

Statistical analysis of counts and spacing of consistent repeating patterns in a set of homologous DNA sequences

D. V. Raje¹, H. J. Purohit^{1,*}, P. Lijnzaad² and R. N. Singh¹

¹Environmental Genomics Unit, National Environmental Engineering Research Institute, Nehru Marg, Nagpur 400 020, India

²Department of Biomedical Genetics, University Medical Center, Utrecht, The Netherlands

Unusual patterns in nucleic acid or protein sequences are often suspected for their biological relevance. Repeating patterns of nucleotides are one such type and are typically searched in large genome sequences. In this exercise, our interest is to look for repeating patterns, which are conserved in a set of homologous DNA sequences, not only in terms of their counts/occurrences, but also their spacing/separating distances. We refer to such patterns as consistent repeating patterns. It becomes desirable to know the probability of multiple occurrence of pattern in sequences and whether the spacing due to occurrences of pattern in the sequence exhibits any statistically significant property. The information derived through statistical analysis may help in planning experiments or even raise new queries that may require attention to better understand the molecular mechanisms. A case study with four hundred 16S rDNA sequences resulted into nine most consistent repeating patterns. The statistical significance of counts of these patterns was studied using Poisson approximation. The spacing analysis of patterns was carried with recourse to uniform probability distribution. The analysis revealed that most of the patterns show significant clustering, with one pattern occurring thrice and evenly dispersed in a sequence. The significance of occurrence and spacing of repeating patterns raised a few queries which requires explanation, perhaps through experimentation.

Keywords: Counts, DNA sequences, patterns, spacing, statistical analysis.

ANALYSING genetic information in terms of patterns of nucleotides or amino acids and relating the findings to either structure or function of a gene is an active field of research since the last decade. Such studies primarily focus on specific patterns in sequences followed by rigorous statistical analysis of their occurrences and inferring their possible structural or functional relevance in the biological system. Although statistical significance and biological significance are not synonymous, the ability to distinguish between that which is likely to occur and that which is unlikely to occur by chance is important in this context.

This may help in identifying sequence features that could be targets of experimental verification¹.

There are some reported studies on these lines. To cite a few: statistical analysis of counts and spacing (separating distances) of 4 to 6 bp palindromes in DNA sequences from a wide range of organisms was carried out and their possible structural and regulatory roles were interpreted². A method was devised to search for the repeated structural patterns and to compute the plausibility of such patterns by checking the frequency of occurrence in random sequences³. These types of patterns are shown to be of relevance in the overall stability of secondary structure of 16S rRNA molecule. Also, a program 'WORDUP' was developed to detect statistically significant oligonucleotide patterns of size 6 to 9 bp in a set of sequences^{4,5}. Further, a program 'PatSearch' was developed to analyse user-submitted sequences and to identify the presence of complex patterns with known functionality and assesses the statistical significance of their occurrences⁶. In addition, a method was proposed to identify significant patterns of arbitrary length in a set of related protein sequences having possible functional or structural relevance⁷.

Repeating patterns of nucleotides are often studied for their biological relevance. Repeats in bacterial genomes have been shown to affect bacterial virulence by acting as the molecular basis of a mechanism⁸. This makes identification of repeats an interesting topic of research. We believe that in a set of homologous sequences, it is quite likely that there are some repeating patterns which are conserved across the set, not only in terms of their occurrences, but also their positions and spacing. The present work aims at identifying such conserved/consistent repeating patterns of maximal length in a set of homologous sequences. It becomes interesting to know whether occurrences of consistent repeats in these sequences have simply arisen by chance. If not, then inferences drawn from the statistical analysis of selected repeats, coupled with the knowledge of the experimenter, may lead to better understanding of the molecular mechanisms. This communication emphasizes only the statistical aspects of the selected repeating patterns using 16S rRNA as model gene.

A repeating pattern is a sub-string of nucleotides in a sequence S , which occurs more than once in S . The origin, evolution and distribution of repetitive elements in genome databases have been the subject of intense study, both experimentally and computationally⁹. There are some algorithms developed to find repeats in a sequence¹⁰⁻¹². Amongst these, the recently developed 'REPuter' is highly efficient and provides exhaustive repeats, even in large genome sequences. The algorithm uses a compact implementation of suffix trees to locate exact repeats in linear space and time for sequences.

