International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

# Linear Prediction of Overlapping Codons in a Genome Sequence

**Venkateswarlu Pedakolmi, E.G. Rajan**

***Abstract:*** ***The concept of overlapping n-ary codons was proposed in this paper as a novel approach to the study of genome sequences in the framework of theoretical computer science. Given a genome sequence of length N, one can have (N/n) non-overlapping n-ary codons with 0 or 1 or up to n-1 unused nucleotides left out in the sequence. Unused nucleotides are not considered in the scheme of genetic code. Alternatively, one can have (N-n+1) overlapping n-ary codons with no unused nucleotide left out in the sequence.***

***Keywords: Linear Perdition, Codons, Linear Boolean Functions, Genome Sequences.***

## I. INTRODUCTION

The experiment due to Crick, Brenner, Barnett and Watts-Tobin revealed that codons consist of three DNA bases. Subsequently in the year 1961, Marshall Nirenberg and Heinrich J. Matthaei described the nature of a codon. In the year 1997, E G Rajan proposed the notion of n-ary codons of overlapping type. For example let us consider a finite sequence **AGTCAGTCG** of length 9. As per Rajan's conjecture, one can consider the following overlapping codons. As per this conjecture, there is no nucleotide left out unused in the cell functioning. In such a case, the plethora of concepts and tools of theoretical computer science could be used in genome study.

**Overlapping codons of AGTCAGTCG**

***Single Nucleotides***

A, G, T, C, A, G, T, C, G

***Overlapping 2- codons***

AG, GT, TC, CA, AG, GT, TC, CG

***Overlapping 3- codons***

AGT, GTC, TCA, CAG, AGT, GTC, TCG

***Overlapping 4- codons***

AGTC, GTCA, TCAG, CAGT, AGTC, GTCG

***Overlapping 5- codons***

AGTCA, GTCAG, TCAGT, CAGTC, AGTCG

By suitably representing these codons as numbers, one can predict such codons using octet linear prediction algorithm. A detailed case study was made on predicting these codons in the genome sequence of Brucella Suis 1330 and the results reported in this paper.

**Revised Manuscript Received on February 28, 2020.**

***** Correspondence Author

**Venkateswarlu Pedakolmi*,** Research Scholar, Dept. of Computer Science, MG-NIRSA, Affiliated to University of Mysore, Manasagangotri, Mysore, Karnataka, India. E-Mail: venkat123.pedakolmi@gmail.com

**Prof. E.G. Rajan,** Director, MG-NIRSA; Director, PRC Global Technologies Inc., Ontario, Canada. E-mail: rajaneg@yahoo.co.in



## II. NUMERICAL REPRESENTATION OF N-ARY CODONS

One can have a unique numerical representation for an n-ary codon set.

**Numerical representation of 2-codon set**

| Number Code | 2-Codons | Number Code | 2-Codons |
|---|---|---|---|
| 1 | AA | 9 | GA |
| 2 | AT | 10 | GT |
| 3 | AG | 11 | GG |
| 4 | AC | 12 | GC |
| 5 | TA | 13 | CA |
| 6 | TT | 14 | CT |
| 7 | TG | 15 | CG |
| 8 | TC | 16 | CC |

**Numerical representation of 3-codon set**

| Number Code | 3-Codons | Number Code | 3-Codons | Number Code | 3-Codons | Number Code | 3-Codons |
|---|---|---|---|---|---|---|---|
| 1 | AAA | 17 | TAA | 33 | GAA | 49 | CAA |
| 2 | AAT | 18 | TAT | 34 | GAT | 50 | CAT |
| 3 | AAG | 19 | TAG | 35 | GAG | 51 | CAG |
| 4 | AAC | 20 | TAC | 36 | GAC | 52 | CAC |
| 5 | ATA | 21 | TTA | 37 | GTA | 53 | CTA |
| 6 | ATT | 22 | TTT | 38 | GTT | 54 | CTT |
| 7 | ATG | 23 | TTG | 39 | GTG | 55 | CTG |
| 8 | ATC | 24 | TTC | 40 | GTC | 56 | CTC |
| 9 | AGA | 25 | TGA | 41 | GGA | 57 | CGA |
| 10 | AGT | 26 | TGT | 42 | GGT | 58 | CGT |
| 11 | AGG | 27 | TGG | 43 | GGG | 59 | CGG |
| 12 | AGC | 28 | TGC | 44 | GGC | 60 | CGC |
| 13 | ACA | 29 | TCA | 45 | GCA | 61 | CCA |
| 14 | ACT | 30 | TCT | 46 | GCT | 62 | CCT |
| 15 | ACG | 31 | TCG | 47 | GCG | 63 | CCG |
| 16 | ACC | 32 | TCC | 48 | GCC | 64 | CCC |

**Numerical representation of 4-codon set**

| Number Code | 4-Codon | Number Code | 4-Codon | Number Code | 4-Codon | Number Code | 4-Codon |
|---|---|---|---|---|---|---|---|
| 1 | AAAA | 17 | ATAA | 33 | AGAA | 49 | ACAA |
| 2 | AAAT | 18 | ATAT | 34 | AGAT | 50 | ACAT |
| 3 | AAAG | 19 | ATAG | 35 | AGAG | 51 | ACAG |
| 4 | AAAC | 20 | ATAC | 36 | AGAC | 52 | ACAC |
| 5 | AATA | 21 | ATTA | 37 | AGTA | 53 | ACTA |
| 6 | AATT | 22 | ATTT | 38 | AGTT | 54 | ACTT |
| 7 | AATG | 23 | ATTG | 39 | AGTG | 55 | ACTG |
| 8 | AATC | 24 | ATTC | 40 | AGTC | 56 | ACTC |
| 9 | AAGA | 25 | ATGA | 41 | AGGA | 57 | ACGA |
| 10 | AAGT | 26 | ATGT | 42 | AGGT | 58 | ACGT |
| 11 | AAGG | 27 | ATGG | 43 | AGGG | 59 | ACGG |
| 12 | AAGC | 28 | ATGC | 44 | AGGC | 60 | ACGC |
| 13 | AACA | 29 | ATCA | 45 | AGCA | 61 | ACCA |
| 14 | AACT | 30 | ATCT | 46 | AGCT | 62 | ACCT |
| 15 | AACG | 31 | ATCG | 47 | AGCG | 63 | ACCG |
| 16 | AACC | 32 | ATCC | 48 | AGCC | 64 | ACCC |

