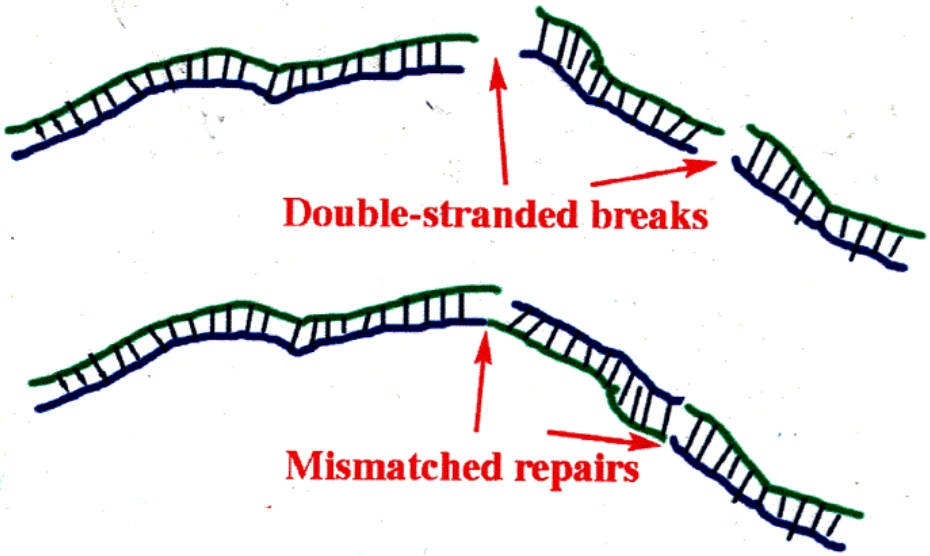
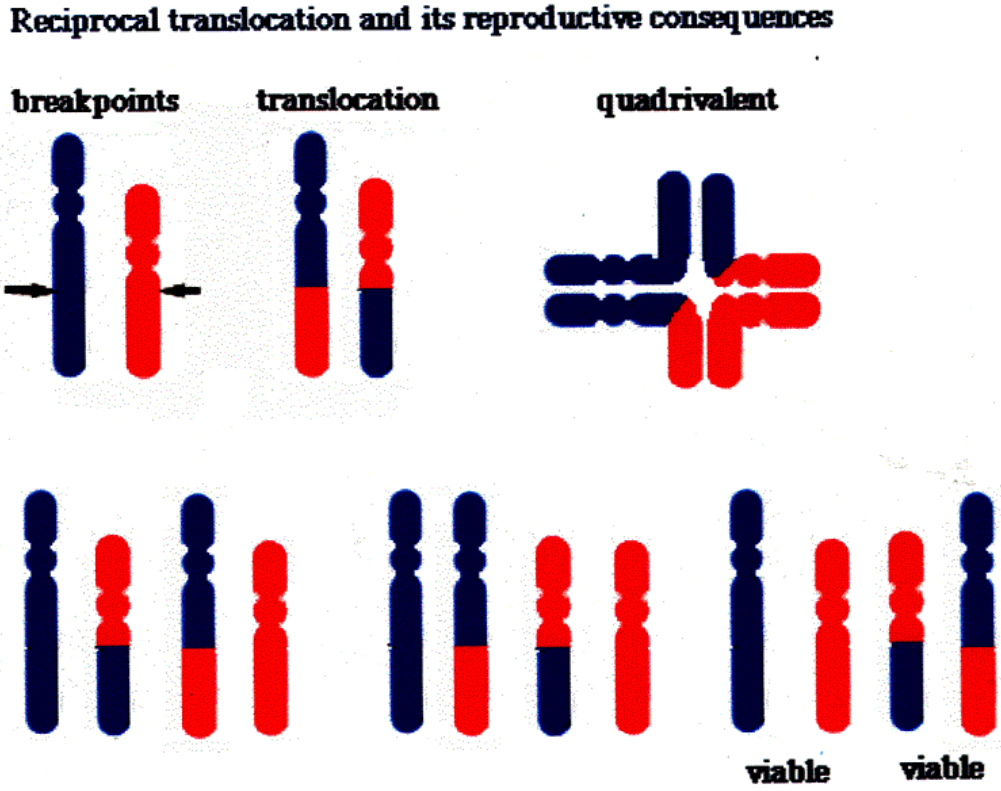
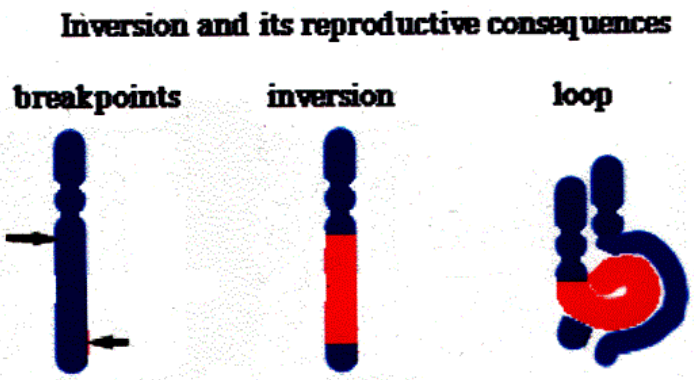
Genome rearrangement