Although the above programs provide a list of repeats in a sequence, they do not have an option to automatically provide repeating patterns which are conserved across the input set of homologous sequences by simultaneously con-

*For correspondence. (e-mail: hemantdrd@hotmail.com)

sidering the spacing criterion. In fact, processing data on repeats from different input sequences can be a small extension to many of these programs. We have developed a program ‘Repeat Tuple Search’ (www.ebi.ac.uk/~lijnzaad/RepeatTupleSearch), which has this additional feature to determine the consistent repeating patterns across the set of sequences.

There are some reported studies on identifying over- and under-represented patterns in genome sequences¹³. Typical models used for DNA sequences are homogeneous m -order Markov chains in which the probability of occurrence of any letter at a particular position depends only on the previous m letters in the sequence. The appropriate choice of m rightly predicts $(m + 1)$ -letter patterns in the biological sequence.

Let S be a finite sequence of length l consisting of letters from set $\{A, C, G, T\}$. Let $w = w_1w_2 \dots w_h$ be a h -letter pattern/word in a sequence where $w_1 \in \{A, C, G, T\}$. Let $\mu_m(w)$ be the equilibrium probability of occurrence of word w in a sequence and $N(w)$ be the number of occurrences of w in a sequence. If a sequence is a stationary m -order Markov chain with transition probability matrix $\Pi = (\pi(a_1a_2 \dots a_{m+1}))_{a_1a_2 \dots a_{m+1} \in \{A, C, G, T\}}$ and stationary distribution $\mu(a_1a_2 \dots a_{m+1})_{a_1a_2 \dots a_m \in \{A, C, G, T\}}$, then probability of word w is

$$\mu_m(w) = \mu(w_1 \dots w_m) \prod_{j=1}^{h-m} \pi(w_j \dots w_{j+m-1}, w_{j+m}). \quad (1)$$

Also the Gaussian approximation of $N(w)$ was proposed where, with the increase in the length of the sequence l , the expected value of $N(w)$ grows to infinity. If the expectation of $N(w)$ is bounded with the increase in l , then the word is rare¹⁴. Alternatively, if $E(N(w)) \ll 1$, then the word is said to be rare. A rare word in a sequence may self-overlap and occur in clumps of size greater than one. For example, $w = \text{CTCACTC}$ can have two possible overlapping occurrences in a sequence like CTCACTCACTC or CTCACTCTCACTC. In the first case, the triplet CTC is overlapped, while in the other, only letter C is overlapped. Such a word is said to be periodic with periods 4 and 6 respectively, constituting $P(w)$, a period set of w . Further, the overlapping words may occur in clumps. For example, in the sequence ATGTCTCACTCACTC GGTACTCACTCTT, there are two clumps of CTCACT. The first one is of size 2 starting at position 5, while the other is of size 1 starting at position 20. The expected number of counts $N(w)$ in this case can be approximated by compound Poisson variable.

The approximate count of w is given by

$$N(w) = \sum_{k \geq 1} k N_k(w).$$

where $N_k(w)$ is the number of k -clumps. The process $N_k(w)$ is approximated by a Poisson process Z_k , such that $Z = \sum k Z_k$

follows compound Poisson distribution with parameter $\Lambda = (l - h + 1)\mu_k$, where

$$\mu_k = (1 - A)^2 A^{k-1} \mu(w), \quad (2)$$

and

$$A = \sum_{p \in P(w)} \prod_{j=1}^p \pi(w_j, w_{j+1}).$$

If a word self-overlaps and occur in clumps of size 1, then the expected number of words can be approximated by Poisson variable (Z). Also, occurrences of non-overlapping word can be approximated by Poisson distribution with parameter $(l - h + 1)\mu'_m(w)$, where for $m = 1$

$$\mu'_1(w) = \mu(w) - \sum_{p \in P(w)} \mu_1(w_1, \dots, w_p w_1 \dots w_h). \quad (3)$$

Accordingly, over- or under-representation of a word in a sequence can be decided using the following criterion: if $P(Z \geq N^{\text{obs}}(w))$ is close to 1, w is under-represented; and if close to 0, w is over-represented¹⁵.

