Analytic combinatorics for bioinformatics I: applications to seeding methods

March 14, 2017

Abstract

Text of the abstract.

1 Introduction

Computing the best alignment between two sequences is carried out by dynamic programming. However, the running time of these algorithms scale $O(mn)$, where m and n are the sequence lengths, so it is only feasible for short sequences. For long sequences, heuristics are used for speed up, and the most popular belong to "seeding" methods.

This document is written primarily for readers without a strong background in analytic combinatorics

2 Weighted generating functions

Definition 1. *Let $\mathcal{A}$ be a set of combinatorial objects characterized by a size and a weight. The weighted generating function of $\mathcal{A}$ is defined as*

$$A(z) = \sum_{a \in A} w(a) z^{|a|}, \tag{1}$$

where $|a|$ and $w(a)$ denote the size and weight of an object a, respectively. This also defines a sequence $(a_k)_{k \geq 0}$ such that

$$A(z) = \sum_{k=0}^{\infty} a_k z^k. \tag{2}$$

By definition $a_k = \sum_{a \in A_k} w(a)$, where A_k is the class of objects of size k. The number a_k is called the total weight of objects of size k.

Note that in definition 1, the weight is a property of combinatorial objects and the total weight is a property of a classes of ojects. Expressions (1) and (2) are equivalent. Depending on the context, we will use one or the other.

The essence of analytic combinatorics is that some operations on combinatorial objects correspond to some operations on their generatintg function. If $A(z)$ and $B(Z)$ are the weighted generating functions of two mutually exclusive sets $\mathcal{A}$ and $\mathcal{B}$, the weighted generating function of $\mathcal{A} \cup \mathcal{B}$ is $A(z) + B(z)$, as appears immediately from expression (1). Size and weight can be extended to pairs of objects in $\mathcal{A} \times \mathcal{B}$ by defining $|(a, b)| = |a| + |b|$ and $w(a, b) = w(a)w(b)$. With this convention, the weighted generating function of $\mathcal{A} \times \mathcal{B}$ is $A(z)B(z)$, as shown by expression (1) once again

$$A(z)B(z) = \sum_{a \in \mathcal{A}} w(a) z^{|a|} \sum_{b \in \mathcal{B}} w(b) z^{|b|} = \sum_{(a,b) \in \mathcal{A} \times \mathcal{B}} w(a)w(b) z^{|a|+|b|}.$$

Example 1. *Assume that $\mathcal{A}$ contains a single object a of size 1 and of weight p. The weighted generating function of $\mathcal{A}$ is pz. The set $\mathcal{A}^2$ contains a single object (a, a) of size 2 and weight p^2. Its weighted generating function is $p^2 z^2 = pz \cdot pz$.*

The definition of size and weight can be further extended to any finite Cartesian product in the same way. The generating function of a cartesian product then comes as the product of their generating functions.

Example 2. *Following up on example 1, the set $\mathcal{A}^k$ contains a single object of size k and weight p^k, and its weighted generating function is $p^k z^k$. Since the sets $\mathcal{A}, \mathcal{A}^2, \mathcal{A}^3, \ldots$ are mutually exclusive, the weighted generating function of their union is*

$$pz + p^2 z^2 + p^3 z^3 \ldots$$

For any given k, $pz + p^2 z^2 + \ldots + p^k z^k = (pz - p^{k+1} z^{k+1})/(1 - z)$. If $|z| < 1/p$ the term $p^{k+1} z^{k+1}$ vanishes as k increases. So the weighted generating function is defined for $|z| < 1/p$ and is equal to

$$pz + p^2 z^2 + p^3 z^3 \ldots = \frac{pz}{1 - pz}.$$

Example 2 can be generalized. For any set $\mathcal{A}$, objects of $\mathcal{A}^+ = \cup_{k=1}^{\infty} \mathcal{A}^k$ are called nonempty sequences of objects of $\mathcal{A}$. By defining $\mathcal{A}^0$ as the set containg only ε, the "empty" object of size 0 and weight 1, we can also define $\mathcal{A}^* = \cup_{k=0}^{\infty} \mathcal{A}^k$ as the set of sequences of objects of $\mathcal{A}$.

