BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(11)

CARd: Carbon distribution analysis program for protein sequences

Ekambaram Rajasekaran*

Department of Bioinformatics, School of Biotechnology and Health Sciences, Karunya University, Karunya Nagar, Coimbatore 641114; Ekambaram Rajasekaran – Email: ersekaran@gmail.edu; * Corresponding author

Received December 27, 2011; Accepted May 11, 2012; Published June 16, 2012

Abstract:
Carbon distribution is responsible for stability and structure of proteins. Arrangement of carbon along the protein sequence is depends on how the amino acids are organized and is guided by mRNAs. An atomic level revision is important for understanding these codes. This will ultimately help in identification of disorders and suggest mutations. For this purpose a carbon distribution analysis program has been developed. This program captures the hydrophobic / hydrophilic / disordered regions in a protein. The program gives accurate results. The calculations are precise and sensitive to single amino acid resolution. This program is to help in mutational studies leading to protein stabilisation.

Keywords: carbon distribution; mutational study, hydrophobic hydrophilic, program, protein disorder, hydropathy

Background:

Proteins evolve in response to the nature of interaction and stability. Hydrophobic force considered to be the major force involved in protein folding and action. Carbon is the element contributes towards this dominant force. Arrangement of this carbon along the sequence depends on how the amino acids are organized. Recent studies on this matter find that proteins prefer to have 31.45% of carbon **[1, 2]** for stability in its structure and sequence. Given this standard, a carbon analysis program **[2]** has been developed to study the carbon distribution profile of protein sequences. This is not sensitive to single amino acid level. However it program can be used to see carbon along the sequence **[3]**. This program helps in identification of ligand binding sites **[4, 5]** and to understand the problem of protein-protein and protein-DNA interactions **[6].** However protein disorders in a short stretch of sequence and possible mutations are not possible to predict. It is reported that allotment of carbon along the sequence is responsible for disorders in proteins **[7]**. With this idea and carbon scale, a new program CARd has been developed for sensitive measure of hydrophobic quantity at amino acid level.

Methodology:

Dataset

The sample sequence of super oxide dismutase (SOD) was taken from Swissprot database. The UniprotKB ID number of the protein is P00441.

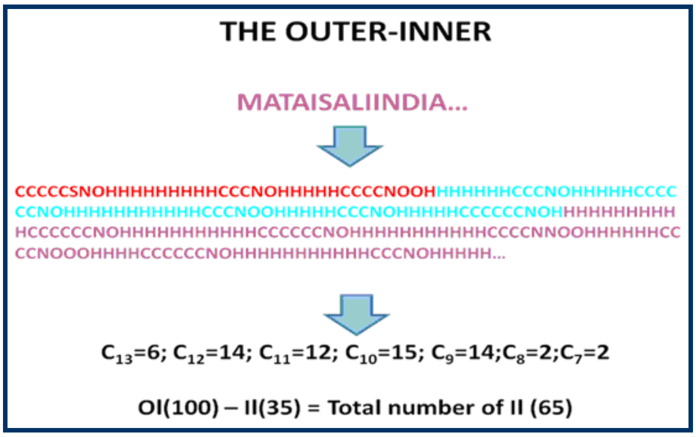
THE OUTER-INNER
MATAISALIINDIA...
CCCCCSNOHHHHHHHHHHCCCNOHHHHHHCCCCNOOHHHHHHHCCCNOHHHHHHCCCC
CCNOHHHHHHHHHHHHCCCNOOHHHHHCCCNOHHHHHHCCCCCCNOHHHHHHHHHH
HCCCCCCNOHHHHHHHHHHHCCCCCCNOHHHHHHHHHHHHCCCCNNOOHHHHHHCC
CCNOOOHHHHCCCCCCNOHHHHHHHHHHHHCCCNOHHHHH...
C_{13}=6; C_{12}=14; C_{11}=12; C_{10}=15; C_9=14;C_8=2;C_7=2
OI(100) – II(35) = Total number of II (65)

Figure 1: Flow Diagram showing how carbon distribution obtained with Outer-Inner length method

