Natural Parameter Values for Generalized Gene Adjacency

Zhenyu Yang and David Sankoff

Department of Mathematics and Statistics, University of Ottawa
{sankoff,zyang009}@uottawa.ca

Abstract. Given the gene orders in two modern genomes, it may be difficult to decide if some genes are close enough in both genomes to infer some ancestral proximity or some functional relationship. Current methods all depend on arbitrary parameters. We explore a two-parameter class of gene proximity criteria, and find natural values for these parameters. One has to do with the parameter value where the expected information contained in two genomes about each other is maximized. The other has to do with parameter values beyond which all genes are clustered. We analyse these using combinatorial and probabilistic arguments as well as simulations.

1 Introduction

As genomes of related species diverge through rearrangement mutations, groups of genes once tightly clustered on a chromosome will tend to disperse to remote locations on this chromosome or even onto other chromosomes. Even if most rearrangements are local, e.g., small inversions or transpositions, after a long enough period of time their chromosomal locations may reflect little or none of their original proximity. Given the gene orders in two modern genomes, then, it may be difficult to decide if some set of genes are close enough in both genomes to infer some ancestral proximity or some functional relationship.

There are a number of formal criteria for gene clustering in two or more organisms, giving rise to cluster detection algorithms and statistical tests for the significance of clusters. These methods, comprehensively reviewed by Hoberman and Durand [7], all depend on one or more arbitrary parameters as well as n, the number of genes in common in the two genomes. The various parameters control, in different ways, the proximity of the genes on the chromosome in order to be considered a cluster. Change the parameters and the number of clusters may change, as may the content of each cluster.

In this paper, we define a two-parameter class of gene proximity criteria, where two genes are said to be (i, j)-adjacent if they are separated by $i - 1$ genes on a chromosome in either one of the genomes and $j - 1$ genes in the other. We define a (θ, ψ) cluster in terms of a graph where the genes are vertices and edges are drawn between those (i, j)-adjacent gene pairs where $\min(i, j) < \min(\theta, \psi)$ and $\max(i, j) < \max(\theta, \psi)$. Then the connected components of the

F.D. Ciccarelli and I. Miklós (Eds.): RECOMB-CG 2009, LNBI 5817, pp. 13–23, 2009.

graph are the (θ, ψ) clusters. These definitions extend our previous notions [10,9] of generalized adjacency, which dealt only with (i, i) adjacency and (θ, θ) clusters (in the present terminology). The virtue of generalized adjacency clusters is that they embody gene order considerations within the cluster. In contrast to r-windows [3] and max-gap clusters [1,8], generalized adjacency cannot have two genes close together in one genome but far apart in the other, although the cluster could be very large.

Of particular interest are $(i, 1)$ adjacencies and $(\theta, 1)$ clusters. These are special in that they pertain to how far apart are genes in one genome that are strictly adjacent in the other genome.

As with other criteria, the quantities θ and ψ would seem arbitrary parameters in our definition of a cluster. The main goal of this paper is to remove some of this arbitrariness, by finding "natural values" for θ and ψ as a function of n, the total number of genes in the genomes. We find two such functions; the first trades off the expected number, across all pairs of genomes, of generalized adjacencies against the parameters θ and ψ, with lower parameter values considered more desirable, i.e., it is good to find a large number of generalized adjacencies, but not at the cost of including unreasonably remote adjacencies.

To do this we first define a wide class of similarities (or equivalently, distances) between two genomes in terms of weights on the (i, j)-adjacencies, namely any system of fixed-sum, symmetric, non-negative weights ω non-increasing in i and j. This is the most general way of representing decreasing weight with increasing separation of the genes on the chromosome. In any pair of genomes, we then wish to maximize the sum of the weights, which essentially maximizes the sensitivity of the criterion. Our main result is a theorem showing that the solution reduces to a uniform weight on gene separations up to a certain value of both θ and ψ, and zero weight on larger separations.