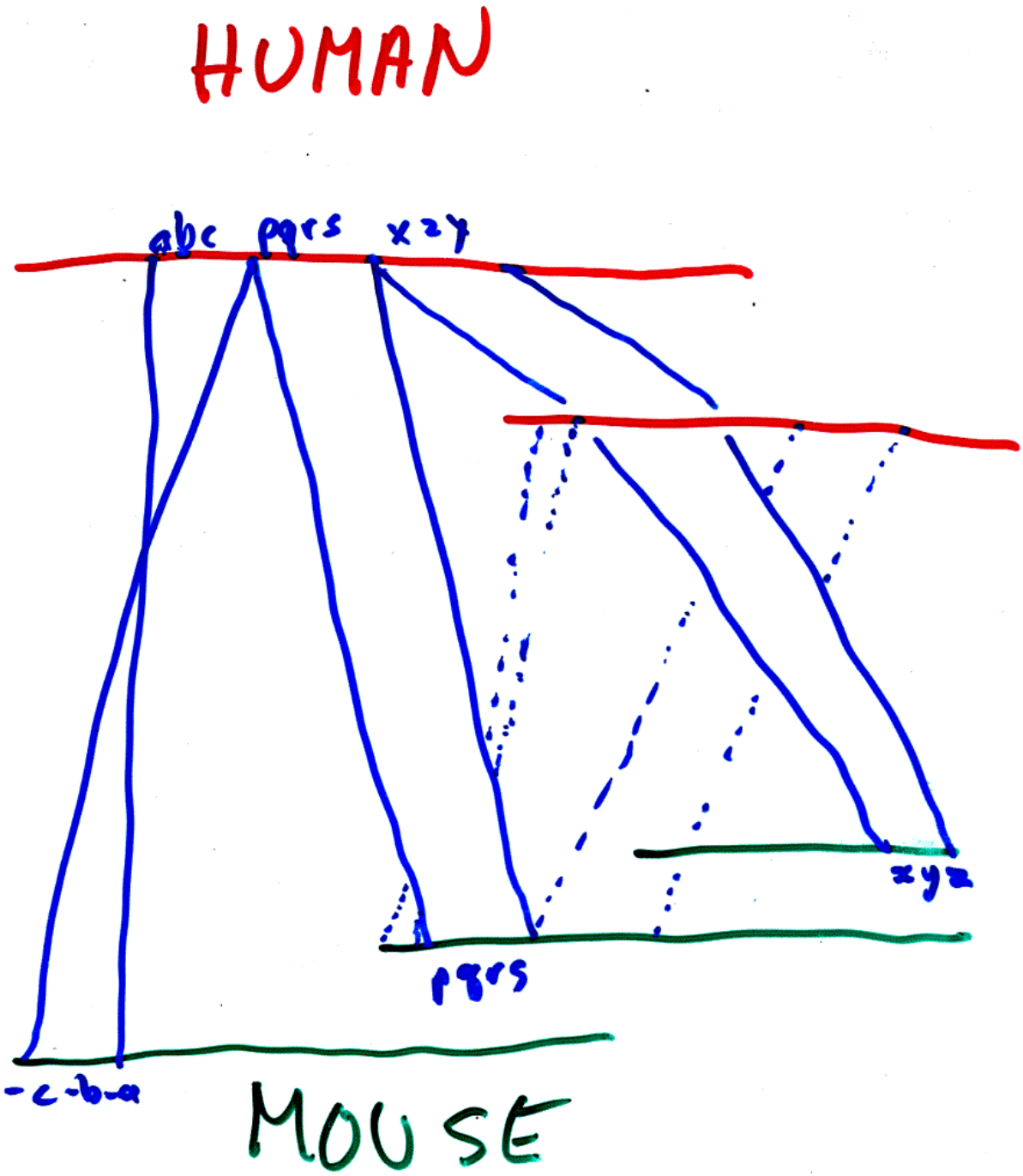
genome G: a b c d e f g h i
genome H: -e -d -c a f g h -b i

