

DETERMINING THE STATISTICAL SIGNIFICANCE OF OBSERVED FREQUENCIES OF SHORT DNA MOTIFS IN A GENOME

Philip E. Pfeiffer
Department of Computer Science
E-mail: pfeiffpe@notes.udayton.edu

Peter W. Hovey
Department of Mathematics
E-mail: Peter.Hovey@notes.udayton.edu

Sudhindra R. Gadagkar
Department of Biology
E-mail: gadagkar@notes.udayton.edu

University of Dayton
Dayton, OH 45469

Abstract: Until recently over 90 percent of the DNA in the human genome was considered junk DNA, with no known function. However, this non-coding DNA is now known to harbor elements that perform important functions in gene regulation. In particular, there is currently much interest in the search for short DNA motifs collectively known as *cis*-regulatory elements. Most studies attempt to identify these elements by means of cross-species comparisons. We have approached the problem of finding *cis*-regulatory elements by searching for conserved DNA motifs within genomes. This requires searching for DNA motifs that are repeated in the genomes either more or less frequently than expected by random chance. However, the usual chi-squared test cannot be used to test for the statistical significance of any observed frequency since overlapping regions of the genome are checked for DNA motif matches. We present here a statistical measure that has been developed to quantify the expectation and variance of the frequency of a given DNA motif in a given target sequence that may contain overlapping regions.

Introduction

The regulation of protein-coding gene expression is very tightly controlled in eukaryotes, necessitated partly by the spatial and temporal intricacies involved in the making of a given

gene product. The process of regulation is itself also very complex, involving the precise interaction among dozens of proteins and between the proteins and non-coding DNA in the vicinity of the protein-coding gene. These proteins, known as transcription factors (TF), are products of other protein-coding genes elsewhere in the genome that trigger a change in the expression of the gene in question. The production and binding of the TFs to the DNA near the gene and the subsequent cascade of reactions that takes place in the vicinity of the gene constitute the *trans* and *cis* parts, respectively, of gene regulation in eukaryotes (Carroll et al., 2005).

Cis gene regulatory elements are modules of noncoding DNA responsible for binding with transcription factors to influence how RNA polymerase II transcribes genes. These elements are typically found in a region of the gene known as the promoter. The most elemental member of regulatory noncoding DNA is the transcription factor binding site (TFBS). A TFBS is a stretch of 5-20 nucleotides that allows for the binding of a transcription factor protein (Rombauts et al., 2003). Transcription factors have specific binding motifs, although some degeneracy within the TFBS is tolerated (Carroll et al., 2005).

It is now known that noncoding regulatory DNA allows for changes in regulation of a given gene in different species without requiring significant alteration of the gene's protein coding DNA. For example, regulatory sequences such as those found in upstream promoter regions, and occasionally in introns and downstream regions, lack the constraints of coding regions and are therefore free to evolve more rapidly, thus permitting other changes to the gene product (such as the timing, quantity or location) without affecting the integrity of the protein itself.

Discovering novel TFBSs or even recognizing known k -mers (a short DNA fragment of length k nucleotides) in a genome using standard bench techniques is a difficult and tedious process that is exacerbated by species- and tissue-specific complexities of the protein-protein and DNA-protein interactions and the extremely short length of TFBSs (Cvekl et al., 2004; Xu et al., 1997). Rather, this is an area that would benefit greatly by the use of computer-intensive tools to narrow down the list of nucleotide k -mers that are likely to be important in *cis*-regulation. Realizing this, there have been several recent studies that have used computational techniques to study *cis*-regulation (Elemento and Tavazoie, 2005; Ettwiller et al., 2005; Jones and Pevzner, 2006; Xie et al., 2005; Xie et al., 2007). Most computational studies infer function from noncoding DNA that is conserved across species. Furthermore, this function is likely to be regulatory. For example, recently Xie et al. (2007) searched for conserved long oligonucleotides within a population of human CNEs (conserved noncoding elements) obtained from an alignment of 12 mammal species, and were able to find both experimentally validated k -mers with function and novel ones.