Moreover, the theorem specifies that the optimizing value of $\frac{\theta^2}{2}$ is the record time of a series of $\frac{n^2}{2}$ random variables, where n is the length of the genomes. These are not i.i.d, random variables, however, being highly dependent and, more important, of decreasing mean and variance. We use simulations to investigate the expected value of the record time under a uniform measure on the space of genomes, finding that these increase approximately as $\sqrt{n}$, in contrast with the value $\frac{n^2}{4}$ to be expected if these were i.i.d. variables. Thus it turns out that with genomes lengths of order 10^3 or 10^4, the optimal value of θ is in the range of 8 – 15.

If we are willing to accept the loss of sensitivity, and prefer to search for clusters more widely dispersed on chromosomes, there is a second set of "natural" parameter values that serve as an upper bound on the meaningful choices θ and ψ. These values are the percolation thresholds of the (θ, ψ) clusters. Beyond these values, tests of significance are no longer meaningful because all clusters rapidly coalesce together. It is no longer surprising, revealing or significant to find large groups of genes clustering together, even in pairs of random genomes.

Percolation has been studied for max-gap clusters [8], but the main analytical results on percolation pertain to completely random (Erdös-Rényi) graphs. The

graphs associated with (θ, ψ) clusters manifest delayed percolation, so the use of Erdös-Rényi percolation values would be a "safe" but conservative way of avoiding dangerously high values of the parameters. We show how to translate known results on Erdös-Rényi percolation back to generalized adjacency clusters. We also introduce random bandwidth-limited graphs and use simulations to compare the delays of generalized adjacency and bandwidth-limited percolation with respect to Erdös-Rényi percolation in order to understand what structural properties of generalized adjacency are responsible for the delay.

2 Definitions

Let S be a genome with gene set $V = \{1, \ldots, n\}$. These genes are partitioned among a number of total orders called **chromosomes**. Two genes g and h on the same chromosome are i**-adjacent** in S if there are $i-1$ genes between them in S. E.g., 1 and 4 are 2-adjacent on the chromosome 2134.

Let E_S^θ is the set of all i-adjacencies in S, where $1 \leq i \leq \theta$. We define a subset of $C \subseteq V$ to be a **generalized adjacency cluster**, or (θ, ψ) cluster, if it consists of the vertices of a connected component of the graph $G_{ST}^{\theta\psi} = (V, (E_S^\theta \cap E_T^\psi) \cup (E_S^\psi \cap E_T^\theta))$. Fig. 1 illustrates how genomes $S = 123456789$ and $T = 215783649$ determine the $(1,3)$ clusters $\{1,2\}$ and $\{3,4,5,6,7,8\}$.

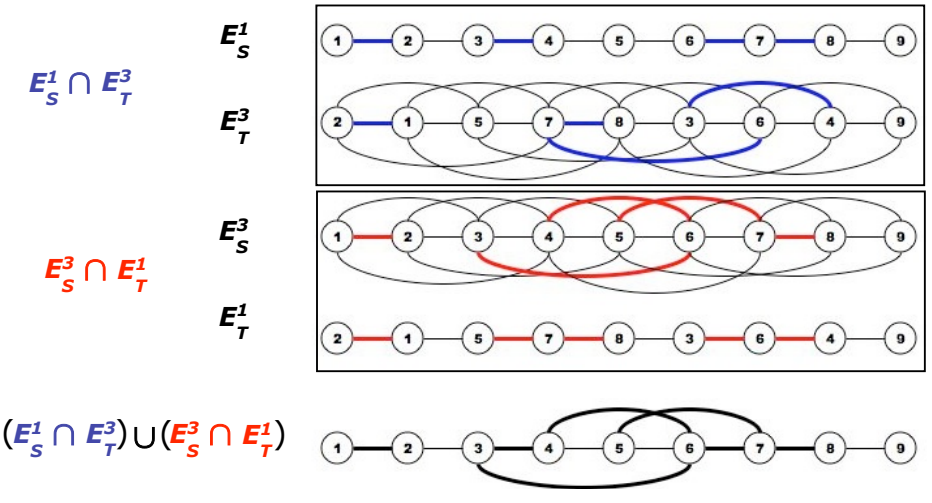

Fig. 1. Determination of $(1,3)$ clusters (or (3,1) clusters)

