1380       1390       1400
atggaatgcc ccactcctcc atgtgttatg atttatctca catcttatag ttaataggag taagtaacaa gctacttttt tcatattata gttcgtttga

      1410       1420       1430       1440       1450       1460       1470       1480       1490       1500
tttttttttt ttaaagtttt tttagtttta tccaaattta ttgaaaaact tagcaacgtt tataatacca aattagtctc atttagttta atattgtata

      1510       1520       1530       1540       1550       1560       1570       1580       1590       1600
tattttgata atatatttat gttatattaa aaatattact atatttttct ataaacatta ttaaaagcca tttataatat aaaatggaag gagtaattaa

      1610       1620       1630       1640       1650       1660       1670       1680       1690       1700
tatggatctc ccccgacatg agaatatttt ccgatggtgt gacgacgcca tgtaagcttc ggtgggcctg gacggccaga ggtgccaaca gccacgtcca

      1710       1720       1730       1740       1750       1760       1770       1780       1790       1800
acaacccctg ggtccccccc taacactcca aacagtagtg agtagtgtct cgtcgcgttt tagtatttga tgacaaacaa agtgtgagtt gagttagcca

      1810       1820       1830       1840       1850       1860       1870       1880       1890       1900
ccaccaactt gcacacgagc acatacattt gtgtccattc tcgccagtca cttccatctc tagtcctaac tcctatctag cgatgtaagc ggataatttc

      1910       1920       1930       1940       1950       1960       1970       1980       1990       2000
atcatccgta tataaacctg tttgttatag ttaatttcct atataatact ataacagtat acattttaaa agaaaacaaa attaggataa acaggccctg

      2010       2020       2030       2040       2050       2060       2070       2080       2090       2100
ctcctatcca tccatggcac ttggaaggac cagactcggt catgccatgc caagccaaga tatgggttat ggaagagtag agaagaggag agatgagaga

      2110       2120       2130       2140       2150       2160       2170       2180       2190       2200
taagcatgcg ttctcctcct cgttggatgt gtattttgga gggatttgtg tagtagtagc agcggcgccg cggggacgga tgcggatggt ggcgctttcg

      2210       2220       2230       2240       2250       2260       2270       2280       2290       2300
gtggcgtttt cccggggggg ttttggtttg gcgcttgggg gggatggcat ggcgcggcgt gcggctgcac gccacacaca cgcgcgcgca cgcacgtacg
```

```
(EMBL)    length:  2900   current pos:   401
```

GA20oxidase2-cDNA

```
ID   C20 oxidase for a candidate gen of sd1
XX
SQ   Sequence  1170 BP;  171 A;  455 C;  152 T;  392 G.
```

ある配列のcDNAの相補鎖

```
        10         20         30         40         50         60         70         80         90        100
atggtggccg agcaccccac gccaccacag ccgcaccaac caccgcccat ggactccacc gccggctctg gcattgccgc cccggcggcg gcggcggtgt

       110        120        130        140        150        160        170        180        190        200
gcgacctgag gatggagccc aagatcccgg agccattcgt gtggccgaac ggcgacgcga ggccggcgtc ggcggcggag ctggacatgc ccgtggtcga

       210        220        230        240        250        260        270        280        290        300
cgtgggcgtg ctccgcgacg gcgacgccga ggggctgcgc cgcgccgcgg cgcaggtggc cgccgcgtgc gccacgcacg ggttcttcca ggtgtccgag

       310        320        330        340        350        360        370        380        390        400
cacggcgtcg acgccgctct ggcgcgcgcc gcgctcgacg gcgccagcga cttcttccgc ctcccgctcg ccgagaagcg ccgcgcgcgc cgcgtcccgg

       410        420        430        440        450        460        470        480        490        500
gcaccgtgtc cggctacacc agcgcccacg ccgaccgctt cgcctccaag ctcccatgga aggagaccct ctccttcggc ttccacgacc gcgccgccgc

       510        520        530        540        550        560        570        580        590        600
ccccgtcgtc gccgactact tctccagcac cctcggcccc gacttcgcgc caatggggag ggtgtaccag aagtactgcg aggagatgaa ggagctgtcg

       610        620        630        640        650        660        670        680        690        700
ctgacgatca tggaactcct ggagctgagc ctgggcgtgg agcgaggcta ctacagggag ttcttcgcgg acagcagctc aatcatgcgg tgcaactact

       710        720        730        740        750        760        770        780        790        800
acccgccatg cccggagccg gagcggacgc tcggcacggg cccgcactgc gaccccaccg ccctcaccat cctcctccag gacgacgtcg gcggcctcga

       810        820        830        840        850        860        870        880        890        900
ggtcctcgtc gacggcgaat ggcgccccgt cagccccgtc cccggcgcca tggtcatcaa catcggcgac accttcatgg cgctgtcgaa cgggaggtat

       910        920        930        940        950        960        970        980        990       1000
aagagctgcc tgcacagggc ggtggtgaac cagcggcggg agcggcggtc gctggcgttc ttcctgtgcc cgcgggagga caggtggtg  cggccgccgc

      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
cgagcgccgc cacgccgcag cactacccgg acttcacctg ggccgacctc atgcgcttca cgcagcgcca ctaccgcgcc gacacccgca cgctcgacgc

      1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
cttcacgcgc tggctcgcgc cgccggccgc cgacgccgcc gcgacggcgc aggtcgaggc ggccagctga
```

```
(EMBL)    length:  1170    current pos:    1171
```

ある配列のGenomicDNA

```
ID
XX
SQ   Sequence 2900 BP; 608 A; 864 C; 784 G; 644 T;

        10         20         30         40         50         60         70         80         90        100
tctcccctgt tacaaatacc ccaccctcct gcccagacag ctcgccctgc acacacacac acactcacac tcacacacgc tctcaactca ctcccgctca
       110        120        130        140        150        160        170        180        190        200
acacagcgct cacttctcat ctccaatctc atggtggccg agcaccccac gccaccacag ccgcaccaac caccgcccat ggactccacc gccggctctg
       210        220        230        240        250        260        270        280        290        300
gcattgccgc cccggcggcg gcggcggtgt gcgacctgag gatggagccc aagatcccgg agccattcgt gtggccgaac ggcgacgcga ggccggcgtc
       310        320        330        340        350        360        370        380        390        400
ggcggcggag ctggacatgc ccgtggtcga cgtgggcgtg ctccgcgacg gcgacgccga ggggctgcgc cgcgccgcgg cgcaggtggc cgccgcgtgc
       410        420        430        440        450        460        470        480        490        500
gccacgcacg ggttcttcca ggtgtccgag cacggcgtcg acgccgctct ggcgcgcgcc gcgctcgacg gcgccagcga cttcttccgc ctcccgctcg
       510        520        530        540        550        560        570        580        590        600
ccgagaagcg ccgcgcgcgc cgcgtcccgg gcaccgtgtc cggctacacc agcgcccacg ccgaccgctt cgcctccaag ctcccatgga aggagaccct
       610        620        630        640        650        660        670        680        690        700
ctccttcggc ttccacgacc gcgccgccgc ccccgtcgtc gccgactact tctccagcac cctcggcccc gacttcgcgc caatggggta attaaaacga
```

DNA ---GTCTGC**ATGCCATTG**GTTTCTGAG**GTCGTGTGA**AACAATAAAGGCA

```
       910        920        930        940        950        960        970        980        990       1000
ggacagcagc tcaatcatgc ggtgcaacta ctacccgcca tgcccggagc cggagcggac gctcggcacg ggcccgcact gcgaccccac cgccctcacc
      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
atcctcctcc aggacgacgt cggcggcctc gaggtcctcg tcgacggcga atggcgcccc gtcagccccg tccccggcgc catggtcatc aacatcggcg
      1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
acaccttcat ggtaaaccat ctcctattct cctctcctct gttctcctct gcttcgaagc aacagaacaa gtaattcaag cttttttttc tctctcgcgc
      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300
gaaattgacg agaaaaataa gatcgtggta ggggcggggc tttcagctga aagcgggaag aaaccgacct gacgtgattt ctctgttcaa atcacaaaca
      1310       1320       1330       1340       1350       1360       1370       1380       1390       1400
atggaatgcc ccactcctcc atgtgttatg atttatctca catcttatag ttaataggag taagtaacaa gctacttttt tcatattata gttcgtttga
      14
tttttttt                        tccc
      15
tatttgc                         aaat
      1610       1620       1630       1640       1650       1660       1670       1680       1690       1700
tatggatctc ccccgacatg agaatatttt ccgatggtgt gacgacgcca tgtaagcttc ggtgggcctg gacggccaga ggtgccaaca gccacgtcca
      1710       1720       1730       1740       1750       1760       1770       1780       1790       1800
acaacccctg ggtccccccc taacactcca aacagtagtg agtagtgtct cgtcgcgttt tagtatttga tgacaaacaa agtgtgagtt gagttagcca
      1810       1820       1830       1840       1850       1860       1870       1880       1890       1900
ccaccaactt gcacacgagc acatacattt gtgtccattc tcgccagtca cttccatctc tagtcctaac tcctatctag cgatgtaagc ggataatttc
      1910       1920       1930       1940       1950       1960       1970       1980       1990       2000
atcatccgta tataaacctg tttgttatag ttaatttcct atataatact ataacagtat acattttaaa agaaaacaaa attaggataa acaggccctg
      2010       2020       2030       2040       2050       2060       2070       2080       2090       2100
ctcctatcca tccatggcac ttggaaggac cagactcggt catgccatgc caagccaaga tatgggttat ggaagagtag agaagaggag agatgagaga
      2110       2120       2130       2140       2150       2160       2170       2180       2190       2200
taagcatgcg ttctcctcct cgttggatgt gtattttgga gggatttgtg tagtagtagc agcggcgccg cggggacgga tgcggatggt ggcgctttcg
      2210       2220       2230       2240       2250       2260       2270       2280       2290       2300
gtggcgtttt cccggggggg ttttggtttg gcgcttgggg gggatggcat ggcgcggcgt gcggctgcac gccacacaca cgcgcgcgca cgcacgtacg

(EMBL)    length:  2900    current pos:    401
```

ある配列のcDNAの相補差

```
ID   C20 oxidase for a candidate gen of sd1
XX
SQ   Sequence  1170 BP;  171 A;  455 C;  152 T;  392 G.

        10         20         30         40         50         60         70         80         90        100
atggtggccg agcaccccac gccaccacag ccgcaccaac caccgcccat ggactccacc gccggctctg gcattgccgc cccggcggcg gcggcggtgt
       110        120        130        140        150        160        170        180        190        200
gcgacctgag gatggagccc aagatcccgg agccattcgt gtggccgaac ggcgacgcga ggccggcgtc ggcggcggag ctggacatgc ccgtggtcga
       210        220        230        240        250        260        270        280        290        300
cgtgggcgtg ctccgcgacg gcgacgccga ggggctgcgc cgcgccgcgg cgcaggtggc cgccgcgtgc gccacgcacg ggttcttcca ggtgtccgag
       310        320        330        340        350        360        370        380        390        400
cacggcgtcg acgccgctct ggcgcgcgcc gcgctcgacg gcgccagcga cttcttccgc ctcccgctcg ccgagaagcg ccgcgcgcgc cgcgtcccgg
       410        420        430        440        450        460        470        480        490        500
gcaccgtgtc cggctacacc agcgcccacg ccgaccgctt cgcctccaag ctcccatgga aggagaccct ctccttcggc ttccacgacc gcgccgccgc
       510        520        530        540        550        560        570        580        590        600
ccccgtcgtc gccgactact tctccagcac cctcggcccc gacttcgcgc caatggggag ggtgtaccag aagtactgcg aggagatgaa ggagctgtcg
       610        620        630        640        650        660        670        680        690        700
                                                                                                        tact
                                                                                                         800
                                                                                                        tcga
                                                                                                         900
                                                                                                         tat
       910        920        930        940        950        960        970        980        990       1000
aagagctgcc tgcacagggc ggtggtgaac cagcggcggg agcggcggtc gctggcgttc ttcctgtgcc cgcgggagga cagggtggtg cggccgccgc
      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
cgagcgccgc cacgccgcag cgctacccgg acttcacctg ggccgacctc atgcgcttca cgcagcgcca ctaccgcgcc gacacccgca cgctcgacgc
      1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
cttcacgcgc tggctcgcgc cgccggccgc cgacgccgcc gcgacggcgc aggtcgaggc ggccagctga

(EMBL)    length:  1170    current pos:    1171
```

cDNA GTCTGC**ATGCCATTG****GTCGTGTGA**AACAATAAAGGCAAAAAAAA

例

GA20oxidase2-genomic DNA

ある配列のGenomicDNA

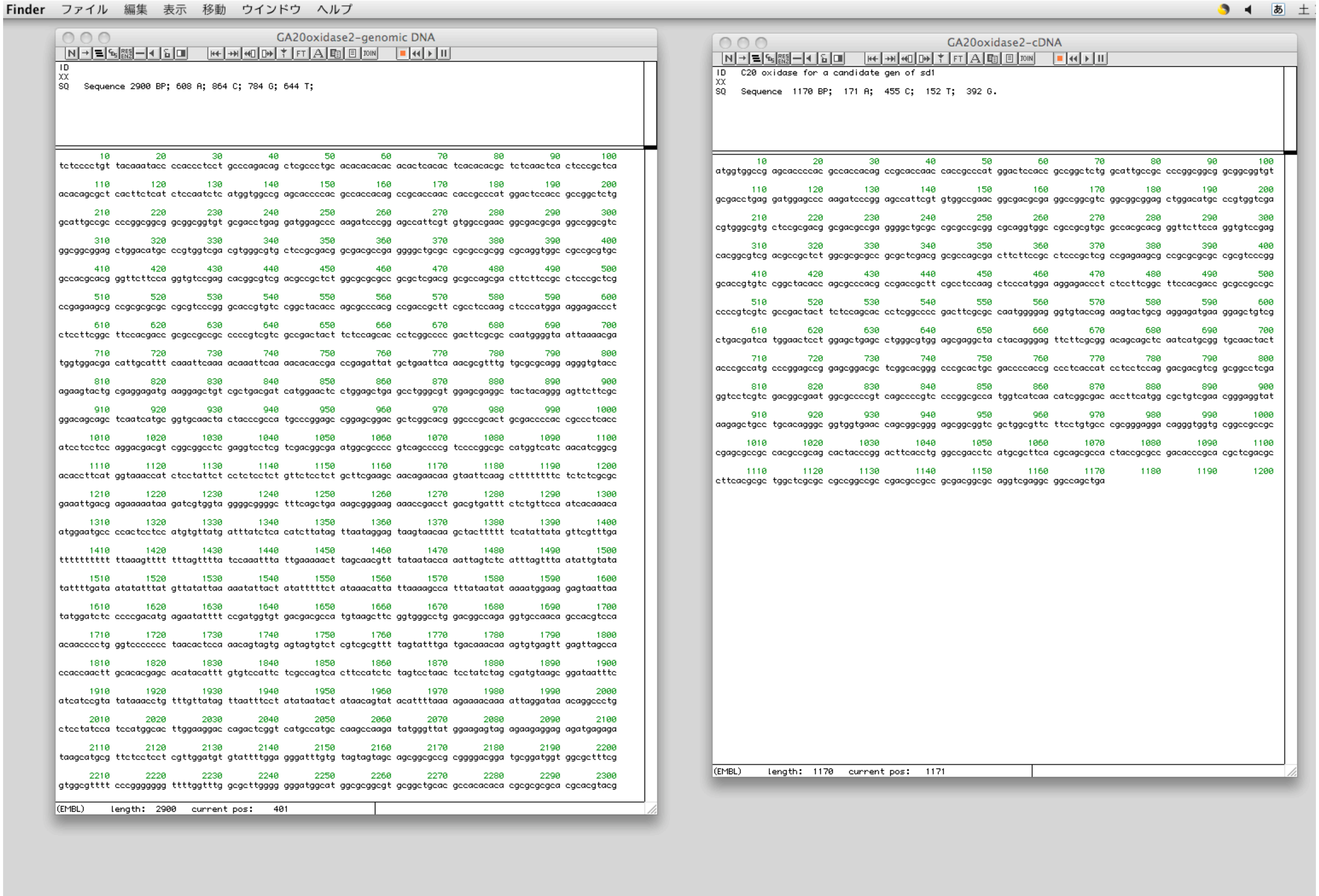
```
ID
XX
SQ   Sequence 2900 BP; 608 A; 864 C; 784 G; 644 T;

        10         20         30         40         50         60         70         80         90        100
tctcccctgt tacaaatacc ccaccctcct gcccagacag ctcgccctgc acacacacac acactcacac tcacacacgc tctcaactca ctcccgctca
       110        120        130        140        150        160        170        180        190        200
acacagcgct cacttctcat ctccaatctc atggtggccg agcaccccac gccaccacag ccgcaccaac caccgcccat ggactccacc gccggctctg
       210        220        230        240        250        260        270        280        290        300
gcattgccgc cccggcggcg gcggcggtgt gcgacctgag gatggagccc aagatcccgg agccattcgt gtggccgaac ggcgacgcga ggccggcgtc
       310        320        330        340        350        360        370        380        390        400
ggcggcggag ctggacatgc ccgtggtcga cgtgggcgtg ctccgcgacg gcgacgccga ggggctgcgc cgcgccgcgg cgcaggtggc cgccgcgtgc
       410        420        430        440        450        460        470        480        490        500
gccacgcacg ggttcttcca ggtgtccgag cacggcgtcg acgccgctct ggcgcgcgcc gcgctcgacg gcgccagcga cttcttccgc ctcccgctcg
       510        520        530        540        550        560        570        580        590        600
ccgagaagcg ccgcgcgcgc cgcgtcccgg gcaccgtgtc cggctacacc agcgcccacg ccgaccgctt cgcctccaag ctcccatgga aggagaccct
       610        620        630        640        650        660        670        680        690        700
ctccttcggc ttccacgacc gcgccgccgc ccccgtcgtc gccgactact tctccagcac cctcggcccc gacttcgcgc caatggggta attaaacga
       710        720        730        740        750        760        770        780        790        800
tggtggacga cattgcattt caaattcaaa acaaattcaa aacacaccga ccgagattat gctgaattca aacgcgtttg tgcgcgcagg agggtgtacc
       810        820        830        840        850        860        870        880        890        900
agaagtactg cgaggagatg aaggagctgt cgctgacgat catggaactc ctggagctga gcctgggcgt ggagcgaggc tactacaggg agttcttcgc
       910        920        930        940        950        960        970        980        990       1000
ggacagcagc tcaatcatgc ggtgcaacta ctacccgcca tgcccggagc cggagcggac gctcggcacg ggcccgcact gcgaccccac cgccctcacc
      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
atcctcctcc aggacgacgt cggcggcctc gaggtcctcg tcgacggcga atggcgcccc gtcagccccg tccccggcgc catggtcatc aacatcggcg
      1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
acaccttcat ggtaaaccat ctcctattct cctctcctct gttctcctct gcttcgaagc aacagaacaa gtaattcaag ctttttttc tctctcgcgc
      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300
gaaattgacg agaaaaataa gatcgtggta ggggcggggc tttcagctga aagcgggaag aaaccgacct gacgtgattt ctctgttcca atcacaaaca
      1310       1320       1330       1340       1350       1360       1370       1380       1390       1400
atggaatgcc ccactcctcc atgtgttatg atttatctca catcttatag ttaataggag taagtaacaa gctactttt tcatattata gttcgtttga
      1410       1420       1430       1440       1450       1460       1470       1480       1490       1500
tttttttttt ttaaagtttt tttagtttta tccaaattta ttgaaaaact tagcaacgtt tataatacca aattagtctc atttagttta atattgtata
      1510       1520       1530       1540       1550       1560       1570       1580       1590       1600
tattttgata atatatttat gttatattaa aaatattact atatttttct ataaacatta ttaaaagcca tttataatat aaaatggaag gagtaattaa
      1610       1620       1630       1640       1650       1660       1670       1680       1690       1700
tatggatctc ccccgacatg agaatatttt ccgatggtgt gacgacgcca tgtaagcttc ggtgggcctg gacggccaga ggtgccaaca gccacgtcca
      1710       1720       1730       1740       1750       1760       1770       1780       1790       1800
acaaccccctg ggtccccccc taacactcca aacagtagtg agtagtgtct cgtcgcgttt tagtatttga tgacaaacaa agtgtgagtt gagttagcca
      1810       1820       1830       1840       1850       1860       1870       1880       1890       1900
ccaccaactt gcacacgagc acatacattt gtgtccattc tcgccagtca cttccatctc tagtcctaac tcctatctag cgatgtaagc ggataatttc
      1910       1920       1930       1940       1950       1960       1970       1980       1990       2000
atcatccgta tataaacctg tttgttatag ttaatttcct atataatact ataacagtat acattttaaa agaaaacaaa attaggataa acaggccctg
      2010       2020       2030       2040       2050       2060       2070       2080       2090       2100
ctcctatcca tccatggcac ttggaaggac cagactcggt catgccatgc caagccaaga tatgggttat ggaagagtag agaagaggag agatgagaga
      2110       2120       2130       2140       2150       2160       2170       2180       2190       2200
taagcatgcg ttctcctcct cgttggatgt gtattttgga gggatttgtg tagtagtagc agcggcgccg cggggacgga tgcggatggt ggcgctttcg
      2210       2220       2230       2240       2250       2260       2270       2280       2290       2300
gtggcgtttt cccggggggg ttttggtttg gcgcttgggg gggatggcat ggcgcggcgt gcggctgcac gccacacaca cgcgcgcgca cgcacgtacg

(EMBL)     length:  2900    current pos:    401
```

GA20oxidase2-cDNA

ある配列のcDNAの相補鎖

```
ID   C20 oxidase for a candidate gen of sd1
XX
SQ   Sequence  1170 BP;  171 A;  455 C;  152 T;  392 G.

        10         20         30         40         50         60         70         80         90        100
atggtggccg agcaccccac gccaccacag ccgcaccaac caccgcccat ggactccacc gccggctctg gcattgccgc cccggcggcg gcggcggtgt
       110        120        130        140        150        160        170        180        190        200
gcgacctgag gatggagccc aagatcccgg agccattcgt gtggccgaac ggcgacgcga ggccggcgtc ggcggcggag ctggacatgc ccgtggtcga
       210        220        230        240        250        260        270        280        290        300
cgtgggcgtg ctccgcgacg gcgacgccga ggggctgcgc cgcgccgcgg cgcaggtggc cgccgcgtgc gccacgcacg ggttcttcca ggtgtccgag
       310        320        330        340        350        360        370        380        390        400
cacggcgtcg acgccgctct ggcgcgcgcc gcgctcgacg gcgccagcga cttcttccgc ctcccgctcg ccgagaagcg ccgcgcgcgc cgcgtcccgg
       410        420        430        440        450        460        470        480        490        500
gcaccgtgtc cggctacacc agcgcccacg ccgaccgctt cgcctccaag ctcccatgga aggagaccct ctccttcggc ttccacgacc gcgccgccgc
       510        520        530        540        550        560        570        580        590        600
ccccgtcgtc gccgactact tctccagcac cctcggcccc gacttcgcgc caatggggag ggtgtaccag aagtactgcg aggagatgaa ggagctgtcg
       610        620        630        640        650        660        670        680        690        700
ctgacgatca tggaactcct ggagctgagc ctgggcgtgg agcgaggcta ctacagggag ttcttcgcgg acagcagctc aatcatgcgg tgcaactact
       710        720        730        740        750        760        770        780        790        800
acccgccatg cccggagccg gagcggacgc tcggcacggg cccgcactgc gaccccaccg ccctcaccat cctcctccag gacgacgtcg gcggcctcga
       810        820        830        840        850        860        870        880        890        900
ggtcctcgtc gacggcgaat ggcgccccgt cagccccgtc cccggcgcca tggtcatcaa catcggcgac accttcatgg cgctgtcgaa cgggaggtat
       910        920        930        940        950        960        970        980        990       1000
aagagctgcc tgcacagggc ggtggtgaac cagcggcggg agcggcggtc gctggcgttc ttcctgtgcc cgcgggagga cagggtggtg cggccgccgc
      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
cgagcgccgc cacgccgcag cactacccgg acttcacctg ggccgacctc atgcgcttca cgcagcgcca ctaccgcgcc gacacccgca cgctcgacgc
      1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
cttcacgcgc tggctcgcgc cgccggccgc cgacgccgcc gcgacggcgc aggtcgaggc ggccagctga

(EMBL)     length:  1170    current pos:    1171
```

マニュアルで配列を解析してみよう

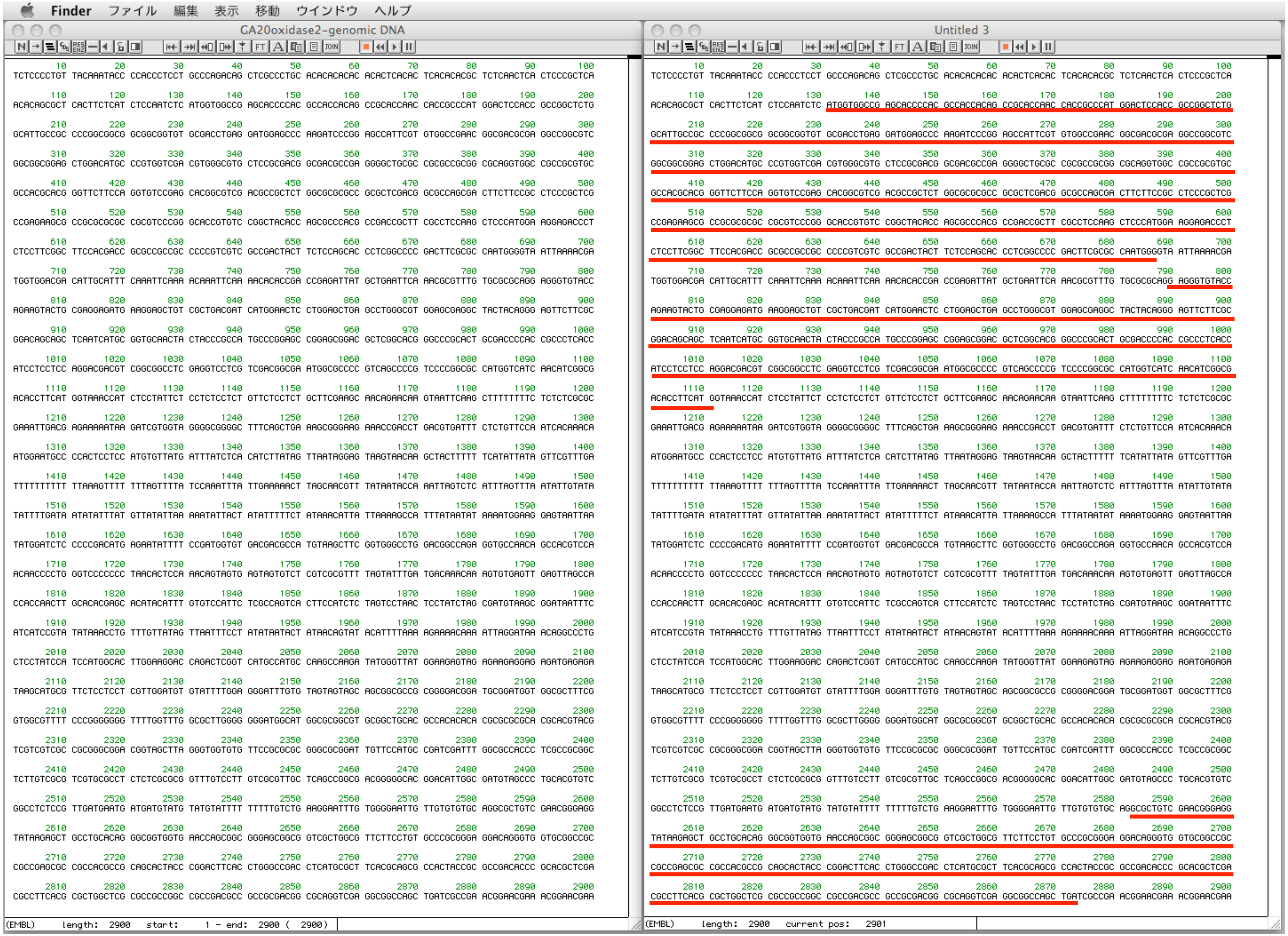


GA20oxidase2-genomic DNA

```
ID
XX
SQ   Sequence 2900 BP; 608 A; 864 C; 784 G; 644 T;

        10         20         30         40         50         60         70         80         90        100
tctcccctgt tacaaatacc ccaccctcct gcccagacag ctcgccctgc acacacacac acactcacac tcacacacgc tctcaactca ctcccgctca

       110        120        130        140        150        160        170        180        190        200
acacagcgct cacttctcat ctccaatctc atggtggccg agcaccccac gccaccacag ccgcaccaac caccgcccat ggactccacc gccggctctg

       210        220        230        240        250        260        270        280        290        300
gcattgccgc cccggcggcg gcggcggtgt gcgacctgag gatggagccc aagatcccgg agccattcgt gtggccgaac ggcgacgcga ggccggcgtc

       310        320        330        340        350        360        370        380        390        400
ggcggcggag ctggacatgc ccgtggtcga cgtgggcgtg ctccgcgacg gcgacgccga ggggctgcgc cgcgccgcgg cgcaggtggc cgccgcgtgc

       410        420        430        440        450        460        470        480        490        500
gccacgcacg ggttcttcca ggtgtccgag cacggcgtcg acgccgctct ggcgcgcgcc gcgctcgacg gcgccagcga cttcttccgc ctcccgctcg

       510        520        530        540        550        560        570        580        590        600
ccgagaagcg ccgcgcgcgc cgcgtcccgg gcaccgtgtc cggctacacc agcgcccacg ccgaccgctt cgcctccaag ctcccatgga aggagaccct

       610        620        630        640        650        660        670        680        690        700
ctccttcggc ttccacgacc gcgccgccgc ccccgtcgtc gccgactact tctccagcac cctcggcccc gacttcgcgc caatggggta attaaaacga

       710        720        730        740        750        760        770        780        790        800
tggtggacga cattgcattt caaattcaaa acaaattcaa aacacaccga ccgagattat gctgaattca aacgcgtttg tgcgcgcagg agggtgtacc

       810        820        830        840        850        860        870        880        890        900
agaagtactg cgaggagatg aaggagctgt cgctgacgat catggaactc ctggagctga gcctgggcgt ggagcgaggc tactacaggg agttcttcgc

       910        920        930        940        950        960        970        980        990       1000
ggacagcagc tcaatcatgc ggtgcaacta ctacccgcca tgcccggagc cggagcggac gctcggcacg ggcccgcact gcgaccccac cgccctcacc

      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
atcctcctcc aggacgacgt cggcggcctc gaggtcctcg tcgacggcga atggcgcccc gtcagccccg tccccggcgc catggtcatc aacatcggcg

      1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
acaccttcat ggtaaaccat ctcctattct cctctcctct gttctcctct gcttcgaagc aacagaacaa gtaattcaag cttttttttc tctctcgcgc

      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300
gaaattgacg agaaaaataa gatcgtggta ggggcggggc tttcagctga aagcgggaag aaaccgacct gacgtgattt ctctgttcca atcacaaaca

      1310       1320       1330       1340       1350       1360       1370       1380       1390       1400
atggaatgcc ccactcctcc atgtgttatg atttatctca catcttatag ttaataggag taagtaacaa gctacttttt tcatattata gttcgtttga

      1410       1420       1430       1440       1450       1460       1470       1480       1490       1500
tttttttttt ttaaagtttt tttagtttta tccaaattta ttgaaaaact tagcaacgtt tataatacca aattagtctc atttagttta atattgtata

      1510       1520       1530       1540       1550       1560       1570       1580       1590       1600
tattttgata atatatttat gttatattaa aaatattact atatttttct ataaacatta ttaaaagcca tttataatat aaaatggaag gagtaattaa

      1610       1620       1630       1640       1650       1660       1670       1680       1690       1700
tatggatctc ccccgacatg agaatatttt ccgatggtgt gacgacgcca tgtaagcttc ggtgggcctg gacggccaga ggtgccaaca gccacgtcca

      1710       1720       1730       1740       1750       1760       1770       1780       1790       1800
acaaccccctg ggtccccccc taacactcca aacagtagtg agtagtgtct cgtcgcgttt tagtatttga tgacaaacaa agtgtgagtt gagttagcca

      1810       1820       1830       1840       1850       1860       1870       1880       1890       1900
ccaccaactt gcacacgagc acatacattt gtgtccattc tcgccagtca cttccatctc tagtcctaac tcctatctag cgatgtaagc ggataatttc

      1910       1920       1930       1940       1950       1960       1970       1980       1990       2000
atcatccgta tataaacctg tttgttatag ttaatttcct atataatact ataacagtat acattttaaa agaaaacaaa attaggataa acaggccctg

      2010       2020       2030       2040       2050       2060       2070       2080       2090       2100
ctcctatcca tccatggcac ttggaaggac cagactcggt catgccatgc caagccaaga tatgggttat ggaagagtag agaagaggag agatgagaga

      2110       2120       2130       2140       2150       2160       2170       2180       2190       2200
taagcatgcg ttctcctcct cgttggatgt gtattttgga gggatttgtg tagtagtagc agcggcgccg cggggacgga tgcggatggt ggcgctttcg

      2210       2220       2230       2240       2250       2260       2270       2280       2290       2300
gtggcgtttt cccggggggg ttttggtttg gcgcttgggg gggatggcat ggcgcggcgt gcggctgcac gccacacaca cgcgcgcgca cgcacgtacg

(EMBL)      length:  2900    current pos:    401
```

GA20oxidase2-cDNA

```
ID   C20 oxidase for a candidate gen of sd1
XX
SQ   Sequence  1170 BP;  171 A;  455 C;  152 T;  392 G.