Our procedure involves determining the number of exact matches for a given k -mer within a genome, chromosome, or other target sequence, and then infers the functional significance of the motif by comparing its frequency to statistical expectations for the target database. However, determining the statistical significance of the frequency of a given k -mer in a target sequence (such as a chromosome) is hampered by several factors, notably that multiple matches for a k -mer may overlap in the target sequence, thus rendering the chi-squared test inappropriate. In fact, even computation of the expected frequency is not straightforward since the nucleotides are not distributed uniformly across each chromosome. However, for the purpose of this paper, we make the simplifying assumption of a uniform probability among the four nucleotides, in order to develop a simple statistical test of an observed frequency of k -mer matches in a target sequence. Work is underway to incorporate this and other biologically realistic scenarios into the framework of the model developed here.

The sample space of our theoretical genome (Γ)

Let λ denote a target sequence (e.g., a chromosome) of length l that is a string of symbols (e.g., the nucleotides A, C, G, and T) drawn from the set Φ , and let Γ be a set of sequences, each representing one replicate of the target sequence. We then define our sample space Γ as the set of all possible strings λ of length l composed of the symbols from set of symbols Φ :

$$\Gamma \equiv \{\lambda \mid \lambda \text{ is a string of symbols } q_i, i = 1, \dots, l, \text{ and } q_i \in \Phi\}.$$

Note that in this paper, we assume that each symbol occurs with equal probability along the entire length of the sequence λ – a simplifying assumption regarding the structure of molecular sequences.

We now define ζ as a k -mer – a short string of symbols of length k (where $k \ll l$), from our symbol set Φ . For example, ζ could represent a DNA fragment that is in the size range of most transcription factor binding sites (TFBSs). The problem consists of testing for the statistical significance of the frequency of exact matches of a given k -mer (candidate TFBS) in a target sequence λ . However, since two or more matches of a given k -mer may overlap in λ , these events are not independent of each other. Under these circumstances, the usual chi-squared test cannot be used. Therefore, we need to compute the expectation and the variance of the number of matches in order to develop a Z -test to evaluate the statistical significance of an observed frequency of matches for a given candidate TFBS.

The Random variable X_i defined on Γ and its expected value and variance

We define X_i as a Bernoulli variable that gets a value of 1 if a given k -mer, ζ , matches exactly the target sequence λ (denoting success) starting at position i , and a value of 0 if it does not (Figure 1, below):

$$X_i = \begin{cases} 1 & \text{for } w_j = q_{i+(j-1)}, j = 1, 2, 3, \dots, k \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The subscript i will vary over all potential match positions in the target sequence, ($i = 1, \dots, N$ where $N = l - k + 1$). For a given target sequence, then, we will then have N random variables, each representing a position in the target sequence beginning at which there is a possible match with each of the corresponding symbols in the given k -mer of interest. In equation (1), q_i is the symbol at position i in a random sequence, λ , from our sample space Γ , w_j is the symbol at position j in the k -mer, ζ , and $1 \leq j \leq k$. In other words, if all the symbols in ζ match the corresponding symbols, starting at position i , in λ , then $X_i = 1$, else $X_i = 0$. For example, Figure 1 shows ζ , a 5-mer, in which all the symbols match exactly the corresponding symbols in the target sequence, λ , starting at position i , and therefore, $X_i = 1$.

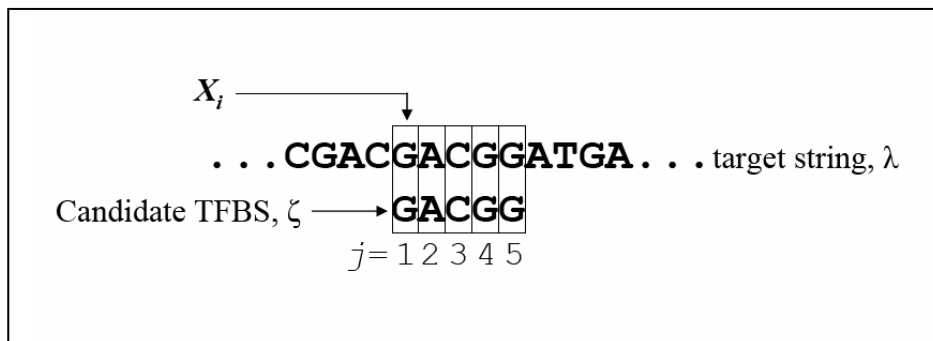


Figure 1. A perfect match between a candidate TFBS, ζ (= GACGG) and the target string, λ , at position i . Here, the Bernoulli variable $X_i = 1$ because a perfect match is found.

