ניתוח השוואתי של מוטיבים גנטיים

זיהוי מוטיבים ושימור גנטי ביונקים באמצעות מבני נתונים למחרוזות

מגישות:

ימית מורנו

שקד טדסה

מרצה מנחה :

סמאח אידריס גזאוי

הכרת הנתונים

- אדם (Human): ~5.5 מיליון בסיסים.
- שימפנזה (Chimpanzee): ~3.2 מיליון בסיסים.
- כלב (Dog): ~1.6 מיליון בסיסים.
- האלפבית: {A, C, G, T}.
- מהות הנתונים: אלה החלקים ב-DNA שהם המתכון לחלבונים, כלומר הקטעים שבאמת מתורגמים לחלבון
- מסר עיקרי: המקורות שונים באורכם, מה שמחייב ניתוח סטטיסטי מנורמל (ממוצעים וחציונים).

המתודולוגיה – איך חיפשנו

אנחנו מתייחסים לגנום כמחרוזת ארוכה של A,C,G,T.

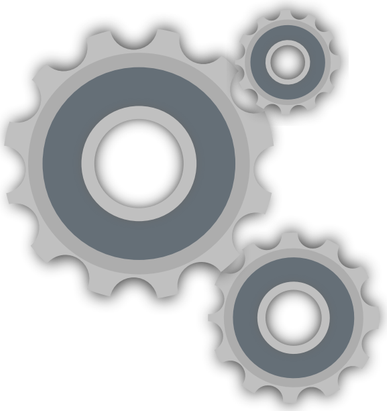
בונים Suffix Array: רשימה של כל הסיומות של המחרוזת, ממוין אלפביתית.

בגלל המיון, סיומות שמתחילות באותו רצף יוצאות צמודות.

מחשבים LCP בין כל שתי סיומות שכנות: כמה אותיות ראשונות הן זהות.

את מערך ה-LCP מחשבים בעזרת אלגוריתם Kasai, שמחשב את כל ערכי ה-LCP ביעילות בזמן ליניארי O(n) (במקום להשוות נאיבית כל פעם מהתחלה).

LCP גדול אומר שמצאנו מקטע חוזר/משותף באורך הזה — ומשם מפיקים מוטיבים, תדירויות ומיקומים.



חזרתיות פנימית – "טביעת האצבע" של המין

ניתוח מבני של שכפולים עצמיים בתוך הגנום

<u>מה בדקנו?</u>

סקאלה משתנה: חיפשנו מוטיבים חוזרים בשלושה אורכים:

- L = 30 (רמת ה"אות"): מייצג "מילות מפתח" בגנום. אלו בדרך כלל אתרי קישור קטנים או בקרים שקובעים אם גן יפעל או לא.
- L = 70 (רמת ה"פסקה"): מייצג דומיינים (Domains). אלו יחידות מבנה עצמאיות בתוך חלבון. רצף באורך כזה כבר מקודד לצורה פיזית מסוימת.
- L = 120 (רמת ה"פרק"): מייצג משפחות גנים או מקטעים פונקציונליים שלמים. כאן אנחנו מחפשים את ה"העתק-הדבק" המשמעותי ביותר של האבולוציה.

סף רגישות: הגדרנו חזרתיות כרצף המופיע 3 פעמים ומעלה בתוך אותו המין.

הגנום האנושי

LRS

ATGGTGCACGTAGCCAGGCTGCTGCTGCTGCTCCTCAC
TTTCTTCCTCCGCACGGATGCTGAGACACCTCCAAGGTT
TACACGAACACCCGTTGATCAGACAGGGGTCTCTGGCG
GAGTTGCCTCTTTCATCTGCCAAGCTACGGGAGACCCA
AGACCTAAAATTGTCTGGAACAAAAAAGGAAAGAAAGTC
AGCAATCAGAGATTTGAGGTAATAGAGTTTGACGATGGG
TCTGGATCAGTTCTCAGAATACAACCCTTACGGACTCCG
AGGGATGAGGCCATTTATGAATGTGTGGCCTCAAATAAT
GTGGGAGAAATAAGTGTATCCACCAGACTCACAGTTTTG
CGGGAAGATCAAATTCCCAGGGGCTTCCCTACCATTGA
CATGGGCCCACAGTTGAAGGTGGTTGAGCGTACTCGCA
CGGCCACCATGCTTTGTGCAGCCAGTGGTAATCCGGAT
CCAGAAATCACTTGGTTTAAAGATTTCTTACCTGTGGACA
CAAGCAACAACAATGGTCGTATTAAGCAGTTACGATCAG
AATCTATTGGTGGTACACCAATAAGAGGAGCCCTTCAGA
TTGAGCAGAGTGAAGAGTCTGACCAAGGAAAATATGAGT
GTGTTGCCACCAACAGCGCGGGCACTCGCTATTCCGCT
CCTGCCAATTTATATGTCAGAGAGCTGCGAGAAGTTCGC
CGTGTCCCACCAAGATTCTCTATCCCACCCACTAATCAT
GAAATCATGCCAGGCGGAAGCGTTAATATCACCTGTGTG
GCCGTGGGGTCACCAATGCCTTATGTAAAGTGGATGTTG
GGGGCAGAAGATCTGACACCTGAAGATGATATGCCAATA
GGAAGAAATGTGCTAGAACTGAATGATGTAAGACAGTCA
GCAAATTACACCTGTGTTGCTATGTCAACACTGGGTGTC
ATTGAAGCAATAGCACAGATCACTGTCAAAGCCTTACCC

ה-LRS הענקי (29,533 בסיסים)

מצאנו שני מקטעי DNA זהים באורך של גנים שלמים
זוהי תופעה שנקראת Segmental Duplication.
בניגוד לחיות פחות מורכבות, הגנום האנושי עשיר בשכפולים כאלה, שנחשבים מנוע
אבולוציוני: עותק אחד נשאר פעיל והעותק השני חופשי להשתנות ולפתח תכונות חדשות.

```
EXECUTIVE REPORT: human
=================================
DATA: Total Length: 5543336
ACTUAL LRS: ATGGTGCACGTAGCCAGGCTGCT
Length: 29533

ANALYSIS RUNTIME: 151788.98 s
PEAK MEMORY: 4698907 KB
```

הגנום האנושי

חזרות קצרות: המוח והמבנה של התא

```
--- L = 30 ---
Total Unique Motifs: 775176
Avg Frequency: 5.67
Median Frequency: 4
Most Frequent Motif: AAGTGGAGATGCTGACATGAAGGGACATTT (117 times)
Top-5 motifs (motif | count):
  AAGTGGAGATGCTGACATGAAGGGACATTT | 117
  ACTGCTGCTGCTCTTGGTGGCAAGTGGAGA | 117
  AGATGCTGACATGAAGGGACATTTTGATCC | 117
  AGTGGAGATGCTGACATGAAGGGACATTTT | 117
  ATCTTTACTGCTGCTGCTCTTGGTGGCAAG | 117
Rarest repeating motif:     0ATGAATGGCACCTACAACACCTGTGGCT (3 times)
Cartesian variants: 774624
Parameterized variants: 775172
-----------------------------

--- L = 70 ---
Total Unique Motifs: 756743
Avg Frequency: 5.56
Median Frequency: 4
Most Frequent Motif: ACCAGAGGCCCTGTCATCTTTACTGCTGCTGCTCTTGGTGGCAAGTGGAGATGCTGACAT
Top-5 motifs (motif | count):
  ACCAGAGGCCCTGTCATCTTTACTGCTGCTGCTCTTGGTGGCAAGTGGAGATGCTGACATGAAGGGACAT | 116
  AGAGGCCCTGTCATCTTTACTGCTGCTGCTCTTGGTGGCAAGTGGAGATGCTGACATGAAGGGACATTTT | 116
  AGGCCCTGTCATCTTTACTGCTGCTGCTCTTGGTGGCAAGTGGAGATGCTGACATGAAGGGACATTTTGA | 116
  ATGGGACCAGAGGCCCTGTCATCTTTACTGCTGCTGCTCTTGGTGGCAAGTGGAGATGCTGACATGAAGG | 116
  CAGAGGCCCTGTCATCTTTACTGCTGCTGCTCTTGGTGGCAAGTGGAGATGCTGACATGAAGGGACATTT | 116
Rarest repeating motif:     0ATGAATGGCACCTACAACACCTGTGGCTCCAGCGACCTCACCTGGCCCCCAC
Cartesian variants: 756271
Parameterized variants: 756741
-----------------------------
```

ממצא כללי

מצאנו חזרתיות גבוהה מאוד (מעל 115 פעמים) ברצפים באורך **30 ו־70 בסיסים**. למרות הדמיון בכמות החזרות, כל אורך מייצג תפקיד ביולוגי שונה לחלוטין.

רצפי 30 בסיסים – בקרה גנטית (Regulatory Motifs)

רצפים קצרים אלו משמשים כמתגים שמפעילים או מכבים גנים באמצעות קישור חלבונים ל־DNA. ריבוי החזרות מאפשר הפעלה מתואמת של קבוצות גנים שלמות בו־זמנית, למשל בתגובה ללחץ או מחלה.

רצפי 70 בסיסים – מבנה חלבונים (Protein Domains)

רצפים אלו מקודדים לדומיינים – יחידות תפקוד בסיסיות בתוך חלבונים. חזרות שלהן משקפות שימוש אבולוציוני יעיל ב"חלקים מוצלחים", שמאפשר יצירת חלבונים מורכבים ורב־תכליתיים.