3 A Class of Genome Distances

Rather than looking directly for natural values of θ and ψ, we first remark that counting all (i,j)-adjacencies (with $i \leq \theta$ and $j \leq \psi$) on the same footing when

defining clusters may be giving undue weight to pairs of genes that are remote on one or both genomes, relative to genes that are directly adjacent on both. Thus we are led to consider a general system of (i, j)-dependent weights and to try to optimize this weighting, instead of just the "cut-off" values θ and ψ.

Given two genomes S and T with the same genes. Let ω_{ij} be the **weight** on two genes that are (i, j)-adjacent, i.e., i-adjacent in one of the genomes and j-adjacent in the other, such that

1. $0 \leq \omega_{ij} = \omega_{ji},\ i, j \in \{1, 2, \ldots, n-1\}$
2. $\sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \omega_{ij} = 1$
3. $\omega_{i,j} \geq \omega_{k,l}$ if
   (a) $\max(i, j) < \max(k, l)$
   (b) $\max(i, j) = \max(k, l)$ and
   $\min(i, j) < \min(k, l)$

We define the **distance** between two genomes S and T as

$$d(S,T) = 2(n-1) - \sum_{i=1}^{n-1} \left(n_{ii}\omega_{ii} + \sum_{j=1}^{n-1} n_{ij}\omega_{ij} \right). \tag{1}$$

where n_{ij} is the total number of gene pairs (x, y) that are i-adjacent in S and j-adjacent in T.

4 The Optimum Weight System Is Uniform on $\{i, j\} \leq \theta$

We wish to find a system of weights ω that tends to allocate, inasmuch as possible, higher weight to (i, j)-adjacencies with small i and j, thus emphasizing the local similarities of the two genomes, but not excluding moderate values of i and j, to allow some degree of genome shuffling.

Our strategy is to examine individual pairs of genomes (S, T) first, and that is the topic of this section. In the next section, we will study the consequences of introducing a uniform measure on the space genomes of length n.

Before stating our main results on (i, j)-adjacencies, we prove a special case. We require that every adjacency with non-zero weight be a $(1, j)$-adjacency. Then the definitions of weight and genome distance in Section 2 can be rephrased as following: Let the **weight** ω_i be any non-negative, non-increasing, function on the positive integers such that $\sum_{i=1}^{n-1} \omega_i = 1$. The weight ω induces a **distance** between genomes S and T with the same genes as follows:

$$d(S,T) = 2(n-1) - \sum_{i=1}^{n-1} (n_i^S + n_i^T)\omega_i \tag{2}$$

where n_j^X is the number of pairs of genes that are j-adjacent on genome X and 1-adjacent on the other genome.

Theorem 1. *For genomes S and T, the weight $\boldsymbol{\omega}$ that minimizes the distance* (2) *has*

$$\omega_i = \begin{cases} \frac{1}{k^*}, & \text{if } 1 \leq i \leq k^* \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where k^* maximizes the function

$$f(k) = \frac{\sum_{i=1}^{k}(n_i^S + n_i^T)}{k} \tag{4}$$

Proof. Let $n_i = n_i^S + n_i^T$. Based on Equation (2), minimizing $d(S,T)$ is equivalent to maximizing the summation

$$R = \sum_{i=1}^{n-1} n_i \omega_i \tag{5}$$

We first note that a uniform upper bound on ω_i is $\frac{1}{i}$. i.e. $0 \leq \omega_i \leq \frac{1}{i}$. This follows from the non-increasing condition on ω and its sum over all i being 1. Moreover, if $\omega_i = \frac{1}{i}$ for some value of i, then $\omega_1 = \omega_2 = \cdots = \omega_{i-1} = \omega_i = \frac{1}{i}$ and $\omega_{i+1} = \cdots = \omega_{n-1} = 0$, by the same argument.