Proposition 1. *Let $\mathcal{A}$ be a set with weighted generating function $A(z)$. The generating functions of $\mathcal{A}^+$ and $\mathcal{A}^*$ are defined for $|A(z)| < 1$ and are respectively equal to*

$$\frac{A(z)}{1 - A(z)}, \text{ and}$$
$$\frac{1}{1 - A(z)}.$$

Proof. For $k \geq 1$, the generating function of $\mathcal{A}^k$ is $A(z)^k$ and since the sets are mutually exclusive, the weighted generating function of their union $\mathcal{A}^+$ is $A(z) + A(z)^2 + \ldots = A(z)/(1 - A(z))$, provided $|A(z)| < 1$.

The generating function of $\mathcal{A}^0$ is 1 so the weighted generating function of $\mathcal{A}^*$ is $1 + A(z) + A(z)^2 + \ldots = 1/(1 - A(z))$, provided $|A(z)| < 1$. □

Remark 1. *These expressions are not defined for $A(z) = 1$, i.e. when $\mathcal{A}$ contains only the empty object. In other words, one cannot construct sequences of empty ojects. From here on, we will not state the conditions of definition of the generating functions. We will simply assume that $|z|$ is lower than the radius of convergence of the given expression.*

3 Exact seeding

Every sequencing technology makes occasional errors, so that the sequence of a read does not always correspond to the biological molecule that was sequenced. Errors can be inertions, deletions (together referred to as indels) or subsitutions.

Figure 1: **Structure of reads**. Here we consider sequencing reads that can have any type of error (insertions, deletions or substitutions). The reads have size k, and the errors are represented as grey squares. A read is composed of error-free intervals and error-only intervals. Note that a deletion is an error-only interval of size 0, so two error-free intervals can be contiguous (between position 3 and 4 in this example). However error-free intervals have size at least 1, so two error-only intervals cannot be contiguous.

Figure 1 gives a graphical representation. It is clear that a read can be segmented in intervals of different nature. Error-only intervals are consecutive nucleotides of the read that are all errors. Importantly, these intervals can have length 0 because of deletions. On the opposite, error-free intervals are consecutive nucleotides of the read that all correspond to the biological molecule being sequcned. Importantly, such intervals cannot have length 0, because there must be at least one correct nucleotide to call an interval error-free. This motivates the following combinatorial definition of reads.

Definition 2. *A read is an alternating sequence of error-only intervals and error-free intervals. The set of reads is denoted $\mathcal{R}$ and its weighted generating function is $R(z)$.*

There are obviously 2^k possible reads of size k, but they are not equally likely because the frequency of errors is different from 1/2 (the error rate of all modern technologies is lower). The weights will come in handy to record the

probability of ocurrence of reads. In what follows, weights can be thought of as frequencies of certain events, or as quantities proportional to these frequencies.

Proposition 2. *Denote the weighed generating function of error-only and of error-free intervals as $E(z)$ and $F(z)$, respectively. Reads have weighted generating function*

$$R(z) = \frac{\big(1 + E(z) - E(0)\big)^2 F(z)}{1 - E(z)F(z)} + E(z) - E(0) + 1. \tag{3}$$

Formula (3) is somewhat cumbersome because reads can neither start nor end with an error-only interval of size 0 (deletions cannot be happen before the reads starts or after it ends). We must proceed with care in order to count every read only once, and exclude those impossible configurations.

Proof. As we have seen in section 2, the weighted generating function of an error-only interval followed by an error-free interval is $E(z)F(z)$. We will call sequences of such pairs u-reads, and their weighted generating function is

$$U(z) = \frac{1}{1 - E(z)F(z)}.$$

Nonempty u-reads are the reads that start with an error-only interval and end with an error-free interval. A pitfall is that u-reads can start with an error-only interval of size 0, *i.e.* a deletion, which can never happen on real reads. We thus need to ensure that such reads are disallowed. Writing $E(z) = e_0 + e_1 z + e_2 z^2 + \ldots$, we see that the weighted generating function of error-only intervals of size greater than 0 is $E(z) = e_1 z + e_2 z^2 + \ldots = E^+(z) = E(z) - E(0)$.

