Geometrical constraints on Bennett's predictions of chromosome order

D. Dorninger and
W. Timischl

Division of Mathematical Biology,
Technical University of Vienna,
Wiedner Hauptstrasse 8-10/1183, A-1040 Vienna,
Austria.

Two aspects of the Bennett model are of interest from a mathematical point of view. First there is the question whether Bennett's "ranking method" for predicting the order of chromosomes will always work. The answer depends on the number n of chromosomes: If n is odd, predictions (being not necessarily unique) are possible in most cases. Secondly there is Bennett's procedure for determining the arrangement of chromosomes. It is shown that the method of minimising the perimeter of the polygon obtained by connecting the centromeres is only applicable if the positions of the n centromeres do not deviate too much from an arrangement along a regular n-gon.

INTRODUCTION

Much attention has been paid to Bennett's work on the disposition of chromosomes in a haploid genome which essentially includes (Bennett, 1982; Heslop-Harrison and Bennett, 1983a, b):

(a) The formulation of a method for predicting chromosome order. This method (shortly called method A) is based on the assumption that a haploid chromosome set is arranged in a chain in which "similar"-size long or short arms are paired.

(b) An approach for testing method A (and the underlying model). For this purpose a second method (shortly called method B) for determining the arrangement of chromosomes (by taking the 3-dimensional coordinates of each centromere and connecting the centromeres by a polygon such that the perimeter is minimized) is applied and a "statistical test" is performed in order to compare the results following from method A and B, respectively.

Different aspects of the statistical work have been examined in a recent publication, and it is obvious that the statistical analysis performed is not satisfactory (Callow, 1985). The problem is, however, not only if there is a correspondence between two experimental methods; there are other fundamental questions remaining unsolved in the formulation of the model. In particular it is not clear from Bennett's paper if the application of method A will always (*i.e.*, for all possible observation data) enable a prediction of chromosome disposition. It is also not obvious to see if method B is suited at all to detect the "real" ordering of centromeres (otherwise there is no justification to test method A against method B). The following sections contain a discussion of these points and it is hoped that the results obtained may help to clarify some formal aspects of Bennett's model.

INTERPRETATION OF METHOD A

Bennett summarises his method for predicting the order of n chromosomes in a haploid set as follows: "... all the long arms of chromosomes and all the short arms are separately ranked in descending order of size and those that are similar are then paired.... The pairing of arms is such that it is possible to arrange the chromosomes in a single unbroken chain (or polygon) with arms of each chromosome joined at their centromere and placed adjacent to other chromosome arms with most similar size." From this text which is taken from Heslop-Harrison and Bennett (1983a, p. 213) it obviously follows that the chromosomes should be arranged in such a way that either short or long chromosome arms are paired, *i.e.*, placed adjacent (hypothesis 1). Assuming this as a principle for chromosome disposition means that for an even n there are $(n-1)!/2$ different ways of ordering a

haploid set along a closed chain (*e.g.*, $(n-1)!/2 = 60$ for $n = 6$); each chain then defines a partition in the set of long (short) arms into $n/2$ pairs. If n is an odd number no closed loop-arrangement of chromosomes is possible; there is a "discontinuity" in the chain because one short and one long chromosome arm are left which cannot be paired; each chain now defines a partition in the set of long arms as well as in the set of short arms into $(n-1)/2$ pairs and one single arm. The number of different chromosome arrangements is now given by $n!$ (*e.g.*, $n! = 5040$ for $n = 7$).

The number of possible orders under hypothesis 1 is reduced if as a second principle (hypothesis 2) only "similar" short and "similar" long arms are allowed to be paired. In order to find the "similar" arms two ranking sequences of chromosomes are introduced by Bennett, one with respect to the short arms of the chromosomes and another with respect to the long arms. Ranking of chromosomes (in descending order) with respect to their short arms yields a certain sequence S of short chromosome arms. Of course, we are free to number the chromosomes in such a way that chromosome 1 has the longest short arm, chromosome 2 the second longest short arm etc.; thus S can be written as $S = 1, 2, \ldots, n$. A second sequence L is obtained by ranking the chromosomes with respect to their long arms. If l_1 is the identifier of the chromosome with the longest long arm, l_2 the identifier of the chromosome with the second longest long arm *etc.*, L can symbolically be expressed by $L = l_1, l_2, \ldots, l_n$.

