BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 13 2005, pages 3051–3052
doi:10.1093/bioinformatics/bti451

Sequence analysis

S-SPatt: simple statistics for patterns on Markov chains

Grégory Nuel
Laboratoire Statistique et Génome, 523 place des terrasses de l'Agora, 91000 Evry, France

Received on February 15, 2005; revised on March 29, 2005; accepted on April 13, 2005
Advance Access publication April 19, 2005

ABSTRACT

Summary: S-SPatt allows the counting of patterns occurrences in text files and, assuming these texts are generated from a random Markovian source, the computation of the P-value of a given observation using a simple binomial approximation.

Availability: S-SPatt is available at: http://stat.genopole.cnrs.fr/spatt

Contact: spatt@genopole.cnrs.fr

1 PREVIOUS WORK

It is well known that patterns frequencies are highly dependent on their overlapping structures as well as on the composition bias in the considered sequences. In order to compare patterns efficiently, we need more than their simple counts. A common solution to this problem consists in using a Markov chain to modelize the sequence thereby ranking patterns according to their P-values.

Several software propose solutions to compute these P-values: QuickScore using exact computation with generative series as proposed in Régnier (2000) (under development but not yet available), RMES (Schbath, 1997) for Gaussian and Compound Poisson approximations and LD-SPatt for large deviations approximations (Nuel, 2004). Unfortunately, all these methods suffer severe drawbacks: QuickScore can not consider Markovian model of order higher than 1 or 2 and is limited to short sequences (for exact computations), Gaussian approximations are wrong for tail distribution events (which are of course the events of interest), Compound Poisson approximations are designed only for rare patterns (and have some numerical issues in the RMES implementation making its use difficult anyway) and, finally, large deviations techniques are limited to short patterns (less than 10 letters long on a DNA alphabet, for example).

The Regulatory Sequence Analysis tool (RSAT) proposes to use simple binomial approximations to compute these P-values. Simulations have shown (van Helden *et al*., 1998) that such approximations are wrong but very close to the optimal solution when the considered patterns are not self-overlapping too much.

As the RSAT package is only available on-line, we decided to implement the same method in a stand-alone GPL program called S-SPatt to compare its overall performances with other similar programs.

2 METHOD

2.1 Statistics

Let $X = X_1, \ldots, X_n$ be an order m stationary Markov chain on a size k finite alphabet $\mathcal{A}$, with Π ($k^m \times k^m$ dimension) sparse transition matrix and μ (k^m dimension vector) as stationary distribution.

We consider a pattern $\mathcal{W} = \{W_1, \ldots, W_r\}$ as a set of r words (of respective length $h_1, \ldots, h_r$).

We count the number of occurrences of a given word $W = w_1, \ldots, w_h$ using $N(W)$ as defined by

$$N(W) = \sum_{i=1}^{n-h+1} I_{\{W \text{ starts in } i\}} \triangleq \sum_{i=1}^{n-h+1} Y_i. \tag{1}$$

According to the model, Y_i has a Bernoulli distribution of parameter

$$P(W) = \mu(W_1^m)\Pi(W_1^m, W_2^{m+1}), \ldots, \Pi(W_{h-m}^{h-1}, W_{h-m+1}^h), \tag{2}$$

where $W_i^j = w_i, \ldots, w_j$ for all $1 \leq i \leq j \leq h$.

Couples (Y_i, Y_j) are clearly not independent nevertheless, we can use the following heuristic distribution for a single word:

$$N(W) \sim \mathcal{B}(n - h + 1, P(W)), \tag{3}$$

which hence gives the following distribution for a pattern:

$$N(\mathcal{W}) \sim \mathcal{B}\left(n - \max_{1 \leq i \leq r} h_i + 1, \sum_{i=1}^{r} P(W_i)\right). \tag{4}$$

2.2 Algorithms

2.2.1 Stationary distribution According to the theorem of Perron–Frobénius, for any irreducible Markov chain, the stationary distribution can be computed by solving an eigenvalue problem. Growing exponentially with the Markov order, the scale of this problem is usually very large. To solve this problem, we propose to use an explicitly restarted Arnoldi's algorithm (Stewart, 1994) which is known to be very efficient with large, sparse eigenproblems (for an order m Markov model we have a non zero density of k^{1-m}).

