Generating Functions in Computational Biology

Mireille Régnier
Algorithms Project — Inria

March 3, 97

[summary by Mireille Régnier]

Abstract

We present a few enumeration problems that arose in computational biology. We point out on several examples how symbolic enumeration methods allow for simplifying and extending previous results. We also present some asymptotics.

1. Secondary Structures

As a first example, we enumerate here combinatorial structures associated to RNA sequences, e.g., secondary structures, hairpins, cloverleaves, ...). This study has been started in [11, 13, 8, 4, 12] with inductions on the size of the RNA sequences. We show here that symbolic enumeration methods allow to find directly equations on generating functions and extend previous results. More precisely, these inductions may deeply depend on the values of the parameters involved such as the minimal size h of the helices, the minimal size b of the loops, ... Due to the number of initial conditions, the recurrence relations may become very intricate. We avoid this unnecessary step.

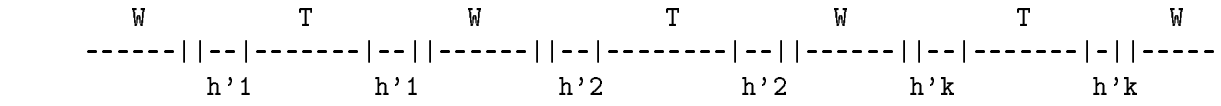
A secondary structure is a set of *helices*, i.e., paired subsequences of the same size. Two paired subsequences are separated by an embedded secondary structure or by a non-paired subsequence, called a *loop*. Secondary structures satisfy three additional conditions. First, the helices must have a minimal size, h. Second, the loops must have a minimal size, b. Third, two helices cannot overlap.

We now count the number of different secondary structures that may be built on the RNA sequences of a given size:

Theorem 1. *Let $\mathcal{S}_n^{[h,b]}$ be the set of RNA secondary structures of size n, where the minimal helix size and loop size are h and b. The generating function $S^{[h,b]}(z)$satisfies the* second degree *equation:*

$$(S^{[h,b]})^2(z)z^{2h} + S^{[h,b]}(z)[(z-1)(1-z^2+z^{2h}) - z^{2h}\frac{1-z^b}{1-z}] + 1 - z^2 + z^{2h} = 0. \quad (1)$$

We consider the external helices, and get $S^{[h,b]}$ from the following decomposition:



where $\mathcal{W}$ is the set of of non-paired subsequences, and $\mathcal{T}^{[h,b]}$ the set of secondary structures that do not start or end with a pairing. Otherwise, such a pairing would be part of the external helix. In

this scheme, we assume there exist k external helices of sizes $h'_1, \ldots, h'_k$. Two external helices are separated by non-paired positions, which is coded by language $\mathcal{W}$. We get the set decomposition:

$$\mathcal{S}^{[h,b]} = \mathcal{W} \times \cup_{k\geq 0}[\mathcal{A}^{[h]} \times \mathcal{T}^{[h,b]} \times \mathcal{W}]^k \tag{2}$$

where $\mathcal{A}^{[h]}$ is the set of couples of subsequences of the same size $h' \geq h$. One can prove a simple relation between $\mathcal{T}^{[h,b]}$ and $\mathcal{S}^{[h,b]}$. Namely:

$$\mathcal{S}^{[h,b]} = \mathcal{A}^{[h]} \times \mathcal{T}^{[h,b]} + \mathcal{T}^{[h,b]} + \mathcal{Y}^{[b]} \tag{3}$$

where $\mathcal{Y}^{[b]}$ counts sequences of length smaller than b (no structure is possible). I.e. $Y^{[b]}(z) = \sum_{i=0}^{b-1} z^i$. We plug (3) into (2) and, applying translation rules, we get (1) after simplification.

Remark. For $h = 1$ and $b = 1$, we get the equation in [10].

The next theorem directly follows from Darboux's theorem and equation (1).

Theorem 2. *When $n \longrightarrow \infty$,*

$$S_n^{[h,b]} \sim \frac{1}{4\sqrt{\pi n^3}}[-\rho\Delta'_{h,b}(\rho)]^{\frac{1}{2}} \cdot \rho_{h,b}^{-(n+2h)} \tag{4}$$

where ρ is the smallest positive root of the discriminant $\Delta_{h,b}(z)$ of (1).

Remark. The location of the roots is discussed in [9]. Notably, it is proven that $1/\rho_{h,b}$ decreases when h and b increase and that, for any h and b, $\triangle_{h,b}(z)$ has one root $\rho_{h,b}$ in $]0,1[$ and that $\Delta_{h,b}(z)$ has one root in $]0,1/2[$. It follows that $S_n^{[h,b]}$ grows exponentially and that $S_n^{[1,b]} \geq 2^n$. Numerical values of ρ have been computed using Maple.