It is obvious that "similar" arms should be adjoining in S and L, respectively. However, what "similar" really means is nowhere precisely expressed in Bennett's papers. We use the following definition: Two short or long arms are similar if one arm succeeds or precedes the other in S and L, respectively. It follows that the first (last) arms can form a similar pair only with the second largest (shortest) arm. All other arms have two neighbours, *i.e.*, they can be combined with the preceding or succeeding arm in order to give a similar pair.

PREDICTIONS FOR EVEN AND ODD NUMBER OF CHROMOSOMES

For an even number n of chromosomes there is exactly one partition into $n/2$ similar pairs in both chains L and S. In this case a pairing procedure based on "most similar" instead of "similar" pairs cannot be accepted as a general rule, because then no partition into $n/2$ "most similar" pairs might

exist (and consequently no arrangement of chromosomes along a closed chain). If, *e.g.*, $n = 4$ and the differences in size within $L = l_1, l_2, l_3, l_4$ are 2, 1 and 3 units between l_1, l_2 and l_2, l_3 and l_3, l_4 respectively, then l_3 is most similar to l_2 and hence no partition into pairs of two neighbouring elements is possible. Even if there exists a partition into $n/2$ similar pairs, this does not imply that the required arrangement of chromosomes is possible. If *e.g.*, $S = 1, 2, 3, 4$ and $L = 4, 3, 2, 1$ then we have the similar pairs $(1, 2)$, $(3, 4)$ of short arms and the similar pairs $(4, 3)$, $(2, 1)$ of long arms. According to this pairing, chromosomes 1 and 2 and chromosomes 3 and 4 form two distinct closed chains and not one single circle as required. Therefore, in order to satisfy hypothesis 1, it is necessary to use another concept of similarity in hypothesis 2 (*e.g.*, one might regard an arm as being similar to the two preceding or succeeding arms within the ranking sequences).

We now turn to (haploid) genomes with an odd number of chromosomes (all examples in Bennett's papers refer to this case). The essential question again is if for $S = 1, 2, \ldots, n$ and any given sequence $L = l_1, l_2, \ldots, l_n$ the chromosomes can be arranged in a chain according to hypotheses 1 and 2. A chain of this type will be represented by $H = h_1, h_2, \ldots, h_n$ where h_i denotes the number of the chromosome at position i and h_1 is assumed (without loss of generality) to be placed with its short arm being unpaired (thus the discontinuity of the chain is between chromosomes h_1 and h_n). The situation can be expressed more precisely by introducing a graph G which consists of n vertices $1, 2, \ldots, n$ (representing the identifiers of the chromosomes) and two types of edges referring to the kind of linkage between arms in S and L. If the chromosomes i and j are adjacent in S the vertices i and j are connected by an "S-edge", and if they are adjacent in L, there will be an "L-edge". Each chain of type H can then be regarded as a Hamiltonian path (cf., *e.g.*, Christofides, 1975) in G which is alternating in L- and S-edges. The problem stated above is to check if G has an alternating Hamiltonian path H. If such a path exists, the graph G (and also the sequence L) will be called admissible.

Every alternating Hamiltonian path $H = h_1, h_2, \ldots, h_n$ (being assumed w.l.o.g. to start with an L-edge) defines partitions of S and L into $(n-1)/2$ pairs of similar arms and one single arm each (the remaining short arm belongs to chromosome h_1, the long arm to chromosome h_n). Because of the fact that chromosomes with similar arms have subsequent identifiers in S and H starts with