| Number Code | 4-Codon | Number Code | 4-Codon | Number Code | 4-Codon | Number Code | 4-Codon |
|---|---|---|---|---|---|---|---|
| 65 | TAAA | 81 | TTAA | 97 | TGAA | 113 | TCAA |
| 66 | TAAT | 82 | TTAT | 98 | TGAT | 114 | TCAT |
| 67 | TAAG | 83 | TTAG | 99 | TGAG | 115 | TCAG |
| 68 | TAAC | 84 | TTAC | 100 | TGAC | 116 | TCAC |
| 69 | TATA | 85 | TTTA | 101 | TGTA | 117 | TCTA |
| 70 | TATT | 86 | TTTT | 102 | TGTT | 118 | TCTT |
| 71 | TATG | 87 | TTTG | 103 | TGTG | 119 | TCTG |
| 72 | TATC | 88 | TTTC | 104 | TGTC | 120 | TCTC |
| 73 | TAGA | 89 | TTGA | 105 | TGGA | 121 | TCGA |
| 74 | TAGT | 90 | TTGT | 106 | TGGT | 122 | TCGT |
| 75 | TAGG | 91 | TTGG | 107 | TGGG | 123 | TCGG |
| 76 | TAGC | 92 | TTGC | 108 | TGGC | 124 | TCGC |
| 77 | TACA | 93 | TTCA | 109 | TGCA | 125 | TCCA |
| 78 | TACT | 94 | TTCT | 110 | TGCT | 126 | TCCT |
| 79 | TACG | 95 | TTCG | 111 | TGCG | 127 | TCCG |
| 80 | TACC | 96 | TTCC | 112 | TGCC | 128 | TCCC |

| Number Code | 4-Codon | Number Code | 4-Codon | Number Code | 4-Codon | Number Code | 4-Codon |
|---|---|---|---|---|---|---|---|
| 129 | GAAA | 145 | GTAA | 161 | GGAA | 177 | GCAA |
| 130 | GAAT | 146 | GTAT | 162 | GGAT | 178 | GCAT |
| 131 | GAAG | 147 | GTAG | 163 | GGAG | 179 | GCAG |
| 132 | GAAC | 148 | GTAC | 164 | GGAC | 180 | GCAC |
| 133 | GATA | 149 | GTTA | 165 | GGTA | 181 | GCTA |
| 134 | GATT | 150 | GTTT | 166 | GGTT | 182 | GCTT |
| 135 | GATG | 151 | GTTG | 167 | GGTG | 183 | GCTG |
| 136 | GATC | 152 | GTTC | 168 | GGTC | 184 | GCTC |
| 137 | GAGA | 153 | GTGA | 169 | GGGA | 185 | GCGA |
| 138 | GAGT | 154 | GTGT | 170 | GGGT | 186 | GCGT |
| 139 | GAGG | 155 | GTGG | 171 | GGGG | 187 | GCGG |
| 140 | GAGC | 156 | GTGC | 172 | GGGC | 188 | GCGC |
| 141 | GACA | 157 | GTCA | 173 | GGCA | 189 | GCCA |
| 142 | GACT | 158 | GTCT | 174 | GGCT | 190 | GCCT |
| 143 | GACG | 159 | GTCG | 175 | GGCG | 191 | GCCG |
| 144 | GACC | 160 | GTCC | 176 | GGCC | 192 | GCCC |

| Number Code | 4-Codon | Number Code | 4-Codon | Number Code | 4-Codon | Number Code | 4-Codon |
|---|---|---|---|---|---|---|---|
| 193 | CAAA | 209 | CTAA | 225 | CGAA | 241 | CCAA |
| 194 | CAAT | 210 | CTAT | 226 | CGAT | 242 | CCAT |
| 195 | CAAG | 211 | CTAG | 227 | CGAG | 243 | CCAG |
| 196 | CAAC | 212 | CTAC | 228 | CGAC | 244 | CCAC |
| 197 | CATA | 213 | CTTA | 229 | CGTA | 245 | CCTA |
| 198 | CATT | 214 | CTTT | 230 | CGTT | 246 | CCTT |
| 199 | CATG | 215 | CTTG | 231 | CGTG | 247 | CCTG |
| 200 | CATC | 216 | CTTC | 232 | CGTC | 248 | CCTC |
| 201 | CAGA | 217 | CTGA | 233 | CGGA | 249 | CCGA |
| 202 | CAGT | 218 | CTGT | 234 | CGGT | 250 | CCGT |
| 203 | CAGG | 219 | CTGG | 235 | CGGG | 251 | CCGG |
| 204 | CAGC | 220 | CTGC | 236 | CGGC | 252 | CCGC |
| 205 | CACA | 221 | CTCA | 237 | CGCA | 253 | CCCA |
| 206 | CACT | 222 | CTCT | 238 | CGCT | 254 | CCCT |
| 207 | CACG | 223 | CTCG | 239 | CGCG | 255 | CCCG |
| 208 | CACC | 224 | CTCC | 240 | CGCC | 256 | CCCC |