The probability function for X_i over our sample space Γ can be defined as follows:

$$p(1) = P(X_i = 1) = \prod_{j=1}^k P(w_j = v_{i+(j-1)}) = \left(\frac{1}{s}\right)^k \quad (2)$$

$$p(0) = P(X_i = 0) = \prod_{j=1}^k P(w_j \neq v_{i+(j-1)} \text{ for some } j) = 1 - P(X_i = 1) = 1 - \left(\frac{1}{s}\right)^k, \quad (3)$$

where p is the probability function of our random variable X_i , k is the length of the given k -mer, and s is the cardinality of our symbol set ($|\Phi| = s$). Note that we have the same function p for each of the random variables when $X_i = 1$, and similarly when $X_i = 0$. This results from our simplifying assumption that each symbol in Φ occurs with equal probability along the entire length of the sequence λ . Thus, the probability of a match of the j^{th} position of the k -mer at a given position i in a random sequence is independent of any other position in the random sequence giving us, $P(w_j = v_{i+(j-1)}) = \left(\frac{1}{s}\right)$. This is true for all i ($1 \leq i \leq N$) in our domain of consideration. This results in a probability dependant only on the number of symbols we have in our symbol set (a constant w.r.t. the subscript i .) In addition, the probability of a given symbol in ζ matching the corresponding symbol in λ is independent of the probability of matching of any other symbol in the two strings. This allows us to use the product law of probability and obtain the quantity $\left(\frac{1}{s}\right)^k$ in (2) and (3) above.

This nature of our random variables causes the expectation for any X_i to become:

$$E(X_i) = \sum_{x=0,1} x \cdot p(x) = [0 \cdot p(0) + 1 \cdot p(1)] = \left(\frac{1}{s}\right)^k, \quad \text{where } s = |\Phi| \quad (4)$$

$$V(X_i) = E[X_i^2] - \mu^2 = 1 \cdot p(1^2) - \left(\left(\frac{1}{s}\right)^k\right)^2 = \left(\frac{1}{s}\right)^k - \left(\frac{1}{s}\right)^{2k} = \left(\frac{1}{s}\right)^k \left[1 - \left(\frac{1}{s}\right)^k\right]. \quad (5)$$

Equations (4) and (5) are consistent with the well known results for a Bernoulli random variable.

The total matches $T(X_1, \dots, X_N)$ on a genome for a given k -mer

In order that a k -mer, ζ , be considered a good candidate TFBS or element with regulatory function, it must be present in the target sequence, λ , at a frequency that is statistically greater than (or less than) that expected by random chance. The number of occurrences of the k -mer in a given λ from our sample space can be obtained by a summation of all the values of all of the random variables X_i over all positions in λ . In other words, the total number of matches T , is a function of all of the random matching variables X_i and is expressed as follows:

$$T = f(X_1, \dots, X_N) = \sum_{i=1}^N X_i. \quad (6)$$

The random variable T expresses the total number of matches of a given k -mer for our target genome. Therefore, the expected value of T will represent the expected value for the total number of matches of our k -mer:

$$E[T] = E[f(X_1, \dots, X_N)] = E\left[\sum_{i=1}^N X_i\right] = \sum_{i=1}^N E[X_i] = \sum_{i=1}^N \left(\frac{1}{s}\right)^k = N\left(\frac{1}{s}\right)^k. \quad (7)$$

The variance of T

As mentioned earlier, the probability of a given symbol in ζ matching the symbol in the corresponding position in λ is the same for all positions (because of our simplifying assumption of uniform probabilities for all s symbols across λ) and thus, independent of another symbol matching in another position. However, this does not exclude the possibility of overlapping matches for multiple instances of the same k -mer in a given target sequence, rendering two random variables, X_i and X_j , for a given k -mer non-independent if there is an overlap (Figure 2 below).

