Find-A-Gene Project Assignment

Name: Steven Gan   Quarter: Fall
PID: A59020397   Course: BGGN-213
E-mail: digan@ucsd.edu   Instructor: Dr. Barry Grant

Q1:

Name: CCCTC-binding factor (CTCF)

Accession: isoform 1   isoform 2   isoform 3
              NP_006556.1   NP_001177951.1   NP_001350845.1

Species: *Homo Sapiens*

Q2:

Isoform 1 will be used for downstream analysis, as it is the longest isoform.

Query: NP_006556.1

Method: TBLASTN (2.13.0+) search against with default parameters

Database: Expressed Sequence Tags (est)

Organism: Papilionoidea (taxid:37572)

TBLASTN Setting:

Chosen Match: Accession GE695232.1, a 750 base pair clone from *Bicyclus anynana*. Alignment details see below.

| Descriptions | Graphic Summary | Alignments | Taxonomy |

hover to see the title · click to show alignments

Alignment Scores: ■ < 40 ■ 40 - 50 ■ 50 - 80 ■ 80 - 200 ■ >= 200

100 sequences selected

Distribution of the top 510 Blast Hits on 100 subject sequences

CBUN9529.b1 B.anynana_wing.0-3d_2-10kb Bicyclus anynana..

Score:142 Evalue:1.5e-37   Accession:GE695232.1
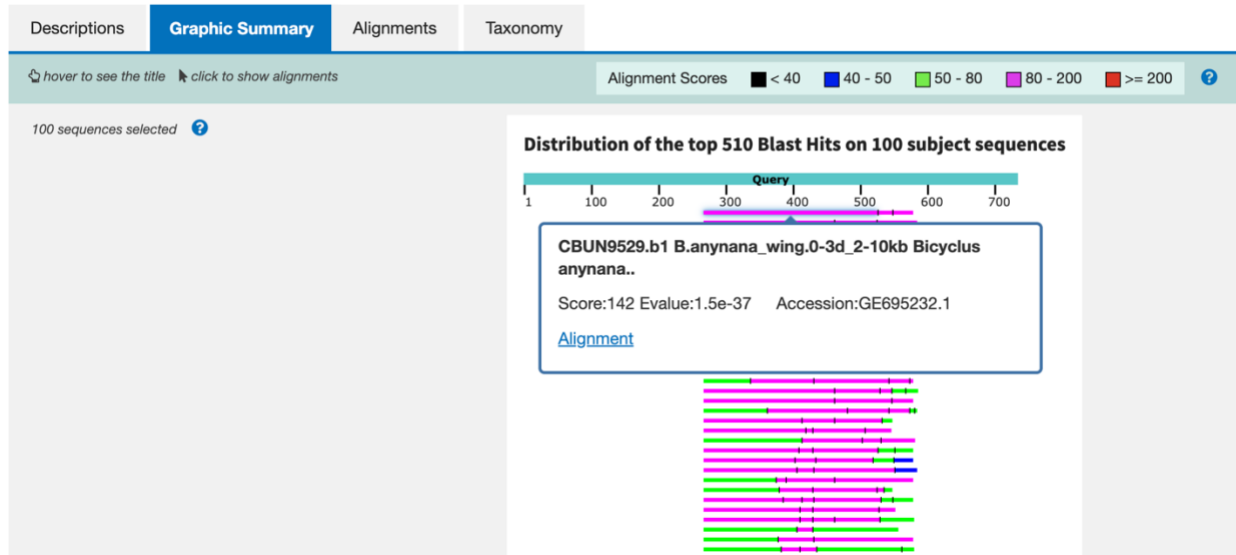
Alignment

| Descriptions | Graphic Summary | **Alignments** | Taxonomy |

Alignment view: Pairwise   Download

100 sequences selected

Download · GenBank Graphics · Sort by: E value   ▼Next ▲Previous ◄Descriptions

CBUN9529.b1 B.anynana_wing.0-3d_2-10kb Bicyclus anynana cDNA clone CBUN9529, mRNA sequence

