

On Eulerian Circuits and Words with Prescribed Adjacency Patterns

JOAN P. HUTCHINSON AND HERBERT S. WILF*

*Department of Mathematics, Dartmouth College, Hanover, New Hampshire 03755,
Department of Mathematics, University of Pennsylvania, Philadelphia, Pa. 19104, and
Department of Mathematics, Rockefeller University, New York, N.Y. 10021*

Communicated by Mark Kac

Received January 2, 1974

In the study of the biochemistry of the DNA molecule [1-3], of the statistical mechanics of large molecules in general [4], and elsewhere, one is led to postulate models of behavior in which the molecule is treated like a "word," and the individual bases arranged on the molecule are the "letters." A useful simplifying assumption is then that the information-carrying properties of these molecules depend only on (a) the number of letters of each type and (b) a nearest-neighbor interaction in which the frequency of each letter pair is relevant, but triples,..., can be ignored. In such a model, one is soon led to consideration of the following purely combinatorial question:

Let v_i ($i = 1, \dots, n$) be given positive integers, and let v_{ji} ($i, j = 1, \dots, n$) be given nonnegative integers. How many words can be made from an alphabet of n letters, in such a way that the letter i appears exactly v_i times in the word ($i = 1, \dots, n$), and exactly v_{ij} times the letter i is followed by the letter j ($i, j = 1, \dots, n$)?

We deal with this question both in the form given above, in which case we are able to give an exact, closed solution, and in the symmetric form, in which the matrix elements v_{ij} represent the number of occurrences of the *unordered* adjacent pair ij in the word, so that $v_{ij} = v_{ji}$ ($i, j = 1, \dots, n$), where we cannot give a complete solution, but can only relate the solution to a well-known unsolved problem of considerable difficulty.

With reference to the problem stated above, our solution is that the number of words satisfying the conditions is exactly

$$N = \left\{ \prod_{i=1}^n (v_i - 1)! \right\} \left\{ \prod_{i,j=1}^n (v_{ij}!) \right\}^{-1} \det(v_i \delta_{ij} - v_{ij})_{i,j=1}^n, \quad (1)$$

* Research carried out under John Simon Guggenheim Memorial Fellowship.

if one of the consistency conditions (5), (9) is met, and there are no solutions otherwise.

As an example, suppose

$$\begin{aligned} \nu_1 = 4, \quad \nu_2 = 2, \quad \nu_3 = 1, \quad \nu_{11} = 1, \quad \nu_{12} = 2, \quad \nu_{13} = 0, \\ \nu_{21} = 1, \quad \nu_{22} = 0, \quad \nu_{23} = 1, \quad \nu_{31} = 1, \quad \nu_{32} = 0, \quad \nu_{33} = 0. \end{aligned}$$

Then (1) yields

$$\begin{aligned} N &= \{6\}\{2\}^{-1} \det \begin{pmatrix} 3 & -2 & 0 \\ -1 & 2 & -1 \\ -1 & 0 & 1 \end{pmatrix} \\ &= 6. \end{aligned}$$

The six solution-words are:

AABABCA,
AABCABA,
ABAABCA,
ABABCAA,
ABCAABA,
ABCABAA.

To prove (1), we first reduce the problem to the case where all $\nu_{ii} = 0$, i.e., blocks of letters of length >1 do not appear. Indeed, for each solution of the reduced problem we obtain

$$\prod_{i=1}^n \binom{\nu_i - 1}{\nu_{ii}} \tag{2}$$

solutions of the original problem by replacing the j th appearance of letter i by a block of i 's of length r_j where $r_j > 0$ ($j = 1, \dots, \nu_i - \nu_{ii}$) and

$$\nu_i = r_1 + r_2 + \dots + r_l \quad (l = \nu_i - \nu_{ii}). \tag{3}$$

The number of representations (3) with all $r_i \geq 1$ is well known to be $\binom{\nu_i - 1}{\nu_{ii}}$, and (2) follows.

We therefore concern ourselves only with the reduced problem.

Next, with the given data we associate a directed multigraph G , as follows: The vertices of G are the n letters $1, 2, \dots, n$. There are exactly ν_{ij} directed arcs from i to j ($i, j = 1, \dots, n$) in G .