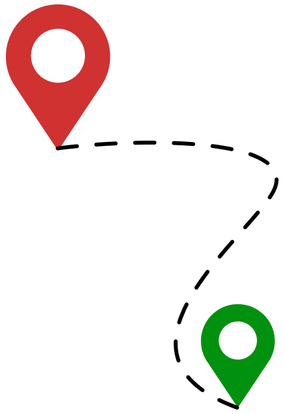
הגנום האנושי

המוטיב הנפוץ (114 חזרות)

מדובר ברצף **באורך 120 בסיסים** שחוזר מעל **100** פעמים, מה שמעיד כנראה על משפחות גנים או דומיינים חלבוניים.

חזרות כאלה מאפשרות לגוף לייצר במהירות ובכמות גדולה חלבונים חיוניים, למשל למערכת החיסון או לשריר הלב.

```
--- L = 120 ---
Total Unique Motifs: 733418
Avg Frequency: 5.44
Median Frequency: 4
Most Frequent Motif: ACCAGAGGCCCTGTCATCTTTACTGCTGCTGCTCTTGGTGGCAAGTGGAGATGCTGACATGAAGGGACATTTTGATCCTGCCAAGTGCCGCTATGCCCTGGGCATGCAGGACCGGACCAT (114 times)
Top-5 motifs (motif | count):
  ACCAGAGGCCCTGTCATCTTTACTGCTGCTGCTCTTGGTGGCAAGTGGAGATGCTGACATGAAGGGACATTTTGATCCTGCCAAGTGCCGCTATGCCCTGGGCATGCAGGACCGGACCAT | 114
  ACTGCTGCTGCTCTTGGTGGCAAGTGGAGATGCTGACATGAAGGGACATTTTGATCCTGCCAAGTGCCGCTATGCCCTGGGCATGCAGGACCGGACCATCCCAGACAGTGACATCTCTGC | 114
  ATCTTTACTGCTGCTGCTCTTGGTGGCAAGTGGAGATGCTGACATGAAGGGACATTTTGATCCTGCCAAGTGCCGCTATGCCCTGGGCATGCAGGACCGGACCATCCCAGACAGTGACAT | 114
  ATGGGACCAGAGGCCCTGTCATCTTTACTGCTGCTGCTCTTGGTGGCAAGTGGAGATGCTGACATGAAGGGACATTTTGATCCTGCCAAGTGCCGCTATGCCCTGGGCATGCAGGACCGG | 114
  CTGCTCTTGGTGGCAAGTGGAGATGCTGACATGAAGGGACATTTTGATCCTGCCAAGTGCCGCTATGCCCTGGGCATGCAGGACCGGACCATCCCAGACAGTGACATCTCTGCTTCCAGC | 114
Rarest repeating motif:     0ATGAATGGCACCTACAACACCTGTGGCTCCAGCGACCTCACCTGGCCCCCAGCGATCAAGCTGGGCTTCTACGCCTACTTGGGCGTCCTGCTGGTGCTAGGCCTGCTGCTCAACAGCC (3 times)
Cartesian variants: 732901
Parameterized variants: 733416
```



מיקומים

דפוס ה"צברים" (Clustering)

החזרות מופיעות בקבוצות צפופות במרחקים קצרים, מה שמעיד על Gene Clusters. אזורים כאלה מכילים עותקים רבים של אותו רכיב, ונפוצים בגנים מבניים או במערכת החיסון, שבהם נדרש מגוון גדול במקום אחד.

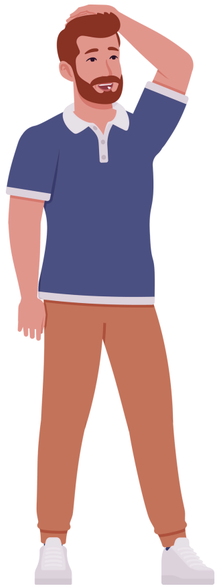
דפוס הקפיצות (Large Genomic Gaps)

בין קבוצות החזרות יש פערים גנומיים עצומים, של מאות אלפי עד מיליוני בסיסים. זהו סימן מובהק ל־Segmental Duplications – שכפול של בלוק DNA שלם והעתקתו למיקום רחוק בגנום.

דפוס החזרות המחזוריות (Periodic Repeats)

באזורים מסוימים החזרות מופיעות במרווח קבוע וסדיר מאוד. דפוס כזה מאפיין Tandem Arrays ומעיד לרוב על אזורים בעלי תפקיד מבני ב־DNA, כמו הצנטרומר, החיוני לחלוקת תא תקינה.