Sequence ID: GE695232.1 Length: 750 Number of Matches: 3

Range 1: 3 to 734 GenBank Graphics   ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
154 bits(389)	1e-41	Compositional matrix adjust.	95/284(33%)	144/284(50%)	47/284(16%)	+3

```
Query  267  QCELCSYTCPRRSNLDRHMKSHTDERPHKCHL----CGRAFRTVTLLRNHLNTHTGTRPH  322
            C++C Y C +R NL  H+++HT E+P+ C +    C R  R    LR+H+ THTG +P
Sbjct  3    SCDICHYKCAQRGNLVCHIRTHTCEKPYSCEMYNYKCARKSR----LRHHMTTHTGEKPF  170

Query  323  KCPDCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCS  382
            C  C+     LV H R  HT EKPF C +C+Y      L  H+++HTGE+PF C
Sbjct  171  SCGICNYKTGVKNSLVCHLR-THTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCG  347

Query  383  LCSYASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHC  442
            +C+Y +     L  H+RTH+GEKP+ C IC+ +F
Sbjct  348  ICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKF--------------------------  449

Query  443  DTVIARKSDLGVHLRKQHSYIEQGKK---CRYCDAVFHERYALIQHQKSHKNEKRFKCDQ  499
            A K +L  H++     I  G+K   C  C+     + +L+ H ++H  EK F C+
Sbjct  450  ----ALKHNLVNHMK-----IHTGEKPFSCEICNYKTRVKNSLVSHLRTHTGEKPFSCEI  602

Query  500  CDYACRQERHMIMHKRTHTGEKPYACSHCDKTFRQKQLLDMHFK  543
            C+Y   ++R+++ H +THTGEKP++C  C+     K  L  H +
Sbjct  603  CNYKSARKRYLLNHMKTHTGEKPFSCDICNYKTGIKNSLVRHMR  734
```

Range 2: 84 to 749 GenBank Graphics   ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
142 bits(358)	1e-37	Compositional matrix adjust.	89/255(35%)	125/255(49%)	33/255(12%)	+3

```
Query  266  FQCELCSYTCPRRSNLDRHMKSHTDERPHKCHLCGRAFRTVTLLRNHLNTHTGTRPHKCP  325
            + CE+ +Y C R+S L  HM +HT E+P  C +C       L  HL THTG +P  C
Sbjct  84   YSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCE  263

Query  326  DCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCSLCS  385
            C+  F    L+ H +  HT EKPF C +C+Y +   + L  H+R+HTGE+PF C +C+
Sbjct  264  ICNYKFALKRNLLNHMK-THTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICN  440

Query  386  YASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHCDTV  445
            Y    + L  HM+ H+GEKP+ C IC+ +     ++  H L+ HT    F C  C+
Sbjct  441  YKFALKHNLVNHMKIHTGEKPFSCEICNYKTRVKNSLVSH-LRTHTGE-KPFSCEICNYK  614

Query  446  IARKSDLGVHLRKQHSYIEQGKKCRYCDAVFHERYALIQHQKSHKNEKRFKCDQCDYACR  505
            ARK                   RY          L+ H K+H  EK F CD C+Y
Sbjct  615  SARK--------------------RY----------LLNHMKTHTGEKPFSCDICNYKTG  704

Query  506  QERHMIMHKRTHTGE  520
            +  ++ H R HTGE
Sbjct  705  IKNSLVRHMRIHTGE  749
```