One also derives functional equations satisfied by the generating functions of specific structures: the so-called *hairpins* and *cloverleaves*. Asymptotics follow similarly.

2. Counting alignments

2.1. Language decomposition.

Definition 1. An alignment of k sequences of size $(n_i)_{i=1,k}$ is a sequence of k-tuples

$$(\alpha_1^{[j]}, \ldots, \alpha_i^{[j]}, \ldots, \alpha_k^{[j]})$$

such that:

- each subsequence $(\alpha_i^{[j]})$ is increasing from 0 to n_i;
- two successive k-tuples are not equal.

An aligned position is a an integer j such that

$$\alpha_i^{[j]} = \alpha_i^{[j-1]} + 1, \qquad i = 1, \ldots, k.$$

A b-block is a subsequence of aligned positions of length at least b.

The number of k-alignments is counted in [2]. A first generalization is proposed by [3], for $k = 2$. Blocks of size below some threshold b are considered as non-significant, and one eliminates all alignments containing such small blocks. Nevertheless, the authors still count separately non-aligned letters. E.g. $\begin{smallmatrix} C & \\ - & G \end{smallmatrix} \neq \begin{smallmatrix} - & C \\ G & - \end{smallmatrix}$. We extend this counting for any $k \geq 2$.

A second generalization is proposed in [11] for $k = 2$ and $b = 1$. One identifies $\begin{smallmatrix} C & \\ - & G \end{smallmatrix}$ and $\begin{smallmatrix} - & C \\ G & - \end{smallmatrix}$. We extend it for matching blocks of size b greater than 1. The motivation is the following. In [3], one considers only matching blocks of size at least b as significant. It follows that the difference

between $\begin{smallmatrix} C & - \\ - & G \end{smallmatrix}$ and $\begin{smallmatrix} - & C \\ G & - \end{smallmatrix}$ should be considered as non significant. Hence, in this section, we identify all alignments that differ by a set of positions without a b matching block. When $b = 1$ and $k = 2$, this is the generalization described in [11].

Theorem 3. *Let $f(n_1, \ldots, n_k)$ be the number of alignments between k sequences of lengths $n_1, \ldots, n_k$. The associated multivariate generating function is, in the ordered case:*

$$\phi(z_1, \ldots, z_k) = \frac{1 - t + t^b}{1 - s(1 - t + t^b) - t} \tag{5}$$

where $t = \prod_{i=1}^{k} z_i$ and $s = \sum_{i=1}^{k} z_i$. The bivariate generating function is, in the ordered case:

$$\phi(z_1, \ldots, z_k) = (1 - t + t^b)\frac{1}{1 - (p-1)(1 - t + t^b) - t} \tag{6}$$

where $p = \prod_{i=1}^{k}(1 - z_i)$.

When $b = 1$ and $k = 2$, this simplifies into $\phi(z_1, z_2) = 1/[1 - (z_1 + z_2)]$. It follows that $g_1(n, m) = \binom{n+m}{n}$ which is the result proved in [3] by a combinatorial approach.

The proof relies on the derivation of a coding language for sequences $f(n_1, \ldots, n_k)$. For example, one proves, for $b = 1$, that $\mathcal{L}_1 = [\prod_{i=1}^{k}(1 + Z_i) - 1]^*$ [2]. Let us build an alignment from left to right. In each position, one may choose, for each sequence, either to align one character or to introduce a gap. This contributes either by Z_i or by 1, and we get, for independent choices, $\prod_{i=1}^{k}(1 + Z_i)$. The only choice to be excluded is the choice of a gap in all sequences, which is associated to $1^k = 1$. Hence, we get Equation (5). It is worth noticing this is a bivariate function of s and t.

2.2. **Asymptotics.** A possible, and usual, assumption is that aligned sequences have the same size. Hence, one is interested in

$$f(n) = [z_1^n \cdots z_k^n]\phi(z_1, \ldots, z_k).$$

The generating function $\sum_n f(n)z^n$ is called the *diagonal* of the generating function $\phi(z_1, \ldots, z_k)$. One can find general results on diagonals in [1, 5]. We provide here a simplified scheme of the approach in the particular case of the alignment of two sequences. We prove:

Theorem 4. *Let us define:*

$$\Delta_1(t) = (1 - t)^2 - 4t(1 - t + t^b)$$
$$\Delta_2(t) = (1 - t^2 + t^{b+1}) - 4t(1 - t + t^b).$$

We have $f(n) \sim \alpha n^{-1/2}\rho^{-n}$ where ρ is the (positive) root of smallest modulus of $\Delta_1(t)$ (respectively $\Delta_2(t)$) in the non-ordered (respectively ordered) case.

