Hamming distance as a concept in DNA molecular recognition

Mina Mohammadi-Kambs[1], Kathrin Hölz[2], Mark M. Somoza[2] and Albrecht Ott[1]*

[1]Biological Experimental Physics, Saarland University, Campus B2.1, 66123 Saarbrücken, Germany

[2]Institute of Inorganic Chemistry, Faculty of Chemistry, University of Vienna, Althanstraße 14 (UZA II), 1090 Vienna, Austria

S1 Number of sequences with at least one run of 4G

In the following we show how to derive Equation (1).

To avoid 4G sequences any run of 'G' must be interrupted by a non-'G' base after at maximum 3 G. Sequences containing no runs of 4G can be constructed in the following way:

Each sequence of length L contains m guanine and k non-guanine bases. Therefore, $L=k+m$. For a given k there are no more than $k+1$ groups of 'G' bases, (0,1,…,k) corresponding to A, B, C, D etc.

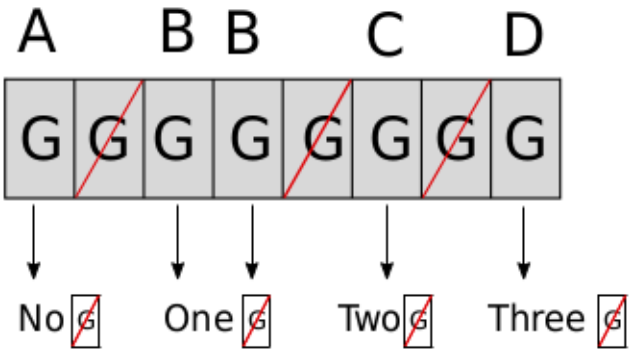

Figure S1. All bases of a sequence are either guanine or not guanine. Each guanine can then be categorized in how many non-guanine bases occurred before it, reading the sequence from left to right. This sequence can be represented by ABBCD.

To determine the number of possible sequences without runs of 4G at the given k we first need to know how many of the letter sequences made of (A, B, C,..) of length L-k (here 5) we can construct, given the condition that each letter appears at most 3 times. This is equivalent to taking out L-k elements from three identical sets that each contains $k+1$ elements {A, B, C, D, …}. The number of possibilities to pick $m=L-k$ elements out of 3 identical sets consisting of $k+1$ elements, is then given by quadrinomial coefficient [1-2]:

$$P(L)=\binom{k+1}{L-k}_3 \tag{1}$$

To determine the number of possibilities for all sequences without at least one run of ‘4G’ we sum up $P(L)$ over all k for a given length and multiply by 3^k to cover all possible combinations of non-‘G’ bases.

$$N_{4G}(L) = 4^L - \sum_{k\,\min}^{L} \underbrace{\binom{k+1}{L-k}_3}_{\text{quadrinomial coefficient}} \times 3^k \qquad k_{\min} = \underbrace{\left\lceil \frac{L-3}{4} \right\rceil}_{\text{ceil-function}} \tag{2}$$

For a given length, $K_{\min}$ is the minimum number of non-‘G’ bases necessary to avoid runs of 4G.

S2 Maximal independent set used in experiment

Table S1. Independent set with L=7, d=5

sequence list from 3'-5'	
1	CTAATTGACTC
2	CTACACCCTTC
3	CTACTGTGGTC
4	CTAGCGAAATC
5	CTAGGATCCTC
6	CTCATACCGTC
7	CTCCATGTATC
8	CTCGGCATTTC
9	CTCTAGTACTC
10	CTGAACTTGTC
11	CTGCGTAGCTC
12	CTGGCTGCTTC
13	CTGGTGCTCTC
14	CTGTGACATTC
15	CTGTTCACATC
16	CTTAAGAGTTC
17	CTTCCAATGTC
18	CTTCGCTAATC
19	CTTGATCAGTC
20	CTTGTAGGATC
21	CTTTCCCGCTC
22	CTTTGGGCGTC
23	CTTTTTTTTTC

S3 Comparison of set sizes

Here we present a summary of the set sizes achieved by our algorithm. Both tables S2 and S3 show all sizes for $L = 4$ until $L = 7$ with $d = 2 \ldots L\text{-}1$. In contrast to table S2, table S3 contains the set sizes for the case that the available sequences are restricted to a 50 % GC content. All sizes are compared to literature values[3-4], if available.

Table S2: Set sizes achieved by the algorithm in comparison to literature values. Set sizes for removing and not removing 4G and 4C sequences are the same except for L=5, d=2, L=6, d=2. The numbers after backslash correspond to set sizes after crossing 4G and 4C.

Sequence length L	Minimum Hamming distance d	Maximal set size	Set size from literature
4	3	16	-
4	2	64	-
5	4	16	16^3
5	3	64	-
5	2	256/252	-
6	5	9	9^3
6	4	64	64^3
6	3	114	-
6	2	1024/1001	-
7	6	6	8^3
7	5	23	23^3
7	4	83	78^3
7	3	364	-
7	2	4096	-

Table S3: Set sizes achieved by the algorithm in comparison to literature values, if the sequences are restricted to a 50 % GC content.

