Metrics for comparing regulatory sequences on the basis of pattern counts

Jacques van Helden

SCMBB, Université Libre de Bruxelles, Campus Plaine CP 263, Boulevard du Triomphe, B-1050 Bruxelles, Belgium

Received on June 18, 2002; revised on October 17, 2002; accepted on August 6, 2003

ABSTRACT
Motivation: Upstream sequences contain short motifs, which mediate transcriptional regulation by specifically binding different transcription factors. The presence of common motifs in the regulatory regions of two genes might be considered as a clue for a potential co-regulation. A pattern count-based (dis)similarity metric between sequences could thus be used to classify genes according to their putative regulatory properties.

Results: We present here several metrics which rely on probability theory, and which aim at comparing sequences on the basis of pattern counts. We compare these metrics to several classical dissimilarity and similarity metrics, and illustrate their behaviour with a biological example.

Contact: jvanheld@ucmb.ulb.ac.be

Supplementary information: The data, results, and R routines used in this paper are freely available at http://rsat.ulb.ac.be/rsat/published_data/pattern_count_metrics_2003/

INTRODUCTION

Regulatory regions are characterized by the presence of short motifs involved in the sequence-specific binding of transcription factors. Several methods have been developed to detect putative regulatory motifs in regulatory regions, on the basis of pattern counts. These string-based approaches have been used mainly in a context of pattern discovery: starting from a set of co-expressed genes, it is possible to detect either oligonucleotides (van Helden et al., 1998), or spaced dyads (van Helden et al., 2000b), which are significantly over-represented in their upstream region. String-based pattern discovery programs (oligo-analysis and dyad-analysis) return a set of significant patterns, which are often related by a strong mutual overlap, and can be assembled to form larger or partially degenerated motifs (van Helden, 2003). It has been shown with yeast regulons that this approach is efficient for detecting the transcription factor binding sites (van Helden et al., 1998, 2000b).

Significant patterns can then be used in a context of pattern matching, to predict the putative positions of cis-acting elements. Each single pattern is poorly informative by itself. Indeed, given the short size of the binding sites (5–10 conserved residues) and the length of the regions (~800 bp in yeast), the presence of a single pattern could be due to chance, and is not sufficient to predict a regulatory response. However, each region usually contains multiple cis-acting elements, reflecting the combinatorial and synergic aspects of transcriptional regulation: it is common to find multiple-binding sites for the same transcription factor in an upstream region, and, in addition, a gene is generally regulated by various factors (Jones et al., 1992).

Upstream sequences can thus be characterized by counts of pattern occurrences, which can then be used to calculate pairwise similarities between sequences, in order to infer possible co-regulations. This requires to choose a metric which takes into account the following aspects: (i) count-based comparison (the number of copies of each pattern should be reflected); (ii) multi-variate comparison (several distinct patterns are considered) and (iii) pattern-specific prior probabilities (some patterns are expected to occur by chance more frequently than others).

Park et al. (2002) recently proposed a heuristics to measure pairwise dissimilarities between sequences on the basis of pattern counts. Their metrics is based on the optimization of two parameters, on the basis of a training set (e.g. binding sites from a transcription factor database). This approach is however not always possible, since training sets are not available for all organisms, and, when they are, the number of entries is generally too weak to obtain a robust estimation of the parameters.

We present here several similarity and dissimilarity metrics, which are based on simple concepts of probability theory, and fulfill the requirements listed above. These metrics do not rely on any prior optimization. We compare these metrics with some classical dissimilarity and similarity metrics on the basis of a random dataset, and we illustrate their potential application with a biological example.

SYSTEMS AND METHODS

Data description
The data set is a matrix ($N$), containing $n$ rows (one per sequence) and $p$ columns (one per pattern). $N_{ij}$ is the number of occurrences of pattern $i$ in sequence $a$. 

Published by Oxford University Press.
Fig. 1. Illustration of the Poisson-based metrics for a single variable. (A) The similarity metric is based on the probability of common occurrences, i.e. the square of the surface marked with shaded bars. The first dissimilarity metric \( D_{\text{poisson.distinct}} \) is the difference between the left tails of curves A and B (plain bars). (B) The second dissimilarity metrics \( D_{\text{poisson.over}} \) is based on the difference between the right tails of the curves A and B (plain bars).

