STATISTICS OF FRAGMENTS FROM ENZYME-INDUCED BREAKUP OF CO-POLYMER CHAINS

JOHN M. BLATT [1]

(Received 1 June 1966)

1. Introduction

A number of studies [1] have concerned themselves with properties of artificial poly-peptide chains, which differ from naturally occurring poly-peptides in two ways: There is only one kind of amino-acid in each polymer chain, but that acid occurs both in its right-handed and in its left-handed (D and L) forms, with some random order of L and D constituents.

An earlier paper [2] has been concerned with the statistical problem which arises when an assembly of such molecules is attacked by an enzyme, which can catalyze the breaking of a bond between two adjacent L constituents, but cannot affect LD or DD bonds. The long molecules break up into a number of smaller pieces, and we are interested in the average weight distribution of the resulting pieces. This weight distribution can be measured experimentally, and one hopes, from such measurements, to be able to reason back to the constitution of the original chains as well as to properties of the enzyme. For example, if the chains are composed, on the average, of equal numbers of L and D, it is still possible that there is a bias against LD or DL neighbours, as compared to LL and DD neighbours. Such a bias would affect the eventual weight distribution of fragments. Again, it is conjectured that the enzyme has a "groove" into which a portion of the chain molecule must fit properly in order to make the enzymatic action possible. If the breaking rules deduced from the experiments indicate that one needs at least three adjacent L's to get a break, one would then deduce that the "groove" is at least three constituents long (D constituents would not fit into the groove of a natural, L, enzyme).

In the earlier paper [2] we had to make rather stringent assumptions about the breaking rules. The present paper, which is written in such a way

[1] On study leave from the University of New South Wales, Sydney, N.S.W., Australia.

as not to require knowledge of the other, represents a significant generalization. The essential difference is that we are now able to allow for "memory" during the breaking process. To explain, let us pick a specific example, namely the chain: *DDLLLLD*. We shall assume, for this discussion, that the "*L*-string" of length 4 in this chain can be attacked by the enzyme, and that a break may occur either in the middle, with probability ψ, say, or just before the right-most *L*, with probability $1-\psi$. Thus the initial event leads to

$$DDLL+LLD \text{ with probability } \psi$$

or to

$$DDLLL+LD \text{ with probability } 1-\psi.$$

Let us assume, furthermore, that the pieces *DDLL*, *LLD*, and *LD* are "unbreakable", but that the piece *DDLLL* can be broken into *DDLL*+*L*, and will be so broken eventually if we let the enzyme act long enough. The eventual outcome is therefore as follows:

Piece	Weight of piece	Probability of eventual formation	Average weight fraction
DDLL	4	$\psi+(1-\psi)=1$	4/7
LLD	3	ψ	$(3/7)\psi$
LD	2	$1-\psi$	$(2/7)(1-\psi)$
L	1	$1-\psi$	$(1/7)(1-\psi)$

If we had a large number of chains *DDLLLLD*, the study of the resulting weight fractions would clearly suffice to determine the probability ψ. In practice, except for quite short chains, the internal constitution of the chains themselves is not known, only their overall weight. In that case, the weight fractions in the table would have to be multiplied by the probability that a chain of length 7 has the actual constitution *DDLLLLD*, and we would have to work out similar tables, and constitutional probabilities, for all the 2^7 possible chains of length 7, and combine the results to get average weight fractions.

In reference [2] the resulting statistical problem was simplified by assuming the absence of "memory" in the breaking rules; that is, we assumed that the eventual pattern of breaks which is observed does not depend on the time sequence of the breaking process. If this assumption is satisfied, we can in principle specify all the eventual breakpoints by looking directly at the original, unbroken chain. If the enzyme acts long enough, breaks will occur at all those points, and at no other points.

The rules illustrated in our example, however, do exhibit memory. If the process starts in a certain way, we get a stable product *LLD*, which

https://doi.org/10.1017/S1446788700004614 Published online by Cambridge University Press

does not break into $L+LD$. If the process starts the other way, the final result contains $L+LD$. This kind of breaking with memory does appear to occur in nature, and we require a statistical theory to cope with breaking rules of this kind.

There are clearly two distinct statistical problems here (a) the constitution of the chains to be broken, and (b) the results of the breakup of breakable segments of a chain. Problem (a) was only mentioned above, in connection with the probability of finding a chain of the precise constitution *DDLLLLD* among all possible chains of length 7. Problem (b) was solved above by explicit enumeration of the individual breakup processes in their time sequence, the "breakable segment" being the L-string of length 4.

2. Constitution of the polymer chain

Since the enzyme requires at least two adjacent L's (usually more) in order to produce a break, each copolymer chain can be divided into "breakable regions" and "unbreakable regions". A "breakable region" is always a number of adjacent L's, which we call an "L-string". There are four distinct types of L-string, depending upon what is on either side of the string. We list the types below:

Types of L-Strings

Type number	Description	Example ($n = 7$)
1	Piece of pure L	*LLLLLL*
2	Right wing	..*DLLLLLL*
3	Left wing	*LLLLLLD*..
4	L-hole	..*DLLLLLLD*..

We note that the "length" of the L-string is defined to be the number of adjacent L's it contains, exclusive of possible D's at either end. Thus, for example, the smallest type 4 L-string of "length" 7 contains 9 amino-acids, *DLLLLLLLD*.

Each type of L-string may or may not be breakable by the enzyme, depending upon its length. We define

(2.1) M_μ = minimum breakable length of an L-string of type μ.

Of particular interest is M_4, the minimum breakable length of an L-hole. We shall *define* an L-hole to be breakable, i.e., a segment of less than M_4 adjacent L's with a D on both sides will *not* be classified as an L-hole at all. Rather, such a segment is part of the *interior of a D-string*.

https://doi.org/10.1017/S1446788700004614 Published online by Cambridge University Press

By definition, a D-string starts with a D, ends with a D, and contains no interior L-hole; i.e., a D-string is an inherently unbreakable region of the original chain.

A typical copolymer chain can now be described as follows: reading from left to right, we start with a left wing (L-string of type 3) of length m, say. There follows a D-string of length d_1, then an L-hole of length l_1, then a D-string of length d_2, then an L-hole of length l_2, etc. Finally, we have the last D-string, of length d_k, say, followed by the final right wing (L-string of type 2) of length n. The total length N of the chain is the sum of the separate lengths, i.e.,

$$m+d_1+l_1+d_2+l_2+\cdots+l_{k-1}+d_k+n=N. \tag{2.2}$$

The set of numbers appearing on the left side of (2.2) defines the "configuration" of the chain. These numbers are non-negative integers with the following additional conditions:

$$d_i \geqq 1 \qquad i=1, 2, \cdots, k, \tag{2.3a}$$

$$l_i \geqq M_4 \qquad i=1, 2, \cdots, k-1. \tag{2.3b}$$

This classification of polymer chains works in all but one case. The exceptional case is a chain of N adjacent L's, with no D's at all. Although such chains are extremely improbable for large N, they must be allowed for in the calculation, and we shall take care of them by a special term.