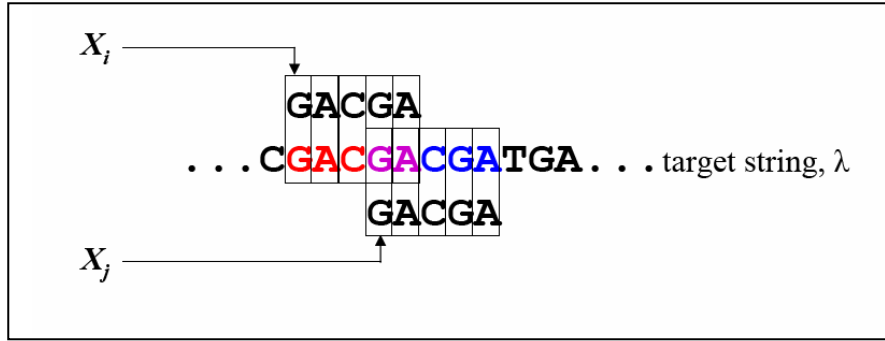


Figure 2. Overlapping exact matches for a 5-mer on a target sequence.

Recalling the definition of T as the sum of non-independent random variables, we obtain the following:

$$V(T) = \sum_{i=1}^N V(X_i) + 2 \sum_{i < j} Cov(X_i, X_j), \quad N = (l - k + 1), \tag{8}$$

$$\text{where } Cov(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j]. \tag{9}$$

Therefore, $V(T)$ can be represented by the variance co-variance matrix of the random variables X_1, \dots, X_N , as follows:

$$V = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \sigma_{13}^2 & \cdots & \sigma_{1k}^2 & 0 & \cdots & 0 & 0 \\ \sigma_{21}^2 & \sigma_{22}^2 & \sigma_{23}^2 & \cdots & \sigma_{2k}^2 & \sigma_{2(k+1)}^2 & 0 & \vdots & \vdots \\ \sigma_{31}^2 & \sigma_{32}^2 & \sigma_{33}^2 & \cdots & \sigma_{3k}^2 & \sigma_{3(k+1)}^2 & \sigma_{3(k+2)}^2 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & 0 \\ \sigma_{k1}^2 & \sigma_{k2}^2 & \sigma_{k3}^2 & \cdots & \bullet & \bullet & \bullet & \bullet & \bullet \\ 0 & \sigma_{(k+1)2}^2 & \sigma_{(k+1)3}^2 & \cdots & \bullet & \bullet & \bullet & \bullet & \bullet \\ \vdots & 0 & \sigma_{(k+2)3}^2 & \cdots & \bullet & \bullet & \bullet & \bullet & \bullet \\ \vdots & \vdots & \ddots & \ddots & \bullet & \bullet & \bullet & \bullet & \bullet \\ 0 & 0 & \cdots & 0 & \bullet & \bullet & \bullet & \bullet & \bullet \end{pmatrix}. \tag{10}$$

Here, V is an $N \times N$ square matrix where $N = (l - k + 1)$, the number of positions in λ that can be compared with a given k -mer for a match. The sum of all the elements in the matrix

is the variance of T , with the main diagonal elements being $V(X_i)$ and the off-diagonal elements representing the covariance terms of X_i and X_j . Note that the sum of the main diagonal elements is equal to the first term in equation (8) and the sum of the off-diagonal elements is the double sum term of $Cov(X_i, X_j)$, for $i < j$.

The $E[X_i]E[X_j]$ terms and non-zero elements in V

Let us first consider the terms of the variance co-variance matrix, V , where X_i and X_j represent non-overlapping segments of the target genome, noting that this condition arises when $|j - i| \geq k$. Under this condition of no overlap, X_i and X_j are independent. Therefore, $E(X_i X_j) = E[X_i]E[X_j]$. The corresponding elements of V then become zero, and we need only consider the $Cov(X_i, X_j)$ elements, where $|j - i| < k$.