**Probabilistic model**

In order to define the dissimilarity between two genes \((a\) and \(b)\), the simple difference between the occurrences \(|N_i^a - N_i^b|\) is not satisfying, because (i) it assigns the same weight to each pattern, despite the fact that some have a higher prior probability than others and (ii) it does not take into account the fact that, for a given pattern, the probability is not a linear function of the number of occurrences. These two aspects can be treated by using a probabilistic model of pattern occurrences. The most immediate model is the binomial distribution, which, in our working conditions, can be approximated by a Poisson distribution (Fig. 1).

Each pattern \(i\) is characterized by a prior probability \(f_i\), indicating the probability to find an occurrence at any position of a sequence. Prior probabilities can be calculated either on the basis of the data set itself, or on the basis of an external background model (e.g. the set of all upstream sequences of a genome, or a transcription factor database), depending on the underlying model. The expected number of occurrences \(m_i\) is obtained by multiplying the prior probability \(f_i\) by the number of possible positions \((T)\) for that pattern:

\[
m_i = f_i T = f_i(L - w + 1),
\]

where \(L\) is the length of the sequence and \(w\) the length of the pattern. For the sake of simplicity, we consider that all the sequences have the same length. When this is not the case, the formula can be adapted by using a sequence-specific estimate of expected occurrences (i.e. replacing \(L\) by \(L_a, L_b, \ldots\)).

The Poisson distribution function \(F(x, m)\) gives the probability to observe \(\leq x\) occurrences when the expected value is \(m\). Thus, for a single gene \(a\) and a single pattern \(i\), the probability to observe at least \(N_i^a\) occurrences is obtained by

\[
P(x \geq N_i^a) = \begin{cases} 1 - F(N_i^a - 1, m_i) & \text{if } N_i^a > 0, \\ 1 & \text{otherwise.} \end{cases}
\]

Low values of \(P(x \geq N_i^a)\) correspond to over-represented patterns.

**Similarity metrics**

**Principle** A univariate (single pattern) similarity is calculated on the basis of the probability of common occurrences (Fig. 1). Single-pattern probabilities are then combined to obtain a multi-variate similarity metric.

**Calculation** For a pair of genes \(a\) and \(b\), let \(C_{ab}^i\) be the \(p\)-element vector containing the common counts for each pattern. \(C_{ab}^i\) is the number of common counts for pattern \(i\), i.e. the number of occurrences found in both sequences.

\[
C_{ab}^i = \min(N_i^a, N_i^b).
\]

The probability to observe at least \(C_{ab}^i\) common occurrences is the joint probability of observing at least \(C_{ab}^i\) occurrences of pattern \(i\) in the upstream sequence of each gene. Under the assumption of independence, this is obtained by the product of the probabilities:

\[
P(x \geq C_{ab}^i) = \left\{ \begin{array}{ll} [1 - F(C_{ab}^i - 1, m_i)]^2 & \text{if } C_{ab}^i > 0, \\ 1 & \text{otherwise.} \end{array} \right.
\]

This probability can be converted into a similarity metric as follows:

\[
s_{ab}^i = 1 - P(x \geq C_{ab}^i),
\]

\(s_{ab}^i\) indicates how exceptional it is to find at least \(C_{ab}^i\) common occurrences of pattern \(i\) in a pair of sequences.

Classically, a multi-variate similarity is obtained by averaging the single-variate similarities (Gordon, 1999), assuming
an additive effect of the variables. The multi-variate similarity between sequences \( a \) and \( b \) is thus obtained by

\[
s_{\text{add}}^{ab} = \frac{1}{p} \sum_{i=1}^{p} s_{i}^{ab}.
\]  

(1)

Alternatively, one could consider the joint probability of observing simultaneously \( C_{1}^{ab} \) common occurrences of pattern 1, \( C_{2}^{ab} \) common occurrences of pattern 2, \ldots and \( C_{p}^{ab} \) common occurrences of pattern \( p \), which is obtained by the product of probabilities for each independent pattern. We take the \( p \)th root of this product, which is the geometric mean of the single-variable probabilities. Small values of the joint probability reflect sequences sharing a significant number of common motifs. The similarity is obtained by substracting the joint probability from Equation (1)

\[
s_{\text{prod}}^{ab} = 1 - \sqrt[p]{\prod_{i=1}^{p} P(x \geq C_{i}^{ab})}.
\]  