an L-edge, H is characterised by the condition

$$|h_{2k} - h_{2k+1}| = 1 \ (k = 1, 2, \ldots, (n-1)/2). \quad (1)$$

For a specified n there are

$$N_H = \frac{n+1}{2} \left(\frac{n+1}{2}\right)! \, 2^{(n-1)/2}$$

permutations of the numbers $1, 2, \ldots, n$ satisfying (1) and thus N_H different alternating Hamiltonian paths H. The corresponding number of different sequences L is given by $N_L = n!$. It should be observed that for an admissible sequence L different Hamiltonian paths can exist: e.g., for a haploid complement of Sultan barley, Heslop-Harrison and Bennett (1983a) obtained the sequence $L = 6, 5, 2, 1, 3, 4, 7$ ($S = 1, 2, 3, 4, 5, 6, 7$) from which the alternating Hamiltonian path $H = 3, 1, 2, 5, 4, 7, 6$ was derived. (Attention should be paid to the fact that our chromosome numbers 1, 2, 3, 4, 5, 6, 7 correspond to the numbers 1, 2, 4, 6, 3, 7, 5 in Bennett's notation.) Another alternating Hamiltonian path is given by $5, 2, 1, 3, 4, 7, 6$. On the other hand two different sequences L (e.g., $L_1 = 6, 5, 2, 1, 3, 4, 7$ and $L_2 = 7, 4, 3, 1, 2, 5, 6$) can generate the same alternating Hamiltonian path (e.g., $H = 3, 1, 2, 5, 4, 7, 6$).

Condition (1) can be used to examine if a chain H according to hypothesis 1 and hypothesis 2 exists for a specified number n of chromosomes and a given sequence L. In order to find the number N_N of inadmissible sequences L for a given n we established a computer algorithm which is based on (1). From this algorithm the results in table 1 were obtained.

For $n < 7$ all sequences L are admissible; for $n \geqq 7$ there are always inadmissible sequences L (which can be proved by theoretical arguments), and the percentage of inadmissible sequences remains small for low haploid chromosome numbers (which is shown by computer calculations). An inadmissible sequence for $n = 7$ is e.g., $L = 6, 7, 3, 4, 5, 1, 2$. Thus it can be concluded that method A—provided n is an odd number—will practically

Table 1 Number N_L of possible sequences L, number N_N of inadmissible L, percentage P_N of inadmissible L and number N_H of different Hamiltonian paths for some different chromosome numbers n

n	N_L	N_N	P_N	N_H
5	120	0	0·0	24
7	5,040	8	0·16	192
9	362,880	896	0·25	1,920
11	39,916,800	135,712	0·34	138,240

work in most situations that occur. It is, however, in general not possible to get at a uniquely determined prediction of chromosome order without applying a further hypothesis which e.g., could be a minimum principle.

It was pointed out by Bennett that "the sum of the differences in size between arms predicted to be adjacent, other than the two arms at the discontinuity, ... should usually be the minimum possible for all ways of putting corresponding long and short arms together". Indeed, if all similar arms are assumed to have the same difference in size, the sum of differences is at a minimum for the chromosome order(s) following from hypotheses 1 and 2. This explains why Bennett's minimum principle holds at least approximately in many cases.

APPLICABILITY OF METHOD B

The following considerations refer to the problem of predicting the order of centromeres within a haploid set of n chromosomes on the basis of measurements of the positions of the centromeres. According to method B the spatial order of chromosomes at metaphase is reconstructed by taking the three dimensional coordinates (with respect to an arbitrary coordinate system) of each centromere and performing a principal component analysis. The first two principal components are introduced as new coordinate axes which are assumed to define the plane on which the centromeres are arranged at metaphase. If x_i is the position vector of centromere i with respect to the new system and if $p(j)$ represents the jth element in an arbitrary permutation p of $1, 2, \ldots, n$, then Bennett suggests to calculate the sum $D(p)$ of all distances $|x_{p(j)} - x_{p(j+1)}|$ with j running from 1 to n and $p(n+1)$ being $p(1)$. The permutation p for which $D(p)$ is at a minimum should reflect the "real" order of centromeres. (For the procedure described not all $n!$ permutations have to be calculated; there are only $(n-1)!/2$ different ways of ordering n centromeres in a closed polygon.)

