Quick Permutation Test: feature filtering of n-gram data

Piotr Sobczyk[1]*, Michał Burdukiewicz[2], Chris Lauber[3], Paweł Mackiewicz[2]
*Piotr.Sobczyk@pwr.edu.pl

[1]Wrocław University of Technology, Department of Mathematics, Poland
[2]University of Wrocław, Department of Genomics, Poland
[3]Dresden University of Technology, Institute of Medical Informatics and Biometry, Poland

Introduction

N-grams (k-tuples) are vectors of n characters derived from input sequence(s). They may form continuous sub-sequences or be discontinuous. Another important n-gram parameter is its position. Instead of just counting n-grams, one may want to count how many n-grams occur at a given position in multiple (e.g. related) sequences.
Originally developed for natural language processing, n-grams are also used in proteomics, genomics and transcriptomics.

	P1	P2	P3	P4	P5	P6
S1	3	4	4	3	4	1
S2	2	4	3	3	3	3
S3	1	3	1	4	3	1

Sample sequences.

1	2	3	4
1	0	2	3
0	1	4	1
3	0	2	1

Unigram counts.

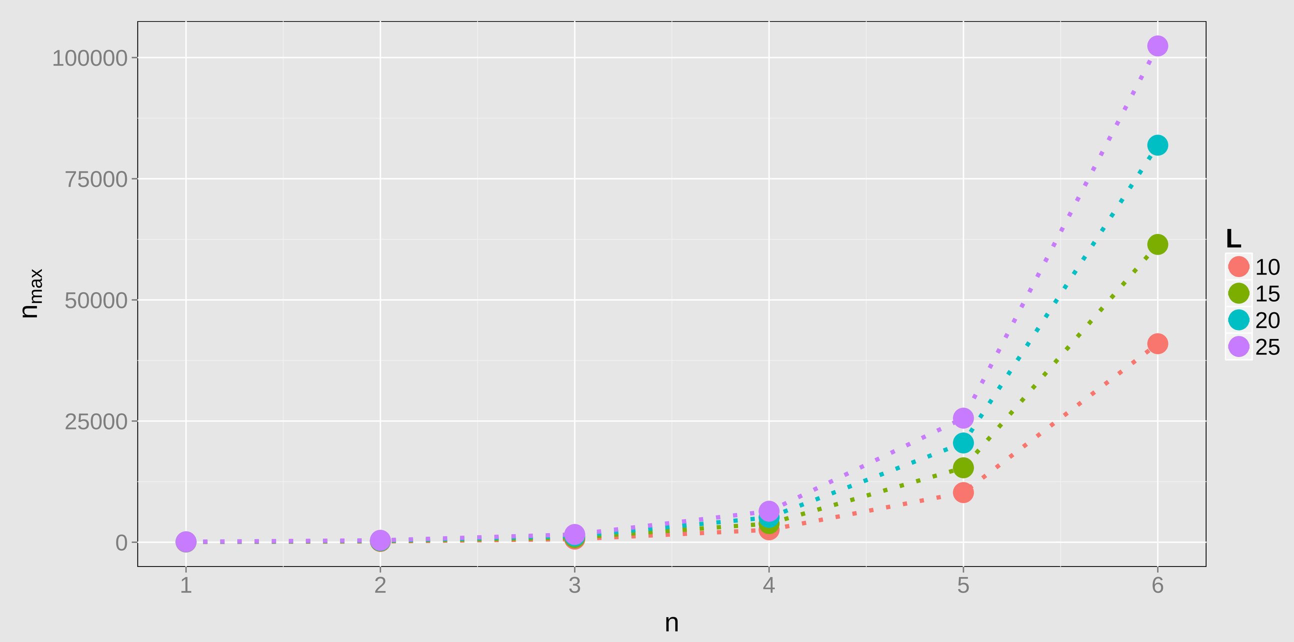
P1_1	P2_1	P3_1	P4_1	P5_1	P6_1	P1_2	P2_2	P3_2	P4_2	P5_2	P6_2	P1_3
0	0	0	0	0	1	0	0	0	0	0	0	1
0	0	0	0	0	0	1	0	0	0	0	0	0
1	0	1	0	0	1	0	0	0	0	0	0	0

A fraction of possible unigrams with position information.