Now we show that for any solution, i.e., an $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_{n-1})$ that maximizes equation (5), there must be one weight in $\boldsymbol{\omega}$ which attains this upper bound.

To prove this, let weights $\omega_1, \omega_2, \ldots, \omega_{n-1}$ maximize the equation (5) for given values of $n_1, n_2, \ldots, n_{n-1}$, such that $\zeta = \max R$. If all the n_i's are equal or all the ω_i are equal, the theorem holds trivially.

For all other cases, assume that there is no weight in $\boldsymbol{\omega}$ that attains its upper bound. We define the set $\mathcal{C} = \{\ i \mid \omega_i > \omega_{i+1},\ 1 \leq i \leq n-2\} \neq \varnothing$. Let $\xi = \min_{\mathcal{C}}(\min(\omega_i - \omega_{i+1}, \frac{1}{i} - \omega_i)) > 0$, by assumption. We select two weights ω_i and ω_j where $n_i \neq n_j$. Without loss of generality, we fix $i < j$ and $n_i < n_j$. Then we define

$$\zeta' = \sum_{k=1}^{i-1} n_k\omega_k + n_i(\omega_i - \xi) + \sum_{k=i+1}^{j-1} n_k\omega_k + n_j(\omega_j + \xi) + \sum_{k=j+1}^{n-1} n_k\omega_k \tag{6}$$

$$= \zeta + (n_j - n_i)\xi \tag{7}$$

$$> \zeta. \tag{8}$$

Then ζ is not the maximal value, contradicting the assumption about $\boldsymbol{\omega}$. Hence, there must exist a weight ω_i in $\boldsymbol{\omega}$ attaining its upper-bound $\frac{1}{i}$. Then the optimal weight is $\omega_1 = \omega_2 = \cdots = \omega_{i-1} = \omega_i = \frac{1}{i}$ and $\omega_{i+1} = \cdots = \omega_{n-1} = 0$.

Substituting this $\boldsymbol{\omega}$ in (5), produces the expression of form (4). So maximizing (5) is the same as maximizing (4).

Thus if we set $\theta = k^*$, we should find a large number of generalized adjacencies, but not at the cost of unreasonably increasing the number of potential adjacencies. The cut-off k^* differs widely of course according to the pair of genomes S

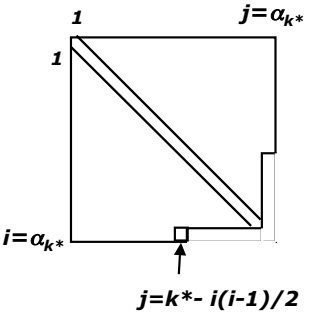

Fig. 2. k is augmented from left to right, starting at the top row, in the lower triangle including the diagonal. Values of ω_{ij} in the upper triangle determined by symmetry.

and T being compared, and this variation increases with n. However, under the uniform measure on the set of permutations, $f(k)$ does not vary much, in the statistical sense, at least as n gets large. Thus we use $\mathrm{E}[k^*]$, as function of n, to find the natural value for the cut-off parameters in the uniform weight-based distance.

Having proved the special case of $(1, j)$-adjacencies, we state the general result for (i, j)-adjacencies. The proof follows the same line as Theorem 1, but its presentation will be postponed to the full version of this paper.

