ACCURACY AND ROBUSTNESS OF ANALYSES BASED ON NUMBERS OF GENES IN OBSERVED SEGMENTS

David Sankoff

Marie-Noelle Parent

David Bryant

We analyze a model of the distribution of linkage disruption points (breakpoints) along the chromosome. We calculate the variance of an estimator of the number of breakpoints and use this to assess whether or not Nadeau and Taylor were simply beneficiaries of the "luck of the draw". In addition, we analyze a possible source of error due to the availability of chromosomal assignment only rather than mapping data on humans. Finally, given evidence of frequent local rearrangements (e.g. inversions) of chromosomal segments, we suggest a method for evaluating the pertinence of comparative maps for the interchromosomal (translocational) history of a genome.

1 Introduction

Based on the very small number of homologous genes (only 83) whose chromosomal assignments were known in both mouse and man at the time, Nadeau and Taylor (1984) estimated the number of linkage disruptions which have occurred since the divergence of the two lineages. This estimate, 178 $\pm$39, has proved remarkably accurate, with modern estimates ranging from 130 (Sankoff et al., 1997b) to around 200 (Seldin, 1999), depending on definitional criteria[1]. The Nadeau-Taylor results, among the most significant in "...the history and development of the mouse as a research tool", established that "...the mouse genome is an extremely good model for the human genome..." (Pennisi, 2000).

In contrast to the 83 genes available in 1984, most of which were not mapped within the human chromosome to which they were assigned and many not even within the corresponding mouse chromosome, recent work is based on 1500 or more genes mapped in both genomes, e.g. Seldin (1999). The visionary result of Nadeau

[1] cf. Sankoff et al. (1997a) for a discussion of problems in delineating conserved segments.

and Taylor seems all the more remarkable in the intellectual climate of the early eighties, when there was little consensus about the existence of cross-species linkage conservation, and none about its quantitative characterization. Indeed, a skeptic might suggest that such an accurate prediction based on so few data attests more to fortuitously placed observation points rather than the inherent well-foundedness of the method (though no such suggestion seems to have found its way into the literature). In this note, we analyze a simple model (Sankoff et al., 1997b) of the distribution of linkage disruption points (breakpoints) along the chromosome. We calculate the variance of an estimator of the number of breakpoints and use this to assess whether or not Nadeau and Taylor were simply beneficiaries of the "luck of the draw". In addition, we analyze a possible source of error due to the availability of chromosomal assignment only rather than mapping data on humans. Finally, given evidence of frequent local rearrangements (e.g. inversions) of chromosomal segments, we suggest a method for evaluating the pertinence of comparative maps for the interchromosomal (translocational) history of a genome.

2 Conserved segments

In comparing two divergent genomes, a contiguous stretch of chromosome in which the number and order of homologous genes is the same in both species, i.e. where linkage has not been interrupted by any of the translocations, inversions or transpositions that may have occurred in either lineage, is called a *conserved segment*. We will refer to the two manifestations of each segment, one in each genome, as *reflexes* of each other. The number of conserved segments increases as they are disrupted by these types of inter- and intra-chromosomal rearrangement events, so that they tend to become smaller over time. Note that the segments are analytical constructs; they are defined only through the comparison of two genomes and have no physical manifestation in a single genome; indeed the set of conserved segments in a genome depends entirely on the reference genome to which it is being compared.

Ideally, conserved segments are discovered experimentally through the identification of one or more pairs of homologous genes in the two species, ordered in the same or reverse way in both genomes, bounded at both ends in each genome by genes whose known homologs do not continue this order in the other genome, and (if there are two or more genes in the segment) uninterrupted in each genome by other genes whose known homolog is absent from, or does not respect the same order in, the putative segment in the other genome.

3 The Nadeau-Taylor data

In counting the number of conserved segments, we must deal with underestimation due to conserved segments in which genes have not yet been identified in one or both species. This is particularly important if there are relatively few homologous genes in the data sets for a pair of species, so that many or most of the conserved

segments are not represented in the comparison. For example, out of the approximately 200 segments now thought to exist, Nadeau and Taylor could detect less than a third. As summarized in Table 1, their 83 genes fell into at least 46 (and at most 65) segments.

observation in mouse	number of genes	number of linkage groups*
mapped	54	36
chromosome assignment only	29	$\geq 10^{\dagger}$, ≤ 29
total	83	≥ 46, ≤ 65

Table 1: Level of detail of observation in mouse genome of 83 genes used in the Nadeau-Taylor study. Human homologs specified as to chromosomal assignment only.

* We use the terminology *conserved segment* in the text as it does not carry the connotation of any particular procedure, e.g. recombination experimentation, as a definitional criterion.

†The 29 genes unmapped in the mouse belong to only 10 different pairs of mouse and human chromosomes.