© 2012 Biomedical Informatics

Method

The flow diagram **(Figure 1)** outlines how the distribution of inner lengths based on carbon fraction is counted in a outer length. The pink coloured sequence is protein sequence which is converted into atomic sequence (shown in multiple colours (red+blue+pink). The red portion is an inner length. The blue portions are outer length which includes the red portion as well. The entire atomic sequences are given in pink colour which includes both blue and red. Here the outer and inner lengths are taken as 100 and 35 atoms. There are 65 (100-35) inner lengths for statistics. These 65 inner lengths are grouped based on carbon fraction. The C_{11} has (11/35) 0.31 carbon fraction. The flow chart **(Figure 2**) on the other side shows how the algorithm works in calculation. A program has been written to do all this calculations. The program reads the protein sequence, converts into atomic sequence, takes a length (anything from 77 to 700 atoms) of sequence, split into small lengths (from 32 to -350 atoms) of equal sizes, finds fraction of carbon atoms in all this small lengths and counts number of small lengths that contain a defined fraction of carbon. There are small lengths with 0.25 / 0.45 carbon but maximum at around 0.31. A distribution of range of small lengths based on carbon fraction appears like a normal distribution curve. This distribution curve is obtained for all possible outer length. The outer lengths can be any length between 77 and 700 corresponds to 5 and 45 amino acids. Any outer length chosen between 100 and 150 atoms is sufficient for most of the observations. Here it is chosen as 150 in all calculations. The inner length can be between 32 and 350 which are not exceeding half of the outer length. Inner length of 35 atoms is chosen here in all calculations as it is the smallest unit with 11 carbons which can produce fraction of 0.31. The outer length is moved with step value of selected atoms. Normally it is half of the outer length. Here it is 75 atoms as the outer length is 150. A carbon distribution profile is obtained for all possible outer length.

Generally normal profiles will have a Gaussian distribution curve with maximum frequency at 0.3145. Any shift from this maximum is considered as hydrophilic (negative side) or hydrophobic (positive side). Difference in normal distribution is considered as disordered outer length which contains improper amino acid distribution. When the outer length contains a proper arrangement of amino acids then there is normal distribution. That is improper arrangements can be identified from this calculations means that a stretch of sequence which are hydrophobic or hydrophilic or unstable can be predicted. The statistical mean, median, mode and standard deviation of the distribution profile curve can be obtained to check whether the distribution is normal or not. While the mean, median and mode are equal for a given stretch, it is considered as normal distribution. Then the stretch of sequence is in proper mode of arrangement. CARd program computes for every (outer length) distribution profile and prints at the end of the profile results. CARd, the flexible program has option to output mean, median, mode and standard deviation at every amino acid site. It also gives simple average carbon fraction at an amino acid site and for given output length. This replaces our earlier CARBANA program **[2]** for carbon analysis. A plot of mean value versus amino acid number can be plotted. This plot is to see the overall hydrophobic or hydrophilic regions in the sequence. It is similar to hydropathy plot.

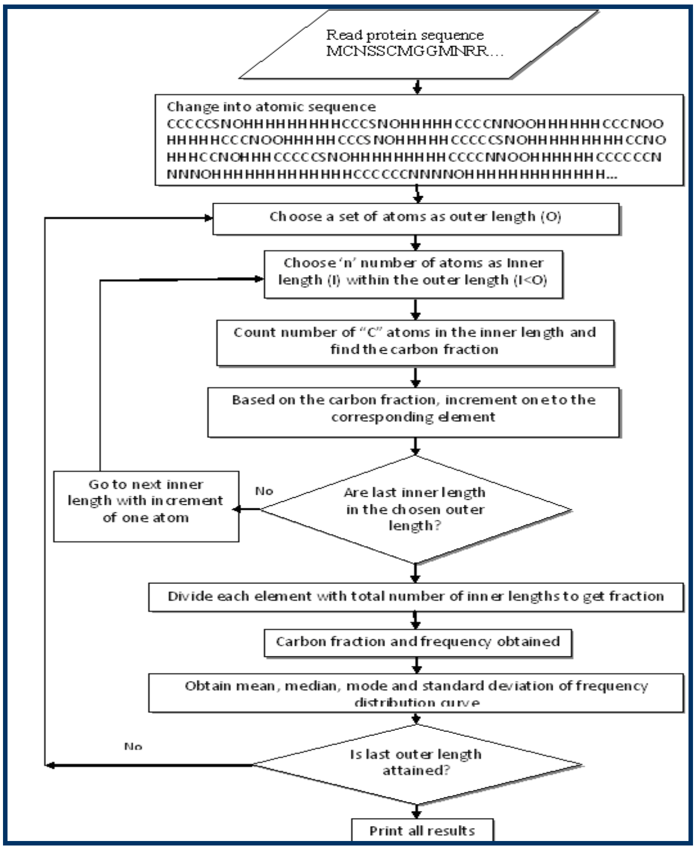
Read protein sequence
MCNSSCMGGMNRR…
Change into atomic sequence
CCCCCSNOHHHHHHHHHHHCCCSNOHHHHHCCCCNNOOHHHHHHCCCNOO
HHHHHCCCNOOHHHHHCCCSNOHHHHHCCCCCSNOHHHHHHHHHCCNO
HHHCCNOHHHCCCCCSNOHHHHHHHHHCCCCNNOOHHHHHCCCCCCN
NNNOHHHHHHHHHHHHHCCCCCCNNNNOHHHHHHHHHHHHH…
Choose a set of atoms as outer length (O)
Choose 'n' number of atoms as Inner length (I) within the outer length (I<O)
Count number of "C" atoms in the inner length and find the carbon fraction
Based on the carbon fraction, increment one to the corresponding element
Go to next inner length with increment of one atom
No
Are last inner length in the chosen outer length?
Divide each element with total number of inner lengths to get fraction
Carbon fraction and frequency obtained
Obtain mean, median, mode and standard deviation of frequency distribution curve
No
Is last outer length attained?
Print all results