3. Probability of a given configuration

In the theory we are about to develop, we shall assume that there is no more than nearest-neighbour memory during the process of polymerization. The present experimental evidence is consistent with no memory at all, so nearest-neighbour memory is a reasonable assumption.

Let δ and λ be the probabilities that any animo-acid be D or L, respectively, with, of course.

$$\delta+\lambda=1. \tag{3.1}$$

In the absence of any memory during the build-up process, the probability of finding a randomly selected pair of neighbours to be LL is λ^2, to be LD is $\lambda\delta$, and so on. We use a subscript 1 to denote an L, a subscript 2 to denote a D, and we define π_{ij} to be the conditional probability, given that the left-hand constituent of a pair of neighbours is "i", to find the right-hand neighbour to be "j". In the absence of correlations, π_{ij} is completely independent of the first index i, and is given by

$$\pi_{11}=\pi_{21}=\lambda, \quad \pi_{12}=\pi_{22}=\delta \text{ (no correlations)}. \tag{3.2}$$

https://doi.org/10.1017/S1446788700004614 Published online by Cambridge University Press

These numbers satisfy the relationships for conditional probabilities:

$$\pi_{11}+\pi_{12} = 1, \tag{3.3}$$

$$\pi_{21}+\pi_{22} = 1, \tag{3.4}$$

as well as the condition that there are as many pairs LD as pairs DL, on the average:

$$\lambda\pi_{12} = \delta\pi_{21}. \tag{3.5}$$

Conditions (3.3)—(3.5) are satisfied also in the presence of nearest-neighbour correlations, and therefore provide three conditions on the four numbers π_{ij}, leaving only one free parameter, h. In terms of this free parameter, the π_{ij} are

$$\begin{pmatrix}\pi_{11} & \pi_{12}\\ \pi_{21} & \pi_{22}\end{pmatrix} = \begin{pmatrix}1-h\delta & h\delta\\ h\lambda & 1-h\lambda\end{pmatrix}. \tag{3.6}$$

These values reduce to (3.2) for the special case $h = 1$, no correlations. They satisfy (3.3)—(3.5) for all values of h. In order for all the π_{ij} to be positive numbers (probabilities), h must be positive and no larger than the smaller one of $1/\lambda$ and $1/\delta$.

Given these basic probability assumptions, let us now compute the probability of finding a chain of length N and given configuration $(m, d_1, l_1, d_2, l_2, \cdots, d_k, n)$ as described in section 2. For simplicity of explanation, let us start with m and n not equal to zero. Reading from left to right, we first encounter an L; the probability of this is λ. We then have $m-1$ further L's in succession, the probability being $(\pi_{11})^{m-1}$. This is followed by a D, with probability π_{12}. Next we have a D-string of length d_1, with probability p_{d_1}, say. The last D of the D-string is followed by an L, probability $= \pi_{21}$, this is followed by l_1-1 further L's, probability $= (\pi_{11})^{l_1-1}$, then by a D, probability $= \pi_{12}$, and so on.

The probability of the given configuration is therefore a product of factors, as follows:

$$P(m, d_1, l_1, d_2, l_2, \cdots, d_k, n) = \hat{w}_m p_{d_1} v_{l_1} p_{d_2} v_{l_2} \cdots p_{d_k} w_n \tag{3.7}$$

where

$$\hat{w}_m = \begin{cases}\lambda(\pi_{11})^{m-1}\pi_{12} & \text{for } m \geqq 1\\ \delta & \text{for } m = 0\end{cases} \quad \text{(Left wing factor)} \tag{3.8}$$

$$v_l = \pi_{21}(\pi_{11})^{l-1}\pi_{12}, \quad (L\text{-hole factor, note } l \geqq M_4), \tag{3.9}$$

$$w_n = \begin{cases}\pi_{21}(\pi_{11})^{n-1} & \text{for } n \geqq 1\\ 1 & \text{for } n = 0\end{cases} \quad \text{(Right wing factor).} \tag{3.10}$$

https://doi.org/10.1017/S1446788700004614 Published online by Cambridge University Press

In (3.8) and (3.10), we have included the necessary modifications for chains starting or ending with a D, i.e., for wings of length zero.

The only factor not yet written down is the D-string factor p_d, which is the conditional probability that a segment of length d, the left-most member of which is known to be a D, is actually a D-string of length d. This factor was discussed in some detail in reference [2], for the special case $M_4 = 3$, i.e., no more than 2 adjacent L's inside a D-string. We derived a recurrence relation for p_d, by considering the possible structures of the D-string in the immediate neighbourhood of the right-most D. We now generalize the argument, and the resulting recurrence relation, to arbitrary M_4.

A D-string of length d starts and ends with a D. Let us now classify these D-strings by l, the number of consecutive L's immediately to the left of the right-most D. If $l = 0$, the D-string of length d ends with DD, and the first $d-1$ constituents themselves form a D-string of length $d-1$. The probability of this configuration is the product of p_{d-1}, the probability of a D-string of length $d-1$, and π_{22}, the probability of having the last D of that D-string followed by another D. If $l = 1$, the D-string of length d ends with DLD, so that the first $d-2$ elements themselves form a D-string. The probability of this configuration is $p_{d-2}\pi_{21}\pi_{12}$. For general l, the probability of the configuration is

$$p_{d-l-1}\pi_{21}(\pi_{11})^{l-1}\pi_{12} \qquad l = 1, 2, \cdots, M_4-1.$$

The highest value of l here is M_4-1, because M_4 consecutive L's between two D's can be broken by the enzyme, and thus M_4 consecutive L's can never occur in the interior of a D-string.

The configurations which we have listed above are mutually exclusive, and the alternatives $l = 0, 1, 2, \cdots, M_4-1$ between them exhaust all possible D-strings of length d. We can therefore add the separate probabilities, to get the recurrence relation

$$(3.11) \quad p_d = p_{d-1}\pi_{22} + p_{d-2}\pi_{21}\pi_{12} + p_{d-3}\pi_{21}\pi_{11}\pi_{12} + \cdots + p_{d-M_4}\pi_{21}(\pi_{11})^{M_4-2}\pi_{12}.$$

As written, this recurrence relation is valid for $d > M_4$. However, it turns out that (3.11) can be used starting with $d = 2$, provided we use the obvious (from the definition) result

$$(3.12a) \qquad p_1 = 1$$

supplemented by the formal definitions

$$(3.12b) \qquad p_d = 0 \text{ for } d \leqq 0.$$

For the special case $M_4 = 3$, (3.11) reduces to the recursion relation derived in the earlier work [2].

https://doi.org/10.1017/S1446788700004614 Published online by Cambridge University Press

We note that (3.11) with the initial conditions (3.12) involves sums of positive terms only, so that there is no loss of numerical accuracy due to cancellations in subtracting one large number from another. Thus (3.11) can be used on an electronic computer to generate all the p_a which are required.

We also note that the coefficients which appear in (3.11) are just the v_l defined as "L-hole factors" in equation (3.9). This is of course no accident: the only difference between an L-hole of length l, and a sequence of l consective L's preceding the last D of a D-string, is the value of l in relation to M_4, the minimum breakable length of such a configuration.