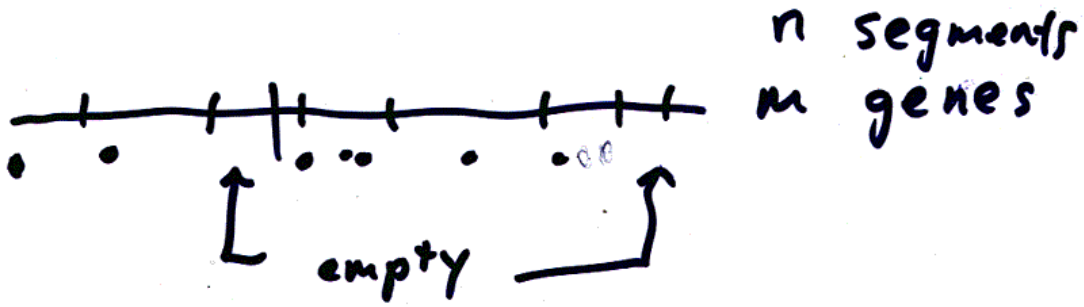
Mechanism




Reciprocal translocation and its reproductive consequences
breakpoints
translocation
quadrivalent
viable
viable


Inversion and its reproductive consequences
breakpoints
inversion
loop


HUMAN
abc pqrs xzy
xyz
pqrs
-c-b-a
MOUSE



Nadeau & Taylor 1984

"observed" 13 segments

estimated 170

$$P(a, m, n) = \frac{\binom{m-1}{a-1}\binom{n+1}{a}}{\binom{n+m}{m}}$$

$$\hat{n} \sim 170$$

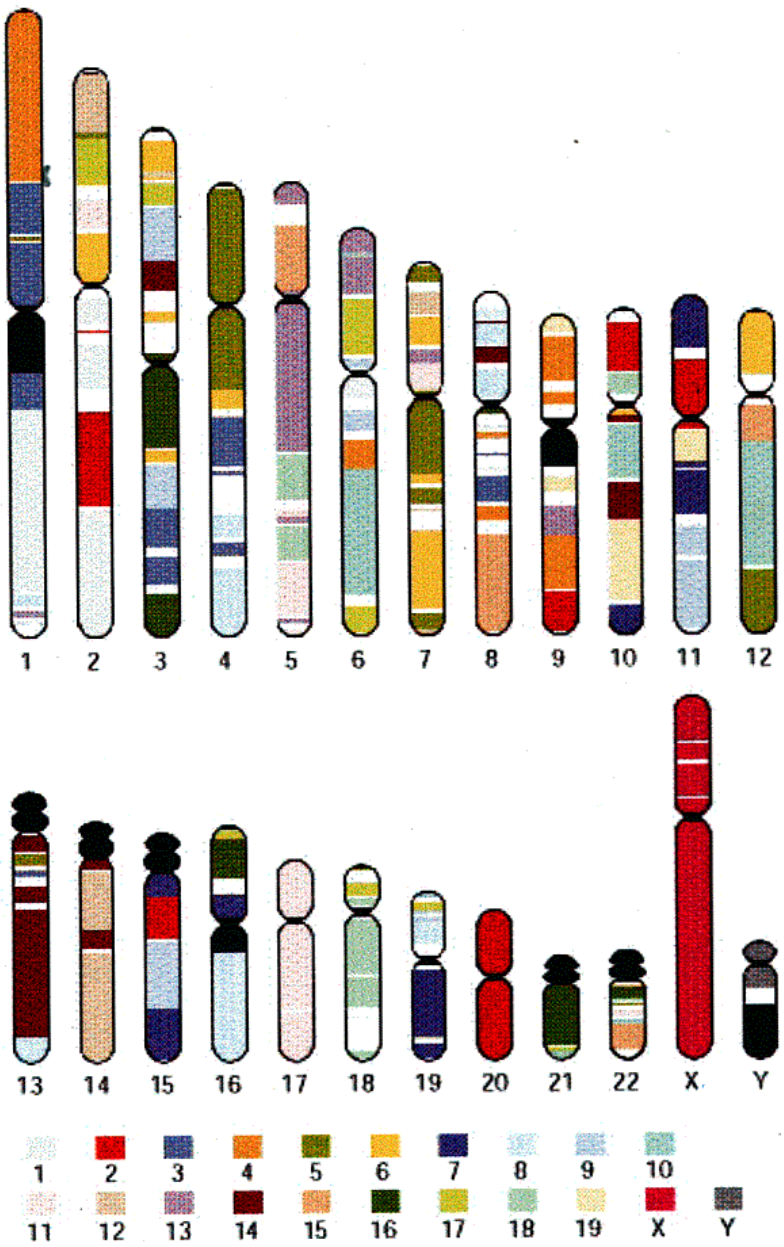


Figure 46 Conserved segments in the human and mouse genome. Human chromosomes, with segments containing at least two genes whose order is conserved in the mouse genome as colour blocks. Each colour corresponds to a particular mouse chromosome. Centromeres, subcentromeric heterochromatin of chromosomes 1, 9 and 16, and the repetitive short arms of 13, 14, 15, 21 and 22 are in black.

Mushegian, A. R. and Koonin, E. V. (1996). Gene order is not conserved in bacterial evolution. *Trends in Genetics*, 12:289–290.

BUT

Lawrence, J. G. and Roth, J. R. (1996). Selfish operons: Horizontal gene transfer may drive the evolution of gene clusters . *Genetics*, 143:1843–1860.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences USA*, 96:2896–2901.

Tamames, J., Casari, G., Ouzounis, C., and Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *Journal of Molecular Evolution*, 44.

etc

Why don't bacteria conserve gene order?

(but conserve gene clusters)

- functional constraints
- horizontal transfer
- 'selfish operons'