## Numerical representation of 5-codon set

| Number Code | 5-Codons | Number Code | 5-Codons | Number Code | 5-Codons | Number Code | 5-Codons |
|---|---|---|---|---|---|---|---|
| 1 | AAAAA | 129 | AGAAA | 257 | TAAAA | 385 | TGAAA |
| 2 | AAAAT | 130 | AGAAT | 258 | TAAAT | 386 | TGAAT |
| 3 | AAAAG | 131 | AGAAG | 259 | TAAAG | 387 | TGAAG |
| 4 | AAAAC | 132 | AGAAC | 260 | TAAAC | 388 | TGAAC |
| 5 | AAATA | 133 | AGATA | 261 | TAATA | 389 | TGATA |
| 6 | AAATT | 134 | AGATT | 262 | TAATT | 390 | TGATT |
| 7 | AAATG | 135 | AGATG | 263 | TAATG | 391 | TGATG |
| 8 | AAATC | 136 | AGATC | 264 | TAATC | 392 | TGATC |
| 9 | AAAGA | 137 | AGAGA | 265 | TAAGA | 393 | TGAGA |
| 10 | AAAGT | 138 | AGAGT | 266 | TAAGT | 394 | TGAGT |
| 11 | AAAGG | 139 | AGAGG | 267 | TAAGG | 395 | TGAGG |
| 12 | AAAGC | 140 | AGAGC | 268 | TAAGC | 396 | TGAGC |
| 13 | AAACA | 141 | AGACA | 269 | TAACA | 397 | TGACA |
| 14 | AAACT | 142 | AGACT | 270 | TAACT | 398 | TGACT |
| 15 | AAACG | 143 | AGACG | 271 | TAACG | 399 | TGACG |
| 16 | AAACC | 144 | AGACC | 272 | TAACC | 400 | TGACC |
| 17 | AATAA | 145 | AGTAA | 273 | TATAA | 401 | TGTAA |
| 18 | AATAT | 146 | AGTAT | 274 | TATAT | 402 | TGTAT |
| 19 | AATAG | 147 | AGTAG | 275 | TATAG | 403 | TGTAG |
| 20 | AATAC | 148 | AGTAC | 276 | TATAC | 404 | TGTAC |
| 21 | AATTA | 149 | AGTTA | 277 | TATTA | 405 | TGTTA |
| 22 | AATTT | 150 | AGTTT | 278 | TATTT | 406 | TGTTT |
| 23 | AATTG | 151 | AGTTG | 279 | TATTG | 407 | TGTTG |
| 24 | AATTC | 152 | AGTTC | 280 | TATTC | 408 | TGTTC |
| 25 | AATGA | 153 | AGTGA | 281 | TATGA | 409 | TGTGA |
| 26 | AATGT | 154 | AGTGT | 282 | TATGT | 410 | TGTGT |
| 27 | AATGG | 155 | AGTGG | 283 | TATGG | 411 | TGTGG |
| 28 | AATGC | 156 | AGTGC | 284 | TATGC | 412 | TGTGC |
| 29 | AATCA | 157 | AGTCA | 285 | TATCA | 413 | TGTCA |
| 30 | AATCT | 158 | AGTCT | 286 | TATCT | 414 | TGTCT |
| 31 | AATCG | 159 | AGTCG | 287 | TATCG | 415 | TGTCG |
| 32 | AATCC | 160 | AGTCC | 288 | TATCC | 416 | TGTCC |
| 33 | AAGAA | 161 | AGGAA | 289 | TAGAA | 417 | TGGAA |
| 34 | AAGAT | 162 | AGGAT | 290 | TAGAT | 418 | TGGAT |
| 35 | AAGAG | 163 | AGGAG | 291 | TAGAG | 419 | TGGAG |
| 36 | AAGAC | 164 | AGGAC | 292 | TAGAC | 420 | TGGAC |
| 37 | AAGTA | 165 | AGGTA | 293 | TAGTA | 421 | TGGTA |
| 38 | AAGTT | 166 | AGGTT | 294 | TAGTT | 422 | TGGTT |
| 39 | AAGTG | 167 | AGGTG | 295 | TAGTG | 423 | TGGTG |
| 40 | AAGTC | 168 | AGGTC | 296 | TAGTC | 424 | TGGTC |
| 41 | AAGGA | 169 | AGGGA | 297 | TAGGA | 425 | TGGGA |
| 42 | AAGGT | 170 | AGGGT | 298 | TAGGT | 426 | TGGGT |
| 43 | AAGGG | 171 | AGGGG | 299 | TAGGG | 427 | TGGGG |
| 44 | AAGGC | 172 | AGGGC | 300 | TAGGC | 428 | TGGGC |
| 45 | AAGCA | 173 | AGGCA | 301 | TAGCA | 429 | TGGCA |
| 46 | AAGCT | 174 | AGGCT | 302 | TAGCT | 430 | TGGCT |
| 47 | AAGCG | 175 | AGGCG | 303 | TAGCG | 431 | TGGCG |
| 48 | AAGCC | 176 | AGGCC | 304 | TAGCC | 432 | TGGCC |
| 49 | AACAA | 177 | AGCAA | 305 | TACAA | 433 | TGCAA |
| 50 | AACAT | 178 | AGCAT | 306 | TACAT | 434 | TGCAT |
| 51 | AACAG | 179 | AGCAG | 307 | TACAG | 435 | TGCAG |
| 52 | AACAC | 180 | AGCAC | 308 | TACAC | 436 | TGCAC |
| 53 | AACTA | 181 | AGCTA | 309 | TACTA | 437 | TGCTA |
| 54 | AACTT | 182 | AGCTT | 310 | TACTT | 438 | TGCTT |
| 55 | AACTG | 183 | AGCTG | 311 | TACTG | 439 | TGCTG |
| 56 | AACTC | 184 | AGCTC | 312 | TACTC | 440 | TGCTC |
| 57 | AACGA | 185 | AGCGA | 313 | TACGA | 441 | TGCGA |
| 58 | AACGT | 186 | AGCGT | 314 | TACGT | 442 | TGCGT |
| 59 | AACGG | 187 | AGCGG | 315 | TACGG | 443 | TGCGG |
| 60 | AACGC | 188 | AGCGC | 316 | TACGC | 444 | TGCGC |
| 61 | AACCA | 189 | AGCCA | 317 | TACCA | 445 | TGCCA |
| 62 | AACCT | 190 | AGCCT | 318 | TACCT | 446 | TGCCT |
| 63 | AACCG | 191 | AGCCG | 319 | TACCG | 447 | TGCCG |
| 64 | AACCC | 192 | AGCCC | 320 | TACCC | 448 | TGCCC |
| 65 | ATAAA | 193 | ACAAA | 321 | TTAAA | 449 | TCAAA |
| 66 | ATAAT | 194 | ACAAT | 322 | TTAAT | 450 | TCAAT |
| 67 | ATAAG | 195 | ACAAG | 323 | TTAAG | 451 | TCAAG |
| 68 | ATAAC | 196 | ACAAC | 324 | TTAAC | 452 | TCAAC |
| 69 | ATATA | 197 | ACATA | 325 | TTATA | 453 | TCATA |
| 70 | ATATT | 198 | ACATT | 326 | TTATT | 454 | TCATT |
| 71 | ATATG | 199 | ACATG | 327 | TTATG | 455 | TCATG |
| 72 | ATATC | 200 | ACATC | 328 | TTATC | 456 | TCATC |
| 73 | ATAGA | 201 | ACAGA | 329 | TTAGA | 457 | TCAGA |
| 74 | ATAGT | 202 | ACAGT | 330 | TTAGT | 458 | TCAGT |
| 75 | ATAGG | 203 | ACAGG | 331 | TTAGG | 459 | TCAGG |
| 76 | ATAGC | 204 | ACAGC | 332 | TTAGC | 460 | TCAGC |
| 77 | ATACA | 205 | ACACA | 333 | TTACA | 461 | TCACA |
| 78 | ATACT | 206 | ACACT | 334 | TTACT | 462 | TCACT |
| 79 | ATACG | 207 | ACACG | 335 | TTACG | 463 | TCACG |
| 80 | ATACC | 208 | ACACC | 336 | TTACC | 464 | TCACC |
| 81 | ATTAA | 209 | ACTAA | 337 | TTTAA | 465 | TCTAA |
| 82 | ATTAT | 210 | ACTAT | 338 | TTTAT | 466 | TCTAT |
| 83 | ATTAG | 211 | ACTAG | 339 | TTTAG | 467 | TCTAG |
| 84 | ATTAC | 212 | ACTAC | 340 | TTTAC | 468 | TCTAC |
| 85 | ATTTA | 213 | ACTTA | 341 | TTTTA | 469 | TCTTA |
| 86 | ATTTT | 214 | ACTTT | 342 | TTTTT | 470 | TCTTT |
| 87 | ATTTG | 215 | ACTTG | 343 | TTTTG | 471 | TCTTG |
| 88 | ATTTC | 216 | ACTTC | 344 | TTTTC | 472 | TCTTC |
| 89 | ATTGA | 217 | ACTGA | 345 | TTTGA | 473 | TCTGA |
| 90 | ATTGT | 218 | ACTGT | 346 | TTTGT | 474 | TCTGT |
| 91 | ATTGG | 219 | ACTGG | 347 | TTTGG | 475 | TCTGG |
| 92 | ATTGC | 220 | ACTGC | 348 | TTTGC | 476 | TCTGC |
| 93 | ATTCA | 221 | ACTCA | 349 | TTTCA | 477 | TCTCA |
| 94 | ATTCT | 222 | ACTCT | 350 | TTTCT | 478 | TCTCT |
| 95 | ATTCG | 223 | ACTCG | 351 | TTTCG | 479 | TCTCG |
| 96 | ATTCC | 224 | ACTCC | 352 | TTTCC | 480 | TCTCC |
| 97 | ATGAA | 225 | ACGAA | 353 | TTGAA | 481 | TCGAA |
| 98 | ATGAT | 226 | ACGAT | 354 | TTGAT | 482 | TCGAT |
| 99 | ATGAG | 227 | ACGAG | 355 | TTGAG | 483 | TCGAG |
| 100 | ATGAC | 228 | ACGAC | 356 | TTGAC | 484 | TCGAC |
| 101 | ATGTA | 229 | ACGTA | 357 | TTGTA | 485 | TCGTA |
| 102 | ATGTT | 230 | ACGTT | 358 | TTGTT | 486 | TCGTT |
| 103 | ATGTG | 231 | ACGTG | 359 | TTGTG | 487 | TCGTG |
| 104 | ATGTC | 232 | ACGTC | 360 | TTGTC | 488 | TCGTC |
| 105 | ATGGA | 233 | ACGGA | 361 | TTGGA | 489 | TCGGA |
| 106 | ATGGT | 234 | ACGGT | 362 | TTGGT | 490 | TCGGT |
| 107 | ATGGG | 235 | ACGGG | 363 | TTGGG | 491 | TCGGG |
| 108 | ATGGC | 236 | ACGGC | 364 | TTGGC | 492 | TCGGC |
| 109 | ATGCA | 237 | ACGCA | 365 | TTGCA | 493 | TCGCA |
| 110 | ATGCT | 238 | ACGCT | 366 | TTGCT | 494 | TCGCT |
| 111 | ATGCG | 239 | ACGCG | 367 | TTGCG | 495 | TCGCG |
| 112 | ATGCC | 240 | ACGCC | 368 | TTGCC | 496 | TCGCC |
| 113 | ATCAA | 241 | ACCAA | 369 | TTCAA | 497 | TCCAA |
| 114 | ATCAT | 242 | ACCAT | 370 | TTCAT | 498 | TCCAT |
| 115 | ATCAG | 243 | ACCAG | 371 | TTCAG | 499 | TCCAG |
| 116 | ATCAC | 244 | ACCAC | 372 | TTCAC | 500 | TCCAC |
| 117 | ATCTA | 245 | ACCTA | 373 | TTCTA | 501 | TCCTA |
| 118 | ATCTT | 246 | ACCTT | 374 | TTCTT | 502 | TCCTT |
| 119 | ATCTG | 247 | ACCTG | 375 | TTCTG | 503 | TCCTG |
| 120 | ATCTC | 248 | ACCTC | 376 | TTCTC | 504 | TCCTC |
| 121 | ATCGA | 249 | ACCGA | 377 | TTCGA | 505 | TCCGA |
| 122 | ATCGT | 250 | ACCGT | 378 | TTCGT | 506 | TCCGT |
| 123 | ATCGG | 251 | ACCGG | 379 | TTCGG | 507 | TCCGG |
| 124 | ATCGC | 252 | ACCGC | 380 | TTCGC | 508 | TCCGC |
| 125 | ATCCA | 253 | ACCCA | 381 | TTCCA | 509 | TCCCA |
| 126 | ATCCT | 254 | ACCCT | 382 | TTCCT | 510 | TCCCT |
| 127 | ATCCG | 255 | ACCCG | 383 | TTCCG | 511 | TCCCG |
| 128 | ATCCC | 256 | ACCCC | 384 | TTCCC | 512 | TCCCC |