מיקומים



```
Pos: [339850, 340480, 340982, 341427, 341846, 343177, 345810, 348443, 351187, 353931, 358045, 358779, 361577, 364210, 1550969, 1553602, 1556235, 1558979, 1561723, 1564356, 1566467, 1566912, 1567331, 1568662, 1569396,
Pos: [339871, 340501, 341003, 341448, 341867, 343198, 345831, 348464, 351208, 353952, 358066, 358800, 361598, 364231, 1550990, 1553623, 1556256, 1559000, 1561744, 1564377, 1566488, 1566933, 1567352, 1568683, 1569417,
Pos: [339865, 340495, 340997, 341442, 341861, 343192, 345825, 348458, 351202, 353946, 358060, 358794, 361592, 364225, 1550984, 1553617, 1556250, 1558994, 1561738, 1564371, 1566482, 1566927, 1567346, 1568677, 1569411,
Pos: [339845, 340475, 340977, 341422, 341841, 343172, 345805, 348438, 351182, 353926, 358040, 358774, 361572, 364205, 1550964, 1553597, 1556230, 1558974, 1561718, 1564351, 1566462, 1566907, 1567326, 1568657, 1569391,
Pos: [339878, 340508, 341010, 341455, 341874, 343205, 345838, 348471, 351215, 353959, 358073, 358807, 361605, 364238, 1550997, 1553630, 1556263, 1559007, 1561751, 1564384, 1566495, 1566940, 1567359, 1568690, 1569424,
Pos: [339875, 340505, 341007, 341452, 341871, 343202, 345835, 348468, 351212, 353956, 358070, 358804, 361602, 364235, 1550994, 1553627, 1556260, 1559004, 1561748, 1564381, 1566492, 1566937, 1567356, 1568687, 1569421,
Pos: [339872, 340502, 341004, 341449, 341868, 343199, 345832, 348465, 351209, 353953, 358067, 358801, 361599, 364232, 1550991, 1553624, 1556257, 1559001, 1561745, 1564378, 1566489, 1566934, 1567353, 1568684, 1569418,
Pos: [339867, 340497, 340999, 341444, 341863, 343194, 345827, 348460, 351204, 353948, 358062, 358796, 361594, 364227, 1550986, 1553619, 1556252, 1558996, 1561740, 1564373, 1566484, 1566929, 1567348, 1568679, 1569413,
Pos: [339849, 340479, 340981, 341426, 341845, 343176, 345809, 348442, 351186, 353930, 358044, 358778, 361576, 364209, 1550968, 1553601, 1556234, 1558978, 1561722, 1564355, 1566466, 1566911, 1567330, 1568661, 1569395,
Pos: [339877, 340507, 341009, 341454, 341873, 343204, 345837, 348470, 351214, 353958, 358072, 358806, 361604, 364237, 1550996, 1553629, 1556262, 1559006, 1561750, 1564383, 1566494, 1566939, 1567358, 1568689, 1569423,
Pos: [339874, 340504, 341006, 341451, 341870, 343201, 345834, 348467, 351211, 353955, 358069, 358803, 361601, 364234, 1550993, 1553626, 1556259, 1559003, 1561747, 1564380, 1566491, 1566936, 1567355, 1568686, 1569420,
Pos: [339848, 340478, 340980, 341425, 341844, 343175, 345808, 348441, 351185, 353929, 358043, 358777, 361575, 364208, 1550967, 1553600, 1556233, 1558977, 1561721, 1564354, 1566465, 1566910, 1567329, 1568660, 1569394,
Pos: [339847, 340477, 340979, 341424, 341843, 343174, 345807, 348440, 351184, 353928, 358042, 358776, 361574, 364207, 1550966, 1553599, 1556232, 1558976, 1561720, 1564353, 1566464, 1566909, 1567328, 1568659, 1569393,
Pos: [339870, 340500, 341002, 341447, 341866, 343197, 345830, 348463, 351207, 353951, 358065, 358799, 361597, 364230, 1550989, 1553622, 1556255, 1558999, 1561743, 1564376, 1566487, 1566932, 1567351, 1568682, 1569416,
Pos: [339866, 340496, 340998, 341443, 341862, 343193, 345826, 348459, 351203, 353947, 358061, 358795, 361593, 364226, 1550985, 1553618, 1556251, 1558995, 1561739, 1564372, 1566483, 1566928, 1567347, 1568678, 1569412,
Pos: [339879, 340509, 341011, 341456, 341875, 343206, 345839, 348472, 351216, 353960, 358074, 358808, 361606, 364239, 1550998, 1553631, 1556264, 1559008, 1561752, 1564385, 1566496, 1566941, 1567360, 1568691, 1569425,
Pos: [339876, 340506, 341008, 341453, 341872, 343203, 345836, 348469, 351213, 353957, 358071, 358805, 361603, 364236, 1550995, 1553628, 1556261, 1559005, 1561749, 1564382, 1566493, 1566938, 1567357, 1568688, 1569422,
Pos: [339873, 340503, 341005, 341450, 341869, 343200, 345833, 348466, 351210, 353954, 358068, 358802, 361600, 364233, 1550992, 1553625, 1556258, 1559002, 1561746, 1564379, 1566490, 1566935, 1567354, 1568685, 1569419,
Pos: [339846, 340476, 340978, 341423, 341842, 343173, 345806, 348439, 351183, 353927, 358041, 358775, 361573, 364206, 1550965, 1553598, 1556231, 1558975, 1561719, 1564352, 1566463, 1566908, 1567327, 1568658, 1569392,
Pos: [339869, 340499, 341001, 341446, 341865, 343196, 345829, 348462, 351206, 353950, 358064, 358798, 361596, 364229, 1550988, 1553621, 1556254, 1558998, 1561742, 1564375, 1566486, 1566931, 1567350, 1568681, 1569415,
Pos: [339868, 340498, 341000, 341445, 341864, 343195, 345828, 348461, 351205, 353949, 358063, 358797, 361595, 364228, 1550987, 1553620, 1556253, 1558997, 1561741, 1564374, 1566485, 1566930, 1567349, 1568680, 1569414,
Pos: [339853, 340483, 340985, 341430, 341849, 343180, 345813, 348446, 351190, 353934, 358048, 358782, 361580, 364213, 1550972, 1553605, 1556238, 1558982, 1561726, 1564359, 1566470, 1566915, 1567334, 1568665, 1569399,
Pos: [339855, 340485, 340987, 341432, 341851, 343182, 345815, 348448, 351192, 353936, 358050, 358784, 361582, 364215, 1550974, 1553607, 1556240, 1558984, 1561728, 1564361, 1566472, 1566917, 1567336, 1568667, 1569401,
Pos: [339852, 340482, 340984, 341429, 341848, 343179, 345812, 348445, 351189, 353933, 358047, 358781, 361579, 364212, 1550971, 1553604, 1556237, 1558981, 1561725, 1564358, 1566469, 1566914, 1567333, 1568664, 1569398,
Pos: [339864, 340494, 340996, 341441, 341860, 343191, 345824, 348457, 351201, 353945, 358059, 358793, 361591, 364224, 1550983, 1553616, 1556249, 1558993, 1561737, 1564370, 1566481, 1566926, 1567345, 1568676, 1569410,
Pos: [339851, 340481, 340983, 341428, 341847, 343178, 345811, 348444, 351188, 353932, 358046, 358780, 361578, 364211, 1550970, 1553603, 1556236, 1558980, 1561724, 1564357, 1566468, 1566913, 1567332, 1568663, 1569397,
Pos: [339858, 340488, 340990, 341435, 341854, 343185, 345818, 348451, 351195, 353939, 358053, 358787, 361585, 364218, 1550977, 1553610, 1556243, 1558987, 1561731, 1564364, 1566475, 1566920, 1567339, 1568670, 1569404,
Pos: [339859, 340489, 340991, 341436, 341855, 343186, 345819, 348452, 351196, 353940, 358054, 358788, 361586, 364219, 1550978, 1553611, 1556244, 1558988, 1561732, 1564365, 1566476, 1566921, 1567340, 1568671, 1569405,
Pos: [339860, 340490, 340992, 341437, 341856, 343187, 345820, 348453, 351197, 353941, 358055, 358789, 361587, 364220, 1550979, 1553612, 1556245, 1558989, 1561733, 1564366, 1566477, 1566922, 1567341, 1568672, 1569406,
Pos: [339854, 340484, 340986, 341431, 341850, 343181, 345814, 348447, 351191, 353935, 358049, 358783, 361581, 364214, 1550973, 1553606, 1556239, 1558983, 1561727, 1564360, 1566471, 1566916, 1567335, 1568666, 1569400,
Pos: [339857, 340487, 340989, 341434, 341853, 343184, 345817, 348450, 351194, 353938, 358052, 358786, 361584, 364217, 1550976, 1553609, 1556242, 1558986, 1561730, 1564363, 1566474, 1566919, 1567338, 1568669, 1569403,
Pos: [339856, 340486, 340988, 341433, 341852, 343183, 345816, 348449, 351193, 353937, 358051, 358785, 361583, 364216, 1550975, 1553608, 1556241, 1558985, 1561729, 1564362, 1566473, 1566918, 1567337, 1568668, 1569402,
Pos: [339862, 340492, 340994, 341439, 341858, 343189, 345822, 348455, 351199, 353943, 358057, 358791, 361589, 364222, 1550981, 1553614, 1556247, 1558991, 1561735, 1564368, 1566479, 1566924, 1567343, 1568674, 1569408,
Pos: [339863, 340493, 340995, 341440, 341859, 343190, 345823, 348456, 351200, 353944, 358058, 358792, 361590, 364223, 1550982, 1553615, 1556248, 1558992, 1561736, 1564369, 1566480, 1566925, 1567344, 1568675, 1569409,
Pos: [339861, 340491, 340993, 341438, 341857, 343188, 345821, 348454, 351198, 353942, 358056, 358790, 361588, 364221, 1550980, 1553613, 1556246, 1558990, 1561734, 1564367, 1566478, 1566923, 1567342, 1568673, 1569407,
Pos: [339881, 340511, 341013, 341458, 341877, 343208, 345841, 348474, 351218, 353962, 358076, 358810, 361608, 364241, 1551000, 1553633, 1556266, 1559010, 1561754, 1564387, 1566498, 1566943, 1567362, 1568693, 1569427,
Pos: [339883, 340513, 341015, 341460, 341879, 343210, 345843, 348476, 351220, 353964, 358078, 358812, 361610, 364243, 1551002, 1553635, 1556268, 1559012, 1561756, 1564389, 1566500, 1566945, 1567364, 1568695, 1569429,
Pos: [339880, 340510, 341012, 341457, 341876, 343207, 345840, 348473, 351217, 353961, 358075, 358809, 361607, 364240, 1550999, 1553632, 1556265, 1559009, 1561753, 1564386, 1566497, 1566942, 1567361, 1568692, 1569426,
Pos: [339882, 340512, 341014, 341459, 341878, 343209, 345842, 348475, 351219, 353963, 358077, 358811, 361609, 364242, 1551001, 1553634, 1556267, 1559011, 1561755, 1564388, 1566499, 1566944, 1567363, 1568694, 1569428,
Pos: [339885, 340515, 341017, 341462, 341881, 343212, 345845, 348478, 351222, 353966, 358080, 358814, 361612, 364245, 1551004, 1553637, 1556270, 1559014, 1561758, 1564391, 1566502, 1566947, 1567366, 1568697, 1569431,
Pos: [339884, 340514, 341016, 341461, 341880, 343211, 345844, 348477, 351221, 353965, 358079, 358813, 361611, 364244, 1551003, 1553636, 1556269, 1559013, 1561757, 1564390, 1566501, 1566946, 1567365, 1568696, 1569430,
Pos: [339890, 340520, 341022, 341467, 341886, 343217, 345850, 348483, 351227, 353971, 358085, 358819, 361617, 364250, 1551009, 1553642, 1556275, 1559019, 1561763, 1564396, 1566507, 1566952, 1567371, 1568702, 1569436,
Pos: [339889, 340519, 341021, 341466, 341885, 343216, 345849, 348482, 351226, 353970, 358084, 358818, 361616, 364249, 1551008, 1553641, 1556274, 1559018, 1561762, 1564395, 1566506, 1566951, 1567370, 1568701, 1569435,
Pos: [339886, 340516, 341018, 341463, 341882, 343213, 345846, 348479, 351223, 353967, 358081, 358815, 361613, 364246, 1551005, 1553638, 1556271, 1559015, 1561759, 1564392, 1566503, 1566948, 1567367, 1568698, 1569432,
Pos: [339887, 340517, 341019, 341464, 341883, 343214, 345847, 348480, 351224, 353968, 358082, 358816, 361614, 364247, 1551006, 1553639, 1556272, 1559016, 1561760, 1564393, 1566504, 1566949, 1567368, 1568699, 1569433,
Pos: [339888, 340518, 341020, 341465, 341884, 343215, 345848, 348481, 351225, 353969, 358083, 358817, 361615, 364248, 1551007, 1553640, 1556273, 1559017, 1561761, 1564394, 1566505, 1566950, 1567369, 1568700, 1569434,
Pos: [339892, 340522, 341024, 341469, 341888, 343219, 345852, 348485, 351229, 353973, 358087, 358821, 361619, 364252, 1551011, 1553644, 1556277, 1559021, 1561765, 1564398, 1566509, 1566954, 1567373, 1568704, 1569438,
Pos: [339898, 340528, 341030, 341475, 341894, 343225, 345858, 348491, 351235, 353979, 358093, 358827, 361625, 364258, 1551017, 1553650, 1556283, 1559027, 1561771, 1564404, 1566515, 1566960, 1567379, 1568710, 1569444,
Pos: [339893, 340523, 341025, 341470, 341889, 343220, 345853, 348486, 351230, 353974, 358088, 358822, 361620, 364253, 1551012, 1553645, 1556278, 1559022, 1561766, 1564399, 1566510, 1566955, 1567374, 1568705, 1569439,
Pos: [339900, 340530, 341032, 341477, 341896, 343227, 345860, 348493, 351237, 353981, 358095, 358829, 361627, 364260, 1551019, 1553652, 1556285, 1559029, 1561773, 1564406, 1566517, 1566962, 1567381, 1568712, 1569446,
Pos: [339891, 340521, 341023, 341468, 341887, 343218, 345851, 348484, 351228, 353972, 358086, 358820, 361618, 364251, 1551010, 1553643, 1556276, 1559020, 1561764, 1564397, 1566508, 1566953, 1567372, 1568703, 1569437,
Pos: [339897, 340527, 341029, 341474, 341893, 343224, 345857, 348490, 351234, 353978, 358092, 358826, 361624, 364257, 1551016, 1553649, 1556282, 1559026, 1561770, 1564403, 1566514, 1566959, 1567378, 1568709, 1569443,
```