Theorem 2. *Let* $\alpha_k = \lfloor \frac{\sqrt{1+8(k-1)}+1}{2} \rfloor$. *The weight* $\boldsymbol{\omega}$ *that minimizes* $d(S,T)$ *has*

$$\omega_{ij} = \begin{cases} \frac{1}{k^*}, \text{ if } i < \alpha_{k^*},\ j \le i, \\ \quad \text{or } i = \alpha_{k^*},\ j \le k^* - \frac{i(i-1)}{2} \\ 0, \quad \text{otherwise} \end{cases} \tag{9}$$

where k^* *is a natural number and maximizes the function*

$$f(k) = \frac{1}{k}\left[\sum_{i=1}^{\alpha_k - 1}\sum_{j=1}^{i}(n_{ij}+n_{ji}) + \sum_{j=1}^{k-\frac{1}{2}\alpha_k(\alpha_k-1)}(n_{\alpha_k j}+n_{j\alpha_k})\right], \tag{10}$$

where n_{ij} *is the number of gene pairs* i*-adjacent on* S *and* j*-adjacent on* T. *(See Fig. 2 for 2-dimensional area measured by* k^**.)*

In analogy to the $(1, \theta)$ clusters mentioned above, we can set $\theta = \psi = \lfloor \frac{\sqrt{1+8(k^*-1)}+1}{2} \rfloor \approx \sqrt{2k^*}$ and use $\mathrm{E}[k^*]$, as function of n, to find the natural value for the cut-off parameters in the uniform weight-based distance.

5 Finding E[k^*]

We will retain the mean of the weight systems over all random pairs of genomes (permutations of length n) as the most natural to use for studying pairs (S, T)

of experimental genomes. The justification of this approach is that it will pick up the maximum resemblance between any two genomes, even random ones, i.e., it maximizes sensitivity. Then the contribution of false positives in reducing $d(S,T)$ for real S and T may be controlled by comparing and "subtracting" the patterns derived from random genomes.

The search for k^* in Theorem 2 over all pairs of random genomes, or even a sample, may seem awkward, but it can be substantially simplified. First we can show the n_{ij} are Poisson distributed, with easily calculated parameters (proof omitted).

Theorem 3. *For two random genomes S and T with n genes the number of pairs of genes n_{ij} that are i-adjacent in S and j-adjacent in T converges to a Poisson distribution with parameter*

$$E(n_{ij}) = \frac{2(n-i)(n-j)}{n(n-1)} \tag{11}$$

Moreover, the number of gene pairs, $N(n,\theta,\psi)$, where the distances are no larger than θ in either of the genomes and no larger than ψ in the other is

$$\begin{aligned} E[N(n,\theta,\psi)] &= \sum_{i=1}^{\theta}\sum_{j=1}^{\psi} E(n_{ij}+n_{ji}) - \sum_{j=2}^{min(\theta,\psi)}\sum_{i=1}^{j-1} E(n_{ij}+n_{ji}) - \sum_{i=1}^{min(\theta,\psi)} E(n_{ii}) \\ &= (4\psi\theta - 2\theta^2) - \frac{2\theta(\psi^2-\theta^2+\psi\theta)}{n} + \frac{\theta(\theta-1)(2\psi^2-2\psi-\theta^2+\theta)}{2n(n-1)}, \end{aligned} \tag{12}$$

where $\theta \le \psi$.

The $n_{ij}+n_{ji}$ $(1 \le j \le i \le n-1)$ are asymptotically independent, so that we may profit from:

Theorem 4. *Let n_{ij} be the number of gene pairs i-adjacent on S and j-adjacent on T, then the $f(k)$ in Theorem 2 satisfies*

$$\begin{aligned} \mathrm{E}[f(k)] &\to \left(2-\frac{\alpha}{n}\right)^2 \\ \mathrm{Var}[f(k)] &\to \frac{8}{\alpha^2}(1+\frac{2}{\alpha}-\frac{2}{\alpha^2}) \end{aligned} \tag{13}$$

as $n \to \infty$, where $\alpha = \lfloor \frac{\sqrt{1+8(k-1)}+1}{2} \rfloor$.

Proof. Using the Poisson distributions in Theorem 3 in Equation 10 leads to the desired result.