We now proceed by partitioning the reads on their first and last intervals. We call type (i) reads those that start with an error-free interval and end with an error-free interval. They are obtained by concatenating an error-free interval and a u-read, so their weighted generating function is $F(z)U(z)$.

Type (ii) reads are those that start with an error-free interval and end with a nonempty error-only interval. They are obtained by concatenating a type (i) read and a nonempty error-only interval so their weighted generating function is $F(z)U(z)E^+(z)$.

Type (iii) reads those that start with a nonempty error-only interval and end with an error-free interval. They are obtained by concatenating a nonempty error-only interval, an error-free interval, and a u-read, so their weighted generating function is $E^+(z)F(z)U(z)$.

Type (iv) reads those that start with a nonempty error-only interval and end with a nonempty error-only interval. They are obtained either by concatenating a type (iii) and an nonempty error-only interval or as just a nonempty error-only interval, so their weighted generating function is $E^+(z)F(z)U(z)E^+(z) + E^+(z)$.

The only missing read is ε, the read of size 0 and whose weighted generating function is 1.

Summing the weighted generarting functions of reads of types (i), (ii), (iii), (iv) and of ε, we obtain expression (3). □

Remark 2. *In the proof of proposition 2, we did not need to add a sparate term for reads that consists of just an error-free interval, because they are a type (i) reads (the concatenation of an error-free interval and the empty u-read).*

Error-free intervals are important for many analyses. For instance, when the read has to be aligned to a reference genome, the search space is usually reduced by searching short sequences with perfect identity, called "seeds". The success of this approach depends on the longest error-free interval (assuming that sequencing errors are the only differences with the reference genome). If all the error-free intervals of the read are shorter than the seeds, then no hit will be found. It is thus useful to study the distribution of the longest error-free interval of a read.

Definition 3. *An exact d-seed is an error-free interval of size at least d.*

We will consider that the integer d is fixed, and we refer to d-seeds as seeds most of the time. We will also call "seedless" the reads that contain no seed. This qualification also depends on a certain value of d, but the context will never be ambiguous.

Proposition 3. *Denote the weighted generating fuctions of error-only intervals as $E(z)$ and denote as $F_d(z)$ the weighted generating function of error-free intervals of size less than d. The weighted generating function of the set of seedless reads is*

$$S(z) = \frac{(1 + E(z) - E(0))^2 F_d(z)}{1 - E(z)F_d(z)} + E(z) - E(0) + 1. \tag{4}$$

Proof. Proceed as in proposition 2. □

Note that if the weighted generating function of error-free intervals $F(z)$ is equal to $f_1 z + f_2 z^2 + \ldots$, then $F_d(z) = f_1 z + f_2 z^2 + \ldots + f_{d-1} z^{d-1}$.

The probability that the seeding approach will fail for a read of length k is the total weight of seedless reads of size k, divided by the total weight of reads of size k. Recalling expression (2), this probability is equal to the coefficients of z^k in $S(z)$, divided by the coefficient of z^k in $R(z)$.

To find this probability, we thus need to define $E(z)$ and $F(z)$, write the expressions of $R(z)$ and $S(z)$, and extract their coefficients. This, in a nutshell, is the standard analytic combinatorics approach. We will see below that in general we cannot obtain exact solutions, but we can get very accurate asymptotic approximations.

3.1 Substitution only

One of the simplest and yet useful models to consider is that errors consist of substitutions only, and that they occur with the same probability p for every decoded nucleotide. This describes reasonably well the error model of the Illumina platforms, where p is around 0.01.