        10         20         30         40         50         60         70         80         90        100
atggtggccg agcaccccac gccaccacag ccgcaccaac caccgcccat ggactccacc gccggctctg gcattgccgc cccggcggcg gcggcggtgt

       110        120        130        140        150        160        170        180        190        200
gcgacctgag gatggagccc aagatcccgg agccattcgt gtggccgaac ggcgacgcga ggccggcgtc ggcggcggag ctggacatgc ccgtggtcga

       210        220        230        240        250        260        270        280        290        300
cgtgggcgtg ctccgcgacg gcgacgccga ggggctgcgc cgcgccgcgg cgcaggtggc cgccgcgtgc gccacgcacg ggttcttcca ggtgtccgag

       310        320        330        340        350        360        370        380        390        400
cacggcgtcg acgccgctct ggcgcgcgcc gcgctcgacg gcgccagcga cttcttccgc ctcccgctcg ccgagaagcg ccgcgcgcgc cgcgtcccgg

       410        420        430        440        450        460        470        480        490        500
gcaccgtgtc cggctacacc agcgcccacg ccgaccgctt cgcctccaag ctcccatgga aggagaccct ctccttcggc ttccacgacc gcgccgccgc

       510        520        530        540        550        560        570        580        590        600
ccccgtcgtc gccgactact tctccagcac cctcggcccc gacttcgcgc caatggggag ggtgtaccag aagtactgcg aggagatgaa ggagctgtcg

       610        620        630        640        650        660        670        680        690        700
ctgacgatca tggaactcct ggagctgagc ctgggcgtgg agcgaggcta ctacagggag ttcttcgcgg acagcagctc aatcatgcgg tgcaactact

       710        720        730        740        750        760        770        780        790        800
acccgccatg cccggagccg gagcggacgc tcggcacggg cccgcactgc gaccccaccg ccctcaccat cctcctccag gacgacgtcg gcggcctcga

       810        820        830        840        850        860        870        880        890        900
ggtcctcgtc gacggcgaat ggcgccccgt cagccccgtc cccggcgcca tggtcatcaa catcggcgac accttcatgg cgctgtcgaa cgggaggtat

       910        920        930        940        950        960        970        980        990       1000
aagagctgcc tgcacagggc ggtggtgaac cagcggcggg agcggcggtc gctggcgttc ttcctgtgcc cgcgggagga cagggtggtg cggccgccgc

      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
cgagcgccgc cacgccgcag cactacccgg acttcacctg ggccgacctc atgcgcttca cgcagcgcca ctaccgcgcc gacacccgca cgctcgacgc

      1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
cttcacgcgc tggctcgcgc cgccggccgc cgacgccgcc gcgacggcgc aggtcgaggc ggccagctga

(EMBL)      length:  1170    current pos:    1171
```



```
        10         20         30         40         50         60         70         80         90        100
TCTCCCCTGT TACAAATACC CCACCCTCCT GCCCAGACAG CTCGCCCTGC ACACACACAC ACACTCACAC TCACACACGC TCTCAACTCA CTCCCGCTCA

       110        120        130        140        150        160        170        180        190        200
ACACAGCGCT CACTTCTCAT CTCCAATCTC ATGGTGGCCG AGCACCCCAC GCCACCACAG CCGCACCAAC CACCGCCCAT GGACTCCACC GCCGGCTCTG

       210        220        230        240        250        260        270        280        290        300
GCATTGCCGC CCCGGCGGCG GCGGCGGTGT GCGACCTGAG GATGGAGCCC AAGATCCCGG AGCCATTCGT GTGGCCGAAC GGCGACGCGA GGCCGGCGTC

       310        320        330        340        350        360        370        380        390        400
GGCGGCGGAG CTGGACATGC CCGTGGTCGA CGTGGGCGTG CTCCGCGACG GCGACGCCGA GGGGCTGCGC CGCGCCGCGG CGCAGGTGGC CGCCGCGTGC

       410        420        430        440        450        460        470        480        490        500
GCCACGCACG GGTTCTTCCA GGTGTCCGAG CACGGCGTCG ACGCCGCTCT GGCGCGCGCC GCGCTCGACG GCGCCAGCGA CTTCTTCCGC CTCCCGCTCG

       510        520        530        540        550        560        570        580        590        600
CCGAGAAGCG CCGCGCGCGC CGCGTCCCGG GCACCGTGTC CGGCTACACC AGCGCCCACG CCGACCGCTT CGCCTCCAAG CTCCCATGGA AGGAGACCCT

       610        620        630        640        650        660        670        680        690        700
CTCCTTCGGC TTCCACGACC GCGCCGCCGC CCCCGTCGTC GCCGACTACT TCTCCAGCAC CCTCGGCCCC GACTTCGCGC CAATGGGGTA ATTAAAACGA

       710        720        730        740        750        760        770        780        790        800
TGGTGGACGA CATTGCATTT CAAATTCAAA ACAAATTCAA AACACACCGA CCGAGATTAT GCTGAATTCA AACGCGTTTG TGCGCGCAGG AGGGTGTACC

       810        820        830        840        850        860        870        880        890        900
AGAAGTACTG CGAGGAGATG AAGGAGCTGT CGCTGACGAT CATGGAACTC CTGGAGCTGA GCCTGGGCGT GGAGCGAGGC TACTACAGGG AGTTCTTCGC

       910        920        930        940        950        960        970        980        990       1000
GGACAGCAGC TCAATCATGC GGTGCAACTA CTACCCGCCA TGCCCGGAGC CGGAGCGGAC GCTCGGCACG GGCCCGCACT GCGACCCCAC CGCCCTCACC

      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
ATCCTCCTCC AGGACGACGT CGGCGGCCTC GAGGTCCTCG TCGACGGCGA ATGGCGCCCC GTCAGCCCCG TCCCCGGCGC CATGGTCATC AACATCGGCG

      1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
ACACCTTCAT GGTAAACCAT CTCCTATTCT CCTCTCCTCT GTTCTCCTCT GCTTCGAAGC AACAGAACAA GTAATTCAAG CTTTTTTTTC TCTCTCGCGC

      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300
GAAATTGACG AGAAAAATAA GATCGTGGTA GGGGCGGGGC TTTCAGCTGA AAGCGGGAAG AAACCGACCT GACGTGATTT CTCTGTTCCA ATCACAAACA

      1310       1320       1330       1340       1350       1360       1370       1380       1390       1400
ATGGAATGCC CCACTCCTCC ATGTGTTATG ATTTATCTCA CATCTTATAG TTAATAGGAG TAAGTAACAA GCTACTTTTT TCATATTATA GTTCGTTTGA

      1410       1420       1430       1440       1450       1460       1470       1480       1490       1500
TTTTTTTTTT TTAAAGTTTT TTTAGTTTTA TCCAAATTTA TTGAAAAACT TAGCAACGTT TATAATACCA AATTAGTCTC ATTTAGTTTA ATATTGTATA

      1510       1520       1530       1540       1550       1560       1570       1580       1590       1600
TATTTTGATA ATATATTTAT GTTATATTAA AAATATTACT ATATTTTTCT ATAAACATTA TTAAAAGCCA TTTATAATAT AAAATGGAAG GAGTAATTAA

      1610       1620       1630       1640       1650       1660       1670       1680       1690       1700
TATGGATCTC CCCCGACATG AGAATATTTT CCGATGGTGT GACGACGCCA TGTAAGCTTC GGTGGGCCTG GACGGCCAGA GGTGCCAACA GCCACGTCCA

      1710       1720       1730       1740       1750       1760       1770       1780       1790       1800
ACAACCCCTG GGTCCCCCCC TAACACTCCA AACAGTAGTG AGTAGTGTCT CGTCGCGTTT TAGTATTTGA TGACAAACAA AGTGTGAGTT GAGTTAGCCA

      1810       1820       1830       1840       1850       1860       1870       1880       1890       1900
CCACCAACTT GCACACGAGC ACATACATTT GTGTCCATTC TCGCCAGTCA CTTCCATCTC TAGTCCTAAC TCCTATCTAG CGATGTAAGC GGATAATTTC

      1910       1920       1930       1940       1950       1960       1970       1980       1990       2000
ATCATCCGTA TATAAACCTG TTTGTTATAG TTAATTTCCT ATATAATACT ATAACAGTAT ACATTTTAAA AGAAAACAAA ATTAGGATAA ACAGGCCCTG

      2010       2020       2030       2040       2050       2060       2070       2080       2090       2100
CTCCTATCCA TCCATGGCAC TTGGAAGGAC CAGACTCGGT CATGCCATGC CAAGCCAAGA TATGGGTTAT GGAAGAGTAG AGAAGAGGAG AGATGAGAGA

      2110       2120       2130       2140       2150       2160       2170       2180       2190       2200
TAAGCATGCG TTCTCCTCCT CGTTGGATGT GTATTTTGGA GGGATTTGTG TAGTAGTAGC AGCGGCGCCG CGGGGACGGA TGCGGATGGT GGCGCTTTCG

      2210       2220       2230       2240       2250       2260       2270       2280       2290       2300
GTGGCGTTTT CCCGGGGGGG TTTTGGTTTG GCGCTTGGGG GGGATGGCAT GGCGCGGCGT GCGGCTGCAC GCCACACACA CGCGCGCGCA CGCACGTACG

      2310       2320       2330       2340       2350       2360       2370       2380       2390       2400
TCGTCGTCGC CGCGGGCGGA CGGTAGCTTA GGGTGGTGTG TTCCGCGCGC GGGCGCGGAT TGTTCCATGC CGATCGATTT GGCGCCACCC TCGCCGCGGC

      2410       2420       2430       2440       2450       2460       2470       2480       2490       2500
TCTTGTCGCG TCGTGCGCCT CTCTCGCGCG GTTTGTCCTT GTCGCGTTGC TCAGCCGGCG ACGGGGGCAC GGACATTGGC GATGTAGCCC TGCACGTGTC

      2510       2520       2530       2540       2550       2560       2570       2580       2590       2600
GGCCTCTCCG TTGATGAATG ATGATGTATG TATGTATTTT TTTTTGTCTG AAGGAATTTG TGGGGAATTG TTGTGTGTGC AGGCGCTGTC GAACGGGAGG

      2610       2620       2630       2640       2650       2660       2670       2680       2690       2700
TATAAGAGCT GCCTGCACAG GGCGGTGGTG AACCAGCGGC GGGAGCGGCG GTCGCTGGCG TTCTTCCTGT GCCCGCGGGA GGACAGGGTG GTGCGGCCGC

      2710       2720       2730       2740       2750       2760       2770       2780       2790       2800
CGCCGAGCGC CGCCACGCCG CAGCACTACC CGGACTTCAC CTGGGCCGAC CTCATGCGCT TCACGCAGCG CCACTACCGC GCCGACACCC GCACGCTCGA

      2810       2820       2830       2840       2850       2860       2870       2880       2890       2900
CGCCTTCACG CGCTGGCTCG CGCCGCCGGC CGCCGACGCC GCCGCGACGG CGCAGGTCGA GGCGGCCAGC TGATCGCCGA ACGGAACGAA ACGGAACGAA

(EMBL)     length:  2900   start:      1 - end:  2900 (  2900)
```

```
        10         20         30         40         50         60         70         80         90        100
TCTCCCCTGT TACAAATACC CCACCCTCCT GCCCAGACAG CTCGCCCTGC ACACACACAC ACACTCACAC TCACACACGC TCTCAACTCA CTCCCGCTCA

       110        120        130        140        150        160        170        180        190        200
ACACAGCGCT CACTTCTCAT CTCCAATCTC ATGGTGGCCG AGCACCCCAC GCCACCACAG CCGCACCAAC CACCGCCCAT GGACTCCACC GCCGGCTCTG

       210        220        230        240        250        260        270        280        290        300
GCATTGCCGC CCCGGCGGCG GCGGCGGTGT GCGACCTGAG GATGGAGCCC AAGATCCCGG AGCCATTCGT GTGGCCGAAC GGCGACGCGA GGCCGGCGTC

       310        320        330        340        350        360        370        380        390        400
GGCGGCGGAG CTGGACATGC CCGTGGTCGA CGTGGGCGTG CTCCGCGACG GCGACGCCGA GGGGCTGCGC CGCGCCGCGG CGCAGGTGGC CGCCGCGTGC

       410        420        430        440        450        460        470        480        490        500
GCCACGCACG GGTTCTTCCA GGTGTCCGAG CACGGCGTCG ACGCCGCTCT GGCGCGCGCC GCGCTCGACG GCGCCAGCGA CTTCTTCCGC CTCCCGCTCG

       510        520        530        540        550        560        570        580        590        600
CCGAGAAGCG CCGCGCGCGC CGCGTCCCGG GCACCGTGTC CGGCTACACC AGCGCCCACG CCGACCGCTT CGCCTCCAAG CTCCCATGGA AGGAGACCCT

       610        620        630        640        650        660        670        680        690        700
CTCCTTCGGC TTCCACGACC GCGCCGCCGC CCCCGTCGTC GCCGACTACT TCTCCAGCAC CCTCGGCCCC GACTTCGCGC CAATGGGGTA ATTAAAACGA

       710        720        730        740        750        760        770        780        790        800
TGGTGGACGA CATTGCATTT CAAATTCAAA ACAAATTCAA AACACACCGA CCGAGATTAT GCTGAATTCA AACGCGTTTG TGCGCGCAGG AGGGTGTACC

       810        820        830        840        850        860        870        880        890        900
AGAAGTACTG CGAGGAGATG AAGGAGCTGT CGCTGACGAT CATGGAACTC CTGGAGCTGA GCCTGGGCGT GGAGCGAGGC TACTACAGGG AGTTCTTCGC

       910        920        930        940        950        960        970        980        990       1000
GGACAGCAGC TCAATCATGC GGTGCAACTA CTACCCGCCA TGCCCGGAGC CGGAGCGGAC GCTCGGCACG GGCCCGCACT GCGACCCCAC CGCCCTCACC

      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
ATCCTCCTCC AGGACGACGT CGGCGGCCTC GAGGTCCTCG TCGACGGCGA ATGGCGCCCC GTCAGCCCCG TCCCCGGCGC CATGGTCATC AACATCGGCG

      1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
ACACCTTCAT GGTAAACCAT CTCCTATTCT CCTCTCCTCT GTTCTCCTCT GCTTCGAAGC AACAGAACAA GTAATTCAAG CTTTTTTTTC TCTCTCGCGC

      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300
GAAATTGACG AGAAAAATAA GATCGTGGTA GGGGCGGGGC TTTCAGCTGA AAGCGGGAAG AAACCGACCT GACGTGATTT CTCTGTTCCA ATCACAAACA

      1310       1320       1330       1340       1350       1360       1370       1380       1390       1400
ATGGAATGCC CCACTCCTCC ATGTGTTATG ATTTATCTCA CATCTTATAG TTAATAGGAG TAAGTAACAA GCTACTTTTT TCATATTATA GTTCGTTTGA

      1410       1420       1430       1440       1450       1460       1470       1480       1490       1500
TTTTTTTTTT TTAAAGTTTT TTTAGTTTTA TCCAAATTTA TTGAAAAACT TAGCAACGTT TATAATACCA AATTAGTCTC ATTTAGTTTA ATATTGTATA

      1510       1520       1530       1540       1550       1560       1570       1580       1590       1600
TATTTTGATA ATATATTTAT GTTATATTAA AAATATTACT ATATTTTTCT ATAAACATTA TTAAAAGCCA TTTATAATAT AAAATGGAAG GAGTAATTAA

      1610       1620       1630       1640       1650       1660       1670       1680       1690       1700
TATGGATCTC CCCCGACATG AGAATATTTT CCGATGGTGT GACGACGCCA TGTAAGCTTC GGTGGGCCTG GACGGCCAGA GGTGCCAACA GCCACGTCCA

      1710       1720       1730       1740       1750       1760       1770       1780       1790       1800
ACAACCCCTG GGTCCCCCCC TAACACTCCA AACAGTAGTG AGTAGTGTCT CGTCGCGTTT TAGTATTTGA TGACAAACAA AGTGTGAGTT GAGTTAGCCA

      1810       1820       1830       1840       1850       1860       1870       1880       1890       1900
CCACCAACTT GCACACGAGC ACATACATTT GTGTCCATTC TCGCCAGTCA CTTCCATCTC TAGTCCTAAC TCCTATCTAG CGATGTAAGC GGATAATTTC

      1910       1920       1930       1940       1950       1960       1970       1980       1990       2000
ATCATCCGTA TATAAACCTG TTTGTTATAG TTAATTTCCT ATATAATACT ATAACAGTAT ACATTTTAAA AGAAAACAAA ATTAGGATAA ACAGGCCCTG

      2010       2020       2030       2040       2050       2060       2070       2080       2090       2100
CTCCTATCCA TCCATGGCAC TTGGAAGGAC CAGACTCGGT CATGCCATGC CAAGCCAAGA TATGGGTTAT GGAAGAGTAG AGAAGAGGAG AGATGAGAGA

      2110       2120       2130       2140       2150       2160       2170       2180       2190       2200
TAAGCATGCG TTCTCCTCCT CGTTGGATGT GTATTTTGGA GGGATTTGTG TAGTAGTAGC AGCGGCGCCG CGGGGACGGA TGCGGATGGT GGCGCTTTCG

      2210       2220       2230       2240       2250       2260       2270       2280       2290       2300
GTGGCGTTTT CCCGGGGGGG TTTTGGTTTG GCGCTTGGGG GGGATGGCAT GGCGCGGCGT GCGGCTGCAC GCCACACACA CGCGCGCGCA CGCACGTACG

      2310       2320       2330       2340       2350       2360       2370       2380       2390       2400
TCGTCGTCGC CGCGGGCGGA CGGTAGCTTA GGGTGGTGTG TTCCGCGCGC GGGCGCGGAT TGTTCCATGC CGATCGATTT GGCGCCACCC TCGCCGCGGC

      2410       2420       2430       2440       2450       2460       2470       2480       2490       2500
TCTTGTCGCG TCGTGCGCCT CTCTCGCGCG GTTTGTCCTT GTCGCGTTGC TCAGCCGGCG ACGGGGGCAC GGACATTGGC GATGTAGCCC TGCACGTGTC

      2510       2520       2530       2540       2550       2560       2570       2580       2590       2600
GGCCTCTCCG TTGATGAATG ATGATGTATG TATGTATTTT TTTTTGTCTG AAGGAATTTG TGGGGAATTG TTGTGTGTGC AGGCGCTGTC GAACGGGAGG

      2610       2620       2630       2640       2650       2660       2670       2680       2690       2700
TATAAGAGCT GCCTGCACAG GGCGGTGGTG AACCAGCGGC GGGAGCGGCG GTCGCTGGCG TTCTTCCTGT GCCCGCGGGA GGACAGGGTG GTGCGGCCGC

      2710       2720       2730       2740       2750       2760       2770       2780       2790       2800
CGCCGAGCGC CGCCACGCCG CAGCACTACC CGGACTTCAC CTGGGCCGAC CTCATGCGCT TCACGCAGCG CCACTACCGC GCCGACACCC GCACGCTCGA

      2810       2820       2830       2840       2850       2860       2870       2880       2890       2900
CGCCTTCACG CGCTGGCTCG CGCCGCCGGC CGCCGACGCC GCCGCGACGG CGCAGGTCGA GGCGGCCAGC TGATCGCCGA ACGGAACGAA ACGGAACGAA
```

```
(EMBL)     length:  2900   current pos:   2901
```

解析ソフトを用いて遺伝子を探す

cDNAが有る場合

生物情報工学II (担当　芦苅）遺伝子の配列解析

第６回目-1（11/1）遺伝子の予測と機能解析

　　-2（11/1）演習

第７回目-1（11/8）DNA配列から遺伝子を探す

　　-2（11/8）DNA、アミノ酸のアラインメント、系統樹解析

　　-3（11/8）演習

第８回目-1（11/15）Genetyxを用いた解析

　　-2（11/15）演習

第９回目-1（11/22）KEGGを用いた解析

連鎖解析

出席と宿題の提出　提出方法

リンク集

データベース検索：

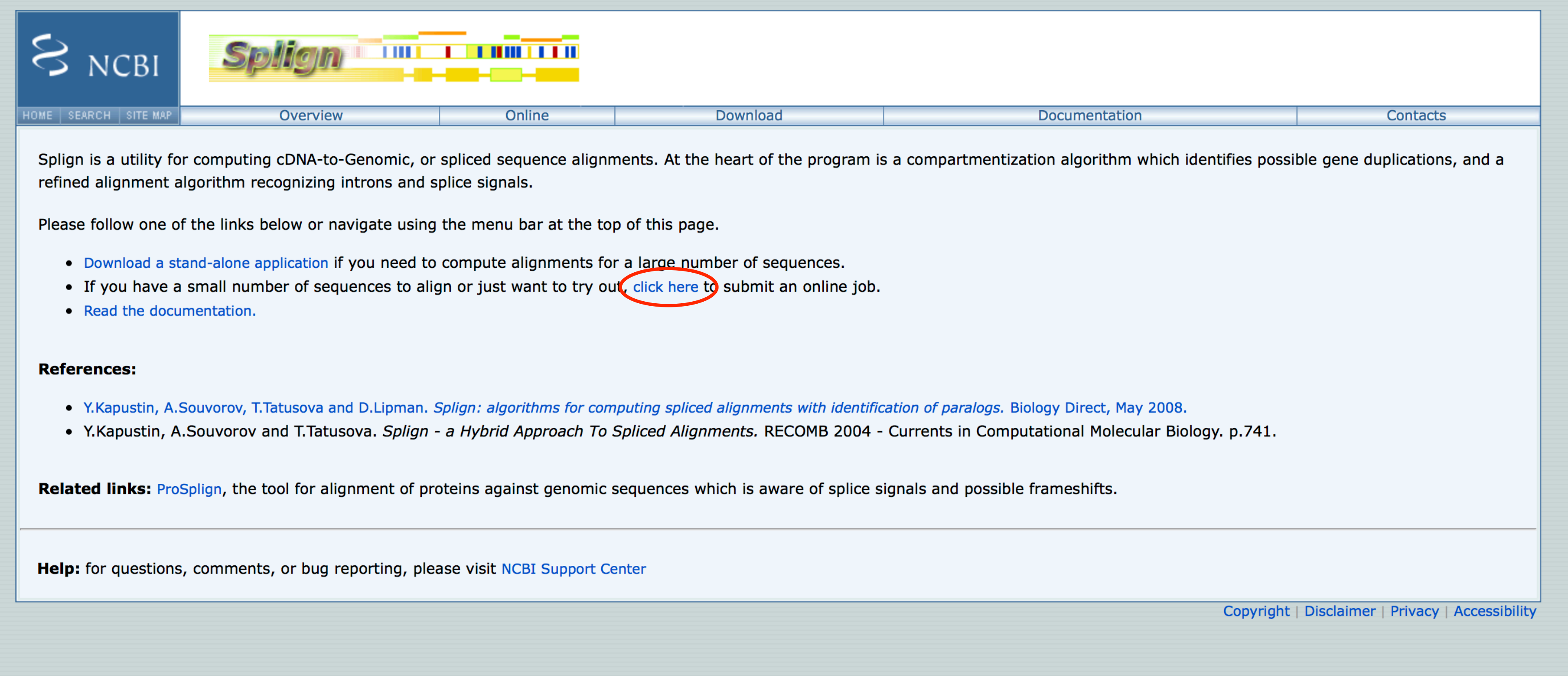
1. PubMed: 論文検索
2. Entrez：総合データベース
3. Google：何でも検索
4. Google Scholar：科学に関する検索
5. 特許データベース

ホモロジー検索：

1. BLAST [GenomeNET]
2. FASTA [GenomeNET]

配列解析：

1. Splign ：DNAとcDNAのマッチング
2. Blast align ：DNAとcDNAのマッチング
3. EMBOSS program" ：DNAとcDNAのマッチング
4. EMBOSS program" ：DNA配列をアミノ酸配列に変換
5. BLAST(NCBI) ：遺伝子のホモロジー検索
6. BLAST(DDBJ) ：遺伝子のホモロジー検索
7. Softberry/FGENESH ：遺伝子の予測
8. GeneMark ：遺伝子の予測
9. GENESCAN ：遺伝子の予測
10. clustalW： アライメント
11. InterProScan： ドメイン解析
12. PSORT： 細胞内局在性予測
13. GEO：遺伝子発現解析
14. Gene Ontology：用語を用いた検索
15. SSPN：スプライシングの予測



Splign is a utility for computing cDNA-to-Genomic, or spliced sequence alignments. At the heart of the program is a compartmentization algorithm which identifies possible gene duplications, and a refined alignment algorithm recognizing introns and splice signals.

Please follow one of the links below or navigate using the menu bar at the top of this page.

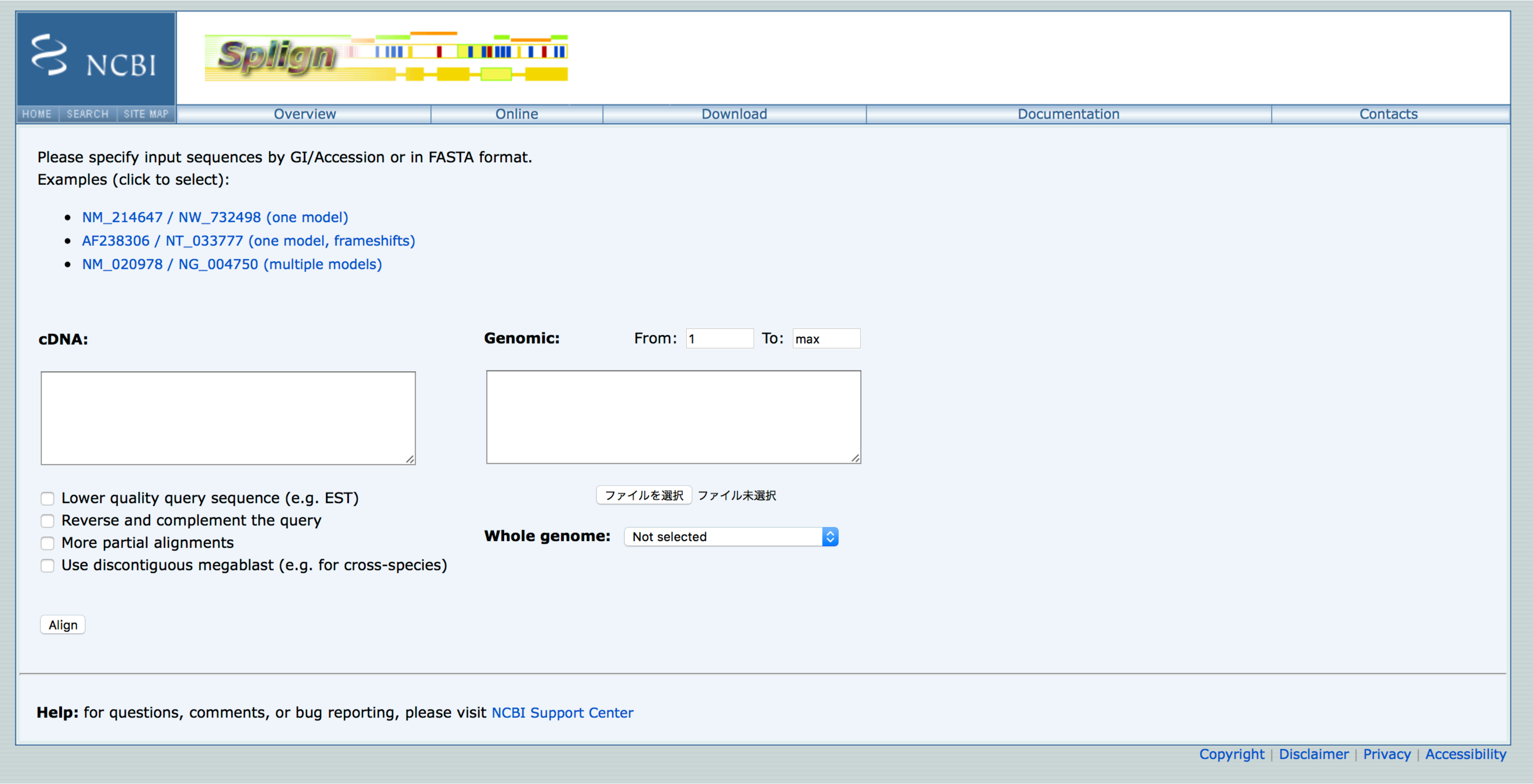
- Download a stand-alone application if you need to compute alignments for a large number of sequences.
- If you have a small number of sequences to align or just want to try out, click here to submit an online job.
- Read the documentation.

References:

- Y.Kapustin, A.Souvorov, T.Tatusova and D.Lipman. *Splign: algorithms for computing spliced alignments with identification of paralogs.* Biology Direct, May 2008.
- Y.Kapustin, A.Souvorov and T.Tatusova. *Splign - a Hybrid Approach To Spliced Alignments.* RECOMB 2004 - Currents in Computational Molecular Biology. p.741.

Related links: ProSplign, the tool for alignment of proteins against genomic sequences which is aware of splice signals and possible frameshifts.

Help: for questions, comments, or bug reporting, please visit NCBI Support Center



Please specify input sequences by GI/Accession or in FASTA format.
Examples (click to select):

- NM_214647 / NW_732498 (one model)
- AF238306 / NT_033777 (one model, frameshifts)
- NM_020978 / NG_004750 (multiple models)

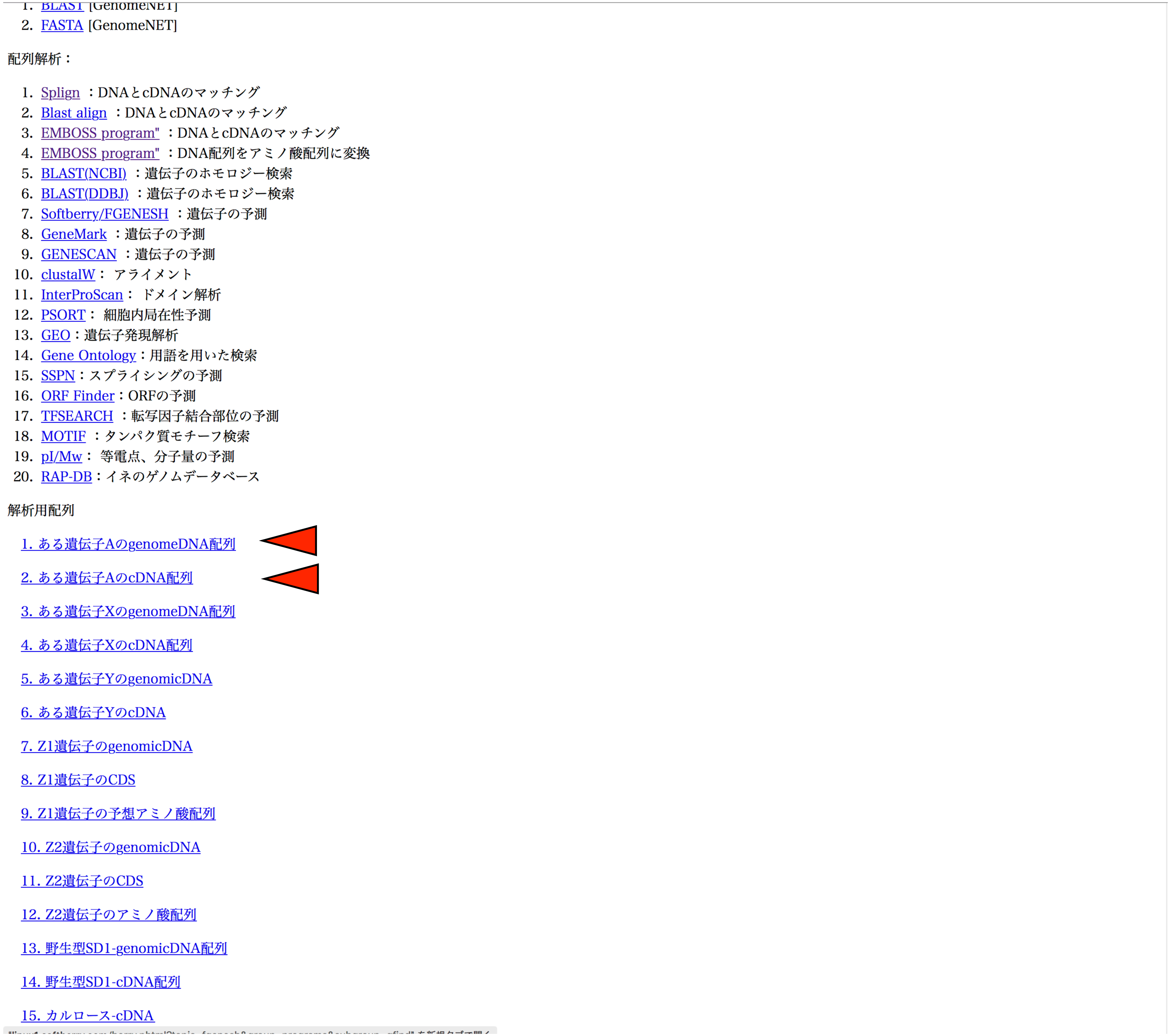
cDNA:

Genomic: From: 1 To: max

☐ Lower quality query sequence (e.g. EST)
☐ Reverse and complement the query
☐ More partial alignments
☐ Use discontiguous megablast (e.g. for cross-species)

Whole genome: Not selected

Help: for questions, comments, or bug reporting, please visit NCBI Support Center



1. BLAST [GenomeNET]
2. FASTA [GenomeNET]

配列解析：

1. Splign ：DNAとcDNAのマッチング
2. Blast align ：DNAとcDNAのマッチング
3. EMBOSS program" ：DNAとcDNAのマッチング
4. EMBOSS program" ：DNA配列をアミノ酸配列に変換
5. BLAST(NCBI) ：遺伝子のホモロジー検索
6. BLAST(DDBJ) ：遺伝子のホモロジー検索
7. Softberry/FGENESH ：遺伝子の予測
8. GeneMark ：遺伝子の予測
9. GENESCAN ：遺伝子の予測
10. clustalW： アライメント
11. InterProScan： ドメイン解析
12. PSORT： 細胞内局在性予測
13. GEO：遺伝子発現解析
14. Gene Ontology：用語を用いた検索
15. SSPN：スプライシングの予測
16. ORF Finder：ORFの予測
17. TFSEARCH ：転写因子結合部位の予測
18. MOTIF ：タンパク質モチーフ検索
19. pI/Mw： 等電点、分子量の予測
20. RAP-DB：イネのゲノムデータベース

解析用配列

1. ある遺伝子AのgenomeDNA配列

2. ある遺伝子AのcDNA配列

3. ある遺伝子XのgenomeDNA配列

4. ある遺伝子XのcDNA配列

5. ある遺伝子YのgenomicDNA

6. ある遺伝子YのcDNA

7. Z1遺伝子のgenomicDNA

8. Z1遺伝子のCDS

9. Z1遺伝子の予想アミノ酸配列

10. Z2遺伝子のgenomicDNA

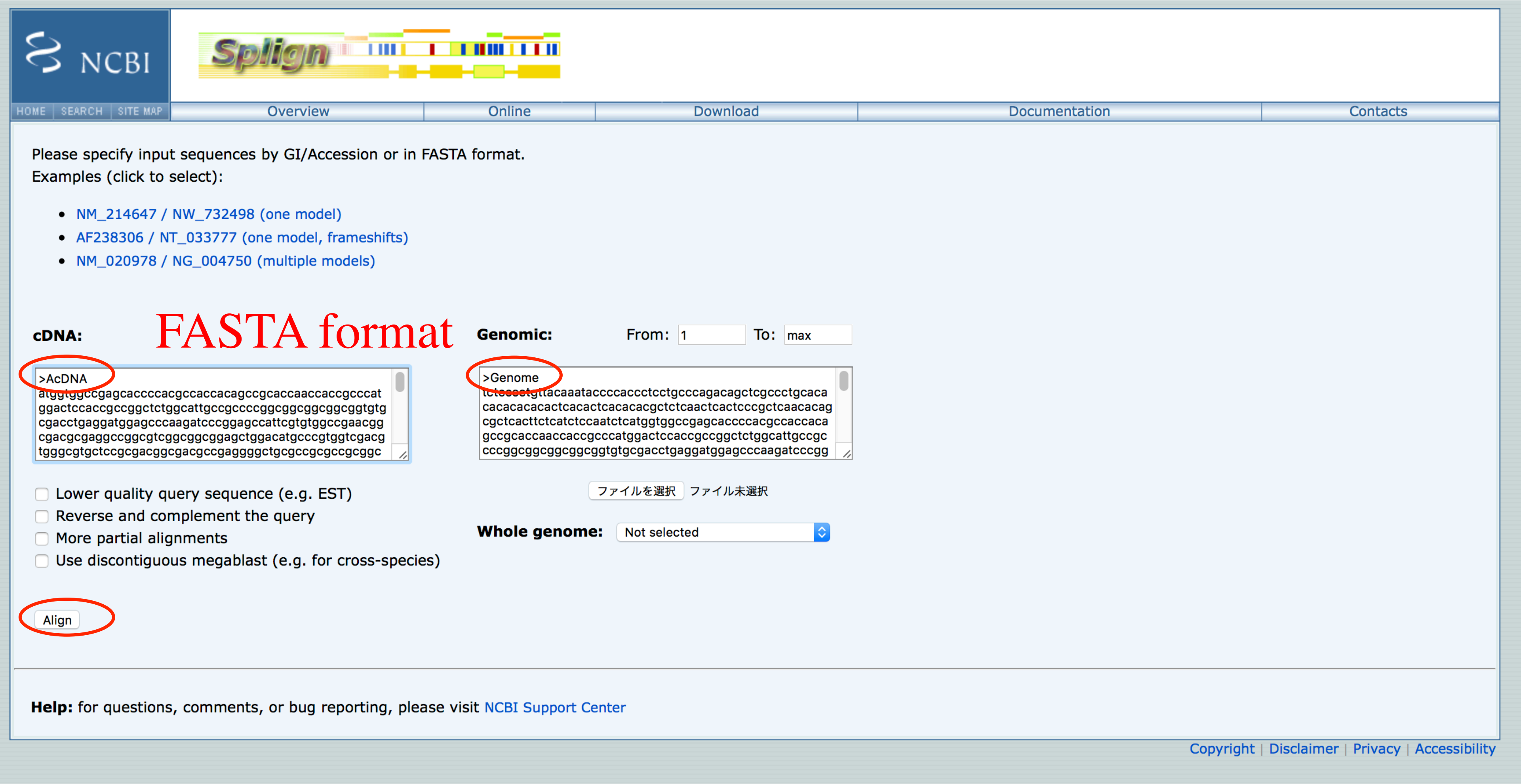
11. Z2遺伝子のCDS

12. Z2遺伝子のアミノ酸配列

13. 野生型SD1-genomicDNA配列

14. 野生型SD1-cDNA配列

15. カルロース-cDNA



Please specify input sequences by GI/Accession or in FASTA format.
Examples (click to select):

- NM_214647 / NW_732498 (one model)
- AF238306 / NT_033777 (one model, frameshifts)
- NM_020978 / NG_004750 (multiple models)

FASTA format

cDNA:

```
>AcDNA
atggtggccgagcaccccacgccaccacagccgcaccaaccaccgcccat
ggactccaccgccggctctggcattgccgccccggcggcggcggcggtgtg
cgacctgaggatggagcccaagatcccggagccattcgtgtggccgaacgg
cgacgcgaggccggcgtcggcggcggagctggacatgcccgtggtcgacg
tgggcgtgctccgcgacggcgacgccgaggggctgcgccgcgccgcggc
```

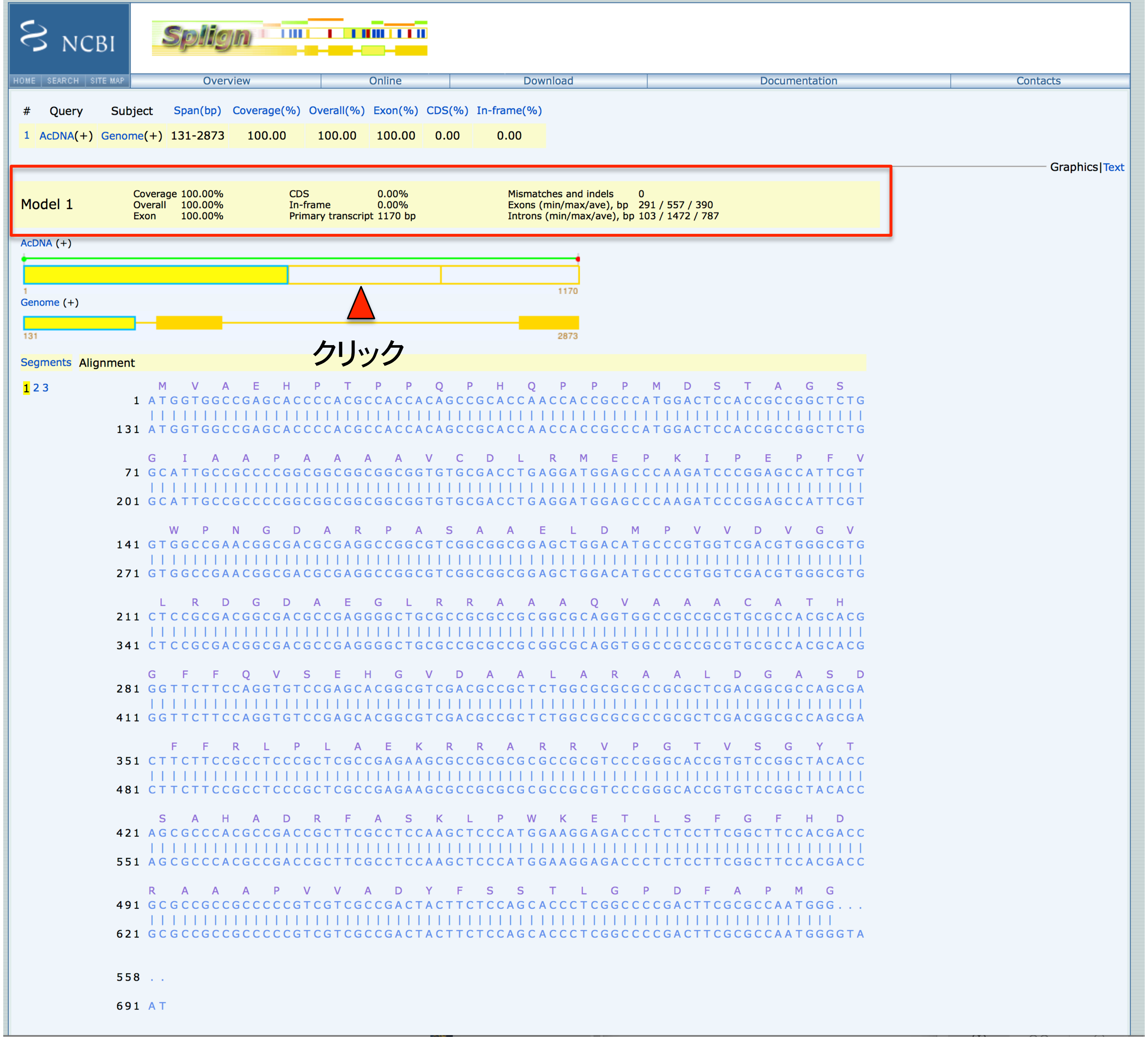
☐ Lower quality query sequence (e.g. EST)
☐ Reverse and complement the query
☐ More partial alignments
☐ Use discontiguous megablast (e.g. for cross-species)

Genomic: From: 1 To: max

```
>Genome
tctcccctgttacaaataccccaccctcctgcccagacagctcgccctgcaca
cacacacacactcacactcacacacgctctcaactcactcccgctcaacacag
cgctcacttctcatctccaatctcatggtggccgagcaccccacgccaccaca
gccgcaccaaccaccgcccatggactccaccgccggctctggcattgccgc
cccggcggcggcggcggtgtgcgacctgaggatggagcccaagatcccgg
```

Whole genome: Not selected

Help: for questions, comments, or bug reporting, please visit NCBI Support Center



#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	AcDNA(+)	Genome(+)	131-2873	100.00	100.00	100.00	0.00	0.00

Model 1

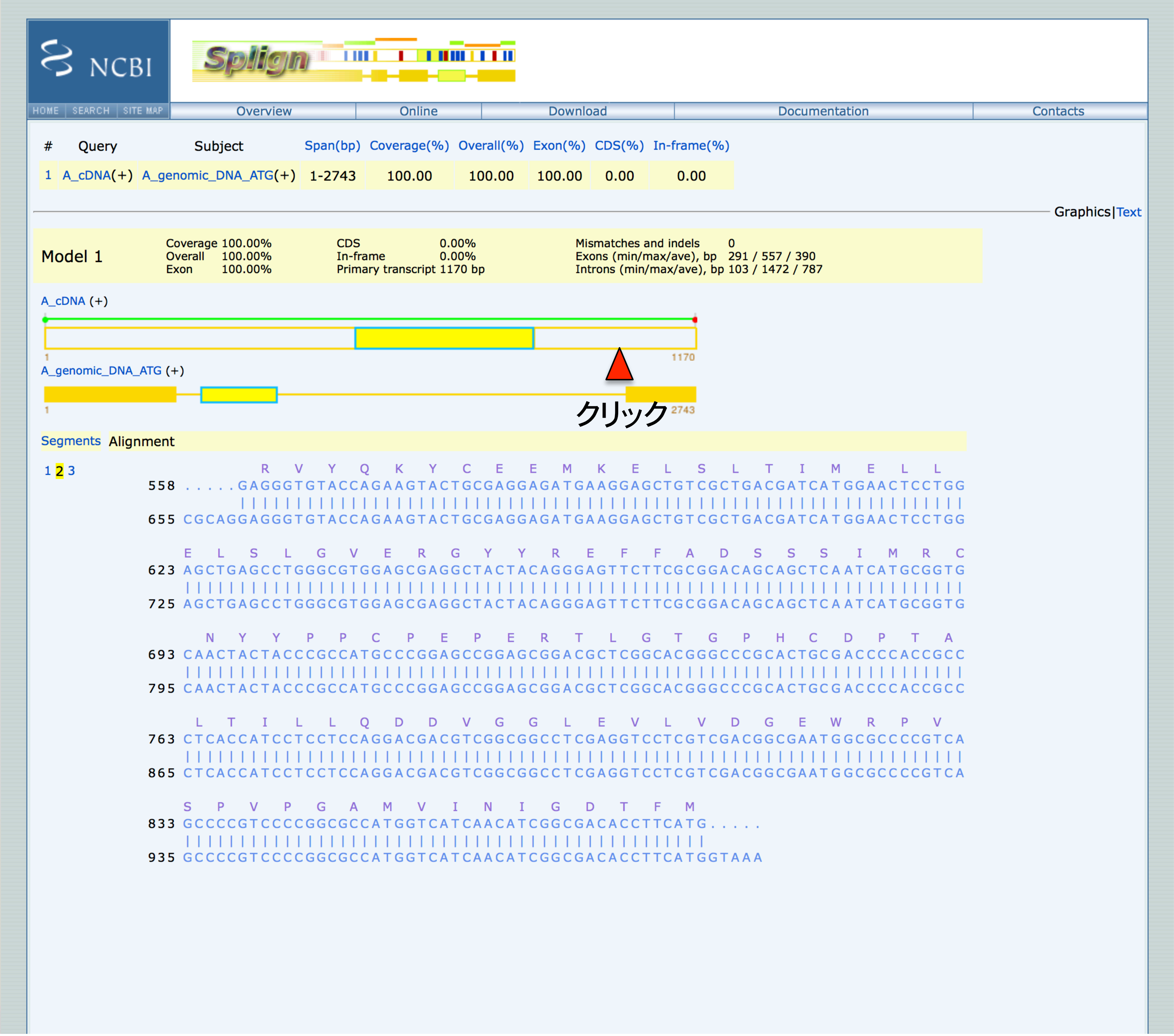
Coverage 100.00%
Overall 100.00%
Exon 100.00%

CDS 0.00%
In-frame 0.00%
Primary transcript 1170 bp

Mismatches and indels 0
Exons (min/max/ave), bp 291 / 557 / 390
Introns (min/max/ave), bp 103 / 1472 / 787

クリック

```
     M  V  A  E  H  P  T  P  P  Q  P  H  Q  P  P  P  M  D  S  T  A  G  S
  1 ATGGTGGCCGAGCACCCCACGCCACCACAGCCGCACCAACCACCGCCCATGGACTCCACCGCCGGCTCTG
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
131 ATGGTGGCCGAGCACCCCACGCCACCACAGCCGCACCAACCACCGCCCATGGACTCCACCGCCGGCTCTG

    G  I  A  A  P  A  A  A  A  V  C  D  L  R  M  E  P  K  I  P  E  P  F  V
 71 GCATTGCCGCCCCGGCGGCGGCGGCGGTGTGCGACCTGAGGATGGAGCCCAAGATCCCGGAGCCATTCGT
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
201 GCATTGCCGCCCCGGCGGCGGCGGCGGTGTGCGACCTGAGGATGGAGCCCAAGATCCCGGAGCCATTCGT

      W  P  N  G  D  A  R  P  A  S  A  A  E  L  D  M  P  V  V  D  V  G  V
141 GTGGCCGAACGGCGACGCGAGGCCGGCGTCGGCGGCGGAGCTGGACATGCCCGTGGTCGACGTGGGCGTG
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
271 GTGGCCGAACGGCGACGCGAGGCCGGCGTCGGCGGCGGAGCTGGACATGCCCGTGGTCGACGTGGGCGTG

     L  R  D  G  D  A  E  G  L  R  R  A  A  A  Q  V  A  A  A  C  A  T  H
211 CTCCGCGACGGCGACGCCGAGGGGCTGCGCCGCGCCGCGGCGCAGGTGGCCGCCGCGTGCGCCACGCACG
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
341 CTCCGCGACGGCGACGCCGAGGGGCTGCGCCGCGCCGCGGCGCAGGTGGCCGCCGCGTGCGCCACGCACG

    G  F  F  Q  V  S  E  H  G  V  D  A  A  L  A  R  A  A  L  D  G  A  S  D
281 GGTTCTTCCAGGTGTCCGAGCACGGCGTCGACGCCGCTCTGGCGCGCGCCGCGCTCGACGGCGCCAGCGA
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
411 GGTTCTTCCAGGTGTCCGAGCACGGCGTCGACGCCGCTCTGGCGCGCGCCGCGCTCGACGGCGCCAGCGA

      F  F  R  L  P  L  A  E  K  R  R  A  R  R  V  P  G  T  V  S  G  Y  T
351 CTTCTTCCGCCTCCCGCTCGCCGAGAAGCGCCGCGCGCGCCGCGTCCCGGGCACCGTGTCCGGCTACACC
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
481 CTTCTTCCGCCTCCCGCTCGCCGAGAAGCGCCGCGCGCGCCGCGTCCCGGGCACCGTGTCCGGCTACACC

     S  A  H  A  D  R  F  A  S  K  L  P  W  K  E  T  L  S  F  G  F  H  D
421 AGCGCCCACGCCGACCGCTTCGCCTCCAAGCTCCCATGGAAGGAGACCCTCTCCTTCGGCTTCCACGACC
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
551 AGCGCCCACGCCGACCGCTTCGCCTCCAAGCTCCCATGGAAGGAGACCCTCTCCTTCGGCTTCCACGACC

    R  A  A  A  P  V  V  A  D  Y  F  S  S  T  L  G  P  D  F  A  P  M  G
491 GCGCCGCCGCCCCCGTCGTCGCCGACTACTTCTCCAGCACCCTCGGCCCCGACTTCGCGCCAATGGG...
    |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
621 GCGCCGCCGCCCCCGTCGTCGCCGACTACTTCTCCAGCACCCTCGGCCCCGACTTCGCGCCAATGGGGTA


558 ..

691 AT
```



#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	A_cDNA(+)	A_genomic_DNA_ATG(+)	1-2743	100.00	100.00	100.00	0.00	0.00

Model 1

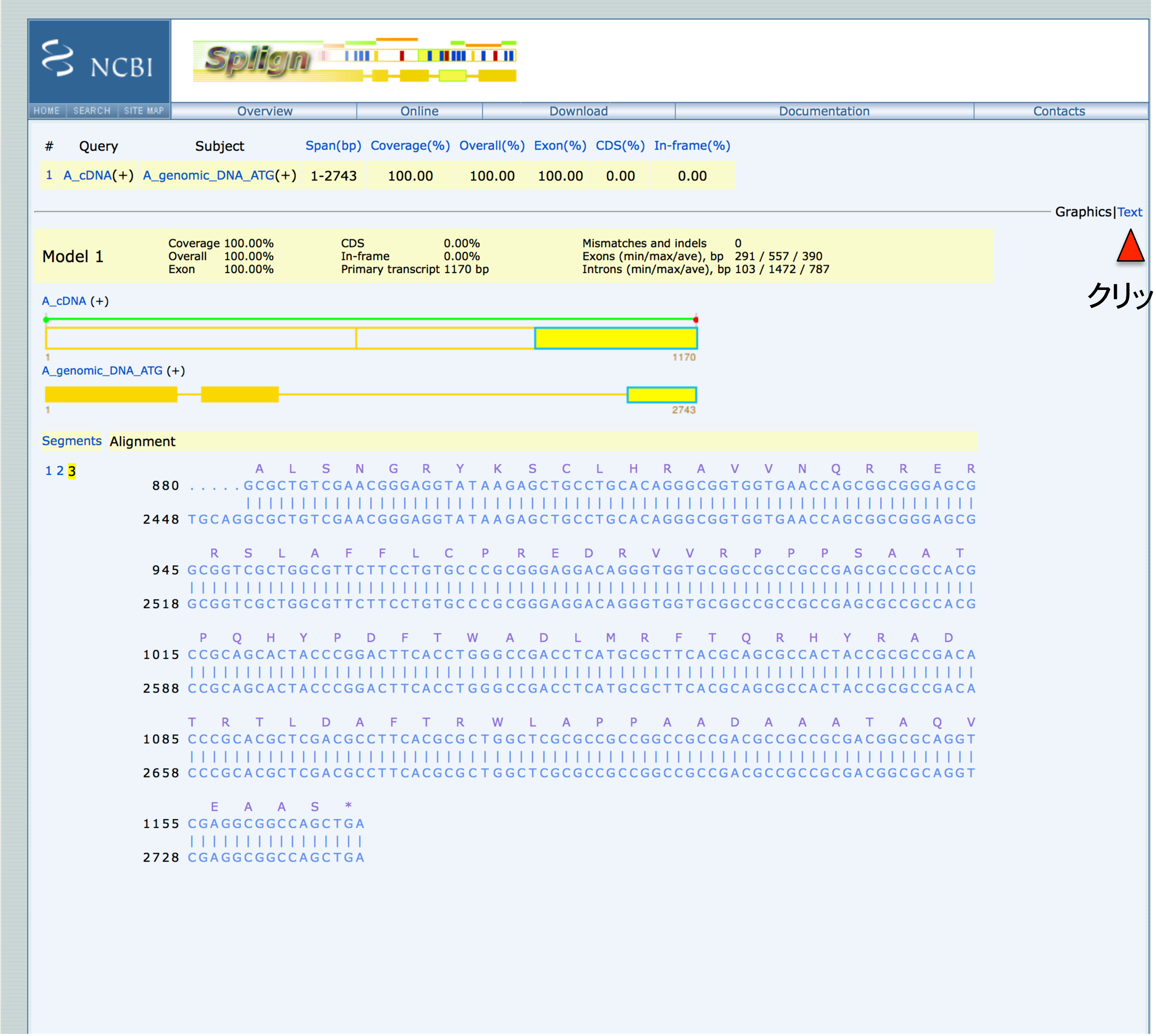
Coverage 100.00%
Overall 100.00%
Exon 100.00%

CDS 0.00%
In-frame 0.00%
Primary transcript 1170 bp

Mismatches and indels 0
Exons (min/max/ave), bp 291 / 557 / 390
Introns (min/max/ave), bp 103 / 1472 / 787

クリック

```
           R  V  Y  Q  K  Y  C  E  E  M  K  E  L  S  L  T  I  M  E  L  L
558 .....GAGGGTGTACCAGAAGTACTGCGAGGAGATGAAGGAGCTGTCGCTGACGATCATGGAACTCCTGG
         |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
655 CGCAGGAGGGTGTACCAGAAGTACTGCGAGGAGATGAAGGAGCTGTCGCTGACGATCATGGAACTCCTGG

    E  L  S  L  G  V  E  R  G  Y  Y  R  E  F  F  A  D  S  S  S  I  M  R  C
623 AGCTGAGCCTGGGCGTGGAGCGAGGCTACTACAGGGAGTTCTTCGCGGACAGCAGCTCAATCATGCGGTG
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
725 AGCTGAGCCTGGGCGTGGAGCGAGGCTACTACAGGGAGTTCTTCGCGGACAGCAGCTCAATCATGCGGTG

      N  Y  Y  P  P  C  P  E  P  E  R  T  L  G  T  G  P  H  C  D  P  T  A
693 CAACTACTACCCGCCATGCCCGGAGCCGGAGCGGACGCTCGGCACGGGCCCGCACTGCGACCCCACCGCC
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
795 CAACTACTACCCGCCATGCCCGGAGCCGGAGCGGACGCTCGGCACGGGCCCGCACTGCGACCCCACCGCC

     L  T  I  L  L  Q  D  D  V  G  G  L  E  V  L  V  D  G  E  W  R  P  V
763 CTCACCATCCTCCTCCAGGACGACGTCGGCGGCCTCGAGGTCCTCGTCGACGGCGAATGGCGCCCCGTCA
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
865 CTCACCATCCTCCTCCAGGACGACGTCGGCGGCCTCGAGGTCCTCGTCGACGGCGAATGGCGCCCCGTCA

    S  P  V  P  G  A  M  V  I  N  I  G  D  T  F  M
833 GCCCCGTCCCCGGCGCCATGGTCATCAACATCGGCGACACCTTCATG.....
    |||||||||||||||||||||||||||||||||||||||||||||||
935 GCCCCGTCCCCGGCGCCATGGTCATCAACATCGGCGACACCTTCATGGTAAA
```



#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	A_cDNA(+)	A_genomic_DNA_ATG(+)	1-2743	100.00	100.00	100.00	0.00	0.00

Graphics|Text

Model 1

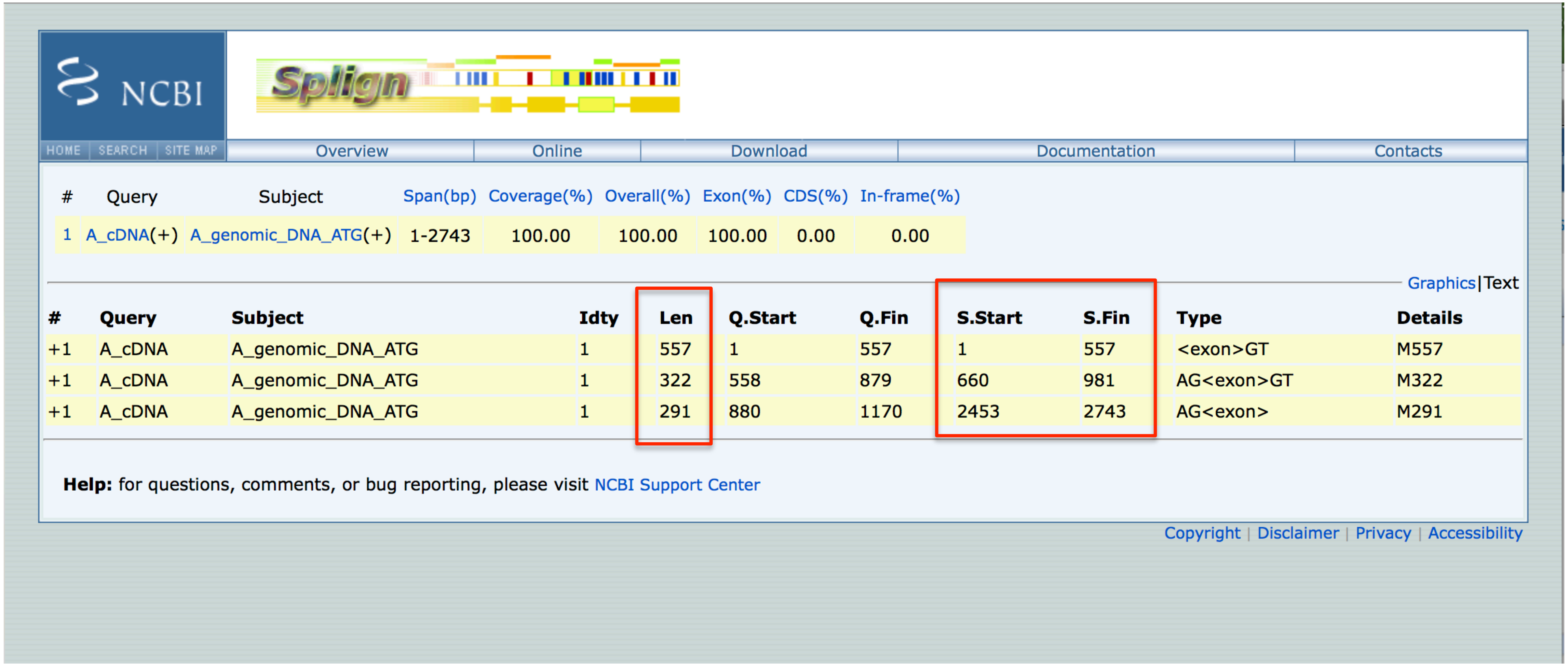
Coverage 100.00%
Overall 100.00%
Exon 100.00%

CDS 0.00%
In-frame 0.00%
Primary transcript 1170 bp

Mismatches and indels 0
Exons (min/max/ave), bp 291 / 557 / 390
Introns (min/max/ave), bp 103 / 1472 / 787

```
           A  L  S  N  G  R  Y  K  S  C  L  H  R  A  V  V  N  Q  R  R  E  R
 880 .....GCGCTGTCGAACGGGAGGTATAAGAGCTGCCTGCACAGGGCGGTGGTGAACCAGCGGCGGGAGCG
          |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
2448 TGCAGGCGCTGTCGAACGGGAGGTATAAGAGCTGCCTGCACAGGGCGGTGGTGAACCAGCGGCGGGAGCG

       R  S  L  A  F  F  L  C  P  R  E  D  R  V  V  R  P  P  P  S  A  A  T
 945 GCGGTCGCTGGCGTTCTTCCTGTGCCCGCGGGAGGACAGGGTGGTGCGGCCGCCGCCGAGCGCCGCCACG
     ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
2518 GCGGTCGCTGGCGTTCTTCCTGTGCCCGCGGGAGGACAGGGTGGTGCGGCCGCCGCCGAGCGCCGCCACG

      P  Q  H  Y  P  D  F  T  W  A  D  L  M  R  F  T  Q  R  H  Y  R  A  D
1015 CCGCAGCACTACCCGGACTTCACCTGGGCCGACCTCATGCGCTTCACGCAGCGCCACTACCGCGCCGACA
     ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
2588 CCGCAGCACTACCCGGACTTCACCTGGGCCGACCTCATGCGCTTCACGCAGCGCCACTACCGCGCCGACA

     T  R  T  L  D  A  F  T  R  W  L  A  P  P  A  A  D  A  A  A  T  A  Q  V
1085 CCCGCACGCTCGACGCCTTCACGCGCTGGCTCGCGCCGCCGGCCGCCGACGCCGCCGCGACGGCGCAGGT
     ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
2658 CCCGCACGCTCGACGCCTTCACGCGCTGGCTCGCGCCGCCGGCCGCCGACGCCGCCGCGACGGCGCAGGT