Because it determines a maximum, looking for k^* is similar to the *upper record problem*, i.e., for a series of random variables $X_1, X_2, \ldots$, we consider the new sequence $L(m)$, $(m = 1, 2, \ldots)$, defined in the following manner:

$$L(1) = 1;\ L(m) = \min\{j : X_j > X_{L(m-1)}\}\ (m \ge 2) \tag{14}$$

where $L(m)$ is the index of the m^{th} upper record (or m^{th} record time), while the corresponding r.v. $X_{L(m)}$ is the value of the m^{th} record (or m^{th} record value).

Well-known properties of record times for i.i.d. random variables are:

- The probability that the i-th random variable attains a record is $\frac{1}{i}$.
- The expected number of records up to the i-th random variable is $\log i$.
- the average time at the record for n random variables is $\frac{n}{2}$.

The quantity k^* in Theorem 2 is a record time over $\frac{n(n-1)}{2}$ values of $f(k)$, though these are clearly neither identical nor independent random variables. That both the mean and variance of $f(k)$ are decreasing functions of n means that records become increasingly harder to attain.

This is illustrated in Fig. 3, which compares the proportion of record values at each (i, j) in 50,000 pairs of random genomes of size $n = 10, 30, 100, 300, 1000$, and 3000, and the accumulated number of record values up to this point, with the corresponding values of i.i.d. random variables. Note that the horizontal axis is k, which maps to $i = \sqrt{2k}$ as a position on the genome.

More important for our purposes is that the average record time is nowhere near half the number of random variables ($\frac{n^2}{4}$ in our case). Fig. 4 clearly shows that k^* is approximately $\sqrt{n}$, so that the cut-off position on the genome of the maximizing weight system will be $o(\sqrt[4]{n})$, actually about $(\sqrt{2}\sqrt[4]{n}$. For genomes of size $n = 12,000$, the expected value of k^* is around 110, so that the cut-off θ for generalized adjacency need not be greater than 15.

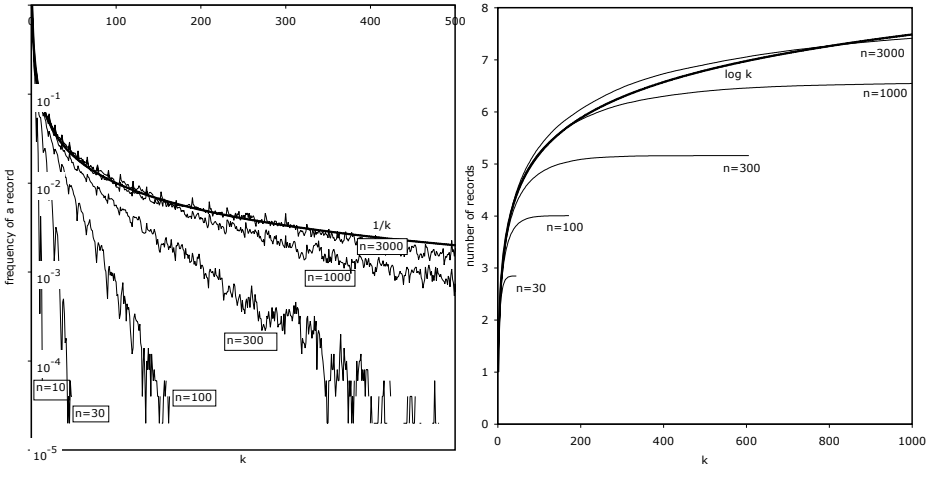

Fig. 3. Comparison of mean optimal k values, over 50,000 pairs of random genomes, with the record behaviour of i.i.d. random variables. Proportion of cases where k is optimal (left) and number of records attained (right), for (i, j) adjacencies as a function of genome size n. As $n \to \infty$, for any k', all curves approach the record time curves for all $k < k'$, but even at $n = 3000$, there is an eventual drop off, due to the declining mean expectations and variances of the n_{ij}.