Proof. When $k = 2$, s and p depend only on two variables, t and z_1. Namely, one has: $s = (z_1 + t/z_1)$ and $p = 1 + t + z_1 + t/z_1$. In both cases, one has:

$$F(t) = \sum_n f(n)t^n = [z_1^0]f(z_1, t/z_1).$$

Applying the Cauchy formula, we get:

$$F(t) = \frac{1}{2i\pi}\int \frac{f(z_1, t/z_1)}{z_1}\, dz_1 = \frac{1}{2i\pi}\int \frac{\psi(t)}{P(t, z_1)}\, dz_1$$

where $P(t, x)$ is a polynomial in t and x of degree 2 with respect to the second variable x. One proves that the discriminant of $P(t, x)$ with respect to x is $\Delta_1(t)$ (respectively $\Delta_2(t)$) in the non-ordered (respectively ordered) case. We first compute the integral. Then, applying Darboux's

theorem, we get that $f(n)/\rho^n$ has a polynomial growth, where ρ is the smallest positive root of $\delta_1(t)$ (respectively $\Delta_2(t)$). Darboux's theorem also provides the dominating term of $f(n)/\rho^n$. We refer the reader to [11] where this term is explicitly given for the non-ordered case. □

3. Miscellaneous problems

Many other parameters of interest to biologists can be studied through a generating function approach. One can cite the longest runs [11] or filtration methods such as the statistical distance [6]. This talk presented new results for the statistical distance in a non-uniform probability model. Finally, statistics for the number of occurrences of a given set of words is also of great interest. We presented a generating function approach [7]. We proved the limiting distribution is asymptotically normal and provided formulæ to compute the moments or the probability of occurrences in the finite range, for Bernoulli and Markov models.

References

[1] Furstenberg (Harry). – Algebraic functions over finite fields. *Journal of Algebra*, vol. 7, 1967, pp. 271–277.

[2] Griggs (J. R.), Hanlon (P.), Odlyzko (A. M.), and Waterman (M. S.). – On the number of alignments of k sequences. *Graphs and Combinatorics*, vol. 6, n° 2, 1990, pp. 133–146.

[3] Griggs (Jerrold R.), Hanlon (Philip J.), and Waterman (Michael S.). – Sequence alignments with matched sections. *SIAM Journal on Algebraic Discrete Methods*, vol. 7, n° 4, 1986, pp. 604–608.

[4] Howell (J. A.), Smith (T. F.), and Waterman (M. S.). – Computation of generating functions for biological molecules. *SIAM Journal on Applied Mathematics*, vol. 39, n° 1, 1980, pp. 119–133.

[5] Litow (B.) and Dumas (Ph.). – Additive cellular automata and algebraic series. *Theoretical Computer Science*, vol. 119, n° 2, October 1993, pp. 345–354.

[6] Pevzner (P. A.). – Statistical distance between texts and filtration methods in sequence comparison. *CABIOS*, vol. 8, n° 2, 1992, pp. 121–127.

[7] Régnier (M.) and Szpankowski (W.). – On the approximate pattern occurences in a text, 1997. In Proceeding SEQUENCE'97, Positano.

[8] Stein (P. R.) and Waterman (M. S.). – On some new sequences generalizing the Catalan and Motzkin numbers. *Discrete Mathematics*, vol. 26, n° 3, 1978, pp. 261–272.

[9] Tahi (F.). – *Méthodes formelles d'analyse des séquences de nucléotides.* – Thèse de 3e cycle, Université de Paris XI, Orsay, 1997. 155 pages.

[10] Viennot (X. G.) and Vauchaussade de Chaumont (M.). – Enumeration of RNA's secondary structures by complexity. In Capasso (V.), Grosso (E.), and Paven-Fontana (S. L.) (editors), *Mathematics in Medicine and Biology. Lecture Notes in Biomathematics*, vol. 57. – Springer-Verlag, 1985.

[11] Waterman (M.). – *Introduction to Computational Biology.* – Chapman and Hall, London, 1995.

[12] Waterman (Michael S.). – *Secondary structure of single-stranded nucleic acids*, pp. 167–212. – Academic Press, 1978, *Advances in Mathematics Supplementary Studies.*

[13] Waterman (Michael S.). – Combinatorics of RNA hairpins and cloverleaves. *Studies in Applied Mathematics*, vol. 60, n° 2, 1979, pp. 91–96.