According to the model, the weight of every error-free nucleotide is $1-p=q$, and the weight of every error is p. The weighted generating functions of error-free intervals and error-only intervals are thus respectively

$$E(z) = pz + p^2z^2 + \ldots = \frac{pz}{1-pz}, \text{ and}$$
$$F(z) = qz + q^2z^2 + \ldots = \frac{qz}{1-qz}.$$

Note that $E(0) = 0$. Substituting the expressions above in (3), the weighted generating function of reads has a relatively simple expression, namely

$$R(z) = \frac{\big(1+E(z)\big)^2 F(z)}{1-E(z)F(z)} + E(z) + 1 = \frac{1}{1-z}. \tag{5}$$

Setting $p = 1$ in example 2 shows that $z/(1-z) = z + z^2 + z^3 + \ldots$, so the coefficient of z^k is equal to 1 for all $k \geq 1$.

We now write the generating function of seedless reads. For this, we only need to truncate $F(z)$ at z^{d-1} as shown below

$$F_d(z) = qz + (qz)^2 + \ldots + (qz)^{d-1}.$$

Substituting this in expression (4), we now obtain the weighted generating function of seedless reads as

$$S(z) = \frac{\big(1+E(z)\big)^2 F_d(z)}{1-E(z)F_d(z)} + E(z) + 1 = \frac{1+qz+\ldots+(qz)^{d-1}}{1-pz(1+qz+\ldots+(qz)^{d-1})}. \tag{6}$$

Remark 3. *Expression (6) suggests an alternative definition of seedless reads. $S(z)$ is the product of $1+qz+\ldots+(qz)^{d-1}$ and $1/\big(1-pz(1+qz+\ldots+(qz)^{d-1})\big)$. Those weighted generating functions represent error-free intervals of size less than d, and sequences of substitutions followed by error-free intervals of size less than d. In both cases, the error-free intervals are possibly empty. In other words, there is a unique decomposition of seedless reads in segments that contain a single error at the beginning (with the possible exception of the first segment) and whose size is at most d (including the error).*

This alternative definition is not only valid for seedless reads, but also for other reads. Expression (5) just does not make it as explicit as expression (6).

The task is now to extract the coefficient of z^k in the series expansion of $S(z)$. These coefficients have no known explicit expression, but one of the crown jewels of analytic combinatorics is that we can approximate them very accurately from the weighted generating function (6). In order to show how to do this, we will first prove a more general result.

Proposition 4. *If a weighted generating function is the ratio of two polynomials $P(z)/Q(z)$, then the coefficient of z^k in its series expansion is asymptotically equivalent to*

$$-\frac{P(z_1)}{Q^{(m)}(z_1)}(k+m-1)(k+m-2)\dots(k+1)\frac{m}{z_1^{k+1}}, \tag{7}$$

where z_1 is the root of smallest modulus of Q, m is its multiplicity, and $Q^{(m)}$ is the m-th derivative of Q.

The roots of Q are called the "singularities" of the rational function (they are values where the function is not defined). Proposition 4 says that asymptotic behavior of the coefficients is dictated by the singularity of smallest modulus, which we will refer to as the "main singularity" of the function. We will first prove a lemma that will also be important to improve the asymptotic approximation of the coefficients later on.

Lemma 1. *Let n be a positive integer and a be a complex number. For $|z| < a$ we have*

$$\frac{1}{(1-z/a)^n} = \sum_{k=0}^{\infty} \binom{k+n-1}{n-1} \frac{z^k}{a^k}. \tag{8}$$

Proof. Example 2 showed that (8) holds for $n = 1$. Proceed by induction and assume that (8) holds for n. Differentiate both sides of (8) to obtain

$$\frac{n}{a}\frac{1}{(1-z/a)^{n+1}} = \sum_{k=0}^{\infty} k\binom{k+n-1}{n-1}\frac{z^{k-1}}{a^k}.$$

Recalling $\frac{k}{n}\binom{k+n-1}{n-1} = \binom{k-1+n}{n}$ and rearranging the terms, we obtain the same expression as (8) where n is replaced by $n+1$. □

We now prove proposition 4.

