Find-A-Gene Project Assignment

|  |  |  |  |
| --- | --- | --- | --- |
| Name: | Steven Gan | Quarter: | Fall |
| PID: | A59020397 | Course: | BGGN-213 |
| E-mail: | [digan@ucsd.edu](mailto: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*

Function: DNA insulation; RNA binding; RNA splicing; DNA loop extrusion; transcriptional regulation; genome instability.

# 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:

Graphical user interface, application

Description automatically generated

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

Graphical user interface, text, application, email, website

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Alignment Details:

**CBUN9529.b1 B.anynana\_wing.0-3d\_2-10kb Bicyclus anynana cDNA clone CBUN9529, mRNA sequence**

**Sequence ID:**[**GE695232.1**](https://www.ncbi.nlm.nih.gov/nucleotide/GE695232.1?report=genbank&log$=nuclalign&blast_rank=1&RID=NCCEP04Y013) **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 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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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 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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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:**

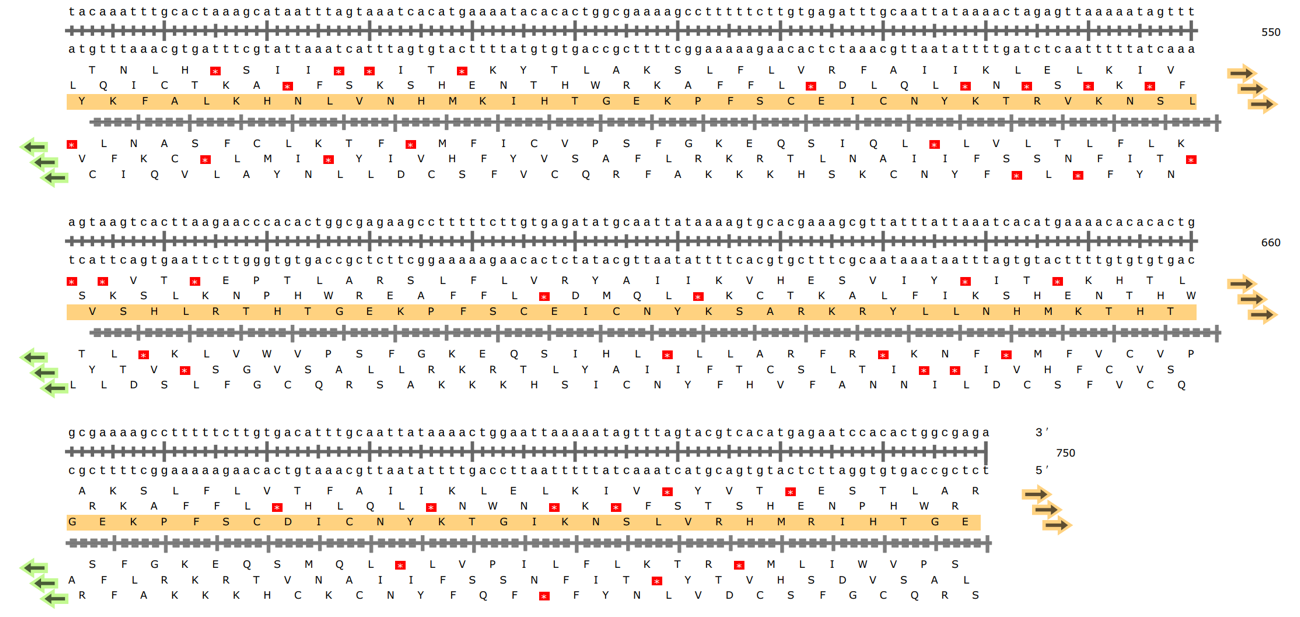
>B. anynana protein (from BLAST results)

SCDICHYKCAQRGNLVCHIRTHTCEKPYSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEICNYKTRVKNSLVSHLRTHTGEKPFSCEICNYKSARKRYLLNHMKTHTGEKPFSCDICNYKTGIKNSLVRHMRIHTG

All six reading frame:

A picture containing diagram

Description automatically generated



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: Graphical user interface, application

Description automatically generated

BLASTP results: Graphical user interface, application

Description automatically generated

Alignment details:

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

# Q5:

**Sequences for multiple alignment:**

>Bicyclus\_anynana (from BLAST results)