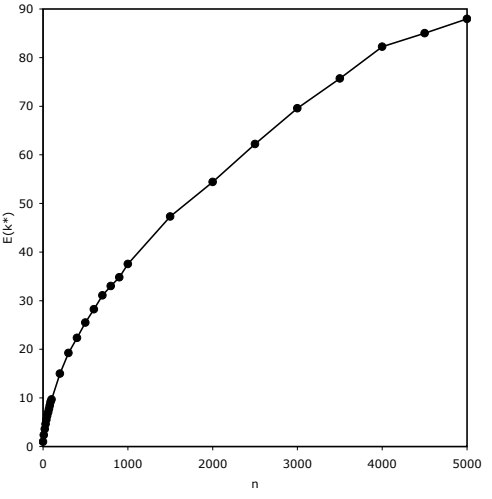

Fig. 4. Average record time as a function of genome length

6 Percolation

Clustering procedures based on parameterized adjacency criteria, e.g., as in Refs. [8,9], can have pathological behaviour as the criteria become less restrictive. At some point, called a *percolation threshold* , instead of large clusters being rare, they suddenly start to predominate and it becomes unusual *not* to find a large cluster. At this point, or earlier, it becomes meaningless to test that the numbers or sizes of clusters exceed those predicted by the null hypothesis of random genomes!

It was established by Erdös and Rényi [4,5,6] that for random graphs where edges are independently present between pairs of the n vertices with probability p, the percolation threshold is $p = \frac{1}{n}$.

When simulating the size of the largest (θ, ψ) cluster as a function of θ and ψ, we obtain graphs like that in Fig. 5.

We note that the percolation of the generalized adjacency graph is delayed considerably compared to unconstrained Erdös-Rényi graphs with the same number of edges, as may be seen in Fig. 6. To understand what aspect of the generalized adjacency graphs is responsible for this delay, we also simulated random graphs of bandwidth $\leq \theta$, since this constraint is an important property of generalized adjacency. It can be seen in Fig. 6, that the limited bandwidth graphs also show delayed percolation, but less than half that of generalized adjacency graphs.

As a control on our simulations, it is known (cf. [2]) that Erdös-Rényi graphs with rn edges, with r somewhat larger than $\frac{1}{2}$ have a cluster of size $(4r-2)n$. Our percolation criterion is that one cluster must have at least $\frac{n}{2}$ vertices. Solving this, we get $r = 0.625$. This means that the $2\theta^2$ edges we use in each of our simulated graphs must be the same as $0.625n$, suggesting that $\theta = 0.56\sqrt{n}$, compared to the $0.61\sqrt{n}$ we found in our limited simulations.

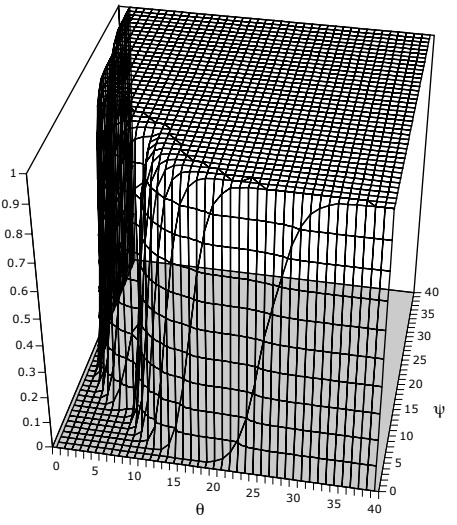

Fig. 5. Proportion of simulations where size of the largest cluster $> \frac{n}{2}$, based on a samples of $50,000$ random permutations for $\theta, \psi = 1, 2, \ldots, 99$ and genome size $n = 100$