In addition to the significance of pattern counts in a sequence, the significance of spacing between the repeats in a sequence can also be studied. The analysis of spacing between consecutive occurrences of patterns suggests the extent of homogeneity of their distribution. The analysis can lead to detecting patterns with significant clustering, dispersion or excessive regularity, and a method to assess these properties based on minimal and maximal spacing has been proposed¹⁶. It rests on the assumption of n randomly distributed words over a sequence of length l . These occurrences induce $n + 1$ spacings (U_0, U_1, \dots, U_n) , where U_0 is the distance before the first occurrence, U_i is the distance between the i th and $i + 1$ th occurrence and U_n is the distance after the last occurrence. The distances are rescaled on a unit interval. The statistical analysis focuses on the extremal spacing like $m^* = \min\{U_0, U_1, \dots, U_n\}$ and $M^* = \max\{U_0, U_1, \dots, U_n\}$. Under the assumption that the variable m^* , which is the smallest interval between two consecutive words is uniformly distributed over the interval $(0, 1/(n + 1))$, its distribution function is given by

$$F(a) = P(m^* < a) = 1 - (1 - (n + 1)a)^n \quad \text{for } 0 < a \leq 1/(n + 1), \quad (4)$$

while the distribution function for M^* is given by

$$G(a) = P(M^* < b) = 1 - \sum_{i=0}^{n+1} \binom{n+1}{i} (-1)^i [\delta(1 - i\delta)]^n, \quad \text{for } 1/(n + 1) \leq b < 1, \quad (5)$$

where $\delta = 1$ if $ib \leq 1$ and $\delta = 0$ otherwise.

Table 1. Forty different bacterial genera

Sr. No.	Bacterial genera	Sr. No.	Bacterial genera	Sr. No.	Bacterial genera	Sr. No.	Bacterial genera
1.	<i>Acetobacter</i>	11.	<i>Enterobacter</i>	21.	<i>Mythylotrophs</i>	31.	<i>Rhizobium</i>
2.	<i>Acinetobacter</i>	12.	<i>Flavobacterium</i>	22.	<i>Mycobacterium</i>	32.	<i>Salmonella</i>
3.	<i>Aeromonas</i>	13.	<i>Gluconobacter</i>	23.	<i>Nocardia</i>	33.	<i>Sphingomonas</i>
4.	<i>Alcaligenes</i>	14.	<i>Haemophilus</i>	24.	<i>Nitrobacter</i>	34.	<i>Streptococcus</i>
5.	<i>Azospirillum</i>	15.	<i>Halobacterium</i>	25.	<i>Nitrosomonas</i>	35.	<i>Streptococcus</i>
6.	<i>Bacillus</i>	16.	<i>Lactobacillus</i>	26.	<i>Photorhabdus</i>	36.	<i>Sulfolobus</i>
7.	<i>Burkholderia</i>	17.	<i>Micrococcus</i>	27.	<i>Pseudomonas</i>	37.	<i>Thermus</i>
8.	<i>Clostridium</i>	18.	<i>Moraxella</i>	28.	<i>Ralstonia</i>	38.	<i>Thiobacillus</i>
9.	<i>Commamonas</i>	19.	<i>Methanococcus</i>	29.	<i>Rhodococcus</i>	39.	<i>Vibrio</i>
10.	<i>Desulfovibrio</i>	20.	<i>Methanosarcia</i>	30.	<i>Rhodospirillum</i>	40.	<i>Xanthomonas</i>

The evaluation of an extremal minimum at some level of significance α depends on a , such that $F(a) = \alpha/100$. The observed m^* is considered to be significantly small if $m^* < a$. Similarly, the largest spacing is considered statistically significant if the observed M^* exceeds b , where b satisfies $G(b) = \alpha/100$. For a too large m^* , i.e. say $m^* \geq c$, where $F(c) = 0.99$ or too small M^* , i.e. $M^* < d$, where $G(d) = 0.99$, the spacings are considered to be excessively regular. The expressions are applicable for small n , while for large n asymptotic expressions based on r -scan statistics have been suggested².