SCDICHYKCAQRGNLVCHIRTHTCEKPYSCEMYNYKCARKSRLRHHMTTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKRNLLNHMKTHTGEKPFSCGICNYKTGVKNSLVCHLRTHTGEKPFCCEICNYKFALKHNLVNHMKIHTGEKPFSCEICNYKTRVKNSLVSHLRTHTGEKPFSCEICNYKSARKRYLLNHMKTHTGEKPFSCDICNYKTGIKNSLVRHMRIHTG

>Homo\_sapiens ref|NP\_006556.1| transcriptional repressor CTCF isoform 1 [Homo sapiens]

MEGDAVEAIVEESETFIKGKERKTYQRRREGGQEEDACHLPQNQTDGGEVVQDVNSSVQMVMMEQLDPTLLQMKTEVMEGTVAPEAEAAVDDTQIITLQVVNMEEQPINIGELQLVQVPVPVTVPVATTSVEELQGAYENEVSKEGLAESEPMICHTLPLPEGFQVVKVGANGEVETLEQGELPPQEDPSWQKDPDYQPPAKKTKKTKKSKLRYTEEGKDVDVSVYDFEEEQQEGLLSEVNAEKVVGNMKPPKPTKIKKKGVKKTFQCELCSYTCPRRSNLDRHMKSHTDERPHKCHLCGRAFRTVTLLRNHLNTHTGTRPHKCPDCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCSLCSYASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHCDTVIARKSDLGVHLRKQHSYIEQGKKCRYCDAVFHERYALIQHQKSHKNEKRFKCDQCDYACRQERHMIMHKRTHTGEKPYACSHCDKTFRQKQLLDMHFKRYHDPNFVPAAFVCSKCGKTFTRRNTMARHADNCAGPDGVEGENGGETKKSKRGRKRKMRSKKEDSSDSENAEPDLDDNEDEEEPAVEIEPEPEPQPVTPAPPPAKKRRGRPPGRTNQPKQNQPTAIIQVEDQNTGAIENIIVEVKKEPDAEPAEGEEEEAQPAATDAPNGDLTPEMILSMMDR

>Maniola\_jurtina ref| XP\_045782119.1 | transcriptional repressor CTCF-like isoform X2 [Maniola jurtina]

MAGICCVDGCDPTAEDVTYFKFPNSRTLRRKWLDAINNSVKVTLDTAVCSRHFLPNQYEVIRGKKRLKAKVVPSVFDNITKPTSPQKEKTDSSDGEDSVPLQKVKSVATDNTDTGQSKQSPDHRLEDRQDSEASVRNGVKEDDIVKRKQPDIVSITDSESNKDIEDIITHYQIKQIRPLHKPTDRTTPDMSIEIEVPLAMDGEMEIGRNREMEMGRNGEIDDVIDVDEEAEPVFIEVAVGKGGGVEETTNEDCMMLLESVQCEVDPSCLMFPPEDPGNDPGDDAGNDSDVIDLGEKKEDPVSLLTSSDEDEVIIEEPKYDMVEVSDETDEDDVPLVRLVDKPSQNKFPKNTKNTDILSETNLTKLLWGRLCEYYCLECRFTSTSNAELRKHMQEHSTQVIQVCEICSYTTSSKHQYIRHKRKHKEDKRFKCHLCKYSARHNMSLIYHLKSHDNGQFVSDMSVFKCEKCNFETDYKVSLMKHIRICSSKSKRYSCAKCSYETDRRSDLKRHKARKHNTGKDGDYEPPAWVTRAKKPKCDK

>Nilaparvata\_lugens ref|XP\_039287301.1| transcriptional repressor CTCF [Nilaparvata lugens]