Range 3: 6 to 728 GenBank Graphics   ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
136 bits(343)	2e-35	Compositional matrix adjust.	87/250(35%)	126/250(50%)	9/250(3%)	+3

```
Query  324  CPDCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCSL  383
            C  C      G LV H R  HT EKP+ C M +Y     S+L+ H+ +HTGE+PF C +
Sbjct  6    CDICHYKCAQRGNLVCHIR-THTCEKPYSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGI  182

Query  384  CSYASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHCD  443
            C+Y +     L  H+RTH+GEKP+ C IC+ +F     + H ++ HT    F C  C+
Sbjct  183  CNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKRNLLNH-MKTHTGE-KPFSCGICN  356

Query  444  TVIARKSDLGVHLRKQHSYIEQGKKCRYCDAVFHERYALIQHQKSHKNEKRFKCDQCDYA  503
            K+ L  HLR     E+   C  C+  F  ++ L+ H K H  EK F C+ C+Y
Sbjct  357  YKTGVKNSLVCHLRTHTG--EKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEICNYK  530

Query  504  CRQERHMIMHKRTHTGEKPYACSHCDKTFRQKQLLDMHFKRYHDPNFVPAAFVCSKCGKT  563
            R + ++ H RTHTGEKP++C  C+    +K+ L  H K +    F C  C
Sbjct  531  TRVKNSLVSHLRTHTGEKPFSCEICNYKSARKRYLLNHMKTHTG----EKPFSCDICNYK  698

Query  564  FTRRNTMARH  573
            +N++ RH
Sbjct  699  TGIKNSLVRH  728
```

Alignment Details:

CBUN9529.b1 B.anynana_wing.0-3d_2-10kb Bicyclus anynana cDNA clone CBUN9529, mRNA sequence

Sequence ID: GE695232.1   **Length:** 750   **Number of Matches: 3**

Range 1: 3 to 734

Alignment statistics for match #1

Score	Expect	Method	Identities	Positives	Gaps	Frame
154 bits (389)	1e-41	Compositional matrix adjust.	95/284(33%)	144/284(50%)	47/284(16%)	+3

```
Query  267  QCELCSYTPCRRSNLDRHMKSHTDERPHKCHL----CGRAFRTVTLLRNHLNTHTGTRPH  322
            C++C Y C +R NL  H+++HT E+P+ C +     C R  R    LR+H+ THTG +P
Sbjct  3    SCDICHYKCAQRGNLVCHIRTHTCEKPYSCEMYNYKCARKSR----LRHHMTTHTGEKPF  170

Query  323  KCPDCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCS  382
             C  C+       LV H R  HT EKPF C +C+Y       L  H+++HTGE+PF C
Sbjct  171  SCGICNYKTGVKNSLVCHLR-THTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCG  347

Query  383  LCSYASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHC  442
            +C+Y +     L  H+RTH+GEKP+ C IC+ +F
Sbjct  348  ICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKF-------------------------  449

Query  443  DTVIARKSDLGVHLRKQHSYIEQGKK---CRYCDAVFHERYALIQHQKSHKNEKRFKCDQ  499
              A K +L  H++      I  G+K   C  C+     + +L+ H ++H  EK F C+
Sbjct  450  ----ALKHNLVNHMK-----IHTGEKPFSCEICNYKTRVKNSLVSHLRTHTGEKPFSCEI  602

Query  500  CDYACRQERHMIMHKRTHTGEKPYACSHCDKTFRQKQLLDMHFK  543
            C+Y   ++R+++ H +THTGEKP++C  C+      K  L  H +
Sbjct  603  CNYKSARKRYLLNHMKTHTGEKPFSCDICNYKTGIKNSLVRHMR  734
```