We have reported that the information within the repeating pattern in 16S rDNA sequences carries the signature, which could be genus-specific¹⁷. For example, in case of genus *Pseudomonas*, the information content analysis carried out for sub-sequence within the repeating pattern CAGCAG yielded a genus-specific signature, which was validated using PCR primer designed from that region¹⁸.

The 16S rRNA gene has been reported to be highly conserved, but still diverse enough to classify eubacteria¹⁹. Thus if repeating patterns exist in these sequences, then some of them may be conserved across different sequences, and also their spacing may be preserved. With this assumption, a sample of 400 sequences belonging to forty different bacterial genera was retrieved from NCBI gene databank. Table 1 provides the list of organisms selected in the study, which are typically observed in the microbial community of effluent treatment plants. From each genus, ten sequences representing different species were selected, each with size around 1.5 kb. The sequences were used in the program 'Repeat Tuple Search' to obtain repeating patterns in each sequence and subsequently the most consistent repeating patterns across sequences. Tables 2 and 3 presents the list of most consistent repeats of size six and seven respectively, observed in more than 50% of the bacterial groups. The patterns in both the tables have been arranged in decreasing order of their occurrences (out of 400). It is evident from Table 2 that CAGCAG tops the list, repeating twice in 325 sequences. In 27 cases, the pattern repeated thrice, showing an additional occurrence towards the 3'-end of the sequences. The pattern has also been found repeating predominantly in 18S

rRNA of Eucarya²⁰. The single or no occurrence of pattern was also recorded as shown in Table 2. Numbers in the square brackets represent bacterial groups in which majority of sequences have either a single or no occurrence. It is evident that most of the selected patterns do not repeat in organisms like *Halobacterium*, *Methanogenes* and *Sulfolobus*. To ensure whether the repetitiveness of any pattern in a particular sequence is a chance phenomenon, the significance analysis was carried out.

As shown in Table 2, the patterns CAGCAG, CGCAAC, TGTCGT and CACACC self-overlap with periods 3,5,5 and 5 respectively, while the rest do not self-overlap. Further, the overlapping or non-overlapping patterns occur in clumps of size one; hence they were approximated by Poisson distribution with parameter $(l-h+1)\mu'(w)$, where $\mu'(w)$ was estimated using eq. (3). In each sequence, the Poisson probability estimates were obtained for the observed occurrences of the selected words. Table 2 provides distribution of sequences into two categories based on P values. If $P \leq 0.05$, the occurrence of a word in a sequence has been considered as significant. For instance, CAGCAG repeats twice in 325 sequences. In 222 cases, such an occurrence was found to be statistically insignificant, while in the remaining 103 cases it was significant. It is obvious that smaller the P value, larger is the difference between the observed and expected occurrences of the pattern in a sequence. As evident from Table 2, single occurrence ($n = 1$) or absence ($n = 0$) of any pattern in a sequence is statistically insignificant with probability values lose to 1. But for $n = 2$ or more, i.e. for repetitive occurrences, statistical significance is observed. Also it is evident that for TGTCGT, CACACC and GCTACA, majority of the sequences have P values less than 0.05 corresponding to $n = 2$, implying that the likelihood of getting a repeat of these patterns in sequences is less, considering the dinucleotide composition of the respective sequences. On the other hand, for CAGCAG, CTGAGA and CGCAAC, most of the sequences show P values above 0.05, indicating that the chance of getting a repeat of these patterns in sequences is relatively more compared to the other three repeats. Similarly, the repetitive occurrences of patterns GTGGGA, TGGGAG and TACACAC (Table 3) are quite

Table 2. Significance of occurrence of repeats of size six

Repeats of size six	Occurrence of pattern in a sequence (<i>n</i>)	No. of sequences*	Significance of repeats in sequences	
			<i>P</i> > 0.05	<i>P</i> ≤ 0.05
cagacag	0	12 _[21]	13	0
	1	30 _[9]	30	0
	2	325	222	103
	3	27	0	27
ctgaga	0	7	7	0
	1	59 _[4,21,22,27,33]	59	0
	2	278	232	46
	3	56	0	56
cgcaac	0	31 _[8,9,21,27]	31	0
	1	40 _[16,36]	39	1
	2	315	240	75
	3	14	0	14
tgtcgt	0	43 _[8,9,21,27]	43	0
	1	75 _[5,10,21,33,36,37]	73	2
	2	253	31	222
	3	28	0	28
	4	1	0	1
cacacc	0	34 _[14]	34	0
	1	92 _[6,9,11,12,19,21,27,36,38]	91	1
	2	243	11	232
	3	30	0	30
gctaca	0	53 _[14,17,21,34]	53	0
	1	100 _[4,6,11,12,19,27,29,33,38]	100	0
	2	225	49	176
	3	22	0	22