Proof. Let $z_1, z_2, \dots, z_n$ be the complex roots of Q sorted by increasing order of modulus, and $m_1, m_2, \dots, m_n$ be their multiplicity. It is well known that there exists complex numbers $s_{i,j}$ such that $P(z)/Q(z)$ can be written as

$$\sum_{j=1}^{n}\sum_{i=1}^{m_i}\frac{s_{i,j}}{(z-z_j)^i} = \sum_{j=1}^{n}\sum_{i=1}^{m_i}\frac{s_{i,j}/z_j}{(1-z/z_j)^i} \tag{9}$$

Here we assumed without loss of generality that the degree of P is lower than the degree of Q. If this is not the case, the decomposition above also contains a polynomial whose coefficients are eventually null and which do not affect the asymptotics. Using lemma 1, we see that the coefficient of z^k in the series expansion of $S(z)$ is equal to

$$-\sum_{j=1}^{n}\sum_{i=1}^{m_i}\binom{k+i-1}{i-1}\frac{s_{i,j}}{z_j^{k+1}}.$$

Since z_1 is the root with smallest modulus and that its multiplicity is m_1, the leading term of this sum is

$$-\binom{k+m_1-1}{m_1-1}\frac{s_{m_1,1}}{z_1^{k+1}}.$$

To find the value of $s_{m_1,1}$, we factorize $Q(z)$ as $(z-z_1)^{m_1}Q_1(z)$, which is possible because z_1 is a root of Q, and we write

$$\frac{P(z)}{Q(z)} = \frac{P(z)}{(z-z_1)^{m_1}Q_1(z)} = \frac{s_{m_1,1}}{(z-z_1)^{m_1}} + \varepsilon(z).$$

Multiplying both sides by $(z-z_1)^{m_1}$ and setting $z = z_1$ we obtain the expression

$$\frac{P(z_1)}{Q_1(z_1)} = s_{m_1,1}.$$

Differentiating m_1 times the equality $Q(z) = (z-z_1)^{m_1}Q_1(z)$ shows that $Q^{(m_1)}(z_1) = m_1!Q_1(z_1)$, so $s_{m_1,1} = m_1!P(z_1)/Q^{(m)}(z_1)$, which concludes the proof. □

Remark 4. *Expression (9) is not an approximation, it is the exact value of the coefficient. By keeping more than one leading term, we can obtain more accurate estimates, and by keeping all the terms we obtain the exact number.*

Returning to the case of seedless reads, proposition 4 invites us to find the roots of the singularities of $S(z)$. The left panel of Figure 2 shows the values of the denominator of $S(z)$ around 0 for $p = .01$ and $d = 17$. The root with smallest modulus is clearly visible, and by the angle of the curve at $Q(z) = 0$, it looks like this is a simple root (*i.e.* that is has multiplicity 1).

The remaining singularities of $S(z)$ are complex and they seem to be evenly spaced on the same circle, as can be seen on the right panel of Figure 2. This is only a visual impression. In fact the singularities are not exactly on the same circle and their rotation angles are not exactly regular.

Proposition 5. *S has only one real positive singularity, which is the main singularity.*

Proof. Write $S(z) = P(z)/Q(z)$ and search the roots of Q. Let x denote a real number. Since $Q(1) = q^d > 0$ and $\lim_{x\to\infty} Q(x) = -\infty$, Q vanishes for a real number greater than 1. □

We can now have all the tools to approximate the coefficients of $S(z)$ using proposition 4 and thus obtain the approximate probability that a read contains no seed.

Proposition 6. *The probability that a read of of size k is seedless is asymptotically equivalent to*