Range 2: 84 to 749

Alignment statistics for match #2

Score	Expect	Method	Identities	Positives	Gaps	Frame
142 bits (358)	1e-37	Compositional matrix adjust.	89/255(35%)	125/255(49%)	33/255(12%)	+3

```
Query  266  FQCELCSYTPCRRSNLDRHMKSHTDERPHKCHLCGRAFRTVTLLRNHLNTHTGTRPHKCP  325
            + CE+ +Y C R+S L  HM +HT E+P  C +C          L  HL THTG +P  C
Sbjct  84   YSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCE  263

Query  326  DCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCSLCS  385
             C+  F      L+ H +  HT EKPF C +C+Y +     + L  H+R+HTGE+PF C +C+
Sbjct  264  ICNYKFALKRNLLNHMK-THTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICN  440

Query  386  YASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHCDTV  445
```

```
        Y     + L  HM+ H+GEKP+ C IC+ +     ++  H L+ HT     F C  C+
Sbjct  441  YKFALKHNLVNHMKIHTGEKPFSCEICNYKTRVKNSLVSH-LRTHTGE-KPFSCEICNYK  614

Query  446  IARKSDLGVHLRKQHSYIEQGKKCRYCDAVFHERYALIQHQKSHKNEKRFKCDQCDYACR  505
             ARK                    RY        L+ H K+H  EK F CD C+Y
Sbjct  615  SARK-------------------RY---------LLNHMKTHTGEKPFSCDICNYKTG  704

Query  506  QERHMIMHKRTHTGE  520
             +  ++ H R HTGE
Sbjct  705  IKNSLVRHMRIHTGE  749
```

Range 3: 6 to 728

Alignment statistics for match #3

Score	Expect	Method	Identities	Positives	Gaps	Frame
136 bits (343)	2e-35	Compositional matrix adjust.	87/250(35%)	126/250(50%)	9/250(3%)	+3

```
Query  324  CPDCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCSL  383
             C C     G LV H R  HT EKP+ C M +Y     S+L+ H+ +HTGE+PF C +
Sbjct  6    CDICHYKCAQRGNLVCHIR-THTCEKPYSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGI  182

Query  384  CSYASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHCD  443
             C+Y +     L H+RTH+GEKP+ C IC+ +F    + H ++ HT     F C  C+
Sbjct  183  CNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKRNLLNH-MKTHTGE-KPFSCGICN  356

Query  444  TVIARKSDLGVHLRKQHSYIEQGKKCRYCDAVFHERYALIQHQKSHKNEKRFKCDQCDYA  503
              K+ L  HLR       E+  C  C+  F  ++ L+ H K H  EK F C+ C+Y
Sbjct  357  YKTGVKNSLVCHLRTHTG--EKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEICNYK  530

Query  504  CRQERHMIMHKRTHTGEKPYACSHCDKTFRQKQLLDMHFKRYHDPNFVPAAFVCSKCGKT  563
             R +  ++ H RTHTGEKP++C  C+     +K+ L  H K +          F C  C
Sbjct  531  TRVKNSLVSHLRTHTGEKPFSCEICNYKSARKRYLLNHMKTHTG----EKPFSCDICNYK  698

Query  564  FTRRNTMARH  573
             +N++ RH
Sbjct  699  TGIKNSLVRH  728
```

Q3:

Chosen sequence:

>B. anynana protein (from BLAST results)
SCDICHYKCAQRGNLVCHIRTHTCEKPYSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGICNYKTGVKNSLVCHLRT
HTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHN
LVNHMKIHTGEKPFSCEICNYKTRVKNSLVSHLRTHTGEKPFSCEICNYKSARKRYLLNHMKTHTGEKPFSCDICNY
KTGIKNSLVRHMRIHTG