We have now defined, either explicitly or by means of a numerically useful recursion relation, all the factors which appear in the configuration probability (3.7). The only configuration not accounted for by (3.7) is the (very unlikely) case that all N constituents of the chain are L's. The probability of this is

$$P_0(N) = \lambda(\pi_{11})^{N-1}. \tag{3.13}$$

For a given value of N, the sum of (3.13) and all the configuration probabilities (3.7) for possible configurations (i.e., configurations satisfying conditions (2.2) and (2.3)) must add to unity. We do not give the explicit proof here.

The result that the configuration probabilities for all configurations except the pure-L chain are simply products of independent factors is a very simple result, which makes the subsequent work quite straightforward. It is therefore worthwhile to enquire to what we owe this remarkable simplification. There are two features of the problem involved here:

(1) We have assumed at most nearest-neighbour memory during the process of building up the co-polymer;

(2) We have assumed that the enzyme, during the breakup process, "sees" only one breakable region at a time, so that two breakable regions (L-holes) separated by a D-string, can be treated independently.

These assumptions appear reasonable on present evidence. If one or both of them must be relaxed as a result of further experimental work, the theory would become appreciably more complicated. A likely trouble spot is assumption (2) in the case that the intervening D-string is just a single D. The enzyme may act differently on the chain

DDDDLLLLLDLLLLLDDDD

from the way it acts on

DDLLLLLDDDDDLLLLLDD

https://doi.org/10.1017/S1446788700004614 Published online by Cambridge University Press

even though both chains contain exactly two L-holes of length 5. If this turns out to be true, we shall have to amend our definition of "breakable region" to arrange it so that two different breakable regions are always broken in statistically independent fashion. This would mean declaring the $LLLLLDLLLLL$ in the first chain above to be *one* "breakable region", not two breakable regions separated by a D-string. A number of other complications would also have to be introduced into the theory concurrently. It is likely that such a theory could be carried through along the lines of the present theory, but this author hopes that nature will not turn out to be so very nasty.

4. Expected numbers of various configurations

We define $Y(N, n, \mu)$ to be the expected value of the number of L-strings of length n and type μ ($\mu = 1, 2, 3, 4$, see § 2), in a chain of length N.

The simplest one to evaluate is for $\mu = 1$, the pure-L piece with no D's anywhere. The only way we can encounter this in a chain of length N is if the entire chain is pure L. The probability of this $= P_0(N)$, equation (3.13). Thus the expected value is given by

$$Y(N, n, 1) = P_0(N)\Delta(N-n), \tag{4.1}$$

where

$$\Delta(k) = \begin{cases} 1 & \text{for } k = 0, \\ 0 & \text{for } k \neq 0. \end{cases} \tag{4.2}$$

Other expected values can be obtained by the use of generating functions. To reduce the length of this paper, only the results are quoted here.

We introduce the function $E(k)$ by

$$E(k) = \begin{cases} 1 & \text{for } k \geqq 1, \\ 0 & \text{for } k \leqq 0, \end{cases} \tag{4.3}$$

to write the result

$$Y(N, n, 2) = \delta w_n E(N-n). \tag{4.4}$$

We can understand that $Y(N, n, 2) = 0$ for $N < n+1$, since the smallest chain with a right wing of length n is a chain consisting of a single D, followed by n L's and thus of total length $n+1$. Next we quote the result

$$Y(N, n, 3) = \hat{w}_n E(N-n). \tag{4.5}$$

Last, we define

$$H(k) = \begin{cases} k, & k \geqq 1, \\ 0, & k \leqq 0, \end{cases} \tag{4.6}$$

to write down the formula

https://doi.org/10.1017/S1446788700004614 Published online by Cambridge University Press

$$Y(N, n, 4) = \delta v_n H(N-n-1). \tag{4.7}$$

We note that the first non-zero value of $Y(N, n, 4)$ occurs for $N = n+2$. This is understandable because the smallest chain N which can contain an L-hole of length n at all, is a chain of length $N = n+2$, namely one D on either side of the L-hole.

We observe that $Y(N, n, \mu)$ has quite different behaviour in the limit of large chain length N, and constant L-string n, for different values of μ. For $\mu = 1$, equations (4.1) and (3.13) show that $Y(N, n, 1)$ approaches zero in that limit. Equations (4.4) and (4.5) show that $Y(N, n, 2)$ and $Y(N, n, 3)$ approach constant values in the same limit (in fact, they are constants, independent of both N and n, as soon as N exceeds its minimum permissible value, $N = n+1$). Finally, $Y(N, n, 4)$ becomes proportional to the chain length N for large N, according to (4.6) and (4.7). All these results are plausible on intuitive grounds: $\mu = 1$ are pure-L chains, which become increasingly improbable; $\mu = 2$ L-strings can come only from the right end of any one long chain, and thus their expected number should be independent of chain length N for a sufficiently long chain; similarly, $\mu = 3$ L-strings can come only from the left end of a long chain; finally the L-holes, $\mu = 4$, can occur anywhere inside a long chain, and thus their expected number increases linearly with the chain length N, for sufficiently large N.

A "D-string", as defined in § 2, cannot be broken by the enzyme. The fragment which emerges, however, is generally longer than the D-string which it contains, having further L's attached to its right and to its left, by "unbreakable" bonds. The probability of getting various combinations of attached L's in turn depends on the original environment of the D-string in question, i.e., on the lengths and types of the L-strings on either side of the D-string.

We classify "clothed D-strings" by:

m, the length of the L-string to the left;
μ, the type of the L-string to the left;
d, the length of the D-string;
n, the length of the L-string to the right;
ν, the type of the L-string to the right.

Each quintuple (m, μ, d, n, ν) defines a particular class of clothed D-strings; not all combinations of μ and ν are possible, the possible combinations being

(4.8a) 1) Interior clothed D-string: $(m, 4, d, n, 4)$ $m \geqq M_4$ and $n \geqq M_4$;
(4.8b) 2) Right-most D-string: $(m, 4, d, n, 2)$ $m \geqq M_4$ and $n \geqq 0$;
(4.8c) 3) Left-most D-string: $(m, 3, d, n, 4)$ $m \geqq 0$ and $n \geqq M_4$;
(4.8d) 4) Sole D-string: $(n, 3, d, n, 2)$ $m+d+n = N$.

https://doi.org/10.1017/S1446788700004614 Published online by Cambridge University Press

In the last case, N is the length of the entire chain, which contains only one D-string in its interior.

We define $Z(N; m, \mu, d, n, \nu)$ to be the expected number of clothed D-strings of type (m, μ, d, n, ν) contained in a chain of length N.