Figure 2: Flow chart showing how the algorithm works in the CARd program.

Discussion:

Carbon distribution profile

CARd analysis is carried out for super oxide dismutase sequence. An outer length of 150 atoms and inner length of 35 atoms are used. The carbon distribution plot **(Figure 3)** is obtained for every outer length with a gap of 75 atoms. The plots show the frequency versus carbon fraction. Each plot is labeled with the range of amino acids involved in the particular outer length. The first plot shows distribution profile for first 10 amino acids and the next one is for 11 amino acids from 6 to 16. Third outer length is from 10 to 21 amino acids and 11 amino acids long. This way the distribution plot till last possible outer length is obtained. Each plot shows a different distribution profile. Outer lengths 72-82, 77-88, 82-93, 88-99, 114-123, 119-129 and 141-152 are having normal distribution and considered as stable regions. Outer lengths 88-99, 114-123 and 129-141 are in perfect distribution. Amino acids in these lengths are arranged with perfect carbon distribution. Outer lengths 93-104 and 99-110 are in order based on carbon distribution but hydrophobic in nature. Outer lengths 50-63, 110-119, 123-135, 129-141 and 135-146 are having normal carbon distribution but hydrophilic. Similarly the outer lengths 56-68, 63-72, 68-77 and 104-114 are hydrophilic plus disorder. Infact 56-68, 63-72 and 68-77 are metal binding sites. There are metal binding sites such as 46-56, 77-88, 82-93 and 114-123 are having normal carbon distribution. The outer length 50-63 and 56-68 has a disulphide bond that stabilise the structure but carbon distribution is not in order. The disordered outer lengths need to be taken for mutational study. Similar plot on abnormal proteins reveal that most of the stretches are in disrorder regions. So this carbon distribution analysis program can find disordered proteins, small stretch that are disorder and amino acid responsible for disorder.This mathod is sensitive to single amino acid level. This can be better exploited for mutational study leading to stabilisation of proteome.