*Number(s) in the square bracket indicates bacterial group(s) as ordered in Table 1. In these groups, there are single or no occurrence of pattern in more than 60% of the sequences belonging to that group.

Table 3. Significance of occurrence of repeats of size seven

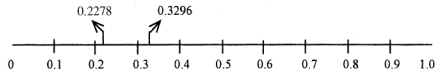
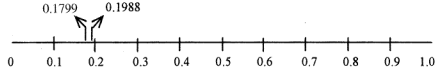
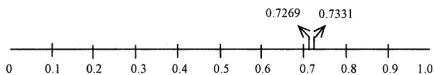
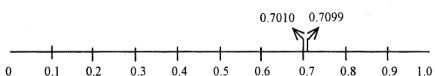
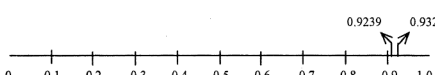
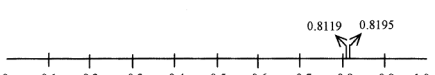
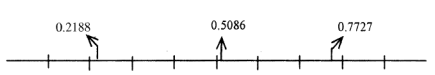
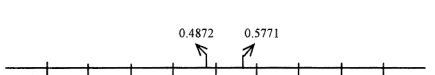
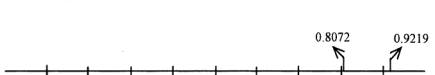
Repeats of size seven	Occurrence of pattern in a sequence (<i>n</i>)	No. of sequences*	Significance of repeats in sequences	
			<i>P</i> > 0.05	<i>P</i> ≤ 0.05
gtggga	0	31 _[8,21,27]	13	0
	1	53 _[2,9,28,38]	30	0
	2	93	2	91
	3	217	0	217
	4	6	0	6
tggggag	0	33 _[9,27]	33	0
	1	77 _[8,12,19,26,36]	77	0
	2	274	0	274
	3	14	0	14
	4	2	0	2
tacacac	0	58 _[8,9,14,21,27]	58	0
	1	97 _[4,6,10,11,12,19,34,38]	97	0
	2	242	0	242
	3	3	0	3

*Number(s) in the square bracket indicates bacterial group(s) as ordered in Table 1. In these groups, there are single or no occurrence of pattern in more than 60% of the sequences belonging to that group.

unusual as indicated by their low probability values in most of the sequences. Especially, the *P* values were much lower for TACACAC compared to others (data not shown).

To study whether the distribution of repeating patterns in a sequence exhibits any statistically significant property, spacing analysis was carried out. The identified repeating

Table 4. Spacing between consistent repeating patterns and their positions in 16S rDNA sequence

Repeating pattern	Separating distance	Positions in a sequence on a [0–1] scale*
cagcag	167	
ctgaga	27	
cgcaac	9	
tgtcgt	13	
cacacc	12	
gctaca	11	
gtgggga	414, 380	
tggggag	110	
tacacac	168	

*Actual separating distance between the repeats in a sequence, re-scaled to 0–1.

patterns were tested for three properties, viz. clustering, over-dispersion and regularity of distribution in a sequence. The positions of each repeating pattern in each sequence were obtained on a (0–1) scale. For example, in a sequence with accession number AB013253 (size 1531 bp), the repeating pattern CAGCAG appears at locations 340 and 507 with reference to the 5'-end. These positions are re-scaled to 0.2220 and 0.3311. Likewise, the re-scaled locations of this pattern in different sequences were obtained and averaged. Table 4 displays the average positional values for the selected patterns. The results of significance analysis of spacing are shown in Table 5.