All six reading frame:
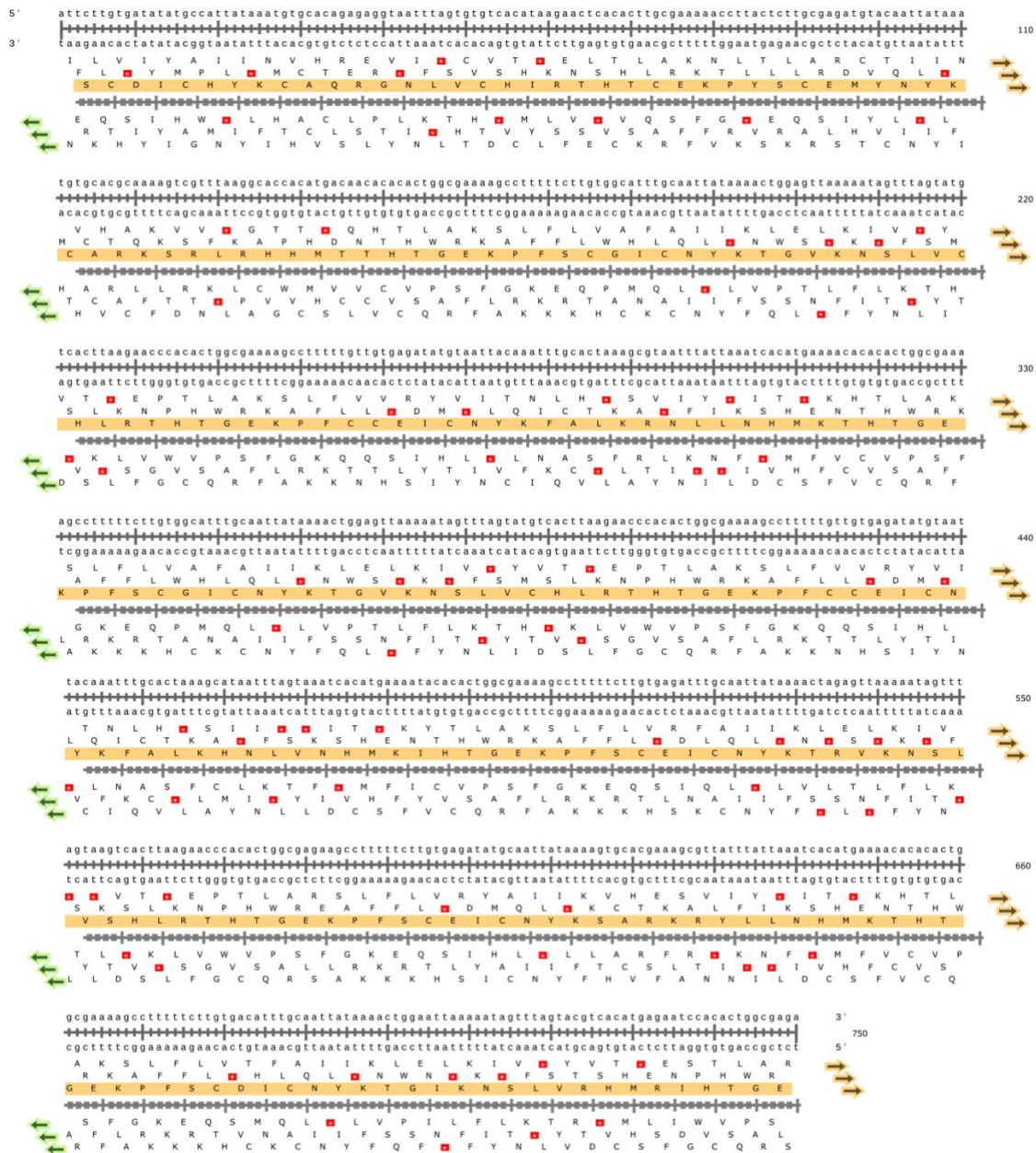
Name: *Bicyclus* CTCF-like protein

Species: *Bicyclus anynana*

Taxonomy: Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Papilionoidea; Nymphalidae; Satyrinae; Satyrini; Mycalesina; Bicyclus.

Q4:

BLASTP search on non-redundant protein sequences (nr) hits top on zinc finger protein 84-like proteins on *Bicyclus anynana*, with identity percentage of 95.16%, suggesting a possible novel protein. See details below.

