A New Tool for Enumerative Combinatorics?

James Nulton[1], Peter Salamon[2], Mya Breitbart[3], Joe Mahaffy[4], Ben Felts[5], Beltran Rodriguez Brito[6], David Bangor[7], Forest Rohwer[8]

Keywords: shotgun sequencing, contigs, bernoulli polynomials, convolution

1 Introduction.

In 1988 Lander and Waterman [1] presented a mathematical framework for approximating various statistics related to shotgun sequencing, a tool used in planning projects related to the physical mapping of the genome. More recently [2], that framework has been used as part of a scheme for inferring population structural features of a community of marine bacteriophage from a statistical analysis of a shotgun library assembled from that community. In a shotgun sample for a single genome a maximal assembly of q sequences that covers a contiguous portion of the genome map is called a *q-contig*. Good estimates for the expected number of q-contigs is critical for this recent application. Lander/Waterman does very well for lower values of q, but Monte Carlo simulations show that, as q increases, the quality of the required estimates deteriorate.

The result showcased below was developed in an effort to analyze the shotgun sampling experiment by exact combinatorial counts. A special class of basis events was identified in terms of which all other events of interest (including contig events) could be expressed by standard combinatorial methods. The method of counting the events in the special class uses a construct that appears to be new to the field of combinatorics. It is a type of convolution in the ring of polynomials.

2 Sampling, Basis Events, and the B-convolution.

A sample (with replacement) of K numbers, called *points* is selected with uniform probability from the set $[G] = \{1, \ldots, G\}$. The sample space consists of G^K equally likely outcomes and can be identified with the set of maps $\{f : [K] \to [G]\}$. We are interested in clustering patterns among the points of an outcome. Let p and p' be a pair of sample points with no sample points between them. If $|p - p'| \geq C$, the pair forms a *gap*, otherwise it forms a *link*. The number C is called the *gap threshold*. In the context of a shotgun experiment, $[G]$ is the genome map, the sample points mark the starts (on the map) of the sequenced fragments, and C is the effective fragment length.

List the K points of an outcome (with possible repetitions) in non-decreasing order and consider consecutive pairs in the list. Each outcome has an ordered pattern of gaps and

[1]Department of Mathematics, San Diego State University, San Diego, California, 92182-7720. E-mail: jnulton@mail.sdsu.edu

[2]Department of Mathematics, San Diego State University, San Diego, California, 92182-7720. E-mail: salamon@saturn.sdsu.edu

[3]Department of Biology, San Diego State University, San Diego, California, 92182-4614. E-mail: mya@sunstroke.sdsu.edu

[4]Department of Mathematics, San Diego State University, San Diego, California, 92182-7720.

[5]Department of Mathematics, San Diego State University, San Diego, California, 92182-7720.

[6]Department of Mathematics, San Diego State University, San Diego, California, 92182-7720.

[7]Department of Mathematics, San Diego State University, San Diego, California, 92182-7720.

[8]Department of Biology, San Diego State University, San Diego, California, 92182-4614. E-mail: forest@sunstroke.sdsu.edu

links that can be represented as a sequence of $K-1$ binary digits, each of which records the status of a consecutive pair, "0" for gap, "1" for link. The jth digit records the status of the jth consecutive pair in the sorted list. Conversely, such a binary string denotes the event consisting of all outcomes with such a pattern of gaps and links. For example, 1001011 is an event ($K=8$) in which the first consecutive pair in the sorted list of points is a link, the second and third are gaps, etc. More generally, a symbol such as $XX0110X$ represents an event in which we are indifferent to the status of pairs in positions 1, 2, and 7, but the status of the others are specified. Incidentally, this event has a 3-contig (2 links) starting at position 4. The problem is to determine the cardinality of such a set. Standard counting methods (principle of inclusion/exclusion) allow us to reduce the problem to a simpler class of *basis events* whose binary symbols use only "X" and "0". Our method of counting such sets uses the following construct on the ring of polynomials.

Let g and h be polynomials of degrees m and n. The equation below defines a polynomial of degree $m+n+1$.

$$f(y)=\sum_{x=0}^{y} g(y-x)h(x) \tag{1}$$

We call f the *B-convolution* of g and h and denote it by $g\#h$. "B" is for Bernoulli, whose polynomials are required for the reduction of the right side to polynomial form.

We now illustrate with an example how this convolution is used to count the number of outcomes in a basis event. To that end, for each positive integer j, define the special polynomials: $b_j(x)=x^j$ and $a_j(x)=(x+1)^j-x^j$. Consider the basis event $E=X0XXX00XX$ (for $K=10$). Introduce 10 asterisks to set off the 9 symbols: $*X*0*X*X*X*0*0*X*X*$. Now remove the X's: $**0****0*0***$. The 10 asterisks are partitioned into $2+4+1+3$. Construct the polynomial $\pi_E=b_2\#a_4\#a_1\#a_3$. Count the number of terms, M, in the partition: $M=4$. The cardinality of the event E is $\pi_E(G-(M-1)C)$. It is a theorem that π_E is independent of the partition order, despite the apparent broken symmetry with a "b" as the first factor in the convolution.

3 The Expected Number of q-Contigs.

Finally, we summarize the result for the number, x_q, of q-contigs in a sample of size K. For every q there is a set of polynomials, $\{\,p_q^{(j)}\,|\,j=0,\ldots,q+1\}$, all of degree $q-1$, and *depending only on* q, in terms of which the expected number of q-contigs is given by

$$E[x_q]=G^{-K}\sum_{j=1}^{q+1}(2p_q^{(j)}\#b_{K-q}+p_q^{(j-1)}\#d_{K-q})(G-jC), \tag{2}$$

where $d_i(x)=x(x+1)a_{i-1}(x)$ are polynomials of degree i. All of the functions represented here as polynomials must, nevertheless, be taken to vanish on negative arguments.

We remark that the p's are integral linear combinations of π's associated with basis events, and there are recursion relations among them. We emphasize that the p's are computed independently of the parameters K, G and C.

References

[1] Lander E. S. and Waterman M. S. 1988. Genomic mapping by fingerprinting random clones. *Genomics* 2:231–239.

[2] Breitbart, M. et al. 2002. Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences USA* 99:14250–14255.