MSPPDKVQVQTEIKLEDGVTIVPENVTDIQNYLDTFNKEIQGGEQVVQQVGVVAADEGGSEEGTYYVDQAGQYYYQSASCDGQQVMTVVSGLPGASGESGESFVALPASAASSQRDNVGGSAPLLIQAATGGGASASGGGTTVGGATVGGASAEGGGGATYQTVTIVPSETNPGELSYLLIVQQPGDEDEGEDGQKDKDEDDDHDLTVYDFDDAEDVGTVSGMESGDEDDKSKIVKFMPKKSQTVTQAHMCNYCNYTSPKRYLLSRHMKSHSEERPHKCSVCERGFKTLASLQNHVNTHTGTKPHRCKHCDSAFTTSGELVRHVRYKHTHEKPHKCTICDYASVELSKMRNHMRCHTGERPYQCPHCTYASPDTFKLKRHLRIHTGEKPYECDICHARFTQSNSLKAHKLIHSGQLHTSISPPSLLGVLTENSFNDFEDWNDFHFRVDNFSMSALQFAKFSVHSKSHEGEKCWRCELCPYASVSQRHLESHMLIHTDQKPYQCDQCDQSFRQKQLLRRHQNLYHNPNYVPPPPREKTHECPECQRAFRHKGNLIRHLSVHDPESLAQERQMLLKQGRQRKLQNINGQRVEVIPGDEDDEDEDDELNGQVMAVEGSDGQQYVVLEVIQLQDDNGQEQAVAVMAADGGLEQAVAALHGAGAEDDEEELDEEDEEEDDVHITPDMVEDHEMMSSLRHQTRQKTQHDMANCFGFDDDEEEEEEDDIGISLKQSNTKSIHLLRSGLQ

>Mus\_musculus ref|NP\_001390652.1| transcriptional repressor CTCF isoform 1 [Mus musculus]

MEGEAVEAIVEESETFIKGKERKTYQRRREGGQEEDACHLPQNQTDGGEVVQDVNSSVQMVMMEQLDPTLLQMKTEVMEGTVAPEAEAAVDDTQIITLQVVNMEEQPINIGELQLVQVPVPVTVPVATTSVEELQGAYENEVSKEGLAESEPMICHTLPLPEGFQVVKVGANGEVETLEQGELPPQEDSSWQKDPDYQPPAKKTKKTKKSKLRYTEEGKDVDVSVYDFEEEQQEGLLSEVNAEKVVGNMKPPKPTKIKKKGVKKTFQCELCSYTCPRRSNLDRHMKSHTDERPHKCHLCGRAFRTVTLLRNHLNTHTGTRPHKCPDCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYASVEVSKLKRHIRSHTGERPFQCSLCSYASRDTYKLKRHMRTHSGEKPYECYICHARFTQSGTMKMHILQKHTENVAKFHCPHCDTVIARKSDLGVHLRKQHSYIEQGKKCRYCDAVFHERYALIQHQKSHKNEKRFKCDQCDYACRQERHMIMHKRTHTGEKPYACSHCDKTFRQKQLLDMHFKRYHDPNFVPAAFVCSKCGKTFTRRNTMARHADNCAGPDGVEGENGGETKKSKRGRKRKMRSKKEDSSDSEENAEPDLDDNEEEEEPAVEIEPEPEPQPQPPPPPQPVAPAPPPAKKRRGRPPGRTNQPKQNQPTAIIQVEDQNTGAIENIIVEVKKEPDAEPAEGEEEEAQAATTDAPNGDLTPEMILSMMDR

>Drosophila\_melanogaster ref|NP\_648109.1| CTCF [Drosophila melanogaster]

MPRRTKKDEDPEDLQTFLNNFHKEIEGNSDEKVVNTILEAISAEAIDLDENGAEAGGSKPMEEAEADLDHAEEAEEEEEDDEDKYFIDDEGNCYIKTTPKKQKELQKKLKQAAAKPGKATRSVVSTATNKSINLRPAKSTPKATTSKPPPEPKAISVRPARAAAAKAKQSAMPPPPALVVKVPAPRGRPRKNPVIPKPEPMDLERELEELVDEPDISSMVTELSDYTVDEAAVEAATATLTPNEAEVYEFEDNATTEDENADKKDVDFVLSNKEVKLKTASSTSQNSNASGHKYSCPHCPYTASKKFLITRHSRSHDVEPSFKCSICERSFRSNVGLQNHINTHMGNKPHKCKLCESAFTTSGELVRHTRYKHTKEKPHKCTECTYASVELTKLRRHMTCHTGERPYQCPHCTYASQDMFKLKRHMVIHTGEKKYQCDICKSRFTQSNSLKAHKLIHSVVDKPVFQCNYCPTTCGRKADLRVHIKHMHTSDVPMTCRRCGQQLPDRYQYKLHVKSHEGEKCYSCKLCSYASVTQRHLASHMLIHLDEKPFHCDQCPQAFRQRQLLRRHMNLVHNEEYQPPEPREKLHKCPSCPREFTHKGNLMRHMETHDDSANAREKRRRLKLGRNVRLQKDGTVITLIKDQYVDMDRDQEENEEDDNPESYDLAEIEPENSEAEDADDDVETIVSDPIRQRIKPAPIIINKQARLAASEKQPMIINQRLRSQRGTKTFHIKEEPDNSDFTVEWQGDDGEVMVVELVNGDEEVLVKHEP