- ordinary neutral, null hypothetical, evolution by random mutation, inheritance and fixation

SHORT INVERSIONS

References

Andersson, S. G. E. and Eriksson, K. (2000). Dynamics of gene order stuctures and genomic architectures. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics: Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*, volume 1 of *Series in Computational Biology*, Dordrecht, NL. Kluwer Academic Press.

Dalevi, D. (2000). Bioinformatic methods for quantifying genome divergence. Master's thesis, University of Uppsala.

Dalevi, D., Eriksen, N., Eriksson, K., and Andersson, S. (2000). Genome comparison: The number of evolutionary events separating *C. pneumoniae* and *C. trachomatis*. Technical report, University of Uppsala.

Eriksen, N., Dalevi, D., Andersson, S., and Eriksson, K. (2000). Gene order rearrangeents with derange: weights and reliability. Technical report, University of Uppsala.

McLysaght, A., Seoighe, C., and Wolfe, K. H. (2000). High frequency of inversions during eukaryote gene order evolution. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map alignment and the Evolution of Gene Families*, volume 1 of *Series in Computational Biology*, Dordrecht, NL. Kluwer Academic Press.

Seoighe, C., Federspiel, N., Jones, T., Hansen, N., Bivolarovic, V., Surzycki, R., Tamse, R., Komp, C., Huizar, L., Davis, R. W., Scherer, S., Tait, E., Shaw, D. J., Harris, D., Murphy, L., Oliver, K., Taylor, K., Rajandream, M.-A., Barrell, B. G., , and Wolfe, K. H. (2000). Prevalence of small inversions in yeast gene order evolution. *PNAS*, 97:14433–14437.