The simplest case is the "sole D-string". Since there is at most one such clothed D-string in any one chain, the expected value equals the probability of occurrence of the event in question. This latter is given directly by (3.7). Using the notation $\Delta(k)$ of (4.2) to incorporate the condition $m+d+n = N$, we get

$$Z(N; m, 3, d, n, 2) = \hat{w}_m p_d w_n \Delta(N-m-d-n). \tag{4.9}$$

Next, let us consider case 3 above, the left-most D-string. In order for a chain to contain a left-most D-string, it must contain at least two D-strings. It can be shown (e.g., by the use of generating functions) that

$$Z(N; m, 3, d, n, 4) = \hat{w}_m p_d v_n E(N-m-d-n). \tag{4.10}$$

We note that $E(N-m-d-n)$ is zero until N is at least equal to $m+d+n+1$. This is indeed the minimum length of a chain which can contain a left-most D-string of type $(m, 3, d, n, 4)$: $m+d+n$ constituents are necessary for the clothed D-string itself, and the last constituent is a sole D, i.e., a D-string of length 1. Furthermore, it is reasonable to expect that for large N, the number of these particular D-strings becomes independent of N: any one chain can contain at most one clothed D-string of this left-most type.

The expected value of the number of right-most D-strings can be shown to be

$$Z(N; m, 4, d, n, 2) = \delta v_m p_d w_n E(N-m-d-n). \tag{4.11}$$

Finally, the expected number of interior clothed D-strings is

$$Z(N; m, 4, d, n, 4) = \delta v_m p_d v_n H(N-m-d-1). \tag{4.12}$$

We note that (4.12) gives zero until the chain length N is at least equal to $m+d+n+2$. This is a chain containing exactly three D-strings, the first and last of which are single D's, and the middle one of which is our clothed interior D-string. We also note that (4.12) becomes proportional to the chain length N for large N, as we would expect.

5. The break-up of L-strings, definitions and preliminary discussion

The four types of L-strings were defined in § 2. Most of the L-strings in a long chain are type 4, i.e., "L-holes". The break-up of an L-hole

https://doi.org/10.1017/S1446788700004614 Published online by Cambridge University Press

proceeds in a number of steps, the first of which results in one right wing and one left wing:

$$(5.1) \qquad (n, 4) \to (k, 2)+(n-k, 3), \ n \geqq M_4.$$

Here (m, μ) denotes an L-string of length m and type μ. The right wing may be breakable, in which case it breaks into a shorter right-wing plus a pure-L piece:

$$(5.2) \qquad (n, 2) \to (k, 2)+(n-k, 1), \ n \geqq M_2.$$

Similarly, a breakable left wing breaks into a piece of pure-L plus a shorter left wing

$$(5.3) \qquad (n, 3) \to (k, 1)+(n-k, 3), \ n \geqq M_3.$$

Finally, if any of the pure-L pieces generated in (5.2) or (5.3) are themselves breakable, they break into smaller pieces of pure-L:

$$(5.4) \qquad (n, 1) \to (k, 1)+(n-k, 1), \ n \geqq M_1.$$

For each of these types of enzyme-induced break-up, there is a parameter k giving the location of the first break. We define $\psi(n, \mu, k)$ to be the probability that the first break in an L-string of length n and type μ occurs in position k. The "position" is defined to be the number of L's in that piece of the original L-string immediately to the left of this first break. Thus, in (5.4) for example, k is at least 1 and at most $n-1$. If breaks can occur right next to a D, then k may be as low as zero in (5.1) and (5.2), and may be as high as n in (5.1) and (5.3). The probabilities $\psi(n, \mu, k)$ satisfy:

$$(5.5) \qquad \psi(n, \mu, k) = 0 \text{ for } n < M_\mu,$$

and

$$(5.6) \qquad \sum_k \psi(n, \mu, k) = 1 \text{ for } n \geqq M_\mu.$$

Usually, breaks cannot occur too close to the exterior D's. Thus, in $\psi(n, 4, k)$, for example, we expect $\psi = 0$ for the first few values of k, as well as for the last few values of k. We define $L_\mu(n)$ to be the left-most possible break position for an L-string of type μ and length n, i.e.,

$$(5.7) \qquad \psi(n, \mu, k) = 0 \text{ for } k < L_\mu(n), \quad \psi \neq 0 \text{ for } k = L_\mu(n),$$

and we define $R_\mu(n)$ to be the right-most possible break position, measured from the right hand end:

$$(5.8) \qquad \psi(n, \mu, k) \begin{cases} = 0, \text{ for } n-k < R_\mu(n), \\ \neq 0, \text{ for } n-k = R_\mu(n). \end{cases}$$

https://doi.org/10.1017/S1446788700004614 Published online by Cambridge University Press

It is frequently the case, experimentally, that the numbers $L_\mu(n)$ and $R_\mu(n)$ are independent of n, the length of the L-string.

Next, we define the quantity $X(n, \mu, n', \mu')$ to be the *expected eventual number of L-strings of type* (n', μ') *arising from an initial L-string of type* (n, μ), after an infinite time has elapsed.

We illustrate this definition by using the case discussed in the introduction (§ 1): the break-up of the chain *DDLLLLD*. From our present point of view, this is the break-up of an L-string of length 4 and type 4. The quantity ψ used there is equal to $\psi(4, 4, 2)$ in our present notation, and $1-\psi$ is $\psi(4, 4, 3)$. The other $\psi(4, 4, k)$ vanish, i.e., $L_4 = 2$ and $R_4 = 1$. Looking at the table in § 1, we see that there is unit probability of getting a final piece *DDLL*, i.e., in our present notation, a right wing of length 2. Thus we have, for these particular breaking rules,

$$X(4, 4, 2, 2) = 1.$$

Looking again at the table, we see that there is a probability ψ for a piece *LLD*, i.e., for a left wing of length 2, and a probability $1-\psi$ for a piece *LD*, i.e., for a left wing of length 1. In terms of X-coefficients, we have therefore

$$X(4, 4, 2, 3) = \psi, \quad X(4, 4, 1, 3) = 1-\psi.$$

Finally, the last line of the table in section 1 shows that the emergent pure-L strings have length 1 only, with probability $1-\psi$. Thus,

$$X(4, 4, 1, 1) = 1-\psi.$$

All $X(4, 4, m', \mu')$ not explicitly listed above are zero.

In this simple example, all the expected values X are also probabilities. For L-strings longer than $n = 4$, this is not generally the case; however, $\mu' = 2$ and $\mu' = 3$ do have X-coefficients, $X(m, \mu, m', \mu')$, which are also probabilities:

$$(5.9) \qquad \sum_{m'} X(m, \mu, m', 2) = 1 \text{ for } m \geqq M_\mu, \ \mu = 2 \text{ and } \mu = 4,$$

$$(5.10) \qquad \sum_{m'} X(m, \mu, m', 3) = 1 \text{ for } m \geqq M_\mu, \ \mu = 3 \text{ and } \mu = 4.$$

Both (5.9) and (5.10) are clearly satisfied for the $X(4, 4, m', \mu')$ listed above. The reason for these relations is the fact that an L-string of type $\mu = 2$ or $\mu = 4$ gives rise to exactly one final right wing, $\mu' = 2$. This is equation (5.9). Similarly, a breakable L-string of type $\mu = 3$ or $\mu = 4$ always gives rise to exactly one final left wing, $\mu' = 3$. This is equation (5.10). For $\mu \neq 4$, the condition of breakability, $m \geqq M_\mu$, can be relaxed, as we shall see shortly.