| Number Code | 5-Codon | Number Code | 5-Codon | Number Code | 5-Codon | Number Code | 5-Codon |
|---|---|---|---|---|---|---|---|
| 513 | GAAAA | 641 | GGAAA | 769 | CAAAA | 897 | CGAAA |
| 514 | GAAAT | 642 | GGAAT | 770 | CAAAT | 898 | CGAAT |
| 515 | GAAAG | 643 | GGAAG | 771 | CAAAG | 899 | CGAAG |
| 516 | GAAAC | 644 | GGAAC | 772 | CAAAC | 900 | CGAAC |
| 517 | GAATA | 645 | GGATA | 773 | CAATA | 901 | CGATA |
| 518 | GAATT | 646 | GGATT | 774 | CAATT | 902 | CGATT |
| 519 | GAATG | 647 | GGATG | 775 | CAATG | 903 | CGATG |
| 520 | GAATC | 648 | GGATC | 776 | CAATC | 904 | CGATC |
| 521 | GAAGA | 649 | GGAGA | 777 | CAAGA | 905 | CGAGA |
| 522 | GAAGT | 650 | GGAGT | 778 | CAAGT | 906 | CGAGT |
| 523 | GAAGG | 651 | GGAGG | 779 | CAAGG | 907 | CGAGG |
| 524 | GAAGC | 652 | GGAGC | 780 | CAAGC | 908 | CGAGC |
| 525 | GAACA | 653 | GGACA | 781 | CAACA | 909 | CGACA |
| 526 | GAACT | 654 | GGACT | 782 | CAACT | 910 | CGACT |
| 527 | GAACG | 655 | GGACG | 783 | CAACG | 911 | CGACG |
| 528 | GAACC | 656 | GGACC | 784 | CAACC | 912 | CGACC |
| 529 | GATAA | 657 | GGTAA | 785 | CATAA | 913 | CGTAA |
| 530 | GATAT | 658 | GGTAT | 786 | CATAT | 914 | CGTAT |
| 531 | GATAG | 659 | GGTAG | 787 | CATAG | 915 | CGTAG |
| 532 | GATAC | 660 | GGTAC | 788 | CATAC | 916 | CGTAC |
| 533 | GATTA | 661 | GGTTA | 789 | CATTA | 917 | CGTTA |
| 534 | GATTT | 662 | GGTTT | 790 | CATTT | 918 | CGTTT |
| 535 | GATTG | 663 | GGTTG | 791 | CATTG | 919 | CGTTG |
| 536 | GATTC | 664 | GGTTC | 792 | CATTC | 920 | CGTTC |
| 537 | GATGA | 665 | GGTGA | 793 | CATGA | 921 | CGTGA |
| 538 | GATGT | 666 | GGTGT | 794 | CATGT | 922 | CGTGT |
| 539 | GATGG | 667 | GGTGG | 795 | CATGG | 923 | CGTGG |
| 540 | GATGC | 668 | GGTGC | 796 | CATGC | 924 | CGTGC |
| 541 | GATCA | 669 | GGTCA | 797 | CATCA | 925 | CGTCA |
| 542 | GATCT | 670 | GGTCT | 798 | CATCT | 926 | CGTCT |
| 543 | GATCG | 671 | GGTCG | 799 | CATCG | 927 | CGTCG |
| 544 | GATCC | 672 | GGTCC | 800 | CATCC | 928 | CGTCC |
| 545 | GAGAA | 673 | GGGAA | 801 | CAGAA | 929 | CGGAA |
| 546 | GAGAT | 674 | GGGAT | 802 | CAGAT | 930 | CGGAT |
| 547 | GAGAG | 675 | GGGAG | 803 | CAGAG | 931 | CGGAG |
| 548 | GAGAC | 676 | GGGAC | 804 | CAGAC | 932 | CGGAC |
| 549 | GAGTA | 677 | GGGTA | 805 | CAGTA | 933 | CGGTA |
| 550 | GAGTT | 678 | GGGTT | 806 | CAGTT | 934 | CGGTT |
| 551 | GAGTG | 679 | GGGTG | 807 | CAGTG | 935 | CGGTG |
| 552 | GAGTC | 680 | GGGTC | 808 | CAGTC | 936 | CGGTC |
| 553 | GAGGA | 681 | GGGGA | 809 | CAGGA | 937 | CGGGA |
| 554 | GAGGT | 682 | GGGGT | 810 | CAGGT | 938 | CGGGT |
| 555 | GAGGG | 683 | GGGGG | 811 | CAGGG | 939 | CGGGG |
| 556 | GAGGC | 684 | GGGGC | 812 | CAGGC | 940 | CGGGC |
| 557 | GAGCA | 685 | GGGCA | 813 | CAGCA | 941 | CGGCA |
| 558 | GAGCT | 686 | GGGCT | 814 | CAGCT | 942 | CGGCT |
| 559 | GAGCG | 687 | GGGCG | 815 | CAGCG | 943 | CGGCG |
| 560 | GAGCC | 688 | GGGCC | 816 | CAGCC | 944 | CGGCC |
| 561 | GACAA | 689 | GGCAA | 817 | CACAA | 945 | CGCAA |
| 562 | GACAT | 690 | GGCAT | 818 | CACAT | 946 | CGCAT |
| 563 | GACAG | 691 | GGCAG | 819 | CACAG | 947 | CGCAG |
| 564 | GACAC | 692 | GGCAC | 820 | CACAC | 948 | CGCAC |
| 565 | GACTA | 693 | GGCTA | 821 | CACTA | 949 | CGCTA |
| 566 | GACTT | 694 | GGCTT | 822 | CACTT | 950 | CGCTT |
| 567 | GACTG | 695 | GGCTG | 823 | CACTG | 951 | CGCTG |
| 568 | GACTC | 696 | GGCTC | 824 | CACTC | 952 | CGCTC |
| 569 | GACGA | 697 | GGCGA | 825 | CACGA | 953 | CGCGA |
| 570 | GACGT | 698 | GGCGT | 826 | CACGT | 954 | CGCGT |
| 571 | GACGG | 699 | GGCGG | 827 | CACGG | 955 | CGCGG |
| 572 | GACGC | 700 | GGCGC | 828 | CACGC | 956 | CGCGC |
| 573 | GACCA | 701 | GGCCA | 829 | CACCA | 957 | CGCCA |
| 574 | GACCT | 702 | GGCCT | 830 | CACCT | 958 | CGCCT |
| 575 | GACCG | 703 | GGCCG | 831 | CACCG | 959 | CGCCG |
