Alejandra Eugenia Medina Rivera
RegulonDB database
This presentation aims to show the user the methodology followed for the evaluation of RegulonDB matrices. It presents the statistic and concepts involved.

**Summary:**
- Building model to predict TFBSs
- Matrix-quality program.
- Matrix Score distribution
- Evaluation criteria
A Annotated TrpR binding sites

<table>
<thead>
<tr>
<th>Site ID</th>
<th>Nucleotides</th>
<th>Target Operon</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECK120012844</td>
<td>G T A C T T A G T T T G A T G T A T G</td>
<td>aroL-ycrA-aroM</td>
</tr>
<tr>
<td>ECK120012187</td>
<td>G T A C T T A G T T T G A T G T A T G</td>
<td>aroL-ycrA-aroM</td>
</tr>
<tr>
<td>ECK120012179</td>
<td>G A A C T T A G T T A A C T A G T A C G</td>
<td>trpLEDCA</td>
</tr>
<tr>
<td>ECK120012892</td>
<td>G A A C T T A G T T A A C T A G T A C G</td>
<td>trpLEDCA</td>
</tr>
<tr>
<td>ECK120012181</td>
<td>G A A C T T A G T T A A C T A G T A C G</td>
<td>trpLEDCA</td>
</tr>
<tr>
<td>ECK120012636</td>
<td>G T A C T G A G A A C T A G T G C A</td>
<td>trpLEDCA</td>
</tr>
<tr>
<td>ECK120012183</td>
<td>G T A C T G A G A A C T A G T G C A</td>
<td>trpLEDCA</td>
</tr>
<tr>
<td>ECK120012185</td>
<td>G T A C T G A G A A C T A G T G C A</td>
<td>trpLEDCA</td>
</tr>
<tr>
<td>ECK120012979</td>
<td>G T A C T C C G T G T A C T G T A C A</td>
<td>mtr</td>
</tr>
<tr>
<td>ECK120012894</td>
<td>G T A C T C C G T G T A C T G T A C A</td>
<td>mtr</td>
</tr>
</tbody>
</table>

B Position specific scoring matrix

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
</tbody>
</table>

C Consensus

G w A C T m G t k w r C t r G T r C r

D Sequence logo

- Sequence logo showing the conservation of nucleotides at different positions along the sequence.
Pattern-matching: scanning sequences for putative TFBSs

Position in \textit{trpR} upstream sequence (0 corresponds to TSS)
Matrix quality

- Is the matrix **good to predict** new putative binding sites?
- Which is the **sensitivity** to recover true binding sites?
- Which is the **false positive** rate for a given **sensitivity**?

---

1 Bailey and Elkan. Systems for Molecular Biology (1994)
Combines theoretical and empirical information to evaluate the quality of matrices, with a series of graphs.
Which are all the possible scores that could be generated by a PSSM?

Score distribution: Theoretical distribution

The 'theoretical distribution' provides an estimate of the expected FPR at each possible weight score (WS), based on the prior choice of a relevant background model.

ATATACGTATCTACTACTG = 3.25

Extended to higher markov models: matrix-distrib (RSAT)
Score distribution: Theoretical distribution

Theoretical probability of scores

- E-value = P-value * number of tested positions
- E.G. When scanning all *E. coli* K12 upstream regions (L=603945 bps)
  - Pval = P(weight >= 10) = 2.7e-6 (purple line)
  - Eval = Pval * L * 2
    - = 2.7e-6 FP/bps * 603945 bps * 2 ~ 3 expected False Positives

The ‘empirical score distribution in all upstream non-coding sequences’ of the organism of interest. These sequences are essentially composed of non-binding sites (the non-coding genomic background), interspersed with a few functional binding sites. The empirical distribution typically fits the theoretical distribution for small WS values (the background), but separates at high WS values, most likely corresponding to functional TF-binding sites.
Empirical score distribution in all upstream non-coding sequences

The separation between the right tails of the empirical and theoretical distributions indicates the capability of the matrix to identify a set of high-scoring putative binding sites in the collection of promoters.

Observed Sites with Score 15 = 7 sites

Expected Sites with Score 15 = 0.04 sites

-50 -40 -30 -20 -10 0 10 20 30
matrix score

Expected Sites with Score 15 = 0.04 sites

-30 -20 -10 0 100 200 300
Upstream position

Weight Score

yaaJ
An empirical estimate of the FPR is obtained by scanning all upstream non-coding sequences with column-permuted matrices, which supposedly do not correspond to any TF in the organism under consideration. If the background model has been chosen correctly, the ‘empirical distribution of the permuted matrices’ should fit the theoretical distribution.
The ‘empirical score distribution in the annotated binding sites’ indicates the sensitivity of the matrix, i.e. its capability to recover binding sites above a given WS threshold.

The empirical score distribution in the upstream regions.

Matrix sites

<table>
<thead>
<tr>
<th>Matrix sites</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAACTAGTTAACTAGTACG</td>
<td>19.3</td>
</tr>
<tr>
<td>GTACTCTTTTAGCGAGTACA</td>
<td>14.9</td>
</tr>
</tbody>
</table>

Theoretical score distribution

Negative control: Permuted matrix.
Empirical score distribution in the annotated binding sites
Leave-One-Out (LOO) Test

Empirical score distribution in upstream regions

Theoretical score distribution

Negative control: Permutated matrix.

Matrices are rebuilt and annotated sites are scored using a LOO procedure to reduce over-fitting biases when estimating the capability to detect novel sites.

GAACTAGTTAACTAGTACG  15.1
GTACTCTTTTAGCGAGTACA  9.8
• Receiver Operating Characteristic (ROC) curves’ are drawn to indicate the tradeoff between sensitivity and False Positive Rate (FPR). These curves provide a direct way to estimate the expected cost (in terms of false positives) for achieving a desired sensitivity, or, reciprocally, the sensitivity that can be expected for a given FPR.
matrix-quality in RegulonDB

Evaluation criteria

- Matrices with information.
- Low FPR.
- Detects sites in the genome.
- LOO ROC is not separated by orders of magnitude from the matrix-sites ROC.

- Matrices with poor information.
- High FPR.
- Does not detect sites in the genome.
- LOO ROC is separated by orders of magnitude from the matrix-sites ROC.