For $\mu' = 1$, the expected values $X(m, \mu, m', 1)$ are most definitely

not probabilities; on the contrary, they rapidly become larger than unity. However, there is a sum rule for these $X(m, \mu, m', \mu')$ which expresses the condition that the total number of separate L's cannot change during the breakup process. This sum rule is

$$\sum_{m'} \sum_{\mu'} m' X(m, \mu, m', \mu') = m. \tag{5.11}$$

It holds separately for each m and each μ.

Since we have defined X to be the expected numbers of *final* L-strings of type (m', μ'), it follows that only ultimately stable L-strings can appear with non-zero X:

$$X(m, \mu, m', \mu') = 0 \text{ unless } m' < M_\mu, \text{ and } \mu' = 1, 2, 3 \text{ only.}^2 \tag{5.12}$$

Also, L-strings of size m can obviously not give rise to final L-strings of size m' in excess of m, thus:

$$X(m, \mu, m', \mu') = 0 \text{ unless } m' \leqq m. \tag{5.13}$$

The X values are particularly simple if the initial L-string (m, μ) is itself stable. Such a string survives unaltered for all time, and we obtain

$$X(m, \mu, m', \mu') = \Delta(m-m')\Delta(\mu-\mu') \text{ for } m < M_\mu, \ \mu = 1, 2, 3, ^2 \tag{5.14}$$

This shows that (5.9) and (5.10) for $\mu \neq 4$ are also valid for $m < M_\mu$, as mentioned above.

We have defined the $X(m, \mu, m', \mu')$ to be the ultimate expected values, after an infinite time has elapsed. If we intended to follow the break-up process in strict time sequence, these would be the limiting values, for infinite time, of time-dependent quantities $X_t(m, \mu, m', \mu')$. It is certainly possible to set up a system of differential equations in time for the X_t, but this procedure is unwieldy, inefficient, and quite unnecessary. For one thing, unlike the ultimate values at infinite time, the finite time values are not restricted by condition (5.12). On the contrary, we would have to keep track of all $X_t(m, \mu, m', \mu')$ with m' all the way up to m, even though in the end we are interested only in the lowest few values of m'. The resulting calculation is terribly messy, and is hard to put on a computer, even, because of excessive numbers of things which must be kept in storage at any one time. It is essential for the success of our calculation that we do *not* take this approach.

6. The break-up of L-strings, regeneration point method

To illustrate the method we use, let us consider the break-up of an L-string of type 1, i.e., piece of pure L, according to the pattern (5.4).

[2] We recall that "unbreakable L-holes", i.e., type 4 and length less than M_4, are not classified as L-holes at all, but rather as parts of D-strings.

https://doi.org/10.1017/S1446788700004614 Published online by Cambridge University Press

After this first break has occurred, with probability $\psi(n, 1, k)$, we have two pieces of the same kind, one of length k, the other of length $n-k$. The first of these gives rise to $X(k, 1, n', 1)$ pure-L pieces of length $n' < M_1$, the second of these gives rise to $X(n-k, 1, n', 1)$ pure-L pieces of length n'. Since the events with different first break-point k are mutually exclusive, we can add their probabilities to get the recursion relation:

$$(6.1) \quad X(n, 1, n', 1) = \sum_k \psi(n, 1, k)[X(k, 1, n', 1)+X(n-k, 1, n', 1)], \quad n \geqq M_1.$$

This relates $X(n, 1, n', 1)$ to earlier X-coefficients, with the *same* values of n' and μ', but lower values of the initial length n. The sum over k goes from $k = L_1(n)$ to $k = n-R_1(n)$, inclusive. The relation (6.1) holds for breakable pieces of type 1, i.e., for $n \geqq M_1$. For unbreakable pieces, $n < M_1$, we know the answer from (5.14). Between them, (5.14) and (6.1) provide a recursive definition of all $X(n, 1, n', 1)$.

Unlike the procedure of following the break-up process in its actual time sequence, the regeneration point method does not require us to keep track of time-dependent intermediate numbers of pieces which turn out to break up eventually. On the contrary, (6.1) is a relation between numbers we want and need, numbers referring only to the eventual situation after an infinite time has elapsed.

We note also that the recursion in (6.1) is not on n', the length of the product piece, but rather on n, the length of the initial piece. Different final lengths n' are "decoupled" in this system of recursion relations.

We now apply the same reasoning to the other break-up processes (5.1)—(5.3). The break-up of a right wing, pattern (5.2), gives rise to the recursion relations:

$$(6.2) \quad X(n, 2, n', 2) = \sum_k \psi(n, 2, k)X(k, 2, n', 2) \text{ for } n \geqq M_2,$$

and

$$(6.3) \quad X(n, 2, n', 1) = \sum_k \psi(n, 2, k)[X(k, 2, n', 1)+X(n-k, 1, n', 1)] \quad \text{for } n \geqq M_2.$$

We note that the $X(n-k, 1, n', 1)$, which appear in (6.3), are known from (6.1) if (6.1) has been solved first for all initial lengths less than or equal to n. The sums over k in (6.2) and (6.3) go from $L_2(n)$ to $n-R_2(n)$, inclusive.

The break-up of a left wing, pattern (5.3), yields the recursion relations:

$$(6.4) \quad X(n, 3, n', 3) = \sum_k \psi(n, 3, k)X(n-k, 3, n', 3) \text{ for } n \geqq M_3,$$

$$(6.5) \quad X(n, 3, n', 1) = \sum_k \psi(n, 3, k)[X(k, 1, n', 1)+X(n-k, 3, n', 1)] \quad \text{for } n \geqq M_3.$$

https://doi.org/10.1017/S1446788700004614 Published online by Cambridge University Press

The sums over k go from $L_3(n)$ to $n-R_3(n)$, inclusive.

Finally, the break-up of an L-hole, type 4, according to the pattern (5.1), gives rise to three distinct recursion relations, namely

$$(6.6)\quad X(n, 4, n', 3) = \sum_k \psi(n, 4, k)X(n-k, 3, n', 3) \text{ for } n \geqq M_4,$$

$$(6.7)\quad X(n, 4, n', 2) = \sum_k \psi(n, 4, k)X(k, 2, n', 2) \text{ for } n \geqq M_4,$$

$$(6.8)\quad X(n, 4, n', 1) = \sum_k \psi(n, 4, k)[X(k, 2, n', 1)+X(n-k, 3, n', 1)] \text{ for } n \geqq M_4.$$

The sums over k range from $L_4(n)$ to $n-R_4(n)$, inclusive.