| 576 | GACCC | 704 | GGCCC | 832 | CACCC | 960 | CGCCC |
| 577 | GTAAA | 705 | GCAAA | 833 | CTAAA | 961 | CCAAA |
| 578 | GTAAT | 706 | GCAAT | 834 | CTAAT | 962 | CCAAT |
| 579 | GTAAG | 707 | GCAAG | 835 | CTAAG | 963 | CCAAG |
| 580 | GTAAC | 708 | GCAAC | 836 | CTAAC | 964 | CCAAC |
| 581 | GTATA | 709 | GCATA | 837 | CTATA | 965 | CCATA |
| 582 | GTATT | 710 | GCATT | 838 | CTATT | 966 | CCATT |
| 583 | GTATG | 711 | GCATG | 839 | CTATG | 967 | CCATG |
| 584 | GTATC | 712 | GCATC | 840 | CTATC | 968 | CCATC |
| 585 | GTAGA | 713 | GCAGA | 841 | CTAGA | 969 | CCAGA |
| 586 | GTAGT | 714 | GCAGT | 842 | CTAGT | 970 | CCAGT |
| 587 | GTAGG | 715 | GCAGG | 843 | CTAGG | 971 | CCAGG |
| 588 | GTAGC | 716 | GCAGC | 844 | CTAGC | 972 | CCAGC |
| 589 | GTACA | 717 | GCACA | 845 | CTACA | 973 | CCACA |
| 590 | GTACT | 718 | GCACT | 846 | CTACT | 974 | CCACT |
| 591 | GTACG | 719 | GCACG | 847 | CTACG | 975 | CCACG |
| 592 | GTACC | 720 | GCACC | 848 | CTACC | 976 | CCACC |
| 593 | GTTAA | 721 | GCTAA | 849 | CTTAA | 977 | CCTAA |
| 594 | GTTAT | 722 | GCTAT | 850 | CTTAT | 978 | CCTAT |
| 595 | GTTAG | 723 | GCTAG | 851 | CTTAG | 979 | CCTAG |
| 596 | GTTAC | 724 | GCTAC | 852 | CTTAC | 980 | CCTAC |
| 597 | GTTTA | 725 | GCTTA | 853 | CTTTA | 981 | CCTTA |
| 598 | GTTTT | 726 | GCTTT | 854 | CTTTT | 982 | CCTTT |
| 599 | GTTTG | 727 | GCTTG | 855 | CTTTG | 983 | CCTTG |
| 600 | GTTTC | 728 | GCTTC | 856 | CTTTC | 984 | CCTTC |
| 601 | GTTGA | 729 | GCTGA | 857 | CTTGA | 985 | CCTGA |
| 602 | GTTGT | 730 | GCTGT | 858 | CTTGT | 986 | CCTGT |
| 603 | GTTGG | 731 | GCTGG | 859 | CTTGG | 987 | CCTGG |
| 604 | GTTGC | 732 | GCTGC | 860 | CTTGC | 988 | CCTGC |
| 605 | GTTCA | 733 | GCTCA | 861 | CTTCA | 989 | CCTCA |
| 606 | GTTCT | 734 | GCTCT | 862 | CTTCT | 990 | CCTCT |
| 607 | GTTCG | 735 | GCTCG | 863 | CTTCG | 991 | CCTCG |
| 608 | GTTCC | 736 | GCTCC | 864 | CTTCC | 992 | CCTCC |
| 609 | GTGAA | 737 | GCGAA | 865 | CTGAA | 993 | CCGAA |
| 610 | GTGAT | 738 | GCGAT | 866 | CTGAT | 994 | CCGAT |
| 611 | GTGAG | 739 | GCGAG | 867 | CTGAG | 995 | CCGAG |
| 612 | GTGAC | 740 | GCGAC | 868 | CTGAC | 996 | CCGAC |
| 613 | GTGTA | 741 | GCGTA | 869 | CTGTA | 997 | CCGTA |
| 614 | GTGTT | 742 | GCGTT | 870 | CTGTT | 998 | CCGTT |
| 615 | GTGTG | 743 | GCGTG | 871 | CTGTG | 999 | CCGTG |
| 616 | GTGTC | 744 | GCGTC | 872 | CTGTC | 1000 | CCGTC |
| 617 | GTGGA | 745 | GCGGA | 873 | CTGGA | 1001 | CCGGA |
| 618 | GTGGT | 746 | GCGGT | 874 | CTGGT | 1002 | CCGGT |
| 619 | GTGGG | 747 | GCGGG | 875 | CTGGG | 1003 | CCGGG |
| 620 | GTGGC | 748 | GCGGC | 876 | CTGGC | 1004 | CCGGC |
| 621 | GTGCA | 749 | GCGCA | 877 | CTGCA | 1005 | CCGCA |
| 622 | GTGCT | 750 | GCGCT | 878 | CTGCT | 1006 | CCGCT |
| 623 | GTGCG | 751 | GCGCG | 879 | CTGCG | 1007 | CCGCG |
| 624 | GTGCC | 752 | GCGCC | 880 | CTGCC | 1008 | CCGCC |
| 625 | GTCAA | 753 | GCCAA | 881 | CTCAA | 1009 | CCCAA |
| 626 | GTCAT | 754 | GCCAT | 882 | CTCAT | 1010 | CCCAT |
| 627 | GTCAG | 755 | GCCAG | 883 | CTCAG | 1011 | CCCAG |
| 628 | GTCAC | 756 | GCCAC | 884 | CTCAC | 1012 | CCCAC |
| 629 | GTCTA | 757 | GCCTA | 885 | CTCTA | 1013 | CCCTA |
| 630 | GTCTT | 758 | GCCTT | 886 | CTCTT | 1014 | CCCTT |
| 631 | GTCTG | 759 | GCCTG | 887 | CTCTG | 1015 | CCCTG |
| 632 | GTCTC | 760 | GCCTC | 888 | CTCTC | 1016 | CCCTC |
| 633 | GTCGA | 761 | GCCGA | 889 | CTCGA | 1017 | CCCGA |
| 634 | GTCGT | 762 | GCCGT | 890 | CTCGT | 1018 | CCCGT |
| 635 | GTCGG | 763 | GCCGG | 891 | CTCGG | 1019 | CCCGG |
| 636 | GTCGC | 764 | GCCGC | 892 | CTCGC | 1020 | CCCGC |
| 637 | GTCCA | 765 | GCCCA | 893 | CTCCA | 1021 | CCCCA |
| 638 | GTCCT | 766 | GCCCT | 894 | CTCCT | 1022 | CCCCT |
| 639 | GTCCG | 767 | GCCCG | 895 | CTCCG | 1023 | CCCCG |
| 640 | GTCCC | 768 | GCCCC | 896 | CTCCC | 1024 | CCCCC |