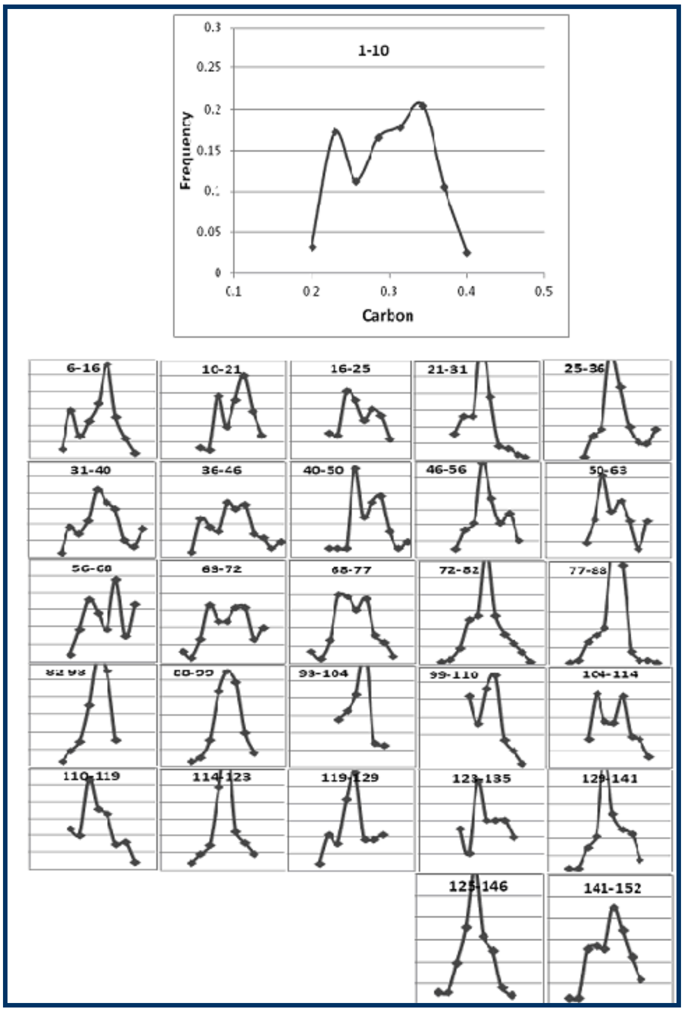

Figure 3: CARd analysis on SOD protein with 150 atoms (10 amino acid) of outer length and 35 atoms of inner length. Gap between each outer length is chosen as75 atoms. The individual plots are frequency versus carbon fraction for the short stretch labeled. For example the first plot is for first 10 amino acids, the second one is for residues 6-16, third one is for residues 10-21 and so on. The distribution plot is shown till last possible outer length. Each plot shows a different distribution curve. If the distribution plot is normal and maximum frequency at 0.3145, then the particular stretch is normal and stable (e.g. 77-88 and 88-99). When the curve is normal and shifted to left or right then the stretch is a hydrophilic (e.g. 125-146) or hydrophobic (e.g. 10-21) region. If the distribution in not normal then it is disordered region (e.g. 104-114).

Mean median, mode and standard deviation of the distribution profile

The statistical mean, median, mode and standard deviation of the distribution curve obtained for SOD. The results are not shown. It gives output of average carbon fraction, mean, median with mode and standard deviation of the distribution profile at every amino acid position as shown in **Table 1** (**see supplementary material).** This is achieved by selecting step size of suitable atoms. A plot of amino acid number versus mean values can be obtained to identify the hydrophobic or hydrophilic regions along the sequence. This is similar to CARBANA output on average carbon fraction along the sequence.

Conclusion:

CARbon distribution (CARd) analysis program has been developed to capture the hydrophobic/hydrophilic regions in proteins. The calculations are precise and sensitive to single amino acid change. This program is hoped to help in mutational studies leading to protein stabilisation. It can capture the essence of hydrophobicity in small stretch of sequence.

References:

[1] Rajasekaran E *et al.* IACSIT-SC, *IEEE* 2009 452

[2] Rajasekaran E & Vijayasarathy M, *Bioinformation.* 2009 **3:** 409 [PMID: 21423892]

[3] Senthil R & Rajasekaran E, *J Adv Biotech.* 2009 **8**: 30

[4] Senthil R *et al. J Adv Bioinfo Appln Res.* 2011 **2:** 98

[5] Princeinal Suganthi S *et al. Int J Bioinfo.* 2011 **4:** 25

[6] Senthil R & Rajasekaran E, *Int J Adv Bioinfo.* 2010 **1**:7

[7] Rajasekaran E *et al. Bioinfo.* 2011 **6:** 291 [PMID: 21769187]

Edited by P Kangueane

Citation: **Rajasekaran,** Bioinformation 8(11): 508-512 (2012)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited

Supplementary material:

Table 1: CARd output for SOD sequence. Amino acid number, average carbon (CARBANA result), mean, median, mode and standard deviation are given. (Input parameters are OL=135, IL=35 and step=15)