4 The model

Our formulation of the Nadeau-Taylor model assumes spatial homogeneity of breakpoints, i.e., that the n endpoints of the $n+1$ conserved segments (the linkage disruption sites or *breakpoints*) are uniformly and independently distributed along the combined length of all the autosomes contained in the genome. Little is lost in not distinguishing between breakpoints separating two segments and concatenation boundaries separating two successive chromosomes (Sankoff and Ferretti, 1995). We also assume spatial homogeneity of gene distribution and independence of gene positions, i.e., that there are m genes distributed uniformly and independently across the genome, and independently of the breakpoints. See Sankoff et al. (1997b) and, especially, Waddington et al. (2000) for the effects of relaxing the independence and homogeneity assumptions.

Recall that the problem is to estimate the number of breakpoints n or the number of conserved segments $n+1$, given only the data on segments that have been "observed" by virtue of containing at least one homologous mapped or sequenced gene in both genomes.

The probability $P(a,m,n)$ of observing a non-empty segments if there are m genes and n breakpoints is:

$$P(a,m,n) = \frac{\left(\begin{array}{c} m-1 \\ a-1 \end{array} \right) \left(\begin{array}{c} n+1 \\ a \end{array} \right)}{\left(\begin{array}{c} n+m \\ m \end{array} \right)} \tag{1}$$

Note that this model is equivalent to a classical occupancy problem of statistical mechanics (Feller, 1965, p. 62). As such, it makes no reference to the linear nature

of chromosomes; gene order enters only during the identification of segments, prior to statistical analysis.

The maximum likelihood estimate $\hat{n}$, given m and a, is the value of n which maximizes P. For given m and n, the expectation and the variance of $\hat{n}$ can be calculated making use of the probability distribution in Eq [1], though for $n \geq m - 1$, they have to be conditioned on $a < m$, since for $a = m$, the estimates are infinite, i.e., if every gene is located in a separate segment, the likelihood increases indefinitely with increasing n.

5 The estimates

If only the segment assignments of the 83 genes available to Nadeau and Taylor were known to us, the value of a would be between 46 and 65, as in Table 1, and $\hat{n}$ would fall between 98 and 275. The uncertainty resides in the 29 genes whose mouse map positions were unknown, but could have been contained in as few as 10 (i.e. if each of the 10 common mouse-human syntenies were each completely linked) and as many as 29 linkage groups (if no two genes were linked in both species), as in Table 2. Since the 54 genes mapped in the mouse data fell into 36 linkage groups (cf Table 1), it would not be unreasonable to expect the 29 unmapped ones to fall, proportionately into 19 segments, in which case $a = 55$ and $\hat{n} = 160$.

More important, were the true value of n equal to 160, the expectation and variance of $\hat{n}$ would be 164 and 1048, respectively, so that $\sigma = 32.4$. We may conclude from this that the accuracy of the Nadeau-Taylor estimate was inherent in their model and the data, and not due to a stroke of luck.

scenario	number of segments among unmapped genes	a	$\hat{n}$	$E(\hat{n}) \pm s.d.(\hat{n})$, assuming $n = \hat{n}$
minimum	10	46	98	99±16.7
proportionate	19	55	160	164± 32.4
maximum	29	65	275	287±71.7

Table 2: Analysis of reconstructed Nadeau-Taylor data according to model in Eq [1]. Last column indicates what could be expected were the true value of n equal to its estimate in the previous column.

6 Underestimation

A source of underestimation in the Nadeau-Taylor procedures follows from the possibility that two or more adjacent segments on a mouse chromosome are counted as one, despite the separate locations of their reflexes remote from each other on a single human chromosome, since this remoteness cannot be inferred from

assignment data alone. In our model, if there are a non-empty segments, the probability $Q(b, m, n, c)$ that only b of these segments will be counted because of this lack of linkage data in humans can be shown to be:

$$Q(b, m, n, c) = \sum_{a=b}^{n+1} P(a, m, n) \binom{a-1}{a-b} (1/c)^{a-b} (\frac{c-1}{c})^{b-1} \tag{2}$$

where $c = 22$ is the number of human autosomes and $P(a, m, n)$ is as in Eq [1]. Then for $m = 83$ and $b = 55$, the maximum likelihood estimator of n, based on Eq [2], is 184. This correction (from 178), while non-negligeable in our model, is not of a magnitude that would affect our evaluation of the Nadeau-Taylor approach. In a comparison of species with very few chromosomes, on the other hand, this correction could become proportionately much larger.

Even if mapping data were available from both of two species being compared, there is the possibility that two separate segments will be counted as one because no homologous pairs of genes have yet been discovered in any of the intervening segments in either genome. For the situation with twenty-odd chromosomes, this is a much smaller effect, and we will not give the details here.

A third concern about the Nadeau-Taylor calculation is that it that it could underestimate the large number of short segments produced by local rearrangement processes such as short inversions. The smaller sizes of newly discovered segments in the mouse-human comparison, however, are exactly in line with what is predicted from our model with complete uniformity of breakpoint and gene distributions (Nadeau and Sankoff, 1998): as m increases, the mean size of the segments remaining to be discovered is

$$\frac{G}{m + n + c}. \tag{3}$$

where G is the total length of the genome. So that as m increases from 100 to 2000, with $n + c = 200$, the length of undiscovered segments drops by a factor of around $\frac{1}{7}$, which parallels experience with mouse-human comparisons over the past 15 years.