## III. LINEAR PREDICTION OF OVERLAPPING N-ARY CODONS

Brucella Suis 1330 genome sequence of length 5806 is used here for case study.

### Linear prediction of overlapping 2-ary codons

The sequence of overlapping 2-ary codons of Brucella Suis 1330 genome sequence is given below. The numerical representation of the sequence of overlapping 2-ary codons is given below. This sequence is of length 5805. Since the sequence length is very large, a part of the total sequence is shown in Fig. 1. Fig. 2 shows its numerical representation. Fig. 3 shows the predicted numerical representation of 2-codons in Brucella Suis 1330 genome sequence (a part is shown here). Fig. 4 shows the error in predicting numerical representation of 2-codons in Brucella Suis 1330 genome sequence (a part is shown here).

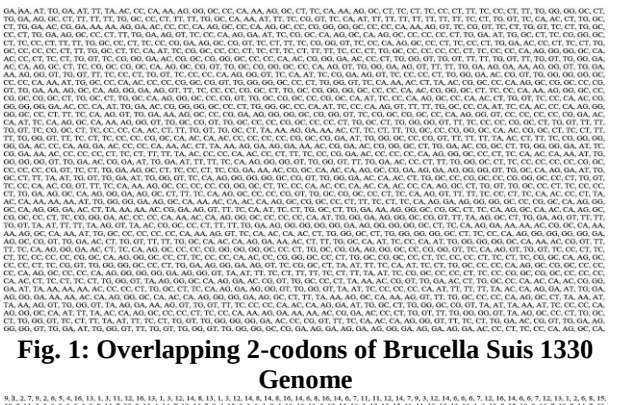


**Fig. 1: Overlapping 2-codons of Brucella Suis 1330 Genome**

**Fig. 2: Numerical representation of 2-codons in Brucella Suis 1330 genome sequence (a part is shown here)**

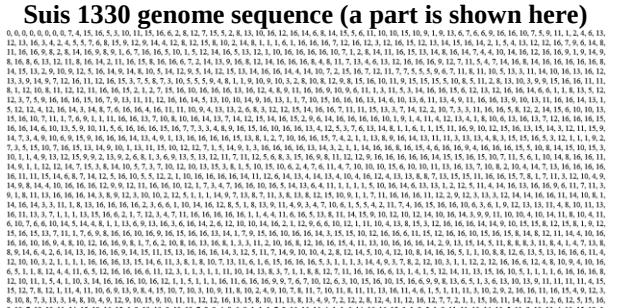


**Fig. 3: Predicted numerical representation of 2-codons in Brucella Suis 1330 genome sequence (a part is shown here)**

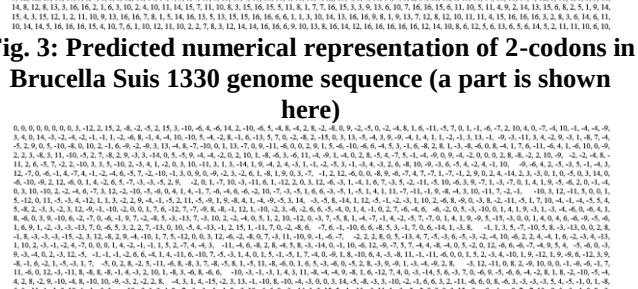
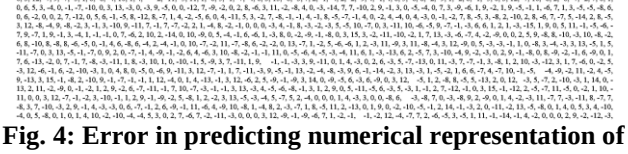
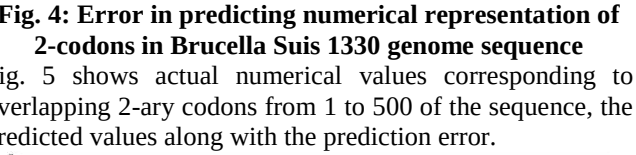


**Fig. 4: Error in predicting numerical representation of 2-codons in Brucella Suis 1330 genome sequence**

Fig. 5 shows actual numerical values corresponding to overlapping 2-ary codons from 1 to 500 of the sequence, the predicted values along with the prediction error.



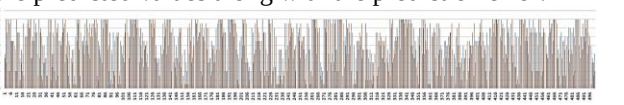
**Actual (Blue) and predicted (red) 2-codons**



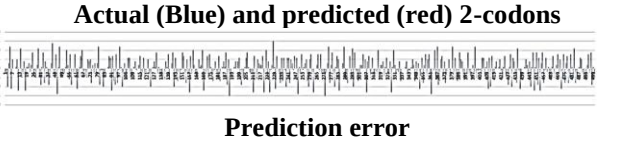
**Prediction error**

**Fig. 5: 1-500 (2-codons)**

Fig. 6 shows actual numerical values corresponding to overlapping 2-ary codons from 501 to 1000 of the sequence, the predicted values along with the prediction error.