$$\frac{C}{z_1^{k+2}},$$

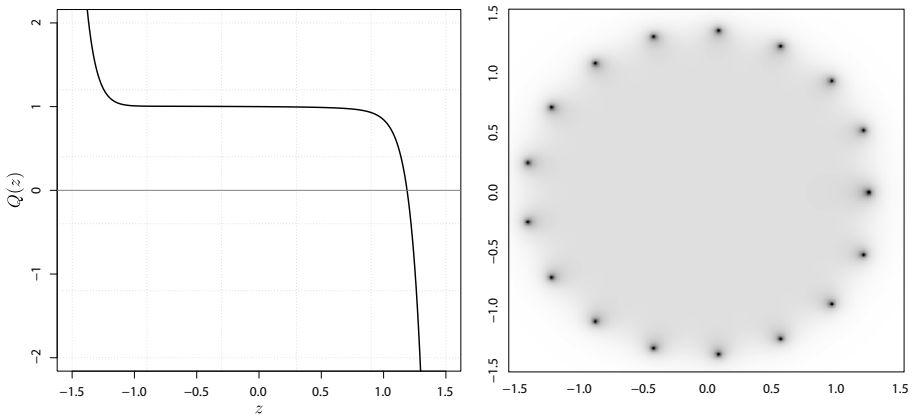

Figure 2: **Singularities of** S. $Q(z)$ denotes the denominator of $S(z)$ from expression (6). *Left*: real representation of Q. The bold line represents the value of $Q(z)$ for $p = 0.01$ and $d = 17$. The root with smallest modulus appears in the interval $(1, 1.5)$. *Right*: complex representation of Q. Shown is the complex plane around the origin. The darker the dot, the higher the value of $1/Q(z)$ at the corresponding value of z. Sixteen singularities of S lie close to a circle. The remaining seventeenth is the one shown on the left panel and it it is the main singularity because it lies slightly closer to the origin.

where z_1 is the only real positive root of the denominator of $S(z)$, i.e. the root of $1 - pz(1 + qz + \ldots + (qz)^{d-1})$ and

$$C = \frac{(1 - qz_1)^2}{p^2\,(1 + d(qz_1)^{d+1} - (d + 1)(qz_1)^d)}. \tag{10}$$

Proof. Apply proposition 4 and proposition 5 to $S(z)$ and also use the fact that $1 + qz_1 + \ldots + (qz_1)^{d-1} = 1/pz_1$. □

Remark 5. *Since $z_1 > 1$, the probability decreases exponentially as k increases.*

We now illustrate proposition 6 with a concrete example explaining how the calculations are done in practice.

Example 3. *Let us approximate the probability that a read of size $k = 50$ is seedless for $d = 17$ and for a substitution rate $p = 0.01$. To find the main singularity of S, we need to solve the equation $1 - pz(1 + qz + \ldots + (qz)^{16}) = 0$. For convenience, we rewrite it as $1 - pz(1 - (qz)^{17})/(1 - qz) = 0$ and solve it numerically with the Newton-Raphson method, yielding $z_1 \approx 1.026886$. Substituting this value in (10) yields $C \approx 1.433681$, so the probability that a read contains no seed is approximately $1.433681/1.026886^{52} \approx 0.3608321$. For comparison, performing 1,000,000 random simulations gives an estimate in the range...*

The demonstration of proposition 4 shows that the analytic combinatorics estimate converges exponentially as k increases because the main singularity has multiplicity 1. The numerical estimates are thus close to the exact values, as illustrated in example 3. Figure 3 shows this on more comprehensive example with longer reads and higher probability of error p.

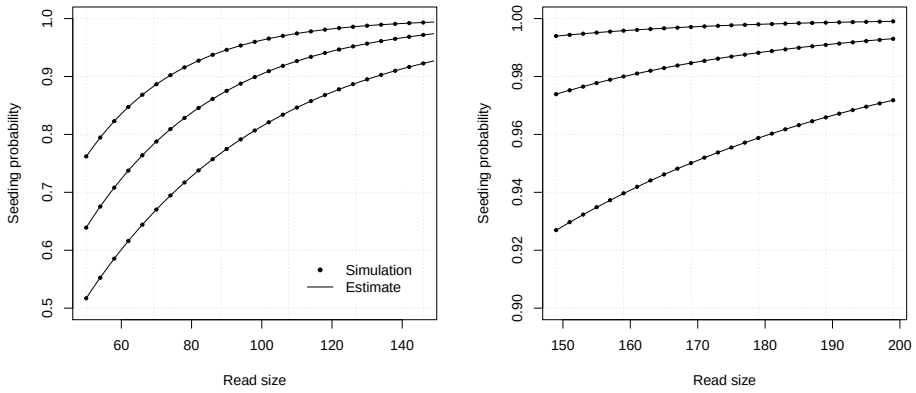

Figure 3: **Accuracy of the estimates**. The analytic combinatorics estimates are compared to random simulations for $d = 17$ and $p = 0.08$, $p = 0.10$ or $p = 0.12$. Shown on both panels are the probablities that a read of given size contains a seed, either estimated by 1,000,000 random simulations (dots), or by the method described above (lines). The analytic combinatorics estimates are slightly under the target values for large values of k, but overall the errors are small (the difference never exceeds 0.00028 on the examples shown here).