       E  A  A  S  *
1155 CGAGGCGGCCAGCTGA
     ||||||||||||||||
2728 CGAGGCGGCCAGCTGA
```

クリック

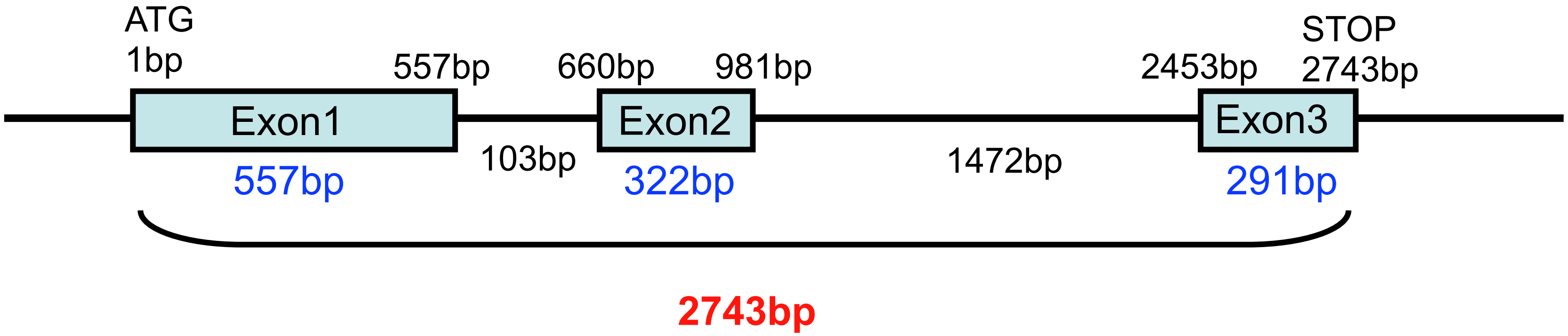


#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	A_cDNA(+)	A_genomic_DNA_ATG(+)	1-2743	100.00	100.00	100.00	0.00	0.00

#	Query	Subject	Idty	Len	Q.Start	Q.Fin	S.Start	S.Fin	Type	Details
+1	A_cDNA	A_genomic_DNA_ATG	1	557	1	557	1	557	<exon>GT	M557
+1	A_cDNA	A_genomic_DNA_ATG	1	322	558	879	660	981	AG<exon>GT	M322
+1	A_cDNA	A_genomic_DNA_ATG	1	291	880	1170	2453	2743	AG<exon>	M291

Help: for questions, comments, or bug reporting, please visit NCBI Support Center

遺伝子構造



生物情報工学II (担当　芦苅）遺伝子の配列解析

第６回目-1（11/1）遺伝子の予測と機能解析

-2（11/1）演習

第７回目-1（11/8）DNA配列から遺伝子を探す

-2（11/8）DNA、アミノ酸のアラインメント、系統樹解析

-3（11/8）演習

第８回目-1（11/15）Genetyxを用いた解析

-2（11/15）演習

第９回目-1（11/22）KEGGを用いた解析

連鎖解析

出席と宿題の提出　提出方法

リンク集

データベース検索：

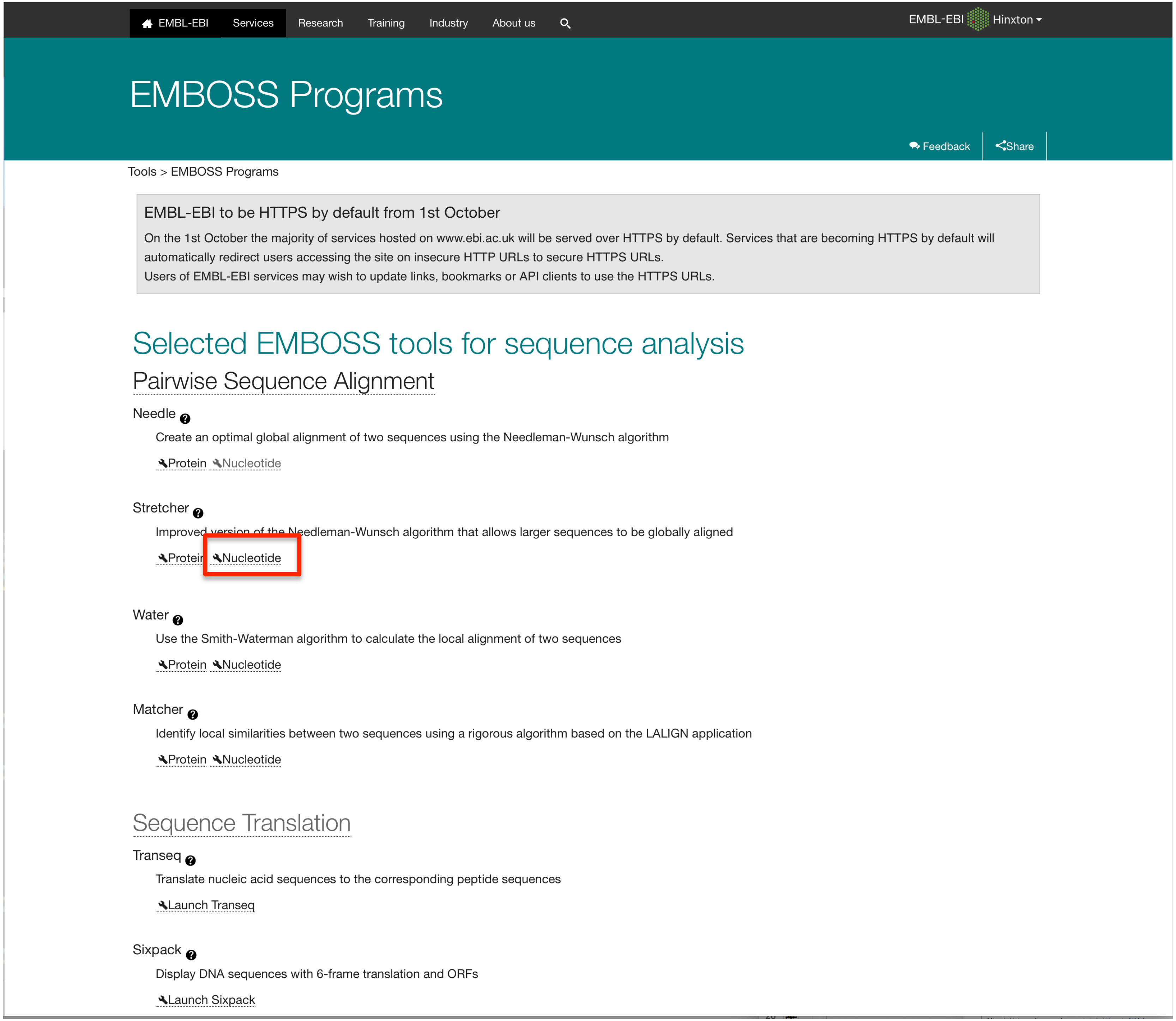
1. PubMed: 論文検索
2. Entrez：総合データベース
3. Google：何でも検索
4. Google Scholar：科学に関する検索
5. 特許データベース

ホモロジー検索：

1. BLAST [GenomeNET]
2. FASTA [GenomeNET]

配列解析：

1. Splign ：DNAとcDNAのマッチング
2. Blast align ：DNAとcDNAのマッチング
3. EMBOSS program" ：DNAとcDNAのマッチング
4. EMBOSS program" ：DNA配列をアミノ酸配列に変換
5. BLAST(NCBI) ：遺伝子のホモロジー検索
6. BLAST(DDBJ) ：遺伝子のホモロジー検索
7. Softberry/FGENESH ：遺伝子の予測
8. GeneMark ：遺伝子の予測
9. GENESCAN ：遺伝子の予測
10. clustalW： アライメント
11. InterProScan： ドメイン解析
12. PSORT： 細胞内局在性予測
13. GEO：遺伝子発現解析
14. Gene Ontology：用語を用いた検索
15. SSPN：スプライシングの予測



EMBOSS Programs

Tools > EMBOSS Programs

EMBL-EBI to be HTTPS by default from 1st October

On the 1st October the majority of services hosted on www.ebi.ac.uk will be served over HTTPS by default. Services that are becoming HTTPS by default will automatically redirect users accessing the site on insecure HTTP URLs to secure HTTPS URLs.
Users of EMBL-EBI services may wish to update links, bookmarks or API clients to use the HTTPS URLs.

Selected EMBOSS tools for sequence analysis

Pairwise Sequence Alignment

Needle

Create an optimal global alignment of two sequences using the Needleman-Wunsch algorithm

Protein Nucleotide

Stretcher

Improved version of the Needleman-Wunsch algorithm that allows larger sequences to be globally aligned

Protein Nucleotide

Water

Use the Smith-Waterman algorithm to calculate the local alignment of two sequences

Protein Nucleotide

Matcher

Identify local similarities between two sequences using a rigorous algorithm based on the LALIGN application

Protein Nucleotide

Sequence Translation

Transeq

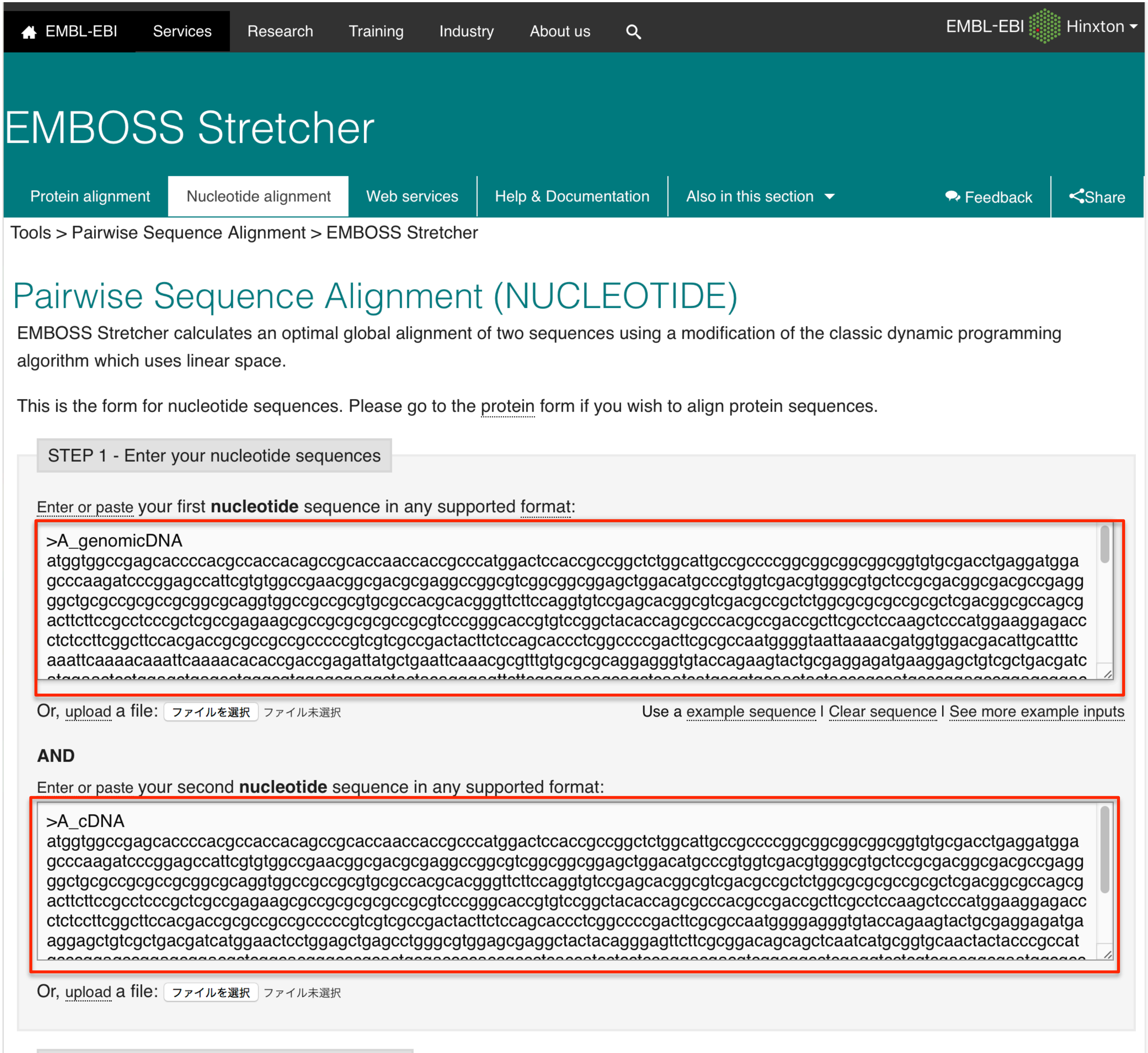
Translate nucleic acid sequences to the corresponding peptide sequences

Launch Transeq

Sixpack

Display DNA sequences with 6-frame translation and ORFs

Launch Sixpack



Pairwise Sequence Alignment (NUCLEOTIDE)

EMBOSS Stretcher calculates an optimal global alignment of two sequences using a modification of the classic dynamic programming algorithm which uses linear space.

This is the form for nucleotide sequences. Please go to the protein form if you wish to align protein sequences.

STEP 1 - Enter your nucleotide sequences

Enter or paste your first **nucleotide** sequence in any supported format:

```
>A_genomicDNA
atggtggccgagcaccccacgccaccacagccgcaccaaccaccgcccatggactccaccgccggctctggcattgccgccccggcggcggcggcggtgtgcgacctgaggatgga
gcccaagatcccggagccattcgtgtggccgaacggcgacgcgaggccggcgtcggcggcggagctggacatgcccgtggtcgacgtgggcgtgctccgcgacggcgacgccgagg
ggctgcgccgcgccgcggcgcaggtggccgccgcgtgcgccacgcacgggttcttccaggtgtccgagcacggcgtcgacgccgctctggcgcgcgccgcgctcgacggcgccagcg
acttcttccgcctcccgctcgccgagaagcgccgcgcgcgcgccgcgtcccgggcaccgtgtccggctacaccagcgcccacgccgaccgcttcgcctccaagctcccatggaaggagacc
ctctccttcggcttccacgaccgcgccgccgcccccgtcgtcgccgactacttctccagcaccctcggccccgacttcgcgccaatggggtaattaaaacgatggtggacgacattgcatttc
aaattcaaaacaaattcaaaacacaccgaccgagattatgctgaattcaaacgcgtttgtgcgcgcaggagggtgtaccagaagtactgcgaggagatgaaggagctgtcgctgacgatc
```

Or, upload a file: ファイルを選択 ファイル未選択

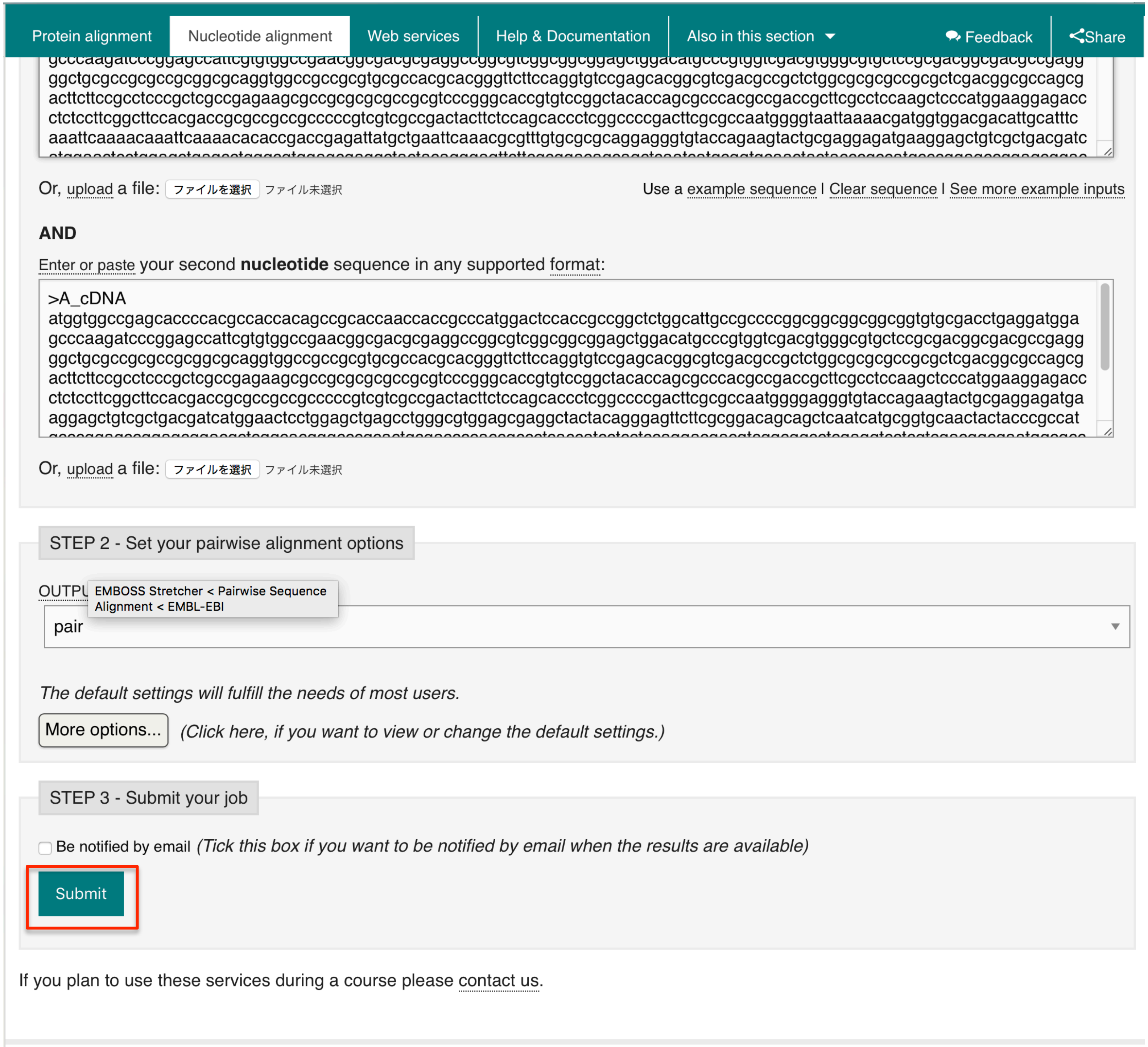
Use a example sequence | Clear sequence | See more example inputs

AND

Enter or paste your second **nucleotide** sequence in any supported format:

```
>A_cDNA
atggtggccgagcaccccacgccaccacagccgcaccaaccaccgcccatggactccaccgccggctctggcattgccgccccggcggcggcggcggtgtgcgacctgaggatgga
gcccaagatcccggagccattcgtgtggccgaacggcgacgcgaggccggcgtcggcggcggagctggacatgcccgtggtcgacgtgggcgtgctccgcgacggcgacgccgagg
ggctgcgccgcgccgcggcgcaggtggccgccgcgtgcgccacgcacgggttcttccaggtgtccgagcacggcgtcgacgccgctctggcgcgcgccgcgctcgacggcgccagcg
acttcttccgcctcccgctcgccgagaagcgccgcgcgcgcgccgcgtcccgggcaccgtgtccggctacaccagcgcccacgccgaccgcttcgcctccaagctcccatggaaggagacc
ctctccttcggcttccacgaccgcgccgccgcccccgtcgtcgccgactacttctccagcaccctcggccccgacttcgcgccaatggggagggtgtaccagaagtactgcgaggagatga
aggagctgtcgctgacgatcatggaactcctggagctgagcctgggcgtggagcgaggctactacagggagttcttcgcggacagcagctcaatcatgcggtgcaactactacccgccat
```

Or, upload a file: ファイルを選択 ファイル未選択



```
gcccaagatcccggagccattcgtgtggccgaacggcgacgcgaggccggcgtcggcggcggagctggacatgcccgtggtcgacgtgggcgtgctccgcgacggcgacgccgagg
ggctgcgccgcgccgcggcgcaggtggccgccgcgtgcgccacgcacgggttcttccaggtgtccgagcacggcgtcgacgccgctctggcgcgcgccgcgctcgacggcgccagcg
acttcttccgcctcccgctcgccgagaagcgccgcgcgcgccgcgtcccgggcaccgtgtccggctacaccagcgcccacgccgaccgcttcgcctccaagctcccatggaaggagacc
ctctccttcggcttccacgaccgcgccgccgcccccgtcgtcgccgactacttctccagcaccctcggccccgacttcgcgccaatggggtaattaaaacgatggtggacgacattgcatttc
aaattcaaaacaaattcaaaacacaccgaccgagattatgctgaattcaaacgcgtttgtgcgcgcaggagggtgtaccagaagtactgcgaggagatgaaggagctgtcgctgacgatc
```

Or, upload a file: ファイルを選択 ファイル未選択

Use a example sequence | Clear sequence | See more example inputs

AND

Enter or paste your second **nucleotide** sequence in any supported format:

```
>A_cDNA
atggtggccgagcaccccacgccaccacagccgcaccaaccaccgcccatggactccaccgccggctctggcattgccgccccggcggcggcggcggtgtgcgacctgaggatgga
gcccaagatcccggagccattcgtgtggccgaacggcgacgcgaggccggcgtcggcggcggagctggacatgcccgtggtcgacgtgggcgtgctccgcgacggcgacgccgagg
ggctgcgccgcgccgcggcgcaggtggccgccgcgtgcgccacgcacgggttcttccaggtgtccgagcacggcgtcgacgccgctctggcgcgcgccgcgctcgacggcgccagcg
acttcttccgcctcccgctcgccgagaagcgccgcgcgcgccgcgtcccgggcaccgtgtccggctacaccagcgcccacgccgaccgcttcgcctccaagctcccatggaaggagacc
ctctccttcggcttccacgaccgcgccgccgcccccgtcgtcgccgactacttctccagcaccctcggccccgacttcgcgccaatggggagggtgtaccagaagtactgcgaggagatga
aggagctgtcgctgacgatcatggaactcctggagctgagcctgggcgtggagcgaggctactacagggagttcttcgcggacagcagctcaatcatgcggtgcaactactacccgccat
```

Or, upload a file: ファイルを選択 ファイル未選択

STEP 2 - Set your pairwise alignment options

OUTPU

pair

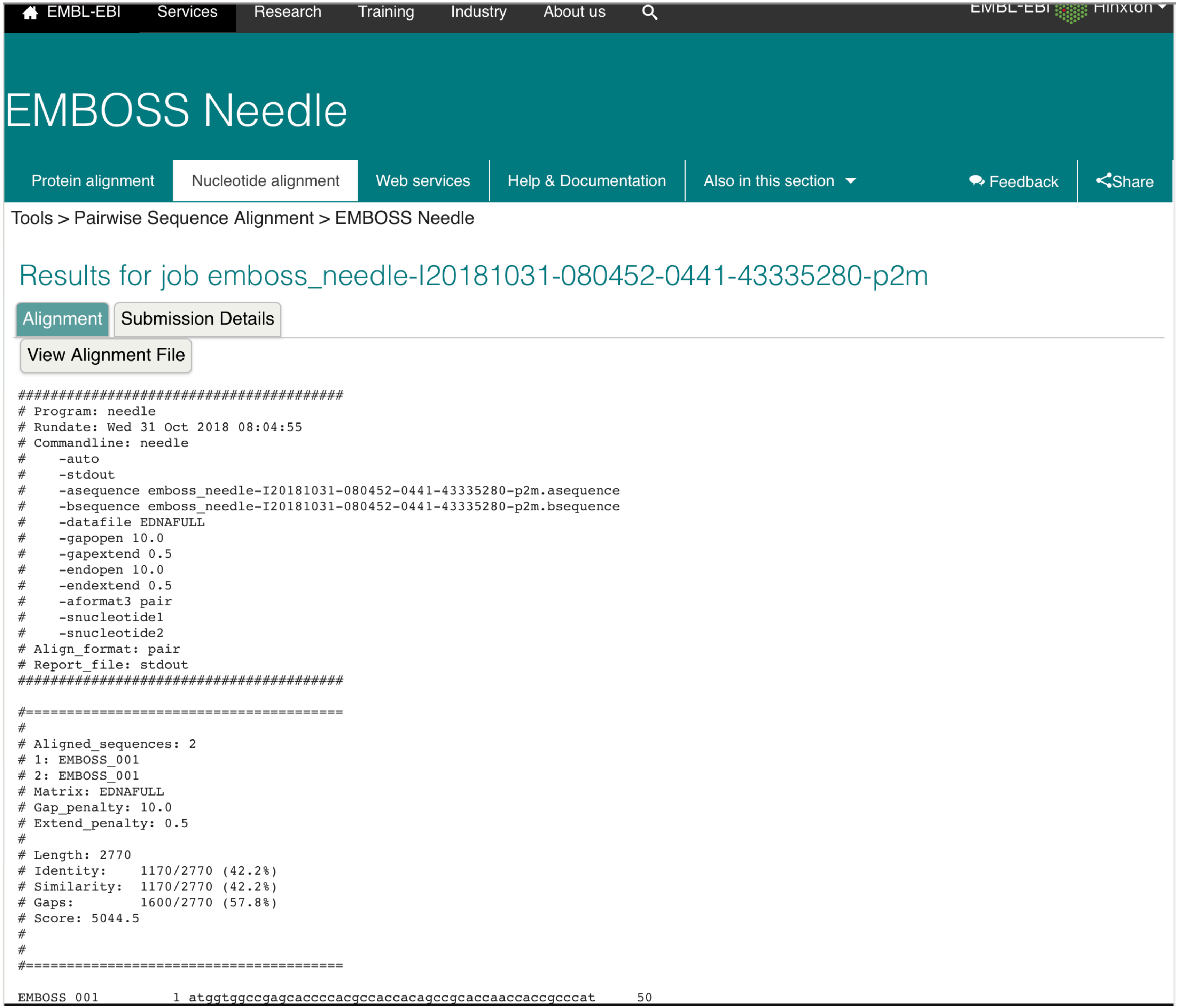
The default settings will fulfill the needs of most users.

(Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

If you plan to use these services during a course please contact us.

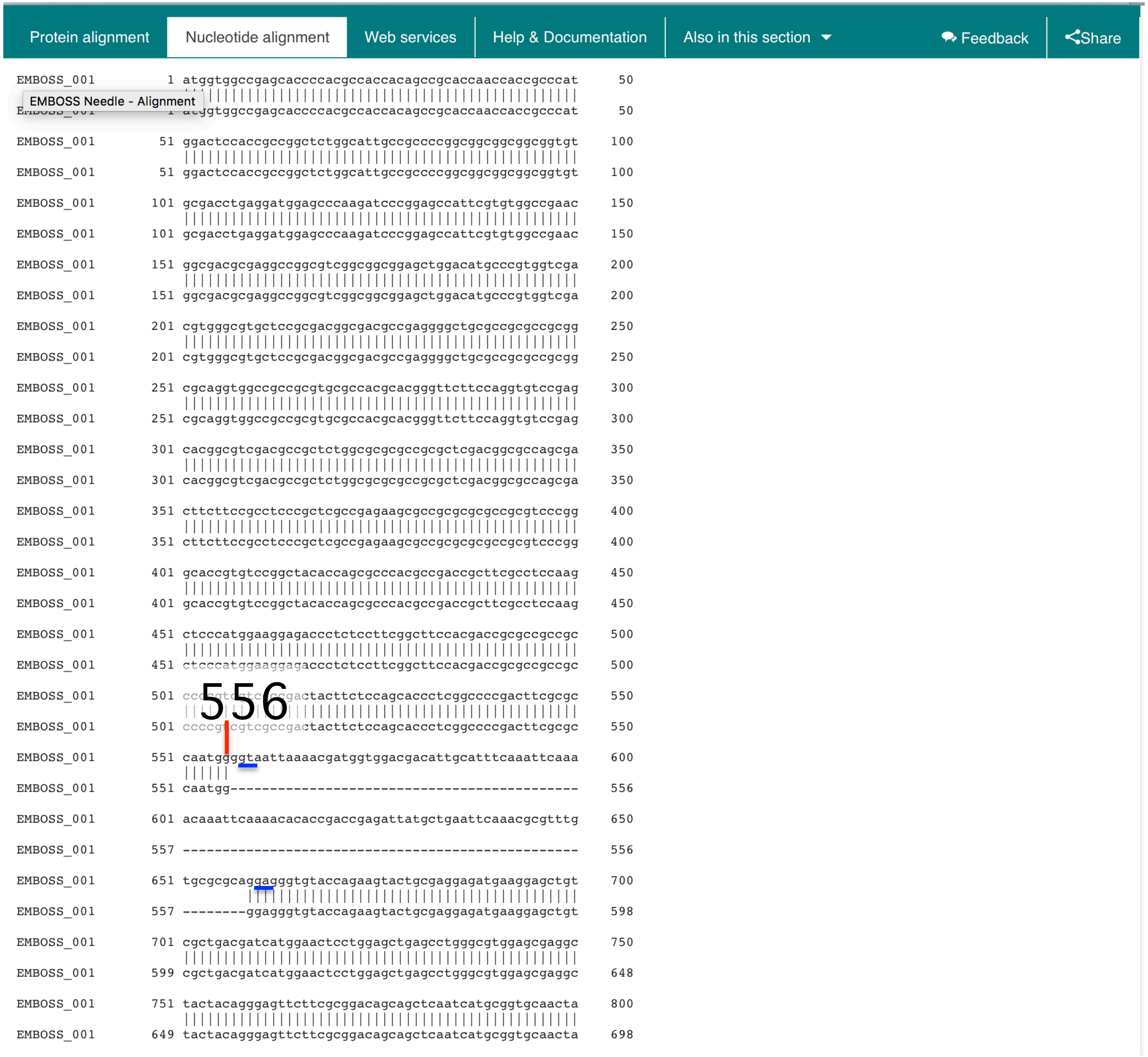


EMBOSS Needle

Tools > Pairwise Sequence Alignment > EMBOSS Needle

Results for job emboss_needle-I20181031-080452-0441-43335280-p2m

```
########################################
# Program: needle
# Rundate: Wed 31 Oct 2018 08:04:55
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20181031-080452-0441-43335280-p2m.asequence
#    -bsequence emboss_needle-I20181031-080452-0441-43335280-p2m.bsequence
#    -datafile EDNAFULL
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -snucleotide1
#    -snucleotide2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 2770
# Identity:    1170/2770 (42.2%)
# Similarity:  1170/2770 (42.2%)
# Gaps:        1600/2770 (57.8%)
# Score: 5044.5
#
#
#=======================================

EMBOSS 001         1 atggtggccgagcaccccacgccaccacagccgcaccaaccaccgcccat     50
```



```
EMBOSS_001         1 atggtggccgagcaccccacgccaccacagccgcaccaaccaccgcccat     50
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001         1 atggtggccgagcaccccacgccaccacagccgcaccaaccaccgcccat     50

EMBOSS_001        51 ggactccaccgccggctctggcattgccgccccggcggcggcggcggtgt    100
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001        51 ggactccaccgccggctctggcattgccgccccggcggcggcggcggtgt    100

EMBOSS_001       101 gcgacctgaggatggagcccaagatcccggagccattcgtgtggccgaac    150
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       101 gcgacctgaggatggagcccaagatcccggagccattcgtgtggccgaac    150

EMBOSS_001       151 ggcgacgcgaggccggcgtcggcggcggagctggacatgcccgtggtcga    200
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       151 ggcgacgcgaggccggcgtcggcggcggagctggacatgcccgtggtcga    200

EMBOSS_001       201 cgtgggcgtgctccgcgacggcgacgccgaggggctgcgccgcgccgcgg    250
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       201 cgtgggcgtgctccgcgacggcgacgccgaggggctgcgccgcgccgcgg    250

EMBOSS_001       251 cgcaggtggccgccgcgtgcgccacgcacgggttcttccaggtgtccgag    300
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       251 cgcaggtggccgccgcgtgcgccacgcacgggttcttccaggtgtccgag    300

EMBOSS_001       301 cacggcgtcgacgccgctctggcgcgcgccgcgctcgacggcgccagcga    350
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       301 cacggcgtcgacgccgctctggcgcgcgccgcgctcgacggcgccagcga    350

EMBOSS_001       351 cttcttccgcctcccgctcgccgagaagcgccgcgcgcgccgcgtcccgg    400
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       351 cttcttccgcctcccgctcgccgagaagcgccgcgcgcgccgcgtcccgg    400

EMBOSS_001       401 gcaccgtgtccggctacaccagcgcccacgccgaccgcttcgcctccaag    450
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       401 gcaccgtgtccggctacaccagcgcccacgccgaccgcttcgcctccaag    450

EMBOSS_001       451 ctcccatggaaggagaccctctccttcggcttccacgaccgcgccgccgc    500
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       451 ctcccatggaaggagaccctctccttcggcttccacgaccgcgccgccgc    500