**Actual (Blue) and predicted (red) 2-codons**



**Prediction error**

**Fig. 6: 501-1000 (2-codons)**

Due to space limitations, the entire sequence of length 5805 is not presented here.

In this case, overlapping 2-ary codon prediction accuracy is

$$\text{Prediction Accuracy} = \frac{479}{5805} \times 100 = \mathbf{8.251507\ \%}$$

## Linear prediction of overlapping 3-ary codons

The sequence of overlapping 3-ary codons of Brucella Suis 1330 genome sequence is given below. This sequence is of length 5804. Since the sequence length is very large, a part of the total sequence is shown in Fig. 7.

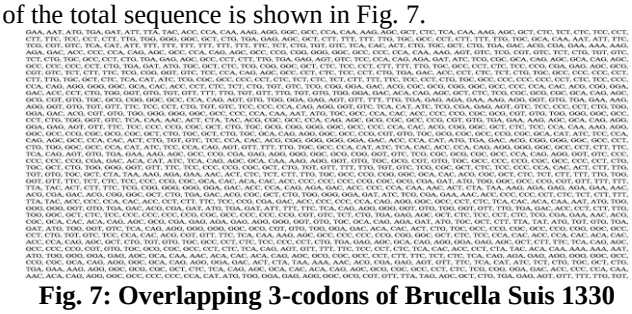


**Fig. 7: Overlapping 3-codons of Brucella Suis 1330 Genome**

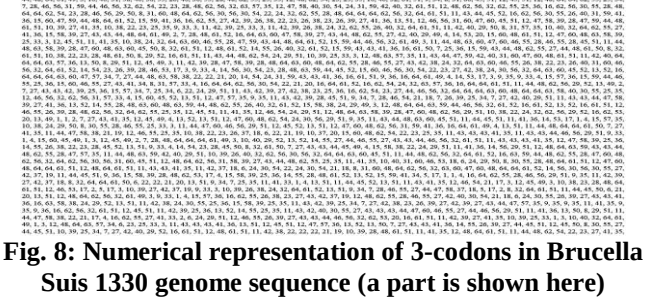


**Fig. 8: Numerical representation of 3-codons in Brucella Suis 1330 genome sequence (a part is shown here)**

Fig. 8 shows its numerical representation. Fig. 9 shows the predicted numerical representation of 3-codons in Brucella Suis 1330 genome sequence (a part is shown here). Fig. 10 shows the error in predicting numerical representation of 3-codons in Brucella Suis 1330 genome sequence (a part is shown here).

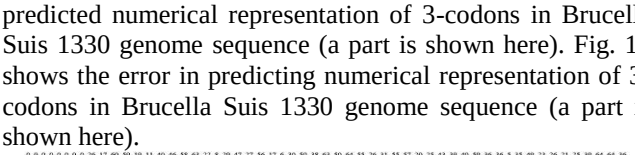


**Fig. 9: Predicted numerical representation of 3-codons in Brucella Suis 1330 genome sequence (a part is shown here)**

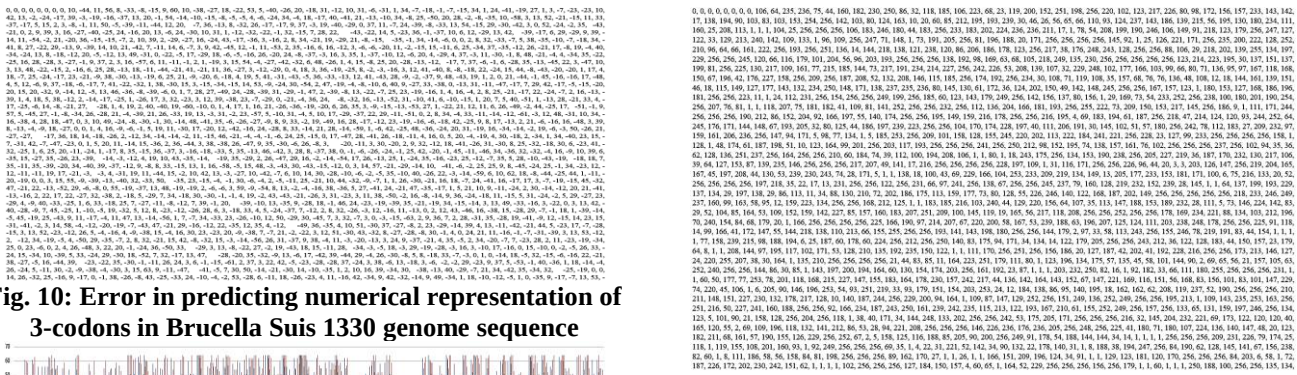
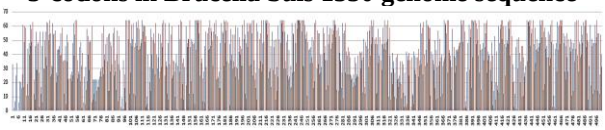
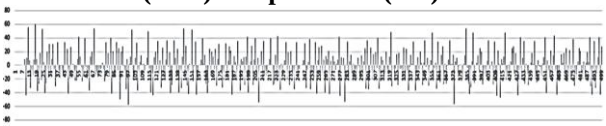


**Fig. 10: Error in predicting numerical representation of 3-codons in Brucella Suis 1330 genome sequence**



**Actual (Blue) and predicted (red) 3-codons**



**Prediction error**

**Fig. 11: 1-500 (3-codons)**

Due to space limitations, the entire sequence of length 5804 is not presented here.

In this case, overlapping 3-ary codon prediction accuracy is

$$\text{Prediction Accuracy} = \frac{107}{5804} \times 100 = \mathbf{1.84323858743\ \%}$$

**Linear prediction of overlapping 4-ary codons**

The sequence of overlapping 4-ary codons of Brucella Suis 1330 genome sequence is given below. This sequence is of length 5803. Since the sequence length is very large, a part of the total sequence is shown in Fig. 12.

**Fig. 12: Overlapping 4-codons of Brucella Suis 1330 Genome**

**Fig. 13: Numerical representation of 4-codons in Brucella Suis 1330 genome sequence (a part is shown here)**

Fig. 14 shows the predicted numerical representation of 4-codons in Brucella Suis 1330 genome sequence (a part is shown here). Fig. 15 shows the error in predicting numerical representation of 4-codons in Brucella Suis 1330 genome



**Fig. 14: Predicted numerical representation of 4-codons in Brucella Suis 1330 genome sequence (a part is shown here)**

**Actual (Blue) and predicted (red) 4-codons**

**Prediction error**

**Fig. 15: 1-500 (4-codons)**

In this case, overlapping 4-ary codon prediction accuracy is

$$\text{Prediction Accuracy} = \frac{28}{5803} \times 100 = \mathbf{0.482342807924\ \%}$$

**Linear prediction of overlapping 5-ary codons**

The sequence of overlapping 5-ary codons of Brucella Suis 1330 genome sequence is given below. This sequence is of length 5802. A part of the total sequence is shown in Fig. 17.