Curse of dimensionality

Even when we limit ourselves to only continuous positioned n-grams, feature space growths rapidly with the number of elements in n-gram (n) and length of the sequence (L).
Number of possible positioned n-grams:

$$n_{\text{max}} = L \times m^n$$



Feature selecting permutation tests

Model and statistic independent permutation tests can be used to filter features obtained through counting n-grams.
During a permutation test class labels are randomly exchanged during computation of significance statistic. p-values are defined as:

$$\text{p-value} = \frac{N_{T_P > T_R}}{N}$$

where $N_{T_P > T_R}$ is number of times when T_P (permuted test statistic) was more extreme than T_R (test statistic for non-permuted data).
Permutation tests are computationally expensive (especially precise estimation of low p-values, because the number of permutations is inversely proportional to the interval between p-values).

QuiPT idea

If probability that target equals 1 is p and probability that feature equals 1 is q then we can compute the probability of given observations, eg.

$$P(Target, Feature) = (1, 1)) = p \cdot q$$

Therefore another view at permutation test is that we get a contingency table, which needs to be tested for independance.

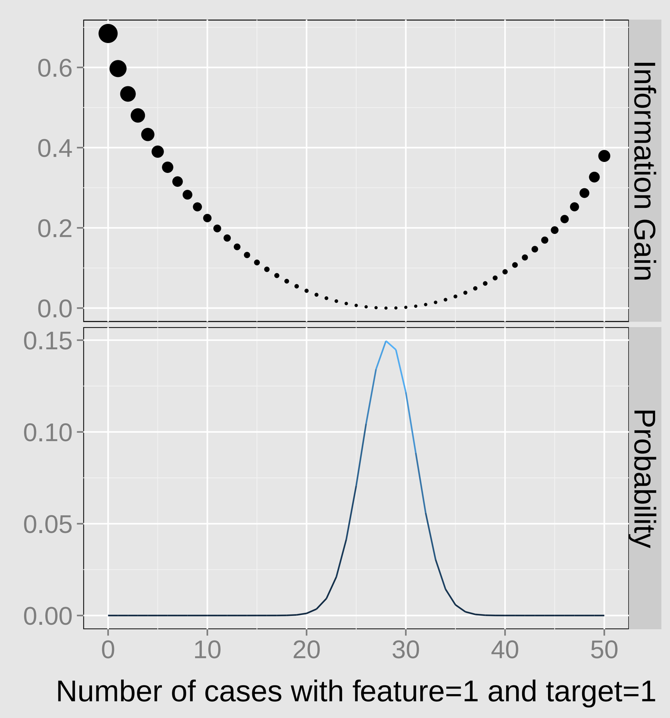
Independence test

F$(n_{1,1}, n_{1,0}, n_{0,1}, n_{0,0}) = \binom{n}{n_{1,1}} (p \cdot q)^{n_{1,1}} \binom{n - n_{1,1}}{n_{1,0}} (p \cdot (1-q))^{n_{1,0}}$
$\binom{n - n_{1,1} - n_{1,0}}{n_{0,1}} ((1-p) \cdot q)^{n_{0,1}} \binom{n - n_{1,1} - n_{1,0} - n_{0,1}}{n_{0,0}} ((1-p) \cdot (1-q))^{n_{0,0}}$
This distribution comes with two constraints: $n_{1,.} = n_{1,1} + n_{1,0}$ and $n_{.,1} = n_{1,1} + n_{0,1}$. Thus, conditioning on $n_{1,1}$, we get hypergeometric distribution.
This is in fact exact two-sided Fisher's test. Information Gain is used here as a way of deciding which contingency tables are more extreme.

Computational cost

The cost of performing QuiPT is equal to computing Information Gain and probability of occurence for $n_{1,1} + n_{0,1}$ contingency tables.
Suppose we consider 6-grams build on sequences of length 25 build of four characters. Then there are around 100,000 n-grams, features to test. This means that for Benjamini-Hochberg procedure, we need to calculate p-values with accuracy of 0.05×10^{-5}. This requires at least 2 million permutations. Each permutation, apart from reshuffling labels, requires computation of IG.
Since n-gram features are very sparse vectors, QuiPT needs to evaluate only few contingency tables.