(2)

Properties of the Poisson-based similarity metrics The two metrics presented in formulae (1) and (2) fit well with the intuitive concept of similarity, in terms of pattern occurrences:

1. If two sequences do not have a single common site, their similarity is 0.
2. The score increases when multiple copies of a given pattern are found in both sequences.
3. The score increases when several patterns are common to both sequences.
4. Patterns with low prior probabilities contribute more than those with high probabilities.

However, these metrics are based on the counts of common occurrences only, and do not reflect the differences, since occurrences found in gene \( a \) but not in gene \( b \) do not affect the score. Such differences are taken into account by the dissimilarity metrics proposed in the next section.

Dissimilarity metrics

Principle A univariate (single pattern-based) dissimilarity \( d_{i}^{ab} \) between sequences \( a \) and \( b \) is estimated as the difference between the \( P \)-values of occurrences found in the respective sequences (Fig. 1A). The multivariate dissimilarity \( D_{i}^{ab} \) is obtained by averaging the single-pattern dissimilarities.

Calculation The simplest way to estimate the dissimilarity is to take the absolute value of the difference between occurrences observed in the two sequences, \( |N_{i}^{a} - N_{i}^{b}| \). This dissimilarity is however not satisfying, for the reasons mentioned above.

A more appropriate metric can be obtained by calculating the probability of the distinct occurrences, i.e. those found in one gene but not the other one (Fig. 1A). If genes \( a \) and \( b \) have \( N_{i}^{a} \) and \( N_{i}^{b} \) occurrences of pattern \( i \) respectively, with \( N_{i}^{a} \leq N_{i}^{b} \), the dissimilarity \( d_{i}^{ab} \) can be estimated as the area comprised below the Poisson curve, within the interval \( [N_{i}^{a}, N_{i}^{b}] \), which is obtained by

\[
d_{i}^{ab} = |F(N_{i}^{b} - m_{i}) - F(N_{i}^{a} - m_{i})|, \quad D_{i}^{ab} = \frac{1}{p} \sum_{i=1}^{p} d_{i}^{ab}.
\]  

(3)

However, in the analysis of regulatory sequences, each pattern is better characterized by its degree of over-representation (Fig. 1B), which is indicated by low values of the probability to observe at least \( x \) occurrences

\[
P(x \geq N_{i}^{a}) = 1 - F(N_{i}^{a} - 1, m_{i}).
\]

The ‘over-representation’ dissimilarity can be calculated on this basis

\[
d_{i}^{over} = |P(x \geq N_{i}^{a}) - P(x \geq N_{i}^{b})| = |F(N_{i}^{b} - 1, m_{i}) - F(N_{i}^{a} - 1, m_{i})|
\]  

(4)

Properties of the Poisson-based dissimilarity \( D_{\text{distinct}} \) and \( D_{\text{over}} \) The metrics presented in formulae (3) and (4) fit with the intuitive concept of dissimilarity between two sequences, in terms of pattern counts:

1. When two sequences have exactly the same counts for all the patterns, their dissimilarity is 0.
2. For each pattern, the score increases with the number of distinct counts.
3. The score increases when differences are observed for multiple patterns.

However, both metrics fail to indicate the significance of the similarity between the two sequences. Thus, if two sequences have exactly the same pattern counts, their dissimilarity will be 0, irrespective of the fact that they contain 100, 5 or 0 common patterns, although these events have very different probabilities.

Mixed metric combining similarity and dissimilarity Ideally, one would like to use a metric which would emphasize the common patterns (similarity), but also include a correction for additional counts found in one element of the pair only (dissimilarity). This can be done by substracting the Poisson-based dissimilarity from the Poisson-based similarity

\[
M^{ab} = S^{ab} - \alpha D^{ab} + \beta,
\]  

(5)

where \( S^{ab} \) is the similarity, calculated with either formula (1) or formula (2), and \( D^{ab} \) is the dissimilarity, calculated with...
either formula (3) or formula (4). \( \alpha \) is a positive weighting parameter, which can be tuned arbitrarily to give more emphasis on the common (low \( \alpha \) values) or distinct (high \( \alpha \) values) occurrences between the two genes, and \( \beta \) is an offset which ensures that the metric is always positive.