The set of equations (6.1)—(6.8), together with (5.14) for unbreakable pieces defines *all* coefficients $X(n, \mu, n', \mu')$ of interest to us, recursively. For a given n, we solve (6.1) to (6.8) in that order. We increase n' one unit at a time, from its minimum value (1 for $\mu' = 1$, 0 for $\mu' = 2$ and $\mu' = 3$) to *one less* than its maximum value $M_{\mu'}$. The last value of $n' = M_{\mu'}$ can be calculated without use of the recursion relations, from the sum rules (5.9), (5.10), and (5.11). This saves machine time. The sequence of equations (6.1)—(6.8) is arranged so as to make this possible. We use the weight sum rule (5.11) to replace (6.1) for $n' = M_1$; we then use the probability sum rule (5.9) to replace (6.2) for $n' = M_2$, and the weight sum rule (5.11) once more to replace (6.3) for $n' = M_1$, and so on. At each stage of this process, all but one of the terms which appear in the given sum rule are already known.

7. The break-up of L-strings, asymptotic expressions

It is apparent from the definition of the $X(n, \mu, n', \mu')$, as well as from the recursion relations used for their evaluation, that the X-coefficients are independent of the average constitution of the long chain molecule, i.e., they are independent of the parameters λ, δ, and π_{ij} which played so prominent a part in §§ 3 and 4. Thus, it saves machine time if one keeps the break-up parameters (which define the elementary break probabilities $\psi(n, \mu, k)$) constant during a series of computer runs during which λ and h are varied. In this case, the X-coefficints need not be recomputed each time.

In practice we need to know the X-coefficients for n up to a value large enough so that the probability of an L-string of this length is negligibly small. This maximum value of n, unlike the X-coefficients themselves, depends upon λ, becoming large when λ approaches unity. Long before this maximum n has been reached, however, the X-coefficients themselves have settled down to predictable values.

As an illustration, let us consider $X(n, 1, n', 1)$. The weight sum rule (5.11) for this case reads simply:

$$\sum_{n'=1}^{M_1-1} n' X(n, 1, n', 1) = n, \tag{7.1}$$

since the sum over μ' in (5.11) gives non-zero values only for $\mu' = 1$ (break-up of a pure-L piece can never lead to right wings or left wings).

For large n, the first break position k is overwhelmingly likely to be somewhere in the middle of the piece, neither close to the left end nor close to the right end. This leaves two pieces, each of which is still large, so that we can expect stabilization of the X-values, or rather, of their ratios from one n' to the next. The actual *values* cannot stabilize, since by (7.1) the sum of these values, weighted with n', must increase linearly with n. If we assume that the *ratios* stabilize after some $n = n_0$, we get the asymptotic approximation:

$$X(n, 1, n', 1) \cong \frac{n}{n_0} X(n_0, 1, n', 1) \text{ for } n \geqq n_0 \gg M_1. \tag{7.2}$$

Another way looking at this equation is to say that, for values of n much larger than the minimum breakable length M_1, the weight fractions from pure-L chains settle down to stable values. Actual computer results show that this stabilization is achieved to very high accuracy for $n_0 = 10M_1$.

Our main interest for later use (to get expressions for the break-up of infinitely long chains) are the X-coefficients for the break-up of L holes, i.e., for $\mu = 4$. For $\mu' = 2$ and $\mu' = 3$, the X-coefficients are themselves probabilities, also, as shown by (5.9) and (5.10). Thus, the actual values can be expected to stabilize, not merely the ratios:

$$X(n, 4, n', \mu') \cong X(n_0, 4, n', \mu') \text{ for } \mu' = 2, 3, \; n \geqq n_0 \gg M_4. \tag{7.3}$$

This is indeed observed in the computer results.

We define the weight sum for these right wing and left wing pieces as

$$s_n = \sum_{n'=0}^{M_2-1} n' X(n, 4, n', 2) + \sum_{n'=0}^{M_3-1} n' X(n, 4, n', 3). \tag{7.4}$$

Since the X-coefficients in this equation all stabilize, so does the sum:

$$s_n \cong s_{n_0} \text{ for } n \geqq n_0 \gg M_4. \tag{7.5}$$

The weight sum rule (5.11) for $\mu = 4$ assumes the form

$$\sum_{n'=1}^{M_1-1} n' X(n, 4, n', 1) = n - s_n. \tag{7.6}$$

We assume that the *ratios* of the $X(n, 4, n', 1)$ for different n' stabilize

https://doi.org/10.1017/S1446788700004614 Published online by Cambridge University Press

for large n, and use (7.5) and (7.6) to normalize the coefficients themselves, to get

$$X(n, 4, n', 1) \cong \frac{n-s_{n_0}}{n_0-s_{n_0}} X(n, 4, n', 1) \text{ for } n \geqq n_0 \gg M_4. \tag{7.7}$$

Formulas (7.3), (7.5), and (7.7) can be expected to be quite reliable approximations for $n_0 = 10M_4$.

8. Weight fractions from finite chains

We are now in a position to tackle the original problem, that is, to determine the weight fractions of fragments of given weight l, arising from the break-up of a chain of initial weight N.

Let us define $R_1(N, l)$ to be the expected number of stable pure-L pieces of length l from our ensemble of chains of length N. Clearly l satisfies the condition $l < M_1$, otherwise $R_1(N, l) = 0$. There are four distinct contributions to R_1:

1) Initial chain is pure-L;

2) The L-piece of length l originates from the initial right wing of the long chain;

3) The L-piece of length l originates from the initial left wing of the long chain;

4) The L-piece of length l originates from one of the L-holes inside the long chain.

Recalling the definitions of $Y(N, n, \mu)$ and $X(n, \mu, n', \mu')$, we see that the contribution number μ above is equal to the sum of the product $Y(N, n, \mu)X(n, \mu, l, 1)$ over all permissible n, that is over n from l to N:

$$R_1(N, l) = \sum_{\mu=1}^{4} \sum_{n=l}^{N} Y(N, n, \mu)X(n, \mu, l, 1). \tag{8.1}$$

We now insert the actual values of the Y-coefficients, obtained from equations (4.1), (4.4), (4.5) and (4.7), to get

$$\begin{aligned} R_1(N, l) = {} & \lambda(\pi_{11})^{N-1} X(N, 1, l, 1) \\ & + \sum_{n=l}^{N} [\hat{w}_n E(N-n)X(n, 2, l, 1) + \delta w_n E(N-n)X(n, 3, l, 1) \\ & + \delta v_n H(N-n-1)X(n, 4, l, 1)]. \end{aligned} \tag{8.2}$$

Although this formula is explicit and simple, it is possible and desirable to simplify the notation somewhat further. We note from (3.8) and (3.10) that

$$\hat{w}_n = \delta w_n, \tag{8.3}$$

https://doi.org/10.1017/S1446788700004614 Published online by Cambridge University Press

and from (3.6) and (3.8) that

$$\lambda(\pi_{11})^{N-1} = w_N/h \text{ for } N = 1, 2, 3 \cdots. \tag{8.4}$$

It is therefore useful to define new coefficients $T(n, \mu, n', \mu')$ by

$$T(n, \mu, n', \mu') = \begin{cases} w_n X(n, \mu, n', \mu') & \text{for } \mu = 1, 2, 3, \\ v_n X(n, \mu, n', \mu') & \text{for } \mu = 4. \end{cases} \tag{8.5}$$