הגנום של השימפנזה

LRS

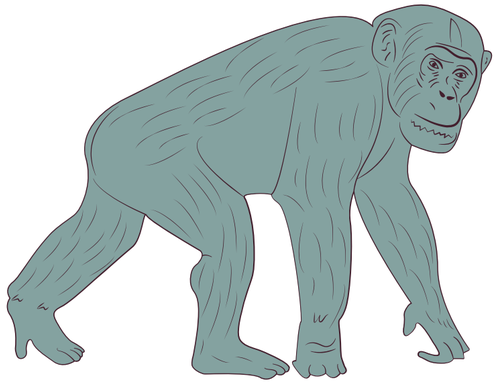
```
GGTCAGAAATCTTGGATAGAGAAAACCTTTTGCAAACGG
GAATGTATCTTTGTAATTCCTAGCACGAAAGACTCTAACA
GGTGTTGCTGTGGCCAGTTCACCAACCAGCATATCCCC
CCTCTGCCAAGTGCAACACCCAGCAAAAATGAAGAGGA
AAACAAACAGGTGGAGACTCAGCCTGAGAAATGGTCTG
TTGCCAAGCACACCCAGAGCTACCCAACAGATTCCTATG
GAGTTCTTGAATTCCAGGGTGGCGGATATTCCAATAAAG
CCATGTATATCCGTGTATCCTATGACACCAAGCCAGACTC
ACTGCTCCATCTCATGGTGAAAGATTGGCAGCTGGAACT
CCCCAAGCTTTTAATATCTGTGCATGGAGGCCTCCAGAA
CTTTGAGATGCAGCCCAAGCTGAAACAAGTCTTTGGGA
AAGGCCTGATCAAGGCTGCCATGACCACCGGGGCCTG
GATCTTCACCGGGGGTGTCAGCACAGGTGTTATCAGCC
ACGTAGGGGATGCCTTGAAAGACCACTCCTCCAAGTCC
AGAGGCCGGGTTTGTGCTATAGGAATTGCTCCATGGGG
CATCGTGGAGAATAAGGAAGACCTGGTTGGAAAGGATG
TAACAAGAGTGTACCAGACCATGTCCAACCCTCTAAGTA
AGCTCTCTGTGCTCAACAACTCCCACACCCACTTCATCC
TGGCTGACAATGGCACCCTGGGCAAGTATGGCGCCGAG
GTGAAGCTGCGAAGGCTGCTGGAAAAGCACATCTCCCT
GCAGAAGATCAACACAAGATTGGGGCAGGGCGTGCCC
CTCGTGGGTCTCGTGGTGGAGGGGGGCCCTAACGTGGT
GTCCATCGTCTTGGAATACCTGCAAGAAGAGCCTCCCGT
CCCTGTGGTGATTTGTGATGGCAGCGGACGTGCCTCGG
ACATCCTGTCCTTTGCGCACAAGTACTGTGAAGAAGGCG
GAATAATAAATGAGTCCCTCAGGGAGCAGCTTCTAGTTA
```

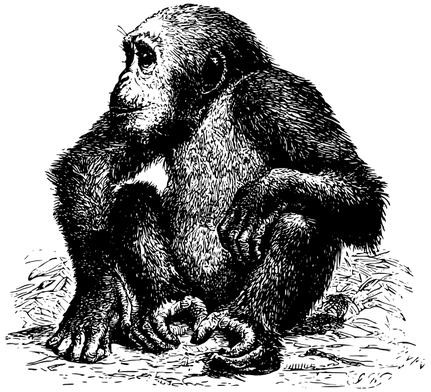
ה-LRS של השימפנזה (9,757 בסיסים)

אורך הרצף שנבדק הוא כ־3.2 מיליון בסיסים, וה־LRS הארוך ביותר עומד על כ־9,757 בסיסים.

בהשוואה לאדם, שבו ה־LRS ארוך בהרבה, הדבר מצביע על כך שהאבולוציה האנושית כללה אירועי Segmental Duplication נרחבים יותר, שתורמים ליצירת גנים חדשים ומורכבות גבוהה יותר.

```
EXECUTIVE REPORT: chimpanzee
==================================
DATA: Total Length: 3256578
ACTUAL LRS: GGTCAGAAATCTTGGATAGAG
Length: 9757

ANALYSIS RUNTIME: 2551.99 s
PEAK MEMORY: 2915504 KB
```





הגנום של השימפנזה

חזרות קצרות: המוח והמבנה של התא

```
--- L = 30 ---
Total Unique Motifs: 475660
Avg Frequency: 4.08
Median Frequency: 3
Most Frequent Motif: NNNNNNNNNNNNNNNNNNNNNNNNNNNN (105 times)
Top-5 motifs (motif | count):
  NNNNNNNNNNNNNNNNNNNNNNNNNNNN | 105
  ATGGCCAAGGCCCTGGTGGCCTGCAAGCTC | 14
  CATGGCCAAGGCCCTGGTGGCCTGCAAGCT | 14
  TGGCCAAGGCCCTGGTGGCCTGCAAGCTCT | 14
  AGCAGCAGCAGCAACAGCAGCAGCAGCAGC | 12
Rarest repeating motif:     0ATGAGGGGCCAGGCCGCCGCCCCGGGCC (3 times)
Cartesian variants: 475591
Parameterized variants: 475659
-----------------------------

--- L = 70 ---
Total Unique Motifs: 459337
Avg Frequency: 4.06
Median Frequency: 3
Most Frequent Motif: NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN (36 ti
Top-5 motifs (motif | count):
  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN | 36
  AAAGCCATCCAGAACACCAGCACCTCACTGCATCTGCAGCTCTCGCTCTGCCTCTTCCTGGCCCACCTCC | 11
  AACACCAGCACCTCACTGCATCTGCAGCTCTCGCTCTGCCTCTTCCTGGCCCACCTCCTCTTCCTCGTGG | 11
  AAGCCATCCAGAACACCAGCACCTCACTGCATCTGCAGCTCTCGCTCTGCCTCTTCCTGGCCCACCTCCT | 11
  AGAACACCAGCACCTCACTGCATCTGCAGCTCTCGCTCTGCCTCTTCCTGGCCCACCTCCTCTTCCTCGT | 11
Rarest repeating motif:     0ATGAGGGGCCAGGCCGCCGCCCCGGGCCCCGTCTGGATCCTCGCCCCGCTGCTGCTGCTGCTGCTGCT
Cartesian variants: 459322
Parameterized variants: 459337
-----------------------------
```

חזרות קצרות (L=30) – "שפת הבקרה"

בשימפנזה המוטיב הנפוץ הופיע 105 פעמים, לעומת 117 אצל האדם. רצפים אלה משמשים כמתגי גנים, והחזרה הגבוהה יותר באדם מצביעה על רשת רגולטורית מורכבת וסנכרון מתקדם של פעילות גנים, במיוחד במוח.

חזרות בינוניות (L=70) – "ארגז הכלים המבני"

בשימפנזה החזרות נמוכות (עד 36), בעוד שבאדם הן גבוהות מאוד (116). רצפים אלה בונים דומיינים חלבוניים, מה שמראה שימוש חוזר אינטנסיבי באדם במודולים מבניים ליצירת חלבונים מורכבים ורב־תכליתיים.

מה זה N (Unknown Base)?

האות N מייצגת בסיס DNA שלא זוהה בוודאות במהלך הריצוף, כאשר המכשיר מעדיף לסמן חוסר מידע במקום לנחש. הופעה חוזרת של N בשימפנזה מעידה על אזורים קשים לריצוף, לרוב Heterochromatin דחוס או חזרתי מאוד.

המשמעות בהשוואה לאדם

החזרות הרבות של N מצביעות על כך שהגנום של השימפנזה פחות מפוענח ומכיל יותר "חורים" מידעיים. לעומת זאת, הגנום האנושי עבר ריצוף והשלמה מעמיקים יותר, ולכן מכיל פחות אזורי N.