**Fig. 17: Overlapping 5-codons of Brucella Suis 1330 Genome**

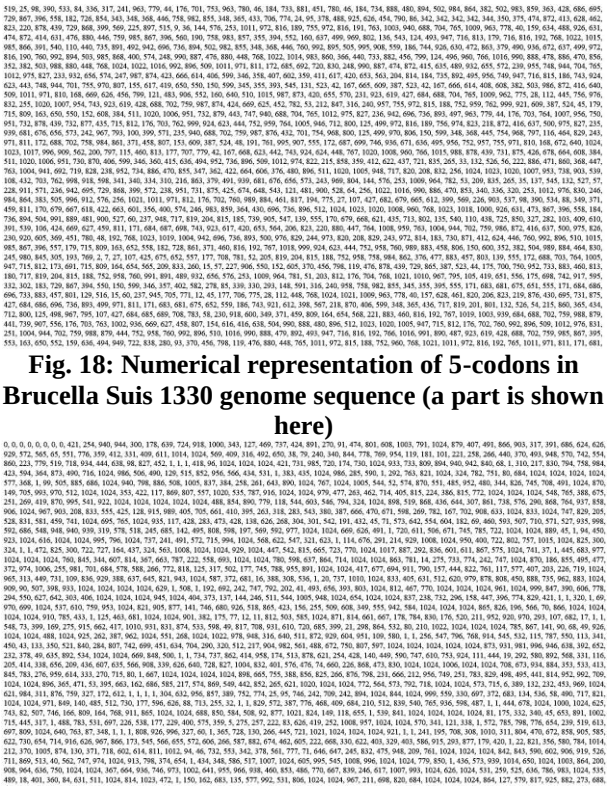


**Fig. 18: Numerical representation of 5-codons in Brucella Suis 1330 genome sequence (a part is shown here)**

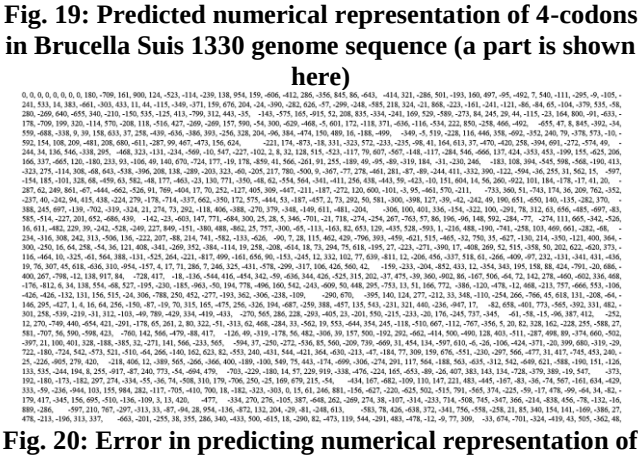


**Fig. 19: Predicted numerical representation of 4-codons in Brucella Suis 1330 genome sequence (a part is shown here)**

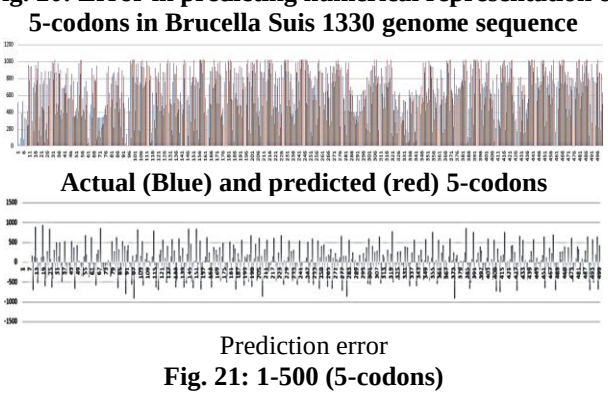


**Fig. 20: Error in predicting numerical representation of 5-codons in Brucella Suis 1330 genome sequence**



**Actual (Blue) and predicted (red) 5-codons**



Prediction error

**Fig. 21: 1-500 (5-codons)**

In this case, overlapping 5-ary codon prediction accuracy is

$$\text{Prediction Accuracy} = \frac{6}{5802} \times 100 = \mathbf{0.10341261633\ \%}$$

## IV CONCLUSION

A genome arrangement from a solitary nucleotide is displayed as a straight dynamical framework like a self-impelled moving body, be it a physical body or an occasion. For instance, an engine vehicle is a self-pushed dynamical framework, not constrained by any outside office yet by the interior office the driver. Another model is the occasion of movement could be applied to a dynamical framework, which is self-impelled. Considering this thought a novel direct estimator called "Octet Algorithm" is presented in this proposal and a considerable report completed to foresee nucleotides and codons of a genome grouping. In this unique circumstance, an original thought of covering codons is presented, which obviously must be approved and approved exactly by hereditary specialists. In any case, the idea works in the system of hypothetical software engineering.

## V. RESULTS

### Overlap Prediction

| Codon Sequence | Prediction Accuracy | No. of Codons | Similar Codons |
|---|---|---|---|
| 1-codon | 28.71662365 | 5806 | 1667 |
| 2-codon | 8.251507234 | 5805 | 479 |
| 3-codon | 1.843238588 | 5804 | 107 |
| 4-codon | 0.482542807 | 5803 | 28 |
| 5-codon | 0.103412616 | 5802 | 6 |

## REFERENCES


1. Sean Eddy (2012) The C-value paradox, junk DNA, and ENCODE, Curr Biol 22(21):R898–R899.
2. Doolittle, W. Ford (2013). "Is junk DNA bunk? A critique of ENCODE". Proc Natl Acad Sci USA **110** (14): 5294–5300. Bibcode:2013PNAS..110.5294D. doi:10.1073/pnas.1221376110. PMC 3619371. PMID 23479647.
3. Palazzo, Alexander F.; Gregory, T. Ryan (2014). "The Case for Junk DNA". PLoS Genetics **10** (5): e1004351. doi:10.1371/journal.pgen.1004351. ISSN 1553-7404.
4. Waterhouse, Peter M.; Hellens, Roger P. (25 March 2015). "Plant biology: Coding in non-coding RNAs". Nature **520** (7545): 41–42. doi:10.1038/nature14378.
5. Tutar, Y. (2012). "Pseudogenes". Comp Funct Genomics **2012**: 424526. doi:10.1155/2012/424526. PMC 3352212. PMID 22611337.
6. Petrov DA, Hartl DL; Hartl (2000). "Pseudogene evolution and natural selection for a compact genome". J. Hered. **91** (3): 221–7. doi:10.1093/jhered/91.3.221. PMID 10833048.
7. Ohno, Susumu (1972). H. H. Smith, ed. So Much "junk" DNA in Our Genome. Gordon and Breach, New York. pp. 366–370. Retrieved 2013-05-15.


## AUTHORS PROFILE

**P Venkateswarlu** completed his M.Tech Computer Science & Engineering and pursuing Ph.D from University of Mysore. His area of interest is in AI, Data Mining, Machine Learning, Software Engineering.

**Dr. E G Rajan,** done his Ph.D form IIT-Kanpur. His area of interest is in Signal and Image Processing, Data Mining.