We note that, unlike the X-coefficients, the T-coefficients depend on the chain constitution parameters λ and h. In terms of the T-coefficients, equation (8.2) becomes:

$$R_1(N, l) = \frac{1}{h} T(N, 1, l, 1) + \delta \sum_{n=l}^{N} \{E(N-n)[T(n, 2, l, 1) + T(n, 3, l, 1)] + H(N-n-1)T(n, 4, l, 1)\}. \tag{8.6}$$

We now return our attention to D-containing fragments. We define $R_0(N, l)$ to be the expected number of stable D-containing fragments of length l from the ensemble of chains of length N. Each such fragment has a D-string of some length d in its interior, preceded by a left-wing L-string of some length m', and followed by a right-wing L-string of some length n'. The total length of the fragment is given by

$$l = m' + d + n'. \tag{8.7}$$

Such a fragment is the result of the enzymatic break-up of a clothed D-string of type (m, μ, d, n, ν) where $m \geqq m'$, $\mu = 3$ or 4, $n \geqq n'$, and $\nu = 2$ or 4. We now use the fact that the X-coefficients $X(m, \mu, m', 3)$ and $X(n, \nu, n', 2)$ are not only expected values but also are themselves probabilities, see the sum rules (5.9) and (5.10). Recalling the definition of $Z(N; m, \mu, d, n, \nu)$ in § 4 as the expected number of clothed D-strings, we see that the contribution of such clothed D-strings to the expected number of fragments of type (8.7) is

$$Z(N; m, \mu, d, n, \nu) X(m, \mu, m', 3) X(n, \nu, n', 2).$$

We now combine the explicit formulas § 4 for the Z-coefficients and the definition (8.5), and sum over all possible values of m, m', n, n', and d consistent with final fragment length l according to (8.7), to get

$$\begin{aligned} R_0(N, l) = \delta \sum_{m, m', n, n', d} & p_d \Delta(l-m'-d-n') \\ & [H(N-m-d-n-1)T(m, 4, m', 3)T(n, 4, n', 2) \\ & + E(N-m-d-n)T(m, 3, m', 3)T(n, 4, n', 2) \\ & + E(N-m-d-n)T(m, 4, m', 3)T(n, 2, n', 2) \\ & + \Delta(N-m-d-n)T(m, 3, m', 3)T(n, 2, n', 2)]. \end{aligned} \tag{8.8}$$

The terms in the square brackets have a simple interpretation: The first is the contribution from "interior" clothed D-strings, the second term is the contribution from the left-most D-string in the original chain, the third term is the contribution from the right-most D-string in the original chain, and the last term is the contribution from those chains which contain exactly one D-string altogether.

The expected number of fragments of length l from break-up of a chain of length N is the sum of (8.6) and (8.8):

$$R(N, l) = R_0(N, l)+R_1(N, l), \tag{8.9}$$

and the fractional weight residing in fragments of length l is related to this quantity by

$$W(N, l) = \frac{l}{N} R(N, l). \tag{8.10}$$

These weight fractions must add to unity

$$\sum_{l=1}^{N} W(N, l) = 1. \tag{8.11}$$

Condition (8.11) provides a useful check on the numerical calculations, since it is very difficult for errors to cancel in such a way as to preserve (8.11) intact.

The form (8.8) for $R_0(N, l)$, though simple to write down, is not actually convenient from the computational point of view. There is a five-fold summation, which remains a true four-fold sum after one allows for the delta-function $\Delta(l-m'-d-n')$ in front of the bracket. Carrying out a four-fold sum for every l from $l = 1$ to $l = N$, and then repeating the process for a range of values of chain lengths N, is an excessive amount of computation even for an electronic computer. Fortunately, there is no need to be that inefficient about the computation. The functions H, E, and Δ inside the bracket are so simple that it is possible to re-order the summations much more efficiently. We collect together terms with the same value of d, all of which get multiplied eventually by p_d. Within that group of terms we collect together terms with the same $k = m+n$, and with the same $k' = m'+n'$. A given d and $k = m+n$ ensures that the functions H, E, and Δ inside the bracket have the same values for this group of terms, and a given d and $k' = m'+n'$ ensures that we are considering a definite fragment length $l = d+k'$. Clearly k is greater than or equal to k'. The details of this re-ordering of summations are in the nature of coding technique and need not detain us here. Suffice it to say that a computer code called SMASH has been written, in FORTRAN, to evaluate these expressions, and gives numerical results in a reasonable amount of computer time.

9. Weight fractions from infinite chains

In this section, we establish the asymptotic forms of the expressions of § 8, in the limit of large chain length N.

In this limit, terms proportional to $H(N-n-1)$ or $H(N-m-d-n-1)$ become proportional to N (see (4.6)), whereas all other terms approach N-independent constant values. Thus we can obtain the limiting expressions formally by replacing H, wherever it appears, by N, and E by zero; we also ignore the last term in the bracket of (8.8).

It is useful to introduce the notation:

$$U(m', \mu') = \sum_{m=m'}^{\infty} T(m, 4, m', \mu,) = \sum_{m=m'}^{\infty} v_m X(m, 4, m', \mu'). \tag{9.1}$$

The sum goes to infinity since, in an infinitely long chain, L-holes of arbitrarily large size m may appear, though with ever-decreasing probability. Only the L-holes, $\mu = 4$, contribute in the limit of infinite N.

The right-hand form (9.1) shows that $U(m', \mu')$ depends on the constitution of the chain through the factors v_m, see (3.9), and on the break-up probabilities through the X-coefficients of §§ 5 and 6. Since the X-coefficients are independent of the chain constitution, and settle down fairly rapidly to predictable values (see § 7), the infinite sums in (9.1) can be approximated effectively by going to m of order $n_0 = 10M_4$, and using the asymptotic formulas (7.3) and (7.7) thereafter. The sums from n_0 to infinity are simply geometric sums.

Using these definitions, we obtain from (8.6)

$$\lim_{N\to\infty} \frac{R_1(N, l)}{N_0} = \delta U(l, 1), \tag{9.2}$$

and from (8.8)

$$\lim_{N\to\infty} \frac{R_0(N, l)}{N} = \delta \sum_{m', d, n'} p_d \Delta(l-m'-n'-d) U(m', 3) U(n', 2). \tag{9.3}$$

As an example of the sort of rearrangement of summations which we mentioned at the end of the preceding section, we give the rearranged form of (9.3):

$$\lim_{N\to\infty} \frac{R_0(N, l)}{N} = \delta \sum_{d=1}^{l} p_d \sum_{m'=0}^{l-d} U(m', 3) U(l-d-m', 2). \tag{9.4}$$

For computational purposes, we evaluate the interior sums in (9.4) once and for all, for all relevant values of $k' = l-d$, and store the results in the fast memory. These stored values are then used for every l, without having to recompute the interior sums. The ranges of summation in (9.4) are overestimates, since we have not taken account of the condition that the final

observed fragments must be stable against enzymatic action. That is,

$$U(m', \mu') = 0 \text{ unless } m' < M_{\mu'}. \tag{9.5}$$

This condition is of course a direct consequence of (9.1) and (5.12), and is thus contained implicitly already. However, its explicit use limits the ranges of the summations in (9.4), as follows: the first factor $U(m', 3)$ in the interior sum vanishes unless:

$$0 \leqq m' \leqq M_3 - 1, \tag{9.6}$$

whereas the second factor $U(l-d-m', 2) = U(k'-m', 2)$ vanishes unless

$$0 \leqq k'-m' \leqq M_2 - 1. \tag{9.7}$$

In both cases, the lower limits can be sharpened up some more if explicit assumptions are made about the breaking rules for L-holes. As written, $m' = 0$ in (9.6) implies a break immediately adjacent to the right-hand D which terminates the L-hole, and $k'-m' = 0$ in (9.7) implies a break immediately adjacent to the left-hand D which terminates the L-hole. We shall not write down the more stringent conditions which arise if such extreme breakpoints are forbidden by the breaking rules.