SANSKISAKNCFGFEDDDDYEEYGDGENEVDGASQEFLQLMDMIEQDS

>Maniola\_hyperantus ref|XP\_034839420.1| transcriptional repressor CTCF-like [Maniola hyperantus]

MPPPDKKSANKCKTILQTYLNSFDQDNEPTTVIVNGDGDEADAGVTYFVDEEGRYYYQPAGDSQNLVSLPIEAEAEDGTEIPQEAQMLVDGDGYQTVTLMPSEEGGELSYVLVMQEETKPVMNIDIKVDQDEEKSSDVYKFEEEEEEDPPIEVSDEVEESIKPKLTFAMKRSKHLRPSFTCSFCPYTSHRRYLLLRHMKTHSEERPHKCNVCERGFKTMASLQNHVNMHNGVKPHVCKYCKSPFTTSGELVRHVRYRHTHEKPHKCTECDYASVELSKLRRHVRCHTGERPYQCPHCTYASQDTFKLKRHMRTHTGEKPYKCDHCNMCFTQSNSLKAHKLIHNVSEKPVFPCELCPTKCGRKTDLRIHVQKLHTSDKPLKCKRCGKSFPDRYSCKIHNKTHEGEKCFKCELCPYASTTLRHLKTHMLKHTDEKPFVCDQCDQSFRQKQLLRRHQNLYHNPNYEPKPPKEKTHTCHECKRTFAHKGNLIRHLAIHDPDSGHQERALALRLGRQKKIKFVDGNVKTDDSDNEPEEIMKLDLGGNQLERGELLTVADNDGQQYVVLEVIQAEDGETQIVSAADYEEEEEEEEEEEDEDDEELDKKEIIYEQIKPKGMMERTIKLESDVDTCFGFDEDEEEPDEDEEGIAYNDKIVLRIV

**Alignment:**

Obtained using MUSCLE (version 3.8) in SnapGene (version 6.0.2)

Calendar

Description automatically generated

# Q6:

Phylogenetic tree: Generated using MEGA (11.0.13), aligned with MUSCLE (3.8)

Graphical user interface

Description automatically generated with medium confidence

# Q7:

Heatmap:

Chart, box and whisker chart

Description automatically generated

# Q8:

Top three hits of blast search based on the consensus sequence:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | Technique | Resolution | Source | E-value | Identity (%) |
| 6QNX\_C | X-ray  Diffraction | 2.700 | *Homo sapiens* | 1.21e-95 | 56.584 |
| 5YEF\_A | X-ray  Diffraction | 2.807 | *Homo sapiens* | 5.49e-66 | 57.071 |
| 5T0U\_A | X-ray  Diffraction | 3.199 | *Homo sapiens* | 1.32e-62 | 61.677 |

Consensus sequence:

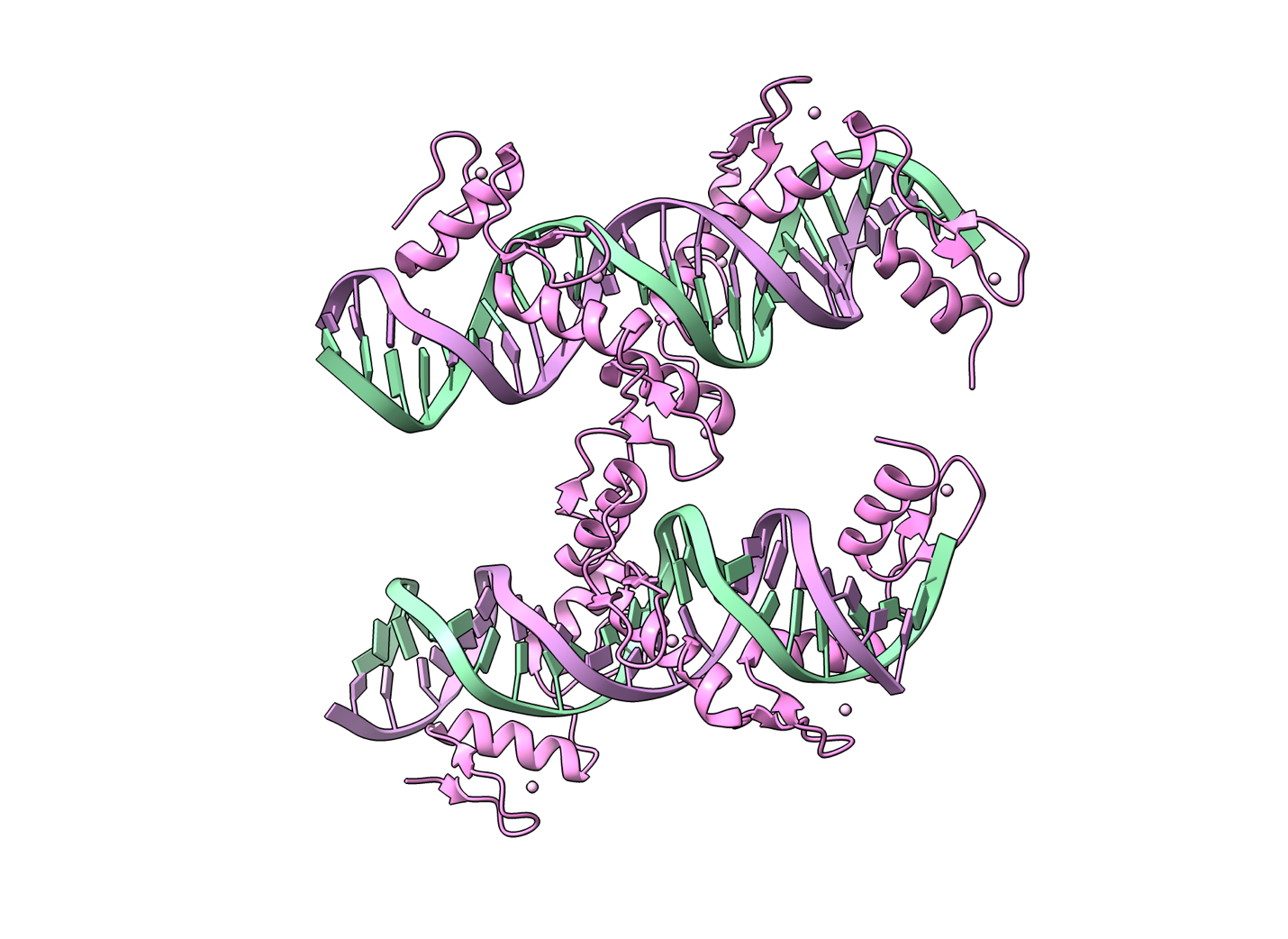
>Aligned Consensus (threshold of >50%)

XCXXCXYTXPXRXXLXRHMKSHXXERPHKCXXCXRXFXTXXXLXNHXNTHTGTKPHXCXXCXXAFXTSGELVRHXRYKHTHEKPXKCXXCDYASVEXSKLRRHMRXHTGERPXQCXXCXYASXDTXKLKRHMRTHTGEKPYXCXICXXRFTQSXSLKXHXLIHXXXXXXXFXCXXCXTXXXRKXDLXXHXXXXHXSXXXXXCXXCXXXXXXRYXXXXHXKSHXGEKXFKCEXCXYASXXXRHLXXHMXXHTXEKPYXCDXCDXXFRQKQLLXRHXXRYHNP

# Q9:

5T0U might not bear too much similarity to *Bicyclus anynana* CTCF like proteins since the identity only go as high as 62%. Figure below is the structure of homodimer CTCF proteins in human binding to two DNA strands.

Structure of 5T0U:



# Q10:

<https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL4523233/>

Target report of CTCF protein in ChEMBL:

Only one chemical, CHEMBL2334661 (C19H16O3; name undefined), is reported to inhibit CTCF in HUVEC (*Homo Sapiens* cell line) by reducing *CTCF* gene transcription.

Assay ID: CHEMBL4421277

Application, table

Description automatically generated with medium confidence