Amino Acid	Carbon (Avg)	Statistical. Mean C	Median C	Mode C	Statistical Deviation
5	0.3037	0.3042	0.3143	0.3429	0.0499
6	0.3037	0.2981	0.3143	0.3429	0.0509
7	0.2889	0.3036	0.3143	0.3429	0.0538
8	0.2815	0.3092	0.3143	0.3429	0.0539
9	0.3037	0.3170	0.3143	0.3429	0.0534
10	0.3037	0.3149	0.3143	0.3429	0.0547
11	0.3111	0.3139	0.3143	0.3429	0.0535
12	0.3037	0.3166	0.3429	0.3429	0.0560
14	0.2889	0.3122	0.3143	0.3429	0.0558
15	0.3185	0.3246	0.3429	0.3714	0.0525
16	0.3037	0.3414	0.3429	0.3714	0.0488
17	0.3481	0.3439	0.3429	0.3714	0.0524
18	0.3481	0.3422	0.3429	0.2857	0.0538
19	0.3481	0.3368	0.3429	0.3714	0.0547
20	0.3481	0.3363	0.3143	0.2857	0.0528
21	0.3185	0.3328	0.3143	0.2857	0.0553
21	0.3185	0.3237	0.3143	0.2857	0.0570
22	0.3333	0.3284	0.3143	0.3143	0.0560
23	0.3185	0.3250	0.3143	0.3143	0.0543
24	0.3259	0.3208	0.3143	0.3143	0.0517
25	0.2963	0.3111	0.3143	0.3143	0.0450
26	0.3037	0.3074	0.3143	0.3143	0.0435
27	0.2963	0.3237	0.3143	0.3143	0.0590
29	0.2963	0.3340	0.3143	0.3143	0.0636
30	0.3407	0.3384	0.3143	0.3143	0.0588
31	0.3333	0.3443	0.3143	0.3143	0.0555
31	0.3556	0.3433	0.3429	0.3143	0.0577
32	0.3333	0.3374	0.3143	0.3143	0.0645
33	0.3259	0.3269	0.3143	0.3143	0.0669
34	0.3259	0.3311	0.3143	0.3143	0.0661
35	0.3037	0.3303	0.3143	0.3143	0.0667
36	0.3111	0.3200	0.3143	0.3143	0.0552
37	0.3185	0.3130	0.3143	0.3143	0.0509
38	0.3111	0.3134	0.3143	0.3143	0.0522
39	0.3185	0.3200	0.3143	0.3429	0.0622
40	0.2889	0.3271	0.3143	0.3143	0.0688
41	0.3481	0.3408	0.3429	0.3143	0.0667
42	0.3407	0.3477	0.3429	0.3143	0.0606
43	0.3333	0.3471	0.3429	0.3143	0.0605
44	0.3333	0.3519	0.3429	0.3143	0.0567
46	0.3481	0.3582	0.3429	0.3143	0.0578
46	0.3481	0.3643	0.3714	0.3143	0.0558
47	0.3556	0.3674	0.3714	0.3143	0.0552
48	0.3481	0.3613	0.3429	0.3143	0.0492
49	0.3778	0.3508	0.3429	0.3143	0.0488
50	0.3407	0.3372	0.3143	0.3143	0.0477
51	0.3333	0.3330	0.3143	0.3143	0.0514
51	0.3481	0.3294	0.3143	0.3429	0.0531
53	0.3407	0.3231	0.3143	0.2857	0.0557
54	0.3185	0.3145	0.3143	0.2857	0.0506
55	0.3185	0.3141	0.3143	0.2857	0.0492
57	0.2889	0.3179	0.3143	0.2857	0.0534
58	0.3185	0.3261	0.3143	0.2857	0.0579
59	0.3333	0.3355	0.3429	0.3714	0.0598
61	0.3259	0.3414	0.3429	0.3714	0.0602
63	0.3333	0.3460	0.3429	0.3714	0.0566
64	0.3481	0.3498	0.3714	0.3714	0.0565
64	0.3259	0.3477	0.3714	0.3714	0.0589
65	0.3556	0.3473	0.3429	0.3714	0.0585
66	0.3333	0.3330	0.3429	0.4286	0.0685
67	0.3333	0.3200	0.3143	0.2571	0.0670
68	0.3185	0.3111	0.3143	0.3429	0.0630
69	0.2889	0.3042	0.3143	0.2571	0.0608
70	0.2963	0.3053	0.3143	0.2571	0.0582
71	0.3037	0.3017	0.2857	0.2857	0.0568
71	0.2815	0.2975	0.2857	0.2571	0.0545
72	0.2963	0.3076	0.3143	0.2857	0.0574
74	0.2815	0.3116	0.3143	0.3143	0.0605
75	0.3037	0.3176	0.3143	0.3143	0.0530
76	0.3037	0.3160	0.3143	0.3143	0.0543
77	0.3185	0.3168	0.3143	0.3143	0.0518
78	0.3185	0.3147	0.3143	0.3143	0.0492