**Properties of the mixed measure** The mixed measure has basic properties of a similarity corrected for dissimilarities.

1. Patterns found in both sequences contribute positively.
2. Patterns found in one sequence but not the other one contribute negatively.
3. A null value indicates either that none of the sequences contains any occurrence of any pattern, or that common and distinct occurrences compensate each other.

A drawback of this measure is that the \( \alpha \) and \( \beta \) parameters have to be tuned arbitrarily, and their choice might drastically affect the result. Optimal values can be estimated on the basis of a training set, as has been done by Park et al. (2002) with their heuristics. However, one does not always dispose of an independent training set. Thus, in order to make our purpose general and to be able to use the metrics in an unsupervised context, the analysis was limited to the most trivial estimates: \( \alpha = 1 \) and \( \beta = -\min(S^{ab}, -D^{ab}) \).

**RESULTS**

We compared the different metrics proposed here with some classical distance and similarity metrics, and with the heuristics proposed by Park et al. (2002), on the basis of two data sets.

The first data set contains pattern counts from randomly generated sequences, and reveals the intrinsic differences between metrics, in the absence of any biological consideration. The second data set contains counts of regulatory patterns in upstream sequences of *Saccharomyces cerevisiae* genes, and illustrates the capability of the different metrics to classify genes according to their biological function.

**Metric comparison on the basis of a random data set**

Fifty random sequences were generated with a Markov chain model of order 5, calibrated to obtain sequences with hexanucleotide frequencies similar to those observed in yeast intergenic sequences (van Helden, 2003). A set of 50 hexanucleotides was randomly selected, and their occurrences were counted in each of the random sequences.

The dissimilarity or similarity between each pair of sequences was calculated with (i) the Poisson-based metrics described above, (ii) a few classical metrics (Gordon, 1999) and (iii) Park’s heuristic (Park et al., 2002).

Table 1 shows the coefficients of correlation between the different metrics. There is a striking correlation (−0.98) between Park’s heuristics and the Poisson mixed ‘over’ metric, which suggests that Park’s heuristic optimization tends to fit the theoretical Poisson distribution. Poisson-based dissimilarities (‘over’ and ‘distinct’) are mutually correlated (0.93), and, to a lesser extent, show some correlation with Manhattan distance.

An interesting difference between Poisson-based similarity metrics and other metrics is the treatment of identity: with classical metrics, identical sequences always have a correlation of 1 (by definition), and distances of 0 (Euclidian, Manhattan, Mahalanobis). On the contrary, Poisson-based similarity metrics, as well as Park’s heuristics, give different scores reflecting the significance of the common patterns.

**Biological application: classification of MET, PHO and NIT genes**

We illustrate the biological relevance of the Poisson-based metrics with a biological example. We selected three sets of genes, each involved in a particular pathway: nitrogen metabolism (NIT family, 31 genes), phosphate utilization (PHO family, 13 genes) and methionine biosynthesis (MET family, 20 genes). We retrieved the 800 bp sequence upstream the start codon of each gene. As negative control, we generated 30 random sequences with a Markov chain model of order 5, calibrated with hexanucleotide frequencies observed in the whole set of yeast intergenic sequences.

Putative regulatory patterns were detected in the upstream sequences of the MET, NIT and PHO families respectively, with the programs oligo-analysis (van Helden et al., 1998) and dyad-analysis (van Helden et al., 2000b). Occurrences of the 44 resulting patterns were counted within each sequence with *dna-pattern* (van Helden et al., 2000a). This returned a multivariate table with 94 rows (one per sequence) and 44 columns (one per pattern), which was used to calculate, for each of the metrics discussed here, a matrix of pairwise dissimilarities between sequences. Similarity metrics were converted into dissimilarities by applying the formula \( d = \max(s) - s \) (Gordon, 1999), which ensures all values to be positive. Dissimilarity matrices were then used as input for hierarchical clustering, using single, average or complete linkage, and Ward clustering respectively (Gordon, 1999). In order to evaluate the correctness of clustering, the tree was pruned at the level of four branches, and each of branch was assigned to one prior class: RAND, NIT, MET or PHO (Fig. 2). The prior and assigned class were compared for each gene. A confusion table was then calculated by comparing the assigned cluster and the known class for each gene (Table 2), and the hit rate was calculated.