The repeating patterns CACACC, GCTACA and TACA CAC are clustered significantly towards the 3'-end of sequences, as indicated by their low probability values (less than 0.05), while the pattern CTGAGA is significantly clustered towards the 5'-end of these sequences. Further, the repeats of patterns CGCAAC, TGTGGT, CACACC and GCTACA are close ($P < 0.05$) compared to others and even the spacing of 27 bases (occurrences of CTGAGA) is significantly close considering the sequence length of approx. 1.5 kb. The significant threshold spacing is 38 bases

at 5% level for approximate sequence length of 1.5 kb, under the assumption of uniform distribution of patterns. Since five out of the eight repeating patterns have their spacing less than 38, clustering in these cases is prominent. For any other set of sequences, the threshold would be different depending upon the length of sequences and thus clustering criterion would be different. Also, high probability values associated with triple occurrence of GTGGGGA indicate regular dispersion of pattern in sequences.

Thus, the present study describes some statistical criteria to assess the significance of occurrences of a pattern and its distribution in a sequence. There is one reported study for protein sequences in which the authors identified compositional extremes, clusters and runs of charge residues and spacing between the identical residue types and evaluated several positional properties based on statistical criteria²¹. On the contrary, our focus is only on the nucleotide repeats and their positions, spacing and statistical characterization.

The positioning of repeats on the secondary structure of *Escherichia coli* 16r RNA²² revealed some interesting observations. The repeats which showed significant clus-

Table 5. Significance of clustering and regularity of dispersion of the selected nine repeating patterns in 16S rDNA sequence

Repeating pattern	Spacing [#]				Probability			
					Clustering of pattern in a sequence		Regularity of dispersion	
	u_1	u_2	u_2	u_4	5'-end	3'-end	$P\{m^* < a\}$	$P\{M^* < b\}$
cagcag	0.2208	0.1088	0.6704	–	0.1086	0.6071	0.5462	0.3259
ctgaga	0.1799	0.0189	0.8012	–	0.0392	0.6725	0.1101	0.1185
cgcaac	0.7269	0.0062	0.2669	–	0.5374	0.0745	0.0368	0.2237
tgctgt	0.7010	0.0089	0.2901	–	0.5039	0.0894	0.0526	0.2682
cacacc	0.9239	0.0082	0.0679	–	0.8688	0.0057	0.0485	0.0173
gctaca	0.8119	0.0076	0.1805	–	0.6715	0.0353	0.0450	0.1061
gtgggga	0.2188	0.2898	0.2441	0.2273	0.4613	0.4767	0.9980	0.9960
tggggag	0.4782	0.0899	0.4229	–	0.3330	0.2629	0.4666	0.7868
tacacac	0.8072	0.1147	0.0781	–	0.8498	0.0371	0.5697	0.1115

tering are seen on the same helix of the secondary structure. For example, the two repeating patterns TGTCGT and CGCAAC are seen on helix 34 and 37 respectively; repeats of GCTACA and CACACC are seen on helix 41 and 44 respectively; while CTGAGA, which is separated by a distance of 27 bp occurs on helix 12. Amongst these, TGTCGT, GCTACA and CACACC even show statistically surprising occurrences in sequences. The presence of these closely spaced repeats, significantly occurring and clustering at either 5' and 3'-ends of the sequences, and conserved across various sequences, requires attention to explore their possible roles in the structure and/or function of the ribosome. Also, the regularly spaced pattern GTGGGGA points towards its possible biological relevance.

It is pertinent to mention here that even though the repeats of TGTCGT, GCTACA, CACACC and GTGGGGA emerged as highlighting features of 16S gene based on statistical criteria, the functional importance is not assured due to lack of firm correspondence between statistical and biological significance. Such analysis can only be exploratory, raising new queries and sometimes providing lines of investigation. For instance, the locations of repeats of the above 6-mers on the secondary structure show that although the repeats of a pattern are on the same helix, their structural contexts are different. The same applies for the remaining identified repeating patterns. Now if the structural contexts of the repeats of a pattern are different, then the functional contexts may presumably differ as well, raising a doubt about the meaningfulness of patterns in the sequence. If so, then why are they so preserved across different genera? Is it because the gene itself is highly preserved and hence may be the repeats of these patterns; but then, why only these repeats and not others? Secondly, what about the evenly spaced pattern showing different structural contexts in different helices? These types of questions require attention of biologists so that their conservation in a set of sequences can

be justified. Such analysis for any other set of homologous sequences would provide relatively important consistent repeating patterns in that group; and the information could provide new insights to better understand molecular mechanisms.