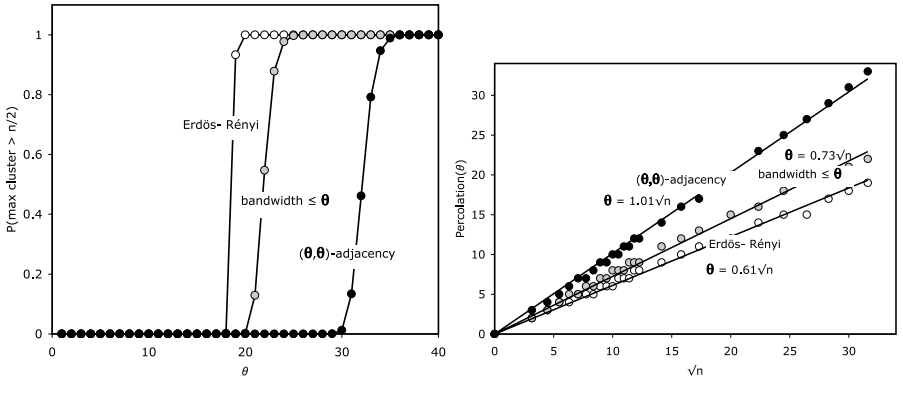

Fig. 6. (left) Simulation with genome length $n = 1000$, with $2\theta^2$ edges in each graph, showing delayed percolation of generalized adjacency graphs with respect to Erdös-Rényi graphs. Bandwidth-limited graphs are also delayed but much less so. (right) Percolation point as a function of $\sqrt{n}$, again with $2\theta^2$ edges per graph. Delay measured by coefficient of $\sqrt{n}$ in equation for trend line.

7 Conclusions

We have defined and explored the notions of generalized gene adjacency and (θ, ψ)-clusters. We have shown that not only simulations, but analytical results are quite feasible. The asymmetry of the criteria in two genomes allows a flexibility not available in previous models.

Of crucial importance is that the natural values of the cut-off we found in Section 5 are less than the percolation points in Section 6.

Future work may help locate k^* analytically as a function of n. Another direction is to pin down other structural properties, beside bandwidth constraints, responsible for the delayed percolation of generalized adjacency graphs.

Acknowledgments

Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). DS holds the Canada Research Chair in Mathematical Genomics. We would like to thank Wei Xu, Ximing Xu, Chunfang Zheng and Qian Zhu for their constant support and help in this work.

References

1. Bergeron, A., Corteel, S., Raffinot, M.: The algorithmic of gene teams. In: Guigó, R., Gusfield, D. (eds.) WABI 2002. LNCS, vol. 2452, pp. 464–476. Springer, Heidelberg (2002)
2. D'Souza, R., Achlioptas, D., Spencer, J.: Explosive percolation in random networks. Science 323, 1453–1455 (2009)
3. Durand, D., Sankoff, D.: Tests for gene clusters. Journal of Computational Biology 10, 453–482 (2003)
4. Erdös, P., Rényi, A.: On random graphs. Publicationes Mathematicae 6, 290–297 (1959)
5. Erdös, P., Rényi, A.: On the evolution of random graphs. Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5, 17–61 (1960)
6. Erdös, P., Rényi, A.: On the strength of connectedness of a random graphs. Acta Mathematica Scientia Hungary 12, 261–267 (1961)
7. Hoberman, R., Durand, D.: The incompatible desiderata of gene cluster properties. In: McLysaght, A., Huson, D.H. (eds.) RECOMB 2005. LNCS (LNBI), vol. 3678, pp. 73–87. Springer, Heidelberg (2005)
8. Hoberman, R., Sankoff, D., Durand, D.: The statistical analysis of spatially clustered genes under the maximum gap criterion. Journal of Computational Biology 12, 1081–1100 (2005)
9. Xu, X., Sankoff, D.: Tests for gene clusters satisfying the generalized adjacency criterion. In: Bazzan, A.L.C., Craven, M., Martins, N.F. (eds.) BSB 2008. LNCS (LNBI), vol. 5167, pp. 152–160. Springer, Heidelberg (2008)
10. Zhu, Q., Adam, Z., Choi, V., Sankoff, D.: Generalized gene adjacencies, graph bandwidth, and clusters in yeast evolution. Transactions on Computational Biology and Bioinformatics 6(2), 213–220 (2009)