79	0.3111	0.3137	0.3143	0.3143	0.0471
80	0.3037	0.3155	0.3143	0.3143	0.0471
80	0.3185	0.3164	0.3143	0.3143	0.0473
81	0.3111	0.3132	0.3143	0.3143	0.0443
82	0.3111	0.3071	0.3143	0.3143	0.0399
84	0.3037	0.3069	0.3143	0.3143	0.0389
85	0.2963	0.3158	0.3143	0.3143	0.0356
85	0.3037	0.3139	0.3143	0.3143	0.0360
87	0.3259	0.3111	0.3143	0.3429	0.0388
88	0.3185	0.3086	0.3143	0.3143	0.0376
89	0.2963	0.3086	0.3143	0.3143	0.0382
90	0.2889	0.3118	0.3143	0.3143	0.0363
92	0.3185	0.3124	0.3143	0.3143	0.0364
92	0.3259	0.3155	0.3143	0.3143	0.0387
93	0.3185	0.3183	0.3143	0.3429	0.0401
95	0.3259	0.3145	0.3143	0.3429	0.0396
96	0.3185	0.3151	0.3143	0.3429	0.0401
97	0.3037	0.3206	0.3143	0.3429	0.0360
98	0.3111	0.3239	0.3429	0.3429	0.0388
100	0.3185	0.3210	0.3143	0.3429	0.0401
100	0.3259	0.3210	0.3143	0.3429	0.0439
101	0.3111	0.3170	0.3143	0.3429	0.0456
102	0.3037	0.3116	0.3143	0.2571	0.0434
104	0.3037	0.3143	0.3143	0.3429	0.0440
105	0.3037	0.3166	0.3143	0.3429	0.0440
105	0.3185	0.3195	0.3143	0.3429	0.0472
107	0.3037	0.3193	0.3143	0.2571	0.0504
107	0.3259	0.3155	0.3143	0.2571	0.0496
109	0.3185	0.3202	0.3143	0.3429	0.0522
110	0.2889	0.3122	0.3143	0.3429	0.0542
111	0.3259	0.3097	0.3143	0.3429	0.0555
112	0.2815	0.3097	0.2857	0.2857	0.0550
113	0.2889	0.3095	0.3143	0.2857	0.0523
114	0.2963	0.3053	0.2857	0.2857	0.0511
115	0.2815	0.3055	0.2857	0.2857	0.0492
116	0.2963	0.3076	0.3143	0.2857	0.0461
117	0.2963	0.3111	0.3143	0.3143	0.0478
118	0.3111	0.3080	0.3143	0.3143	0.0422
119	0.3333	0.3118	0.3143	0.3143	0.0393
120	0.2741	0.3124	0.3143	0.3143	0.0397
121	0.3259	0.3227	0.3143	0.3143	0.0437
121	0.3259	0.3206	0.3143	0.3143	0.0439
122	0.3259	0.3223	0.3143	0.3143	0.0490
123	0.3037	0.3149	0.3143	0.3143	0.0498
124	0.3111	0.3061	0.3143	0.3143	0.0529
125	0.2815	0.3032	0.3143	0.3143	0.0515
127	0.2889	0.3063	0.3143	0.2857	0.0542
127	0.2889	0.3092	0.3143	0.2857	0.0544
129	0.3185	0.3126	0.3143	0.2857	0.0526
129	0.2963	0.3074	0.2857	0.2857	0.0491
131	0.2889	0.3048	0.2857	0.2857	0.0489
133	0.2889	0.2956	0.2857	0.2857	0.0479
134	0.3037	0.2975	0.2857	0.2857	0.0468
135	0.2815	0.3021	0.2857	0.2857	0.0449
136	0.3037	0.3017	0.2857	0.2857	0.0468
137	0.2815	0.2971	0.2857	0.2857	0.0441
137	0.2963	0.2973	0.2857	0.2857	0.0426
138	0.2741	0.2897	0.2857	0.2857	0.0457
140	0.2963	0.2845	0.2857	0.2857	0.0448
141	0.2889	0.2872	0.2857	0.2857	0.0483
143	0.2667	0.2977	0.2857	0.2857	0.0481
144	0.2889	0.2994	0.3143	0.3143	0.0496
145	0.3185	0.3046	0.3143	0.3429	0.0535
145	0.2963	0.3025	0.3143	0.3143	0.0509
147	0.3111	0.3032	0.3143	0.3143	0.0547
148	0.2889	0.2905	0.2857	0.2286	0.0637
149	0.3037	0.2912	0.3143	0.2286	0.0694