Overall, the analytic combinatorics estimates are close to the exact values. Figure 3 also shows that relatively small changes in the prboability of error have a large influence on the probability of that the read contains a seed in the depicted range of read sizes.

3.2 Substitutions and deletions

The uniform substitution model does not describe all sequencing technologies. For instance, long read technologies often have bursts of insertions and deletions, with typical frequencies that differ from substitutions. In order to model more complex behaviors, we need to distinguish the different types of errors.

To not jump too fast into the problem, we will first focus on a fictional case where errors can be deletions or susbtitutions. As in the case of uniform substitutions, we assume that every nucleotide call is false with a probability p and true with a probability $1 - p = q$. But here, we also assume the "space" between consecutive nucleotides can contain a deletion with probability δ.

With this formalism, the weighted generating function of an error-only interval of size 0 is simply δ. Whether the "spaces" between the nucleotides of error-only intervals of size $k > 0$ contain deletions or not is irrelevant. It also does not matter whether the interval starts or ends with a deletion, so the weighted generating function of error-only interval of size $k > 1$ is $p^k z^k$. Finally, the $k-1$ "spaces" between the nucleotides of an error-free interval of size k must not contain any deletion, so the weighted generating function of error-free intervals of size $k > 1$ is $(1-\delta)^{k-1}(pz)^k$. In summary $E(z) = \delta + pz + (pz)^2 + \ldots = \delta + pz/(1-pz)$ and $F(z) = qz + (1-\delta)(qz)^2 + (1-\delta)^2(qz)^3 + \ldots = qz/(1-(1-\delta)qz)$.

Substituting these values in equation (3), the weighted generating function of reads appears as

$$R(z) = \frac{\big(1 + E(z) - E(0)\big)^2 F(z)}{1 - E(z)F(z)} + E(z) - E(0) + 1 = \frac{1}{1-z}.$$

The weighted generating function of the reads is the same as under the constant substitution model, and the total weights are all equal to 1.

The weighted generating function of error-free reads of size less than d is $F_d(z) = qz + (1-\delta)(qz)^2 + \ldots + (1-\delta)^{d-2}(qz)^{d-1}$. Substituting this value in equation (4), we obtain the weighted generating function of seedless reads as

$$\begin{aligned} S(z) &= \frac{\big(1 + E(z) - E(0)\big)^2 F_d(z)}{1 - E(z)F_d(z)} + E(z) - E(0) + 1 \\ &= \frac{1 + (1-\delta)\big(qz + (1-\delta)(qz)^2 + \ldots + (1-\delta)^{d-2}(qz)^{d-1}\big)}{1 - pz - \big(pz(1-\delta) + \delta\big)\big(qz + (1-\delta)(qz)^2 + \ldots + (1-\delta)^{d-2}(qz)^{d-1}\big)}. \end{aligned}$$

Applying proposition 4 to this expression, we obtain the following proposition.

Proposition 7. *The probability that a read of size k is seedless is asymptotically equivalent to*

$$\frac{C}{z_1^k},$$

where z_1 is the only real positive root of the denominator of $S(z)$, i.e. the root of $1 - pz - \big(pz(1-\delta) + \delta\big)\big(qz + (1-\delta)(qz)^2 + \ldots + (1-\delta)^{d-2}(qz)^{d-1}\big)$, and

$$\begin{aligned} C &= \frac{\big(1 - (1-\delta)(1-p)z\big)^2}{\big((p+q\delta)z - \gamma^*(1-\delta)^{d-1}(qz)^d\big)\big(\delta + (1-\delta)pz\big)}, \\ \gamma^* &= d\delta - (1-\delta)\big((d-1)\delta - p((d-1)\delta + d + 1)\big)z - d(1-\delta)^2 pqz^2. \end{aligned} \tag{11}$$