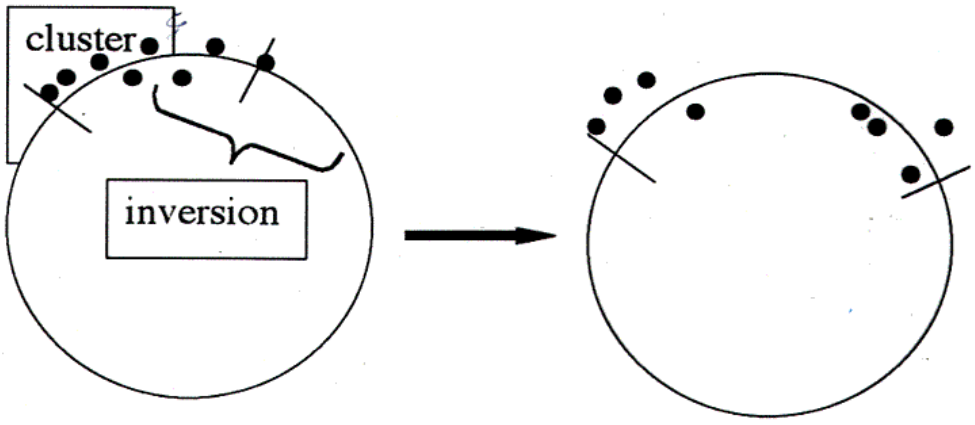
Figure 2: Effect of inversion of length x in lengthening a cluster interval from d to d'

