Optimizing the computing of the oligonucleotide frequencies and the Genomic Signature Distance

To enable faster computation of the Genomic Signature Distance among DNA sequences, the following approaches were applied:

Optimization step 1: computing and storage of oligonucleotide occurrences

The occurrences of all oligonucleotides of length k in a sequence are computed using a sliding window of length k through the end of the sequence. It must be pointed out that, when comparing oligonucleotide frequencies between sequences, better results are obtained by comparing the frequencies in both DNA strands (data not shown).

Due to the complementarity of the two DNA strands, **occurrences of oligonucleotides in one strand may be used to compute the occurrences of oligonucleotides in both strands**. To illustrate this, dinucleotide occurrences are computed in table 1.

a) DNA sequence

Strand A: 5′-GACTCAGGCGTTTAGCCTGGAAGCCGCATCGCCTATCACC-3′
Strand B: 3′-CTGAGTCCGCAAATCGGACCTTCGGCGTAGCGGATAGTGG-5′

b) Dinucleotide	Strand		
	A	B	A+B
AA	1	2	3
AC	2	1	3
AG	3	3	6
AT	2	2	4
CA	3	1	4
CC	4	2	6
CG	3	3	6
CT	3	3	6
GA	2	3	5
GC	5	5	10
GG	2	4	6
GT	1	2	3
TA	2	2	4
TC	3	2	5
TG	1	3	4
TT	2	1	3

c) Ocurrence of dinucleotides in both DNA strands:

Type A	
AA/TT	3
AC/GT	3
AG/CT	6
CA/TG	4
CC/GG	6
GA/TC	5
Type B	
AT	4
CG	6
GC	10
TA	4

Table 1: Occurrences of oligonucleotides in one strand may be used to compute the occurrences of oligonucleotides in both strands, as shown in the example for dinucleotides. a) Example DNA sequences. b) Dinucleotide ocurrences in the DNA strands. c) Summarized dinucleotide occurrences discerning Type A and Type B oligonucleotides (see text); saves storage (to store dinucleotide occurrences only 10 numerical data are used instead of 16).

In the example, two types of dinucleotides are discerned:

- Type A dinucleotides: the dinucleotide and its reverse complement are different (for example, AA and TT). For dinucleotides, 12 dinucleotides are type A. The number of occurrences for this type of dinucleotides in both DNA strands (for example, AA or TT) is equal to the occurrences of the oligonucleotide and its reverse complement in one strand (occurrences of AA and TT in one strand).
- Type B dinucleotides: the dinucleotide and its reverse complement are identical (for example, AT). For dinucleotides, four dinucleotides are type B (dinucleotides AT, CG, GC, and TA). For type B dinucleotides, the number of occurrences in both strands is twice the number of occurrences in one strand.

This classification of oligonucleotides may be applied to longer oligonucleotides, and it may be used to **reduce the amount of information stored in databases**.

When this strategy is applied to longer oligonucleotides, the storage requirements are reduced according to table 2. This approach requires properly controlling the order of oligonucleotides in the database, but it also allows for fast computation of the Genomic Signature Distance (or other distances, such as Pearson's distance or Euclidean distance) as described below.

Oligonucleotide length (k)	All oligonucleotides (4^k)	Type A oligonucleotides	Type B oligonucleotides	Types A+B
2	16	6	4	10
3	64	32	-	32
4	256	120	16	136
5	1024	512	-	512
6	4096	2016	64	2080
7	16384	8192	-	8192
8	65536	32640	256	32896

Table 2: Storage requirements to save k-bases long oligonucleotide occurrences are shown in the last column when data from type A and type B oligonucleotides are discerned. When k is an odd number, no type B oligonucleotides exist.

Optimization step 2: fast computation of the Genomic Signature Distance

To optimize computation, two approaches were applied:

- precomputing the sum of all oligonucleotide occurrences, and
- adapting the Genomic Signature Distance formula to be used with type A and type B oligonucleotides.

The sum of all oligonucleotide occurrences in array X (Σf_{xi}) and array Y (Σf_{yi}) may be considered constants. Those values may be computed alongside the oligonucleotide occurrences, and they may be stored in a database. Consequently, a new formula can be defined for the Genomic Signature Distance:

$$d = \sum_{i=1}^{n} \left| \frac{fx_i}{\sum_{i=1}^{n} fx_i} - \frac{fy_i}{\sum_{i=1}^{n} fy_i} \right| \quad \rightarrow \quad d = \sum_{i=1}^{n} \left| \frac{fx_i}{sumX} - \frac{fy_i}{SumY} \right|$$

where sumX and sumY are precomputed constant values for Σf_{xi} and Σf_{yi} and are not computed each time they are required.

Additionally, the Genomic Signature Distance may be modified so that the type A and type B oligonucleotides described in optimization step1 are used separately:

$$d = 2\sum_{i=1}^{m} \left| \frac{fx_i}{sumX} - \frac{fy_i}{SumY} \right| - \sum_{i=1}^{p} \left| \frac{fx_i}{sumX} - \frac{fy_i}{SumY} \right|$$

where *m* is the number of type A oligonucleotides and *p* is the number of type B oligonucleotides.

For octanucleotides, *m* will correspond to 32,640 octanucleotides and p to 256. The first part of the formula is multiplied by two to include the occurrences of the reverse complement oligonucleotides that were not saved to the database (the missing values are identical to the saved ones). Consequently, to compute octanucleotide-based distances, 32,896 oligonucleotide occurrence pairs will be used in the formula, and this is an important saving compared to using 65,536 pairs of values while computing.