Consider, now, a solution-word

$$w = i_1, i_2, \dots, i_q \tag{4}$$

of our problem, and suppose first that $i_1 \neq i_q$. Then the word w corresponds to an Euler path on the edges of G , beginning at vertex i_1 , and ending at i_q ; i.e., each arc of G is used exactly once in the walk $i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_q$ on the arcs of G . By the well-known theorem of I. J. Good [5], G must satisfy the conditions for the existence of an Euler path, namely

$$(a) \quad \sum_{k=1}^n v_{ik} = \sum_{k=1}^n v_{ki} + \delta_{ii_1} - \delta_{ii_q} \quad (i = 1, 2, \dots, n)$$

$$(b) \quad v_i = \sum_{k=1}^n v_{ki} + \delta_{ii_1} \quad (i = 1, 2, \dots, n)$$

Hence conditions (5) are necessary and sufficient for the existence of a solution-word w in which $i_1 \neq i_q$, and if (5) is satisfied all solution words begin with the same letter i_1 and all end with the same letter i_q .

We remark that conditions (5) are valid for the unreduced problem as well, and can therefore be applied directly to the original data.

An Euler path on G corresponds to a unique solution-word w , namely the vertex-sequence along the path. Conversely, a solution word w corresponds to many Euler paths. In fact, if

$$\mathcal{P}: e_1, e_2, \dots, e_E \quad (6)$$

is an Euler path corresponding to w of (4), fix any pair of letters i, j ($i \neq j$). There are, in the sequence (6), v_{ij} edges which join vertex i to vertex j . The order in which these v_{ij} edges appear in \mathcal{P} may be permuted arbitrarily, and w will remain invariant. Therefore, to w correspond exactly

$$\prod_{i \neq j} (v_{ij}!) \quad (7)$$

different Euler paths.

It remains to count the Euler paths of G . Adjoin to the graph G a new vertex x . Draw a directed edge from x to i_1 and from i_q to x . There is a 1-1 correspondence between Euler paths on G and Euler circuits on this new graph G_x . The number of Euler circuits on a directed graph is given by the theorem of de Bruijn-Ehrenfest, Smith-Tutte [6, 7], the "BEST" theorem, which in this case gives

$$\prod_{i=1}^n (v_i - v_{ii} - 1)! \det(v_i \delta_{ij} - v_{ij})_{i,j=1}^n \quad (8)$$

for the number of Euler circuits on G_x , and therefore for the number of

paths on G , also. If we combine (2), (7), (8), we obtain (1), which settles the case where $i_1 \neq i_q$ in all solution words, i.e., where (5) holds.

Suppose, now, that in a solution word w , of the form (4), we have $i_1 = i_q$. Then corresponding to w is an Euler circuit on G , and Good's [5] theorem implies that the conditions

$$\begin{aligned}
 \text{(a)} \quad & \sum_{k=1}^n \nu_{ik} = \sum_{k=1}^n \nu_{ki} \quad (i = 1, 2, \dots, n), \\
 \text{(b)} \quad & \nu_i = \sum_{k=1}^n \nu_{ik} + \delta_{ii_1} \quad (i = 1, \dots, n),
 \end{aligned}
 \tag{9}$$

hold. Hence, conditions (9) are necessary and sufficient for the existence of a solution-word w in which $i_q = i_1$, and if (9) is satisfied, all solution words begin and end with the same letter i_1 . If the data of the problem satisfy neither (5) nor (9), there are no solutions.

Again, conditions (9) also apply directly to the original, unreduced problem.

To find the number of solution-words in the case (9), we adjoin a new letter x , with the data $\nu_x = 1$, and

$$\nu_{i,x} = 0 \quad (i = 1, \dots, n); \quad \nu_{x,i} = \delta_{i,i_1} \quad (i = 1, \dots, n); \quad \nu_{x,x} = 0.$$

Any solution-word of the modified problem is of the form

$$x, i_1, i_2, \dots, i_q$$

in which we must have $i_q = i_1$. Indeed the associated graph is of Eulerian-path type in which x is the initial vertex and i_1 is the terminal vertex of all paths because it is the unique vertex whose in-valence is 1 greater than its out-valence. Thus any solution of the modified problem is of the form (x, w) , where w is a solution of the original problem. The modified problem is of the type (5) which we have already counted by (1). If we apply this result to the modified problem and expand by minors down column x , we obtain the formula (1) again, completing the proof.

We turn now to the symmetric form of the problem, where our answers are much less complete. As before, we have ν_i appearances of i in our words, and ν_{ij} appearances of the adjacent unordered pair ij ($i, j = 1, \dots, n$). The number of solution-words is now given by

$$N = \left\{ \prod_{i=1}^n \binom{\nu_i - 1}{\nu_{ii}} \right\} \left\{ \prod_{i < j} (\nu_{ij}!) \right\}^{-1} E(G), \tag{10}$$

if the consistency conditions (11) are met ($N = 0$ otherwise), where G

is the undirected multigraph associated with the reduced problem, and $E(G)$ is the number of Euler paths on G between i_1 and i_q (i_1 and i_q defined by (11)), or the number of Euler circuits on G (if $i_1 = i_q$). (We regard Euler circuits which differ by only a cyclic permutation of edges as distinct.)