We now reduce the double-sum into two terms, one for the expectation of the product of two variables and the other the product of the expectations of two variables. We also limit the summation over all non-zero terms in V , that is, where $0 < (j - i) < k$. These are terms where the positions of the k -mer associated with the X_i and X_j trials overlap. Equation (8) now becomes:

$$V(T) = N \left(\frac{1}{s}\right)^k \left[1 - \left(\frac{1}{s}\right)^k \right] + 2 \left[\sum_{0 < (j-i) < k} E[X_i X_j] - \sum_{0 < (j-i) < k} E[X_i]E[X_j] \right]. \tag{11}$$

We further reduce this equation by observing that the last term is a double sum of constant terms, since the expectation of $E[X_i] = \left(\frac{1}{s}\right)^k$ for any i . It is a constant for all our trial variables, dependent on only the number of symbols (s) and the length of $\zeta(k)$. This gives us:

$$V(T) = N \left(\frac{1}{s}\right)^k \left[1 - \left(\frac{1}{s}\right)^k \right] + 2 \left[\left(\sum_{0 < (j-i) < k} E[X_i X_j] \right) - D \left(\frac{1}{s}\right)^{2k} \right], \tag{12}$$

$$\text{where } D = \left[\frac{N(N-1)}{2} \right] - \left[\frac{(N-k+1)(N-k)}{2} \right], \quad N = (l-k+1). \quad (13)$$

Here, D represents the number of terms in the double sum in equation (11) or the non-zero terms in the upper triangle of \mathbf{V} .

Shift Symmetry in ζ

Here we will digress a little from our discussion of variance and the idea of matching a target sequence with a motif and define a property of a given motif or k -mer. We will use this property later in our continuing discussion of the variance of T . A given motif is said to have a shift symmetry of order r , if the first $k-r$ positions of the k -mer match the last $k-r$ positions. Note that a given k -mer can have zero or more such shift symmetries and may have as many as $k-1$ non-trivial shift symmetries. In addition, the sub-sequence associated with the symmetry may be longer than half the length of the k -mer, resulting in an overlap in the leading $k-r$ positions with the trailing $k-r$ positions.

More formally we define the set of shift symmetries Ψ for a given k -mer ζ , as the set of all natural numbers r , such that ζ has a shift symmetry of order r . The set of non-trivial symmetries Ψ is the set of all Natural numbers r , such that

$w_j = w_{j+r}$, $j = 1, \dots, (k-r)$ and $0 < r < k$ for the given k -mer. In set notation:

$$\Psi = \left\{ \forall r \in \mathbb{N} \mid w_j = w_{j+r}, \text{ for } j = 1, \dots, (k-r) \text{ and } 0 < r < k \right\}, \quad (14)$$

where w_j is the symbol in the j th position of the k -mer. Figure 3 below shows an example of a 9-mer with $\Psi = \{3, 6\}$. In addition, our definition of Ψ excludes the trivial symmetry where $r = 0$.

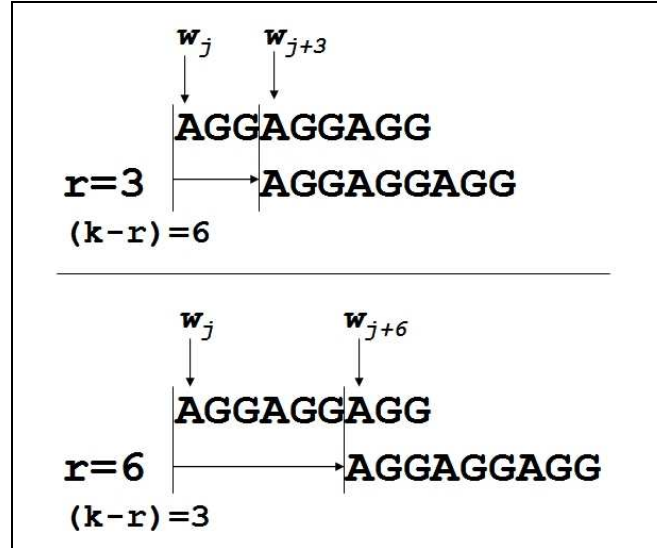


Figure 3. A 9-mer with a shift-symmetry of $\Psi = \{3, 6\}$.