As shown in Table 2, the choice of the metrics and clustering method drastically affects the hit rate. Although some correct grouping is observed within each tree (supplementary material), the Poisson-based metrics are clearly more reflective of the biological properties of the sequences. The best hit rate (76.6%) is obtained with \( S_{\text{product}}, \) [Formula (2)]. The combination of any dissimilarity metric with this similarity metric did not bring any improvement over the similarity alone.
### Table 1. Correlation between the different similarity and dissimilarity metrics (50 random sequences)

<table>
<thead>
<tr>
<th></th>
<th>correlation coefficient</th>
<th>canberra.distance</th>
<th>binary.distance</th>
<th>poisson.mixed.over</th>
<th>park.similarity</th>
<th>poisson.similarity</th>
<th>poisson.similarity.product</th>
<th>poisson.mixed.distinct.product</th>
<th>poisson.mixed.over.product</th>
<th>poisson.dissimilarity.distinct</th>
<th>poisson.dissimilarity.over</th>
<th>manhattan.distance</th>
<th>euclidian.distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>correlation.coefficient</td>
<td>1.00</td>
<td>−0.76</td>
<td>−0.81</td>
<td>−0.80</td>
<td>0.79</td>
<td>0.72</td>
<td>−0.71</td>
<td>0.60</td>
<td>0.28</td>
<td>−0.72</td>
<td>−0.67</td>
<td>−0.56</td>
<td>−0.56</td>
</tr>
<tr>
<td>canberra.distance</td>
<td>−0.76</td>
<td>1.00</td>
<td>0.98</td>
<td>0.91</td>
<td>−0.89</td>
<td>−0.89</td>
<td>−0.88</td>
<td>0.84</td>
<td>−0.73</td>
<td>−0.38</td>
<td>0.75</td>
<td>0.72</td>
<td>0.57</td>
</tr>
<tr>
<td>binary.distance</td>
<td>−0.81</td>
<td>0.98</td>
<td>1.00</td>
<td>0.91</td>
<td>1.00</td>
<td>−0.98</td>
<td>−0.76</td>
<td>0.74</td>
<td>−0.58</td>
<td>−0.12</td>
<td>0.89</td>
<td>0.92</td>
<td>0.78</td>
</tr>
<tr>
<td>poisson.mixed.over</td>
<td>−0.80</td>
<td>0.91</td>
<td>0.91</td>
<td>1.00</td>
<td>−0.98</td>
<td>−0.76</td>
<td>−0.74</td>
<td>0.60</td>
<td>0.60</td>
<td>−0.85</td>
<td>−0.89</td>
<td>−0.79</td>
<td>−0.72</td>
</tr>
<tr>
<td>park.similarity</td>
<td>0.79</td>
<td>−0.89</td>
<td>−0.89</td>
<td>−0.98</td>
<td>1.00</td>
<td>0.76</td>
<td>−0.74</td>
<td>0.60</td>
<td>0.15</td>
<td>−0.85</td>
<td>−0.89</td>
<td>−0.79</td>
<td>−0.72</td>
</tr>
<tr>
<td>poisson.similarity</td>
<td>0.72</td>
<td>−0.85</td>
<td>−0.88</td>
<td>−0.76</td>
<td>0.76</td>
<td>1.00</td>
<td>−0.96</td>
<td>0.92</td>
<td>0.71</td>
<td>−0.50</td>
<td>−0.45</td>
<td>−0.29</td>
<td>−0.29</td>
</tr>
<tr>
<td>poisson.similarity.product</td>
<td>−0.71</td>
<td>0.82</td>
<td>0.84</td>
<td>0.74</td>
<td>−0.74</td>
<td>−0.96</td>
<td>1.00</td>
<td>−0.98</td>
<td>0.76</td>
<td>0.46</td>
<td>0.43</td>
<td>0.28</td>
<td>0.30</td>
</tr>
<tr>
<td>poisson.mixed.distinct.product</td>
<td>0.60</td>
<td>−0.71</td>
<td>−0.73</td>
<td>−0.58</td>
<td>0.60</td>
<td>0.92</td>
<td>−0.98</td>
<td>1.00</td>
<td>0.87</td>
<td>−0.25</td>
<td>−0.24</td>
<td>−0.08</td>
<td>−0.13</td>
</tr>
<tr>
<td>poisson.mixed.over.product</td>
<td>0.28</td>
<td>−0.35</td>
<td>−0.38</td>
<td>−0.12</td>
<td>0.15</td>
<td>0.71</td>
<td>−0.76</td>
<td>0.87</td>
<td>1.00</td>
<td>0.18</td>
<td>0.26</td>
<td>0.26</td>
<td>0.24</td>