The consistency conditions are well-known from Euler's original solution of the "Königsberg bridges" problem. They are

$$\sum_{k=1}^n v_{ik} = 2v_l - v_{il} - \delta_{l,i_1} - \delta_{l,i_q} \quad (l = 1, \dots, n). \quad (11)$$

Equations (11) define the indices i_1, i_q from the input data. If $i_1 = i_q$, all words begin and end with i_1 , while if $i_1 \neq i_q$ half of the solution words will begin with i_1 and end with i_q , and the other half will be obtained from the first by reversing the sequence of letters.

There are no formulas for $E(G)$. Indeed $E(K_{2n+1})$ is unknown. Our contribution to this part of the problem will consist in giving the exact solution for the cases of $n = 2, 3$, or 4 letters. We note further that two-sided *inequalities* for $E(G)$ in the symmetric case can be found in Tutte [8].

First, if $n = 2$ it is easy to see that there are

$$N = \binom{v_1 - 1}{v_{11}} \binom{v_2 - 1}{v_{22}}$$

solutions if $i_1 \neq i_q$ in (11), or

$$N = \binom{v_1 - 1}{v_{11}} \binom{v_2 - 1}{v_{22} - 1}$$

solutions if $i_1 = i_q$. We follow the convention here, that $\binom{n}{m} = 0$ unless $0 \leq m \leq n$.

Now suppose $n = 3$, and consider the reduced problem ($v_{ii} = 0, i = 1, 2, 3$). From a solution-word w we can remove all occurrences of some fixed letter, m . The resulting word of 2 letters will be called a *preword*. That is, a preword is a word of 2 letters to which v_m m 's can be added so as to make a solution-word of the reduced problem on 3 letters. Our aim is to count the number of prewords and then to see in how many ways the missing letter can be added to make a solution-word.

We assert that for $n = 3$, given parameters which satisfy (11) with $i_1 \neq i_q$ and with $v_{kk} = 0, k = 1, 2, 3$, the number of prewords on letters i_1 and i_q is given by

$$\sum_{l=0}^{\infty} \binom{v_{i_1} - 1}{v + l} \binom{v_{i_q} - 1}{v + l}, \quad (12)$$

where $v = \lfloor v_{i_1 i_q} / 2 \rfloor$.

Now a suitable preword is a word in the letters i_1 and i_q which begins with i_1 , ends with i_q , and which has i_1 and i_q adjacent at least $\nu_{i_1 i_q}$ times. Let

$$\nu' = 2[\nu_{i_1 i_q}/2] + 1 = 2\nu + 1.$$

Then $\nu' \geq \nu_{i_1 i_q}$ and we create all possible words which have i_1 and i_q adjacent $\nu', \nu' + 2, \nu' + 4, \dots$ times. There are

$$\sum_{l=0}^{\infty} \binom{\nu_{i_1} - 1}{(\nu' - 1 + 2l)/2} \binom{\nu_{i_q} - 1}{(\nu' - 1 + 2l)/2} = \sum_{l=0}^{\infty} \binom{\nu_{i_1} - 1}{\nu + l} \binom{\nu_{i_q} - 1}{\nu + l}$$

such words.

We must check that every such word p can be extended to a solution word w . We add the third letter, k , to a word p as follows.

- (i) Whenever an i_1 or an i_q is adjacent to itself in p , a k must be inserted between the two identical letters.
- (ii) All the remaining k 's must be inserted between an i_1 and an i_q in p .

In a preword p there are

$$\nu_m - \frac{1}{2}(\nu' + 2l + 1)$$

occurrences of adjacent m 's ($m = i_1, i_q$), and thus

$$x = \nu_k - \nu_{i_1} + \frac{1}{2}(\nu' + 2l + 1) - \nu_{i_q} + \frac{1}{2}(\nu' + 2l + 1)$$

of the k 's are inserted between an i_1 and an i_q .

Thus the resulting word w is a reduced word by (i), and we must only check the adjacencies resulting from (ii). We have

$$i_1 \text{ and } k \text{ adjacent } 2(\nu_{i_1} - \frac{1}{2}(\nu' + 2l + 1)) + x \text{ times,}$$

$$i_q \text{ and } k \text{ adjacent } 2(\nu_{i_q} - \frac{1}{2}(\nu' + 2l + 1)) + x \text{ times, and}$$

$$i_1 \text{ and } i_q \text{ adjacent } \nu' + 2l - x \text{ times.}$$

Using the conditions of (11) we see that these numbers are equal to the three numbers $\nu_{i_1 k}$, $\nu_{i_q k}$, and $\nu_{i_1 i_q}$, respectively. Thus w is a solution-word as claimed.