The 9-mer in Figure 3, **AGGAGGAAG**, has two shift symmetries. The first symmetry where $r = 3$, has the sub-sequence **AGGAGG** occurring at the beginning and repeated at the end with an offset of 3 ($r = 3$). In addition, there is overlap in this symmetry for the given sub-sequence. For this same 9-mer, the symmetry corresponding to $r = 6$, has the sub-sequence **AGG** at the beginning repeated at the end.

Some other examples of 9-mers with non-trivial shift symmetries are:

$$\mathbf{GGGGGGGGG}, \Psi = \{1, 2, 3, 4, 5, 6, 7, 8\}; \mathbf{AGGTCAAGG}, \Psi = \{6\}$$

The $E[X_i X_j]$ terms and Shift Symmetry

Of particular interest is the variance of T defined on our sample space is dependant on the shift symmetries Ψ of the k -mer and more specifically the only terms dependant on Ψ are the $E[X_i X_j]$ terms.

We now show that the $E[X_i X_j]$ terms in equation (12) depend on the shift symmetry Ψ of the specific k -mer of interest. By definition, the expected value of the product of two Bernoulli trials of a k -mer is given as

$$E[X_i X_j] = \sum_{x_i} \sum_{x_j} x_i x_j p(x_i, x_j). \quad (15)$$

Here p is the joint probability function and x_i and x_j represent the values of our random variables, X_i and X_j , respectively. Expanding the sum on the right hand side of equation (15) gives us:

$$\sum_{x_i} \sum_{x_j} x_i x_j p(x_i, x_j) = (0 \cdot 0)p(0,0) + (0 \cdot 1)p(0,1) + (1 \cdot 0)p(1,0) + (1 \cdot 1)p(1,1). \quad (16)$$

The only non-zero term in equation (16) is the last term:

$$E[X_i X_j] = P(X_i = 1, X_j = 1) = p(1,1).$$

Figure 4 reveals that the only condition in which the k -mer will match the target sequence at both overlapping positions i , and j is when the motif has a shift symmetry of order r , that is, when $\Psi \neq \emptyset$ and $j - i = r \in \Psi$, otherwise $p(1,1)$ is zero.

For those cases when $p(1,1) \neq 0$, that is, when $j - i = r \in \Psi$, the probability is that the combined string consisting of the k positions of the k -mer followed by the last r positions of the k -mer match the $k + r$ positions in the target sequence beginning at i (Figure 4).

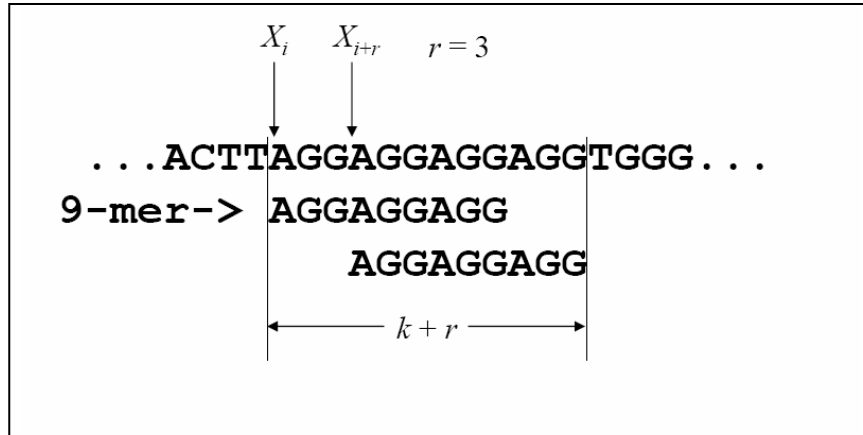


Figure 4. The joint probability, $p(1,1)$.

We now have the expression for the final terms for the variance of T .

$$E[X_i X_{i+r}] = p(1,1) = \begin{cases} \left(\frac{1}{s}\right)^{k+r} & \text{when } w_j = w_{j+r}, \quad j = 1, \dots, (k-r) \text{ and } r \in \Psi \\ 0 & r \notin \Psi \end{cases} \quad (17)$$

When a given k -mer has $\Psi = \emptyset$ (null set) the double sum in equation (18) will be zero.

