## Nucleotide Frequencies Matrices for TATA-box from various sets of plant promoters*

Nucleotide Frequencies Matrix for TATA-box from 345 experimentally verified plant promoters

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.147 | 0.162 | 0.269 | 0.139 | 0.009 | 0.971 | 0.009 | 0.988 | 0.630 | 0.968 | 0.361 | 0.699 | 0.145 | 0.312 | 0.286 | 0.329 |
| C | 0.358 | 0.384 | 0.292 | 0.607 | 0.000 | 0.000 | 0.014 | 0.000 | 0.012 | 0.000 | 0.038 | 0.072 | 0.402 | 0.410 | 0.298 | 0.286 |
| G | 0.116 | 0.165 | 0.168 | 0.081 | 0.003 | 0.000 | 0.003 | 0.003 | 0.003 | 0.012 | 0.020 | 0.101 | 0.303 | 0.153 | 0.173 | 0.197 |
| T | 0.379 | 0.289 | 0.272 | 0.173 | 0.988 | 0.029 | 0.974 | 0.009 | 0.355 | 0.020 | 0.581 | 0.127 | 0.150 | 0.124 | 0.243 | 0.188 |
| | y | y | n | c | T | A | T | A | W | A | W | A | s | m | n | n |

Nucleotide Frequencies Matrix for TATA-box from 256 experimentally verified dicot plant promoters

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.172 | 0.172 | 0.272 | 0.152 | 0.020 | 0.972 | 0.004 | 0.984 | 0.604 | 0.960 | 0.384 | 0.748 | 0.180 | 0.356 | 0.288 | 0.352 |
| C | 0.324 | 0.368 | 0.296 | 0.560 | 0.004 | 0.000 | 0.016 | 0.000 | 0.012 | 0.000 | 0.044 | 0.068 | 0.340 | 0.384 | 0.260 | 0.284 |
| G | 0.120 | 0.136 | 0.120 | 0.080 | 0.004 | 0.000 | 0.000 | 0.004 | 0.004 | 0.016 | 0.012 | 0.072 | 0.300 | 0.112 | 0.184 | 0.152 |
| T | 0.384 | 0.324 | 0.312 | 0.208 | 0.972 | 0.028 | 0.980 | 0.012 | 0.380 | 0.024 | 0.560 | 0.112 | 0.180 | 0.148 | 0.268 | 0.212 |
| | y | y | n | c | T | A | T | A | W | A | W | A | n | m | n | n |

Nucleotide Frequencies Matrix for TATA-box from 84 experimentally verified monocot plant promoters

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.083 | 0.143 | 0.226 | 0.095 | 0.000 | 0.964 | 0.000 | 1.000 | 0.702 | 0.988 | 0.298 | 0.643 | 0.071 | 0.226 | 0.238 | 0.262 |
| C | 0.429 | 0.429 | 0.310 | 0.750 | 0.000 | 0.000 | 0.012 | 0.000 | 0.012 | 0.000 | 0.012 | 0.095 | 0.524 | 0.452 | 0.429 | 0.333 |
| G | 0.119 | 0.262 | 0.274 | 0.071 | 0.000 | 0.000 | 0.012 | 0.000 | 0.000 | 0.000 | 0.060 | 0.167 | 0.333 | 0.238 | 0.167 | 0.321 |
| T | 0.369 | 0.167 | 0.190 | 0.083 | 1.000 | 0.036 | 0.976 | 0.000 | 0.286 | 0.012 | 0.631 | 0.095 | 0.071 | 0.083 | 0.167 | 0.083 |
| | y | y | n | C | T | A | T | A | W | A | W | a | y | v | n | v |

* Positions 5-12 correspond to the core-motif of TATA box. The mean distance between TATA box and TSS is 26 bp.

The following IUPAC codes for nucleotide abbreviations are used: A – Adenine, C – Cytosine, G – Guanine, T – Thymine, Y – C or T, S – G or C, W – A or T, M – A or C, V – A or C or G, N – any base (http://www.bioinformatics.org/sms/iupac.html).

To get consensus sequences, the following thresholds were applied:

- Single nucleotide in lower case – if its frequency >0.5, but <0.65;
- Single nucleotide in upper case – if its frequency ≥0.65;
- Combination of two nucleotides in lower case – if a sum of their frequencies ≥0.65, but <0.9;
- Combination of two nucleotides in upper case – if a sum of their frequencies >0.9;
- Combination of three nucleotides in lower case – if a sum of their frequencies >0.85;
- n - other cases.