There is no doubt that the principle of minimising $D(p)$ defines some "theoretical" order of centromeres which would also be obtained if the chromosomes were randomly dispersed. Then the theoretical order would only be a mathematical result without any real significance. So let us assume that there is a (highly) ordered disposition of chromosomes and ask what kind of "real" orders can be detected by method B. It is obvious that method B is applicable if the real order is such

Table 2 Percentage of the maximum allowable displacement $v(n)/d$ of the corners of a regular n-gon with side d (the percentages are truncated to integers)

n	6	7	8	9	10	15	20	30
$v(n)/d$ in per cent	28	31	34	36	37	41	42	44

that the n centromeres are positioned at the corners of a regular n-gon P_n (all adjacent corners having the same distance d) in a plane. Method B, however, also works if the regular arrangement is distorted up to a certain degree. We derived the following result for $n > 4$:

If n centromeres are arranged at the corners of a n-gon Q_n which is obtained by displacing the corners of P_n in any direction within the plane such that each displacement from the original points remains less than

$$v(n) = \frac{d(-sc + \sqrt{c^2 s^2 + c(1+c)(2+c)^2})}{(2+c)^2} \quad (2)$$

where

$$c = \cos(2\pi/n), \qquad s = \sin(2\pi/n),$$

then method B is applicable to detect the real order. Method B is not applicable if $v(n) > d/2$.

How much the regular arrangement can be distorted depends on the number n of centromeres. Table 2 contains some numerical results for the maximum allowable displacement related to d. It is seen that e.g., for $n = 7$ the displacement can be up to one third of d in any direction which allows a remarkably high deformation of the regular arrangement. It is not required that Q_n is a convex polygon; if, however, only such deformations are allowed which produce a convex Q_n, then according to a known result from discrete geometry displacements up to $d/2$ in magnitude are allowed (Barachet, 1957).

CONCLUSIONS

It was the object of this paper to analyse some aspects of Bennett's model for the disposition of chromosomes from a mathematical point of view. Our conclusions are:

(a) The application of method A will in most cases allow a prediction of chromosome order, if there is an odd number n of chromosomes. The prediction is, however, in general not uniquely determined. If n is an even number, in most cases no prediction is obtained by a rigorous application of method A. The method

depends very much on a suitable definition of "similar" chromosome arms.

(b) Method B is a numerical procedure for ordering n centromeres in a closed polygon. However, before any conclusions with respect to the real order of chromosomes can be drawn, one has to assume that there is such an order and that the extent of deviations from this order (caused, e.g., by the experimental procedure) is known.

All together we regard the Bennett model as an interesting approach for predicting the order of chromosomes on the base of rather formal principles. It is hoped that the mathematical inconsistencies at the present stage of modelling will be resolved by further developments of the model.

APPENDIX: PROOF OF FORMULA (2)

Let P_n be a regular n-gon with corners A_i ($i = 1, \ldots, n$) and assume that by displacing the corners of P_n a new polygon is obtained with corners B_i such that B_i remains within a circle C_i with radius ε around A_i for $i = 1, \ldots, n$ where $\varepsilon \leq v(n)$. In order to prove the applicability of method B it is sufficient to show that the distance b_{ij} of any point B_i from a neighbouring point B_j is always smaller than the distance b_{ik} from a point B_k which is not a neighbour of B_i.

W.l.o.g. we put $d = 1$ and take three subsequent corners of P_n (denoted by A_1, A_2, A_3). The A_i are assumed to be arranged in the (x, y)-plane as shown in fig. 1, which also contains the circles C_i around the A_i. We have to prove that $b_{12} < b_{13}$ for any choice of B_1, B_2, B_3 within C_1, C_2 and C_3, respectively. It is seen by geometrical considerations that $b_{12} < b_{13}$ whenever the same inequality holds for the special locations B_1', B_2' and B_3' (cf.
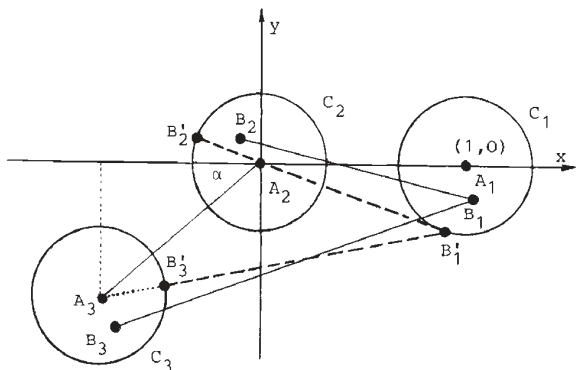
Figure 1 Illustration to the proof of formula (2).

fig. 1). So we can restrict to prove $b'_{12} < b'_{13}$ where b'_{1j} is the distance between B'_1 and B'_j ($j = 2, 3$).