1. Karlin, S. and Altschul, S., Methods for assessing the statistical significance of molecular sequence features for using general scoring schemes. *Proc. Natl. Acad. Sci. USA*, 1990, **87**, 2264–2268.
2. Karlin, S., Brendel, V. and Bucher, P., Significant similarity and dissimilarity in homogeneous proteins. *Mol. Biol. Evol.*, 1992, **9**, 152–167.
3. Bouthinon, D. and Soldano, H., A new method to predict the consensus secondary structure of a set of unaligned RNA sequences. *Bioinformatics*, 1999, **15**, 785–798.
4. Pesole, G., Prunella, N., Liuni, S., Attimonelli, M. and Saccone, C., WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. *Nucleic Acids Res.*, 1992, **20**, 2871–2875.
5. Liuni, S., Prunella, N., Pesole, G., D'Orazio, T., Stella, E. and Distante, A., SIMD parallelization of the WORDUP algorithm for detecting statistically significant patterns in DNA sequences. *CABIOS*, 1993, **9**, 701–707.
6. Pesole, G., Liuni, S. and D'Souza, M., PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics*, 2000, **16**, 439–450.
7. Bejerano, G. and Yona, G., Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics*, 2001, **17**, 23–43.
8. Van Belkum, A., Scherer, S., van Alphen, L. and Verbrugh, H., Short sequence DNA repeats in prokaryotic genomes. *Microbiol. Mol. Biol. Rev.*, 1998, **62**, 275–293.
9. Smit, A., The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.*, 1996, **6**, 743–748.
10. Rigoutsos, I. and Floratos, A., Combinational pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics*, 1998, **14**, 55–67.
11. Kurtz, S. and Schleiermacher, C., REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics*, 1999, **15**, 426–427.
12. Kurtz, S., Choudhari, J., Schleiermacher, C., Stoye, J. and Giegerich, R., REPuter: the manifold applications of repeat analysis on a genome scale. *Nucleic Acids Res.*, 2001, **29**, 4633–4642.

13. Leung, M., Marsh, G. and Speed, T., Over and under representation of short DNA words in Herpesvirus genomes. *J. Comp. Biol.*, 1996, **3**, 345–360.
14. Prum, B., Rodolphe, F. and Turckheim, E. D. D., Finding words with unexpected frequencies in DNA sequences. *J. R. Stat. Soc. Ser. B*, 1995, **57**, 205–220.
15. Schbath, S., Compound Poisson approximation of word counts in DNA sequences. *Probability Stat.*, 1995, **1**, 1–16.
16. Karlin, S. and Brendel, V., Chance and statistical significance in protein and DNA sequence analysis. *Science*, 1992, **257**, 39–49.
17. Liskiewicz, M., Purohit, H. J. and Raje, D. V., Relation of residues in the variable regions of 16S rDNA and their relevance to genus specificity. *Lect. Notes Comput. Sci.*, 2004, **3240**, 362–373.
18. Purohit, H. J., Raje, D. V. and Kapley, A., Identification of signature and primers specific to genus *Pseudomonas* using mismatched patterns of 16S rDNA sequences. *BMC Bioinformatics*, 2003, **4**, 19.
19. Amann, R., Ludwig, W. and Schleifer, K., Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol. Rev.*, 1995, **5**, 143–169.
20. Iyer, D. S., Raje, D. V., Purohit, H. J., Gupta, A. and Singh, R. N., CAGCAG – the most consistent repeating pattern in evolution of small subunit of rRNA gene sequences. *Curr. Sci.*, 2004, **87**, 494–500.
21. Brendel, V., Bucher, P., Nourbakhsh, I., Blaisdell, E. and Karlin, S., Methods and algorithm for statistical analysis of protein sequences. *Proc. Natl. Acad. Sci. USA*, 2000, **89**, 2002–2006.
22. Connor, M., Thomas, C., Zimmermann, R. and Dahlberg, A., Decoding fidelity at the ribosomal A and P sites: influence of mutations in three different regions of the decoding domains in 16S rRNA. *Nucleic Acids Res.*, 1997, **25**, 1185–1193.