Example 4. *Let us approximate the probablity that a read of size $k = 100$ is seedless for $d = 17$ and for substitution and deletion rates $p = 0.05$ and $\delta = 0.15$, respectively. In order to find the main singularity of S, we need to solve the*

equation $1 - pz - \big(pz(1-\delta)+\delta\big)\big(qz + (1-\delta)(qz)^2 + \ldots + (1-\delta)^{15}(qz)^{16}\big) = 0$. *We rewrite it as* $1-pz-\big(pz(1-\delta)+\delta\big)(qz-(1-\delta)^{16}(qz)^{17})/(1-(1-\delta)qz) = 0$ *and solve it numerically with the Newton-Raphson method, yielding* $z_1 \approx 1.006705$. *Substituting this value in (11) yields* $C \approx 1.088876$, *so the probability that a read contains no seed is approximately* $1.088876/1.006705^{100} \approx 0.558141$. *For comparison, performing 1,000,000 random simulations gives an estimate in the range...*

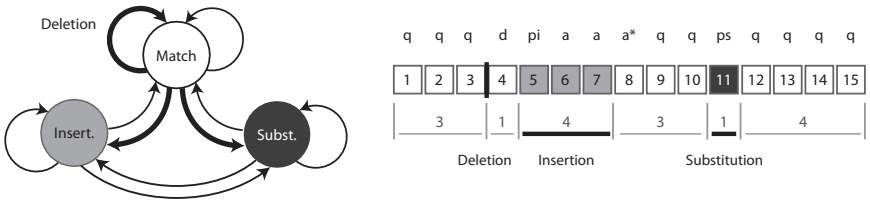

Figure 4: **Complex error models**. Text.

4 Inexact seeding

We now consider a more challenging problem. The ongoing development of algorithms and data structures allows us to use...

Here we assume the constant substitution rate model.

Definition 4. *An inexact d-seed is an interval of size at least d that contains at most a bustitution and no other error. An exact d-seed is an inexact d-seed, but the converse is not true.*

In other words, an inexact seed is either an error-free interval or the concatenation of two error-free intervals and a substitution, each with total size at least d. In this context, a "seedless" read will designate a read that does not contain any inexact seed.

As in the case of exact seeding, we will give a construction of seedless reads, find the corresponding weighted generating function and extract the asymptotic behavior of the coefficients. The main difference with the previous case is that seedless reads are not simple sequences of intervals. The error-free intervals on either side of the substitution are linked by the constraint that their total size cannot be larger than $d-2$ (the read must not contain a inexact d-seed, and the substitution has size 1).

Following remark 3, it is more convenient to segment seedless reads in a slightly different way.

The easiest way to encode this dependence is to design a transition matrix M, such that the entry at position (i, j) is the weighted generating function of a substitution and an error-free intervals of size j on the right-hand-side of an error-free interval of size $i+1$. For instance, the term at position $(1, 1)$ is always

qz (provided $d \geq 3$). This corresponds to appending an error-free interval of size 1 to an error-free interval and a substitution.

The general format of the matrix is

$$M = pz \begin{bmatrix} 1 & qz & (qz)^2 & (qz)^3 & \dots & (qz)^{d-3} & (qz)^{d-2} \\ 1 & qz & (qz)^2 & (qz)^3 & \dots & (qz)^{d-3} & 0 \\ 1 & qz & (qz)^2 & (qz)^3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & qz & (qz)^2 & 0 & \dots & 0 & 0 \\ 1 & qz & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix} \tag{12}$$

The term at position (i, j) in the matrix M^n is the generating function of reads with $n + 1$ error-free intervals (possibly empty), whose first interval has size $i - 1$ and whose last interval has size $j - 1$.

The vector $pz(1, qz, (qz)^2, \dots, (qz)^{d-2})$.