Nevertheless, the possibility of a relatively frequent process of short inversions (McLysaght et al., 2000) in comparison with a slower rhythm of translocation, leads to the question of how to adapt the model account for both the inter- and intrachromosomal rearrangements within the same analysis (cf Schoen (2000)). This question is best addressed at a later stage of data acquisition than the one we have been discussing, when all or almost all the segments have been identified.

7 Translocations versus intrachromosomal rearrangement

Each rearrangement event determined by new breakpoints adds to the number of segments. Let the total number of segments on a mouse chromosome be

$$s = t + u + 1, \tag{4}$$

where t is the number due to translocations, and u the number due to local arrangements, and let c' be the number of human chromosomes that have at least one reflex on that mouse chromosome. Under a random translocation model we can predict how often two or more segments from the same human chromosome show up on the mouse chromosome through separate translocational events. Since

$$E(c') = c[1 - (1 - \frac{1}{c})^t]$$

then

$$\hat{t} = \frac{\log(c) - \log(c - c')}{\log(c) - \log(c - 1)} \tag{5}$$

is an estimator of t. To illustrate, for the 19 mouse autosomes, the data in Seldin (1999) indicate 192 segments with reflexes in human autosomes, while the sum of the c' is 99. Applying Eq [5] to each chromosome and summing the 19 values of $\hat{t}$ gives a total of 112 segments. In other words, in at most 13 cases, two segments from the same human chromosome are found on the same mouse chromosome *because of independent translocational events*. By Eq [4], this leaves unaccounted for $192-112-19 = 61$ segments, which must be attributed to local rearrangements such as inversion.

8 Discussion

In the early eighties, where an alternative hypothesis of random gene scrambling throughout the genome could not be excluded, Nadeau and Taylor were obliged to invoke a number of mathematical assumptions and approximations that, while justifiable, turn out to be unnecessary within our formulation of the key assumptions of spatial homogeneity and independence of breakpoint and gene distributions. For example, they required two or more genes linked on a mouse chromosome as a criterion for the existence of a conserved segment. At least two genes are necessary for the estimation of segment length, in cM, as a step towards the estimation of the number of segments. Only 13 such segments occurred in their data. This removed from consideration segments containing only one of the genes, and those where only chromosomal attribution, but not map position, was available in the mouse data, so in this sense they had even less data than we have used. However, they used segment length, which is ignored in our model, so the amount of information extracted from the data in the two approaches is comparable.

The analytic insights of Nadeau and Taylor and the prophetic accuracy of their estimation of the number of segments conserved between the mouse and human genomes have become increasingly relevant with the recent massive increases in the available genomic data, whether genetic maps, physical maps or complete sequences. Their work serves as a starting point for a variety of algorithmic, probabilistic, statistical and other applications of mathematical science.

Acknowldgments

Research supported by grants to DS from the Natural Sciences and Engineering Research Council (NSERC) and the Medical Research Council of Canada. DS is a Fellow, and DB was a postdoctoral fellow (1999), in the Evolutionary Biology Program of the Canadian Institute for Advanced Research.

References

Feller, W. (1965). *Introduction to Probability Theory and its Applications*, volume 1. John Wiley and Son, New York, second edition.

McLysaght, A., Seoighe, C., and Wolfe, K. H. (2000). High frequency of inversions during eukaryote gene order evolution. In this volume.

Nadeau, J. H. and Sankoff, D. (1998). The lengths of undiscovered segments in comparative maps. *Mammalian Genome*, 9:491–495.

Nadeau, J. H. and Taylor, B. A. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences (U.S.A.)*, 81:814–818.

Pennisi, E. (2000). Mouse economy: A mouse chronology. *Science*, 288:248–257.

Sankoff, D. and Ferretti, V. (1995). Karyotype distributions in a stochastic model of reciprocal translocation. *Genome Research*, 6:1–9.

Sankoff, D., Ferretti, V., and Nadeau, J. H. (1997a). Conserved segment identification. *Journal of Computational Biology*, 4:559–565.

Sankoff, D., Parent, M.-N., Marchand, I., and Ferretti, V. (1997b). On the Nadeau-Taylor theory of conserved chromosome segments. In Apostolico, A. and Hein, J., editors, *Combinatorial Pattern Matching. 8th Annual Symposium.* Lecture Notes in Computer Science 1264, pages 262–274. Springer, New York.

Schoen, D. (2000). Comparative genomics, marker density and statistical analysis of chromosome rearrangements. *Genetics*, 154:943–952.

Seldin, M. F. (1999). The Davis Human/Mouse Homology Map. http://www.ncbi.nlm.nih.gov/Homology/.

Waddington, D., Springbett, A. J., and Burt, D. W. (2000). A chromosome based model to estimate the number of conserved segments between pairs of species from comparative maps. *Genetics*, 154:323–332.

Centre de recherches mathématiques, Université de Montréal, CP 6128 succursale Centre-Ville, Montréal, Québec H3C 3J7.
E-mail: sankoff@ere.umontreal.ca, bryant@crm.umontreal.ca

Statistics Canada
E-mail: marie-noelle.parent@statcan.ca