2.2.2 Number of occurrences All words of length smaller than a given L are first computed ($n \times L$ in time and memory) and deterministic finite state automata are used for larger patterns (Hopcroft and Ullman, 1979).

2.2.3 P-values P-value for the observations are computed with the incomplete Beta function (Press *et al*., 1992).

3 RESULTS

3.1 Numerical performances

The following table gives the computation time T (in seconds) for N computations of simple words' P-values (all words of length $h = 6, \ldots, 10$ in *Mycoplasma genitalium* with an order 1 Markov model; computations are performed on an Intel Pentium 4 processor at 2.8 Gz).

N	4^6	4^7	4^8	4^9	4^{10}
$T(s)$	0.10	0.17	0.42	1.41	5.33

A simple linear regression gives about 200 000 computations per second which is very fast. Let us add, as a remark, that this result is independent of the Markov model order (thanks to the Arnoldi algorithm). In comparison, for words of length 8 on a DNA alphabet, RMES Gaussian performs about 30 000 computations per second, RMES compound Poisson ~2000 and LD-SPatt ~4 (computation time with LD-SPatt grows exponentially with the length of the patterns).

S-SPatt can also be used to compute very efficiently simple word frequencies. For example, counting all words of length 8 on *Escherichia coli* complete genome requires roughly 15 min with the wordcount program from the popular EMBOSS package while S-SPatt can achieve the same task in less than half a second (hence, S-SPatt is ~2000 times faster than EMBOSS).

3.2 Reliability of the heuristic

As Gaussian approximations are expected to be good in the center of the distribution (for high P-values), they are taken as reference for such events while large deviations are used as reference for all tail distribution events (see Nuel, 2004 for more discussion on the subject).

We can see from Table 1 that our simple binomial approximation is the closest to the reference both in terms of relative error and, more important, in terms of rank agreement (see Press *et al*., 1992 for more details on Kendall's Tau).

Table 1. Reliability comparison

Method	Relative error		Kendall's Tau	
LD-SPatt	Ref	0.516	Ref	Ref
RMES G	0.262	Ref	0.636	0.794
RMES CP[a]	0.145	0.076	0.936	0.298
S-SPatt	0.119	0.065	0.932	1.000

Computation done with words of length 8 on *E.coli K12* with an order 1 Markov model estimated by maximum likelihood. Mean relative errors on log P-value are given first for significant words (P-value smaller than 10^{-4}) and then for the non significant ones. Mean Kendall's Tau are given first for the 50 most under-represented words and then for the 50 most over-represented ones.

[a] Indicates that the 20% worse results have been removed for RMES CP (because that software suffers severe open numerical issues in its actual implementation).

4 CONCLUSION

S-SPatt is not only the fastest tool (about 200 000 computations per second) but also the most reliable one (after the reference ones).

Moreover, the proposed implementation have several interesting features:

- Support of any user defined alphabet (regular DNA, purin-pyrimidin, amino acids, group of amino acids, latin, case sensitive, ...) with a simple syntax.
- Markov model parameters are estimated using maximum likelihood or are specified by the user.
- Support for high order Markov chain including computation of the stationary distribution using efficient linear algebra methods.

In conclusion, S-SPatt seems to be a very good heuristic for the computation of pattern statistics on Markov chains.

REFERENCES

van Helden,J. *et al*. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.

Hopcroft,J.E. and Ullman,J.D. (1979) *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, Reading, MA.

Nuel,G. (2004) LD-SPatt: large deviations statistics for patterns on Markov chains. *J. Comp. Biol.*, **11**, 1023–1033.

Press,W.H. *et al*. (1992) *Numerical Recipes in C*. Cambridge University Press.

Régnier,M. (2000) A unified approach to word occurrence probabilities. *Discrete Appl. Math.*, **104**, 259–280.

Schbath,S. (1997) An efficient statistic to detect over- and under-represented words in DNA sequences. *J. Comp. Biol.*, **4**, 189–192.

Stewart,W.J. (1994) *Introduction to Numerical Solution to Markov Chains*. Princeton University Press.