To count prewords which lead to solution-words of the form

$$w = i_1, \dots, i_1$$

the same procedure will produce the number

$$\sum_{l=0}^{\infty} \binom{\nu_{i_1} - 1}{\nu + l} \binom{\nu_j - 1}{\nu - 1 + l}, \tag{13}$$

for the number of prewords which use just the two letters i_1, j , where $j \in \{1, 2, 3\}, j \neq i_1$, and where $\nu = [(\nu_{i_j} + 1)/2]$.

We can now give the complete answer to the symmetric problem when $n = 3$. Given parameters which satisfy (11) with $i_1 \neq i_q$ the number of solution-words is given by

$$N = \left\{ \prod_{i=1}^3 \binom{\nu_i - 1}{\nu_{ii}} \right\} \left\{ \sum_{l=0}^{\infty} \binom{\nu_{i_1} - \nu_{i_1 i_1} - 1}{\nu + l} \binom{\nu_{i_q} - \nu_{i_q i_q} - 1}{\nu + l} \binom{2\nu + 2l + 1}{\nu_{i_1 i_q}} \right\}, \tag{14}$$

where $\nu = [(\nu_{i_1 i_q} + 1)/2]$. When $i_1 = i_q$, we have

$$N = \left\{ \prod_{i=1}^3 \binom{\nu_i - 1}{\nu_{ii}} \right\} \left\{ \sum_{l=0}^{\infty} \binom{\nu_{i_1} - \nu_{i_1 i_1} - 1}{\nu + l} \binom{\nu_m - \nu_{mm} - 1}{\nu + l - 1} \binom{2\nu + 2l}{\nu_{i_1 m}} \right\} \tag{15}$$

in which m is either of the two elements of $\{1, 2, 3\} - \{i_1\}$, and where $\nu = [(\nu_{i_1 m} + 1)/2]$.

In both formulas the initial product of binomial coefficients results from unreduced words as we have seen in (2).

To establish (14) and (15) we must count the number of ways in which a given preword can be extended to a solution-word. Notice that in the count of prewords, only in step (ii) did we have some freedom of choice in inserting the third letter. In the case of (14) for a fixed $l \geq 0$, we have

$$\begin{aligned} x &= \nu_k - \nu_{i_1} + \frac{1}{2}(\nu' + 2l + 1) - \nu_{i_q} + \frac{1}{2}(\nu' + 2l + 1) \\ &= 2\nu + 1 + 2l - \nu_{i_1 i_q} \end{aligned}$$

k 's to insert into $\nu' + 2l = 2\nu + 1 + 2l$ possible positions; thus we have

$$\binom{2\nu + 2l + 1}{\nu_{i_1 i_q}}$$

choices. Similarly in the case of (15) we find we have

$$\binom{2\nu + 2l}{\nu_{i_1 m}}, \quad m \in \{1, 2, 3\}, \quad m \neq i_1$$

choices.

As shown above each such choice gives a solution-word. Furthermore, for a fixed preword each such choice gives a different solution-word, and

no solution-word can be created from 2 different prewords. Thus (14) and (15) follow.

We have also the solution for four letters, obtained by the same method, but we believe it to be too lengthy to warrant its appearance here.

REFERENCES

1. L. L. GATLIN, The Information Content of DNA, II, *J. Theoret. Biol.* **18** (1968), 181-194.
2. G. HUTCHINSON, Evaluation of polymer sequence fragment data using graph theory, *Bull. Math. Biophysics* (1969), 541-562.
3. J. E. MOSIMANN, M. B. SHAPIRO, C. R. MERRIL, D. F. BRADLEY, AND J. E. VINTON, Reconstruction of protein and nucleic acid sequences IV, *Bull. Math. Biophys.* **28** (1966), 235-260.
4. G. S. RUSHBROOKE AND H. D. URSELL, On One-Dimensional Regular Assemblies, *Proc. Cambridge Philos. Soc.* **44** (1948), 263-271.
5. I. J. GOOD, *J. London Math. Soc.* **21** (1947), 167-169.
6. N. G. DE BRUIJN AND T. VAN AARDENNE-EHRENFEST, *Simon Stevin* **28** (1951), 203-217.
7. C. A. B. SMITH AND W. TUTTE, On unicursal paths in a network of degree 4, *Amer. Math. Monthly* **48** (1941), 233-237.
8. W. TUTTE, *Connectivity in Graphs*, University of Toronto Press, Toronto, Canada, 1966.
9. G. KARREMAN, Cooperative specific adsorption, *Ann. N. Y. Acad. Sci.* **204** (1973), 393-409.
10. G. S. RUSHBROOKE, "Introduction to Statistical Mechanics," Clarendon Press, Oxford, 1949.