Validation procedure

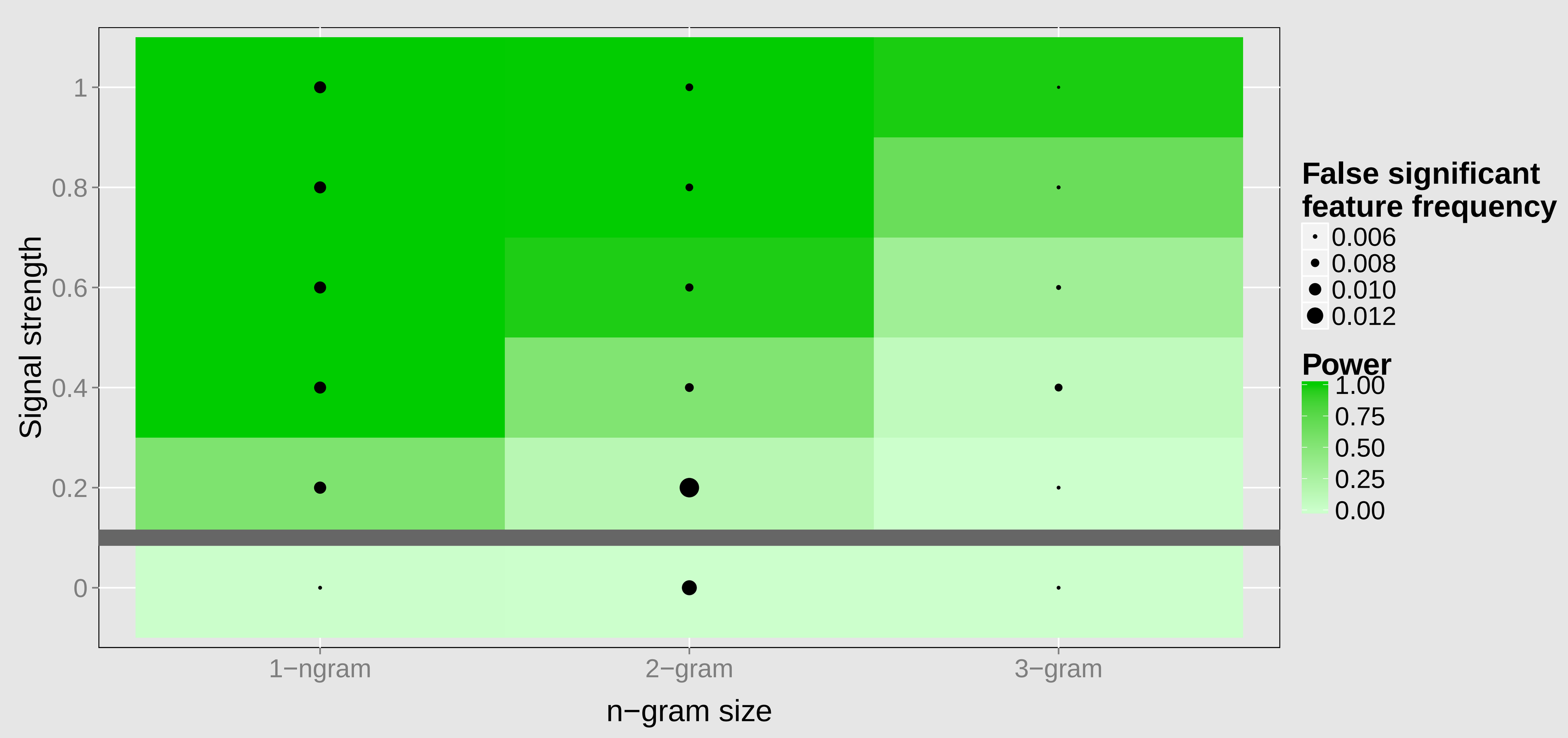


	Target	Feature	Freq
1	0	0	40
2	1	0	10
3	0	1	25
4	1	1	40

Simulation scheme

1. Random 4000 sequences (20 nucleotides each). The half of the sequences has label 0.
2. Choose a single position between 3 and 18 (to avoid border cases).
3. Resample nucleotides at chosen position. The dominant nucleotoide has probabilitiy of occurence $p_d = 0.25$. Other nucleotides have probability of occurence $p_o = (1 - p_d)/3$.
4. Perform QuiPT (Information Gain) and choose significant features (with p-value < 0.001).
5. Iterate steps 1-4 over other values of p_d - 0.38, 0.51, 0.65, 0.78, 0.91.
6. Repeat steps 1-5 200 times.

Power and False discoveries



Summary

Quick permutation test is a powerful and quick equivalent of permutation test in binary feature-binary target testing scenario.

Avaibility

biogram R package:
`http://cran.r-project.org/web/packages/biogram/`

Bibliography