EMBOSS_001       501 ccccgtcgtcgccgactacttctccagcaccctcggccccgacttcgcgc    550
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       501 ccccgtcgtcgccgactacttctccagcaccctcggccccgacttcgcgc    550

EMBOSS_001       551 caatgggtaattaaaacgatggtggacgacattgcatttcaaattcaaa    600
                     ||||||
EMBOSS_001       551 caatgg--------------------------------------------    556

EMBOSS_001       601 acaaattcaaaacacaccgaccgagattatgctgaattcaaacgcgtttg    650

EMBOSS_001       557 --------------------------------------------------    556

EMBOSS_001       651 tgcgcgcaggagggtgtaccagaagtactgcgaggagatgaaggagctgt    700
                             ||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       557 --------ggagggtgtaccagaagtactgcgaggagatgaaggagctgt    598

EMBOSS_001       701 cgctgacgatcatggaactcctggagctgagcctgggcgtggagcgaggc    750
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       599 cgctgacgatcatggaactcctggagctgagcctgggcgtggagcgaggc    648

EMBOSS_001       751 tactacagggagttcttcgcggacagcagctcaatcatgcggtgcaacta    800
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       649 tactacagggagttcttcgcggacagcagctcaatcatgcggtgcaacta    698
```

エキソンとイントロンの境界

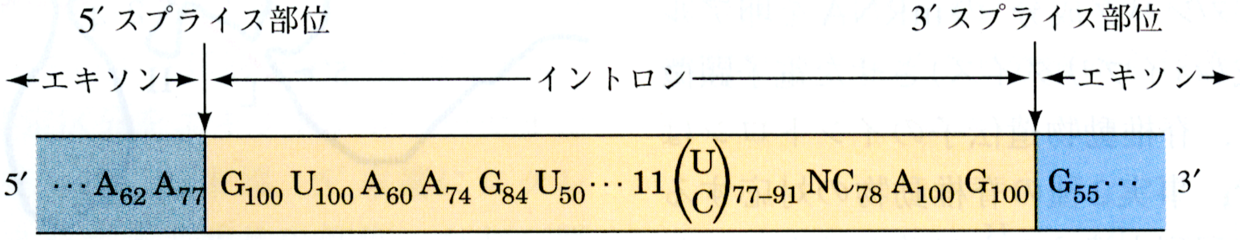
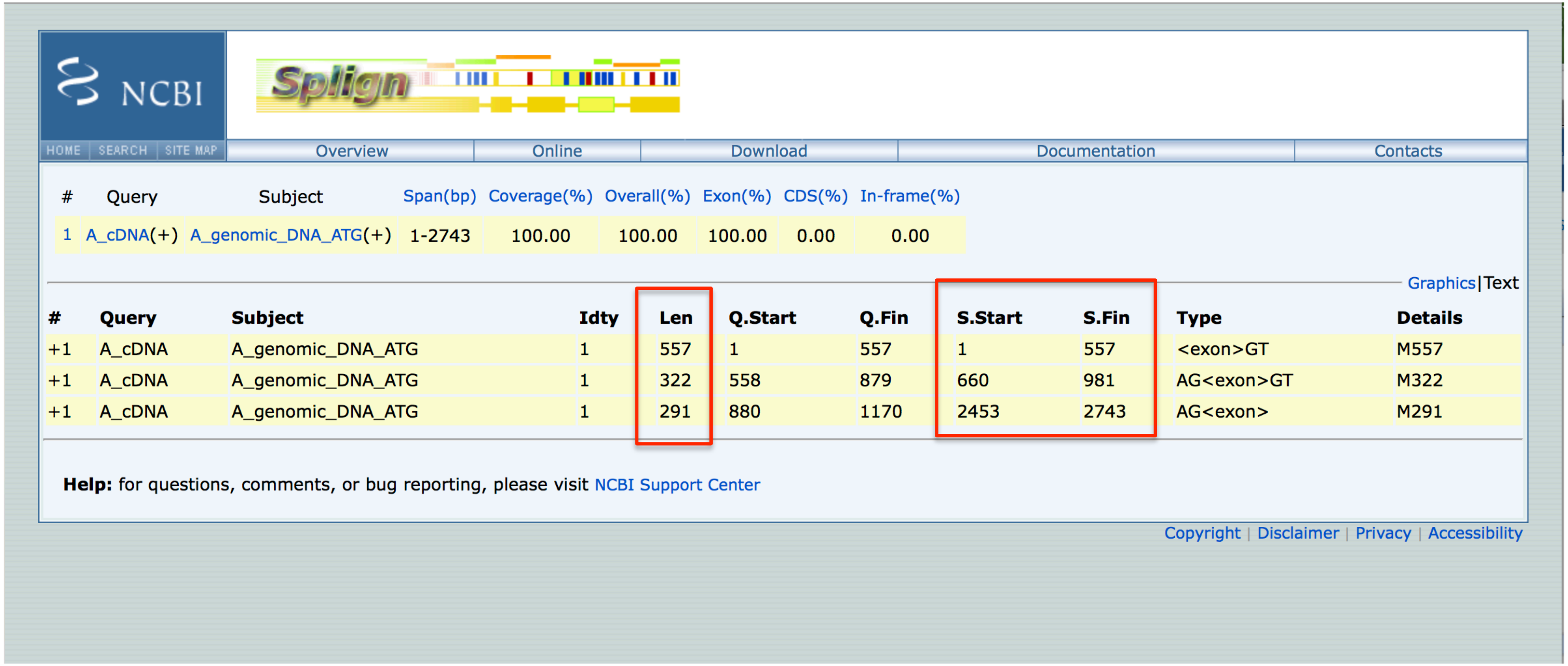


図 25・19　真核細胞mRNA前駆体のエキソン・イントロン接合部にあるコンセンサス配列　各塩基の添え数字はその塩基の出現頻度(%)．3′スプライス部位のすぐ上流にピリミジンに富む11 nt区域がある［R.A. Padgett, P.J. Grabowski, M.M. Konarska, S.S. Seiler, P.A. Sharp, *Annu. Rev. Biochem*. **55**, 1123 (1986) のデータに基づく］



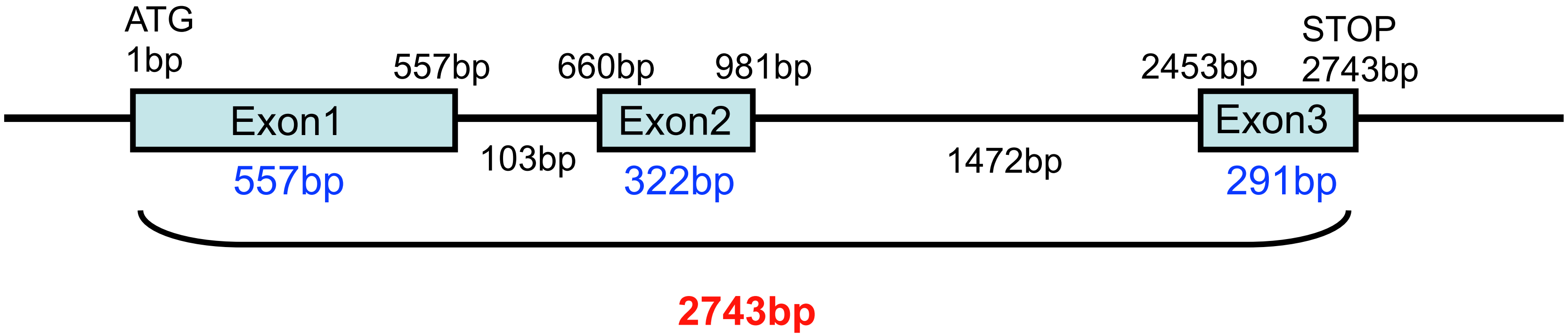
#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	A_cDNA(+)	A_genomic_DNA_ATG(+)	1-2743	100.00	100.00	100.00	0.00	0.00

#	Query	Subject	Idty	Len	Q.Start	Q.Fin	S.Start	S.Fin	Type	Details
+1	A_cDNA	A_genomic_DNA_ATG	1	557	1	557	1	557	<exon>GT	M557
+1	A_cDNA	A_genomic_DNA_ATG	1	322	558	879	660	981	AG<exon>GT	M322
+1	A_cDNA	A_genomic_DNA_ATG	1	291	880	1170	2453	2743	AG<exon>	M291

Help: for questions, comments, or bug reporting, please visit NCBI Support Center

Copyright | Disclaimer | Privacy | Accessibility

遺伝子構造



生物情報工学II (担当　芦苅）遺伝子の配列解析

第６回目-1（11/1）遺伝子の予測と機能解析

-2（11/1）演習

第７回目-1（11/8）DNA配列から遺伝子を探す

-2（11/8）DNA、アミノ酸のアラインメント、系統樹解析

-3（11/8）演習

第８回目-1（11/15）Genetyxを用いた解析

-2（11/15）演習

第９回目-1（11/22）KEGGを用いた解析

連鎖解析

出席と宿題の提出　提出方法

リンク集

データベース検索：

1. PubMed: 論文検索
2. Entrez：総合データベース
3. Google：何でも検索
4. Google Scholar：科学に関する検索
5. 特許データベース

ホモロジー検索：

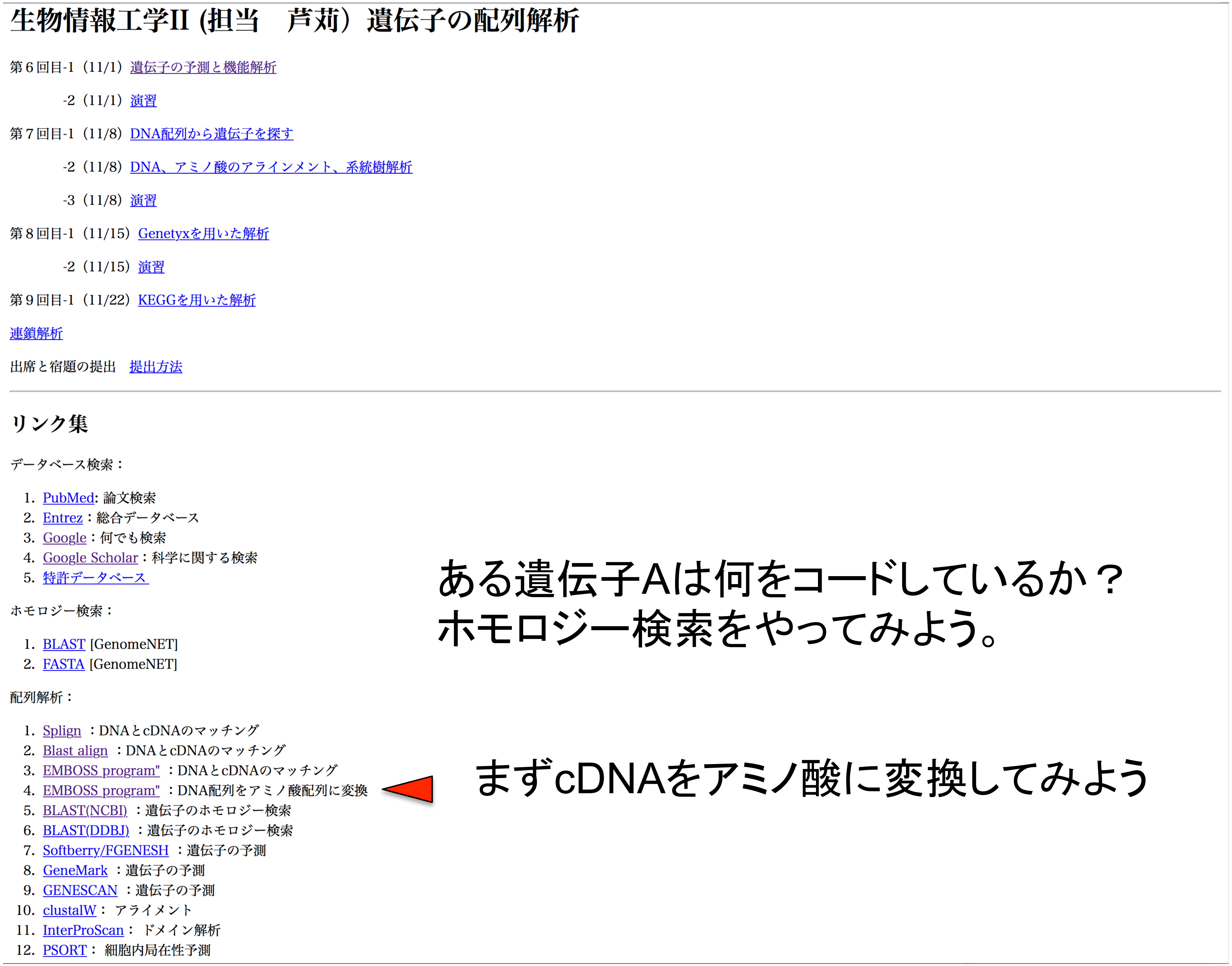
1. BLAST [GenomeNET]
2. FASTA [GenomeNET]

配列解析：

1. Splign ：DNAとcDNAのマッチング
2. Blast align ：DNAとcDNAのマッチング
3. EMBOSS program" ：DNAとcDNAのマッチング
4. EMBOSS program" ：DNA配列をアミノ酸配列に変換
5. BLAST(NCBI) ：遺伝子のホモロジー検索
6. BLAST(DDBJ) ：遺伝子のホモロジー検索
7. Softberry/FGENESH ：遺伝子の予測
8. GeneMark ：遺伝子の予測
9. GENESCAN ：遺伝子の予測
10. clustalW： アライメント
11. InterProScan： ドメイン解析
12. PSORT： 細胞内局在性予測
13. GEO：遺伝子発現解析
14. Gene Ontology：用語を用いた検索
15. SSPN：スプライシングの予測

エクソン-イントロンを正確に知りたいとき

ただ配列比較をしたい場合



生物情報工学II (担当　芦苅）遺伝子の配列解析

第６回目-1（11/1）遺伝子の予測と機能解析

-2（11/1）演習

第７回目-1（11/8）DNA配列から遺伝子を探す

-2（11/8）DNA、アミノ酸のアラインメント、系統樹解析

-3（11/8）演習

第８回目-1（11/15）Genetyxを用いた解析

-2（11/15）演習

第９回目-1（11/22）KEGGを用いた解析

連鎖解析

出席と宿題の提出　提出方法

リンク集

データベース検索：

1. PubMed: 論文検索
2. Entrez：総合データベース
3. Google：何でも検索
4. Google Scholar：科学に関する検索
5. 特許データベース

ホモロジー検索：

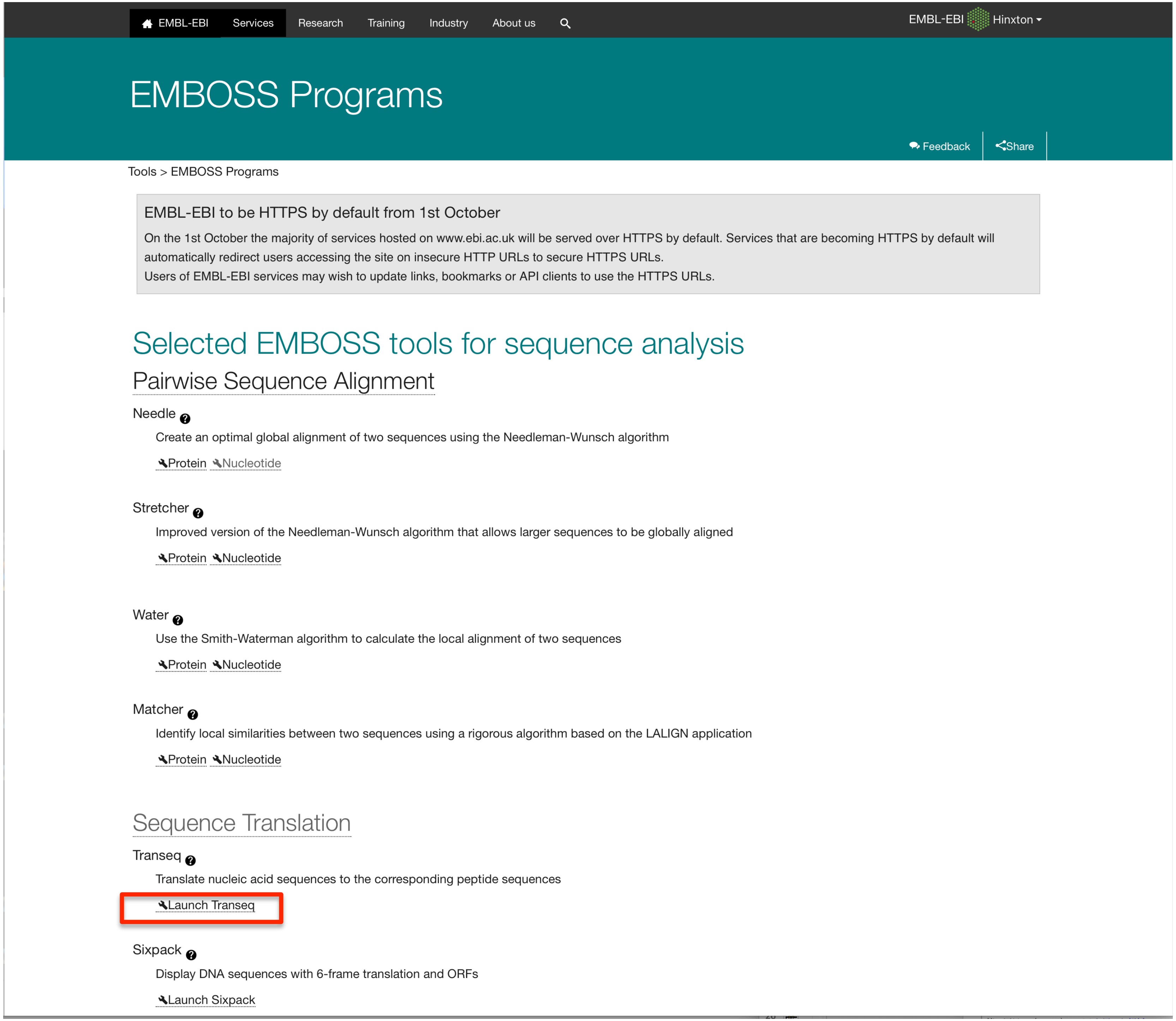
1. BLAST [GenomeNET]
2. FASTA [GenomeNET]

配列解析：

1. Splign ：DNAとcDNAのマッチング
2. Blast align ：DNAとcDNAのマッチング
3. EMBOSS program" ：DNAとcDNAのマッチング
4. EMBOSS program" ：DNA配列をアミノ酸配列に変換
5. BLAST(NCBI) ：遺伝子のホモロジー検索
6. BLAST(DDBJ) ：遺伝子のホモロジー検索
7. Softberry/FGENESH ：遺伝子の予測
8. GeneMark ：遺伝子の予測
9. GENESCAN ：遺伝子の予測
10. clustalW： アライメント
11. InterProScan： ドメイン解析
12. PSORT： 細胞内局在性予測

ある遺伝子Aは何をコードしているか？
ホモロジー検索をやってみよう。

まずcDNAをアミノ酸に変換してみよう



EMBOSS Programs

Tools > EMBOSS Programs

EMBL-EBI to be HTTPS by default from 1st October

On the 1st October the majority of services hosted on www.ebi.ac.uk will be served over HTTPS by default. Services that are becoming HTTPS by default will automatically redirect users accessing the site on insecure HTTP URLs to secure HTTPS URLs.
Users of EMBL-EBI services may wish to update links, bookmarks or API clients to use the HTTPS URLs.

Selected EMBOSS tools for sequence analysis

Pairwise Sequence Alignment

Needle

Create an optimal global alignment of two sequences using the Needleman-Wunsch algorithm

Stretcher

Improved version of the Needleman-Wunsch algorithm that allows larger sequences to be globally aligned

Water

Use the Smith-Waterman algorithm to calculate the local alignment of two sequences

Matcher

Identify local similarities between two sequences using a rigorous algorithm based on the LALIGN application

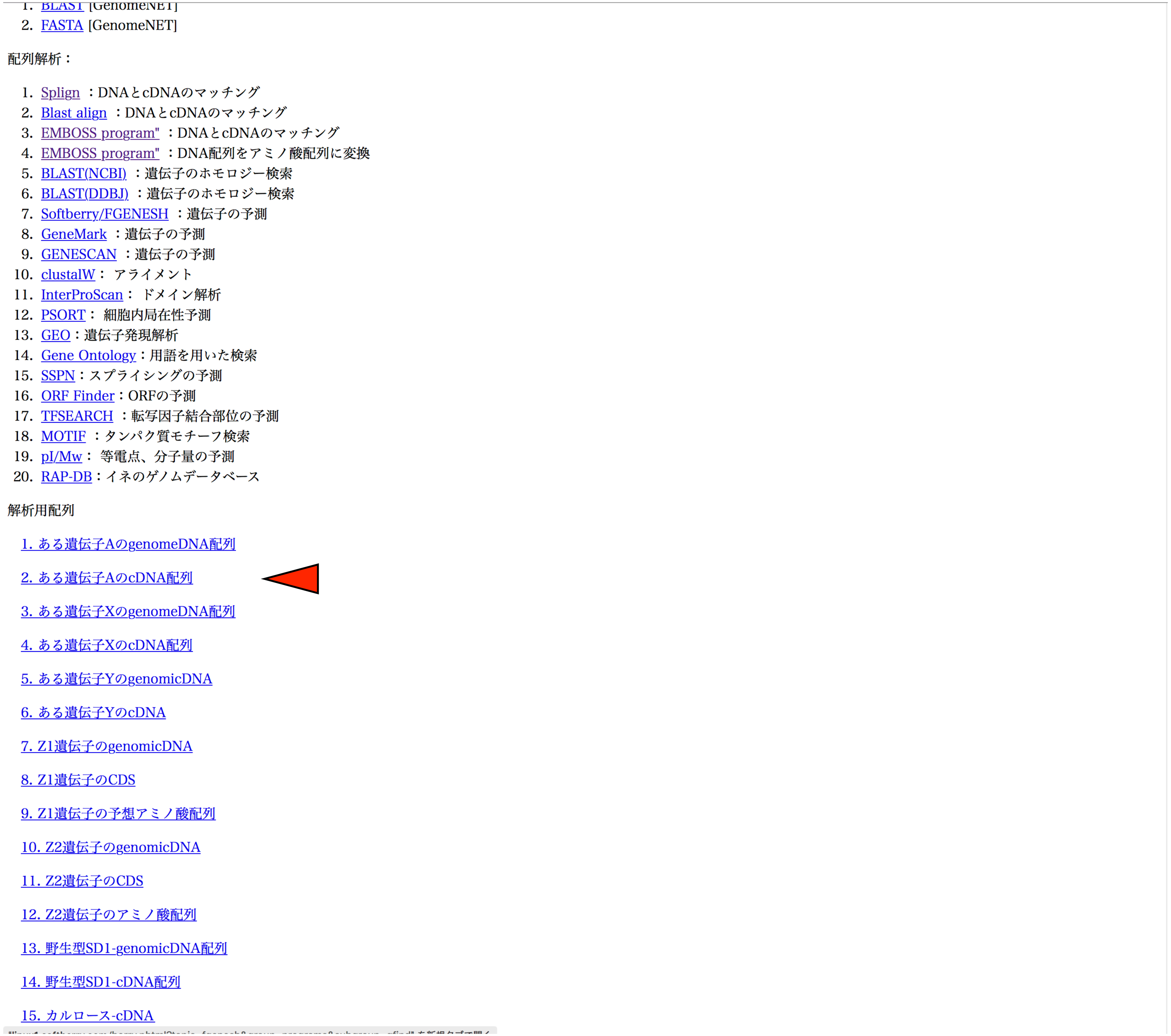
Sequence Translation

Transeq

Translate nucleic acid sequences to the corresponding peptide sequences

Sixpack

Display DNA sequences with 6-frame translation and ORFs



1. BLAST [GenomeNET]
2. FASTA [GenomeNET]

配列解析：

1. Splign ：DNAとcDNAのマッチング
2. Blast align ：DNAとcDNAのマッチング
3. EMBOSS program" ：DNAとcDNAのマッチング
4. EMBOSS program" ：DNA配列をアミノ酸配列に変換
5. BLAST(NCBI) ：遺伝子のホモロジー検索
6. BLAST(DDBJ) ：遺伝子のホモロジー検索
7. Softberry/FGENESH ：遺伝子の予測
8. GeneMark ：遺伝子の予測
9. GENESCAN ：遺伝子の予測
10. clustalW： アライメント
11. InterProScan： ドメイン解析
12. PSORT： 細胞内局在性予測
13. GEO：遺伝子発現解析
14. Gene Ontology：用語を用いた検索
15. SSPN：スプライシングの予測
16. ORF Finder：ORFの予測
17. TFSEARCH ：転写因子結合部位の予測
18. MOTIF ：タンパク質モチーフ検索
19. pI/Mw： 等電点、分子量の予測
20. RAP-DB：イネのゲノムデータベース

解析用配列

1. ある遺伝子AのgenomeDNA配列

2. ある遺伝子AのcDNA配列

3. ある遺伝子XのgenomeDNA配列

4. ある遺伝子XのcDNA配列

5. ある遺伝子YのgenomicDNA

6. ある遺伝子YのcDNA

7. Z1遺伝子のgenomicDNA

8. Z1遺伝子のCDS

9. Z1遺伝子の予想アミノ酸配列

10. Z2遺伝子のgenomicDNA

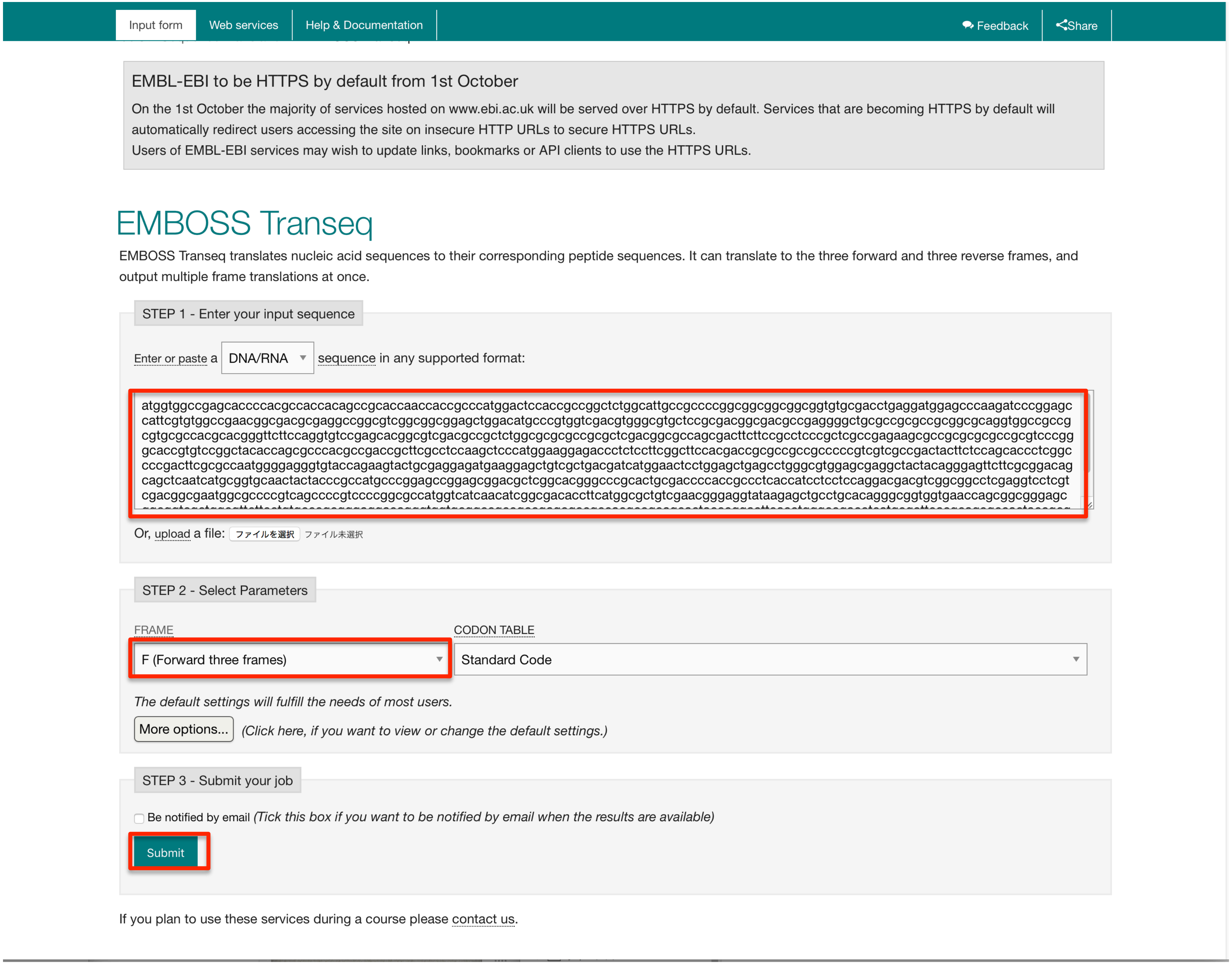
11. Z2遺伝子のCDS

12. Z2遺伝子のアミノ酸配列

13. 野生型SD1-genomicDNA配列

14. 野生型SD1-cDNA配列

15. カルロース-cDNA



EMBL-EBI to be HTTPS by default from 1st October

On the 1st October the majority of services hosted on www.ebi.ac.uk will be served over HTTPS by default. Services that are becoming HTTPS by default will automatically redirect users accessing the site on insecure HTTP URLs to secure HTTPS URLs.
Users of EMBL-EBI services may wish to update links, bookmarks or API clients to use the HTTPS URLs.

EMBOSS Transeq

EMBOSS Transeq translates nucleic acid sequences to their corresponding peptide sequences. It can translate to the three forward and three reverse frames, and output multiple frame translations at once.

STEP 1 - Enter your input sequence

Enter or paste a DNA/RNA sequence in any supported format:

```
atggtggccgagcaccccacgccaccacagccgcaccaaccaccgcccatggactccaccgccggctctggcattgccgccccggcggcggcggcggtgtgcgacctgaggatggagcccaagatcccggagc
cattcgtgtggccgaacggcgacgcgaggccggcgtcggcggcggagctggacatgcccgtggtcgacgtgggcgtgctccgcgacggcgacgccgaggggctgcgccgcgccgcggcgcaggtggccgccg
cgtgcgccacgcacgggttcttccaggtgtccgagcacggcgtcgacgccgctctggcgcgcgccgcgctcgacggcgccagcgacttcttccgcctcccgctcgccgagaagcgccgcgcgcgccgcgtcccgg
gcaccgtgtccggctacaccagcgcccacgccgaccgcttcgcctccaagctcccatggaaggagaccctctccttcggcttccacgaccgcgccgccgcccccgtcgtcgccgactacttctccagcaccctcggc
cccgacttcgcgccaatggggagggtgtaccagaagtactgcgaggagatgaaggagctgtcgctgacgatcatggaactcctggagctgagcctgggcgtggagcgaggctactacagggagttcttcgcggacag
cagctcaatcatgcggtgcaactactacccgccatgcccggagccggagcggacgctcggcacgggcccgcactgcgaccccaccgccctcaccatcctcctccaggacgacgtcggcggcctcgaggtcctcgt
cgacggcgaatggcgccccgtcagccccgtccccggcgccatggtcatcaacatcggcgacaccttcatggcgctgtcgaacgggaggtataagagctgcctgcacagggcggtggtgaaccagcggcgggagc
```

Or, upload a file: ファイルを選択 ファイル未選択

STEP 2 - Select Parameters

FRAME: F (Forward three frames)

CODON TABLE: Standard Code

The default settings will fulfill the needs of most users.

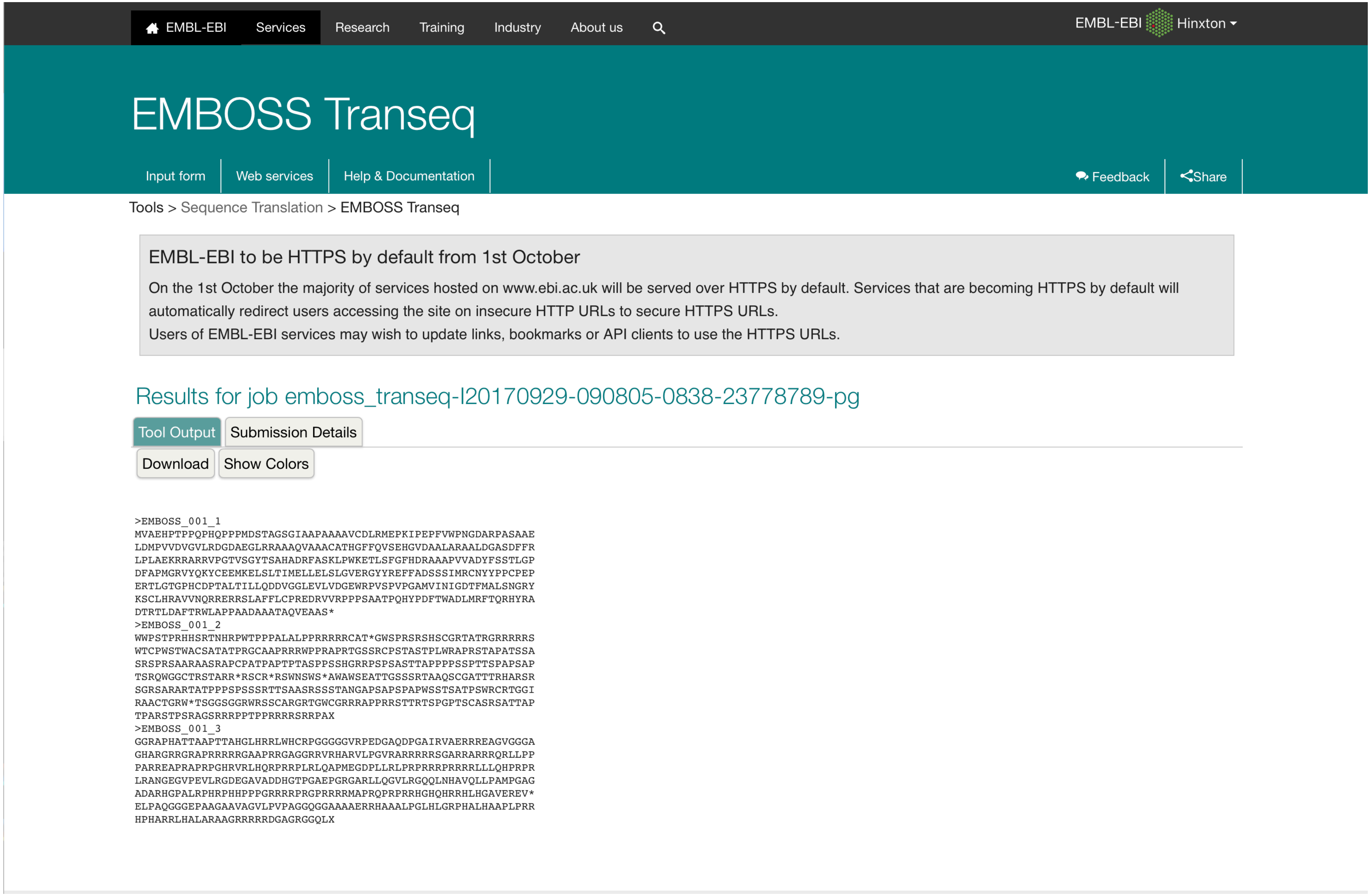
More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

If you plan to use these services during a course please contact us.



EMBOSS Transeq

EMBL-EBI to be HTTPS by default from 1st October

On the 1st October the majority of services hosted on www.ebi.ac.uk will be served over HTTPS by default. Services that are becoming HTTPS by default will automatically redirect users accessing the site on insecure HTTP URLs to secure HTTPS URLs.
Users of EMBL-EBI services may wish to update links, bookmarks or API clients to use the HTTPS URLs.

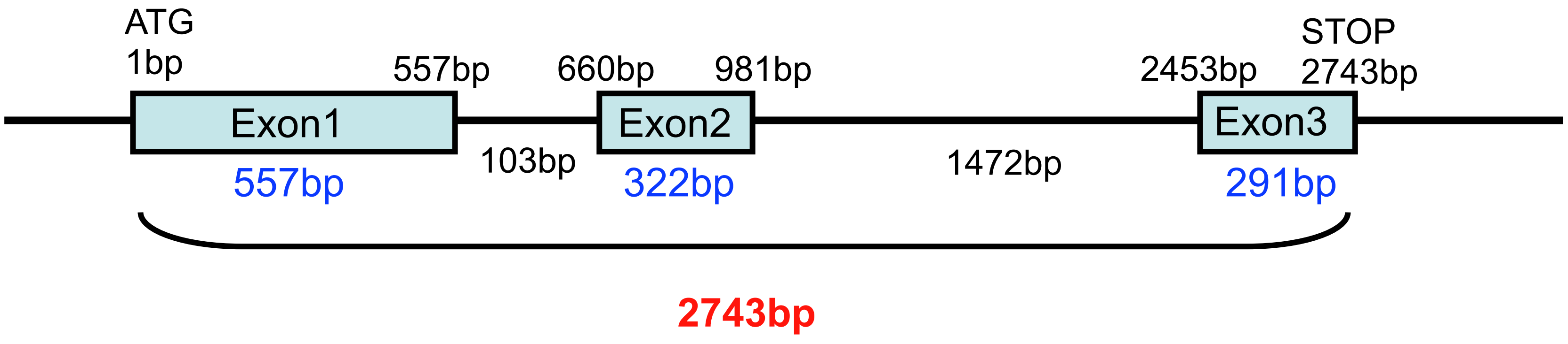
Results for job emboss_transeq-I20170929-090805-0838-23778789-pg

```
>EMBOSS_001_1
MVAEHPTPPQPHQPPPMDSTAGSGIAAPAAAAVCDLRMEPKIPEPFVWPNGDARPASAAE
LDMPVVDVGVLRDGDAEGLRRAAAQVAAACATHGFFQVSEHGVDAALARAALDGASDFFR
LPLAEKRRARRVPGTVSGYTSAHADRFASKLPWKETLSFGFHDRAAAPVVADYFSSTLGP
DFAPMGRVYQKYCEEMKELSLTIMELLELSLGVERGYYREFFADSSSIMRCNYYPPCPEP
ERTLGTGPHCDPTALTILLQDDVGGLEVLVDGEWRPVSPVPGAMVINIGDTFMALSNGRY
KSCLHRAVVNQRRERRSLAFFLCPREDRVVRPPPSAATPQHYPDFTWADLMRFTQRHYRA
DTRTLDAFTRWLAPPAADAAATAQVEAAS*
>EMBOSS_001_2
WWPSTPRHHSRTNHRPWTPPPALALPPRRRRRCAT*GWSPRSRSHSCGRTATRGRRRRRS
WTCPWSTWACSATATPRGCAAPRRRWPPRAPRTGSSRCPSTASTPLWRAPRSTAPATSSA
SRSPRSAARAASRAPCPATPAPTPTASPPSSHGRRPSPSASTTAPPPPSSPTTSPAPSAP
TSRQWGGCTRSTARR*RSCR*RSWNSWS*AWAWSEATTGSSSRTAAQSCGATTTRHARSR
SGRSARARTATPPPSPSSSRTTSAASRSSSTANGAPSAPSPAPWSSTSATPSWRCRTGGI
RAACTGRW*TSGGSGGRWRSSCARGRTGWCGRRRAPPRRSTTRTSPGPTSCASRSATTAP
TPARSTPSRAGSRRRPPTPPRRRRSRRPAX
>EMBOSS_001_3
GGRAPHATTAAPTTAHGLHRRLWHCRPGGGGGVRPEDGAQDPGAIRVAERRREAGVGGGA
GHARGRRGRAPRRRRRGAAPRRGAGGRRVRHARVLPGVRARRRRRSGARRARRRQRLLPP
PARREAPRAPRPGHRVRLHQRPRRPLRLQAPMEGDPLLRLPRPRRRPRRRRLLLQHPRPR
LRANGEGVPEVLRGDEGAVADDHGTPGAEPGRGARLLQGVLRGQQLNHAVQLLPAMPGAG
ADARHGPALRPHRPHHPPPGRRRRPRGPRRRRMAPRQPRPRRHGHQHRRHLHGAVEREV*
ELPAQGGGEPAAGAAVAGVLPVPAGGQGGAAAAERRHAAALPGLHLGRPHALHAAPLPRR
HPHARRLHALARAAGRRRRRDGAGRGGQLX
```

コピー

```
>EMBOSS_001_1
MVAEHPTPPQPHQPPPMDSTAGSGIAAPAAAAVCDLRMEPKIPEPFVWPNGDARPASAAE
LDMPVVDVGVLRDGDAEGLRRAAAQVAAACATHGFFQVSEHGVDAALARAALDGASDFFR
LPLAEKRRARRVPGTVSGYTSAHADRFASKLPWKETLSFGFHDRAAAPVVADYFSSTLGP
DFAPMGRVYQKYCEEMKELSLTIMELLELSLGVERGYYREFFADSSSIMRCNYYPPCPEP
ERTLGTGPHCDPTALTILLQDDVGGLEVLVDGEWRPVSPVPGAMVINIGDTFMALSNGRY
KSCLHRAVVNQRRERRSLAFFLCPREDRVVRPPPSAATPQHYPDFTWADLMRFTQRHYRA
DTRTLDAFTRWLAPPAADAAATAQVEAAS*
>EMBOSS_001_2
WWPSTPRHHSRTNHRPWTPPPALALPPRRRRRCAT*GWSPRSRSHSCGRTATRGRRRRRS
WTCPWSTWACSATATPRGCAAPRRRWPPRAPRTGSSRCPSTASTPLWRAPRSTAPATSSA
SRSPRSAARAASRAPCPATPAPTPTASPPSSHGRRPSPSASTTAPPPPSSPTTSPAPSAP
TSRQWGGCTRSTARR*RSCR*RSWNSWS*AWAWSEATTGSSSRTAAQSCGATTTRHARSR
SGRSARARTATPPPSPSSSRTTSAASRSSSTANGAPSAPSPAPWSSTSATPSWRCRTGGI
RAACTGRW*TSGGSGGRWRSSCARGRTGWCGRRRAPPRRSTTRTSPGPTSCASRSATTAP
TPARSTPSRAGSRRRPPTPPRRRRSRRPAX
>EMBOSS_001_3
GGRAPHATTAAPTTAHGLHRRLWHCRPGGGGVRPEDGAQDPGAIRVAERRREAGVGGA
GHARGRRGRAPRRRRRGAAPRRGAGGRRVRHARVLPGVRARRRRRSGARRARRRQRLLPP
PARREAPRAPRPGHRVRLHQRPRRPLRLQAPMEGDPLLRLPRPRRRPRRRRLLLQHPRPR
LRANGEGVPEVLRGDEGAVADDHGTPGAEPGRGARLLQGVLRGQQLNHAVQLLPAMPGAG
ADARHGPALRPHRPHHPPPGRRRRPRGPRRRRMAPRQPRPRRHGHQHRRHLHGAVEREV*
ELPAQGGGEPAAGAAVAGVLPVPAGGQGGAAAAERRHAAALPGLHLGRPHALHAAPLPRR
HPHARRLHALARAAGRRRRRDGAGRGGQLX
```

ある遺伝子Aの構造



ある遺伝子
3つのエクソン（第1エキソン 557bp, 第2エキソン 322bp, 第1エキソン 291bp）、
2つのイントロン（第1イントロン 103bp, 第2イントロン 1472bp）からなる。
遺伝子の長さは、2743bp(但しプロモーター領域は含まず)

アミノ酸配列

```
MVAEHPTPPQPHQPPPMDSTAGSGIAAPAAAAVCDLRMEPKIPEPFVWPNGDARPASAAE
LDMPVVDVGVLRDGDAEGLRRAAAQVAAACATHGFFQVSEHGVDAALARAALDGASDFFR
LPLAEKRRARRVPGTVSGYTSAHADRFASKLPWKETLSFGFHDRAAAPVVADYFSSTLGP
DFAPMGRVYQKYCEEMKELSLTIMELLELSLGVERGYYREFFADSSSIMRCNYYPPCPEP
ERTLGTGPHCDPTALTILLQDDVGGLEVLVDGEWRPVSPVPGAMVINIGDTFMALSNGRY
KSCLHRAVVNQRRERRSLAFFLCPREDRVVRPPPSAATPQHYPDFTWADLMRFTQRHYRA
DTRTLDAFTRWLAPPAADAAATAQVEAAS*
```

2. 何の遺伝子をコードしているのか?

(1) ホモロジーサーチによって遺伝子を推測する

- Blast searchを行う

生物情報工学II (担当　芦苅）遺伝子の配列解析

第６回目-1（11/1）遺伝子の予測と機能解析

-2（11/1）演習

第７回目-1（11/8）DNA配列から遺伝子を探す

-2（11/8）DNA、アミノ酸のアラインメント、系統樹解析

-3（11/8）演習

第８回目-1（11/15）Genetyxを用いた解析

-2（11/15）演習

第９回目-1（11/22）KEGGを用いた解析

連鎖解析

出席と宿題の提出　提出方法

リンク集

データベース検索：

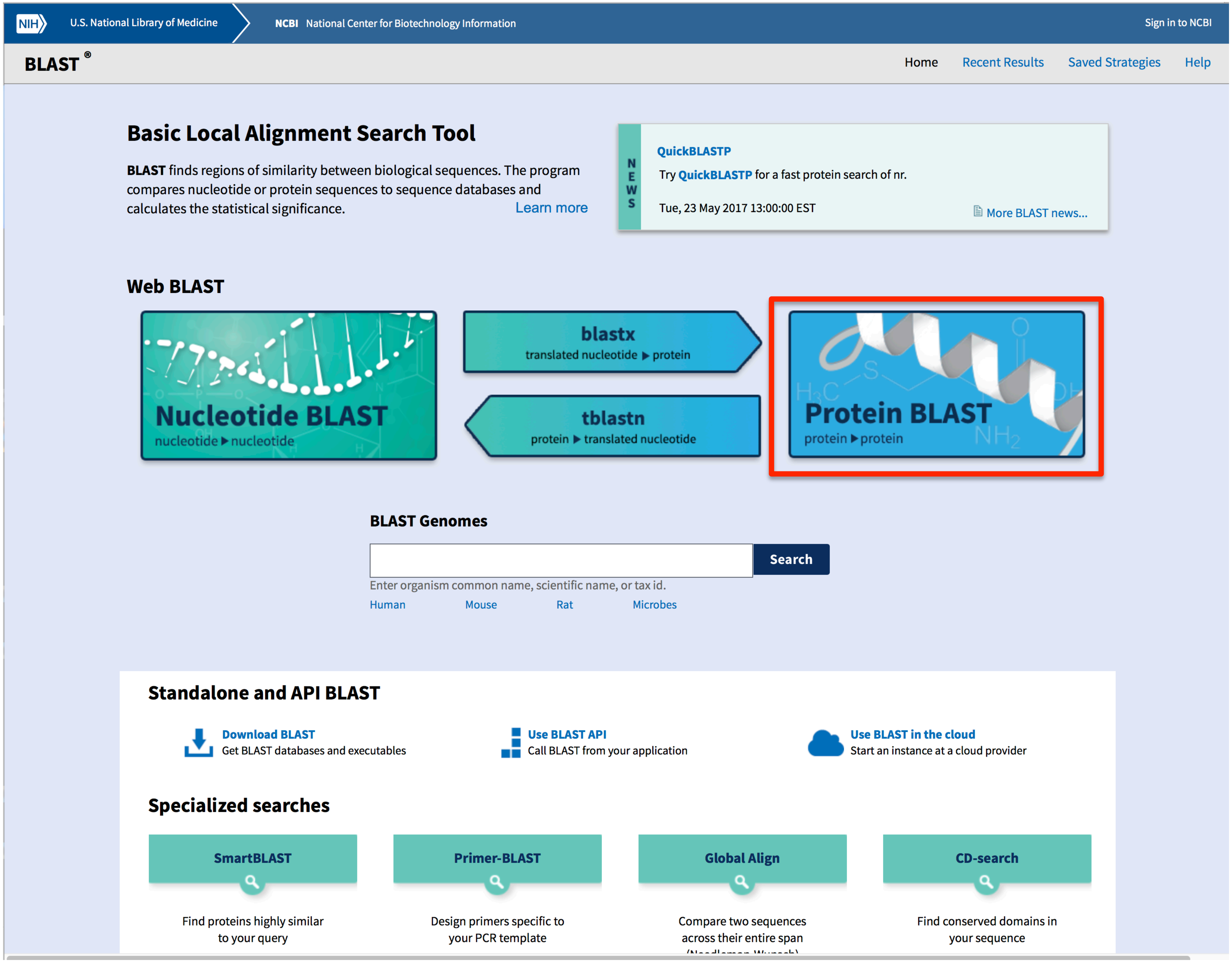
1. PubMed: 論文検索
2. Entrez：総合データベース
3. Google：何でも検索
4. Google Scholar：科学に関する検索
5. 特許データベース

ホモロジー検索：

1. BLAST [GenomeNET]
2. FASTA [GenomeNET]

配列解析：

1. Splign ：DNAとcDNAのマッチング
2. Blast align ：DNAとcDNAのマッチング
3. EMBOSS program" ：DNAとcDNAのマッチング
4. EMBOSS program" ：DNA配列をアミノ酸配列に変換
5. BLAST(NCBI) ：遺伝子のホモロジー検索
6. BLAST(DDBJ) ：遺伝子のホモロジー検索
7. Softberry/FGENESH ：遺伝子の予測
8. GeneMark ：遺伝子の予測
9. GENESCAN ：遺伝子の予測
10. clustalW： アライメント
11. InterProScan： ドメイン解析
12. PSORT： 細胞内局在性予測



U.S. National Library of Medicine

NCBI National Center for Biotechnology Information

Sign in to NCBI

BLAST®

Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

QuickBLASTP

Try **QuickBLASTP** for a fast protein search of nr.

Tue, 23 May 2017 13:00:00 EST

More BLAST news...

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

BLAST Genomes

Search

Enter organism common name, scientific name, or tax id.

Human Mouse Rat Microbes

Standalone and API BLAST

Download BLAST
Get BLAST databases and executables

Use BLAST API
Call BLAST from your application

Use BLAST in the cloud
Start an instance at a cloud provider

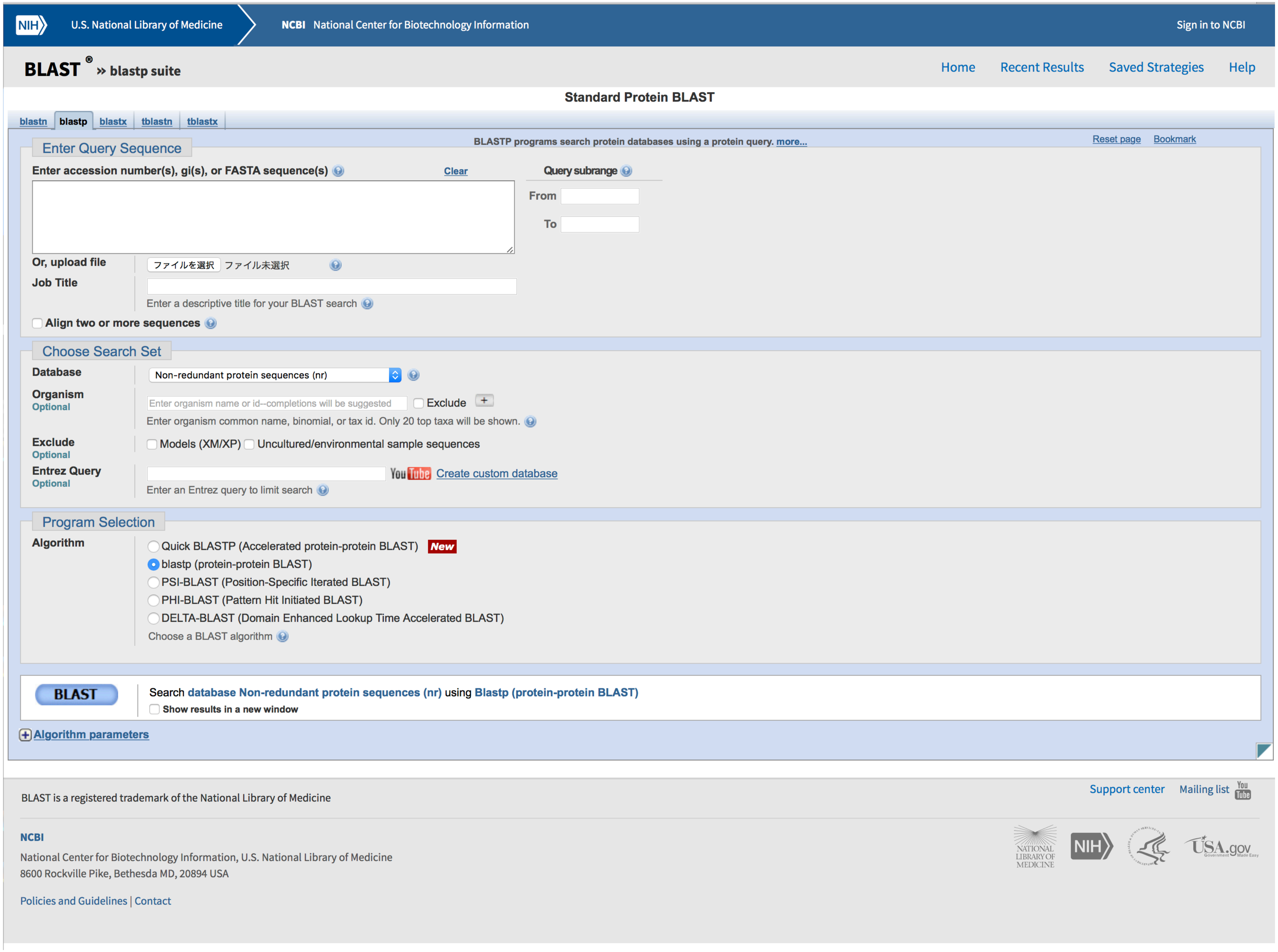
Specialized searches

SmartBLAST
Find proteins highly similar to your query

Primer-BLAST
Design primers specific to your PCR template

Global Align
Compare two sequences across their entire span

CD-search
Find conserved domains in your sequence



NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® » blastp suite

Home Recent Results Saved Strategies Help

Standard Protein BLAST

blastn **blastp** blastx tblastn tblastx

BLASTP programs search protein databases using a protein query. more...

Reset page Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange

From

To

Or, upload file ファイルを選択 ファイル未選択

Job Title

Enter a descriptive title for your BLAST search

☐ **Align two or more sequences**

Choose Search Set

Database Non-redundant protein sequences (nr)

Organism
Optional

Enter organism name or id--completions will be suggested ☐ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude
Optional

☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query
Optional

YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Algorithm

○ Quick BLASTP (Accelerated protein-protein BLAST) *New*
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST

Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

☐ **Show results in a new window**

+ Algorithm parameters

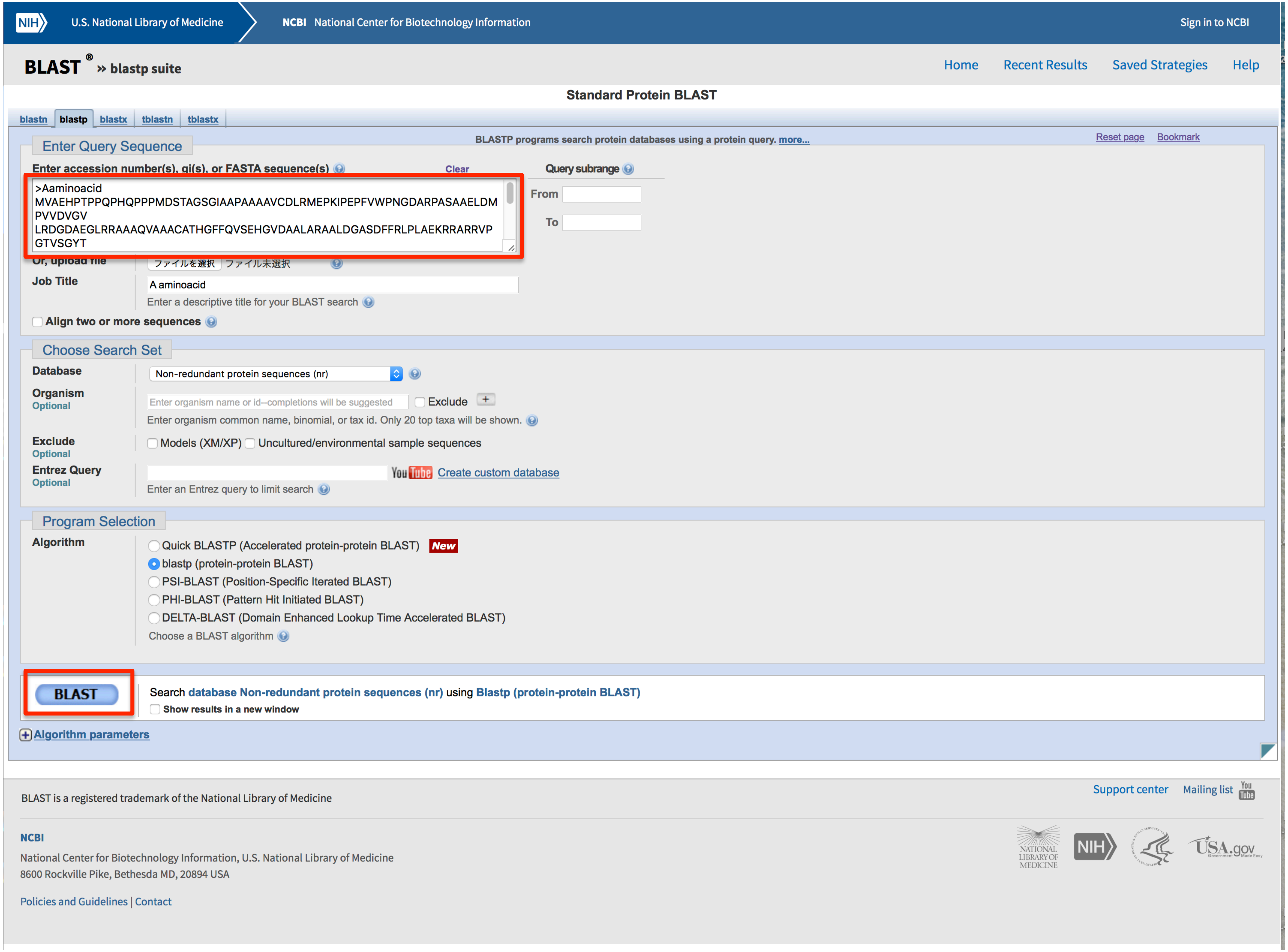
BLAST is a registered trademark of the National Library of Medicine

Support center Mailing list

NCBI

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

Policies and Guidelines | Contact



BLAST® » blastp suite

Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>Aaminoacid
MVAEHPTPPQPHQPPPMDSTAGSGIAAPAAAAVCDLRMEPKIPEPFVWPNGDARPASAAELDM
PVVDVGV
LRDGDAEGLRRAAAQVAAACATHGFFQVSEHGVDAALARAALDGASDFFRLPLAEKRRARRVP
GTVSGYT
```

Or, upload file: ファイルを選択 ファイル未選択

Job Title: A aminoacid

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested ☐ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query (Optional): YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Algorithm

○ Quick BLASTP (Accelerated protein-protein BLAST) New

● blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

☐ Show results in a new window

+ Algorithm parameters

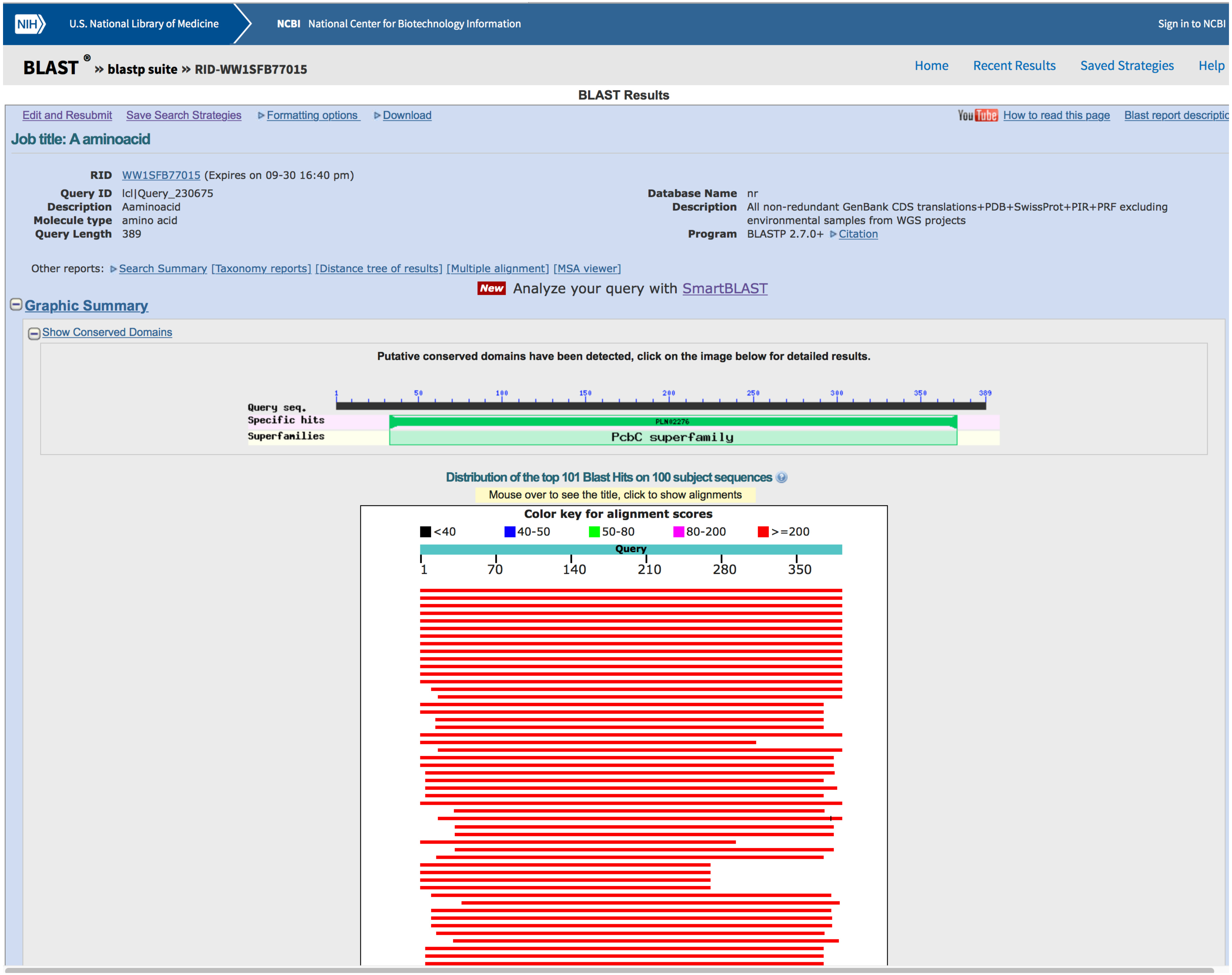
BLAST is a registered trademark of the National Library of Medicine

Support center Mailing list

NCBI

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

Policies and Guidelines | Contact



BLAST ® » blastp suite » RID-WW1SFB77015

BLAST Results

Job title: A aminoacid

RID	WW1SFB77015 (Expires on 09-30 16:40 pm)		
Query ID	lcl|Query_230675	**Database Name**	nr
Description	Aaminoacid	**Description**	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type	amino acid		
Query Length	389	**Program**	BLASTP 2.7.0+ Citation

Other reports: Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] [MSA viewer]