Let $u = b'_{12} - \varepsilon$, $v = b'_{13} + \varepsilon$ and x, y be the coordinates of B'_1. Then $b'_{12} < b'_{13}$ is equivalent to $v > u + 2\varepsilon$. It is obvious from fig. 1 that $x^2 + y^2 = u^2$, $(x - 1)^2 + y^2 = \varepsilon^2$ and $v^2 = (x + \cos \alpha)^2 + (y + \sin \alpha)^2$. It follows that

$$v^2 = u^2 + (u^2 + 1 - \varepsilon^2) \cos \alpha + 2y \sin \alpha + 1. \quad (3)$$

Furthermore, because of $-\varepsilon \leqq y \leqq \varepsilon$, $\alpha = 2\pi/n$ and $n > 4$, the inequality

$$2y \sin \alpha \geqq -2\varepsilon \sin \alpha$$

holds which together with (3) gives

$$v^2 \geqq (u^2 + 1)(1 + \cos \alpha) - \varepsilon^2 \cos \alpha - 2\varepsilon \sin \alpha. \quad (4)$$

Let $R(u)$ denote the right side of inequality (4). In order to prove the inequality $v > u + 2\varepsilon$ we show that $R(u) - (u + 2\varepsilon)^2 > 0$, i.e.,

$$u^2 - \frac{4\varepsilon}{c} u + \frac{1 - 4\varepsilon^2}{c} + (1 - \varepsilon^2) - \frac{2s\varepsilon}{c} > 0, \quad (5)$$

where $c = \cos \alpha > 0$ and $s = \sin \alpha > 0$ because of $0 < \alpha = 2\pi/n < 2\pi/5$. Inequality (5) holds for all u if

$$D(\varepsilon) = \frac{4\varepsilon^2}{c^2} - \frac{1 - 4\varepsilon^2}{c} - (1 - \varepsilon^2) + \frac{2s\varepsilon}{c} < 0.$$

$D(\varepsilon)$ is a quadratic polynom in ε which is negative at $\varepsilon = 0$ and which has exactly one positive root, namely

$$v(n) = \frac{-sc + \sqrt{s^2 c^2 + c(1 + c)(2 + c)^2}}{(2 + c)^2}$$

thus $D(\varepsilon) < 0$ for $0 < \varepsilon < v(n)$.

REFERENCES

BARACHET, L. L. 1957. Graphic solution of the travelling salesman problem. *Oper. Res.*, *5*, 841-845.

BENNETT, M. D. 1982. Nucleotypic basis of the spatial ordering of chromosomes in eukaryotes and the implications of the order for genome evolution and phenotypic variation. Dover, G. A. and Flavell, R. B. (eds.) In *Genome Evolution*, Academic Press, London and New York, pp. 239-261.

CALLOW, R. S. 1985. Comments on Bennett's model of somatic chromosome disposition. *Heredity*, *54*, 171-177.

CHRISTOFIDES, N. 1975. *Graph Theory—An Algorithmic Approach*, Academic Press, London, New York and San Francisco.

HESLOP-HARRISON, J. S. AND BENNETT, M. D. 1983a. Prediction and analysis of spatial order in haploid chromosome complements. *Proc. R. Soc. Lond.*, B, *218*, 211-223.

HESLOP-HARRISON, J. S. AND BENNETT, M. D. 1983b. The spatial order of chromosomes in root-tip metaphases of *Aegilops umbellulata*. *Proc. R. Soc. Lond.*, B, *218*, 225-239.