Sequence length L	Minimum Hamming distance d	Number of GC bases	Maximal set size	Set size from literature
4	3	2	12	-
4	2	2	48	$48^4/48^3$
5	4	2	10	-
5	3	2	27	-
5	2	2	156	$120^4/142^3$
6	5	3	8	-
6	4	3	36	-
6	3	3	82	$56^4/85^3$
6	2	3	640	-
7	6	3	7	-
7	5	3	21	-
7	4	3	65	-
7	3	3	238	$224^4/230^3$
7	2	3	2240	-

Figure S2 depicts some of our set sizes as a function of sequence length for different minimum hamming distance.

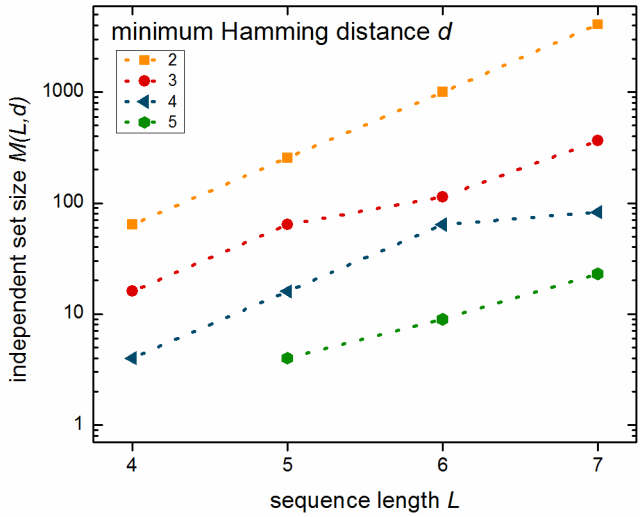

Figure S2. Some of the set sizes of largest independent set *M(L,d)* found by our algorithm as a function of length *L* for different *d*.

To estimate the maximal possible size of these sets *M*(*L*,*d*), coding theory proposes the Singleton and the Gilbert-Varshamov[5-6] as an upper and lower bound, respectively (Equation 3). Figure S3 shows the set sizes that we determined for *L*=4-7 and *d*=4 fulfill these inequalities.

$$\frac{4^L}{\sum_{k=0}^{d-1}\binom{L}{k}\times 3^k} \le M(L,d) \le 4^{L-d+1} \qquad (3)$$

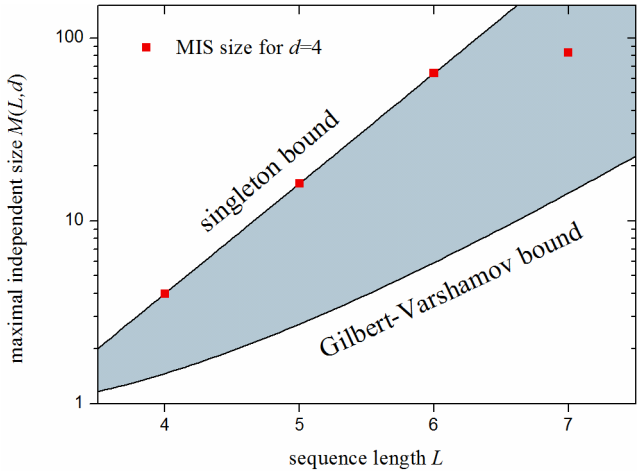

Figure S3. Maximal independent set size determined by our algorithm as a function of sequence length for d=4. For each L, $M(L,d)$ is within the bounds defined by coding theory.

S4 References

1. Provost, S. B.; Ratemi, W. M., Polynomial expansions via embedded Pascal's triangles. *Acta et Commentationes Universitatis Tartuensis de Mathematica* **2011,** *15* (1), 45-60.
2. Smith, C.; HOGGATT, V., STUDY OF THE MAXIMAL VALUES IN PASCALS QUADRINOMIAL TRIANGLE. *FIBONACCI QUARTERLY* **1979,** *17* (3), 264-269.
3. Tulpan, D. C.; Hoos, H. H.; Condon, A. E. In *Stochastic local search algorithms for DNA word design*, International Workshop on DNA-Based Computers, Springer: 2002; pp 229-241.
4. Li, M.; Lee, H. J.; Condon, A. E.; Corn, R. M., DNA word design strategy for creating sets of non-interacting oligonucleotides for DNA microarrays. *Langmuir* **2002,** *18* (3), 805-812.
5. Barg, A.; Guritman, S.; Simonis, J., Strengthening the Gilbert–Varshamov bound. *Linear Algebra Appl.* **2000,** *307* (1), 119-129.
6. El Rouayheb, S.; Georghiades, C. N., Graph theoretic methods in coding theory. In *Classical, Semi-classical and Quantum Noise*, Springer: 2012; pp 53-62.