We rewrite the inequality (9.7) as an inequality for m', and combine the resulting condition with (9.6). This yields the combined inequality

$$\text{Max } (0, k'+1-M_2) \leqq m' \leqq \text{Min } (M_3-k, k'). \tag{9.8}$$

Not only does (9.8) limit the range of summation over m' in (9.4) for a given $k' = l-d$; but it also limits the range of possible k'. If k' becomes large, the lower limit on m', according to (9.8) equals $k'+1-M_2$, whereas the upper limit on m' equals M_3-1. The lower limit exceeds the upper limit (and thus the sum becomes zero) unless k' is limited by

$$k' \leqq M_2+M_3-2. \tag{9.9}$$

Thus the interior sums in (9.4) vanish, and need not be computed or stored, if condition (9.9) is violated. Furthermore, the sum over d in (9.4) is correspondingly limited: for large l, the lower limit on d is not 1, but rather the lowest possible $l-k'$, namely $l-M_2-M_3+2$.

All these limitations have a highly beneficial effect on storage space and computing time for the explicit evaluation of these expressions by SMASH. Corresponding, though somewhat less straightforward, simplifications apply to the evaluation of the expressions of the preceding section.

The conversion of (9.2) and (9.4) to asymptotic expressions for weight fractions is a straightforward matter. We introduce the notation:

$$K = M_2+M_3-2 \tag{9.10}$$

https://doi.org/10.1017/S1446788700004614 Published online by Cambridge University Press

for the maximum number of "dangling L's" which appears in (9.9). Taking the limit of (8.9) and (8.10) as N goes to infinity then gives:

$$\lim_{N\to\infty} W(N, l) = l\delta\left[U(l, 1)+\sum_{k'=0}^{\text{Min }(K,\, l-1)} p_{l-k'}Q_{k'}\right], \tag{9.11}$$

where

$$Q_{k'} = \sum_{m'=\text{Max }(0,\, k'+1-M_3)}^{\text{Min }(k',\, M_3-1)} U(m', 3)U(k'-m', 2). \tag{9.12}$$

In these expressions, k' can be interpreted to mean the total number of outer L's attached to the D-string of length $d = l-k'$. Of these k' L's, m' are to the left of the D-string, and $k'-m'$ are to the right of the D-string. $Q_{k'}$ is the probability, for an infinite chain, that the emergent D-containing fragments will have exactly k' outer L's attached. The contribution $U(l, 1)$ in the bracket of (9.11) accounts for the pure-L fragments; this vanishes for all but the first few values of l, namely for all $l \geqq M_1$.

In the computer programme SMASH, the $U(m', \mu')$ are evaluated first, and stored, making use of the asymptotic forms of § 7 to shorten the labour. Next, the $U(m', 3)$ and $U(k'-m', 2)$ are combined into quantities $Q_{k'}$ according to (9.12), and these $Q_{k'}$ are stored. Finally the limiting weight fractions are evaluated from (9.11). A plotting routine is used to get a rough graph of these asymptotic weight fractions against l.

10. *D*-length probabilities and related quantities for an infinite chain

For the sake of completeness, as well as for the intrinsic usefulness of the results, we conclude this paper by giving generalizations of the expressions of the first paper of this series, reference [2], to the arbitrary breaking rules allowed in the present work.

The work of reference [2] dealt with D-length probabilities p_d, probabilities of gap (i.e., L-hole) lengths, and with expectation values of the D-length and the gap length. It should be noted that all these quantities are independent of all but one of the parameters which are involved in the breaking rules: this one parameter is M_4, the minimum breakable L-hole length. All other parameters in the breaking rules are involved only if we wish to determine probabilities for various numbers of outer L's attached to the D-string in the final fragment, and to determine the distribution-in-weight of the pure-L final fragments. Thus, the only parameters of interest to us in this section are M_4, the minimum breakable L-hole size, and the chain constitution parameters λ and h.

We write the recursion relation (3.11) in the form

https://doi.org/10.1017/S1446788700004614 Published online by Cambridge University Press

$$p_d = \sum_{k=1}^{M_4} c_k p_{d-k}, \quad d = 2, 3, 4, \cdots, \tag{10.1}$$

where the coefficients c_k are given by

$$c_1 = \pi_{22}, \quad c_k = \pi_{21}(\pi_{11})^{k-2}\pi_{12}, \quad k = 2, 3, 4, \cdots. \tag{10.2}$$

The initial conditions for (10.1) are (3.12), i.e., $p_1 = 1$ and $p_n = 0$ for non-positive n.

The quantities p_d are not normalized to unit sum. It can be shown that the normalized D-length probabilities P_d are

$$P_d = \pi_{21}(\pi_{11})^{M_4-1} p_d = P_1 p_d. \tag{10.3}$$

The factor has a simple interpretation: in order that d constituents, of which the left-most is the D which starts a D-string, actually be a D-string of length exactly d, we need two things: (1) the d constituents themselves must form a D-string, factor $= p_d$, and (2) this sequence of d constituents must be followed by an L-hole, i.e., by at least M_4 consecutive L's; this is the other factor in (10.3). The second form uses the fact that $p_1 = 1$.

Next, the average D-length of D-strings is given by

$$\bar{d} = 1 + \frac{1}{P_1} \sum_{k=1}^{M_4} k c_k. \tag{10.4}$$

This reduces, in the special case $M_4 = 3$, to the expression derived in reference [2].

Finally, the average gap length (L-hole length) is given by

$$\bar{l} = \sum l v_l / \sum v_l = \frac{\pi_{11}}{\pi_{12}} + M_4. \tag{10.5}$$

The ratio π_{11}/π_{12} in (10.5) is the "average excess gap length" of reference [2], M_4 being the minimum gap length before a group of adjacent L's is called a gap at all.

We are grateful to Professor A. Berger for an introduction to this problem and for many valuable and informative discussions throughout the course of this work. We wish to thank the Weizmann Institute, Rehovoth, Israel, for the grant of a John F. Kennedy Senior Fellowship during the tenure of which this work was done.

References

[1] I. Schechter and A. Berger, *Israel Journal of Chemistry* **3** (1965), 98.
[2] J. M. Blatt, to appear.

Weizmann Institute
Rehovoth, Israel