הגנום של השימפנזה

רמת "התבניות המורכבות" (L=120) – אדם מול שימפנזה

בשימפנזה נמצאה שכיחות נמוכה מאוד של חזרות באורך זה (עד 10 חזרות) לעומת האדם. מאחר ש־120 בסיסים מייצגים תת־יחידה תפקודית שלמה בחלבון, הפער מעיד שאצל האדם מתרחש שכפול נרחב של מקטעים מורכבים, המאפשר יצירת משפחות גנים גדולות ומערכת חלבונים מתקדמת יותר, בעוד שבשימפנזה רצפים כאלה נדירים ומשמשים בעיקר לתפקידי מבנה כרומוזומלי.

```
--- L = 120 ---
Total Unique Motifs: 439937
Avg Frequency: 4.03
Median Frequency: 3
Most Frequent Motif: AAAAACCAGTATTCAAGACCTCAGGAGTATCGCTGGCCTTACATTTTTACTGGGAATAACTTGGGGCTTTGCCTTCTTTGCCTGGGGACCAGTTAACGTGACCTTCATGTATCTGTTTGC (10 times)
Top-5 motifs (motif | count):
  AAAAACCAGTATTCAAGACCTCAGGAGTATCGCTGGCCTTACATTTTTACTGGGAATAACTTGGGGCTTTGCCTTCTTTGCCTGGGGACCAGTTAACGTGACCTTCATGTATCTGTTTGC | 10
  AAAAACTGAAGAAGCTGGTGGAACAGAATGGAACAGATGACTGGAAAGTTATTGCCAATTATCTCCCGAATCGAACAGATGTGCAGTGCCAGCACCGATGGCAGAAAGTACTAAACCCTG | 10
  AAAAAGAAGAAGCAACTGGGAGCCCAGCGAAAAACCAGTATTCAAGACCTCAGGAGTATCGCTGGCCTTACATTTTTACTGGGAATAACTTGGGGCTTTGCCTTCTTTGCCTGGGGACCA | 10
  AAAAAGAGGAAGAGGACATGGAGCTCACAGCAATGTTGGGACGAAACAACGGGGAGTCCTCCAGGAAGAAGGATGAAGAGGAAGTTCAGAGCAGGCACCGGTTAATCCCCCTCGGCAGAA | 10
  AAAAAGGAAGAAGAAAAGAAGAAGCCCCACATAAAGAAACCTCTTAATGCATTCATGTTGTATATGAAGGAAATGAGAGCAAAGGTCGTAGCTGAGTGCACGTTGAAAGAAAGCGCGGCC | 10
Rarest repeating motif:     0ATGAGGGGCCAGGCCGCCGCCCCGGGCCCCGTCTGGATCCTCGCCCCGCTGCTGCTGCTGCTGCTGCTGGGACGCCGCGCGCGGGCGGCCGCCGGAGCAGACGCGGGGCCCGGGCCCG (3 times)
Cartesian variants: 439925
Parameterized variants: 439937
-----------------------------
```

הגנום של הכלב

LRS

LRS אצל הכלב – רצף חוזר ארוך (6,881 בסיסים)

הגנום של הכלב כולל רצף חוזר ארוך מאוד (6,881 אותיות), המהווה חלק נכבד מהדגימה כולה. זה מעיד על כך שהכלב משתמש בשיטה של הכפלת גנים שלמים כדי ליצור גיבוי או תפקודים חדשים, אך בניגוד לאדם, הוא כמעט ולא משכפל תבניות מורכבות וקצרות יותר.

```
GGTCAGAAATCTTGGATAGAGAAAACCTTTTGCAAACGG
GAATGTATCTTTGTAATTCCTAGCACGAAAGACTCTAACA
GGTGTTGCTGTGGCCAGTTCACCAACCAGCATATCCCC
CCTCTGCCAAGTGCAACACCCAGCAAAAATGAAGAGGA
AAACAAACAGGTGGAGACTCAGCCTGAGAAATGGTCTG
TTGCCAAGCACACCCAGAGCTACCCAACAGATTCCTATG
GAGTTCTTGAATTCCAGGGTGGCGGATATTCCAATAAAG
CCATGTATATCCGTGTATCCTATGACACCAAGCCAGACTC
ACTGCTCCATCTCATGGTGAAAGATTGGCAGCTGGAACT
CCCCAAGCTTTTAATATCTGTGCATGGAGGCCTCCAGAA
CTTTGAGATGCAGCCCAAGCTGAAACAAGTCTTTGGGA
AAGGCCTGATCAAGGCTGCCATGACCACCGGGGCCTG
GATCTTCACCGGGGGTGTCAGCACAGGTGTTATCAGCC
ACGTAGGGGATGCCTTGAAAGACCACTCCTCCAAGTCC
AGAGGCCGGGTTTGTGCTATAGGAATTGCTCCATGGGG
CATCGTGGAGAATAAGGAAGACCTGGTTGGAAAGGATG
TAACAAGAGTGTACCAGACCATGTCCAACCCTCTAAGTA
AGCTCTCTGTGCTCAACAACTCCCACACCCACTTCATCC
TGGCTGACAATGGCACCCTGGGCAAGTATGGCGCCGAG
GTGAAGCTGCGAAGGCTGCTGGAAAAGCACATCTCCCT
GCAGAAGATCAACACAAGATTGGGGCAGGGCGTGCCC
CTCGTGGGTCTCGTGGTGGAGGGGGGCCCTAACGTGGT
GTCCATCGTCTTGGAATACCTGCAAGAAGAGCCTCCCGT
CCCTGTGGTGATTTGTGATGGCAGCGGACGTGCCTCGG
ACATCCTGTCCTTTGCGCACAAGTACTGTGAAGAAGGCG
GAATAATAAATGAGTCCCTCAGGGAGCAGCTTCTAGTTA
```

```
EXECUTIVE REPORT: dog
==============================
DATA: Total Length: 1666555
ACTUAL LRS: CAGGTGATTTCAAATCAGAG
Length: 6881

ANALYSIS RUNTIME: 389.24 s
PEAK MEMORY: 635315 KB
```

הגנום של הכלב

חזרות קצרות: המוח והמבנה של התא

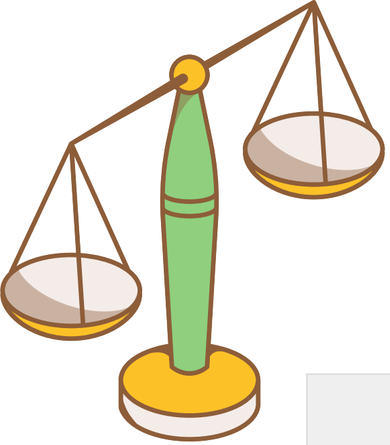
```
--- L = 70 ---
Total Unique Motifs: 30171
Avg Frequency: 3.02
Median Frequency: 3
Most Frequent Motif: AGATAGGCAGAACCAATGAGACTCGCTATGAGGTACAGGCCCTGGAACCCGGGACCCTGTACAACCTCAG (6 times)
Top-5 motifs (motif | count):
  AGATAGGCAGAACCAATGAGACTCGCTATGAGGTACAGGCCCTGGAACCCGGGACCCTGTACAACCTCAG | 6
  CAGATAGGCAGAACCAATGAGACTCGCTATGAGGTACAGGCCCTGGAACCCGGGACCCTGTACAACCTCA | 6
  CCAGATAGGCAGAACCAATGAGACTCGCTATGAGGTACAGGCCCTGGAACCCGGGACCCTGTACAACCTC | 6
  GATAGGCAGAACCAATGAGACTCGCTATGAGGTACAGGCCCTGGAACCCGGGACCCTGTACAACCTCAGC | 6
  AAAAAGGAAGAAGAAAAGAAGAAGCCCCACATAAAGAAACCTCTTAACGCATTCATGTTGTATATGAAGG | 4
Rarest repeating motif: AAAAAAAAAGTGCGTTCGCTACATACAAGGTGAAGGCAGCTGCCTCAGTCCACCCTCTTCAGATGGAAGC (3 times)
Cartesian variants: 30171
Parameterized variants: 30171
-----------------------------

--- L = 120 ---
Total Unique Motifs: 27872
Avg Frequency: 3.02
Median Frequency: 3
Most Frequent Motif: AAAAAGGAAGAAGAAAAGAAGAAGCCCCACATAAAGAAACCTCTTAACGCATTCATGTTGTATATGAAGGAAATGAGGGCCAAGGTCGTGGCCGAGTGCACGTTGAAAGAAAGTG
Top-5 motifs (motif | count):
  AAAAAGGAAGAAGAAAAGAAGAAGCCCCACATAAAGAAACCTCTTAACGCATTCATGTTGTATATGAAGGAAATGAGGGCCAAGGTCGTGGCCGAGTGCACGTTGAAAGAAAGTGCGGCC | 4
  AAAAGAAGAAGCCCCACATAAAGAAACCTCTTAACGCATTCATGTTGTATATGAAGGAAATGAGGGCCAAGGTCGTGGCCGAGTGCACGTTGAAAGAAAGTGCGGCCATCAACCAGATCC | 4
  AAAAGGAAGAAGAAAAGAAGAAGCCCCACATAAAGAAACCTCTTAACGCATTCATGTTGTATATGAAGGAAATGAGGGCCAAGGTCGTGGCCGAGTGCACGTTGAAAGAAAGTGCGGCCA | 4
  AAACAGGAATCCTCCCAGAGTGACGTCGGCTCACTCCACAGCTCAAAGCATCAGGACTCCAAAAAGGAAGAAGAAAAGAAGAAGCCCCACATAAAGAAACCTCTTAACGCATTCATGTTG | 4
  AAACCCACCTCCACACTTACCAGCTGACGTAGACCCCAAAACAGGAATCCCACGGCCTCCGCACCCTCCAGATATATCTCCGTATTACCCGCTATCGCCTGGCACCGTAGGACAAATCCC | 4
Rarest repeating motif: AAAAAAAAAGTGCGTTCGCTACATACAAGGTGAAGGCAGCTGCCTCAGTCCACCCTCTTCAGATGGAAGCTTACTAGACTCGCCTCCTCCCTCCCCCAACCTGCTAGGCTCC
Cartesian variants: 27872
Parameterized variants: 27872
```

השוואת ה"מתגים" (L=30) – היכן הבקרה?