BLASTP setting:

BLASTP results:

Alignment details:

Range 2: 532 to 779 GenPept Graphics   ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
409 bits(1051)	1e-134	Compositional matrix adjust.	200/248(81%)	216/248(87%)	0/248(0%)

```
Query   1    SCDICHYKCAQRGNLVCHIRTHTCEKPYSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGI   60
             SC IC+YK   + +LVCH+RTHT EKP+ CE+ NYK A K  L +HM THTGEKPFSCGI
Sbjct   532  SCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCGI   591

Query   61   CNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCGICNYK   120
             CNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALK NL+NHMK HTGEKPFSC ICNYK
Sbjct   592  CNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEICNYK   651

Query   121  TGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEICNYKTRVK   180
             T VKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMK HTGEKPFSC+ICNYKT +K
Sbjct   652  TRVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKTHTGEKPFSCDICNYKTGIK   711

Query   181  NSLVSHLRTHTGEKPFSCEICNYKSARKRYLLNHMKTHTGEKPFSCDICNYKTGIKNSLV   240
             N+LV H+RTHTGEKPFSCEICN+KSA K  LL+HMKTHTGEKPFSC ICNYK   K  L+
Sbjct   712  NTLVRHMRTHTGEKPFSCEICNHKSALKHSLLSHMKTHTGEKPFSCKICNYKCVRKQHLL   771

Query   241  RHMRIHTG   248
             HM+ HTG
Sbjct   772  GHMKTHTG   779
```

Range 3: 588 to 809 GenPept Graphics   ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
334 bits(857)	8e-106	Compositional matrix adjust.	165/222(74%)	187/222(84%)	0/222(0%)

```
Query   1    SCDICHYKCAQRGNLVCHIRTHTCEKPYSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGI   60
             SC IC+YK   + +LVCH+RTHT EKP+ CE+ NYK A K  L +HM  HTGEKPFSCGI
Sbjct   588  SCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEI   647

Query   61   CNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCGICNYK   120
             CNYKT VKNSLVCHLRTHTGEKPFCCEICNYKFALK NL+NHMKTHTGEKPFSC DICNYK
Sbjct   648  CNYKTRVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKTHTGEKPFSCDICNYK   707

Query   121  TGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEICNYKTRVK   180
             TG+KN+LV H+RTHTGEKPF CEICN+K ALKH+L++HMK HTGEKPFSC+ICNYK   K
Sbjct   708  TGIKNTLVRHMRTHTGEKPFSCEICNHKSALKHSLLSHMKTHTGEKPFSCKICNYKCVRK   767

Query   181  NSLVSHLRTHTGEKPFSCEICNYKSARKRYLLNHMKTHTGEK   222
             L+ H++THTGEK F C++CNYK ARKR L +HMKTHTG K
Sbjct   768  QHLLGHMKTHTGEKSFCCKLCNYKCARKRDLESHMKTHTGGK   809
```

Range 4: 447 to 667 GenPept Graphics   ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
317 bits(813)	3e-99	Compositional matrix adjust.	161/221(73%)	182/221(82%)	3/221(1%)

```
Query   31   EMYN-YKCARKSRLRHH--MTTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCE   87
             ++Y+ +K  +K+ L  +  + T T E +SC IC+YK   + +LVCH+RTHTGEKP+ CE
Sbjct   447  QLYDIFKKPKKTVLDENPRVKTLTNEILYSCDICHYKCAQRGNLVCHIRTHTGEKPYSCE   506

Query   88   ICNYKFALKRNLLNHMKTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNY   147
             +CNYK A K  L +HM THTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNY
Sbjct   507  MCNYKCAHKSRLRHHMTTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNY   566

Query   148  KFALKHNLVNHMKIHTGEKPFSCEICNYKTRVKNSLVSHLRTHTGEKPFSCEICNYKSAR   207
             KFALK NL+NHMK HTGEKPFSC ICNYKT VKNSLV HLRTHTGEKPF CEICNYK A
Sbjct   567  KFALKRNLLNHMKTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFAL   626

Query   208  KRYLLNHMKTHTGEKPFSCDICNYKTGIKNSLVRHMRIHTG   248
             K  L+NHMK HTGEKPFSC+ICNYKT +KNSLV H+R HTG
Sbjct   627  KHNLVNHMKIHTGEKPFSCEICNYKTRVKNSLVCHLRTHTG   667
```