However, when $\Psi \neq \emptyset$ and letting r take on the set of shift symmetry values for the given k -mer, we have:

$$2 \sum_{0 < (j-i) < k} E[X_i X_j] = 2 \sum_{r \in \Psi} A_r \left(\frac{1}{s}\right)^{k+r}, \quad \Psi = \{r_1, r_2, \dots, r_m\}, \quad (18)$$

where,

$$A_r = N - r, \quad r \in \Psi \quad (19)$$

A_r is the number of terms in the upper triangle of the variance co-variance matrix, \mathbf{V} , corresponding to the given shift symmetry element, r . Note that this will be a diagonal of elements σ_{ij}^2 where $0 < (j-i) = r$, for the given symmetry value r .

Note that on the left hand side of equation (18), we are limiting the double sum over the elements where there is overlap; the summation can now be further reduced by observing that the terms in the equation are zero for elements in Ψ that do not match their corresponding overlap positions (Equation 17). This reduction can be noted in the right-hand side of equation (18) where the sum is over $r \in \Psi$.

Conclusion

Expressions for the expectation and variance frequency of occurrence variable T for a specific k -mer in our sample space now become the following:

$$E[T] = N \left(\frac{1}{s} \right)^k, \quad N = (l - k + 1) \tag{20}$$

$$V(T) = N \left(\frac{1}{s} \right)^k \left[1 - \left(\frac{1}{s} \right)^k \right] + 2 \left[\left(\sum_{r \in \Psi} A_r \left(\frac{1}{s} \right)^{k+r} \right) - D \left(\frac{1}{s} \right)^{2k} \right] \tag{21}$$

$$D = \left[\frac{N(N-1)}{2} \right] - \left[\frac{(N-k+1)(N-k)}{2} \right] \tag{22}$$

$$A_r = N - r, \quad r \in \Psi \tag{23}$$

These expressions now allow us to formulate a Z test to evaluate the statistical significance of the size of an observed frequency of matches for a given k -mer in a specified target sequence such as a chromosome. It must be noted, however, that the above model has been developed with a number of simplifying assumptions such as equal probability of occurrence (uniform probability) of each of the four nucleotides. Each position in the

genome is also assumed to have values independent of the other positions (the uniform and independence assumption.) Other complexities needed to be incorporated in future studies include varying probabilities of occurrence of nucleotides along the length of the target sequence, the problem of negotiating boundaries between regions of differing probabilities, and degeneracy in the TFBS (where the match need not be perfect).

References

- Carroll, SB., JK Grenier, SD Weatherbee, 2005. *From DNA to diversity : molecular genetics and the evolution of animal design*, 2nd edn. Malden, MA: Blackwell Pub.
- Cvekl, A., Y Yang, BK Chauhan, K Cveklova, 2004. Regulation of gene expression by Pax6 in ocular cells: a case of tissue-preferred expression of crystallins in lens. *Int J Dev Biol*, **48**:829-44.
- Elemento, O., S Tavazoie, 2005. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol*, **6**:R18.
- Ettwiller, L., B Paten, M Souren, F Loosli, J Wittbrodt, E Birney, 2005. The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biol*, **6**:R104.
- Jones, NC., PA Pevzner, 2006. Comparative genomics reveals unusually long motifs in mammalian genomes. *Bioinformatics*, **22**:e236-42.
- Rombauts, S., K Florquin, M Lescot, K Marchal, P Rouze, Y van de Peer, 2003. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiol*, **132**:1162-76
- Xie, X., J Lu, EJ Kulbokas, TR Golub, V Mootha, K Lindblad-Toh, ES Lander, M Kellis, 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**:338-45.
- Xie, X., TS Mikkelsen, A Gnirke, K Lindblad-Toh, M Kellis, ES Lander, 2007. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A* **104**: 7145-7150.
- Xu, ZP., GF Saunders, 1997. Transcriptional regulation of the human PAX6 gene promoter. *J Biol Chem*, **272**:3430-6