</tr>
<tr>
<td>poisson.dissimilarity.distinct</td>
<td>−0.72</td>
<td>0.76</td>
<td>0.75</td>
<td>0.89</td>
<td>−0.85</td>
<td>−0.50</td>
<td>0.46</td>
<td>−0.25</td>
<td>0.18</td>
<td>1.00</td>
<td>0.93</td>
<td>0.89</td>
<td>0.78</td>
</tr>
<tr>
<td>poisson.dissimilarity.over</td>
<td>−0.67</td>
<td>0.74</td>
<td>0.72</td>
<td>0.92</td>
<td>−0.89</td>
<td>−0.45</td>
<td>0.43</td>
<td>−0.24</td>
<td>0.26</td>
<td>0.93</td>
<td>1.00</td>
<td>0.91</td>
<td>0.76</td>
</tr>
<tr>
<td>manhattan.distance</td>
<td>−0.56</td>
<td>0.62</td>
<td>0.57</td>
<td>0.78</td>
<td>−0.79</td>
<td>−0.29</td>
<td>0.28</td>
<td>−0.08</td>
<td>0.36</td>
<td>0.89</td>
<td>0.91</td>
<td>1.00</td>
<td>0.94</td>
</tr>
<tr>
<td>euclidian.distance</td>
<td>−0.56</td>
<td>0.60</td>
<td>0.54</td>
<td>0.68</td>
<td>−0.72</td>
<td>−0.29</td>
<td>0.30</td>
<td>−0.13</td>
<td>0.24</td>
<td>0.78</td>
<td>0.76</td>
<td>0.94</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Table 2. Confusion table and hit rate obtained when MET, PHO, NIT and RAND sequences are classified with different clustering methods and different metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Model</th>
<th>Clustering method</th>
<th>MET</th>
<th>NIT</th>
<th>PHO</th>
<th>RAND</th>
<th>MET</th>
<th>NIT</th>
<th>PHO</th>
<th>RAND</th>
<th>MET</th>
<th>NIT</th>
<th>PHO</th>
<th>RAND</th>
<th>External True</th>
<th>False</th>
<th>Hit rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>poisson.similarity product</td>
<td>Product</td>
<td>Ward</td>
<td>1</td>
<td>22</td>
<td>0</td>
<td>5</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>9</td>
<td>4</td>
<td>24</td>
</tr>
<tr>
<td>poisson.mixed.distinct</td>
<td>Additive</td>
<td>Complete</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>10</td>
<td>6</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>poisson.mixed.over</td>
<td>Additive</td>
<td>Ward</td>
<td>3</td>
<td>21</td>
<td>1</td>
<td>7</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>poisson.mixed.distinct.product</td>
<td>Additive</td>
<td>Ward</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>5</td>
<td>16</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>poisson.mixed.distinct.product</td>
<td>Additive</td>
<td>Complete</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>3</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>poisson.similarity</td>
<td>Additive</td>
<td>Ward</td>
<td>4</td>
<td>25</td>
<td>2</td>
<td>14</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>poisson.similarity product</td>
<td>Additive</td>
<td>Ward</td>
<td>4</td>
<td>22</td>
<td>2</td>
<td>12</td>
<td>15</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>poisson.mixed.over.product</td>
<td>Additive</td>
<td>Ward</td>
<td>3</td>
<td>23</td>
<td>1</td>
<td>9</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>9</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>correlation.coefficient</td>
<td>Additive</td>
<td>Ward</td>
<td>1</td>
<td>26</td>
<td>1</td>
<td>8</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>poisson.similarity.over</td>
<td>Additive</td>
<td>Ward</td>
<td>9</td>