Received 14 February 2005; revised accepted 9 March 2006

Pollination biology of *Aristolochia tagala*, a rare species of medicinal importance

R. Murugan¹, K. R. Shivanna² and R. R. Rao^{1,*}

¹Central Institute of Medicinal and Aromatic Plants, Resource Centre, Allalasaandra, GKVK Post, Bangalore 560 065, India

²Ashoka Trust for Research in Ecology and the Environment, No. 659, 5th 'A' Main Road, Hebbal, Bangalore 560 024, India

Floral phenology, pollination biology and breeding system were studied in *Aristolochia tagala* Cham. (Aristolochiaceae) grown under *ex situ* conditions. The flower exhibits structural features typical of fly-trap mechanism described for other *Aristolochia* species. Flowers show pronounced protogyny. Stigmas are receptive at anthesis and remain so for 24 h. Anthers dehiscence 45–48 h after anthesis by which time stigma receptivity is lost. Chironomid fly (Diptera) is the pollinator. Attracted by the odour and colour of the flower, the flies enter it and are detained in the chamber of the peri-

anth tube (where the anthers and stigma are located) for nearly 50 h. Their escape is prevented by the presence of dense downward-pointing hairs in the perianth tube. The nectaries provide food to the insects. Following anther dehiscence, the thorax of the flies becomes loaded with sticky pollen grains. Hairs on the inner wall of the perianth tube wither and facilitate the exit of the flies. When a fly carrying the pollen load enters a fresh flower, it brings about pollination. Manual pollinations showed that the species permits geitonogamous pollination. The percentage of fruit set in manually pollinated flowers is higher than that resulting from open pollination, confirming that pollination is a limitation for fruit set in the *ex situ*-grown population. Nevertheless, fruit and seed set is sufficiently high for *ex situ* conservation purposes.

Keywords: *Aristolochia* sp., Chironomid fly, geitonogamy, pollination biology.

ARISTOLOCHIA L. is a large genus of the Aristolochiaceae with about 120 species, distributed throughout the tropical and subtropical countries. *Aristolochia tagala*, a climbing shrub is distributed in India, Sri Lanka, China, Malaysia, Burma, Java and Australia, and is a rare medicinal plant. The roots are strongly aromatic and are used to treat snake bites, bone fracture, malaria, indigestion, rheumatism, toothache and various dermatological conditions by Kani tribe of Thiruvananthapuram and Tirunelveli hills¹. Roots are also used for medicated steam bath 'sudorification'. Leaves are used to treat colic fits and bowel complaints. Due to indiscriminate harvesting of roots for local medicine and trade, the species has become rare in its natural habitat¹. Saplings collected from natural habitats have been introduced at the Conservatory of the Central Institute of Medicinal and Aromatic Plants (CIMAP) Resource Centre at Bangalore (lat. 13°05' N, long. 77°35' E; altitude 930 m asl). They have established well and are flowering regularly. Each plant produces a large number of flowers (>500).

Adequate knowledge on reproductive biology is essential for conservation, management and recovery of rare and endangered species. To our knowledge, there are no studies on the reproductive biology of *A. tagala*. This communication reports the results of our studies on floral phenology, pollination biology and breeding system of *A. tagala* grown under *ex situ* condition in Bangalore.

Flowers are distinctly stalked, bisexual, zygomorphic with inferior ovary and are produced in axillary cymes (Figure 1a). The perianth consists of three united, tubular, 7–8 cm long, purplish-brown lobes. The perianth tube is 2 mm wide and the inner surface is lined with strigose downward-pointed hairs, which facilitate the entry of flies into the chamber of the flower, but restrict their exit. The perianth tube is swollen into a globose chamber (utricle) in the basal part. The inner surface of the chamber is purplish and bears six dark brown, thick secretory nectaries. The perianth tube terminates into an expanded limb

*For correspondence. (e-mail: (email: rr_rao@vsnl.net)