New Analyze your query with SmartBLAST

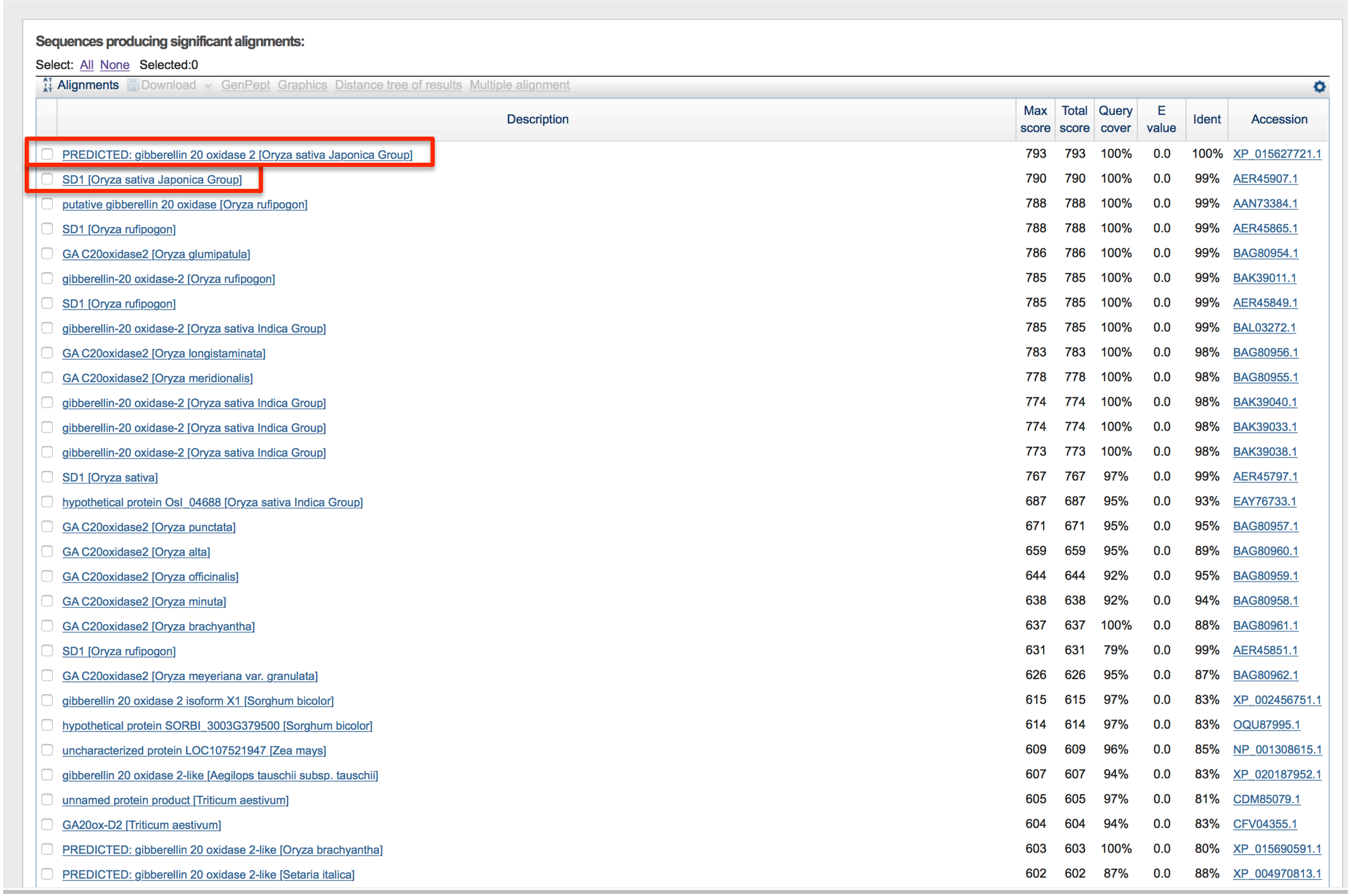
Graphic Summary

Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of the top 101 Blast Hits on 100 subject sequences

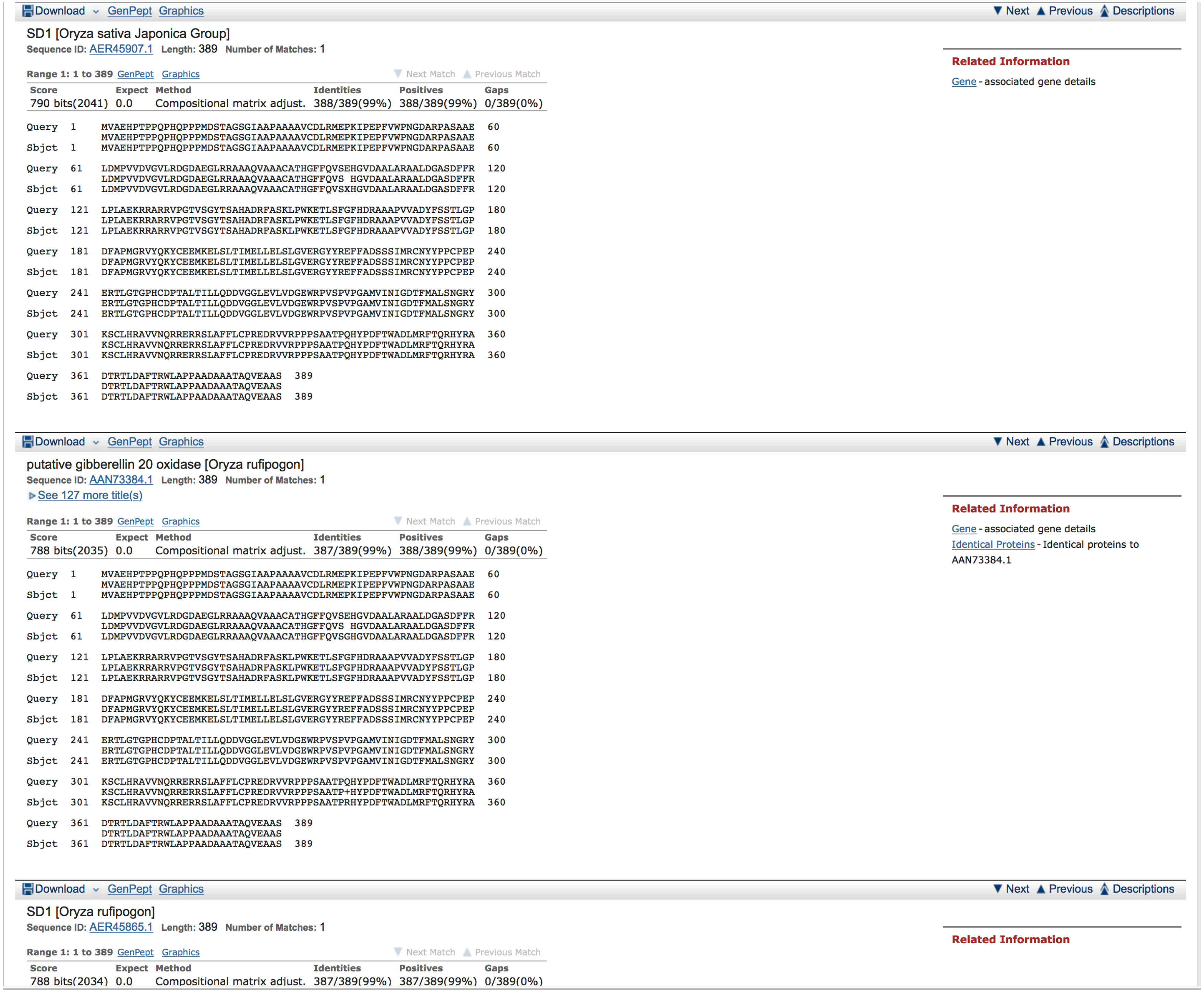
Mouse over to see the title, click to show alignments



Sequences producing significant alignments:

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	PREDICTED: gibberellin 20 oxidase 2 [Oryza sativa Japonica Group]	793	793	100%	0.0	100%	XP_015627721.1
☐	SD1 [Oryza sativa Japonica Group]	790	790	100%	0.0	99%	AER45907.1
☐	putative gibberellin 20 oxidase [Oryza rufipogon]	788	788	100%	0.0	99%	AAN73384.1
☐	SD1 [Oryza rufipogon]	788	788	100%	0.0	99%	AER45865.1
☐	GA C20oxidase2 [Oryza glumipatula]	786	786	100%	0.0	99%	BAG80954.1
☐	gibberellin-20 oxidase-2 [Oryza rufipogon]	785	785	100%	0.0	99%	BAK39011.1
☐	SD1 [Oryza rufipogon]	785	785	100%	0.0	99%	AER45849.1
☐	gibberellin-20 oxidase-2 [Oryza sativa Indica Group]	785	785	100%	0.0	99%	BAL03272.1
☐	GA C20oxidase2 [Oryza longistaminata]	783	783	100%	0.0	98%	BAG80956.1
☐	GA C20oxidase2 [Oryza meridionalis]	778	778	100%	0.0	98%	BAG80955.1
☐	gibberellin-20 oxidase-2 [Oryza sativa Indica Group]	774	774	100%	0.0	98%	BAK39040.1
☐	gibberellin-20 oxidase-2 [Oryza sativa Indica Group]	774	774	100%	0.0	98%	BAK39033.1
☐	gibberellin-20 oxidase-2 [Oryza sativa Indica Group]	773	773	100%	0.0	98%	BAK39038.1
☐	SD1 [Oryza sativa]	767	767	97%	0.0	99%	AER45797.1
☐	hypothetical protein OsI_04688 [Oryza sativa Indica Group]	687	687	95%	0.0	93%	EAY76733.1
☐	GA C20oxidase2 [Oryza punctata]	671	671	95%	0.0	95%	BAG80957.1
☐	GA C20oxidase2 [Oryza alta]	659	659	95%	0.0	89%	BAG80960.1
☐	GA C20oxidase2 [Oryza officinalis]	644	644	92%	0.0	95%	BAG80959.1
☐	GA C20oxidase2 [Oryza minuta]	638	638	92%	0.0	94%	BAG80958.1
☐	GA C20oxidase2 [Oryza brachyantha]	637	637	100%	0.0	88%	BAG80961.1
☐	SD1 [Oryza rufipogon]	631	631	79%	0.0	99%	AER45851.1
☐	GA C20oxidase2 [Oryza meyeriana var. granulata]	626	626	95%	0.0	87%	BAG80962.1
☐	gibberellin 20 oxidase 2 isoform X1 [Sorghum bicolor]	615	615	97%	0.0	83%	XP_002456751.1
☐	hypothetical protein SORBI_3003G379500 [Sorghum bicolor]	614	614	97%	0.0	83%	OQU87995.1
☐	uncharacterized protein LOC107521947 [Zea mays]	609	609	96%	0.0	85%	NP_001308615.1
☐	gibberellin 20 oxidase 2-like [Aegilops tauschii subsp. tauschii]	607	607	94%	0.0	83%	XP_020187952.1
☐	unnamed protein product [Triticum aestivum]	605	605	97%	0.0	81%	CDM85079.1
☐	GA20ox-D2 [Triticum aestivum]	604	604	94%	0.0	83%	CFV04355.1
☐	PREDICTED: gibberellin 20 oxidase 2-like [Oryza brachyantha]	603	603	100%	0.0	80%	XP_015690591.1
☐	PREDICTED: gibberellin 20 oxidase 2-like [Setaria italica]	602	602	87%	0.0	88%	XP_004970813.1



SD1 [Oryza sativa Japonica Group]

Sequence ID: AER45907.1 **Length:** 389 **Number of Matches:** 1

Related Information

Gene - associated gene details

Range 1: 1 to 389 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
790 bits(2041)	0.0	Compositional matrix adjust.	388/389(99%)	388/389(99%)	0/389(0%)

```
Query  1    MVAEHPTPPQPHQPPPMDSTAGSGIAAPAAAAVCDLRMEPKIPEPFVWPNGDARPASAAE  60
            MVAEHPTPPQPHQPPPMDSTAGSGIAAPAAAAVCDLRMEPKIPEPFVWPNGDARPASAAE
Sbjct  1    MVAEHPTPPQPHQPPPMDSTAGSGIAAPAAAAVCDLRMEPKIPEPFVWPNGDARPASAAE  60

Query  61   LDMPVVDVGVLRDGDAEGLRRAAAQVAAACATHGFFQVSEHGVDAALARAALDGASDFFR  120
            LDMPVVDVGVLRDGDAEGLRRAAAQVAAACATHGFFQVS HGVDAALARAALDGASDFFR
Sbjct  61   LDMPVVDVGVLRDGDAEGLRRAAAQVAAACATHGFFQVSXHGVDAALARAALDGASDFFR  120

Query  121  LPLAEKRRARRVPGTVSGYTSAHADRFASKLPWKETLSFGFHDRAAAPVVADYFSSTLGP  180
            LPLAEKRRARRVPGTVSGYTSAHADRFASKLPWKETLSFGFHDRAAAPVVADYFSSTLGP
Sbjct  121  LPLAEKRRARRVPGTVSGYTSAHADRFASKLPWKETLSFGFHDRAAAPVVADYFSSTLGP  180

Query  181  DFAPMGRVYQKYCEEMKELSLTIMELLELSLGVERGYYREFFADSSSIMRCNYYPPCPEP  240
            DFAPMGRVYQKYCEEMKELSLTIMELLELSLGVERGYYREFFADSSSIMRCNYYPPCPEP
Sbjct  181  DFAPMGRVYQKYCEEMKELSLTIMELLELSLGVERGYYREFFADSSSIMRCNYYPPCPEP  240

Query  241  ERTLGTGPHCDPTALTILLQDDVGGLEVLVDGEWRPVSPVPGAMVINIGDTFMALSNGRY  300
            ERTLGTGPHCDPTALTILLQDDVGGLEVLVDGEWRPVSPVPGAMVINIGDTFMALSNGRY
Sbjct  241  ERTLGTGPHCDPTALTILLQDDVGGLEVLVDGEWRPVSPVPGAMVINIGDTFMALSNGRY  300

Query  301  KSCLHRAVVNQRRERRSLAFFLCPREDRVVRPPPSAATPQHYPDFTWADLMRFTQRHYRA  360
            KSCLHRAVVNQRRERRSLAFFLCPREDRVVRPPPSAATPQHYPDFTWADLMRFTQRHYRA
Sbjct  301  KSCLHRAVVNQRRERRSLAFFLCPREDRVVRPPPSAATPQHYPDFTWADLMRFTQRHYRA  360

Query  361  DTRTLDAFTRWLAPPAADAAATAQVEAAS  389
            DTRTLDAFTRWLAPPAADAAATAQVEAAS
Sbjct  361  DTRTLDAFTRWLAPPAADAAATAQVEAAS  389
```

putative gibberellin 20 oxidase [Oryza rufipogon]

Sequence ID: AAN73384.1 **Length:** 389 **Number of Matches:** 1

See 127 more title(s)

Related Information

Gene - associated gene details

Identical Proteins - Identical proteins to AAN73384.1

Range 1: 1 to 389 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
788 bits(2035)	0.0	Compositional matrix adjust.	387/389(99%)	388/389(99%)	0/389(0%)

```
Query  1    MVAEHPTPPQPHQPPPMDSTAGSGIAAPAAAAVCDLRMEPKIPEPFVWPNGDARPASAAE  60
            MVAEHPTPPQPHQPPPMDSTAGSGIAAPAAAAVCDLRMEPKIPEPFVWPNGDARPASAAE
Sbjct  1    MVAEHPTPPQPHQPPPMDSTAGSGIAAPAAAAVCDLRMEPKIPEPFVWPNGDARPASAAE  60

Query  61   LDMPVVDVGVLRDGDAEGLRRAAAQVAAACATHGFFQVSEHGVDAALARAALDGASDFFR  120
            LDMPVVDVGVLRDGDAEGLRRAAAQVAAACATHGFFQVS HGVDAALARAALDGASDFFR
Sbjct  61   LDMPVVDVGVLRDGDAEGLRRAAAQVAAACATHGFFQVSGHGVDAALARAALDGASDFFR  120

Query  121  LPLAEKRRARRVPGTVSGYTSAHADRFASKLPWKETLSFGFHDRAAAPVVADYFSSTLGP  180
            LPLAEKRRARRVPGTVSGYTSAHADRFASKLPWKETLSFGFHDRAAAPVVADYFSSTLGP
Sbjct  121  LPLAEKRRARRVPGTVSGYTSAHADRFASKLPWKETLSFGFHDRAAAPVVADYFSSTLGP  180

Query  181  DFAPMGRVYQKYCEEMKELSLTIMELLELSLGVERGYYREFFADSSSIMRCNYYPPCPEP  240
            DFAPMGRVYQKYCEEMKELSLTIMELLELSLGVERGYYREFFADSSSIMRCNYYPPCPEP
Sbjct  181  DFAPMGRVYQKYCEEMKELSLTIMELLELSLGVERGYYREFFADSSSIMRCNYYPPCPEP  240

Query  241  ERTLGTGPHCDPTALTILLQDDVGGLEVLVDGEWRPVSPVPGAMVINIGDTFMALSNGRY  300
            ERTLGTGPHCDPTALTILLQDDVGGLEVLVDGEWRPVSPVPGAMVINIGDTFMALSNGRY
Sbjct  241  ERTLGTGPHCDPTALTILLQDDVGGLEVLVDGEWRPVSPVPGAMVINIGDTFMALSNGRY  300

Query  301  KSCLHRAVVNQRRERRSLAFFLCPREDRVVRPPPSAATPQHYPDFTWADLMRFTQRHYRA  360
            KSCLHRAVVNQRRERRSLAFFLCPREDRVVRPPPSAATP+HYPDFTWADLMRFTQRHYRA
Sbjct  301  KSCLHRAVVNQRRERRSLAFFLCPREDRVVRPPPSAATPRHYPDFTWADLMRFTQRHYRA  360

Query  361  DTRTLDAFTRWLAPPAADAAATAQVEAAS  389
            DTRTLDAFTRWLAPPAADAAATAQVEAAS
Sbjct  361  DTRTLDAFTRWLAPPAADAAATAQVEAAS  389
```

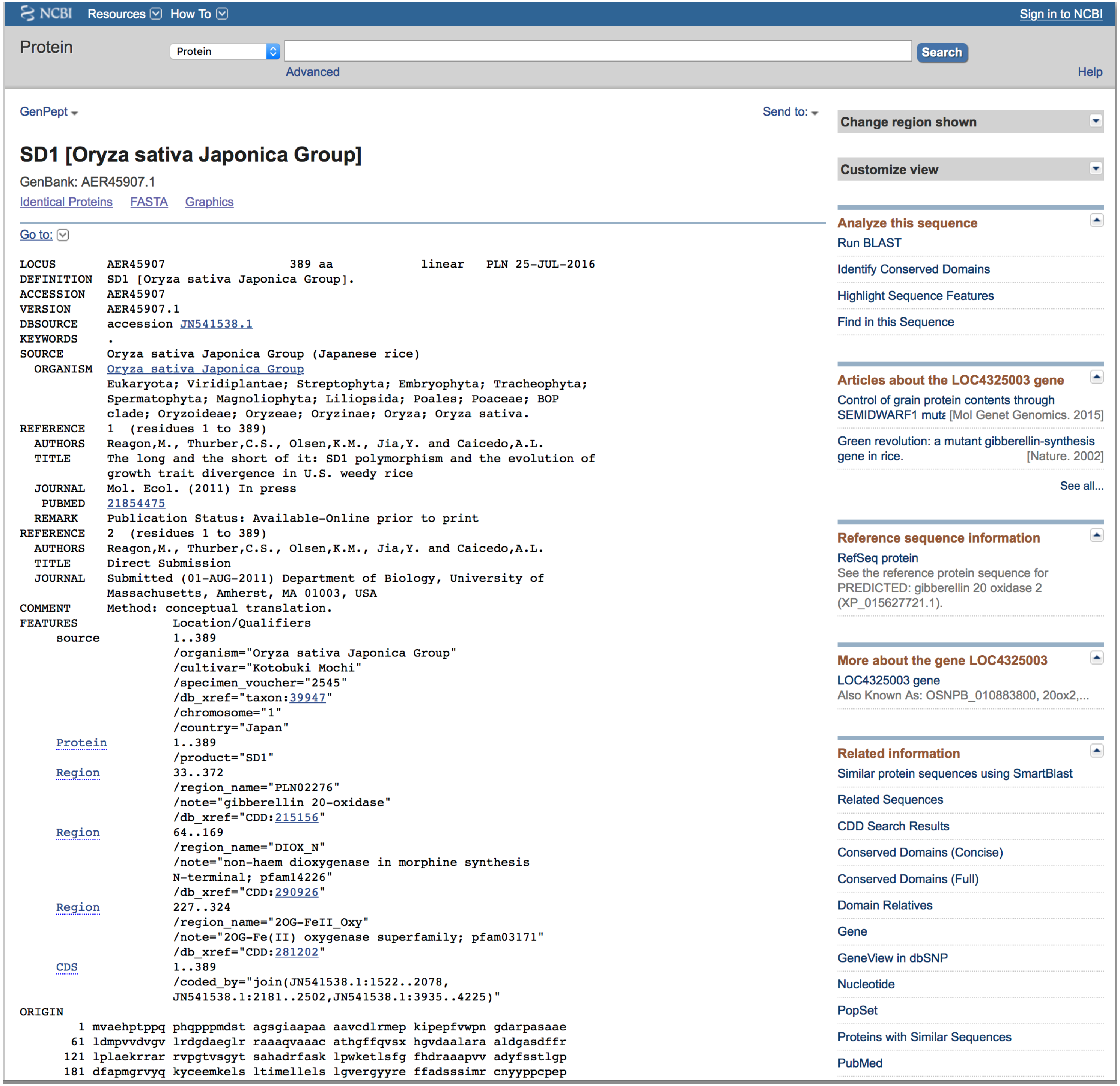
SD1 [Oryza rufipogon]

Sequence ID: AER45865.1 **Length:** 389 **Number of Matches:** 1

Related Information

Range 1: 1 to 389 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
788 bits(2034)	0.0	Compositional matrix adjust.	387/389(99%)	387/389(99%)	0/389(0%)



SD1 [Oryza sativa Japonica Group]

GenBank: AER45907.1

Identical Proteins FASTA Graphics

```
LOCUS       AER45907                 389 aa            linear   PLN 25-JUL-2016
DEFINITION  SD1 [Oryza sativa Japonica Group].
ACCESSION   AER45907
VERSION     AER45907.1
DBSOURCE    accession JN541538.1
KEYWORDS    .
SOURCE      Oryza sativa Japonica Group (Japanese rice)
  ORGANISM  Oryza sativa Japonica Group
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; BOP
            clade; Oryzoideae; Oryzeae; Oryzinae; Oryza; Oryza sativa.
REFERENCE   1  (residues 1 to 389)
  AUTHORS   Reagon,M., Thurber,C.S., Olsen,K.M., Jia,Y. and Caicedo,A.L.
  TITLE     The long and the short of it: SD1 polymorphism and the evolution of
            growth trait divergence in U.S. weedy rice
  JOURNAL   Mol. Ecol. (2011) In press
   PUBMED   21854475
  REMARK    Publication Status: Available-Online prior to print
REFERENCE   2  (residues 1 to 389)
  AUTHORS   Reagon,M., Thurber,C.S., Olsen,K.M., Jia,Y. and Caicedo,A.L.
  TITLE     Direct Submission
  JOURNAL   Submitted (01-AUG-2011) Department of Biology, University of
            Massachusetts, Amherst, MA 01003, USA
COMMENT     Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..389
                     /organism="Oryza sativa Japonica Group"
                     /cultivar="Kotobuki Mochi"
                     /specimen_voucher="2545"
                     /db_xref="taxon:39947"
                     /chromosome="1"
                     /country="Japan"
     Protein         1..389
                     /product="SD1"
     Region          33..372
                     /region_name="PLN02276"
                     /note="gibberellin 20-oxidase"
                     /db_xref="CDD:215156"
     Region          64..169
                     /region_name="DIOX_N"
                     /note="non-haem dioxygenase in morphine synthesis
                     N-terminal; pfam14226"
                     /db_xref="CDD:290926"
     Region          227..324
                     /region_name="2OG-FeII_Oxy"
                     /note="2OG-Fe(II) oxygenase superfamily; pfam03171"
                     /db_xref="CDD:281202"
     CDS             1..389
                     /coded_by="join(JN541538.1:1522..2078,
                     JN541538.1:2181..2502,JN541538.1:3935..4225)"
ORIGIN
        1 mvaehptppq phqpppmdst agsgiaapaa aavcdlrmep kipepfvwpn gdarpasaae
       61 ldmpvvdvgv lrdgdaeglr raaaqvaaac athgffqvsx hgvdaalara aldgasdffr
      121 lplaekrrar rvpgtvsgyt sahadrfask lpwketlsfg fhdraaapvv adyfsstlgp
      181 dfapmgrvyq kyceemkels ltimellels lgvergyyre ffadsssimr cnyyppcpep
```

Analyze this sequence

Run BLAST

Identify Conserved Domains

Highlight Sequence Features

Find in this Sequence

Articles about the LOC4325003 gene

Control of grain protein contents through SEMIDWARF1 mut‹ [Mol Genet Genomics. 2015]

Green revolution: a mutant gibberellin-synthesis gene in rice. [Nature. 2002]

See all...

Reference sequence information

RefSeq protein

See the reference protein sequence for PREDICTED: gibberellin 20 oxidase 2 (XP_015627721.1).

More about the gene LOC4325003

LOC4325003 gene

Also Known As: OSNPB_010883800, 20ox2,...

Related information

Similar protein sequences using SmartBlast

Related Sequences

CDD Search Results

Conserved Domains (Concise)

Conserved Domains (Full)

Domain Relatives

Gene

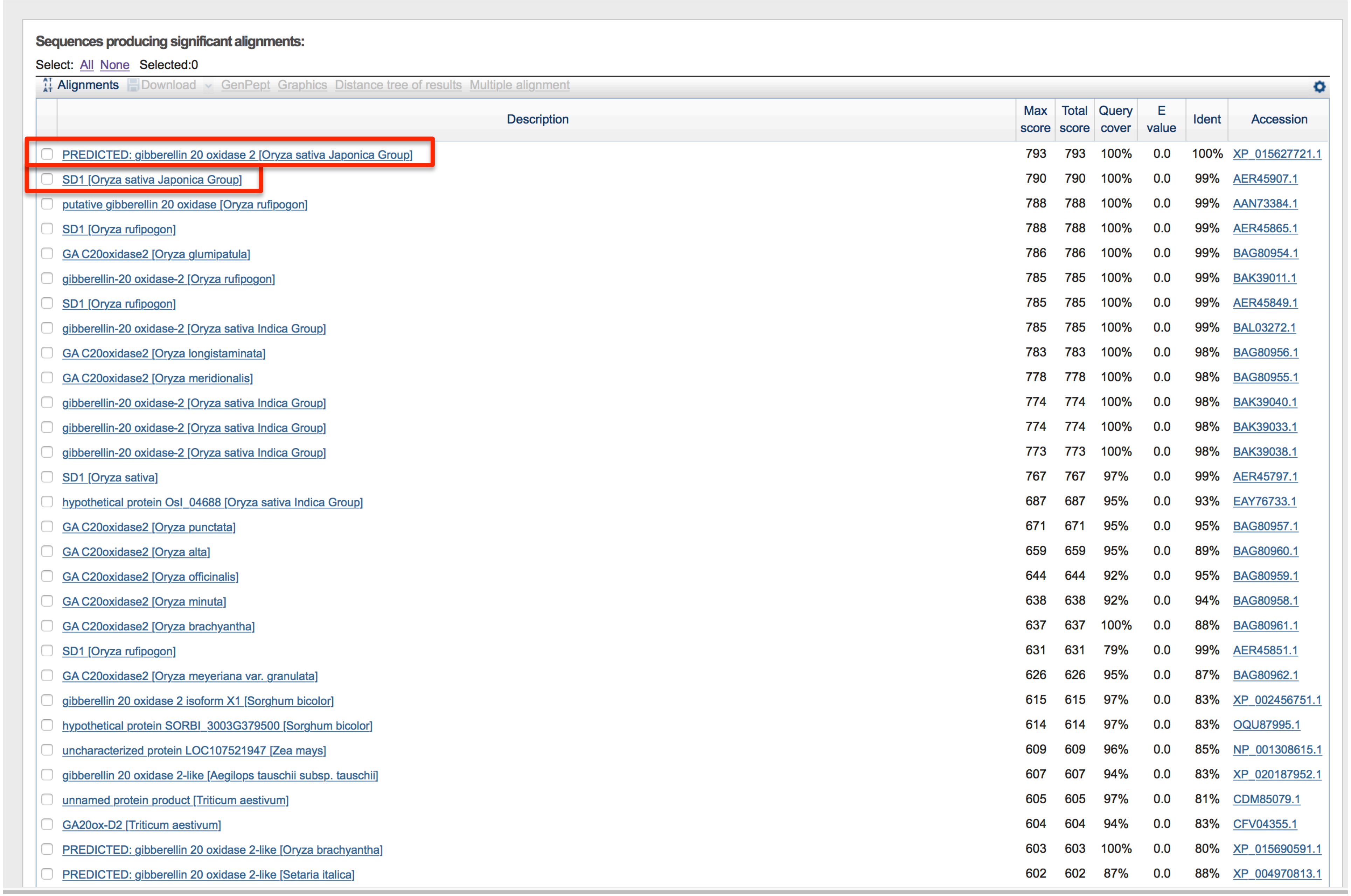
GeneView in dbSNP

Nucleotide

PopSet

Proteins with Similar Sequences

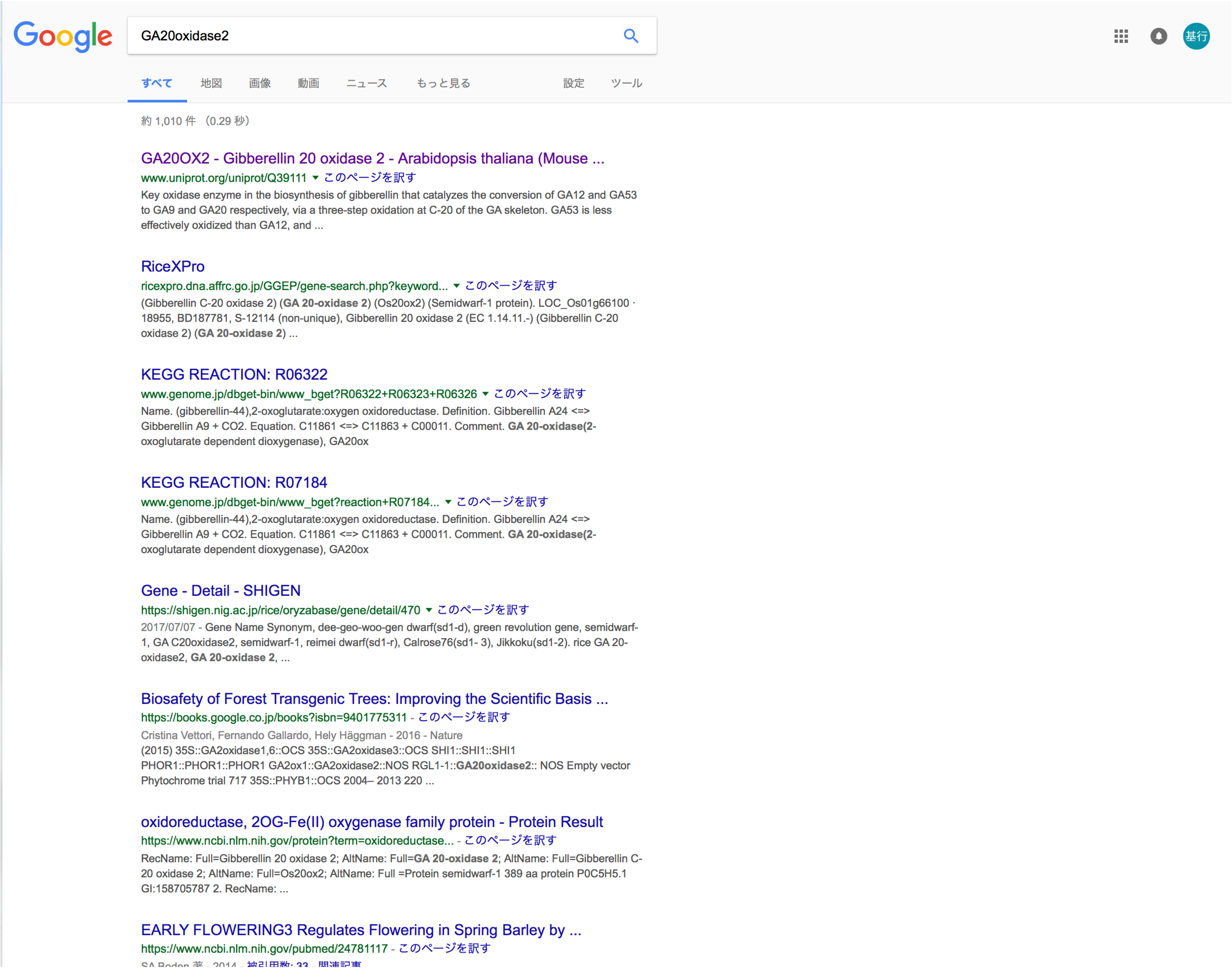
PubMed



Sequences producing significant alignments:

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	PREDICTED: gibberellin 20 oxidase 2 [Oryza sativa Japonica Group]	793	793	100%	0.0	100%	XP_015627721.1
☐	SD1 [Oryza sativa Japonica Group]	790	790	100%	0.0	99%	AER45907.1
☐	putative gibberellin 20 oxidase [Oryza rufipogon]	788	788	100%	0.0	99%	AAN73384.1
☐	SD1 [Oryza rufipogon]	788	788	100%	0.0	99%	AER45865.1
☐	GA C20oxidase2 [Oryza glumipatula]	786	786	100%	0.0	99%	BAG80954.1
☐	gibberellin-20 oxidase-2 [Oryza rufipogon]	785	785	100%	0.0	99%	BAK39011.1
☐	SD1 [Oryza rufipogon]	785	785	100%	0.0	99%	AER45849.1
☐	gibberellin-20 oxidase-2 [Oryza sativa Indica Group]	785	785	100%	0.0	99%	BAL03272.1
☐	GA C20oxidase2 [Oryza longistaminata]	783	783	100%	0.0	98%	BAG80956.1
☐	GA C20oxidase2 [Oryza meridionalis]	778	778	100%	0.0	98%	BAG80955.1
☐	gibberellin-20 oxidase-2 [Oryza sativa Indica Group]	774	774	100%	0.0	98%	BAK39040.1
☐	gibberellin-20 oxidase-2 [Oryza sativa Indica Group]	774	774	100%	0.0	98%	BAK39033.1
☐	gibberellin-20 oxidase-2 [Oryza sativa Indica Group]	773	773	100%	0.0	98%	BAK39038.1
☐	SD1 [Oryza sativa]	767	767	97%	0.0	99%	AER45797.1
☐	hypothetical protein OsI_04688 [Oryza sativa Indica Group]	687	687	95%	0.0	93%	EAY76733.1
☐	GA C20oxidase2 [Oryza punctata]	671	671	95%	0.0	95%	BAG80957.1
☐	GA C20oxidase2 [Oryza alta]	659	659	95%	0.0	89%	BAG80960.1
☐	GA C20oxidase2 [Oryza officinalis]	644	644	92%	0.0	95%	BAG80959.1
☐	GA C20oxidase2 [Oryza minuta]	638	638	92%	0.0	94%	BAG80958.1
☐	GA C20oxidase2 [Oryza brachyantha]	637	637	100%	0.0	88%	BAG80961.1
☐	SD1 [Oryza rufipogon]	631	631	79%	0.0	99%	AER45851.1
☐	GA C20oxidase2 [Oryza meyeriana var. granulata]	626	626	95%	0.0	87%	BAG80962.1
☐	gibberellin 20 oxidase 2 isoform X1 [Sorghum bicolor]	615	615	97%	0.0	83%	XP_002456751.1
☐	hypothetical protein SORBI_3003G379500 [Sorghum bicolor]	614	614	97%	0.0	83%	OQU87995.1
☐	uncharacterized protein LOC107521947 [Zea mays]	609	609	96%	0.0	85%	NP_001308615.1
☐	gibberellin 20 oxidase 2-like [Aegilops tauschii subsp. tauschii]	607	607	94%	0.0	83%	XP_020187952.1
☐	unnamed protein product [Triticum aestivum]	605	605	97%	0.0	81%	CDM85079.1
☐	GA20ox-D2 [Triticum aestivum]	604	604	94%	0.0	83%	CFV04355.1
☐	PREDICTED: gibberellin 20 oxidase 2-like [Oryza brachyantha]	603	603	100%	0.0	80%	XP_015690591.1
☐	PREDICTED: gibberellin 20 oxidase 2-like [Setaria italica]	602	602	87%	0.0	88%	XP_004970813.1



GA20oxidase2

約 1,010 件 （0.29 秒）

GA20OX2 - Gibberellin 20 oxidase 2 - Arabidopsis thaliana (Mouse ...
www.uniprot.org/uniprot/Q39111 ▾ このページを訳す
Key oxidase enzyme in the biosynthesis of gibberellin that catalyzes the conversion of GA12 and GA53 to GA9 and GA20 respectively, via a three-step oxidation at C-20 of the GA skeleton. GA53 is less effectively oxidized than GA12, and ...