<td>16</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>poisson.similarity</td>
<td>Additive</td>
<td>Complete</td>
<td>13</td>
<td>25</td>
<td>2</td>
<td>13</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>correlation.coefficient</td>
<td>Additive</td>
<td>Complete</td>
<td>0</td>
<td>26</td>
<td>3</td>
<td>11</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>5</td>
<td>10</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>poisson.similarity.product</td>
<td>Product</td>
<td>Complete</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>10</td>
<td>1</td>
<td>9</td>
<td>6</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>manhattan.dist</td>
<td>Additive</td>
<td>Ward</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>13</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>0</td>
<td>7</td>
<td>2</td>
<td>8</td>
<td>3</td>
<td>26</td>
</tr>
<tr>
<td>park.similarity (pruning 5)</td>
<td>Additive</td>
<td>Ward</td>
<td>4</td>
<td>14</td>
<td>2</td>
<td>9</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>poisson.similarity.over</td>
<td>Additive</td>
<td>Complete</td>
<td>5</td>
<td>31</td>
<td>4</td>
<td>29</td>
<td>13</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>poisson.mixed.distinct.product</td>
<td>Product</td>
<td>Ward</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>14</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>5</td>
<td>7</td>
<td>2</td>
<td>14</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>poisson.similarity.product</td>
<td>Additive</td>
<td>Complete</td>
<td>5</td>
<td>31</td>
<td>4</td>
<td>30</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>poisson.mixed.over</td>
<td>Additive</td>
<td>Complete</td>
<td>3</td>
<td>28</td>
<td>3</td>
<td>26</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>park.similarity</td>
<td>Additive</td>
<td>Ward</td>
<td>5</td>
<td>29</td>
<td>4</td>
<td>28</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>euclidian.dist</td>
<td>Additive</td>
<td>Ward</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>16</td>
<td>3</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>mahalanobis.dist</td>
<td>Additive</td>
<td>Complete</td>
<td>6</td>
<td>16</td>
<td>5</td>
<td>12</td>
<td>9</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>manhattan.dist</td>
<td>Additive</td>
<td>Complete</td>
<td>6</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>12</td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td>euclidian.dist</td>
<td>Additive</td>
<td>Complete</td>
<td>0</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>14</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>mahalanobis.dist</td>
<td>Additive</td>
<td>Ward</td>
<td>6</td>
<td>15</td>
<td>5</td>
<td>9</td>
<td>11</td>
<td>11</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>park.similarity.d</td>
<td>Additive</td>
<td>Complete</td>
<td>15</td>
<td>30</td>
<td>4</td>
<td>30</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>poisson.mixed.over.product.d</td>
<td>Product</td>
<td>Complete</td>
<td>5</td>
<td>27</td>
<td>9</td>
<td>24</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Pattern count based sequence comparison