בכלב השכיחות המקסימלית של מתגים היא 44 חזרות בלבד, לעומת 105 בשימפנזה ו־117 באדם. משמעות הדבר היא שלכלב יש פחות נקודות בקרה משותפות, והוא מפעיל רשתות גנים קטנות וממוקדות יותר, עם חזרות כמו AGCAGCAGC שמתקשרות ליחידות נוירולוגיות אך בכמות קטנה משמעותית.

רמת ה"מודולים" (L=70 ו-L=120) – יעילות מול מורכבות

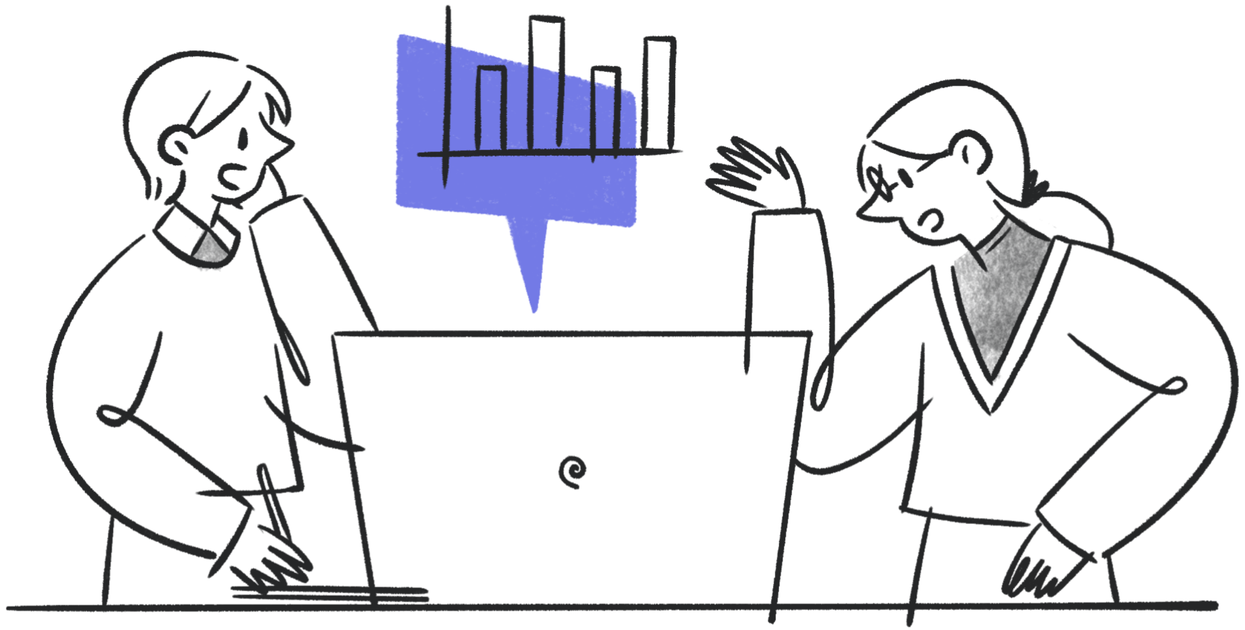
בשכיחות רצפי 70–120 בסיסים הכלב מגיע ל־4–6 חזרות בלבד, בעוד שבאדם השכיחות נשארת גבוהה (114–116). כלב אנחנו רואים הרבה פחות חזרות של מוטיבים ארוכים (70–120), מה שמרמז שיש פחות שכפול של קטעים ארוכים זהים בגנום שניתחנו. כלומר, החזרתיות הארוכה פחות בולטת, בהשוואה לאדם