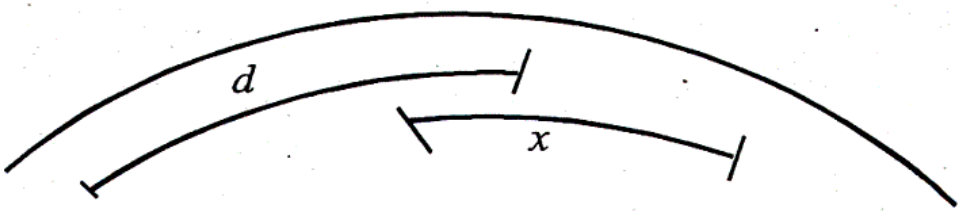
MODEL

Genome: circle of circumference 1

Rearrangement: inversion of length x

Cluster: interval of length d

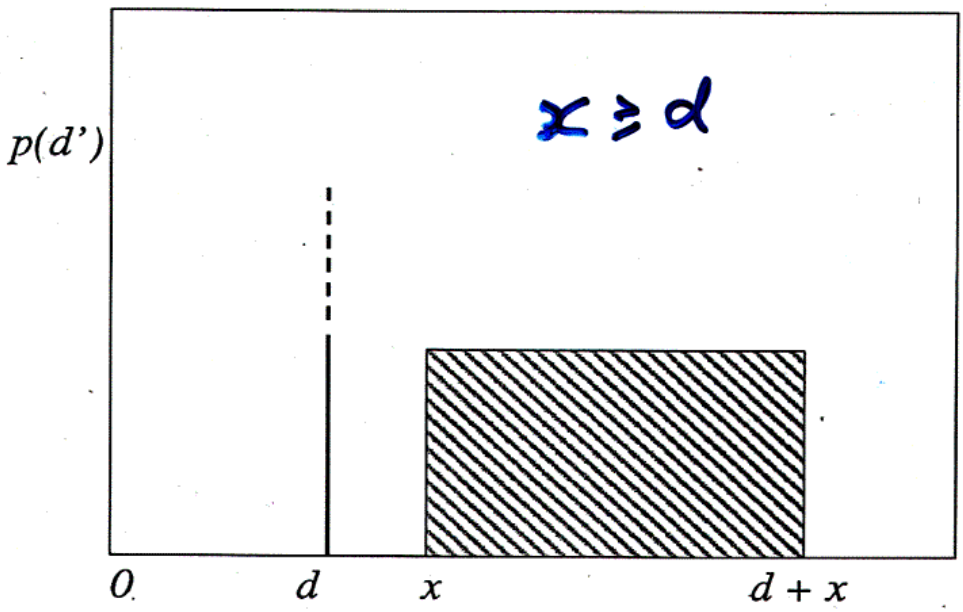


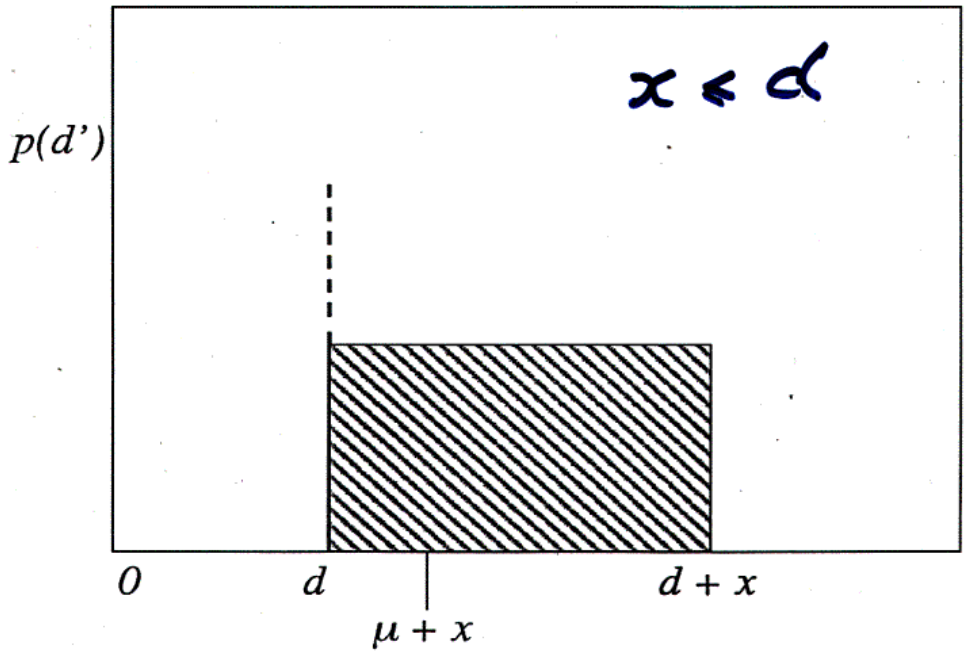


Probability that random intervals of lengths d and x intersect

$= 2 \times$ length of the shorter interval

$= 2 \min[d, x]$


p(d')
x ≥ d
0
d
x
d + x


x ≤ d
p(d')
0
d
d + x
μ + x

Distribution of the interval length d' post-inversion:

$$p(d') = (1-2x)\delta(d) + 2xU[d, d+x], \text{ for } x \leq d$$

$$p(d') = (1-2d)\delta(d) + 2dU[x, d+x], \text{ for } x \geq d$$

where $\delta(d)$ is a unit point mass at d

and $U[a,b]$ is the uniform distribution on the interval $[a,b]$.