RiceXPro
ricexpro.dna.affrc.go.jp/GGEP/gene-search.php?keyword... ▾ このページを訳す
(Gibberellin C-20 oxidase 2) (**GA 20-oxidase 2**) (Os20ox2) (Semidwarf-1 protein). LOC_Os01g66100 · 18955, BD187781, S-12114 (non-unique), Gibberellin 20 oxidase 2 (EC 1.14.11.-) (Gibberellin C-20 oxidase 2) (**GA 20-oxidase 2**) ...

KEGG REACTION: R06322
www.genome.jp/dbget-bin/www_bget?R06322+R06323+R06326 ▾ このページを訳す
Name. (gibberellin-44),2-oxoglutarate:oxygen oxidoreductase. Definition. Gibberellin A24 <=> Gibberellin A9 + CO2. Equation. C11861 <=> C11863 + C00011. Comment. **GA 20-oxidase(2-**oxoglutarate dependent dioxygenase), GA20ox

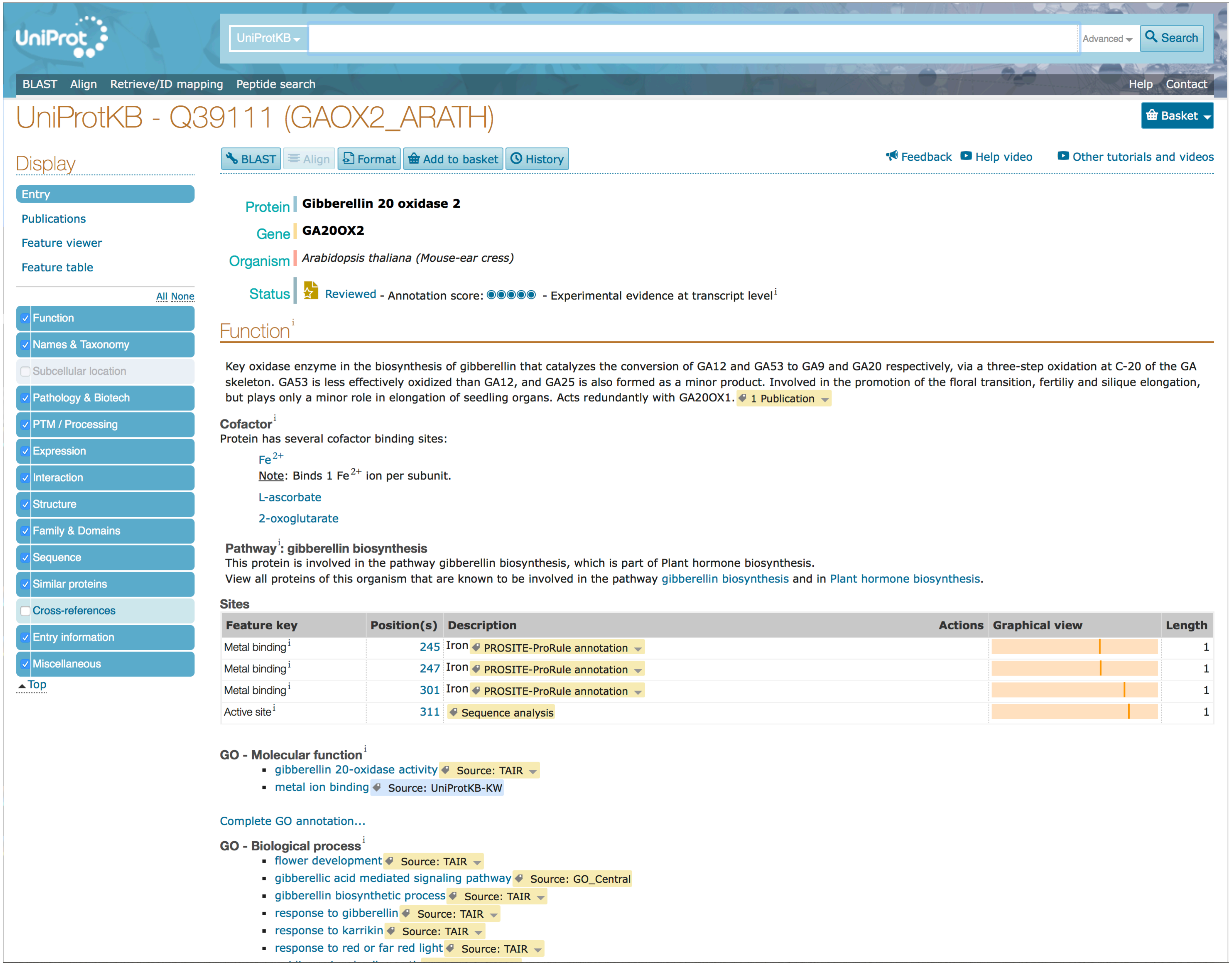
KEGG REACTION: R07184
www.genome.jp/dbget-bin/www_bget?reaction+R07184... ▾ このページを訳す
Name. (gibberellin-44),2-oxoglutarate:oxygen oxidoreductase. Definition. Gibberellin A24 <=> Gibberellin A9 + CO2. Equation. C11861 <=> C11863 + C00011. Comment. **GA 20-oxidase(2-**oxoglutarate dependent dioxygenase), GA20ox

Gene - Detail - SHIGEN
https://shigen.nig.ac.jp/rice/oryzabase/gene/detail/470 ▾ このページを訳す
2017/07/07 - Gene Name Synonym, dee-geo-woo-gen dwarf(sd1-d), green revolution gene, semidwarf-1, GA C20oxidase2, semidwarf-1, reimei dwarf(sd1-r), Calrose76(sd1- 3), Jikkoku(sd1-2). rice GA 20-oxidase2, **GA 20-oxidase 2**, ...

Biosafety of Forest Transgenic Trees: Improving the Scientific Basis ...
https://books.google.co.jp/books?isbn=9401775311 - このページを訳す
Cristina Vettori, Fernando Gallardo, Hely Häggman - 2016 - Nature
(2015) 35S::GA2oxidase1,6::OCS 35S::GA2oxidase3::OCS SHI1::SHI1::SHI1 PHOR1::PHOR1::PHOR1 GA2ox1::GA2oxidase2::NOS RGL1-1::**GA20oxidase2**:: NOS Empty vector Phytochrome trial 717 35S::PHYB1::OCS 2004– 2013 220 ...

oxidoreductase, 2OG-Fe(II) oxygenase family protein - Protein Result
https://www.ncbi.nlm.nih.gov/protein?term=oxidoreductase... - このページを訳す
RecName: Full=Gibberellin 20 oxidase 2; AltName: Full=**GA 20-oxidase 2**; AltName: Full=Gibberellin C-20 oxidase 2; AltName: Full=Os20ox2; AltName: Full =Protein semidwarf-1 389 aa protein P0C5H5.1 GI:158705787 2. RecName: ...

EARLY FLOWERING3 Regulates Flowering in Spring Barley by ...
https://www.ncbi.nlm.nih.gov/pubmed/24781117 - このページを訳す
SA Boden 著 - 2014 - 被引用数: 33 - 関連記事



UniProtKB - Q39111 (GAOX2_ARATH)

Protein **Gibberellin 20 oxidase 2**

Gene **GA20OX2**

Organism *Arabidopsis thaliana (Mouse-ear cress)*

Status Reviewed - Annotation score: ●●●●● - Experimental evidence at transcript level

Function

Key oxidase enzyme in the biosynthesis of gibberellin that catalyzes the conversion of GA12 and GA53 to GA9 and GA20 respectively, via a three-step oxidation at C-20 of the GA skeleton. GA53 is less effectively oxidized than GA12, and GA25 is also formed as a minor product. Involved in the promotion of the floral transition, fertiliy and silique elongation, but plays only a minor role in elongation of seedling organs. Acts redundantly with GA20OX1. 1 Publication

Cofactor
Protein has several cofactor binding sites:

Fe^{2+}
Note: Binds 1 Fe^{2+} ion per subunit.

L-ascorbate

2-oxoglutarate

Pathway: gibberellin biosynthesis
This protein is involved in the pathway gibberellin biosynthesis, which is part of Plant hormone biosynthesis.
View all proteins of this organism that are known to be involved in the pathway gibberellin biosynthesis and in Plant hormone biosynthesis.

Sites

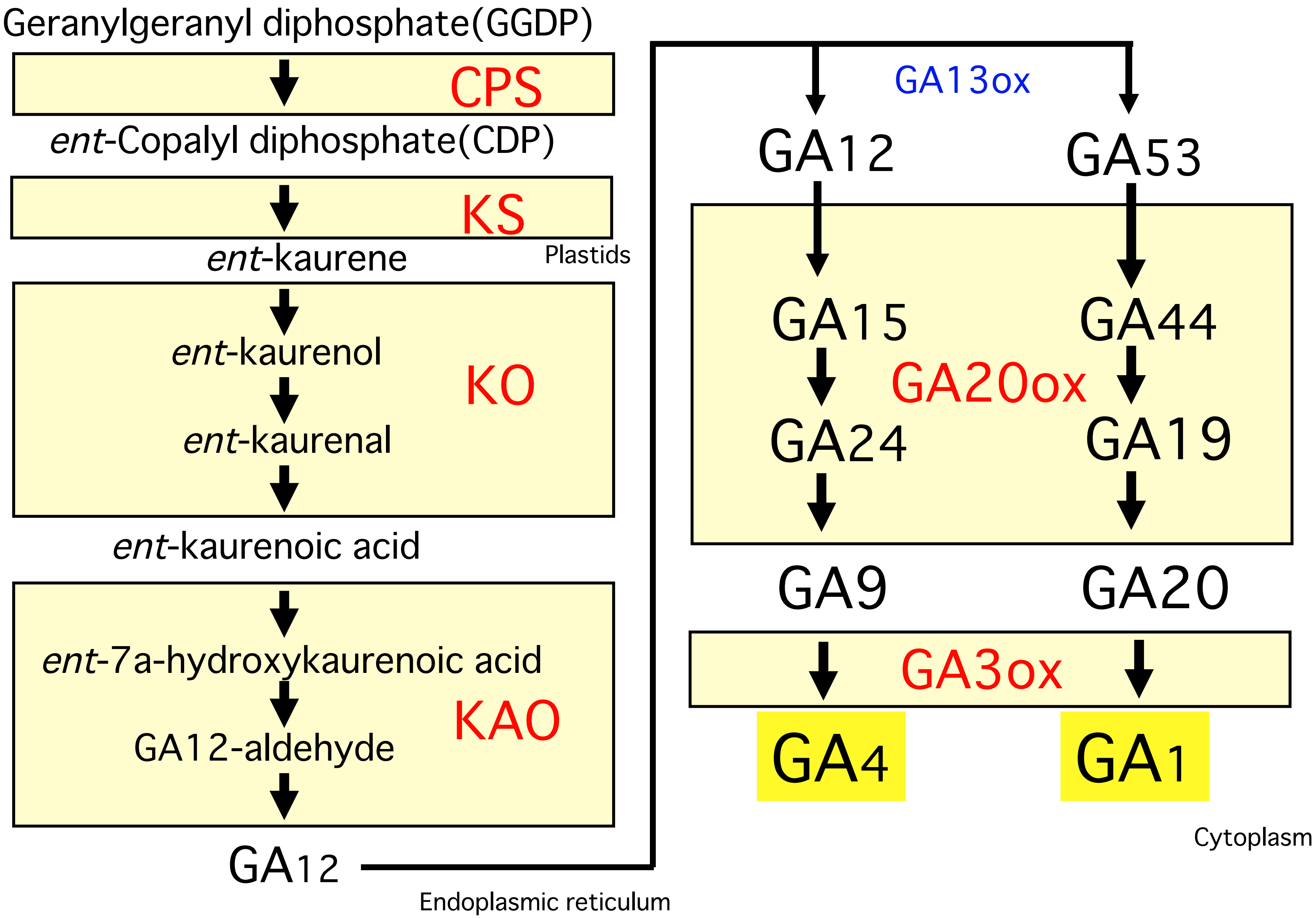
Feature key	Position(s)	Description	Actions	Graphical view	Length
Metal binding	245	Iron PROSITE-ProRule annotation			1
Metal binding	247	Iron PROSITE-ProRule annotation			1
Metal binding	301	Iron PROSITE-ProRule annotation			1
Active site	311	Sequence analysis			1

GO - Molecular function
- gibberellin 20-oxidase activity Source: TAIR
- metal ion binding Source: UniProtKB-KW

Complete GO annotation...

GO - Biological process
- flower development Source: TAIR
- gibberellic acid mediated signaling pathway Source: GO_Central
- gibberellin biosynthetic process Source: TAIR
- response to gibberellin Source: TAIR
- response to karrikin Source: TAIR
- response to red or far red light Source: TAIR

植物におけるGA合成経路



Geranylgeranyl diphosphate(GGDP)

CPS

ent-Copalyl diphosphate(CDP)

KS

ent-kaurene

Plastids

ent-kaurenol

KO

ent-kaurenal

ent-kaurenoic acid

ent-7a-hydroxykaurenoic acid

KAO

GA12-aldehyde

GA12

Endoplasmic reticulum

GA13ox

GA12

GA53

GA15

GA44

GA20ox

GA24

GA19

GA9

GA20

GA3ox

GA4

GA1

Cytoplasm

約 1,710 件 （0.38 秒）

名古屋大学生物機能開発利用研究センター植物分子育種分野
bbc.agr.nagoya-u.ac.jp/~pmb/HANWAISEI.htm ▾
GA20酸化酵素遺伝子はその5番目の反応を触媒する。**sd1**変異体が、ジベレリン合成不全により**イネ**の生長が阻害され矮性になったとしたら、それ以外の合成反応がダメになっても、同じように**イネ**の矮性が引き起こされるはずで、それらの矮性変異体を用いれ ...

半矮性遺伝子 - Wikipedia
https://ja.wikipedia.org/wiki/半矮性遺伝子 ▾
半矮性遺伝子（はんわいせいいでんし、semidwarf gene）とは、植物の背の高さ（草丈） を低くする遺伝子。主に植物育種学で用いられる用語。緑の革命において**イネ**、コムギ品種の改良に用いられ、その子孫に当たる品種に引き継がれている。 ... しかしながら、 先述の**イネ**の半矮性遺伝子**sd1**および小麦農林10号の半矮性遺伝子Rht1,Rht2 の作用機構には植物ホルモンの一種であるジベレリンが関与することが明らかになっている。
矮性と半矮性 · 歴史 · 生理 · 脚注

[PDF] 古代人の 緑の革命 イネにおける 緑の革命 遺伝子の人類への ... - 神...
www.kobe-u.ac.jp/documents/research/news/H23/pp2011_05_06-02.pdf ▾
2011/06/20 - 1940 1960年代に**イネ**やコムギで 緑の革命が起こり、収量が飛躍的に増加しました。 **イネ**で大きく貢献した遺伝子は**SD1**（Semi-dwarf1、半矮性遺伝子、この変異体の草丈は低. くなる）であり、現在でも世界中の多くの品種でこの遺伝子が ...

[PDF] イネの収量を決定する重要遺伝子を同定 - 理化学研究所
www.riken.jp/~/media/riken/pr/press/2005/20050624_1/20050624_1.pdf ▾
2005/06/24 - 一総長）は、**イネ**の収量増加に関わる重要な遺伝子が第一染色体上腕部の「OsCKX2」 . であることを ... さらにこの知見を分子育種に利用し、Gn1 と背丈を低くする QTL(**sd1**) ... し、**イネ**の緑の革命の原因遺伝子であった **sd1**※4 と同一でした。

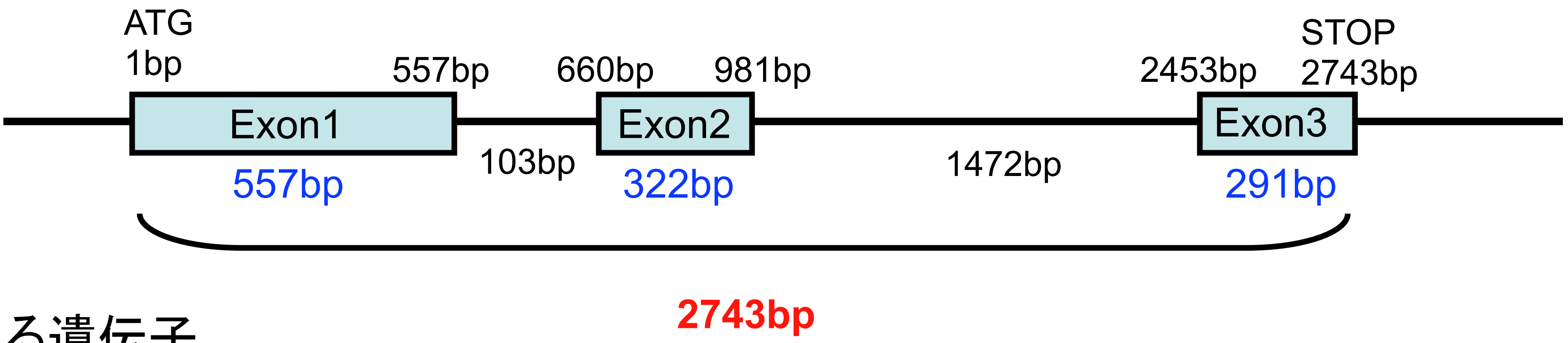
イネの半矮性遺伝子sd-1と 脱粒性遺伝子の連鎖 緒 言 実験材料 ... - J-St...
https://www.jstage.jst.go.jp/article/jsta1957/37/2/37_2_115/_pdf
S OBA 著 - 1993 - 被引用数: 3 - 関連記事
1993. **イネ**の半矮性遺伝子**sd-1**と 脱粒性遺伝子の連鎖. 第2報. インド型半矮性在来品種低脚烏尖の持っ半矮性遺伝子 **sd-1**と 脱粒性遺伝子の連鎖分析. 大 場 伸 哉 ・菊 池 文 雄*. 岐阜大学農学部. 〒501-11岐. 阜市柳戸1-1,. *筑波大学農林学系,〒305つ.

作物ゲノム育種研究センター:Sd1 | 農研機構
www.naro.affrc.go.jp › ... › ゲノム育種支援 › ゲノム育種マーカー情報(イネ) ▾
農研機構は食料・農業・農村に関する研究開発を行う機関です。 マーカー名(遺伝子名、 遺伝子座名) **Sd1** 形質 耐倒伏性 特性 収量性 活用...

短稈遺伝子sd1のアリルにより、イネ玄米のタンパク質含量がコントロ...
www.naro.affrc.go.jp › ... › 中央農業総合研究センター 2015年の成果情報 ▾
農研機構は食料・農業・農村に関する研究開発を行う機関です。玄米のタンパク質含量のQTLが第1染色体に有り、インディカ型劣性で高タンパク質含量を示す。この原因遺伝子は短稈遺伝子**sd1**であり、低脚烏尖およびCalrose 76由来 ...

第1回 組み換えと異なるもうひとつの遺伝子技術 - ナショナルジオグラ

ある遺伝子Aの構造

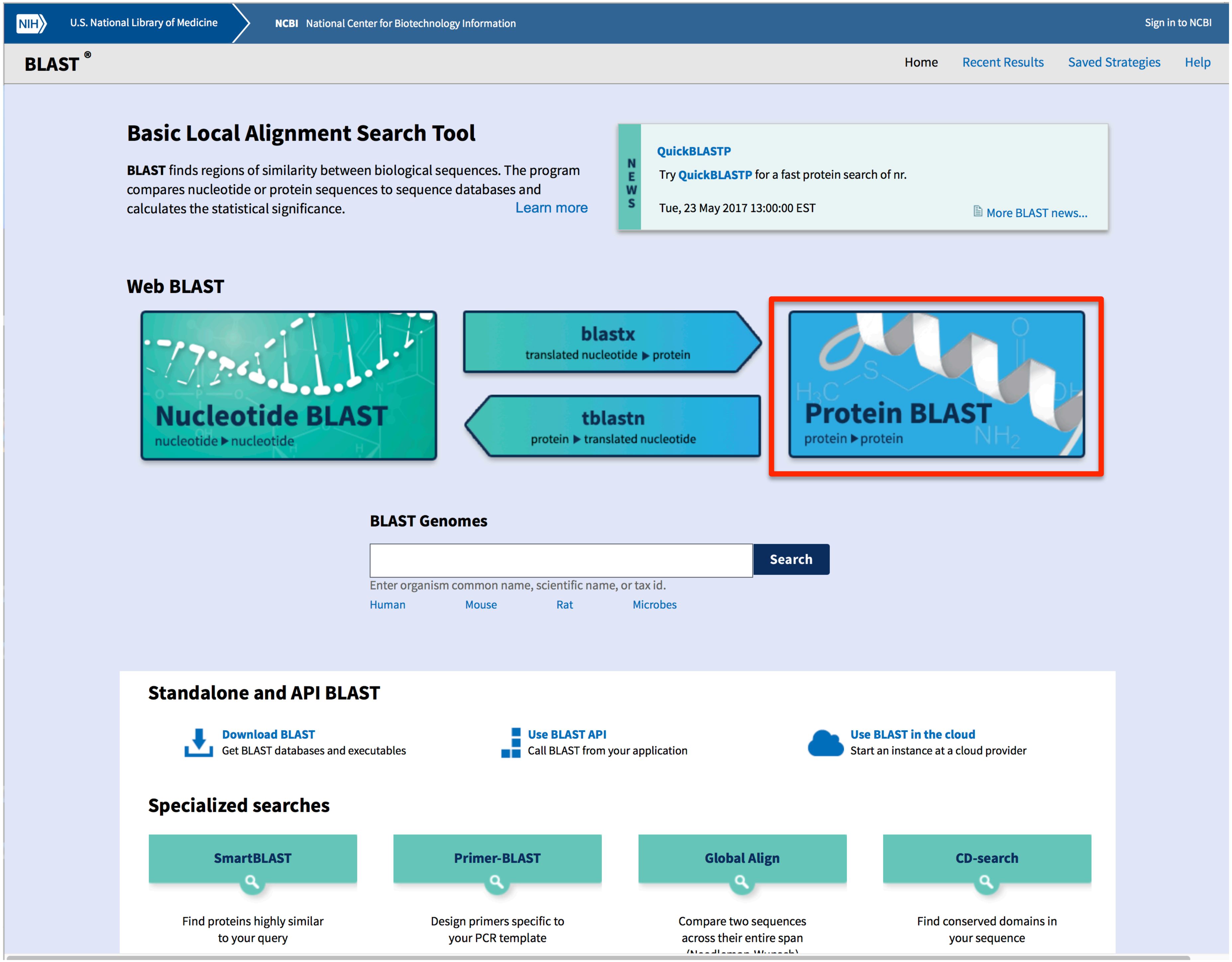


ある遺伝子
3つのエクソン（第1エキソン 557bp, 第2エキソン 322bp, 第1エキソン 291bp）、
2つのイントロン（第1イントロン 103bp, 第2イントロン 1472bp）からなる。
遺伝子の長さは、2743bp(但しプロモーター領域は含まず)

アミノ酸配列

```
MVAEHPTPPQPHQPPPMDSTAGSGIAAPAAAAVCDLRMEPKIPEPFVWPNGDARPASAAE
LDMPVVDVGVLRDGDAEGLRRAAAQVAAACATHGFFQVSEHGVDAALARAALDGASDFFR
LPLAEKRRARRVPGTVSGYTSAHADRFASKLPWKETLSFGFHDRAAAPVVADYFSSTLGP
DFAPMGRVYQKYCEEMKELSLTIMELLELSLGVERGYYREFFADSSSIMRCNYYPPCPEP
ERTLGTGPHCDPTALTILLQDDVGGLEVLVDGEWRPVSPVPGAMVINIGDTFMALSNGRY
KSCLHRAVVNQRRERRSLAFFLCPREDRVVRPPPSAATPQHYPDFTWADLMRFTQRHYRA
DTRTLDAFTRWLAPPAADAAATAQVEAAS*
```

予想される遺伝子：ジベレリンの生合成酵素遺伝子
予想される機能：ジベレリンの合成を担っている。



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

QuickBLASTP

Try **QuickBLASTP** for a fast protein search of nr.

Tue, 23 May 2017 13:00:00 EST

More BLAST news...

Web BLAST

BLAST Genomes

Enter organism common name, scientific name, or tax id.

Standalone and API BLAST

Download BLAST
Get BLAST databases and executables

Use BLAST API
Call BLAST from your application

Use BLAST in the cloud
Start an instance at a cloud provider

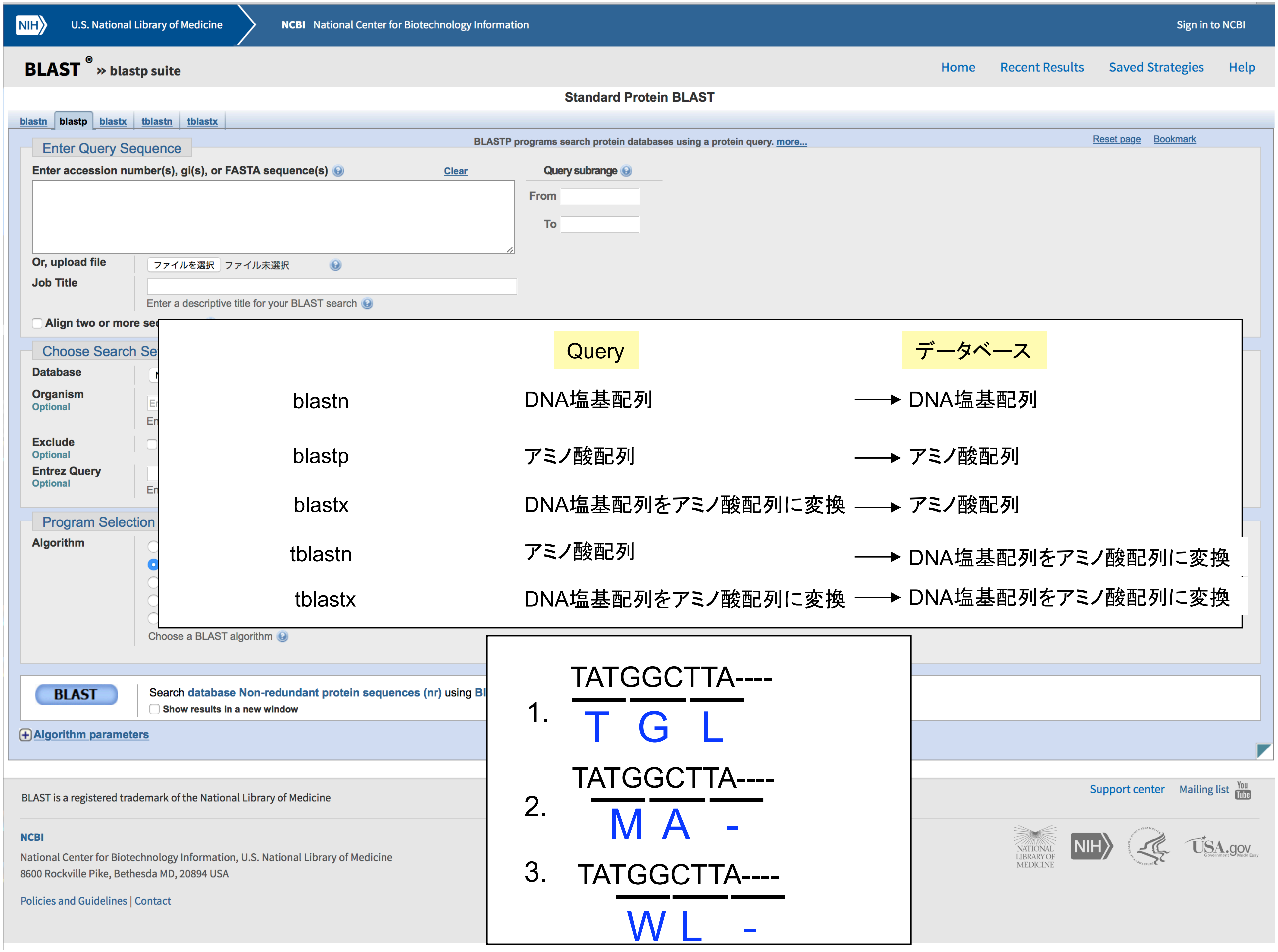
Specialized searches

SmartBLAST
Find proteins highly similar to your query

Primer-BLAST
Design primers specific to your PCR template

Global Align
Compare two sequences across their entire span

CD-search
Find conserved domains in your sequence



	Query		データベース
blastn	DNA塩基配列	→	DNA塩基配列
blastp	アミノ酸配列	→	アミノ酸配列
blastx	DNA塩基配列をアミノ酸配列に変換	→	アミノ酸配列
tblastn	アミノ酸配列	→	DNA塩基配列をアミノ酸配列に変換
tblastx	DNA塩基配列をアミノ酸配列に変換	→	DNA塩基配列をアミノ酸配列に変換

1. TATGGCTTA----
T G L

2. TATGGCTTA----
M A -

3. TATGGCTTA----
W L -