Fig. 2. Hierarchical clustering (ward algorithm) of MET, PHO, NIT and RAND sequences on the basis of pattern counts, using \( S_{\text{prod}} \) [Formula (2)] as similarity metrics. The prior and assigned classes of each gene are compared at the bottom of the tree (empty circle: identical, filled circle: different).

On the contrary, with the additive model, mixed metrics provide better results than similarity or dissimilarity alone, respectively. Among the mixed metrics, the best results (with a hit rate of 72.3%) were obtained by complete linkage on a dissimilarity metrics calculated with the mixed Poisson-based metrics. Other Poisson-based metrics returned slightly weaker results, but clearly supersede Euclidian, Manhattan and Mahalanobis distances.

There is a clear dependency between dissimilarity metrics and clustering method. For example, With Ward clustering, \( M_{\text{over}} \) and \( M_{\text{distinct}} \) provide similar results (70.2 and 69.1% resp.), whereas they behave very differently with the complete linkage (53.2 and 72.3% resp.).

Beyond the raw hit rate, it is interesting to analyse the types of misclassifications. With the highest scoring tree (\( S_{\text{prod}} \)), 15 of the 22 errors are ‘false negative’, i.e. a gene with a known function is assigned to the RAND cluster. This is even stronger with the Poisson mixed metrics and complete clustering, where 22 out of 26 misclassifications are false negatives.

DISCUSSION

Transcriptional regulation is a complex mechanism, and the intention of this paper is certainly not to give the false impression that gene function could be deciphered on the simple basis of the presence/absence of a few patterns in upstream regions. In addition, we only illustrated the metrics on the basis of one example taken from a yeast, where regulation is incomparably simpler than in higher organisms. Being aware of these limitations, our ambition is restricted to emphasize that (i) with compact genomes, simple counts of multiple patterns already provide some functional grouping of genes; (ii) the results are drastically affected by the choice of the metric and clustering method and (iii) if pattern counts have to be compared between sequences, probabilistic metrics provide better results than classical metrics.

It is hard to choose between the different Poisson-based metrics on pure theoretical considerations. We took the pragmatic option to evaluate them on the basis of a biological example. The best results were obtained with \( S_{\text{prod}} \), which reflects the joint probability of multiple pattern occurrences. All additive models return weaker results. However, among them, mixed metrics provide better results than similarity or dissimilarity alone. Another important criterion is the choice of the clustering method. With most metrics, Ward clustering returns better results than complete, average or simple linkage. However, different results might be obtained with other data sets, and it will be interesting to assess the robustness of the results.

It is interesting to note that some metrics may show different behaviors with random or biological sequences, respectively. Although with random data Park’s heuristic shows 98% correlation with the \( M_{\text{over}} \) metrics, its performances on our biological example are clearly weaker. This likely comes from the fact that, for their heuristic, we used \( \alpha \) and \( \beta \) coefficients taken from their publication. A specific optimization would probably improve the performance of Park’s metrics on our data set, but this would require an independent training set. Thus, in a general context of unsupervised classification, a probabilistic metric like the \( S_{\text{prod}} \) seems more appropriate than a training-based heuristic.
Poisson-based metrics are of general application, and can be used in different contexts, but they rely on two assumptions: (i) each variable must be Poisson-distributed and (ii) the variables must be independent from each other.

The first assumption (Poisson distribution) is generally valid, provided the sequence is sufficiently larger than the pattern size. There is however one exception: for self-overlapping patterns (e.g. AAAAAA, TATATA), the random distribution discards from Poisson (and from the binomial), and presents the characteristics of an aggregative distribution: the expected mean is unaffected, but the variance is increased, resulting in a higher probability of either low or high numbers of occurrences (Kleffe and Borodovsky, 1992). This bias can be circumvented in two ways. The simplest approach is to discard overlapping occurrences from the count, i.e. each time a pattern of length \( w \) is encountered at position \( j \), the next \( w - 1 \) positions are invalid for this pattern. We adopted this counting mode in this paper. The second possibility is to calculate probabilities on the basis of an aggregative distribution, like the compound Poisson (Reinert and Schbath, 1998), which permits to introduce a correction for the biased variance.

The second assumption (independence of the variables) can be problematic when the patterns present a high degree of mutual overlap. For example, if one measures the occurrences of cACGTG and ACGTGg, each pattern has a probability of the order of \( 4^{-6} \), but after each occurrence of cACGTG, the probability to observe ACGTGg at the next position raises to \( \sim 4^{-1} \). This problem is not specific for the Poisson-based dissimilarity metrics, and is common to most metrics (Euclidian, Manhattan, correlation, . . .) discussed in this paper. The Mahalanobis distance includes a correction for correlated variables, but this does apparently not improve the results on our data set. A correction for variable correlation in the dissimilarity/similarity metrics would be an interesting perspective to pursue the present work.

Finally, we would like to stress that the biological example presented here relies on hierarchical clustering, an unsupervised classification method. It is however clear that, whenever some prior information is available about class membership of a subset of genes, this information can be used to train a supervised classifier (e.g. discriminant analysis, support vector machines), which can then be used to assign a class to new objects (manuscript in preparation).

ACKNOWLEDGEMENTS

We thank Peter Park for interesting discussions and for comments and suggestions about this paper. This work was funded by the European Communities, grant no. QLR1-199-0133 to the CYGD project.

REFERENCES