$$\mu = d + x^2, \text{ for } x \leq d$$

$$\mu = d + 2xd - d^2, \text{ for } x \geq d$$

$d + x^2 - (x-d)^2$

$$\mu = d + x^2, \text{ for } x \leq d$$

$$\mu = d + 2xd - d^2, \text{ for } x \geq d$$

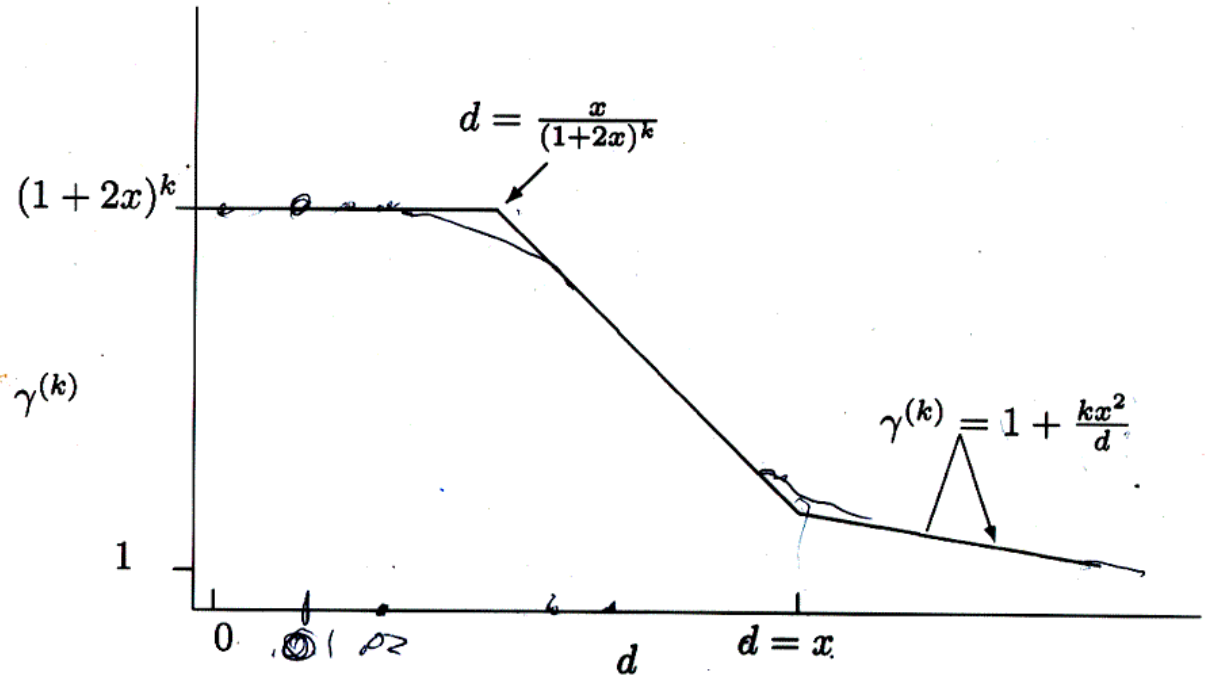
$$\mu_k = d + kx^2, \text{ for } x \leq d$$

$$\mu_k \approx d(1+2x)^k, \text{ for } x \geq d$$

The number of inversions i it would take for an interval to grow to size h in the two cases is

$$i \approx \frac{h-d}{x^2}, \text{ for } x \leq d$$

$$i \approx \frac{\log h - \log d}{\log(1+2x)}, \text{ for } x \geq d.$$



e 3: Factor of expansion after k inversions of length x as a function of initial interval length d.

$$p(d') = (1-2x)\delta(d) + 2xU[d, d+x], \text{ for } x \leq d$$

$$p(d') = (1-2d)\delta(d) + 2dU[x, d+x], \text{ for } x \geq d$$

$f(x) \quad x \geq 0$

General form:

$$p(d') = \delta(d)\,[1 - 2[F(d)\mu(f,d) + d(1-F(d))]]$$

$$+ 2[F(d') - F(d'-d)],$$

where $\delta(d)$ is a unit point mass at d,

F is the cumulative probability corresponding to the density f,

and $\mu(f,d)$ is the mean of f conditioned on $x < d$.

As part of their analysis of gene duplication in the human genome, Venter *et al.* [14] suggest that the probability of a fixed set of m genes occurring in a given order within an interval of r successive gene positions in a random genome of length n is

$$u(n,m,r) = \frac{\Sigma_{i=m-2}^{r-2} \binom{i}{m-2}}{n^{m-1}} \qquad (1)$$

(Assumes $n^{m-1} \approx (n-1)!/(n-m+1)!$)

They also offer a probability in the case of a duplicated set of m genes, "allowing for" the original set of genes "to be spread across r positions":

$$u_2(n,m,r) = \frac{\left[\Sigma_{i=m-2}^{r-2} \binom{i}{m-2}\right]^2}{n^{m-1}} \qquad (2)$$

(Wrong?)

Without gene families: One genome is $1, 2, 3, \cdots, n$ and the other is some permutation on $1, \cdots, n$. Consider any $i+1, \cdots, i+m$ in the first genome. The probability that these m genes, in any order, span exactly r slots in the second genome (i.e. first and last of the r slots contain 2 of the m genes and the remaining $r-2$ slots contain the remaining $m-2$ plus $r-m$ intruders) is:

$$p(n,m,r) = \frac{\binom{r-2}{m-2}}{\binom{n-1}{m-1}} . \qquad (3)$$

Using the identity

$$\sum_{s=m}^{n} \binom{s-2}{m-2} = \binom{n-1}{m-1},$$

the probability that these m genes span at most r slots (the *window*) in the second genome is:

$$q(n,m,r) = \frac{\binom{r-1}{m-1}}{\binom{n-1}{m-1}} . \qquad (4)$$

The corrected form of u in expression 1 is then

$$u(n, m, r) = q(n, m, r)/m! \qquad (5)$$

As n increases, Stirling's approximation shows that

$$q(n, m, r) = \left(\frac{w\theta}{e}\right)^{m-1} \theta^{-(r-\frac{1}{2})} \mathcal{O}(1), \qquad (6)$$

where $w = \frac{r-1}{n-1}$ and $\theta = 1 - \frac{m-1}{r-1}$ are two parameters introduced to represent *window proportion of genome* and *gap proportion of window*, respectively.

The expected number of such clusters of m genes to show up in the second genome, each spanning at most r slots, is $nq(n, m, r)$. There is a range of values of r for which this is non-negligible. For example, if $m = 3$, then it suffices that $r = \mathcal{O}(\sqrt{n})$.

Let H be a set of h putatively associated genes. The probability that a specific subset of H of size m appears in a genome spanning at most r slots is $q(n, m, r)$ and the expected number of such subsets is

$$Q_H(n, m, r) = \binom{h}{m} q(n, m, r). \qquad (7)$$

Of more interest is the probability $U_H(n, m, r)$ that at least one of the $\binom{h}{m}$ subsets appear. Many of the subsets intersect. To correct for this, the dominant term is due to pairs of subsets whose intersections are as large as possible: $m - 1$. Such a pair will occupy at most $2r - m + 1$ slots. Not every window of size at most $2r - m + 1$ containing $m+1$ terms of H will be the union of two windows of size at most r each containing m terms. Nevertheless, if we calculate

$$Q'_H(n, m, r) = \binom{h}{m+1} q(n, m+1, 2r-m+1), \qquad (8)$$

then $Q_H - Q'_H$ represents a first order approximation to U_H, the probability that at least one cluster of m elements from H appears.

For k genomes of same gene content the probability that a specific set of m genes appear in all these genomes spanning at most r slots is $q^k = q(n, m, r)^k$.

In the previous section, we relaxed the condition that all h genes in a set H must be represented in a cluster. Here we relax the requirement that the cluster appears in all k genomes, and ask what is the probability it appear in at least $k' \leq k$ of these genomes, spanning at most r slots in each case. The expression for this is

$$\sum_{j=k'}^{k} \binom{k}{j} q^j (1-q)^{k-j}. \qquad (9)$$

For k genomes of size $n_1, \cdots, n_k$ the probability that at least one subset of H of size $m_1, \cdots, m_k$ appears spanning at most $r_1, \cdots, r_k$ slots, respectively, is

$$\Pi_{i=1}^{k} U_H(n_i, m_i, r_i), \tag{10}$$

where each $m_i \leq r_i$. For uniform n_i, m_i and r_i, this becomes $U_H^k = U_H(n, m, r)^k$.

Combining the analyses in this section with that in the preceding one, what is the probability that subsets of H of size m appear in at least $k' \leq k$ of these genomes, spanning at most r slots in each case. Substituting U_H for q in Equation 9, we have

$$\sum_{j=k'}^{k} \binom{k}{j} U_H^j (1 - U_H)^{k-j}. \tag{11}$$

Now suppose that each gene j among the m consecutive genes in the first genome has f_j homologs in the second genome (gene family j). For each of the $\Pi_{j=1}^{m} f_j$ choices of m genes, one from each family, the probability that it spans at most r slots is $q(n,m,r)$ and the expected number of such choices is

$$S(n,m,r) = q(n,m,r)\Pi_{j=1}^{m} f_j. \qquad (12)$$

Of more interest, however, to know the probability $T(n,m,r)$ that at least one of the $\Pi_{j=1}^{m} f_j$ choices appear.

$$S'(n,m,r) = q(n,m+1,2r-m+1) \sum_{j=1}^{m} \frac{\Pi_{i=1}^{m} f_i}{f_j} \binom{f_j}{2}, \qquad (13)$$

then $S - S'$ represents a first order approximation to T, the probability that at least one cluster of m, containing one member of each family, appears.