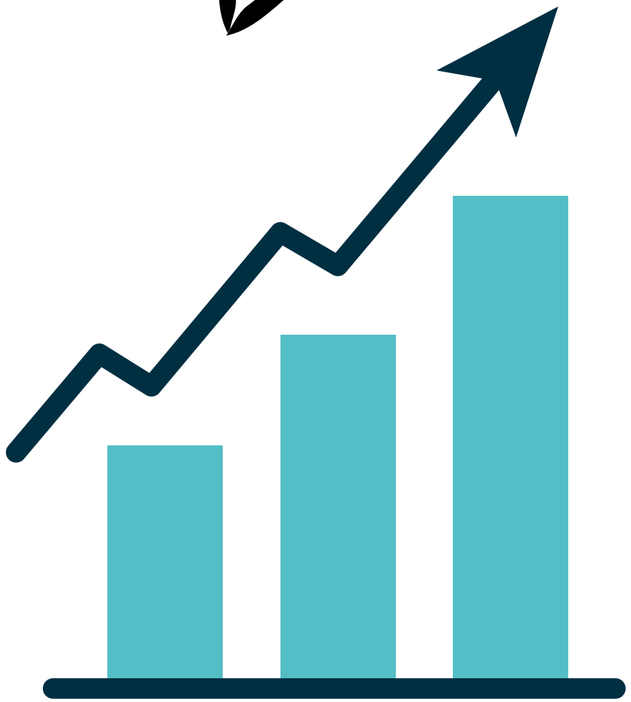


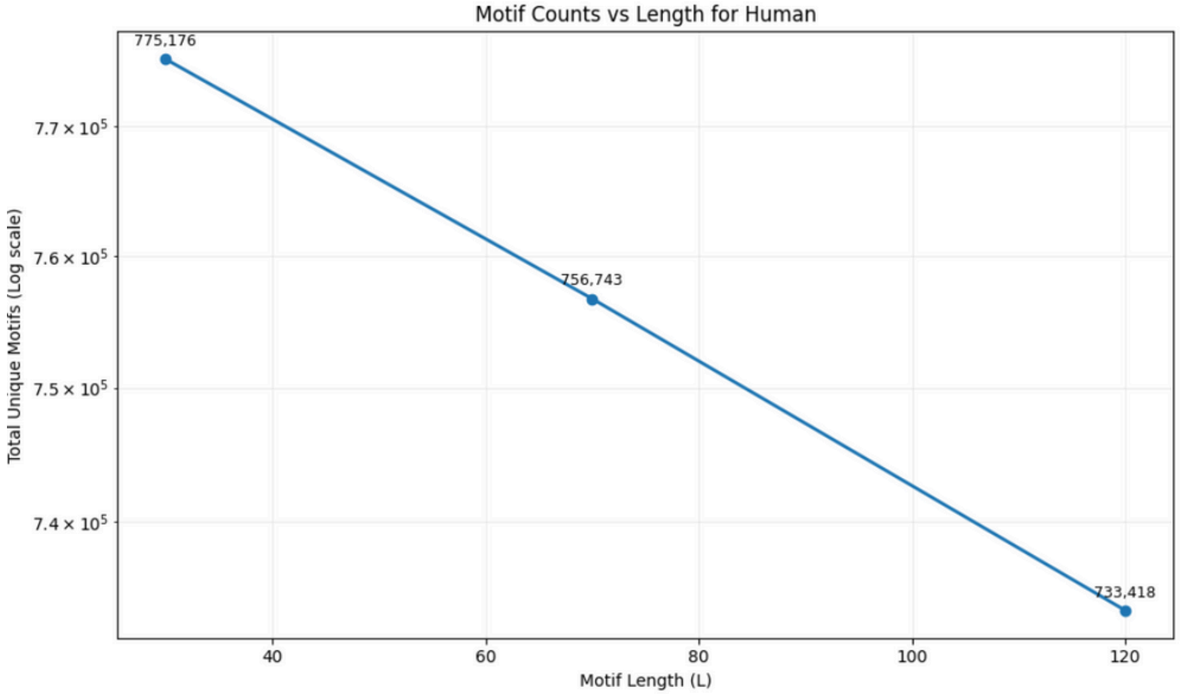
טבלת השוואות

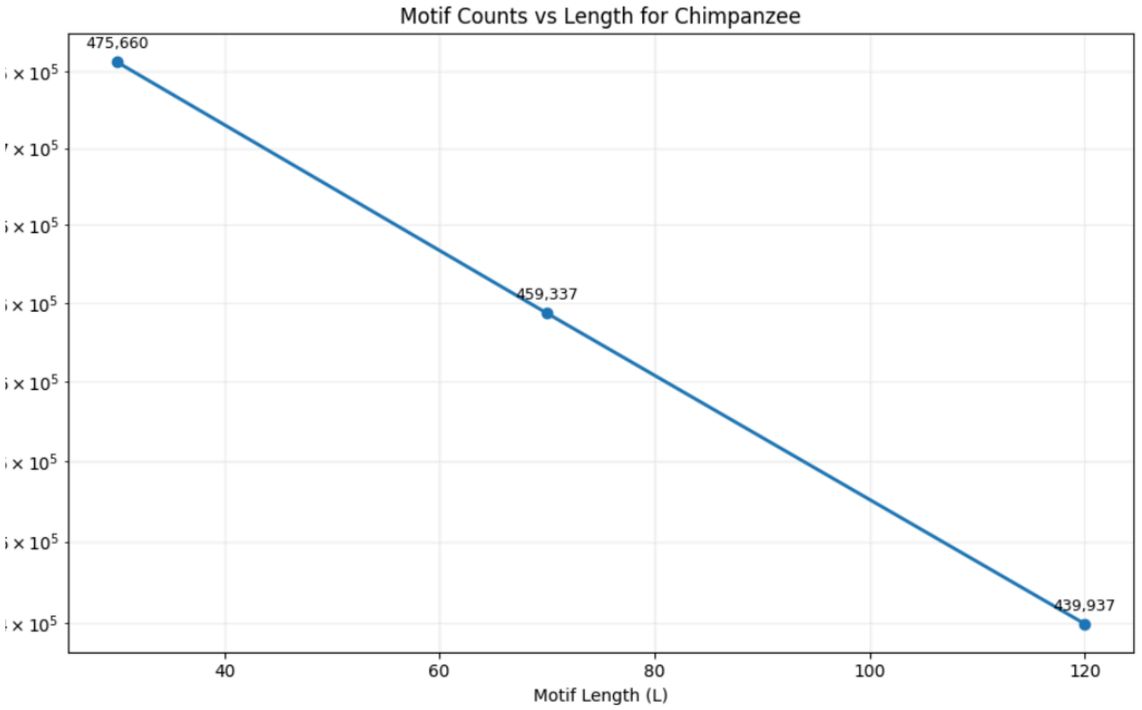
מדד	אדם	שימפנזה	כלב
אורך LRS	~29,500 (שיא המורכבות)	~9,700	~6,800
שכיחות ב-L=30	117 (בקרה מסועפת)	105	44 (בקרה פשוטה)
שכיחות ב-L=120	114 (שכפול תבניות)	10	4 (ייחודיות גבוהה)
שימוש ב-N (חורים)	נמוך מאוד	גבוה (36)	נמוך

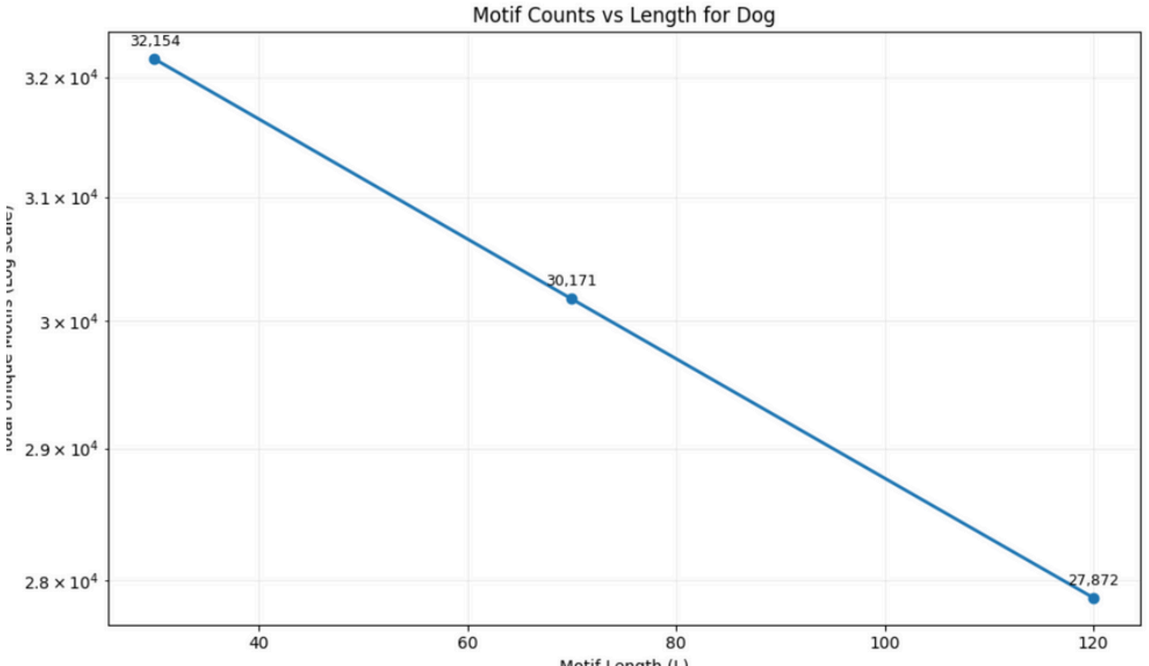
גרפים

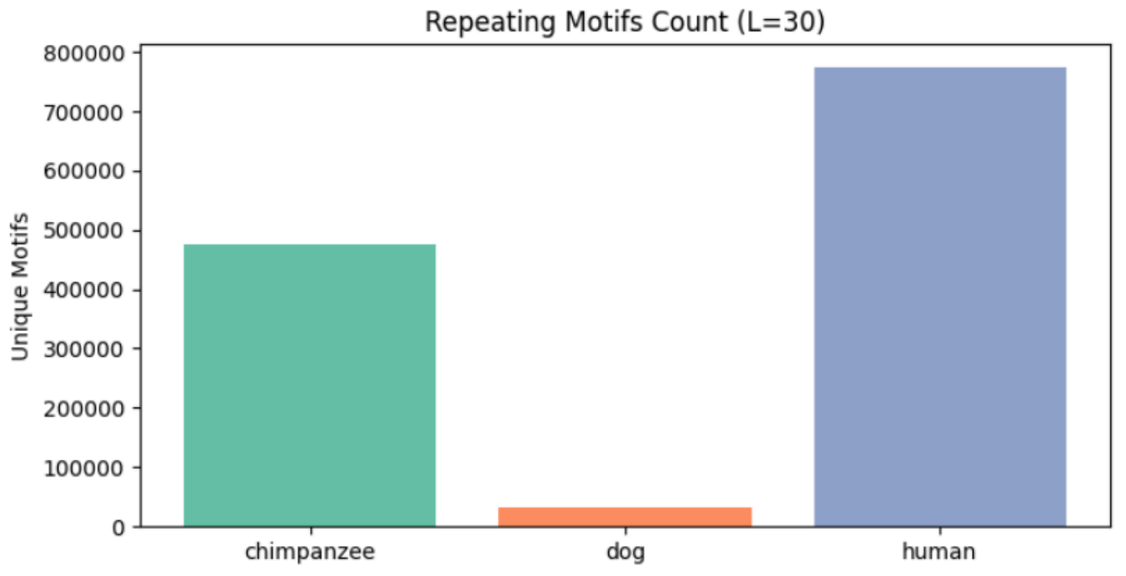


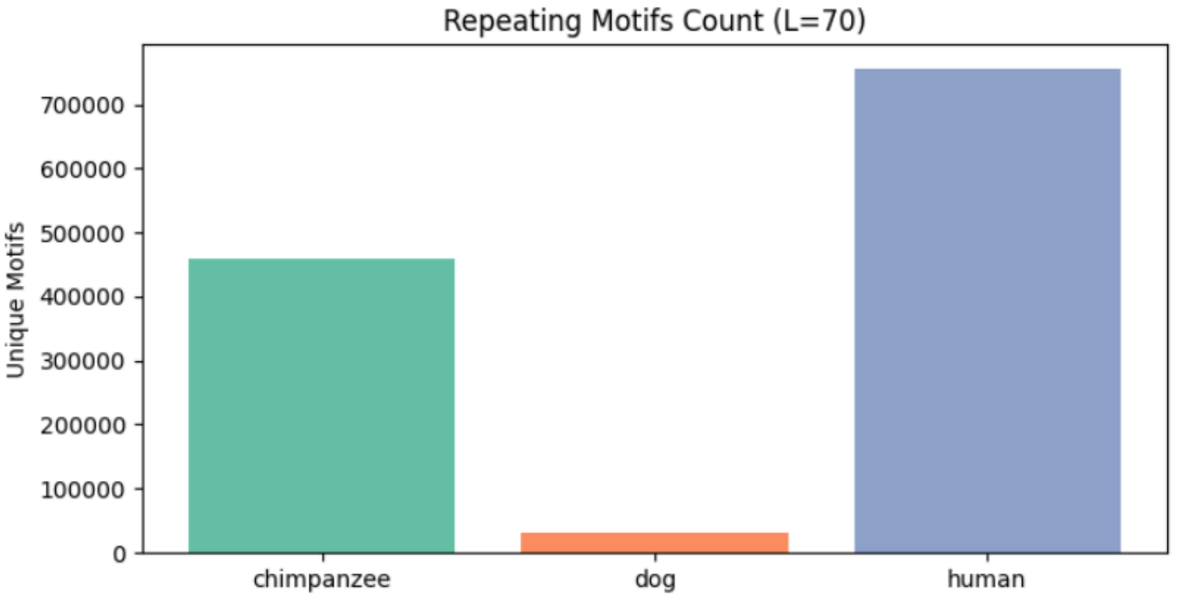


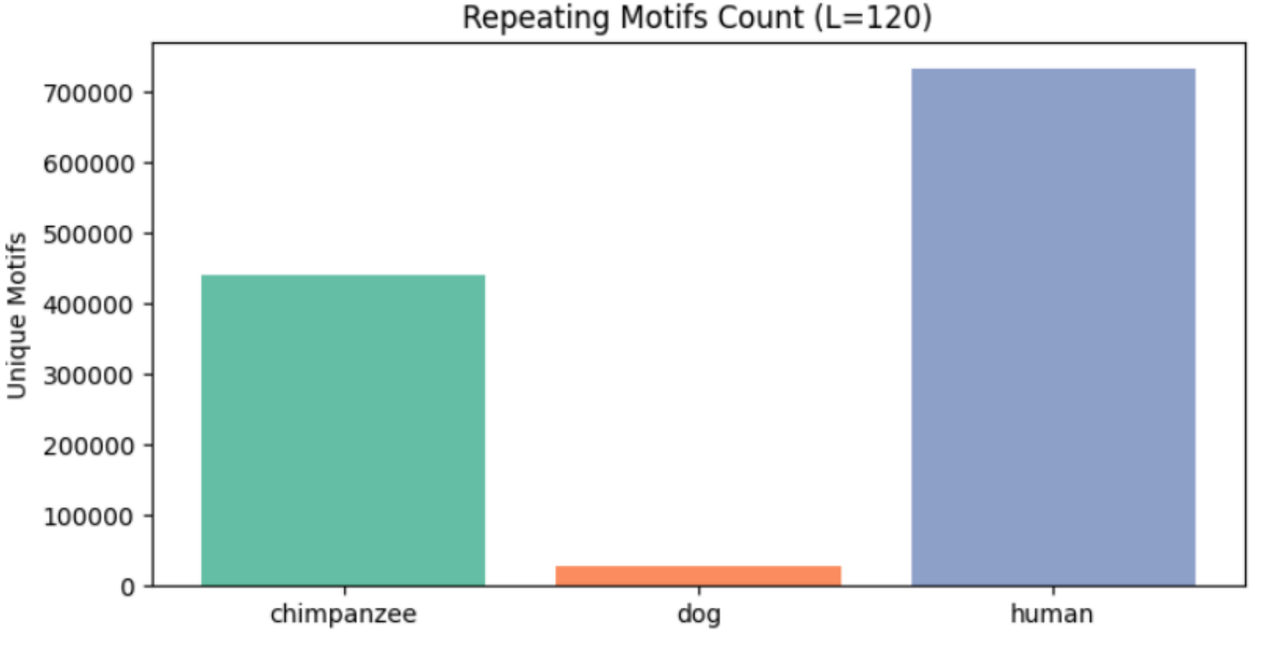

Motif Counts vs Length for Human
775,176
756,743
733,418
7.7 × 10^5
7.6 × 10^5
7.5 × 10^5
7.4 × 10^5
Total Unique Motifs (Log scale)
40
60
80
100
120
Motif Length (L)


Motif Counts vs Length for Chimpanzee
475,660
459,337
439,937
40
60
80
100
120
Motif Length (L)


Motif Counts vs Length for Dog
32,154
30,171
27,872
3.2 × 10^4
3.1 × 10^4
3 × 10^4
2.9 × 10^4
2.8 × 10^4
40
60
80
100
120


Repeating Motifs Count (L=30)
Unique Motifs
800000
700000
600000
500000
400000
300000
200000
100000
0
chimpanzee
dog
human


Repeating Motifs Count (L=70)
Unique Motifs
700000
600000
500000
400000
300000
200000
100000
0
chimpanzee
dog
human


Repeating Motifs Count (L=120)
Unique Motifs
700000
600000
500000
400000
300000
200000
100000
0
chimpanzee
dog
human

```
[ AT LENGTH L = 30 ]
Number of motifs shared by ALL 3 species: 759

Pairwise Shared Counts:
  - chimpanzee vs dog: 1071 motifs
  - chimpanzee vs human: 333742 motifs
  - dog vs human: 1664 motifs

[ AT LENGTH L = 70 ]
Number of motifs shared by ALL 3 species: 62

Pairwise Shared Counts:
  - chimpanzee vs dog: 62 motifs
  - chimpanzee vs human: 265054 motifs
  - dog vs human: 151 motifs

[ AT LENGTH L = 120 ]
Number of motifs shared by ALL 3 species: 0

Pairwise Shared Counts:
  - chimpanzee vs dog: 0 motifs
  - chimpanzee vs human: 202243 motifs
  - dog vs human: 25 motifs
```

1. האדם והשימפנזה – "אחים גנטיים"

ב־L=30 יש 333,742 מוטיבים משותפים; גם ב־L=120 נשארו 202,243 מוטיבים זהים.

מסקנה: רוב מנגנוני הבקרה והחלבונים זהים לחלוטין, מאשש קרבה אבולוציונית מיידית.

2. הכלב – "הזר המוכר"

מול האדם יש רק 1,664 מוטיבים משותפים ב־L=30.

מסקנה: מערכות הבקרה של הכלב נבנו באופן עצמאי ושונה מהפרימטים, והוא חולק איתנו רק את ה"ליבה" של החיים.

3. הליבה המשותפת לכולם (CORE)

759 מוטיבים משותפים ב־L=30, אך 0 ב־L=120.

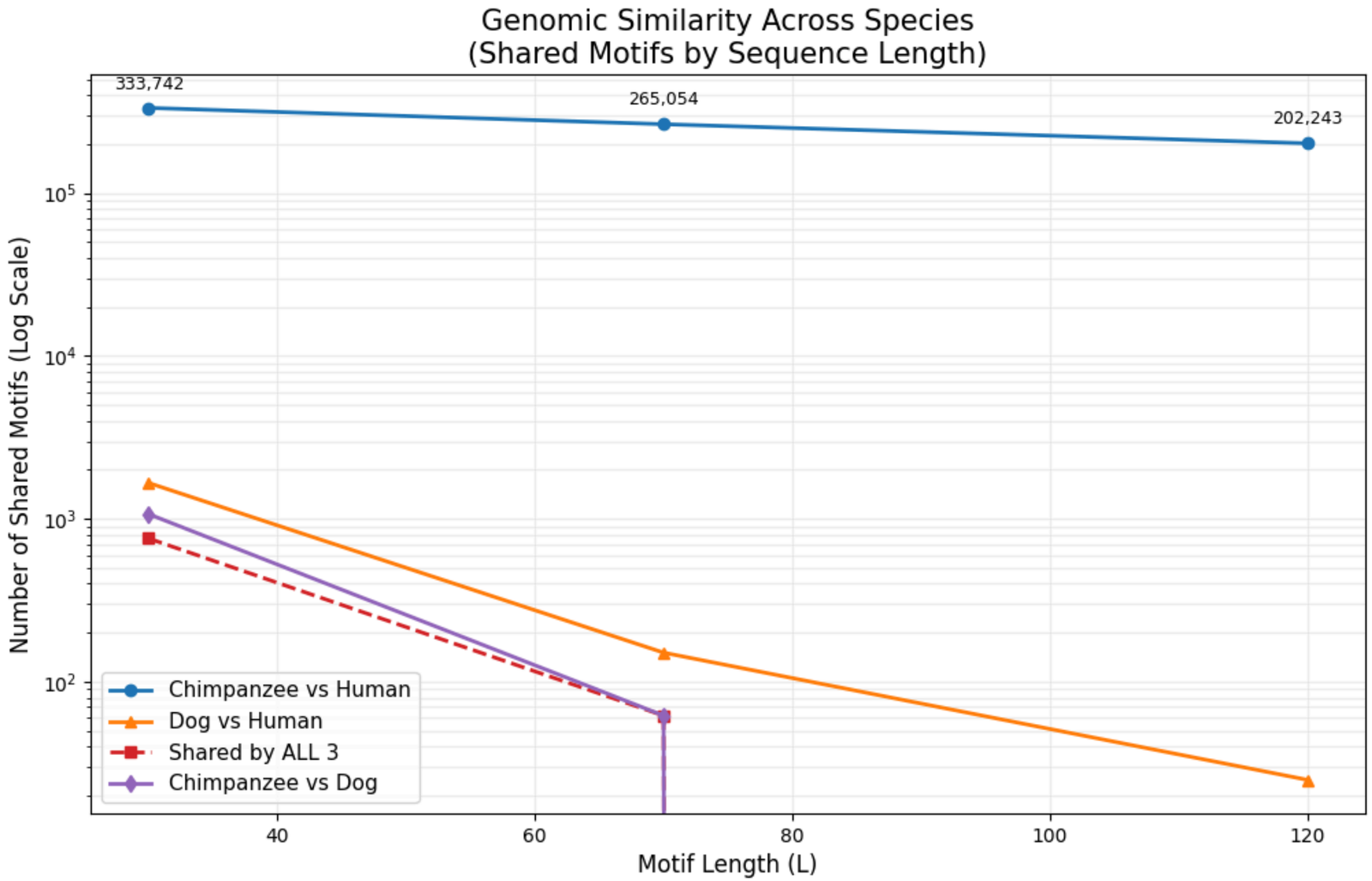
מסקנה: אלו רצפים בסיסיים של יונקים – נשימה תאית, חלוקת תאים ושכפול DNA. רצפים ארוכים שונים לכל מין.

4. חוק ה-120 – הגבול של הדמיון

בין השימפנזה לכלב אין אפילו רצף ארוך זהה (L=120).

מסקנה: פרימטים יוצאים דופן מבחינה גנטית, בעוד שכלב מייצג סדרה אבולוציונית נפרדת.

לסיכום



אדם ושימפנזה:

דמיון מוחלט כמעט בכל אורך.

הכלב:

דמיון נמוך כבר ברצפים קצרים, שוני מוחלט ברצפים ארוכים.

הליבה של 759 מוטיבים:

פרוטוקול בסיסי שמאפיין את כל היונקים.

תודה על ההקשבה!